DATA ANALYTICS OF TIME-SERIES FOR COMPLEX (BIOLOGICAL) SYSTEMS

By

Nimit Dhulekar

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

Major Subject: COMPUTER SCIENCE

Approved by the
Examining Committee:

Bülent Yener, Thesis Adviser

Charles V. Stewart, Member

Malik Magdon-Ismail, Member

Gaurav Pandey, Member

Rensselaer Polytechnic Institute
Troy, New York

April 2015
(For Graduation May 2015)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to begin by thanking my advisor Professor Bülent Yener – without his constant support and guidance this thesis would not have been possible. It has been a wonderful learning experience for me to have had the opportunity to work with him these past five years. He is truly ahead of his time, visualizing solutions to problems years before any of his peers. He has been a role model for me and many others with his professionalism, openness, caring attitude, and big heart. I would consider myself extremely fortunate if I could achieve even a modicum of his success in life.

Another person who has had a significant impact on my time at Rensselaer Polytechnic Institute is Dr. Basak Oztan. He was a great mentor, and was always available whenever I ran into trouble with a project. We spent countless hours learning altogether new techniques and models for various projects. I could not have asked for a better mentor and peer to guide me through my PhD.

I'd also like to thank my committee members Prof. Charles V. Stewart, Prof. Malik Magdon-Ismail, and Prof. Gaurav Pandey for helping me correct my mistakes, and guiding me with this thesis. I am grateful that I have had the opportunity to work with both Prof. Stewart and Prof. Pandey, and learn immensely from both of them.

This thesis is also the aggregation of the efforts of many of my lab-mates and multiple co-authors: Srinivas Nambirajan, Lauren Bange, Shayoni Ray, Cagatay Bilgin, Daniel Yuan, Abhirami Baskaran, Aritra Chowdhury, Dr. Melinda Larsen, and Dr. Gillian Beamer.

I would be remiss if I did not mention my intimate support system including James Thompson, Samara Ahmed, Megan Rogers, Dane Bush, Timothy Breen, Shoshana Rubinstein, Vasudevan Venkateshwaran, Margaret Katz, Divya Gupta, Onkar Bhardwaj, John Postl, John Licato, Beverley Parry, and David Robinson. I would not have made it through these grueling years of a graduate degree without their constant encouragement, support, and occasional motivational speeches. I owe

my successes to you – you have been my family away from home.

There are others I'd like to thank who afforded me opportunities at RPI that have allowed me to grow and mature as an individual. These amazing individuals include Gretchen Sileo, Joseph Cassidy, Martha McElligott, Cameron McLean, Joseph Campo, Erin Amarello, Kyle Keraga, and Jennifer Church. Thank you for believing in me. I would also like to thank my team-mates on "Euler's Zeroes" and its various incarnations, Trudge, Quizbowl Club, and all of the people I've met through my indulgence in pickup games around the capital district. The camaraderie that I have established with these fine folks will extend well beyond my time at RPI.

Finally, I would like to thank my family for giving me the freedom to pursue this wondrous adventure, and for instilling in me the qualities to succeed in not just graduate school, but also life. My parents' endless faith in me is why I stand here today having completed seven years of graduate school and this thesis. Thank you for making my RPI experience the most memorable period of my life.

ABSTRACT

Complex time-series systems such as biological networks have been studied for many years using conventional molecular and cellular techniques. However, the multiscale nature of these networks make these techniques limited in their application. In this thesis, we present coupled interdisciplinary algorithms covering disparate concepts such as graph-theory, level sets, autoregressive modeling, and domain knowledge transfer – with a view to improving the modeling and prediction of the evolution of biological networks. Applying our approaches to various modalities such as image-based and signal-based data, we demonstrate the importance of coupling these various techniques for a much improved holistic algorithm.

We begin by investigating cleft formation in the first round of branching morphogenesis in the mouse submandibular salivary gland. The mouse model has been well-established in biology as a precursor testing ground before human testing. By developing a model that can predict this developmental process, we would be one step closer to building realistic models for human salivary glands. Here, we present a dynamic-graph-based algorithm that not only describes tissue evolution using novel region-of-interest features but also predicts the growth of the tissue under varying concentrations of growth factors.

Sticking with the primary theme of coupling algorithmic techniques, we then combine this graph-theoretical model with level sets to create an improved model. This model takes into account cellular spatial organization and provides a better method for gland evolution. We demonstrate that this coupled cellular level set model simulates the growth of the tissue much better than other models currently in use.

Next, we tackle problems related to epileptic seizures, in particular, seizure localization and early seizure prediction. Epilepsy is a growing concern among physicians, and a concentrated effort is being made to develop algorithms for these problems. Using the notion of a seizure being a synchronous event spreading to all regions of the brain, we build synchronization graphs to quantify this neural

activity. The EEG electrodes inserted into the brain to record neural activity form the vertices of this graph, and edges are added to the graph between two vertices when they record similar neural activity. By calculating features describing these time-evolving synchronization graphs, we are able to localize the seizures temporally. A tensor-based-approach with the three modes consisting of electrodes, time, and features, is used to spatially localize the seizure to a particular side of the brain.

The other critical problem associated with epilepsy is seizure prediction, or early seizure detection. Looking at a raw EEG signal, it is not possible for physicians to identify early indicators of seizures. The problem is further complicated by the addition of two constraints – optimizing seizure prediction horizon and minimizing false positive rate. The seizure prediction horizon is the period during which predictions are made about the impending seizure. This period has to be optimized such that it gives the patient ample time to prepare for the seizure but not an excessively large period of time that would negatively disrupt the patient's life. Also, in terms of the success of the algorithm, missing a seizure completely is much worse than making a few incorrect predictions, and would make the system unusable in a real-world setup. Thus, the false positive rate has to be minimized as well. We tackle these problems by augmenting the synchronization graphs with concepts from linear algebra and machine learning. In particular, we construct an autoregressive process on the features calculated from the synchronization graphs. The autoregression coefficients are then improved on by using transfer learning and manifold alignment. We then used a one-dimensional error profile based on our prediction of the state of the system vs. the actual state of the system.

We then move on to the problem of investigating the susceptibility of heterogeneous mice populations to *Mycobacterium tuberculosis*. Varied mouse populations react in different ways to the bacterial disease. However, this differentiation is not completely obvious, and one of the goals of this project was to identify the important set of features that can perform this separation successfully. We present results on using different feature sets and multiple supervised learning classification algorithms to separate the various mouse populations.

CHAPTER 1
INTRODUCTION

1.1 Overview

At present, the tremendous quantity and diversity of molecular, genetic, cellular, and physiological data on the human body far outstrips our ability to use it for improving human health. For example, despite the abundance of molecular details known about wound healing [1], it is virtually impossible to accurately predict the final functional state of a healing wound. In a recent review [2], a 10-year retrospective analysis of wound healing assessment was performed and it was found that 29 different assessment methods are in use, all of them entirely descriptive. The primary reasoning for this disconnect between data and techniques seems to be the inability of conventional cellular and molecular biological techniques to describe complex biological phenomena; rather a more holistic approach that captures the multi-scale nature of the data is required. Thus, computational approaches have become an important tool in studying biological systems. Complex biological systems with multiple interactions that were hard to comprehend using *in vivo* experiments can now be understood using *in silico* simulations; this also reduces the time and expense of performing *in vivo* experiments. Computational models can also be constructed with incomplete information, using a relevant range of values for the various parameters.

Rcently, the theoretical computer science community has been attracted towards biological modeling because biological networks can be realized by graph-theoretical networks, where the vertices and edges are domain-dependent. A graph consists of vertices (or nodes) representing entities that share similar properties and would belong to the same set, and edges (or links) that define the level of interaction between these vertices [3]. These graphs can be utilized to analyze properties such as clustering, compactness, and topological traits. These properties of the underlying network can help uncover previously unknown information about the system, thereby making a significant contribution to the advancement of biological

knowledge.

Biological networks are inherently multiscale covering scales such as molecular, cellular, genetic, and proteomic levels. To properly model these networks requires merging of computational techniques that can work on different levels, and vertically integrating these methods. For instance, at the tissue level, it is presently very difficult to demonstrate precisely what aspects of molecular and cellular organization directly impact tissue function. Likewise, it is equally difficult to predict the functional outcome when these complex molecular systems are disrupted in injury and disease. A concrete example is the case of modeling branching morphogenesis – a multiscale computational model that jointly predicts cleft formation at the tissue level, and cell-to-cell spatial organization and cellular behavior, needs to be constructed. Thus, one of the underlying core themes of this thesis is an integrated approach which will remedy the shortcomings of existing mathematical modeling paradigms and lead to a new paradigm of hybrid modeling. More specifically, we focus on developing computational models for describing complex time-evolving systems, and in particular biological systems. We utilize various algorithmic techniques such as graph-theoretical models, level sets, and autoregressive processes at multiple scales to build comprehensive models that can accurately represent the underlying biological processes.

1.2 Focus of Thesis

The primary thread tying the different projects that form this thesis is the *analysis of time-series data*. In particular, the focus of this thesis is on the analysis of *multimodal (biological) time-series data*. We identified new *feature spaces* to better *describe* the specific biological time-series, and then performed the *prediction* of important events in these specialized feature spaces. The time-series data examined in the various projects consists of images, electrical signal recordings, and combined image and domain-specific data, respectively.

This thesis looks at developing computational techniques to study three different modalities in biological data sets, namely –

1. Image-driven data sets: We consider branching morphogenesis in the mouse

submandibular salivary gland.

2. Signal-based data sets: We investigate epileptic seizures recorded via Electroencephalograms (EEG).

3. Pseudo-time-series data: We predict the susceptibility of diversity outbred (DO) mice to *Mycobacterium tuberculosis*.

1.3 Challenges

There are multiple important challenges for which we attempt to provide solutions in this thesis. These challenges include -

1. Dissimilar time-series – The time-series data used in the five projects are quite dissimilar from each other. The first two projects use video-data split into images. The time scale for the samples in this data varies from 6 to 10 min. The third and fourth projects utilizes EEG recordings that require signal-processing techniques for analysis purposes. The samples are recorded at 400Hz, and need to be discretized into more manageable epochs. The fifth project is a pseudo-time-series, in the sense that there is no continuous data. The mice die from tuberculosis on different days based on their susceptibility. Thus, the data consists of divergent mice populations that share certain common characteristics; these characteristics are important and we are interested in studying these attributes.

2. Missing and noisy data – The image-driven and signal-driven data sets consist of both noisy and missing data. In the image data, some of the images are not particularly clean because of inherent issues with the recording microscope or human error. In the signal data, there are noisy perturbations in the data caused by patient blinking or patient motion. In the pseudo-time-series, samples arent recorded every day; they are only available for the days that the mice were euthanized. This makes it difficult to understand the changes in the mice population over time. Also, in some cases the physicians were unable to detect certain properties of the blood stream or lungs and this led to missing data.

3. New features spaces – Coming up with new features and feature spaces in which to best examine the data sets. For the image-based data, we calculated contextual biological features to better understand growth patterns of the gland. For the signal-based data, we developed an entirely new graph-based analysis system called synchronization graphs in which to examine the neuronal activity of the brain.

4. Coupling of various algorithmic approaches – Using algorithms or techniques independently usually leads to less effectual results. However, coupling interdisciplinary techniques or algorithms from different conceptual areas can lead to improvements in the results. We merged multiple algorithmic techniques in the various projects to illustrate the use of integrating divergent tools. For the image-based data, we combined dynamic graph-based models with biological knowledge and predictive analysis to build a model capable of predicting the gland development up to 9 hours into the future. We further improved on this model, by merging the graph-theoretical model with level sets, which in turn improved the topological characteristics of the growth and increased the speed of computation. For the signal-based data, we combined the technique of synchronization graphs with autoregressive processes, noise reduction techniques, transfer learning, and manifold alignment to improve on the predictive nature of the algorithm.

1.4 Contributions

We made the following contributions over the course of the five projects discussed in the following chapters:

1. From the *image-based data*, we developed a *dynamic graph-based descriptive model* that simulated the first round of branching morphogenesis with elongation of existing clefts, tissue proliferation, and *de novo* cleft formation. From this graph model, novel *local cleft and bud-based features*, and global morphological features were analyzed that successfully characterized the developmental stages of the submandibular mouse salivary gland (SMG) and simulated

temporal changes of structural properties of the tissue.

2. Given only the initial configuration of an SMG organ explant, the dynamic graph model *predicts the time-evolving development* of the SMG between embryonic days E12 and E13 as a function of initial gland morphology, mitosis and cleft progression rates, and Epidermal Growth Factor (EGF). No apriori knowledge of the target configuration is provided to the model – a significant benefit over most other similar models. Our predictive model does not require intermediate configurations, and as such can deal solves the dual issues of noisy and missing data.

3. We compared the structural characteristics of the tissue evolved with the dynamic graph model to a Monte-Carlo based on-lattice model, the Glazier-Graner-Hogeweg (GGH) model. We enabled the GGH model to work on a global tissue, allowing it to form new clefts during the simulations. Similar morphometric features were extracted from the GGH model and the trends in the features were compared with the dynamic graph-based growth model derived features. We demonstrate that our model is in a better quantitative agreement to the target configuration of the biological data as compared to the GGH model.

4. We improved on the dynamic-graph based descriptive model by merging a graph-theoretical model with *level sets*. Level sets are utilized to model the normal growth of the tissue, whereas the graph model is used to reproduce the behavior of cells within the tissue. This is a much more realistic model than using either of the two techniques individually.

5. From the *signal-based EEG data*, we developed *time-evolving synchronization graphs* as a means to capture the pair-wise correspondence between electrodes recording the EEG signals. These time-evolving synchronization graphs allow us to investigate a rich set of graph-based features that aid in determining the temporal and spatial detection of epileptic seizures. We used a clustering approach on the synchronization graph features to accurately detect seizures in the temporal domain, and a tensor-based approach to accurately detect

seizures in the spatial domain. A *noise reduction technique* in Independent Component Analysis was applied to resolve the issue of missing and noisy data.

6. We merged techniques from control theory and statistical analysis for improving our estimates of predicting seizures. We built an *autoregressive model* of order 1 (AR1) on the features mined from synchronization graphs and the features from the original signal. We then applied a *transfer learning* and *manifold alignment* approach to the AR1 models to further improve the early prediction of the seizure by utilizing data from other patients, and from the particular patient's prior recordings. This model merging is a unique addition to the community, and has not been performed earlier. Also, the transfer learning approach allows us to work with *noisy or missing* data, since we only need a few data samples to build an initial estimate of the forecast operator. Thereafter, we can utilize the transfer sets for the estimates.

7. We examined *divergent mice population* to understand the characteristic *susceptibility profiles* of each population with a view to better understanding the effect of tuberculosis. We were able to mine the data to determine the important set of features for classifying the various mice populations. At the same time, we also made interesting observations about heretofore-unknown features from the blood stream and lungs.

1.5 Thesis Outline

The following chapters detail the state of the art in the various time-series modalities, and our contributions within the specific field. In Chapter 2 , we consider the problem of describing and predicting cleft formation during the early stages of branching morphogenesis in mouse submandibular salivary glands (SMG) under the influence of epidermal growth factors (EGF). We build a descriptive model that captures the underlying biological processes via region of interest features that can characterize the growth of the tissue. We also devise a predictive growth model that simulates EGF-modulated branching morphogenesis using a dynamic graph

algorithm. We compare the prediction accuracy of the dynamic graph model to the Glazier-Graner-Hogeweg (GGH) model, an on-lattice Monte-Carlo simulation model. We demonstrate that our model yields better models of gland growth than the GGH model, at a fraction of the computational complexity. Finally, we enhanced this model to predict the SMG growth for an EGF concentration without the assistance of a ground truth time-lapse biological video data. This is a substantial benefit of our model over other similar models that are guided and terminated by information regarding the final SMG morphology.

In Chapter 3, we present a coupled cellular level sets approach for the same problem as before: describing and predicting cleft formation during the early stages of branching morphogenesis in SMG. Via this approach, we demonstrate the coupling of physics-based-continuous-modeling with discrete empirical models. We formulate a predictive model using the level-set method that simulates branching morphogenesis. This model successfully predicts the topological evolution, however, it is blind to the cellular organization, and to cell-to-cell interactions inside a gland that are available in the image data. This shortcoming of the model can be rectified by merging it with a discrete cellular model. As in chapter 2, we compare the prediction accuracy of our model to the GGH model. The results demonstrate that the coupled model yields better simulations of gland growth to that of the GGH model with a computational complexity even lower than that of the dynamic-graph-based predictive model.

In Chapter 4, we present a graph-based statistical approach for detecting epileptic seizures in EEG recordings. We utilize the pair-wise correspondence between electrodes recording EEG signals to establish edges between the electrodes which then become vertices in a *synchronization graph*. As EEG signals are recorded over time, we discretize the time axis into overlapping epochs, and build a series of time-evolving synchronization graphs for each epoch and for each traditional frequency band. We show that graph theory provides a rich set of graph features that can be used for mining and learning from the EEG signals to determine temporal and spatial localization of epileptic seizures. We present several techniques to capture the pair-wise synchronization and apply unsupervised learning algorithms,

such as k-means clustering and multi-way modeling of third-order tensors, to analyze the labeled clinical data in the feature domain to detect the onset and origin location of the seizure. We use k-means clustering on two-way feature matrices for detection of seizures, and Tucker3 tensor decomposition for localization of seizures. We demonstrate that our algorithm can temporally localize the seizure in 88.24% of the patients, and spatially localize the seizure in 76.47% of the patients.

In Chapter 5, we present a novel approach to predicting epileptic seizures by accurately modeling and predicting non-ictal cortical activity, and using prediction errors to distinguish octal from non-ictal activity. We suppress seizure-related activity by modeling EEG signal acquisition as a cocktail party problem and obtaining seizure-related activity using Independent Component Analysis. As in chapter 4, we construct dynamic EEG synchronization graphs and extract 38 intuitive features from the synchronization graph as well as the original signal. From this, we use a rigorous method of feature selection to determine minimally redundant features that can describe the non-ictal EEG signal maximally. We learn a one-step forecast operator restricted to just these features, using autoregression (AR(1)). We improve this in a novel way by cross-learning common knowledge across patients and recordings using Transfer Learning, and devise a novel transformation to increase the efficiency of transfer learning. We declare imminent seizure based on detecting outliers in our prediction errors using a simple and intuitive method. Our median seizure detection time is 11.04 minutes prior to the labeled start of the seizure compared to a benchmark of 1.25 minutes prior, based on previous work on the topic.

In Chapter 6, we present a supervised learning approach to distinguish the susceptibility of a heterogeneous mice population to *Mycobacterium tuberculosis*. Given a set of properties (such as blood and lung properties) calculated on the mice by the physicians, the goal of the project is to determine the best set of features to accurately classify the mice into 4 classes based on susceptibility. We show that six lung features – Tumor Necrosis Factor (TNF), neutrophil chemokines (CXCL1, CXCL2, CXCL5), Interferon-γ, Interleukin-12 – and two blood features – IL-2 and TNF – were identified as important by machine learning methods. Models with neutrophil chemokines generated the most accurate models, whereas those with

TNF increased the percentage of misclassifications.

In Chapter 7 we present summarized conclusions and discuss relevant future work.

CHAPTER 2
PREDICTION OF GROWTH FACTOR DEPENDENT CLEFT FORMATION DURING BRANCHING MORPHOGENESIS USING A DYNAMIC GRAPH-BASED GROWTH MODEL

2.1 Introduction

Branching morphogenesis is a developmentally conserved process occurring in many organs, including the lungs, pancreas, kidneys, salivary and mammary glands [4, 5]. Branching morphogenesis is temporally regulated highly dynamic, multiscale process involving mRNA modifications, protein signaling pathways and reciprocal interactions between epithelial and mesenchymal cell types; leading to tissue level structural changes affecting organogenesis [5]. Although the branching structures in developing organs have been studied in detail, we are still far from comprehending the integrated process.

Since, the early developmental processes in branching morphogenesis in several branched organs are conserved, we used mouse embryonic submandibular salivary gland (SMG) in our investigations [6]. The ability to produce saliva is important in maintaining oral health, and continued efforts are being targeted to identify methods to restore functionality or design artificial salivary glands. Computational modeling of the developing organ can not only add to the basic knowledge of developmental mechanisms but can also facilitate organ engineering efforts. The SMG has long been used as a biological model system to study the process of branching morphogenesis using embryonic glands that are grown ex vivo as organ explants clearly demonstrate defined branching patterns in culture [6, 7]. The embryonic SMG explants undergo branching morphogenesis when grown on filters at the air/media interface in serum-free medium in a way that reproduces the branching pattern that occurs *in vivo* [8].

This chapter has been submitted to: Dhulekar N et al. (2015) Prediction of growth factor dependent cleft formation during branching morphogenesis using a dynamic graph-based growth model. *IEEE Trans Comp Biol Bioinform* (under review)

The SMG initiates as a thickening of the primitive oral cavity epithelium on embryonic day 11 (E11). At E12 the protrusion of the primitive oral epithelium into the surrounding condensed mesenchyme forms a single cellularized epithelial primary bud on an epithelial stalk, as shown in Fig. 2.1. By E12.5 clefts, or invaginations of the basement membrane begin to form on the epithelial surface of this initial bud, as seen in Fig. 2.1(a). These clefts are stabilized before they begin to progress deeper into the gland and separate the initial bud into multiple secondary buds, as shown in Fig. 2.1(b). Epithelial proliferation occurs during cleft progression aiding in tissue growth [8]. Clefts eventually stop progressing further into the tissue and begin to widen at their base during cleft termination, as seen in the left-most cleft in Fig. 2.1(c), and ultimately transition into newly forming ducts. The gland undergoes multiple rounds of cleft and bud formation, and duct elongation throughout development and, as a result, a progressively complex and highly arborized structure is formed. Detectable epithelial cellular differentiation starts by E15; thereafter creating functional ductal structures to transport saliva and secretory acinar units capable of saliva secretion.

Branching of the salivary gland epithelial tissue is known to be dependent upon growth factors and exogenous basement membrane [9, 10, 11]. Epidermal growth factor (EGF) is one such growth factor, which is known to be involved in the morphogenesis and fetal development of several organs, including the lungs [12], kidney [13], mammary gland [14], pancreas [15] and the submandibular salivary gland (SMG) [16, 17, 18]. The role of several growth factors in SMG branching morphogenesis, including EGF, was previously investigated using mesenchyme-free epithelial rudiments cultured in a basement membrane extract in the presence of exogenously added growth factors [19, 20, 21]. Addition of EGF induced cleft formation and development of a highly lobed structure with little ductal elongation. The EGFR family displays receptor tyrosine kinase activity and ligand binding induces several downstream signaling cascades that modulate EGFR activity affecting global growth patterns in a tissue [22, 23, 24, 25]. EGF is known to activate several developmental processes including growth, survival, migration, and cell-fate determination [26]; however, its exact role in branching morphogenesis and cleft formation

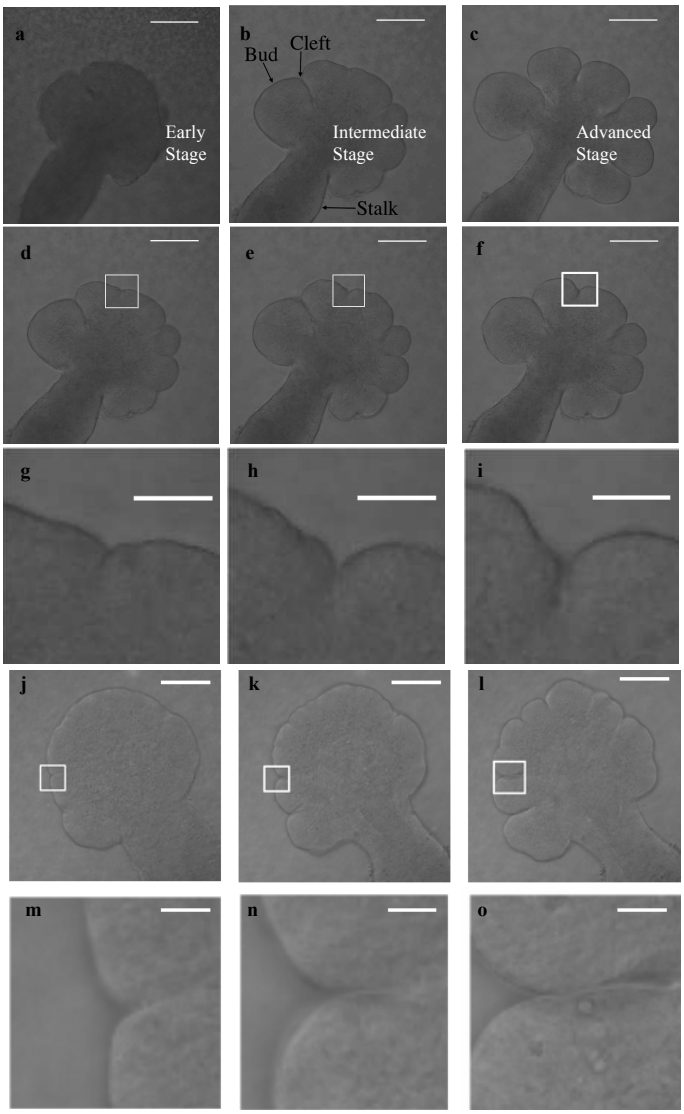


Figure 2.1: Three stages in cleft formation during early branching morphogenesis between embryonic days E12 and E13. Selected frames from the time-lapse confocal microscopy image sets demonstrate progressively deeper clefts and bud outgrowth. In (a), multiple nascent clefts are visible on a single large bud, which deepen in (b), and begin to form buds. By image (c), some clefts have terminated, and the gland has at least partially separated into distinct buds. Scale, 100 μm. Figures (d)-(o) show the progression of individual clefts in two data sets, with the upper rows ((d)-(f) and (j)-(l) Scale, 100 μm) displaying the entire gland and a white rectangle around the particular cleft, and the lower rows ((g)-(i) and (m)-(o) Scale, 50 μm) displaying the zoomed in version of this rectangle.

is not known.

Conventional cellular and molecular biological techniques are limited in their ability to describe complex biological phenomenon, and thus computational approaches have been introduced as a means to model branching morphogenesis [27]. Computational modeling of morphogenesis dates back to the mid 20th century with important mathematical models that advanced our understanding of fundamental properties of clusters of cells [28, 29]. These theories were followed by biochemical and mechanochemical models [30] that led to the use of continuum mechanics which considered a tissue to be composed of cells and extracellular matrix (ECM) and described the stress forces between the cells and the ECM [31, 32, 33]. Such models have also been used for modeling of epithelial morphogenesis in 3D breast culture acini [34], and lung [35, 36, 37] and kidney branching morphogenesis [38]. Each of these models was tailored to the particular biological process in question to account for the structurally different final branching patterns in these organs, even though mechanistic pathways are conserved across several branching organs.

The continuum mechanics models laid the foundation of utilizing computational approaches to model complex biological processes; however, they often made oversimplified assumptions regarding small deformations at the tissue-scale. Also, solutions for continuum mechanical problems have higher computational complexity, and require knowledge of the strength of bonds between cell types or the knowledge of various force fields, which are not known. Although recent studies have considered the tissue to behave as a viscous liquid under the assumption that the epithelium and mesenchyme are immiscible Stokes fluids; these models also fell short in reproducing actual salivary gland cleft shapes [39, 40, 41]. To overcome the shortcomings in these earlier modeling techniques and to better replicate the complicated dynamics governing biological processes, stochastic models were constructed based on Monte-Carlo (MC) methods. MC-based approaches can provide approximate solutions to complex, sometimes intractable mathematical problems when a large percentage of the possible configurations of the system have high energies and thus have a low probability of being attained [42]. The Glazier-Graner-Hogeweg model, an on-lattice MC model, was used to determine cellular parameters regulating cleft

progression during branching morphogenesis in the epithelial tissue of an early embryonic SMG [43]. The disadvantages of on-lattice MC-based approaches include a time-consuming sampling step to reach desired solution, potential negative effects of lattice discretization, and the use of variance-reduction techniques [44].

Over the past 20 years, graph theoretical models [45] have become significantly important in analyzing large-scale networks with complex interactions between multiple participating entities. Biological networks have also commonly benefitted from the advent of network analysis tools and techniques [46] that have been used to model protein-protein interactions [47, 48, 49, 50, 51], metabolic networks [52, 53, 54, 55], genetic and transcriptional regulatory networks [56, 57], disease progression [58, 59], and neuronal connectivity [60]. We previously developed a graph theoretical model called cell-graphs to study the structure of cellular networks [61, 62, 63, 64]. A cell-graph is an unweighted and undirected graph where the topological organization of the cells within tissues is characterized by graph theoretic features. The graph vertices (nodes) represent the cellular nuclei within the tissue and graph edges (links) capture cell-to-cell interactions. Cell-graphs enable quantification of the spatial uniformity, connectedness, and compactness at multiple scales. Conventionally, graph models have been used to depict structural properties of tissues at fixed time-points enabling characterization and quantification of the spatial evolution of tissue shape and integrity, without explicitly addressing the temporal component.

2.1.1 Our Contributions

In this study, we utilize a novel approach towards quantifying the spatio-temporal evolution of tissue shape and growth pattern using a graph-based growth model. We utilized time-lapse confocal images of SMG grown for 12 hours under the influence of varying concentrations of EGF and constant concentration of FGF. The primary contributions of this study are as follows:

- *Descriptive model*: We extracted morphometric parameters from multiple time-lapse confocal images of mesenchyme-free epithelial rudiments. A dynamic graph-based growth model was constructed that simulated the first round of branching morphogenesis with elongation of existing clefts, tissue pro-

liferation and de novo cleft formation. From this graph model, novel local cleft and bud-based features such as median cleft depth and median bud perimeter percentage, and global morphological features such as area and perimeter of the growing epithelial tissue, were analyzed for each concentration of EGF. We show that the dynamic graph-based growth model successfully characterized the developmental stages of the SMG growth pattern and simulated temporal changes of structural properties of the tissue.

- *Predictive growth model*: Given only the initial configuration of an SMG organ explant, the dynamic graph model predicts the time-evolving development of the SMG between E12 and E13 as a function of initial gland morphology, EGF concentration, and mitosis and cleft progression rates. Varying concentrations of EGF was used to adjust rates of epithelial proliferation in the model [16, 17, 18] that accounted for global tissue growth.

- *Interpolation-driven prediction modeling*: We enhanced the dynamic graph-based growth model to predict the time-evolving morphology and expected configuration of an SMG organ explant without (i) time-lapse data obtained from biological experiments, and (ii) apriori knowledge to determine the halting configuration, or cleft termination. Our approach is based on building a linear regression model of cleft deepening dependent on the perimeter of the adjoining buds and the EGF concentration; this dependency allows us to better modulate the growth of the gland. This approach is novel since it automatically calculates the target configuration to determine the halting condition. We present simulation results that demonstrate significant qualitative agreement between the target configuration predicted by our linear regression model and the biological insight.

We also show that while the dynamic graph model is nondeterministic, multiple executions of the model successfully produce deeper clefts with smaller variance across individual executions as compared to a Monte-Carlo-based simulation approach. This is partly due to the fact that it is driven by EGF concentration, and mitosis and cleft progression rates, and does not enforce assumptions regarding

bonds between cell types. These factors also contribute towards a significantly reduced computational complexity in favor of the dynamic graph-based growth model.

This study advances our earlier preliminary work on the dynamic graph-based growth model [65] in terms of extracting novel EGF concentration-dependent local epithelial features, which facilitated characterization of the growth stages of SMG development. In this study, we improve on our preliminary results from [65] in several significant ways:

1. Effect of epidermal growth factor (EGF) on the growth of the SMG – Different EGF concentration levels lead to very different growth patterns in the SMG. Also, the conference proceedings paper used a single data set, whereas this paper looks at six data sets.

2. Dynamic cleft creation – Our preliminary work was limited in that it could only model clefts that were already present in the SMG. We have improved on this preliminary result by introducing dynamic clefts in this study for both the dynamic graph-based growth model and the GGH model. Based on an algorithm that takes EGF concentration level and bud perimeter into account, we can determine when a new cleft could initiate. Please refer to step 4 "Dynamic cleft creation" of the algorithm explained in Section 2.2.10.

3. Linear regression model for cleft deepening – Previously, we used measurements from the image data to deepen clefts. We have improved on that initial approach by creating a model for cleft deepening that uses linear regression on the adjacent bud perimeters of the cleft under investigation, as well as the EGF concentration.

4. Analysis of novel local features such as cleft depth and bud perimeter – In addition to the global morphological features, we have added "local" region of interest features such as median cleft depth and median bud perimeter to our analysis. Our prior work indicates that a complete model that can accurately describe the branched pattern of the SMG requires both global and local features.

5. Interpolation-driven prediction modeling – We also introduce a new algorithm to predict the time-evolving morphology of the SMG organ explant given only its initial configuration and growth rules, without any apriori knowledge of the target configuration. This is an advance on our previous work in that it can predict unknown results. We use a linear regression model on the morphological features to predict the values of these features at a future time point and then execute our dynamic graph-based growth algorithm to reach this predicted target configuration.

2.1.2 Organization of the Paper

The rest of the paper is organized into four primary sections. Following this introduction (Section 2.1), Section 2.2 describes the biological data sets used for our experiments, preliminary image processing techniques applied on these biological data sets, algorithms for detection and development of clefts, and extraction of features including global morphological features, and local region of interest features. Section 2.3 describes the descriptive model based on dynamic graphs, the feature-based clustering of the biological data sets, comparison of the predictive growth model with a Monte-Carlo based on-lattice model, and the augmented prediction model based on dynamic graphs. Section 2.4 summarizes the findings of the study and alludes to potential future directions.

2.2 Materials and Methods

2.2.1 Data Acquisition: *Ex Vivo* Submandibular Salivary Gland Epithelial Organ Cultures

Timed-pregnant female mice (strain CD-1, Charles River Laboratories) at E12, with day of plug discovery designated as E0, were used to obtain SMG rudiments following protocols approved by the National Institute of Dental and Craniofacial Research IACUUC committee, as reported previously [19]. E12 SMGs that contain a single primary epithelial bud were microdissected, the mesenchyme was removed and the epithelial rudiments were cultured in presence of 100ng/ml FGF, as described previously [19]. For three of the glands, the media was also supplemented

with 20 ng/mL epidermal growth factor (EGF), while for the other three 1 ng/mL EGF (R&D Systems) was used. Images were collected as described in the next section 2.2.2. Henceforth, we will refer to the image sets as EGF-20a, EGF-20b and EGF-20c (20 ng/mL EGF), and EGF-1a, EGF-1b and EGF-1c (1 ng/mL EGF).

2.2.2 Data Acquisition: Confocal Time-lapse Series Acquisition

Epithelial rudiments were either labeled with a tracer molecule, Alexa Fluor 647-labeled human plasma fibronectin (FN), to identify the location of the basement membrane, or with green fluorescent protein (GFP) to label a subset of the epithelial cells, and imaged using time-lapse confocal microscopy, as described previously [19]. The glands were imaged using a 20X, 0.8 N.A. objective either with or without an additional level of optical zoom (1.7-2.5X) using a Zeiss 510 Meta confocal microscope. Images of either 7μm or 5μm thick optical sections were captured at either 6 min or 10 min intervals. In some experiments, the 488 nm laser was used to capture an additional brightfield/near Differential Interference Contrast (DIC) microscopy image. Image sets were captured at 512×512 pixel resolution using a scan speed of 8 in line averaging mode. For this study, images from the center of the explant were used for 2D analysis. The *ex vivo* SMG and confocal time-lapse series images are available at `http://dsrc.rpi.edu/cellgraph/SMG_modeling` on the Data Science Research Centers website.

2.2.3 Data Acquisition: Whole-Mount Immunocytochemistry and Confocal Imaging

Whole-mount immunocytochemistry was performed, as previously described [8, 62]. SYBR Green I (1:10000, Invitrogen) was used to detect nuclei, proliferating cells were detected using phospho-Histone H3 (pHH3) antibody (1:100, Cell Signaling Technology) and epithelium was detected using an antibody recognizing Ecadherin (1:250, BD Biosciences). SMGs were imaged on a Zeiss LSM510 confocal microscope at 20X (Plan Apo/0.75 NA), or 63X (Plan Apo/1.4 NA) magnification. To enhance the contrast of the grey-scale pHH3 images and the SYBR Green I images, we applied the contrast-limited adaptive histogram equalization algorithm (CLAHE) [66] to the images. The CLAHE algorithm considers the image as a collection of smaller

regions and applies histogram equalization on these regions. The total area of the connected components in both images is calculated, and the ratio gives us the percentage of SYBR Green I cells (total cells), which are proliferating. We also used the time-lapse images to approximate the average mitosis rate for the EGF-1 ng/mL and EGF-20 ng/mL data sets. The formula used was:

$$MR = \frac{\text{Final area of gland - Initial area of gland}}{\text{Average cell size} \times \text{Size of data set (in time)}}$$

2.2.4 Quantification of Ground Truth: Image Processing and Segmentation

The first step in characterizing the SMG morphology consists of segmenting the SMG regions in the FN (via ImageJ) and GFP (manual segmentation due to noisy images) time-lapse data sets. The FN images were segmented via Otsu's technique [67] by calculating an optimal threshold to separate the tissue (foreground) from the Matrigel medium (background). Biologists visually inspected this manual segmentation to ensure that it correctly captured the morphology. To obtain nuclear information regarding cell distribution and cell morphologies, we referred to the *ex vivo* data set (in Section 2.2.1).

Data sets including the FN label were automatically segmented in ImageJ [68]. These images were preprocessed by increasing the contrast to highlight the FN boundary. An optimal threshold value was then calculated to separate the tissue (foreground) from the Matrigel medium (background) using Otsu's technique [67]. Using Otsu's segmentation technique, the threshold value is iteratively decided by minimizing the inter class variance of the foreground and background objects. Finally, noise and outliers were removed from the images. Data sets that were labeled with GFP had low contrast and high noise, and thus had to be manually traced in GNU Image Manipulation Program (GIMP) [69]. Biologists visually inspected this manual segmentation to ensure that it correctly captured the morphology. To obtain nuclear information regarding cell distribution and cell morphologies, we referred to the *ex vivo* data set.

2.2.5 Quantification of Ground Truth: Detection of Cleft Regions

The first important step in characterizing the SMG morphology is the detection of clefts as they form and deepen. The SMG is comprised of alternating buds and clefts, where clefts are narrow valley-shaped indentations that form in the basement membrane. Figures 2.1(d)-(o) illustrate the progression stages of typical clefts from shallow nascent clefts to narrow and deep progressive clefts in the EGF-1a and EGF-20a data sets. We characterize the cleft region using cleft center defined as the deepest point of the cleft, with the walls of the cleft extending on either side of the surface normal at the cleft center, and the corresponding left and right extrema points that determine the extent of the cleft; the buds are considered to be starting beyond the points marked as cleft extrema. The cleft center and cleft extrema are illustrated in Fig 2.2(a). Automated detection of these key points is carried out as follows:

1. We identify local extrema of the gradient along the SMG boundary by detecting angular variations greater than 35° at regular intervals of 14 successive (x,y) coordinate points measured via Euclidean distance on a rectangular Cartesian grid. This interval constituted by 14 successive boundary points was found to be optimal based on the successful identification of inflection points along the boundary. Angular thresholds lower than 35° identified multiple outliers. The extrema thus identified correspond to potential cleft centers or peaks of boundary irregularities.

2. These peaks are eliminated using the signed area of the triangle formed by the candidate point, t, and two of its immediate 8-connected neighbors, t-1 and t+1, along the boundary ordered in clock-wise direction. This is obtained as $\begin{pmatrix} x_{t-1} & y_{t-1} & 1 \\ x_t & y_t & 1 \\ x_{t+1} & y_{t+1} & 1 \end{pmatrix}$, where (x_t, y_t), (x_{t-1}, y_{t-1}), and (x_{t+1}, y_{t+1}) represent the horizontal and vertical coordinates of the candidate point and its previous and next neighbors along the boundary, respectively. This expression is positive for clefts and negative for peaks.

3. After the peaks are eliminated, we identify the cleft extrema points using

the mean-squared error (MSE) between the best-fit line and SMG boundary points on either side of the potential cleft centers. As points from the curved buds are included in the best-fit line, a higher MSE is obtained in comparison to the steeper cleft walls. The algorithm progresses from the cleft center incrementally adding points on either side of the cleft center to the cleft region. When the MSE exceeds a threshold the boundary point is labeled as a cleft extrema. We set a dynamic threshold for the MSE that is computed as a function of depth of the cleft from the closest convex hull vertices obtained after fitting a convex hull around the SMG. For every cleft center detected by the algorithm, we identify the vertices lying on the convex hull to its immediate left and right. The depth is then calculated as the perpendicular distance from the cleft center to the mid-point of the line segment joining these closest vertices identified on the convex hull. We implement cleft tracking as part of the algorithm to track the progress of the cleft.

4. As a final filtering step to eliminate boundary irregularities or nascent clefts, we exploit the cleft depth and spanning angle as illustrated for a sample cleft in Fig. 2.2(a). Cleft depth is described as the shortest Euclidean distance from the cleft center to the line segment joining the two-extrema points, and the spanning angle is the angle formed by the two line segments joining the extrema points to the cleft center. Indentations that had a depth of less than 9μm and spanning angle greater than 150° were not considered since our analysis of time-lapse data indicated that such regions were boundary irregularities that might not form a stable cleft. These thresholds were decided based on discussions with biologists and measurements from empirical data. Figures 2.2(b)-(e) show original images from four data sets with detected clefts highlighted in green and their cleft centers marked in red (or maroon as in Fig. 2.2(c)).

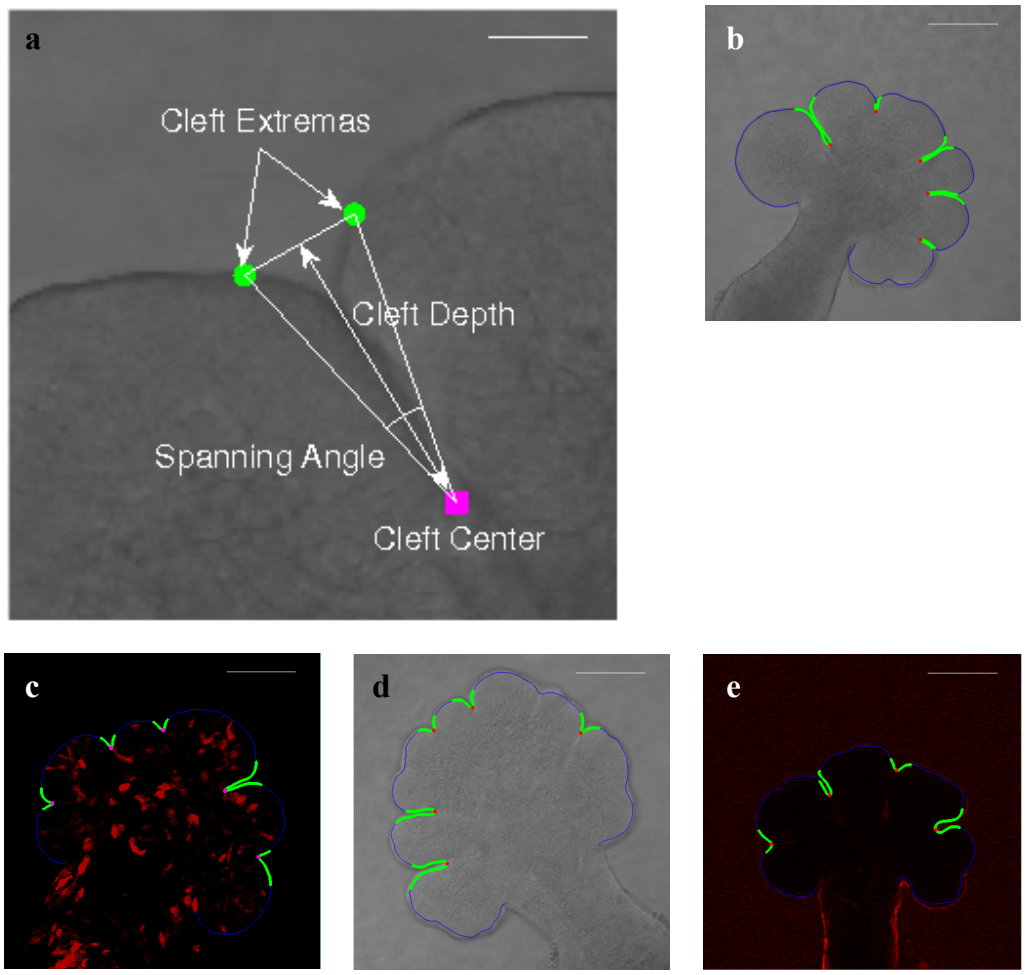


Figure 2.2: Characterization of clefts and sample results of the cleft detection algorithm. The figure shows cleft extrema (marked in green) and cleft center (marked in maroon) points that characterize the cleft in (a) (Scale, 50 μm). Spanning angle and cleft depth are calculated from these points as illustrated. In (b)-(e) (Scale, 100 μm), results of applying the cleft detection algorithm to four different data sets is shown. The cleft regions are highlighted in the DIC microscopy ((b) and (d)), GFP-labeled (c), and FN-labeled (e) images. The cleft centers are highlighted in red (or maroon in (c)) and the cleft regions are marked in green.

2.2.6 Quantification of Ground Truth: Extraction of Global SMG Morphological Features

The morphology of the SMG undergoes quantifiable transformations as a consequence of creating the ramified structure. We capture these transformations by extracting seven morphological features, namely area, perimeter, eccentricity, el-

Table 2.1: Global Morphological Feature Names and Definitions

Morphological Feature	Feature Explanation
Area	The size of the region bounded by the SMG contour.
Perimeter	The size of the region bounded by the SMG contour.
Eccentricity	Ratio of the distance between the foci of the ellipse fitted to the SMG that has the same second-moments as the SMG, and its major axis.
Elliptical Variance	Deviation of the SMG contour from an ellipse with equal covariance matrix (page 401 in [70]).
Convexity	Ratio of perimeters of the SMG convex hull to the SMG.
Solidity	Ratio of areas of the SMG to its convex hull.
Box-count dimension	The SMG is overlaid with boxes of increasing size, recording the number of boxes to cover the boundary. The box-count dimension is taken as the slope of the best-fit line in log-log space with least mean-squared error (MSE) to the vector containing box-count dimension values for square boxes of size $2^n \times 2^n$ where n varies from 0 to 9.

liptical variance, convexity, solidity, and box-count dimension. We label this data matrix of morphological features as $\mathbb{D}$. When referring to a feature matrix in subsequent text, we allude to the data matrix $\mathbb{D} \in \mathbb{R}^{M \times 7}$, consisting of the values of the seven morphological features over M time-steps. Table 2.1 lists the definitions of the various morphological features.

2.2.7 Quantification of Ground Truth: Extraction of Novel Local Cleft and Bud Features

Early branching morphogenesis is characterized primarily by bud outgrowth and cleft deepening. Early clefts tend to get deeper and narrower with time, with the end result being that the initial buds are split into multiple secondary buds. We ascertained novel local cleft and bud-based features to analyze the effects of cleft deepening on early branching morphogenesis. Using QR factorization with column pivoting, we sorted these features in accordance with their ability to capture the variance within the data [71]. This factorization is performed when the feature matrix, A, is not of full rank. QR factorization with column pivoting is given as $A = QRP^T$, where Q is an orthogonal matrix, R is an upper triangular matrix,

and P is a permutation matrix chosen such that the diagonal elements of R are non-increasing $|r_{11}| \geq |r_{22}| \ldots \geq |r_{nn}|$. The selection of features (columns) from A is based on finding the feature with the maximum Euclidean norm, and successively finding the features maximally orthogonal to the subspace spanned by the previously such determined features. The sequence of selection of features is stored in P. Other algorithms including singular value decomposition (SVD) may also be used for feature selection (please refer `http://featureselection.asu.edu/` for other feature selection techniques). The lower computational cost of QR factorization as compared to SVD was the reason we chose it as the feature selection algorithm.

2.2.8 Modeling Cleft Progression as a Function of EGF Concentration and Adjacent Bud Perimeters

We observed that although EGF stimulates branching, higher EGF concentrations produced quantitatively shallower clefts as compared to lower EGF concentrations. We thus determined that cleft depth is a function of the EGF concentration levels. We also found that cleft depth is a function of the perimeter of the adjacent buds. Larger adjacent buds allow the cleft to progress much deeper into the tissue, and higher EGF concentration levels create more buds but shallower clefts. Table 2.2 lists correlation coefficients, represented by ρ, between cleft depth and adjacent bud perimeters for three of the data sets. In all we identified 524 cleft segments, where a cleft segment is defined only for the sequence of images where its adjacent buds do not split. We collected information regarding cleft depth and adjacent bud perimeters from all the data sets to formulate cleft progression as a linear regression model with equations of the form $c = A^{-1}B$, where $c \in \mathbb{R}^{524\times1}$ is a vector of depths attained by the cleft before one of its adjacent bud splits creating a new cleft, $A \in \mathbb{R}^{524\times3}$ is the matrix of equation coefficients, and $B \in \mathbb{R}^{524\times2}$ is the matrix of adjacent bud perimeters. For our simulations, to determine the depth a cleft can achieve, we first calculate distances to every cleft in the database by comparing the adjacent bud perimeters in Euclidean space. We then apply two levels of weights to these Euclidean distances, one weight for the EGF concentration, and the other weight for each cleft in the database. For the EGF concentration, we

Table 2.2: Correlation Coefficients Between Cleft Depth and Adjacent Bud Perimeters. Each Row of the Table Lists the Pearson Correlation Coefficient Between the Cleft Depth and its Adjacent Bud Perimeters, Represented by ρ_{cd-B1} and ρ_{cd-B2}, from the Time of Creation/Identification of the Cleft till Further Splitting of the Bud.

Data Set	Cleft_ID	ρ_{cd-B1}	ρ_{cd-B2}
EGF-1a	1	0.973	0.993
	2	0.989	0.992
	3	0.972	0.931
	4	0.969	0.952
	5	0.939	0.949
EGF-20c	1	0.888	0.872
	2	0.951	0.978
	3	0.802	0.970
EGF-1b	1	0.970	0.769
	1	0.963	0.960

calculate the absolute difference of the EGF concentration from 1 ng/mL and 20 ng/mL (our two sample EGF concentrations), calculate the inverse of these differences and normalize the inverse differences by dividing by their sum. We call these weights as W1. We repeat this procedure for the distances of the bud perimeters adjacent to the cleft under investigation to the corresponding bud perimeters for each cleft in the database: calculate the inverse of the distances and normalize by dividing the inverse distances by their sum. We call these weights W2. We split the equation coefficients (B) into two groups for the two sample EGF concentrations, and weight the coefficients corresponding to each cleft in either group by the appropriate weight in W2. We then take the product of these new coefficients with W1 and sum the cleft depth values estimated by the two EGF concentrations. This gives us the final interpolated cleft depth. The simulated cleft is then assigned this cleft depth, which is updated when there is further splitting of its adjacent buds, or it exceeds the ascribed depth. We decrease the rate of growth of the cleft for higher EGF concentrations, and vice versa for lower EGF concentrations. This allows us to create shallow clefts for higher EGF concentrations, and deeper clefts for lower EGF concentrations.

2.2.9 Quantification of Ground Truth - Local Gell-Graph Features

Given a cell-graph $G(N, E)$ with N representing the set of cellular nuclei vertices (nodes) and E representing the set of edges (links) between these vertices, we extract the following features from the cell-graph.

1. Clustering Coefficient is the average of the local clustering coefficients that represents the ratio between the number of edges between the neighbors of vertex n and the total possible number of edges between the neighbors of vertex n.

$$\text{Clustering Coefficient} = \frac{\sum_{n=1}^{|N|} C_n}{|N|} , \tag{2.1}$$

where the clustering coefficient of each vertex n is

$$C_n = \frac{2|E_n|}{k_n(k_n - 1)} \tag{2.2}$$

2. Average eccentricity represents an average of the eccentricity per vertex in the graph.

$$\text{Average Eccentricity} = \frac{\sum \epsilon_n}{|N|} , \tag{2.3}$$

where eccentricity of the nth vertex ϵ_n, $n = 1, 2, \ldots, |N|$, is the maximum shortest path length from vertex n to all of the reachable vertices.

3. Closeness Centrality uses the sum of the distances of all other vertices v_j in G from a vertex v_i to quantify the centrality or importance of v_i.

$$c(v_i) = \frac{1}{\sum_j d(v_i, v_j)} \tag{2.4}$$

4. Betweenness Centrality measures how many shortest paths between all pairs of vertices include vertex v_i,

$$c(v_i) = \sum_{j \neq i} \sum_{k \neq i, k > j} \frac{n_{jk}(v_i)}{n_{jk}} , \tag{2.5}$$

where n_{jk} denotes the number of shortest paths between vertices v_j and v_k, and $n_{jk}(v_i)$ denotes the number of such paths that include v_i.

5. Edge-length statistics are computed for the spatial-length distribution in the graph in edge-threshold units. The following statistics were computed: mean, standard deviation, skewness, and kurtosis.

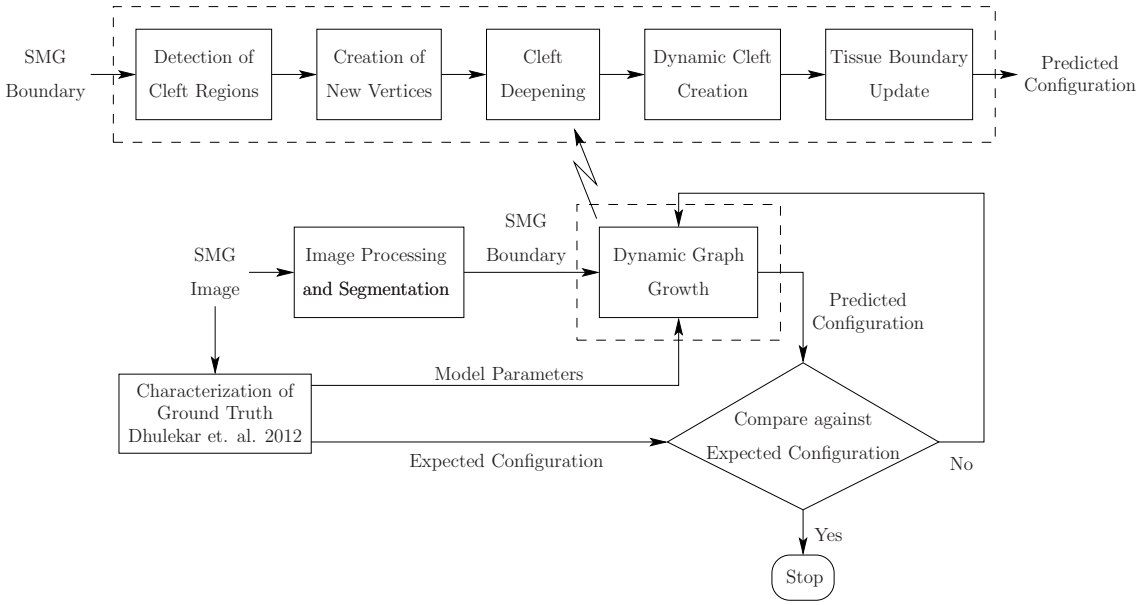
2.2.10 Predictive Dynamic Graph Growth Model: Construction of a Biologically-Data-Driven-Dynamic-Graph-Based Growth Model of Epithelial Branching

This section details one of the most important contributions of this study – the development of a dynamic graph model to describe the first round of cleft formation and progression in SMG branching morphogenesis. The dynamic graph-based growth model advances our prior work on a graph-theoretical model called cell-graphs, which was used for histopathological image analysis [72], tissue modeling [61], and characterization of the response of E13 salivary glands to 24 hours of inhibition of the Rho kinase (ROCK) signaling pathway [62, 64]. A cell-graph $G = (V, E)$ consists of a set of vertices V representing cell nuclei, and a set of edges E representing cell to cell interactions. An edge is present in the graph where the distance between two cell nuclei is less than a predetermined threshold. The model takes the initial gland morphology, nuclei locations, and EGF concentration as input. The initial image in each time-lapse image set was used for gland morphology. The EGF concentration levels determine the mitosis and cleft deepening rates. We assume cells to be circular in shape, and cell size is approximated by the diameter. We start with a uniform grid-graph where V is the intersection of the grid lines, and the grid lines themselves constitute E. The grid is continuously distorted on every iteration of the algorithm.

The outline of the dynamic graph-based growth model is given in Fig 2.3. The steps involved in the algorithm are listed below:

1. Detection of cleft regions: The first step of the model involves identifying clefts. Please refer to Section 2.2.5 for details regarding the methodology of detecting clefts.

2. Creation of new vertices: Each iteration of the growth algorithm divides the cells into two populations based on the distance from the gland boundary,



(a)

Figure 2.3: Overview of dynamic graph-based growth model. We start with acquisition of the biological data that gives us the SMG images. We then quantify the biological data by identifying clefts and computing global tissue-scale morphological features and local features. We run our dynamic graph-based growth model with the model parameters computed from the biological data as shown in the figure. Termination is based on reaching the optimal iteration that minimizes the distance to the target configuration.

namely internal (I) and periphery (P). Subsets $I' \subset I$ and $P' \subset P$ are chosen to undergo a proliferation attempt. Cells in P' that successfully undergo mitosis create new cells (or vertices) V' that are added to V. For I', we compute the shortest distance to the boundary of the gland (not including the cleft region) and find the cell in P closest to that boundary point; new cells $\hat{V}$ thus created are added to V. New edges, E' and $\hat{E}$ for periphery and internal cells respectively, are also constructed based on the distances from the new cells to existing cells in G.

$$\forall I' \subset I : mitosis(I') = \hat{V} \text{ then } \hat{V} \in V, \hat{E} \in E$$

$$\forall P' \subset P : mitosis(P') = V' \text{ then } V' \in V, E' \in E$$

We measured average mitosis rates for EGF-1 and EGF-20 concentration levels as 4 cells/minute and 6 cells/minute, respectively. Additional assumptions that build upon the Eden model [29] were imposed to model mitosis. These assumptions included cells with identical topology and growth permitted only at the gland boundary where a hypothesized "nutrient medium" provided by the mesenchyme is accessible. In our dynamic graph-based growth model, this similarity is enforced via the local structural properties of cell-graphs that maintain consistency in the topology of the SMG throughout the development stages. When first created, potential daughter vertices are placed outside the initial gland boundary in a region within 20° of the surface normal from the parent vertex at a minimum distance of one cell diameter, but less than the specified maximum edge length. Five possible candidate daughter vertices satisfying these spatial and angular constraints are chosen, and the daughter vertex with the closest local cell-graph features to the parent vertex is selected as the optimal daughter vertex. These local structural features (refer to 2.2.9) assess the spatial uniformity (clustering coefficient), connectedness (degree, closeness centrality, betweenness centrality), and compactness (edge length statistics) of the cell-graph. We distribute the extension distances to the neighbors of the parent vertex to model bud outgrowth in a local region and prevent spikes in the gland boundary. Figures 2.4(a-c) show a sample illustration of the configuration with creation of new vertices at different magnification levels. Supplementary Movie V1 shows mitosis events occurring over 3 hours (movie can be downloaded here: `http://dsrc.rpi.edu/cellgraph/SMG_modeling/Supplementary_Video_V1.mp4`). As described above, mitosis events are only allowed to occur on the SMG boundary; this was done primarily to reduce computational overload involving bookkeeping tasks.

3. Cleft deepening: To create dynamic clefts and to deepen existing clefts, we delete edges to vertices that would now lie in the cleft region (C).

$$\forall E(V_i, V_j) : V_i \in C, V_j \in C \text{ then}$$
$$E(V_i, V_j) \notin G, V_i \notin V, V_j \notin V$$

Table 2.3: Observed Cleft Depth Values as Calculated by the Cleft Detection Algorithm vs. the Cleft Depth Values Predicted by the Linear Regression Model. Cleft Depth is reported in μm.

Observed Cleft Depth	Estimated Cleft Depth from Linear Regression Model
17.475	20.296
15.878	21.514
21.808	20.333
22.097	21.935
22.443	21.603
22.550	21.597
22.903	20.931
22.505	20.919
22.654	21.626
22.229	22.732
22.094	22.218

This edge deletion effectively isolates these vertices from the rest of the graph. All edges that have vertices lying on opposite sides of a cleft, i.e. edges that go across the cleft, are also deleted. Figure 2.5 illustrates the cleft deepening algorithm. Once a cleft deepens, the cleft centers (marked in red) are removed from the cell-graph – the edges from these vertices to all other vertices are deleted. The deleted vertices are marked in brown in the second panel. Cleft progression is based on the linear regression model described previously in Eq. 2.2.8. The cleft deepening rate is modulated by the EGF concentration. For lower EGF concentrations, we use a higher cleft deepening rate, thereby producing deeper clefts, and we use a lower cleft deepening rate for higher EGF concentrations to produce shallower clefts. For a sample cleft we list the observed and estimated cleft depth values in Table 2.3, based on the growth coefficients computed by Eq. 2.2.8. We can see that the linear regression model estimates the cleft depth well verifying our observation that cleft depth is indeed a function of the adjacent bud perimeters. Restricting mitosis in the clefts as well as progressively increasing cleft depth causes the cleft to narrow and to deepen, both characteristics of progressive cleft formation.

4. Dynamic cleft creation: We first determined the smallest possible perimeter at which a bud splits, and the percent increase in the perimeter of buds before they split. These values were also found to be dependent on the EGF concentration. Since higher EGF concentrations stimulate branching morphogenesis and create more buds, we use progressively smaller increases in bud perimeter under increased EGF concentration levels. We take the mean μ and standard deviation σ of the percent increases in bud perimeter into consideration, and plug these values into a Gaussian distribution $N(\mu, \sigma^2)$ on the points along the bud boundary to determine the location of the split. We isolate the vertices that lie in the newly created cleft region from the rest of the graph by deleting edges incident to these vertices. All edges that go across the cleft, are also deleted.

5. Maintaining boundary smoothness: We use the spatial orientation of vertices to create a smoother gland boundary. The smoothness algorithm is based on the hypothesis that if daughter vertices are aligned similarly to the parent vertices, then smoothness will be maintained when the daughter vertices are integrated into the boundary. This is accomplished by minimizing the quantity $|\phi_i' - \phi_i|$, where ϕ_i is the angle $\angle p_{i-1}p_ip_{i+1}$, and ϕ_i' is the angle $\angle p_{i-1}p_i'p_{i+1}$, as shown in Fig. 2.4(d). The previous and next vertices p_{i-1} and p_{i+1}, respectively, are fixed, and the position of the daughter vertex p_i' is varied along the line segment p_ip_i'. This process is repeated from the second till the $(n-1)$th daughter vertex, keeping the first and nth daughter vertices fixed.

6. Updating the gland boundary: An interpolating cubic spline curve is used to insert the daughter vertices into the existing gland boundary. If the distance between the current and next daughter vertices is greater than a predetermined threshold, we connect the current daughter vertex to the +3 neighbor of the parent vertex along the SMG boundary.

7. Compare against expected target configuration: The simulation is run for 100 iterations, after which the optimal terminating iteration is selected via post-processing by determining the iteration number that minimizes the weighted

Euclidean distance to the target configuration. From our analysis, we found that about 6 hours of experimental data translates to running 100 iterations of the model. We compute the morphological feature vectors for all the iterations, and project this data matrix into the reduced space (x_j) by post-multiplying by the right singular vectors of the particular ground truth data set. We then compare this projected feature matrix to the projected feature vector corresponding to the target configuration of the corresponding ground truth data set (y_j) using a weighted Euclidean distance given as:

$$d_{xy} = \sqrt{\sum_{j=1}^{N} w_j (x_j - y_j)^2} \tag{2.6}$$

Weights are computed as the square of the positive singular values of the projected ground truth data set divided by the sum of the squares of all positive singular values. We look for the global minima in the distance d_{xy} values (the saddle point) and select this optimal iteration as the terminal configuration for our simulation. Table 2.4 lists biological processes and properties, and the corresponding mechanisms to handle them in the dynamic graph-based growth model.

2.3 Results

We present three different types of results in this section. First, we validate the accuracy of our representation of the ground truth. Next, we compare the prediction accuracy of the dynamic graph-based growth model with a Monte-Carlo-based simulation model. Finally, we present results for the interpolation-driven prediction modeling.

2.3.1 Time-Series Analysis of SMG Structural Features

Based on the features described in the Sections 2.2.6 and 2.2.7, we present in this section the growth trends of the respective features for two of the ground truth data sets. Figure 2.6 displays the trends for the global and local SMG morphological

Table 2.4: Biological Processes and Properties, and their Corresponding Interpretations in the Dynamic Graph-Based Growth Model, and the State-of-art Monte-Carlo-Based-Simulation Model Used for Comparison.

Biology	Dynamic Graph Model	GGH Model
Gland Structure	Graph Geometry	Effective Energy
Mitosis Rate	New Vertex Creation Rate	Mitosis
Cell-cell Adhesion	Maximum Link Length	Contact Energy
Cell Volume	Cell Diameter, or Minimum Link Length	Cell Area
Cell Surface Area	Not Included	Cell Perimeter
Cleft Deepening	Edge Deletion	Manual Specification

features for the EGF-1a and EGF-20b data sets. There are 29 images in the EGF-1a data set stretching over a period of about 3 hours. There are 47 images in the EGF-20b data set stretching over a period of about 8 hours. Area and perimeter (Figures 2.6(a) and (b)) display an increasing trend over time as the SMG matures. Eccentricity (Fig. 2.6(c)) is a measure of the circularity of the ellipse fitted to the SMG that has the same second-moments as the SMG, and quantifies the elongation of the SMG. Eccentricity increases as the SMG becomes more elongated with growth. The drop in eccentricity values for the EGF-20b data set can be attributed to the drop in perimeter values for the same range. Elliptical variance (Fig. 2.6(d)) displays an increasing trend as the clefts deepen since the error of fitting an ellipse to the SMG with deeper clefts would be higher than an SMG with shallow clefts. Convexity (Fig. 2.6(e)) displays a decreasing trend over the time-lapse images. As the clefts progress, the perimeter of the SMG increases, but there is only a minor increase in the perimeter of the convex hull. The perimeter of the convex hull is dependent on bud outgrowth, and this is a much slower process as compared to cleft progression [8]. Solidity (Fig. 2.6(f)) decreases over time since deepening of the clefts reduces the rate of growth of the SMG as compared to its convex hull. The box-count dimension (Fig. 2.6(g)) is a measure of a shape's space-filling capacity. Cleft deepening is expected to increase the box-count dimension. Although the rates of change of the features differ for the EGF-1 and EGF-20 data sets, all the global morphological features follow similar trends, i.e. either the features increase for both

data sets or the features decrease for both data sets.

From the feature analysis, we identified median cleft depth and median bud perimeter percentage as the most important local features. We considered median cleft depth since some clefts progress towards termination faster than other nascent clefts. Median bud perimeter percentage is the median of the percentage of the SMG perimeter belonging to individual buds. Figure 2.6(h) shows the trend in median cleft depth. An increasing trend is seen for this feature since the clefts deepen with time. Sudden dips in the trends indicate formation of nascent clefts. The deeper clefts that are characteristic of EGF-1 data sets are the reason a higher slope is observed in the graph for EGF-1a as compared to the slope for the EGF-20b graph. Median bud perimeter percentage drops over time as more buds are created that are smaller in perimeter than the original buds.

2.3.2 Validation of the Morphological Features

To verify whether the set of global morphological and local features described in the previous section was sufficient to capture the tissue-scale and local changes in the SMG, we attempted to cluster the EGF-1 and EGF-20 biological data sets (ground truth). The hypothesis being that the two EGF concentrations give rise to different morphological changes, and thus would cluster separately. As an illustration, consider the area of the SMG. It is known that the mitosis rate is dependent on EGF concentration [17]. A higher EGF concentration increases mitosis rate as compared to a lower mitosis rate for lower EGF concentration. This in turn implies that the area values for the target configuration of the EGF-1 data sets will differ from the area values for the target configuration of the EGF-20 data sets.

We ran QR factorization with column pivoting to determine the importance of each of the nine morphological features, seven global and two local region of interest features (see Section 2.2.7). We ran the factorization for each of the six data sets and found that no feature was consistently ranked with the least important score indicating that we needed to consider all nine features for our analysis. We ran k-means clustering [73] in the full nine-dimensional space, with k equals to 2, to separate the sets into two classes as shown in Fig. 2.7 with the three EGF-1 data

Table 2.5: Purity Measures for Evaluating the Effectiveness of Clustering

Cluster Numbers	**Purity Measures**			
	Recall / Purity	Precision	F-score	Entropy
Cluster C1	0.87	0.57	0.69	0.98
Cluster C2	0.69	0.92	0.79	0.41

sets listed first, followed by the three EGF-20 data sets. All the data sets are ordered chronologically within themselves from the first frame in the set to the last frame in that set. The EGF-1 data sets are either completely contained in cluster C2, or transition fairly early from cluster C1 to cluster C2 as they develop deeper clefts and larger buds. The EGF-20 data sets are either completely contained in cluster C1 or transition comparatively late from cluster C1 to cluster C2 exhibiting behavior similar to advanced EGF-1 data sets. The majority presence of the EGF-1 data sets in cluster C2 is proof that this cluster signifies larger but fewer buds and deeper clefts, whereas the majority presence of EGF-20 data sets in cluster C1 verifies that this cluster is characteristic of smaller but more buds and shallower clefts. Thus, we verify that our features were able to represent the morphological changes that occur in the SMG during branching morphogenesis. An interesting observation that was revealed via this analysis was that although increased EGF concentration stimulates branching morphogenesis by creating more buds, the clefts are shallower than when compared to lower EGF concentrations. Table 2.5 shows multiple clustering measures including recall and precision, F-score, and entropy to validate that the clustering is able to sufficiently distinguish the two data sets [74].

2.3.3 Performance Evaluation of the Predictive Growth Model on the basis of the Accuracy of Predicted SMG Morphology

To understand the efficacy of the predictive growth model, we compare it with a Monte-Carlo-based simulation model that works on the principle of energy minimization of the combined effective energy function (constructed as a Hamiltonian expression).

2.3.3.1 A Brief Overview of the GGH model

The Glazier-Graner-Hogeweg (GGH) model [75, 76] is built upon the energy minimization-based Ising model [77], using imposed fluctuations via a Monte Carlo (MC) approach. Previously, we performed an in-depth investigation where we constructed a local GGH-based model of epithelial cleft formation and analyzed the ranges of the cellular parameters [43]. In this study, we utilized similar values for each of the parameters in the simulation of cleft formation occurring on a tissue scale. Please refer to Appendix A for further details about implementation of the GGH model. The authors would also like to mention that they discussed the initial development of the specific GGH model parameter set with members of the Glazier Lab at Indiana University.

We compared the ability of the dynamic graph-based growth model to simulate EGF-stimulated cleft formation to that of the GGH model. To make a fair comparison between the dynamic graph-based growth model and the GGH model, we included the dynamic cleft creation module, described in the Section 2.2.10 to our implementation of the GGH model. This ability of the GGH model to generate de novo clefts is an improvement on our earlier work [65, 43].

2.3.3.2 Evaluation of the Predictive Growth Model

The morphological feature-based comparison of the dynamic graph-based growth model with a quantitative analysis of the biological data set (ground truth) and the GGH model under the specific parameter set described above (and in Appendix A and [43]) for organ explants grown in the presence of low or high levels of EGF (EGF-1a and EGF-20b datasets, respectively) is shown in Fig. 2.8. For the sake of brevity, only four features including area, perimeter, eccentricity, and median cleft depth are shown in the figure. Please refer to Figure 2.9 for further feature comparisons. Although, both models are constructed by very different modeling techniques, one is a graph-based model whereas the other minimizes a Hamiltonian formulation, our comparison is solely based on the final shape of the epithelial tissue produced by them. No comparison is done based on the outputs of the models in their original form. The ground truth trends are displayed in green, the dynamic graph-based-

growth-model's trends are displayed in red, and the GGH model's trends are displayed in blue. As observed in the ground truth, an increase in area and perimeter is seen in both models. The dynamic graph-based growth model is able to replicate the increase in area and perimeter more effectively than GGH, and usually remains faithful to the increasing trend. Eccentricity increases for the ground truth as the gland becomes more elongated. This is a trend that the dynamic graph model is also able to replicate, whereas GGH fails to reproduce the appropriate trend. The GGH models inability to properly reproduce the eccentricity trend could be attributed to the fact that it tries to acquire a circular structure because of the anisotropic growth of the model. Both models are able to model the trends in median cleft depth that are observed in the ground truth, although GGH performs better in a few cases since it can exactly specify the number of clefts, location of the clefts, and the depth of each cleft. The sudden drops in the value of the median cleft depth in the ground truth as well as the models signify creation of new clefts.

Although the range of values of elliptical variance is fairly small (10^{-2}), both models show an increasing trend for this feature. Convexity drops as the rate of growth of perimeter of the SMG increases at a faster rate than the perimeter of its convex hull. In keeping with the ground truth, both models show the appropriate decreasing trend. Solidity decreases with deepening of the clefts. While the dynamic graph-based growth model reproduces the appropriate trend, the GGH model tends to grow the gland more circular. This circular growth effects the solidity of the GGH model. Also, the clefts generated by GGH are in constant flux, appearing and disappearing from one MCS step to the next, and this may also be causing the solidity patterns to be modeled incorrectly. In general, the dynamic graph model has a higher box-count dimension than the GGH model. This could be because it tends to create more clefts than the GGH model, thereby creating more area of concavity. These plots illustrate that the dynamic graph-based growth model is able to predict the growth of the clefts during early branching morphogenesis. Figure 2.10 displays target ground truth configuration, and sample terminal configurations for both models for the EGF-1a and EGF-20b data sets. As can be noticed from the terminal configurations, the dynamic graph-based growth model is able to produce

de novo clefts and the final configuration is comparable to that produced by the GGH model.

2.3.3.3 Computational Complexity Comparison

We also compared the time it takes each model to complete one simulation, which is comparable to approximately 3 hours of experimental data. In order to compare the computational time complexities of the two models, we ran 50 experiments on a 2.4 GHz Intel Core 2 Duo processor with 4GB RAM. The dynamic graph model was on average 10.24 times faster than the GGH model, taking an average of 4.83 min $\pm$ 1.41 sec to complete the experiment, whereas GGH took 49.47 min $\pm$ 34.21 sec. The large volume of data generated by the GGH (requiring substantial post-processing and disk storage), as well as the increased complexity from considering cellular-level detail makes it unsurprising that the dynamic graph model takes less execution time.

2.3.4 Interpolation-Driven Prediction Modeling: Prediction of Growth Factor Dependent Branching Morphogenesis

For the majority of comparisons to biological data, both the time-lapse data set and the expected target configurations are available in advance. Our objective was to enhance our dynamic graph growth model such that when presented with an initial SMG boundary image and EGF concentration, the model could predict gland morphologies without the aid of a time-lapse data set, and thus no information regarding the target configuration.

We start by artificially creating ground truth morphological feature vectors from the initial SMG boundary image and EGF concentration, using as a first attempt, linear growth equations under the assumption that all features are uncorrelated. We computed two sets of average linear regression models for all features for EGF-1 and EGF-20 concentrations, given in Table 2.6. The linear growth rates follow the expected behavior as explained in the Section 2.3.1. For any intermediate EGF concentration, we interpolate between these two sets of models, and apply a normalized inverse distance function for the EGF concentration as a weighting factor (similar to the function defined in 2.2.5). With these individual feature growth

Table 2.6: Average Rates of Change of Feature Values Over Time for EGF-1 and EGF-20 Data sets.

	Average Rates of Change	
Morphological Features	EGF-1	EGF-20
Area(μm^2/min)	29.735	43.127
Perimeter(μm/min)	2.335	0.644
Eccentricity	3.068×10^{-4}	1.299×10^{-4}
Solidity	-2.404×10^{-4}	-8.338×10^{-5}
Box-count dimension	2.028×10^{-4}	2.271×10^{-5}
Elliptical Variance	2.591×10^{-4}	6.163×10^{-5}
Convexity	7.802×10^{-4}	-3.198×10^{-4}

equations, we determine the values the features would attain after a certain time interval. For our experiments, we consider time intervals of 3, 4.5, and 6 hours. Since the GGH model requires apriori knowledge of expected target configurations (i.e. final cleft depth) it was not possible to make such predictions with this model.

Using the formula mentioned earlier in the Section 2.2.4, we calculated average mitosis rates (MR) for EGF-1 and EGF-20 concentrations as 4 cells/minute and 6 cells/minute, respectively. Mitosis rates for intermediate EGF concentrations were predicted by interpolating between these two mitosis rates, weighted by a normalized inverse distance function. We also interpolated cleft deepening rates and maximum cleft depth using a linear interpolation scheme (described in Section 2.2.8), and ran our algorithm (Section 2.2.10) based on these interpolated mitosis and cleft deepening rates, and bud-splitting statistics for about 100 iterations, approximately the number of iterations required to simulate 6 hours of growth. For all iterations of the algorithm, we computed the morphological feature vector (y_j) and compared it to the expected target feature vector, or target configuration, as determined by the ground truth generated by the linear growth model (x_j). The comparison is done as following:

$$d_{xy} = \sqrt{\sum_{j=1}^{N} \left(\frac{y_j - x_j}{x_j} \right)^2}$$

The comparison was performed in the original feature space, since the rank of the

Table 2.7: Comparison of Final Configurations of the Three EGF Concentrations as Predicted by the Dynamic Graph-Based Growth Model.

	Area (μm^2)	Perimeter (μm)	Number of clefts	Median cleft depth (μm)
EGF-1	74132.43	1057.03	5	36.42
EGF-10	77473.87	1041.53	6	23.11
EGF-20	80715.32	1067.93	6	17.09

matrix of the linearly generated ground truth is 1. We look for the iteration number that minimizes the distance d_{xy}, and use this iteration number as the terminal configuration for each of the three time-intervals mentioned above.

This interpolation-driven dynamic graph growth model was used to predict gland morphology at specific time points for different EGF concentrations. Figure 2.11 shows the same starting image grown under EGF-1, EGF-10, and EGF-20 concentrations with predicted outcomes using the dynamic graph-based growth model for 3, 4.5, and 6 hours. Table 2.7 lists the area, perimeter, number of clefts, and median cleft depth for the three configurations. We can observe that higher EGF concentrations stimulate branching morphogenesis by creating more buds. As time progresses, the lower EGF concentrations have deeper clefts. This can be observed at 6 hours where EGF-1 has a much higher median cleft depth, whereas EGF-20 has relatively shallower clefts that have hardly progressed beyond 3 hours. Supplementary Movie V2 shows a simulation time-lapse movie produced by the dynamic graph model of all three EGF concentrations together (movie can be downloaded from `http://dsrc.rpi.edu/cellgraph/SMG_modeling/Supplementary_Video_V2.mp4`).

2.4 Discussion

The objective of this study was to quantify and predict the core processes involved in the initial stages of branching morphogenesis to initiate the process of branching morphogenesis: cleft initiation, stabilization, and progression, under different concentrations of EGF. These cleft stages occur simultaneously with bud outgrowth and subsequently lead to cleft termination and duct formation. For

this purpose, we extracted morphometric parameters from time-lapse mouse submandibular salivary gland (SMG) images. We developed a biological data driven descriptive model utilizing dynamic graphs. This dynamic graph model describes and predicts early branching morphogenesis in SMG. Given an initial SMG boundary image and EGF concentration level, the dynamic graph model is able to predict the growth of the SMG between embryonic days E12 and E13. The model probabilistically adds daughter cells and integrates these cells into the SMG by appropriately expanding its boundary. The model also creates new edges between the daughter cells and cells existing in the graph representing the initial gland. This augmented cell-graph (with the daughter cells) maintains the local structural properties of the original cell-graph. Cleft deepening and creation of dynamic clefts are crucial components of the model allowing it to produce more realistic branched structures and deeper clefts, and are based on the rules captured from the time-lapse data sets. The process of *de novo* cleft creation is modeled by first identifying the regions of initiation of the cleft, as well as the increase required in the perimeter of the buds. Given this information, we then use a probabilistic model to create *de novo* clefts.

Our results indicate that the dynamic graph model can correctly capture and represent the tissue-level morphological changes during cleft formation in the developmental stages of the SMG branching morphogenesis. We also showed that cleft progression is linearly dependent on the perimeters of the adjacent buds and is modulated by the EGF concentration. As was expected, higher EGF stimulated branching morphogenesis by producing more clefts, whereas lower EGF concentrations produced fewer clefts. Our analysis revealed an interesting observation regarding the depth of the clefts. Higher EGF concentrations produced shallower clefts, whereas lower EGF concentrations produced deeper clefts, for reasons that remain unclear. Future studies will be required to understand the cellular processes activated by EGF during cleft formation. Since this dynamic graph modeling approach can model cleft formation in response to modulation of EGF signaling, this approach could be employed to evaluate the contribution of other signaling pathways to cleft formation. Similar modeling approaches could be employed towards understanding other developmental processes in which large changes in shape occur.

Interestingly, the morphology of the epithelial buds tends to be slightly circular than the buds produced by the dynamic graph-based growth model. It is not possible to capture this difference in the circularity of buds with the methods used here and would require a more complicated model. Currently, this level of detail in shape modeling is beyond the consideration of the dynamic graph-based growth model, and would be a potential direction of enhancement for the model. We compared our results against a well-known on-lattice Monte-Carlo-based simulation model, the Glazier-Graner-Hogeweg (GGH) model, under a specific parameter set consisting of energy functions that have biologically relevant equivalents, and demonstrated that our results are in a similar quantitative agreement with the biological data as those of the GGH model, but converge significantly faster to the target configuration. The authors would like to point out that the GGH model handles cellular-level changes at a higher resolution than the dynamic graph-based growth model. We also presented a method to introduce *de novo* clefts in the SMG using the GGH model thus adding to the dynamic nature of the GGH model.

We enhanced the dynamic graph-based growth model to predict the growth of the SMG at any specified time between embryonic days E12 and E13 without requiring a time-lapse data set. This is one of the primary benefits of the interpolation-driven prediction modeling approach – it does not require apriori knowledge of the target configuration – the initial configuration and growth rules determined from the biological data are sufficient for the algorithm to predict the gland morphology at a future time point. The predictive nature of the model reduces its dependence on *in vivo* experiments, allowing the biologists to view a simulation of the experiment prior to performing it. Most other computational biological techniques, including the GGH model, require information regarding the final configuration of the ground truth data in advance. Thus, it was not possible to compare our predictive model to other simulative models. We examined the growth trends in the biological data from the viewpoint of morphological features and discovered that individual linear growth models are able to predict the evolution of each feature. This allowed us to identify the expected configuration of the morphological features at different time points.

While analyzing the results of the dynamic-graph growth model, we observed that it has certain shortcomings. Since individual tissue-scale and cellular-scale models have their deficiencies with regards to modeling different aspects of cleft formation in branching morphogenesis, future efforts should be aimed at creating a multi-scale hybrid model that can achieve better tissue-scale modeling, and be able to more realistically capture cellular-level events such as mitosis and cellular reorganization in cleft regions. The dynamic graph model is a generic model and as such can be used in conjunction with other models to create this hybrid model. Another future direction for the dynamic graph-based growth model involves including dynamic cell movement information for accurate construction of cell graphs with a better estimation of the spatial distribution of cells [19].

Acknowledgments

The authors would like to thank Dr. James A. Glazier and his research group members Dr. Abbas Shirinifard and Dr. Srividhya Jeyaraman at Indiana University, Bloomington, IN for discussions, suggestions, and help with the use of GGH model for SMG branching morphogenesis. The authors would also like to thank former research group member, Lauren Bange, for her efforts on the initial understanding and implementation of the GGH model. We also thank Dr. Kenneth Yamada for use of the Zeiss 510 Meta confocal microscope for time-lapse imaging. This work was supported by a grant from the NIH to Melinda Larsen and Bülent Yener (R01 DE019244) and by NIH C06 RR015464 to University at Albany, SUNY.

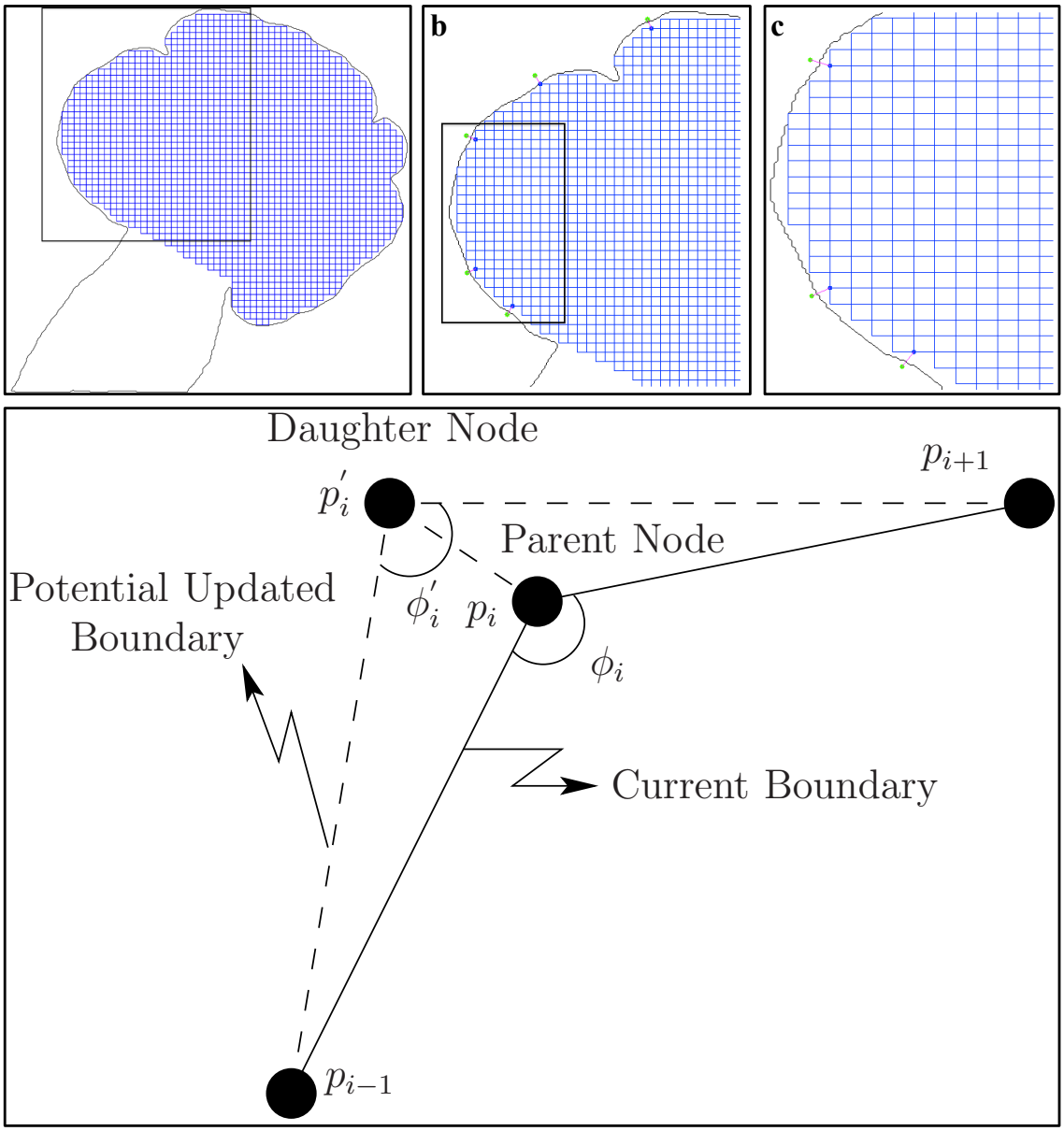


Figure 2.4: Creation of new vertices and maintaining boundary smoothness. The configuration of the cell-graph (initially a grid-graph where vertices are found at the intersection of the grid lines, and the grid lines are the edges of the graph) after a single iteration of creation of new vertices (step 2 of the dynamic graph-based growth algorithm) is shown in (a). The black rectangle in (a) represents the closer snapshot of the sub-graph viewed in (b). A further black rectangle in (b) represents a smaller sub-graph shown in (c). The spatial positions of parent vertices in the current boundary are used to identify the optimal location for daughter vertices, as shown in (d). The uniformity in the grid-graph is distorted with every iteration of the dynamic graph-based growth algorithm.

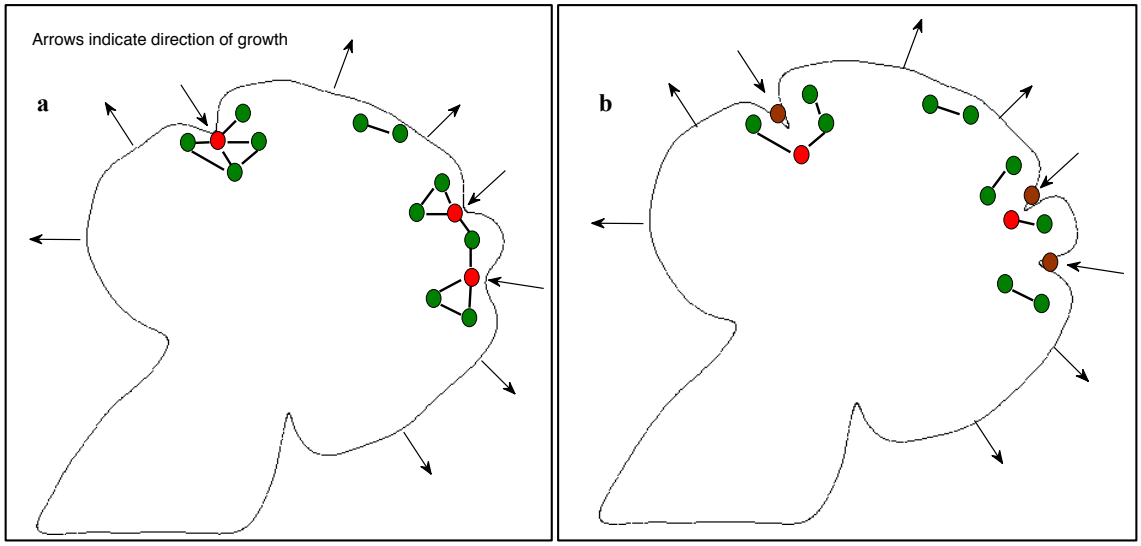
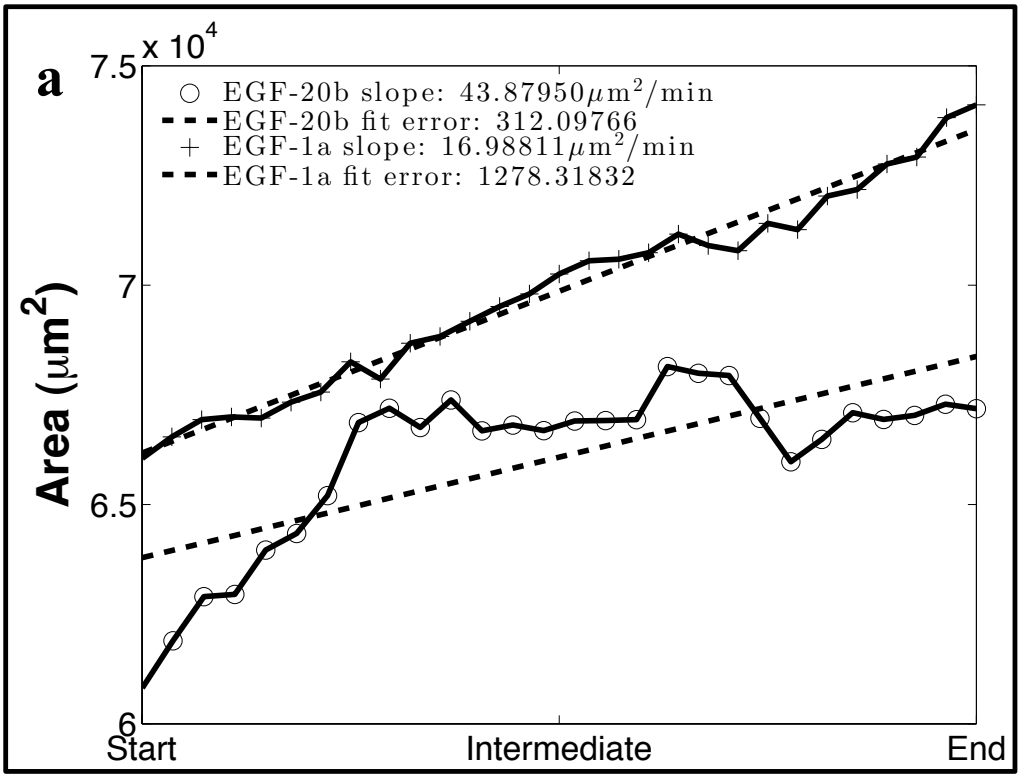
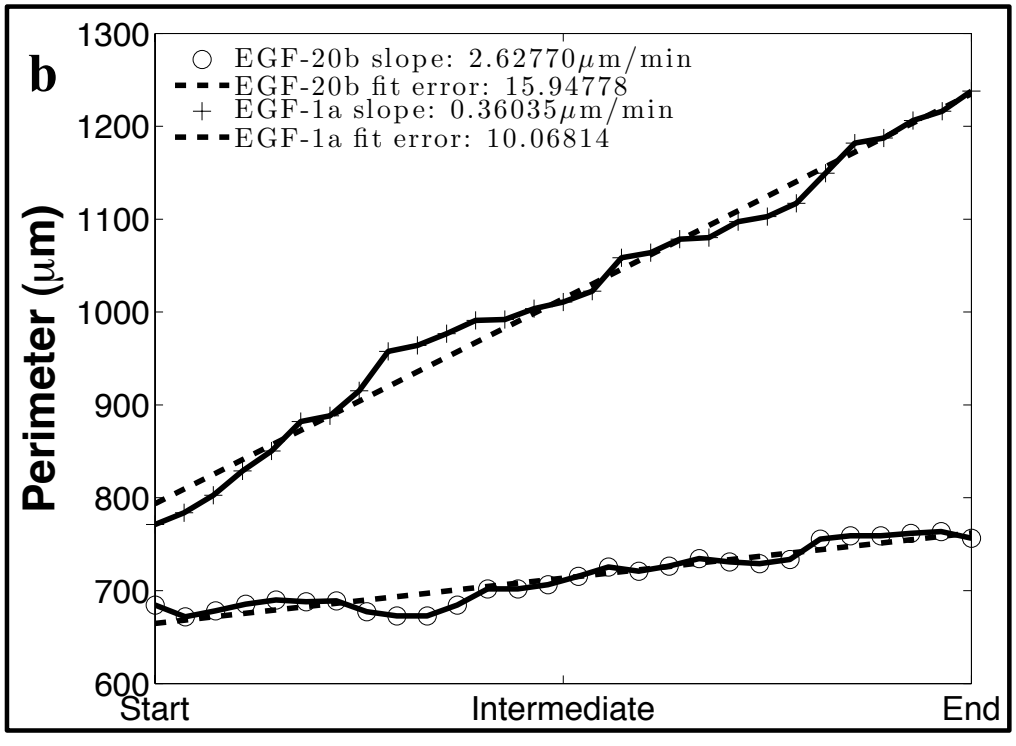


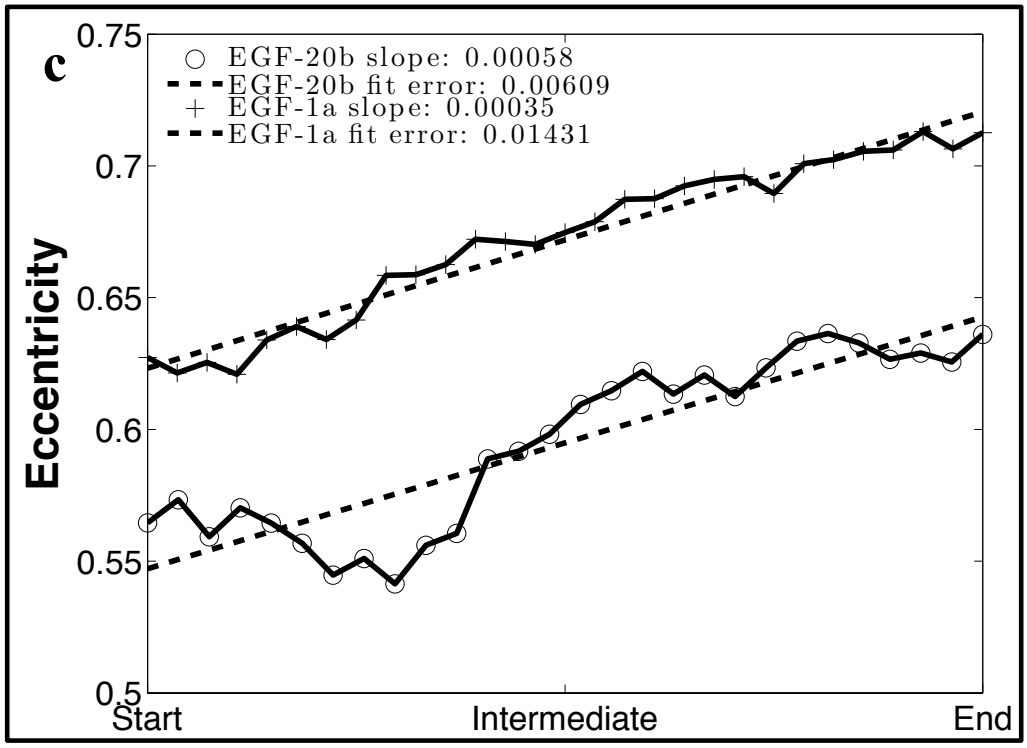
Figure 2.5: Illustration of the algorithm driving cleft deepening and dynamic cleft creation. The two panels show the initial (in a) and final (in b) conditions of the SMG before and after cleft deepening. In (a), vertices that are designated as cleft centers are marked in red. All other vertices are marked in green. After cleft deepening, all cleft centers lie in the cleft region (these original cleft centers are marked in brown) and are replaced by new vertices. The edges from these original cleft centers to other vertices are deleted. This effectively isolates these vertices from the graph. Supplementary Movie V1 shows mitosis events occurring over 3 hours.


a
x 10^4
7.5
7
6.5
6
Area (μm^2)
○ EGF-20b slope: 43.87950μm^2/min
EGF-20b fit error: 312.09766
+ EGF-1a slope: 16.98811μm^2/min
EGF-1a fit error: 1278.31832
Start
Intermediate
End

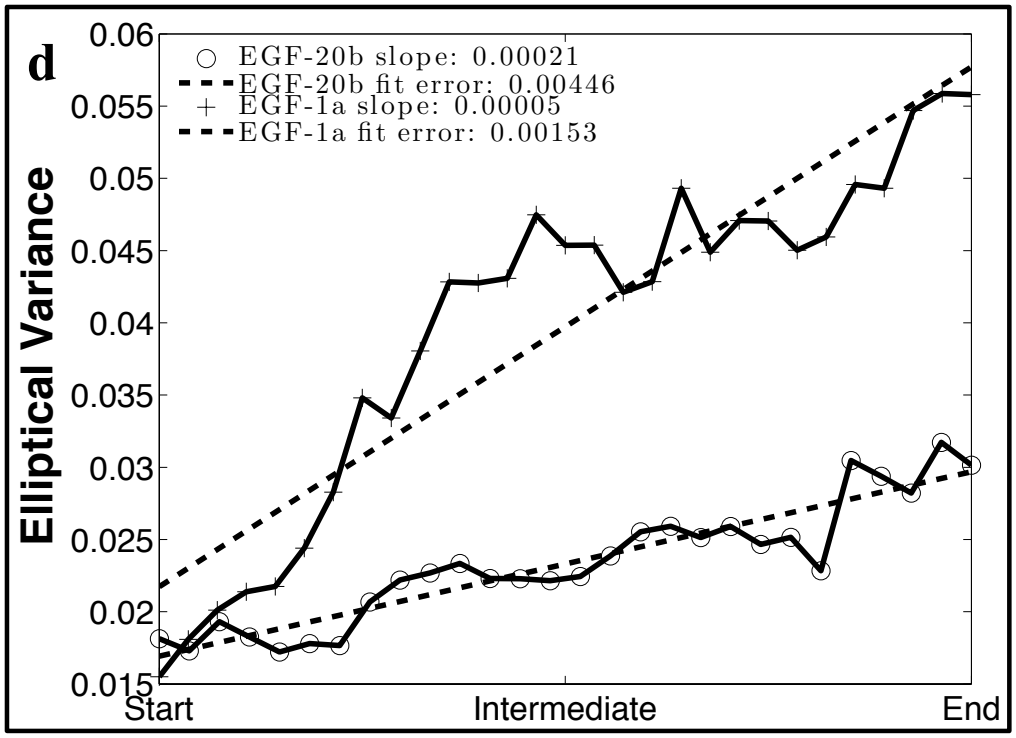
(a)


b
1300
1200
1100
1000
900
800
700
600
Perimeter (μm)
○ EGF-20b slope: 2.62770μm/min
EGF-20b fit error: 15.94778
+ EGF-1a slope: 0.36035μm/min
EGF-1a fit error: 10.06814
Start
Intermediate
End

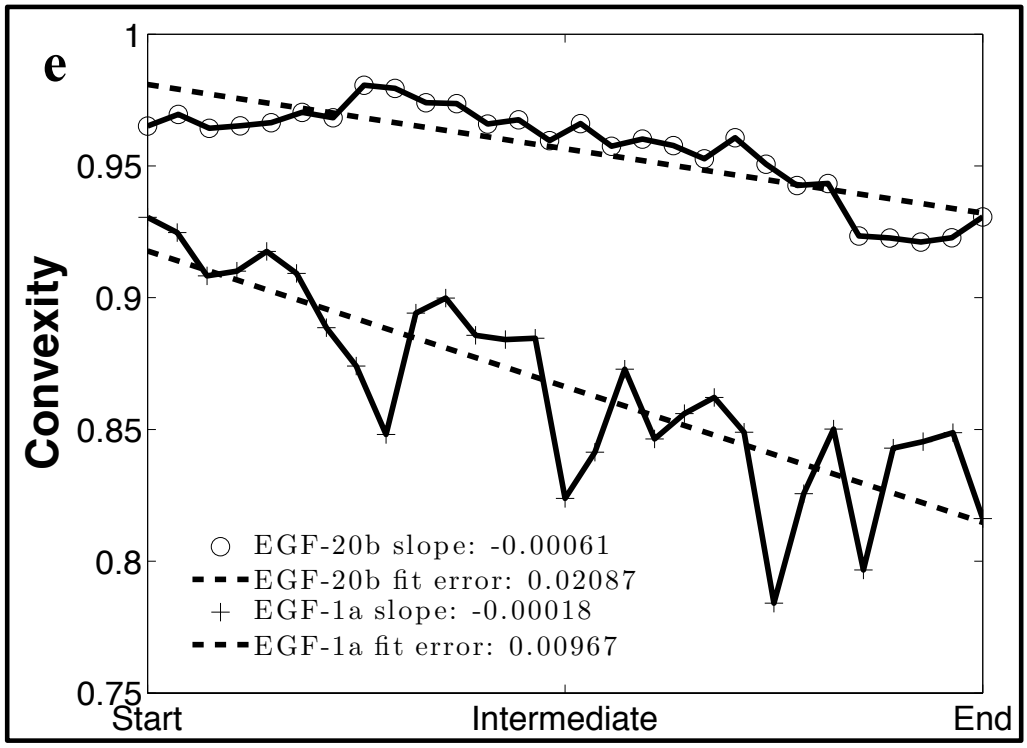
(b)


c
EGF-20b slope: 0.00058
EGF-20b fit error: 0.00609
EGF-1a slope: 0.00035
EGF-1a fit error: 0.01431
Eccentricity
0.75
0.7
0.65
0.6
0.55
0.5
Start
Intermediate
End

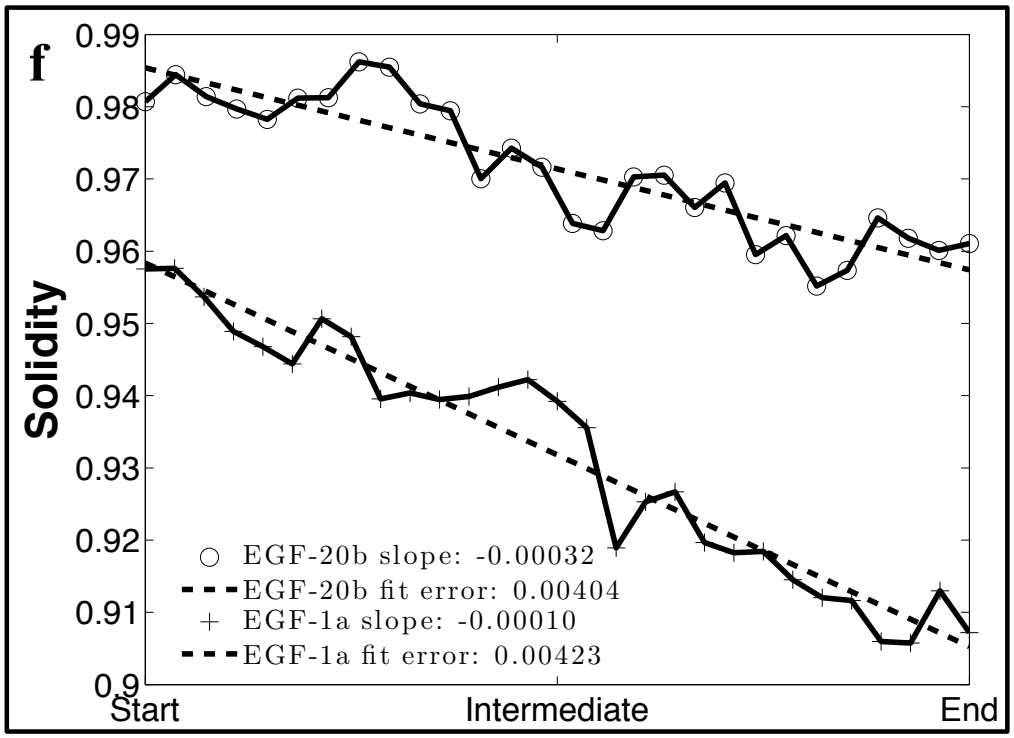
(c)


d
EGF-20b slope: 0.00021
EGF-20b fit error: 0.00446
EGF-1a slope: 0.00005
EGF-1a fit error: 0.00153
Elliptical Variance
0.06
0.055
0.05
0.045
0.04
0.035
0.03
0.025
0.02
0.015
Start
Intermediate
End

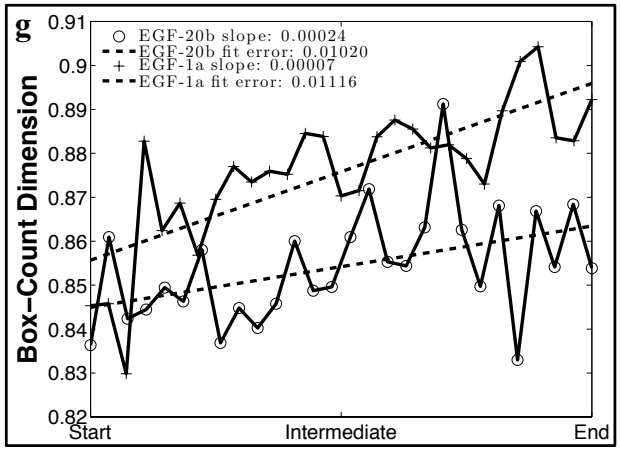
(d)


e
Convexity
1
0.95
0.9
0.85
0.8
0.75
Start
Intermediate
End
EGF-20b slope: -0.00061
EGF-20b fit error: 0.02087
EGF-1a slope: -0.00018
EGF-1a fit error: 0.00967

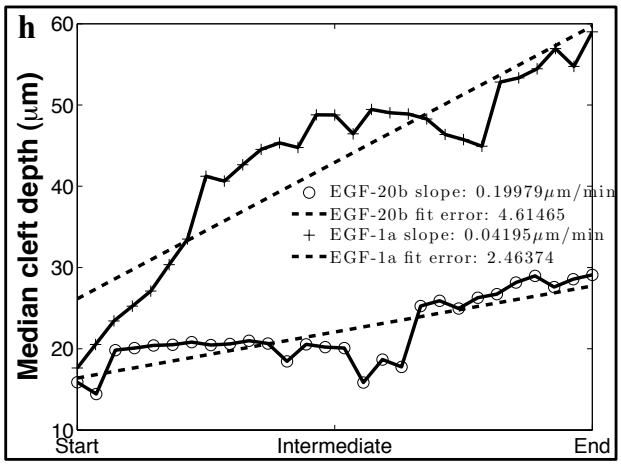
(e)


f
Solidity
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
Start
Intermediate
End
EGF-20b slope: -0.00032
EGF-20b fit error: 0.00404
EGF-1a slope: -0.00010
EGF-1a fit error: 0.00423

(f)



(g)



(h)

Figure 2.6: Global and local features characterizing the morphology of SMG developmental stages. Area, perimeter, eccentricity, elliptical variance, convexity, solidity, box-count dimension, and median cleft depth for EGF-1a and EGF-20b data sets are shown in (a)-(h). There are 29 images in the EGF-1a data set stretching over a period of 3 hours, and 47 images in the EGF-20b data set stretching over a period of 8 hours. Feature values for EGF-1a are shown by solid lines with "o" marker, whereas feature values for EGF-20b are shown by solid lines with "+" marker. The trends in each feature are indicated with the best-fit lines (dashed lines) based on the fitting root-mean square errors. Deeper clefts are observed for EGF-1 data sets (solid line with "o" marker) as compared to EGF-20 data sets (solid line with "+" marker).

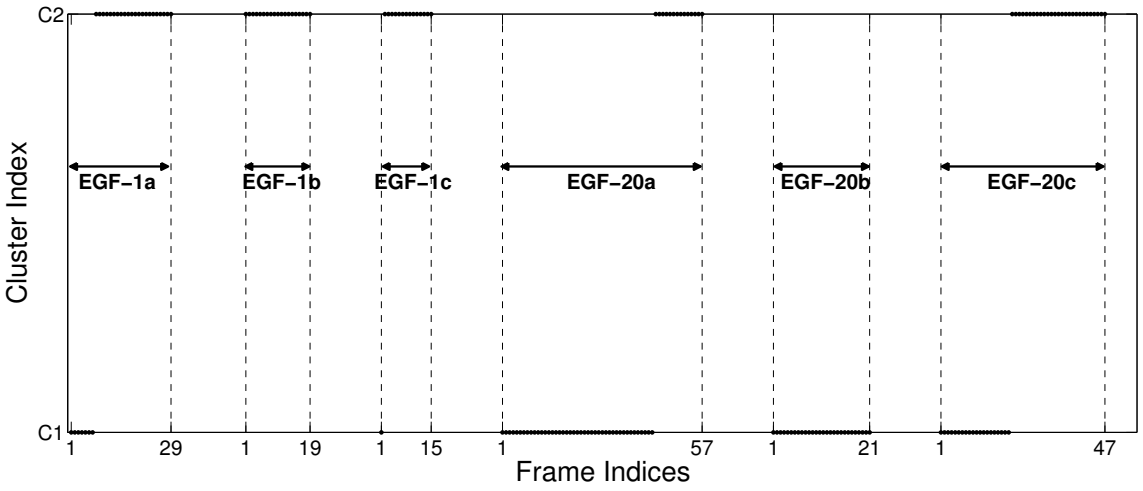
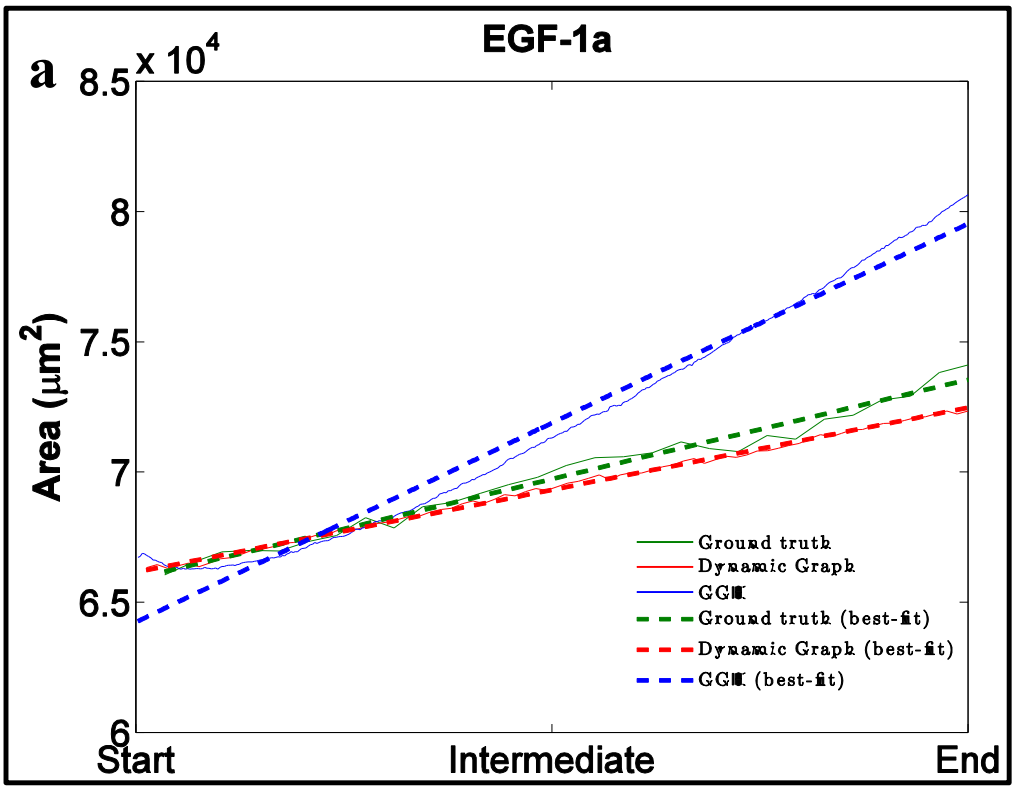
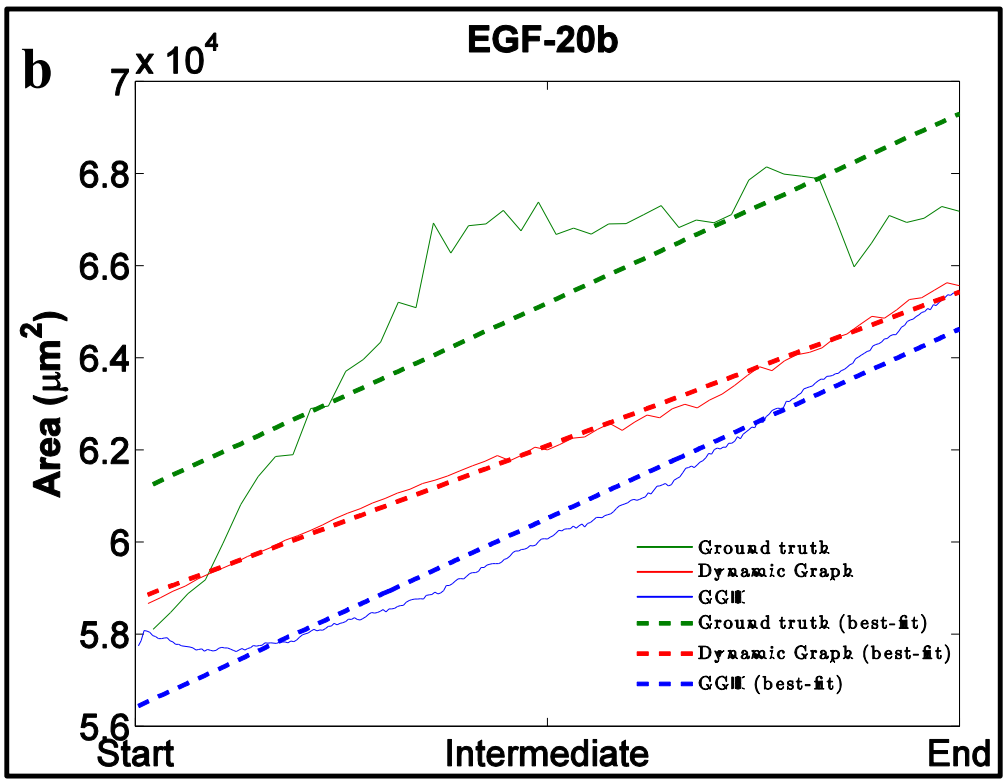


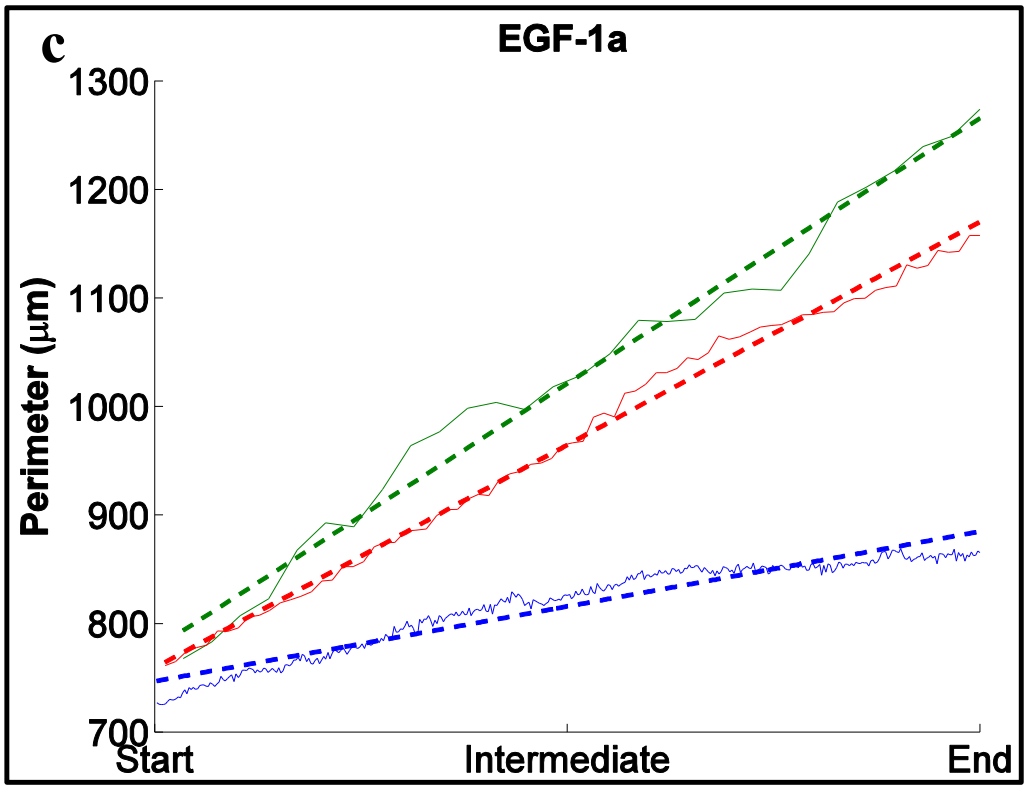
Figure 2.7: k**-means clustering of EGF-1 and EGF-20 data sets based on morphological features. The markers represent the frames that belong to each data set. The EGF-1 data sets are listed first, followed by the EGF-20 data sets. All the data sets are ordered chronologically from the first image in the set to the last image in that set, where the frame indices for the initial and final image are indicated in the abscissa. We utilize the k-means clustering algorithm with k = 2 clusters. It is observed that cluster C1 is characteristic of smaller buds and shallows clefts, whereas cluster C2 is characteristic of larger buds and deeper clefts. The majority of EGF-1 data sets are present in cluster C2 indicating that this cluster represents larger but fewer buds with deep clefts. The majority of EGF-20 data sets are present in cluster C1 indicating that this cluster represents smaller but multiple buds with shallow clefts. Although, increased EGF stimulates branching morphogenesis by creating more buds, clefts are shallow in comparison to lower EGF concentrations.**


a
EGF-1a
x 10^4
8.5
8
7.5
7
6.5
6
Area (μm^2)
Start
Intermediate
End
Ground truth
Dynamic Graph
GGH
Ground truth (best-fit)
Dynamic Graph (best-fit)
GGH (best-fit)

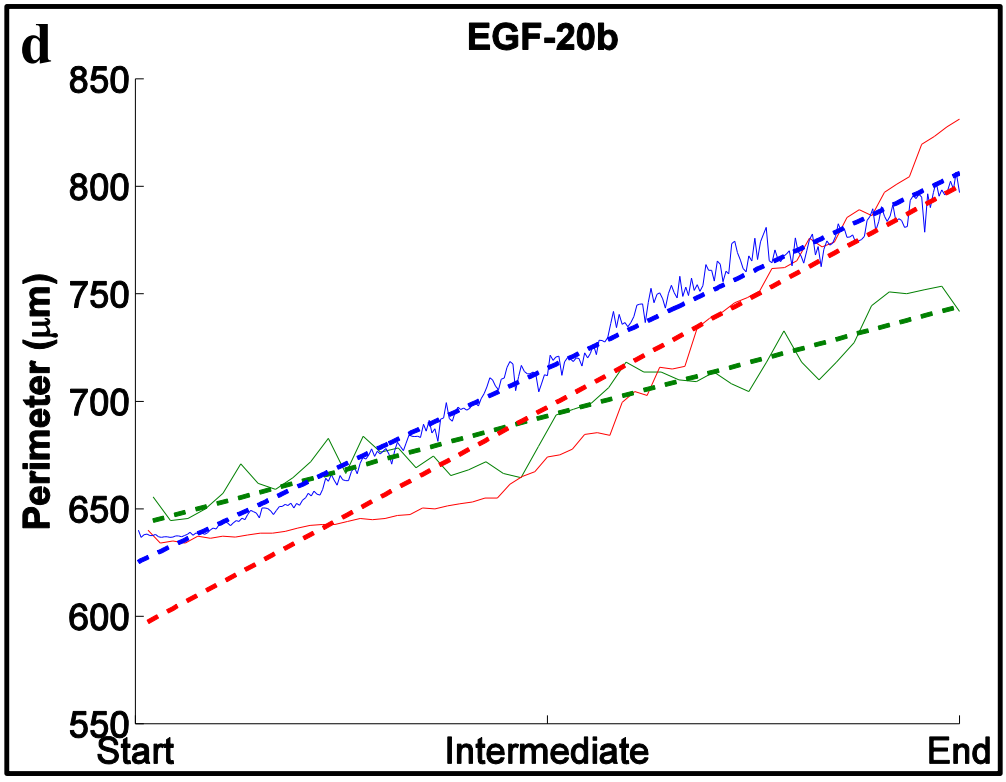
(a)


b
EGF-20b
x 10^4
7
6.8
6.6
6.4
6.2
6
5.8
5.6
Area (μm^2)
Start
Intermediate
End
Ground truth
Dynamic Graph
GGH
Ground truth (best-fit)
Dynamic Graph (best-fit)
GGH (best-fit)

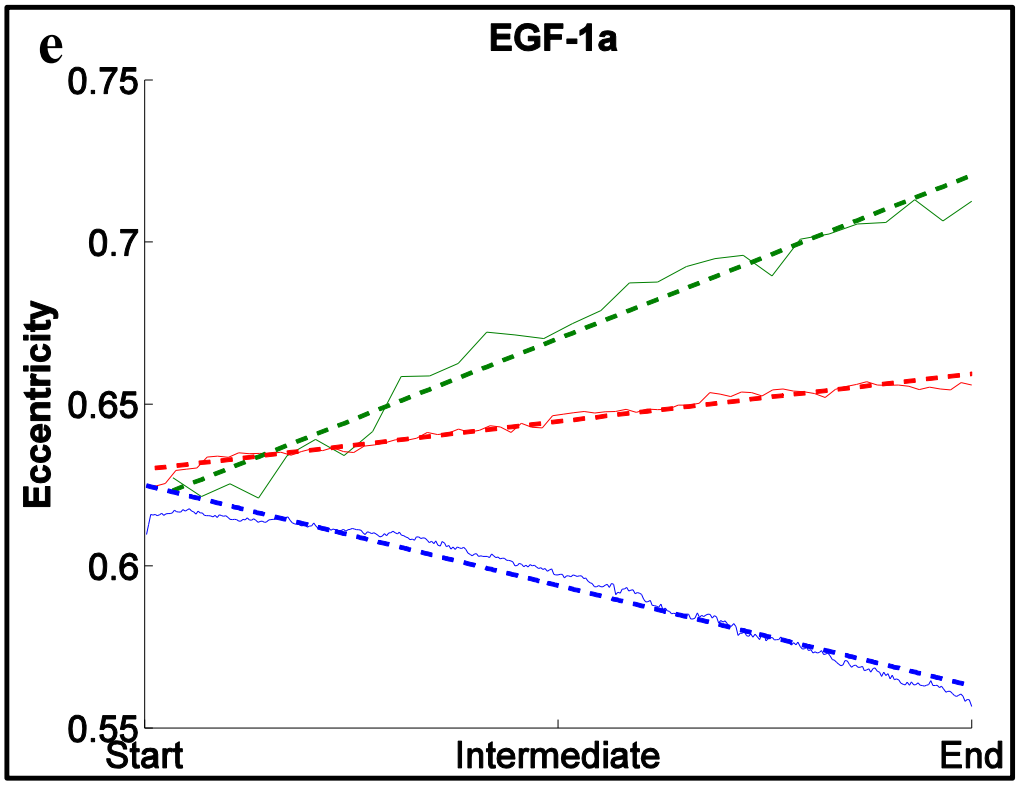
(b)


c
EGF-1a
Perimeter (μm)
1300
1200
1100
1000
900
800
700
Start
Intermediate
End

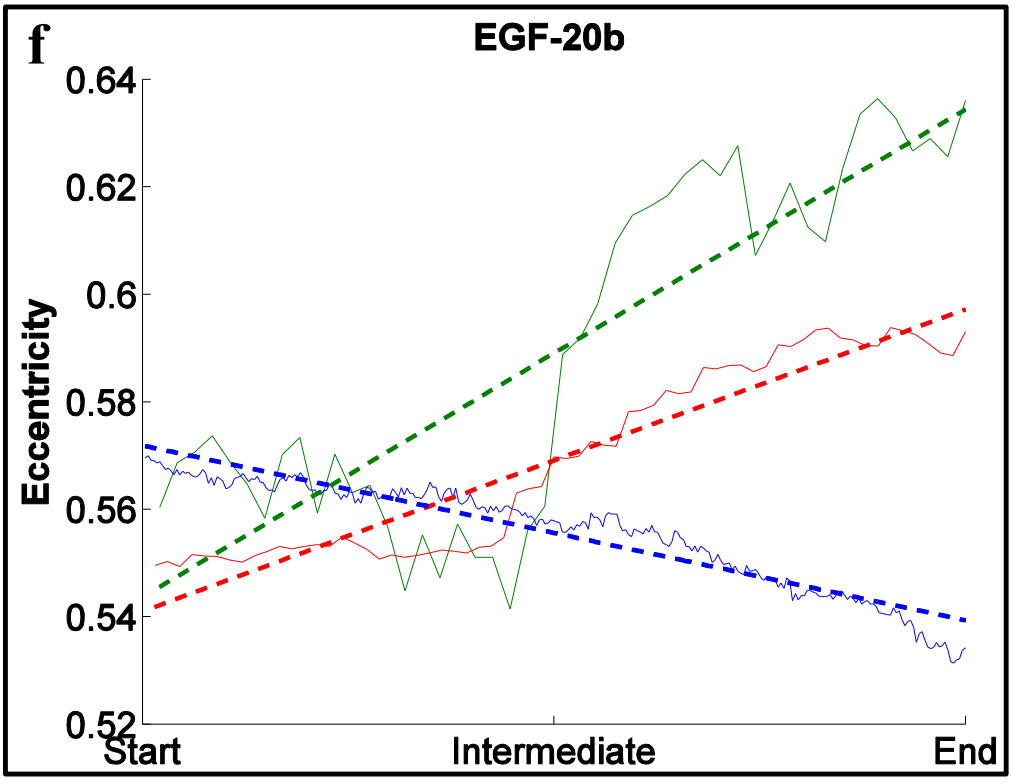
(c)


d
EGF-20b
Perimeter (μm)
850
800
750
700
650
600
550
Start
Intermediate
End

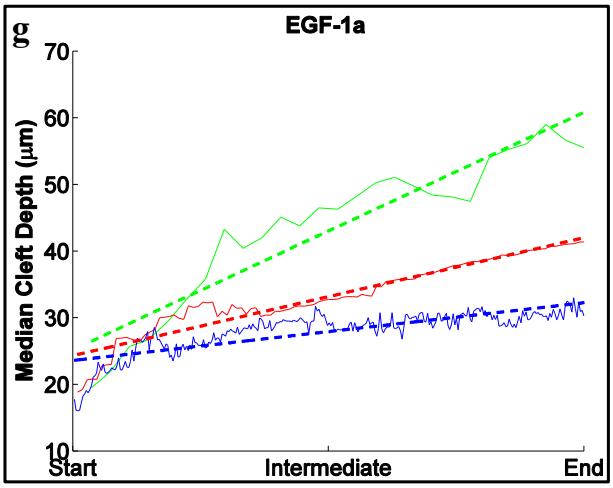
(d)


e
EGF-1a
0.75
0.7
0.65
0.6
0.55
Eccentricity
Start
Intermediate
End

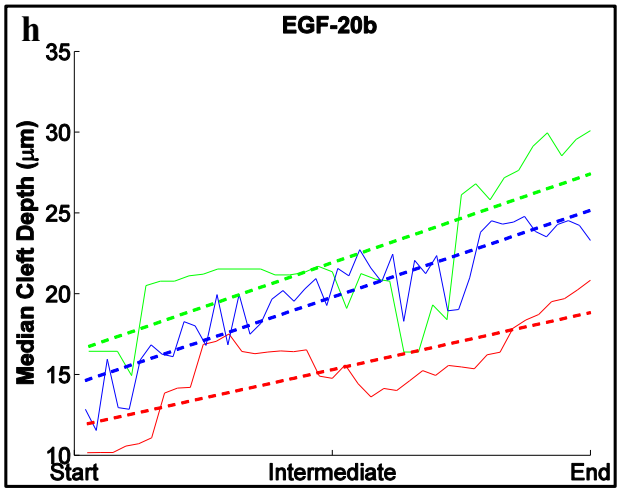
(e)


f
EGF-20b
0.64
0.62
0.6
0.58
0.56
0.54
0.52
Eccentricity
Start
Intermediate
End

(f)

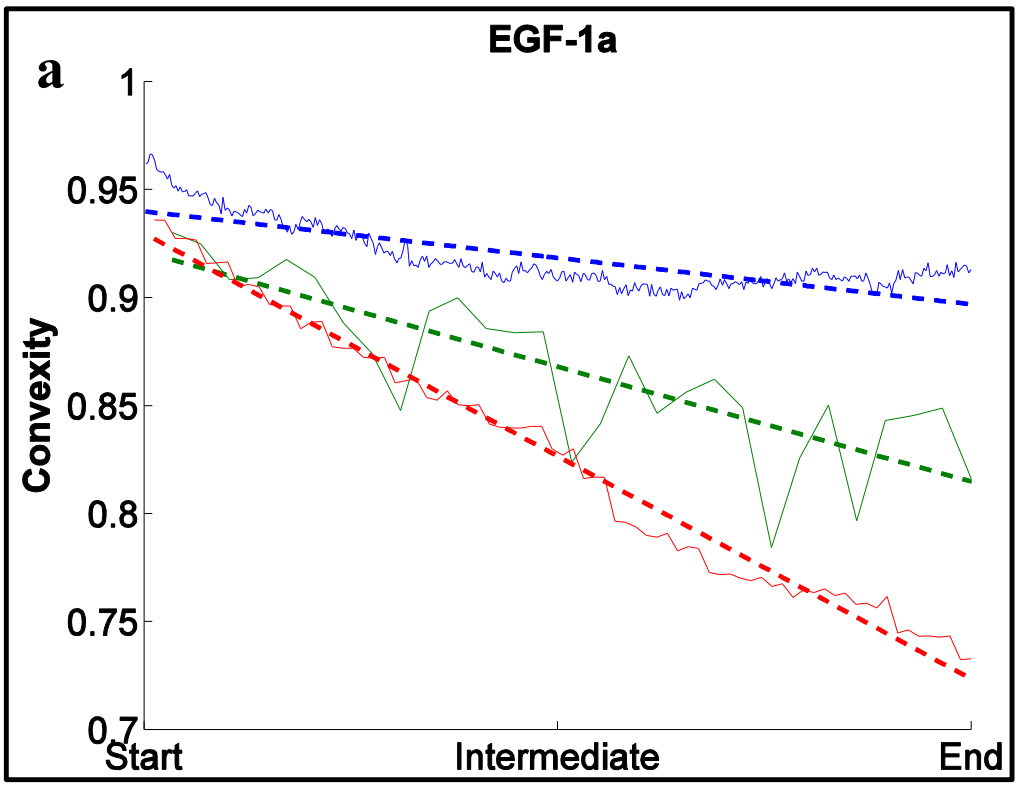


(g)

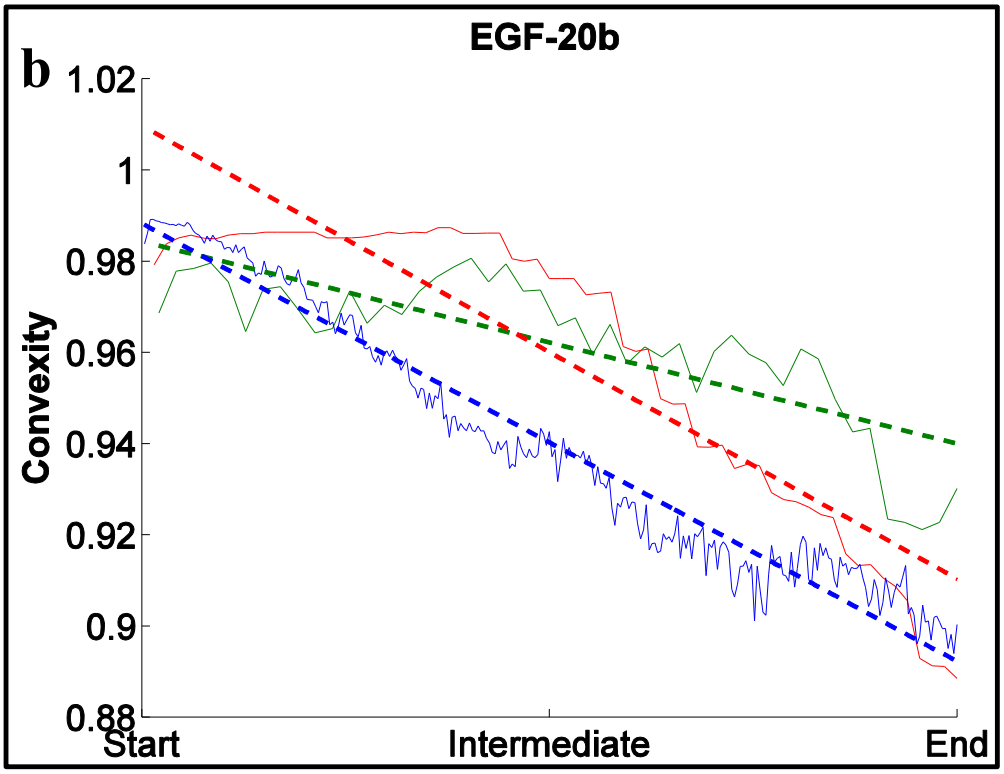


(h)

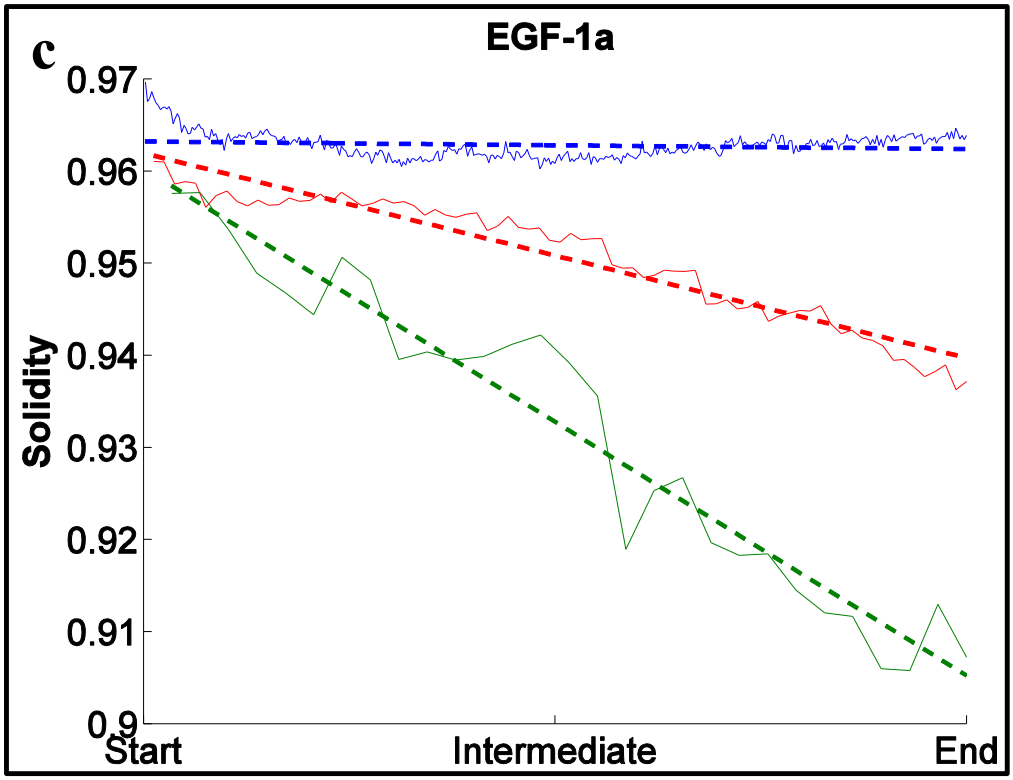
Figure 2.8: Comparison of SMG morphological features between ground truth (green), dynamic graph-based growth model (red), and the GGH simulation (blue) for EGF-1a and EGF-20b data sets. The features are area (a) and (b), perimeter (c) and (d), eccentricity (e) and (f), and median cleft depth (g) and (h). "Start" refers to cleft initiation, "Intermediate" refers to mid cleft progression stage, and "End" refers to beginning of cleft termination. Area and perimeter display increasing trends and similar trends are also seen for the dynamic graph model as well as the GGH simulation. GGH shows the opposite trend for eccentricity as it tries to create circular shapes. Increasing median cleft depth trends are shown by both dynamic graph model and the GGH simulation. For more detailed explanation of the feature trends, please refer to Section 3.3.


a
EGF-1a
Convexity
1
0.95
0.9
0.85
0.8
0.75
0.7
Start
Intermediate
End

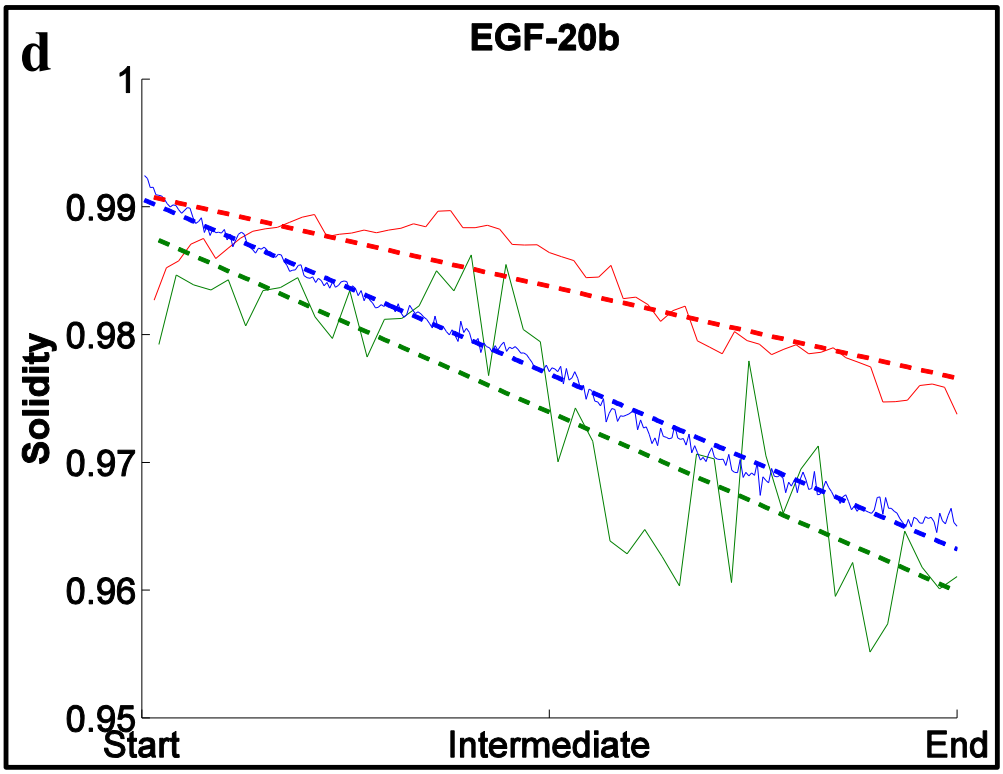
(a)


b
EGF-20b
Convexity
1.02
1
0.98
0.96
0.94
0.92
0.9
0.88
Start
Intermediate
End

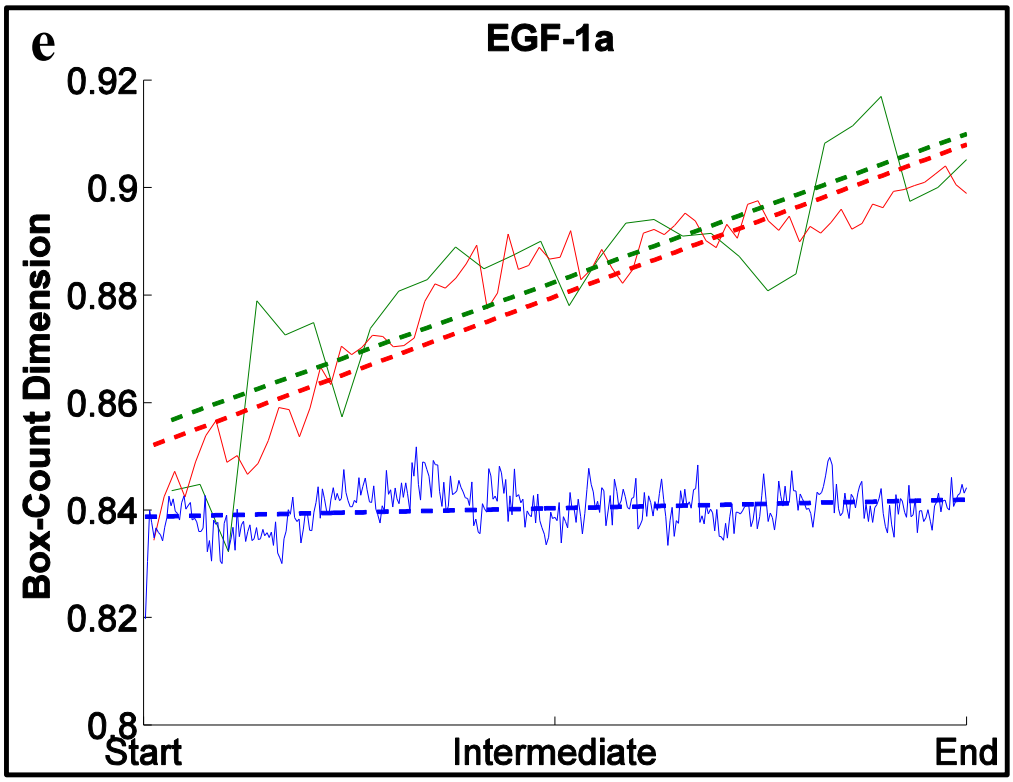
(b)


c
EGF-1a
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.9
Solidity
Start
Intermediate
End

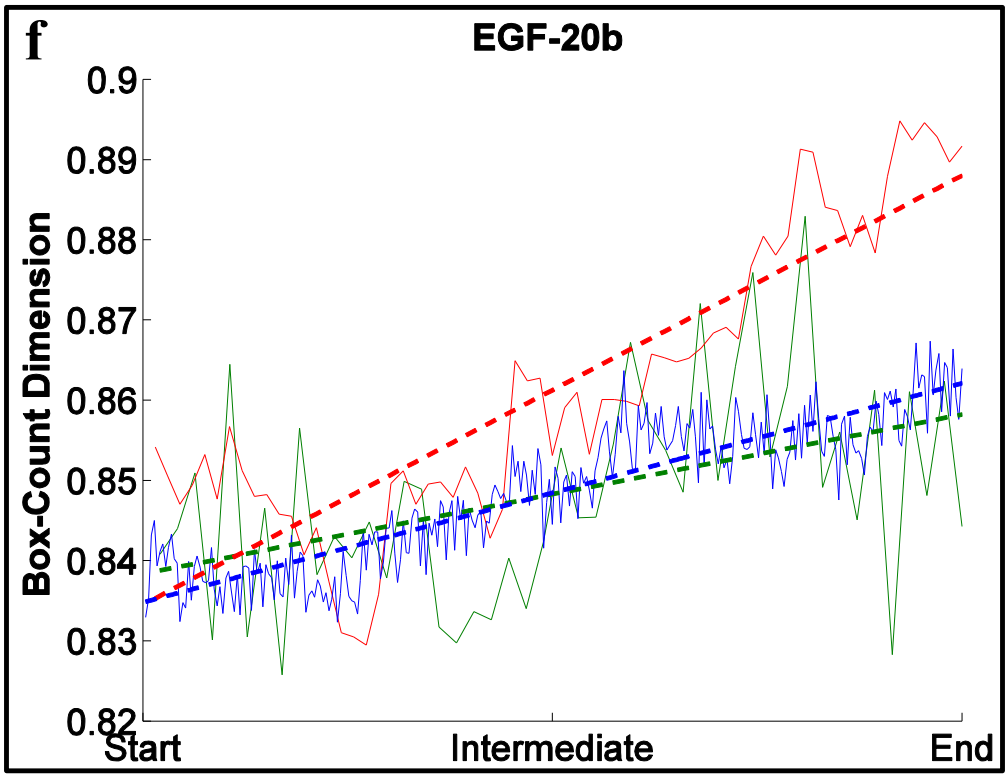
(c)


d
EGF-20b
1
0.99
0.98
0.97
0.96
0.95
Solidity
Start
Intermediate
End

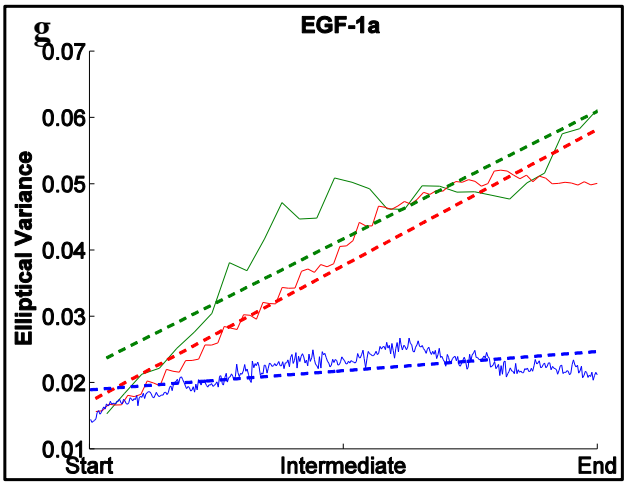
(d)


e
EGF-1a
Box-Count Dimension
0.92
0.9
0.88
0.86
0.84
0.82
0.8
Start
Intermediate
End

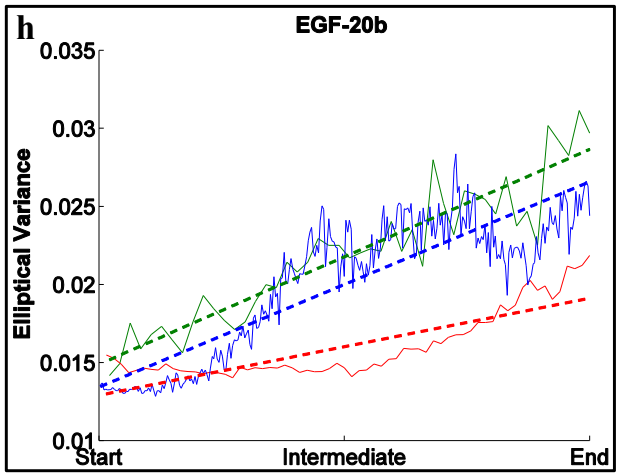
(e)


f
EGF-20b
Box-Count Dimension
0.9
0.89
0.88
0.87
0.86
0.85
0.84
0.83
0.82
Start
Intermediate
End

(f)



(g)



(h)

Figure 2.9: Comparison of SMG morphological features between ground truth (green), dynamic graph-based growth model (red), and the GGH simulation (blue) for EGF-1a and EGF-20b data sets. The features are convexity (a) and (b), solidity (c) and (d), box-count dimension (e) and (f), and elliptical variance (g) and (h). "Start" refers to cleft initiation, "Intermediate" refers to mid cleft progression stage, and "End" refers to beginning of cleft termination. Convexity and solidity display decreasing trends and similar trends are also seen for the dynamic graph model as well as the GGH simulation. Both dynamic graph model and the GGH simulation show increasing box-count dimension and elliptical variance trends.

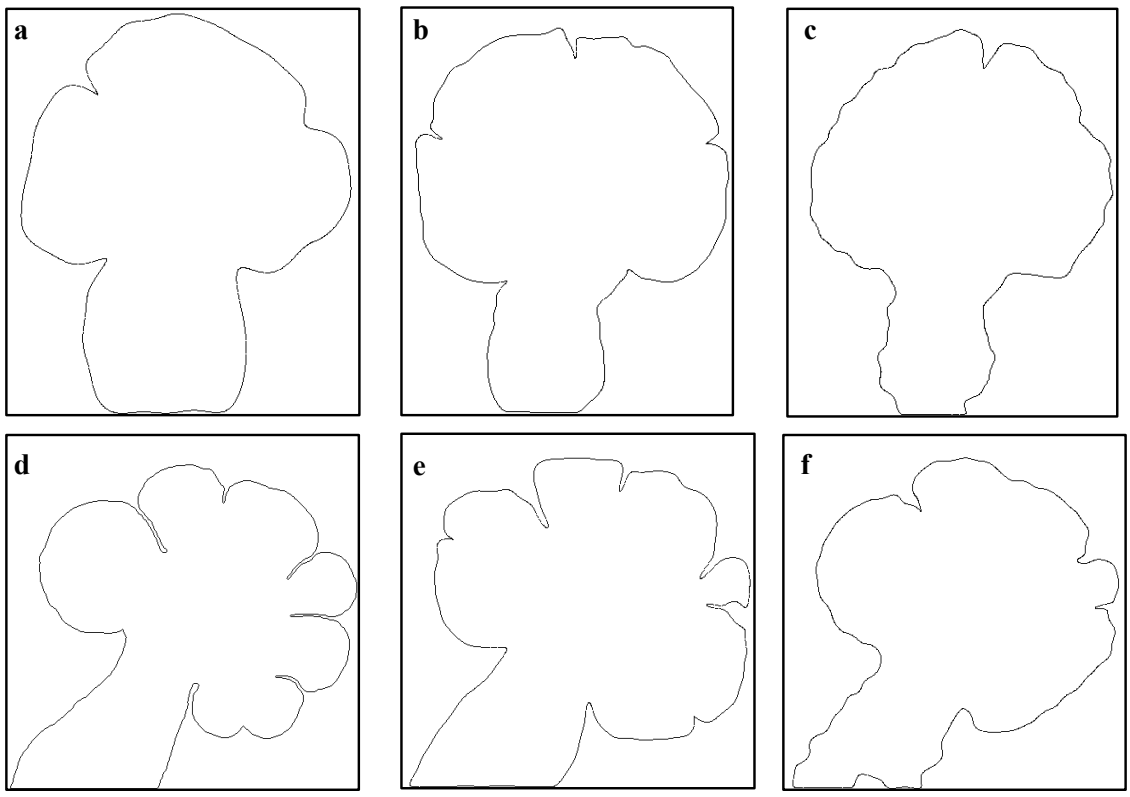


Figure 2.10: Target configuration of the ground truth data sets, and the final configurations reached by the simulations for the dynamic graph-based growth model and the GGH simulation. The target ground truth configuration is shown in (a) and (d), the dynamic graph model's configurations are shown in (b) and (e), and the GGH model's configurations are shown in (c) and (f). Dynamic graph-based growth model creates more clefts than the GGH model and this is one of the primary reasons that it has better quantitative agreement with the ground truth in regards to the morphological features.

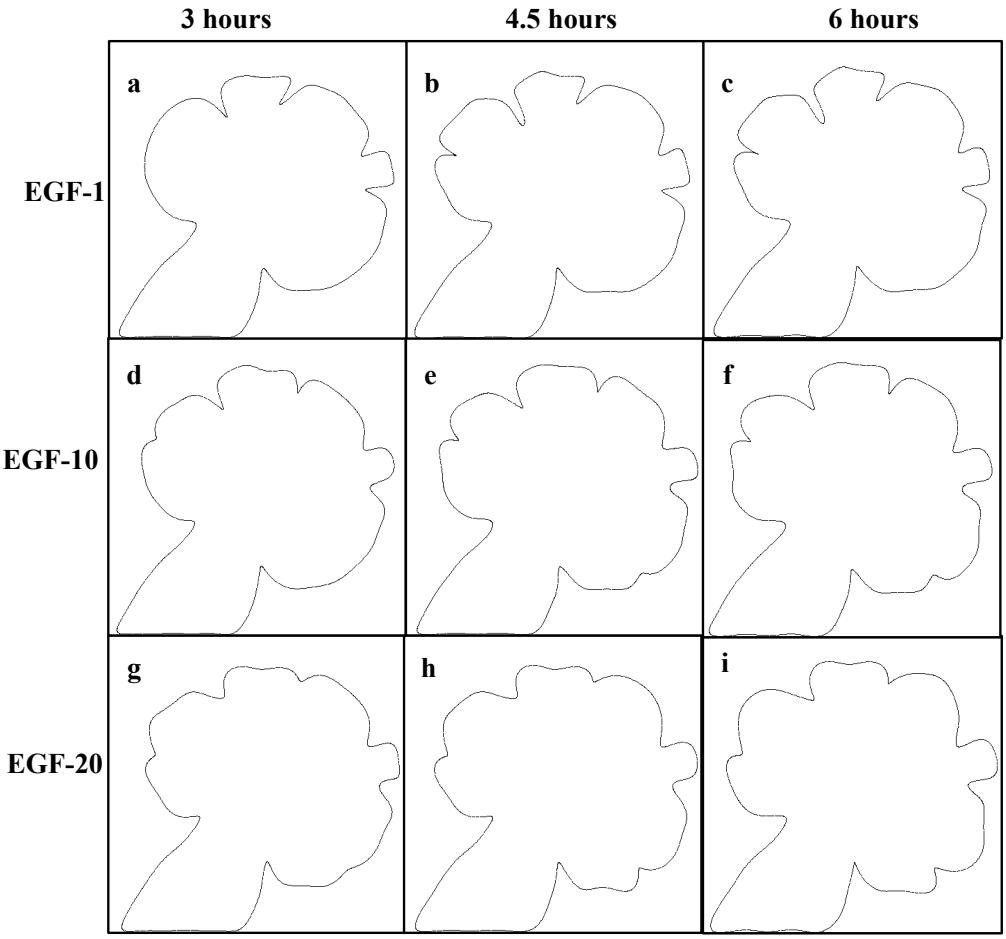


Figure 2.11: Dynamic graph-based prediction model results. Same starting image grown under different EGF concentrations for 3, 4.5, and 6 hours. The configuration of the salivary gland after 3 hours under EGF-1, EGF-10, and EGF-20 concentrations are shown in (a), (d), and (g), respectively. The configuration of the salivary gland after 4.5 hours under under the same concentrations are shown in (b), (e), and (h), respectively. The configuration of the salivary gland after 6 hours under the same concentrations are shown in (c), (f), and (i), respectively. The number of buds increases from EGF-1 to EGF-20, with clefting occurring more frequently in higher EGF concentrations. Buds are larger and clefts are deeper in lower EGF concentrations. Please recall that we do not compare the prediction model to a Monte-Carlo-based simulation (MCS) model. For further information about the prediction model, please refer to Section 2.3.4

CHAPTER 3
MODEL COUPLING FOR PREDICTING A DEVELOPMENTAL PATTERNING PROCESS

3.1 Introduction

Predictive models are founded on the evolution, and parametric optimization, of continuous physical/material properties of tissues over time. These approaches fall into two main categories [78] – (i) approaches that consider simple shapes for the cells and incorporate large populations of cells thereby focussing on the changes at the tissue scale with limited spatial detail at the cellular scale, and (ii) approaches that aim to model the changes at cellular scales by considering deformable shapes for the cells and smaller cellular populations with higher spatial detail. Underlying mathematical formulations in the models are either deterministic (such as ordinary/partial differential equations (ODE/PDE)) or stochastic (such as Monte Carlo simulations). Depending on how the positions of the cells are defined, these models can also be divided into two classes: on- and off-lattice models [78]. In on-lattice models, the locations and the interaction neighborhood of the cells are defined on a regular lattice structure such as a square or hexagonal lattice. This provides a simplification on the computational complexity of the models. The models in this category include cellular automata (CA) models based on a game theoretical approach [79] and cellular Potts models [80, 81]. These can be considered as extension of Ising methods which define a "cost/energy function" and use Monte Carlo simulations to minimize it. Although these agent-based techniques predict some organizational patterns of cells, they oversimplify the problem and require apriori setting of initial simulation parameters (e.g., cell-cell adhesion, target volume, elasticity, surface area). The fundamental problem with these models is that representing cell-cell interactions by interaction potentials vastly simplifies the complex nonlinear anisotropic mechanics between cells.

Off-lattice models use a more realistic representation of cell distributions and interaction neighborhoods, but are typically more computationally complex than

on-lattice models. Off-lattice models include cell-centered and elastic models based on ODEs and PDEs [33, 82, 83]. Yet these models cannot capture tissue structural organization since the parameters do not consider the biological properties of tissues such as intercellular communication and change in cellular phenotype. For instance, physics-based-predictive models solved with finite element methods can predict the force-displacement (i.e., shape) of a gland after three days, but must apriori ignore the functional relationships between elements of a developing organ.

One common property of the physics based models is that they can be data agnostic and still produce reasonable results which may be good enough for certain application domains such as lung branching morphogenesis [37] but would fail in other cases such as salivary gland branching morphogenesis.

In contrast to these continuous models, data-driven discrete models are based on extracting features, and using statistical machine learning algorithms for prediction. However, they suffer from the difficulty of mapping predictions from feature space back to the original domain. For example, it is hard to construct the glandular membrane given only the prediction of spatial cellular organization.

Branching morphogenesis is an evolutionary process that has been studied using continuous physics-based models. This process is responsible for the development of glands such as lungs, kidneys, and pancreas [4]. Via this branching process the gland attains a morphology that maximizes the total area of contact between the metabolic exchange surfaces and the surrounding environment, while minimizing the total volume of the organ. This allows for an efficient exchange of gases, nutrients, metabolites, and wastes. As a consequence of branching morphogenesis, the gland undergoes significant transformation [5]. Starting out from a primary bud, where small clefts (indentations in the basement membrane) appear, the gland obtains a complex structure via bud outgrowth and cleft progression, followed by formation of ducts. Once the new buds are fully formed, cleft formation begins anew, and occurs reiteratively throughout development to create the ramified structure of the adult organ.

Most importantly a modeling approach should be faithful to the biology and capture the underlying processes. For example, branching morphogenesis requires

an understanding of the molecular mechanisms regulating epithelial-mesenchymal interactions during development of the glands [84, 37]. The process is highly dynamic, involving interactions between multiple participating cells and molecules. Thus, our hypothesis is that coupling a pair of models: one predicting the continuous topological evolution and the other predicting cellular organization, would result in more realistic and accurate modeling. This also leads to a model that functions at both the organ and tissue levels, a truly multiscale model.

In this paper we present a case study for coupling continuous-physics-based models with discrete empirical ones to address the prediction of cleft formation during the early stages of branching morphogenesis in mouse submandibular salivary glands (SMG) and make a proof of concept proving the veracity of our hypothesis.

The rest of this paper is organized into four sections. Following this introduction (Section 3.1), Section 3.2 describes the experiments performed to collect the biological data, initial image processing techniques applied to the data, algorithm for cleft detection, extraction of global morphological features, and the details of the coupled model. Section 3.3 presents the results of comparing the predictive capability of the coupled model to the ground truth and a Monte-Carlo-based-simulation model. Section 3.4 summarizes the findings of this study.

3.2 Materials and Methods

3.2.1 Data Acquisition

3.2.1.1 *Ex Vivo* Submandibular Salivary Gland Epithelial Organ Cultures

Experiments were carried out on timed-pregnant female mice (strain CD-1, Charles River Laboratories) at embryonic day 12 (E12), with day of plug discovery designated as E0. The protocols approved by the National Institute of Dental and Craniofacial Research IACUUC committee were followed to obtain these SMG rudiments. These E12 SMG rudiments were cultured in the presence of 100 ng/mL Fibroblast Growth Factor (FGF). The E12 SMG initiates with a single primary epithelial bud, which was microdissected, and the mesenchyme was removed. The media was supplemented with 1 ng/mL EGF (R&D Systems). Images were collected

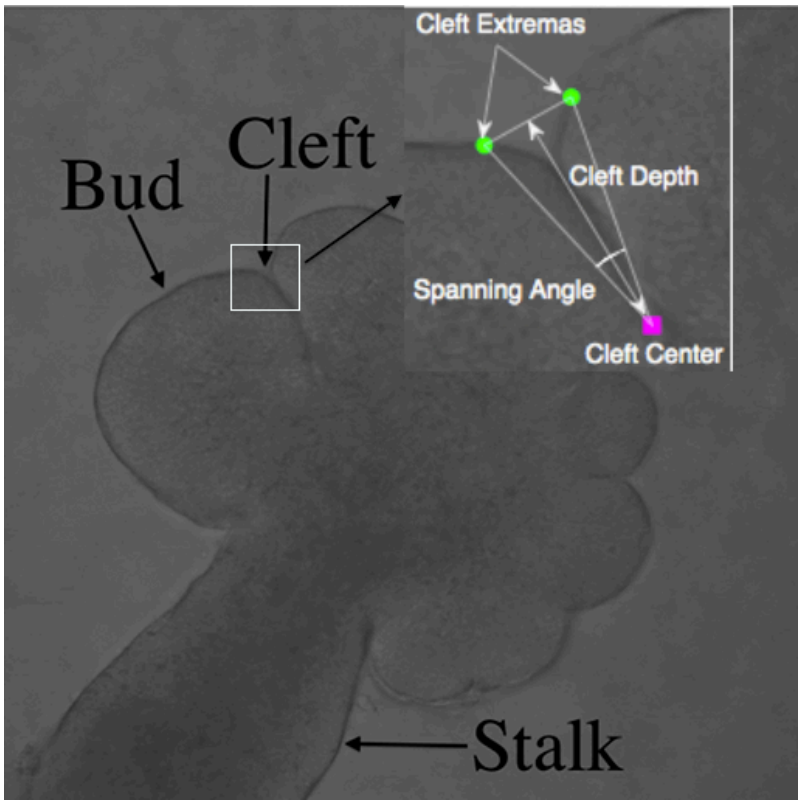


Figure 3.1: Characterization of clefts. The enlarged figure shows the constituent parts of the gland. The area in the white rectangle is blown up and shown in the top right with the cleft extrema marked in green and cleft center marked in maroon. Spanning angle and cleft depth are calculated from these points as illustrated.

as described in the next subsection.

3.2.1.2 Confocal Time-lapse Series Acquisition

A tracer molecule such as Alexa Fluor 647-labeled human plasma fibronectin (FN) was used to label the epithelial rudiments, to identify the location of the basement membrane, or with green fluorescent protein (GFP) to label a subset of the epithelial cells, and imaged using time-lapse confocal microscopy. A 20X, 0.8 N.A. objective with an additional level of optical zoom (1.7-2.5X) Zeiss 510 Meta confocal microscope was used to image these glands. Images of either 7μm or 5μm thick optical sections were captured at either 6 min intervals. In some experiments, the 488 nm laser was used to capture an additional brightfield/near Differential Interference

Contrast (DIC) microscopy image. Image sets were captured at 512 × 512 pixel resolution using a scan speed of 8 in line averaging mode. For this study, images from the center of the explant were used for 2D analysis.

3.2.2 Quantification of Ground Truth

We quantify the ground truth to calculate certain measurable properties of the evolution of the gland that can then help in validating the predictions produced by our coupled model. Since there are 3 data samples each acquired from a different mouse, we calculate all statistics and parameters for our modeling approach on 2 mice, and test on the third mice.

3.2.2.1 Image Processing and Segmentation

The first step in characterizing the SMG morphology consists of segmenting the SMG regions in the FN (via ImageJ) time-lapse data sets. We employed Otsu's technique [67] to segment the FN images. This technique calculates an optimal threshold to differentiate the tissue (foreground) from the Matrigel and messenchyme (background). The segmentation results were visually inspected by biologists to verify that it accurately captured the morphology. To obtain nuclear information regarding cell distribution and cell morphologies, we referred to the *ex vivo* data set (in Section 3.2.1.1).

3.2.2.2 Detection of Cleft Regions

We begin characterizing the SMG morphology by identifying its constituent parts, primarily buds and clefts. The SMG is comprised of alternating buds and clefts, where clefts are narrow valley-shaped indentations that form in the basement membrane. Progressively these clefts become narrow and deep. We characterize the cleft region using cleft center defined as the deepest point of the cleft, with the walls of the cleft extending on either side of the surface normal at the cleft center, and the corresponding left and right extrema points that determine the extent of the cleft; the buds are considered to be starting beyond the points marked as cleft extrema. The cleft center and cleft extrema are illustrated in Fig 3.1.

Automatic detection of the cleft region is carried out by calculating the mean curvature along the boundary of the gland. The mean curvature, κ, of an interface is defined as the divergence of the normal $\vec{N} = (n_x, n_y)$,

$$\kappa = \nabla.\vec{N} = \frac{\partial n_x}{\partial x} + \frac{\partial n_y}{\partial y} \ , \tag{3.1}$$

where ∇ is the gradient operator. κ is positive for convex regions (buds), and negative for concave regions (clefts). By identifying an appropriate threshold τ, we can remove boundary irregularities, and are left with the correct cleft regions. We apply this algorithm to every image in the data set, and thus time is not considered a parameter in our formulation in Eq 3.1.

3.2.2.3 Extraction of Global SMG Morphological Features

The morphology of the SMG undergoes quantifiable transformations as a consequence of creating the ramified structure. We capture these transformations by extracting six morphological features, namely *area*, *perimeter*, *elliptical variance*, *convexity*, *solidity*, and *box-count dimension*. We label this data matrix of morphological features as $\mathbb{D}$. When referring to a feature matrix in subsequent text, we allude to the data matrix $\mathbb{D} \in \mathbb{R}^{M \times 6}$, consisting of the values of the six morphological features over M time-steps. Table 2.1 lists the definitions of the various morphological features.

3.2.3 Predictive Modeling

3.2.3.1 Modeling Shape Formation–Ignoring Cellular Organization

We used the level set method devised by Osher and Sethian [85], which was introduced in the area of fluid dynamics and has spread to many other research areas. The basic level set equation for evolution of an interface ϕ is given as:

$$\phi_t(x, y) + \vec{V}(x, y).\nabla\phi(x, y) = 0 \tag{3.2}$$

where ϕ_t denotes the temporal partial derivative, ∇ is the gradient operator, and $\vec{V}(x, y)$ is the velocity field on the interface. Note that ϕ represents both the interface

and the implicit function.

The central tenet to the level set approach is embedding the interface ϕ as the zero level set $\omega = 0$ of a higher-dimensional function ω. This implicit function ϕ divides $\mathbb{R}^2$ into subdomains with non-zero areas. $\phi(x, y)$ is equal to 0 for all points (x, y) lying on the interface, $\phi(x, y)$ is negative for all points (x, y) lying in the interior region as defined by ω, and $\phi(x, y)$ is positive for all points (x, y) lying in the exterior region as defined by ω.

A commonly used implicit function is the signed distance function. To understand the signed distance function, we first describe a distance function. A distance function $d(\vec{x})$ is defined as: $p(\vec{x}) = \min(|\vec{x} - \vec{x_I}|) \ \ \forall \ \vec{x_I} \in \phi_t$, implying that the Euclidean distance p on the boundary is 0. For all other points, the distance to the closest point on the boundary is calculated. Note that a distance function is positive in the entire domain as evidenced by the presence of the absolute value function. A signed distance function $\phi(x, y)$ is an implicit function with additional properties of distance functions including primarily $|\nabla\phi(x, y)| = 1$.

We define the level set function ω_0 for the initial image, $I_0(x, y)$, of the data set using a signed distance function of the form $\omega_0 = -P(x, y).I_0(x, y)+P(x, y).\overline{I_0(x, y)}$, where P is the Euclidean distance function, i.e. for every pixel it assigns the shortest distance to a point on the boundary of the gland. We take the zero level set ϕ_0 of ω as the starting interface location. Considering only the normal component of the external velocity in Eq. 3.2 (since $\vec{V_t}.\nabla\phi = 0$), we get

$$\phi_t(x, y) + \vec{V}(x, y)_n|\nabla\phi(x, y)| = 0\,. \tag{3.3}$$

Now, the velocity field can be defined by a term for motion in the normal direction

$$\phi_t(x, y) + a(x, y)|\nabla\phi(x, y)| = 0 \tag{3.4}$$

Numerical methods including upwind differencing schemes [86] are utilized to solve this equation and thus evolve the interface over time.

3.2.3.2 Modeling Cellular Organization–Ignoring Shape Formation

We model the cellular structural organization inside a gland by a graph $G = (V, E)$ consisting of a set of vertices V representing cell nuclei, and a set of edges E representing cell to cell interactions. An edge is established if the distance between two cell nuclei is less than a predetermined threshold. We assume cells to be circular in shape, and cell size is approximated by the diameter. The model takes the initial gland morphology and nuclei locations as input. Each iteration of the growth algorithm divides the cells into two populations based on the distance from the gland boundary, namely internal (I) and periphery (P). Subsets $I' \subset I$ and $P' \subset P$ are chosen to undergo a proliferation attempt. Cells in P' that successfully undergo mitosis create new cells (or vertices) V' that are added to V. For I', we compute the shortest distance to the boundary of the gland (not including the cleft region) and find the cell in P closest to that boundary point; new cells $\hat{V}$ thus created are added to V. New edges, E' and $\hat{E}$ for periphery and internal cells respectively, are also constructed based on the distances from the new cells to existing cells in G.

3.2.3.3 Coupling The Models and Enhancing Level Set

The objective behind coupling the two models is to enhance the physics-based-interface-evolution-algorithm by incorporating spatial cellular information, i.e. to allow level sets to produce a better prediction of gland growth by the addition of relevant biological information from within the gland. Coupling of these two models is achieved at the boundary where mesenchymal and epithelial cells are separated by the gland membrane. Intuitively, we aim to capture the opposing forces that the epithelial cells apply on the boundary to grow outwards while the mesenchymal cells push in the opposite direction. Thus we relate the surface tension for growth speed and direction to the cellular type and density at the boundary.

More precisely, we define the normal growth speed $a(x, y)$ as a function of the sign of the mean curvature $S(\kappa)$ and cellular density as

$$a(x, y) \propto f(x, y)\ \Gamma\ S(\kappa(x, y))\ , \tag{3.5}$$

where $\Gamma = \left[\frac{\text{Number of epithelial cells}}{\text{Number of mesenchymal cells}}\right]$. We use a rolling circle of radius equal to

$5 \times$ epithelial cellular nucleus centered at the point (x, y), along the interface to determine the epithelial and mesenchymal cells in the neighborhood of the this point. Note that the coordinates of these cells are determined by the graph algorithm summarized in Section 3.2.3.2.

The function $f(x, y)$ on the interface allows us to shape the buds anisotropically with the the center of the bud pushing outward faster than the extremities. $f(x, y)$ is an exponential function such that the maximum value for the function is found at the center of the bud, and exponentially drops-off towards the clefts where it approaches negative values. The sign of the curvature is defined as:

$$S(\kappa(x,y)) = \begin{cases} 1, & \text{if } S(\kappa(x,y)) \geq 0 \text{ (at convex bud regions)} \\ -1, & \text{if } S(\kappa(x,y)) < 0 \text{ (at non-convex cleft regions)} \end{cases}$$

Note that the curvature is only used to determine areas corresponding to clefts and buds by using the sign of the curvature. Its magnitude is not used in our formulation.

When the size of a bud increases beyond a pre-determined threshold (9% as observed from the biological data), we introduce a dynamic cleft in the bud, effectively creating two new buds. The position of the split is determined probabilistically according to a Gaussian distribution fit to the coordinate locations on the boundary with the center of the bud becoming the mean of this distribution. We then select coordinate points that fall less than one standard deviation away from the mean. These points are in the vicinity of the center of the bud, and we randomly choose a point from this list as the cleft center for the dynamic cleft. Thus, the cleft is not always created in the same position along the boundary.

On analyzing the ground truth, it was determined that cleft depth is a function of the size of the adjacent buds. Larger adjacent buds allow the cleft to progress much deeper into the tissue. Table 2.2 lists correlation coefficients, represented by ρ, between cleft depth and adjacent bud sizes for the data set under investigation.

To create and maintain regions of negative growth, we define triangular cusp regions for the cleft. Whenever a cleft is determined, we use bud size statistics to ascertain the possible depth of the cleft. The cusp is a function of the form $y \propto x^{1.5}$,

where x and y are interface (or boundary) coordinates. This cusp region prevents the buds from getting closer than a pre-determined minimum cleft spanning angle preventing cleft walls from collapsing. This formulation allows the clefts to grow deeper and the buds to grow outwards. Since the evolving interface is the zero level-set of the signed distance function and is reinitialized at every iteration, it is a continuous closed interface. The zero-crossing regions where the normal growth speed crosses from positive to negative (and vice versa, so called inflection points) are small and the change in value is also very gradual. A conservation form of Eq. 3.2 can be used to compute solutions with "shocks" or discontinuities, as a consequence of the Lax-Wendroff Theorem [87]. This theorem states that solutions to such conservative schemes (if possible) would converge to a weak solution of Eq. 3.2. Level sets also follow this scheme and thus allow shocks to propagate correctly.

We terminate our simulations based on externally applied halting criteria synthesized from the image data. These halting criteria include morphological comparisons with the target image, i.e. we use a saddle point curve of Euclidean distances in six-dimensional morphological feature space to terminate our simulation.

3.3 Results

We compare the accuracy of the representation of the coupled level set cellular model by comparing it with a Monte-Carlo-based simulation model in terms of the morphological features and computational complexity.

3.3.1 Performance Evaluation of the Coupled Level-Set-Cellular Model on the Basis of the Accuracy of Predicted SMG Morphology

To understand the efficacy of the predictive growth model, we compare it with a Monte-Carlo-based simulation model that works on the principle of energy minimization of the combined effective energy function (constructed as a Hamiltonian expression).

3.3.1.1 A Brief Overview of the GGH model

The Glazier-Graner-Hogeweg (GGH) model [75] is built upon the energy minimization-based Ising model [77], using imposed fluctuations via a Monte Carlo (MC) approach. An in-depth investigation analyzing the ranges of the cellular parameters for construction of a local GGH-based model of epithelial cleft formation is reported in Appendix A. The authors urge the reader to refer to Appendix 2 for further details about implementation of the GGH model. We also compared the ability of the coupled level set model to simulate cleft formation to that of the GGH model. To make a fair comparison between the two models, we included the dynamic cleft creation module, described in the Section 3.2.3.3 to our implementation of the GGH model.

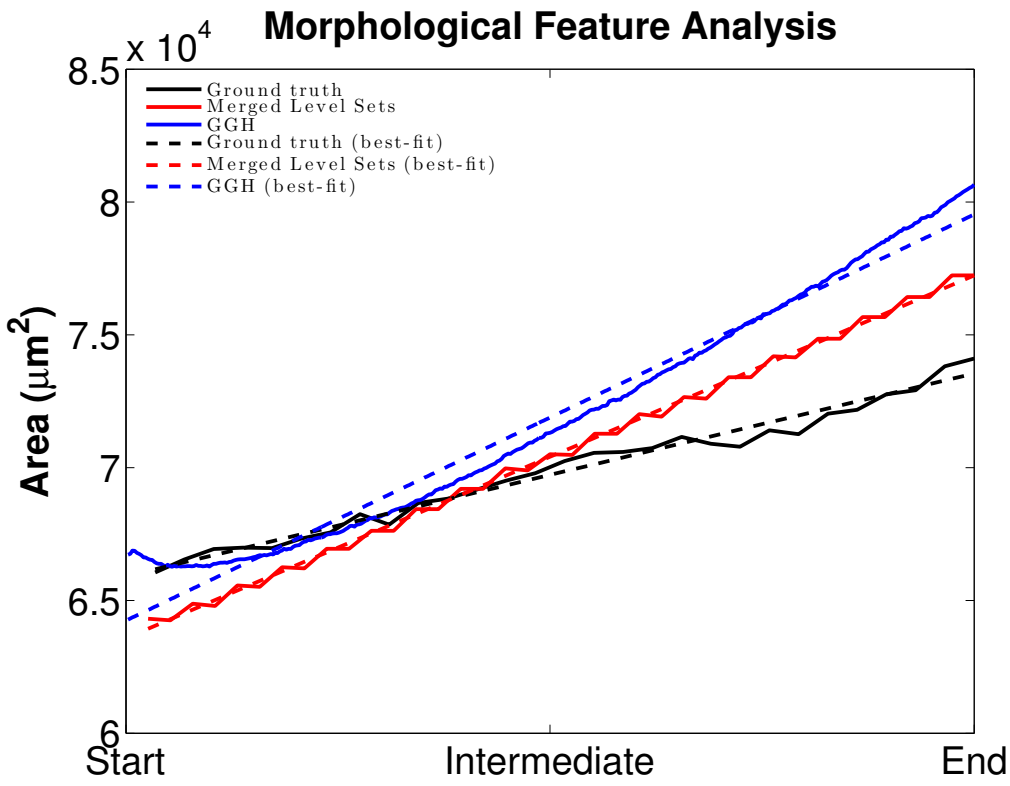
3.3.1.2 Comparison of the Coupled Level-Set-Cellular Model to the Ground Truth and the GGH model

The morphological feature-based comparison of the coupled level set celullar model with a quantitative analysis of the biological data set (ground truth) and the GGH model under the specific parameter set described in Appendix A for organ explants is shown in Fig. 3.2. All six features including *area*, *perimeter*, *elliptical variance*, *box-count dimension*, *solidity*, and *convexity* are shown in the figure. Although, both models are constructed by very different modeling techniques, one is based on interface evolution via level sets whereas the other minimizes a Hamiltonian formulation, our comparison is solely based on the final shape of the epithelial tissue produced by them. No comparison is done based on the outputs of the models in their original forms. The ground truth trends are displayed in black, the coupled-level-set-cellular-model's trends are displayed in red, and the GGH model's trends are displayed in blue. As observed in the ground truth, an increase in area and perimeter is seen in both models. The coupled model is able to replicate the increase in area and perimeter more effectively than GGH, and usually remains faithful to the increasing trend.

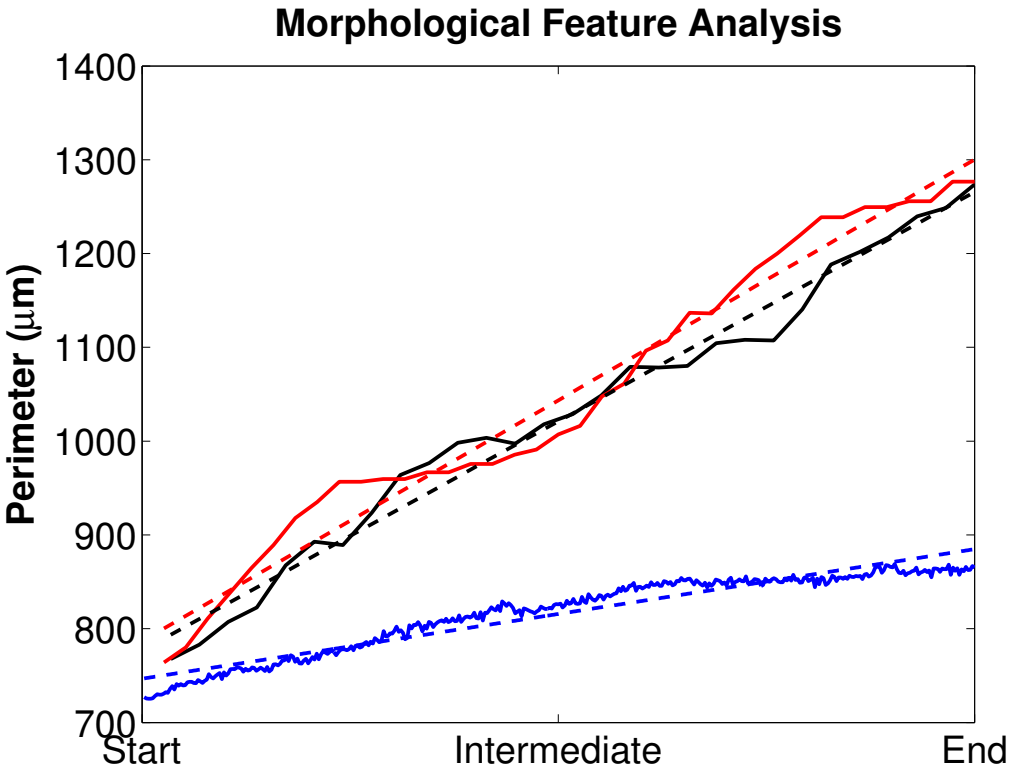
Although the range of values of elliptical variance is fairly small (10^{-2}), both models show a general increasing trend for this feature. In general, the coupled level

set cellular model has a higher box-count dimension than the GGH model. This could be because it tends to create more clefts than the GGH model, thereby creating more area of concavity. Solidity decreases with deepening of the clefts. While the coupled model more closely reproduces the appropriate trend, since the GGH model tends to grow the gland more circular, solidity is affected by this trait of the GGH model. Also, the clefts generated by GGH are in constant flux, appearing and disappearing from one MCS step to the next, and this may also be causing the solidity patterns to be modeled incorrectly. Convexity drops as the rate of increase of perimeter of the SMG increases at a faster rate than the perimeter of its convex hull. In keeping with the ground truth, both models show the appropriate decreasing trend. These plots illustrate that the coupled model is able to predict the growth of the clefts during early branching morphogenesis. Figure 3.3 displays target ground truth configuration, and sample terminal configurations for both models for the data set under investigation. As can be noticed from the terminal configurations, the coupled level set cellular model is able to produce *de novo* clefts and the final configuration is comparable to that produced by the GGH model. The authors urge the reader to refer to the Supplementary Movie (can be downloaded from `http://dsrc.rpi.edu/cellgraph/SMG_modeling/Supplementary_Video_V4.mp4`) for a better illustration of the working of the model.

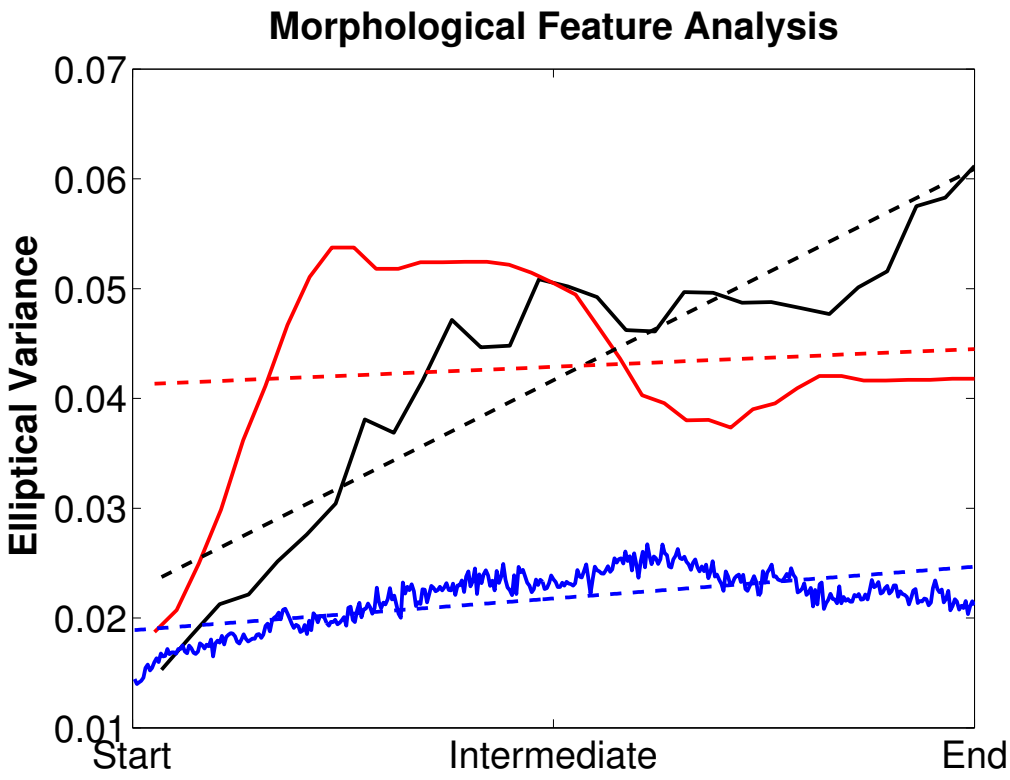
We also compared the time it takes each model to complete one simulation, which is comparable to approximately 3 hours of experimental data. In order to compare the computational time complexities of the two models, we ran 50 experiments on a 2.4 GHz Intel Core 2 Duo processor with 4GB RAM. The coupled level set cellular model was on average 22.79 times faster than the GGH model, taking an average of 2.17 min $\pm$ 0.26 sec to complete the experiment, whereas GGH took 49.47 min $\pm$ 34.21 sec. The large volume of data generated by the GGH (requiring substantial post-processing and disk storage), as well as the increased complexity from considering cellular-level detail makes it unsurprising that the coupled model takes less execution time.


Morphological Feature Analysis
x 10^4
8.5
8
7.5
7
6.5
6
Area (μm^2)
Ground truth
Merged Level Sets
GGH
Ground truth (best-fit)
Merged Level Sets (best-fit)
GGH (best-fit)
Start
Intermediate
End

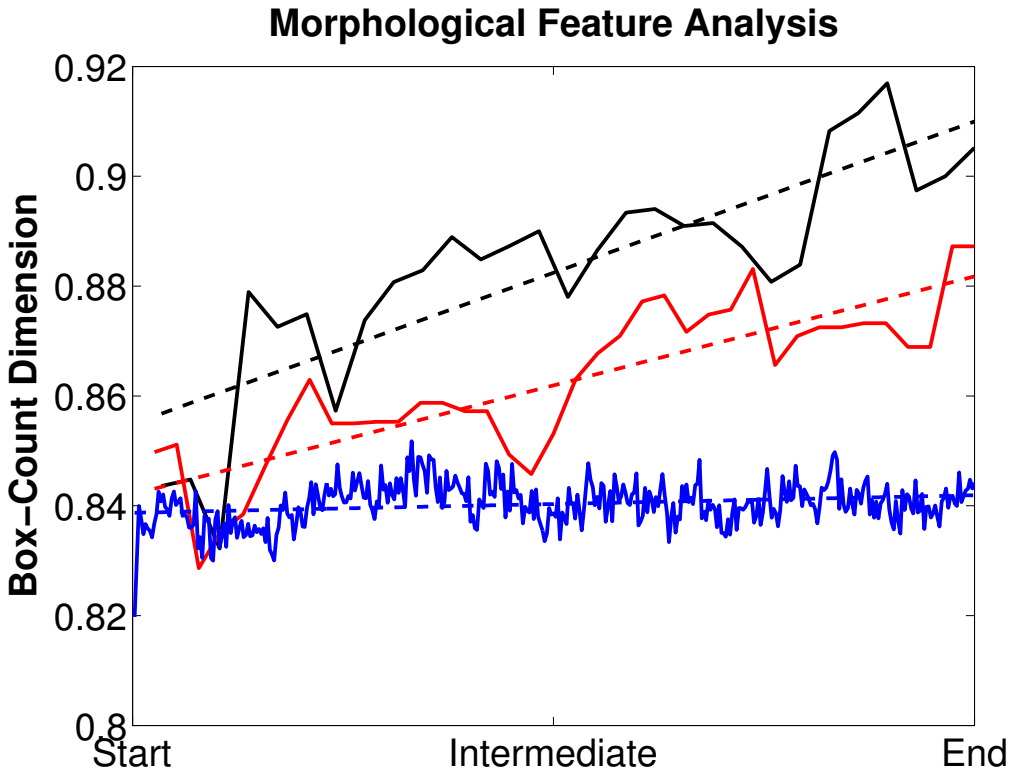
(a)


Morphological Feature Analysis
1400
1300
1200
1100
1000
900
800
700
Perimeter (μm)
Start
Intermediate
End

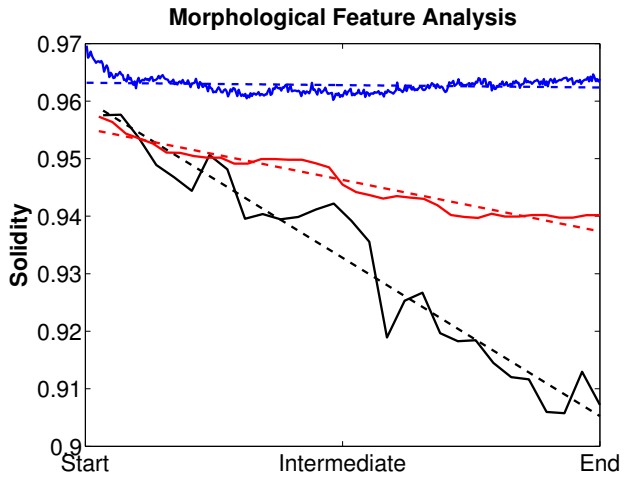
(b)


Morphological Feature Analysis
Elliptical Variance
0.07
0.06
0.05
0.04
0.03
0.02
0.01
Start
Intermediate
End

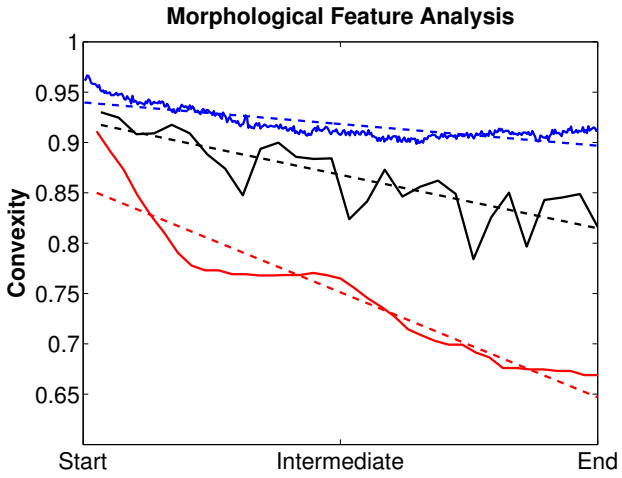
(c)


Morphological Feature Analysis
Box−Count Dimension
0.92
0.9
0.88
0.86
0.84
0.82
0.8
Start
Intermediate
End

(d)



(e)



(f)

Figure 3.2: Comparison of SMG morphological features between ground truth (black), coupled level set cellular model (red), and the GGH simulation (blue) for the particular data set. The features are (a) Area, (b) Perimeter (c) Elliptical Variance, (d) Box-count dimension, (e) Solidity, and (f) Convexity. "Start" refers to cleft initiation, "Intermediate" refers to mid cleft progression stage, and "End" refers to beginning of cleft termination. Area and perimeter display increasing trends and similar trends are also seen for the coupled model as well as the GGH simulation. Increasing trends are also observed for ground truth and the two models for both elliptical variance and box-count dimension. Decreasing trends are displayed by the ground truth and the two models for both solidity and convexity. For more detailed explanation of the feature trends, please refer to Section 3.3.1.2.

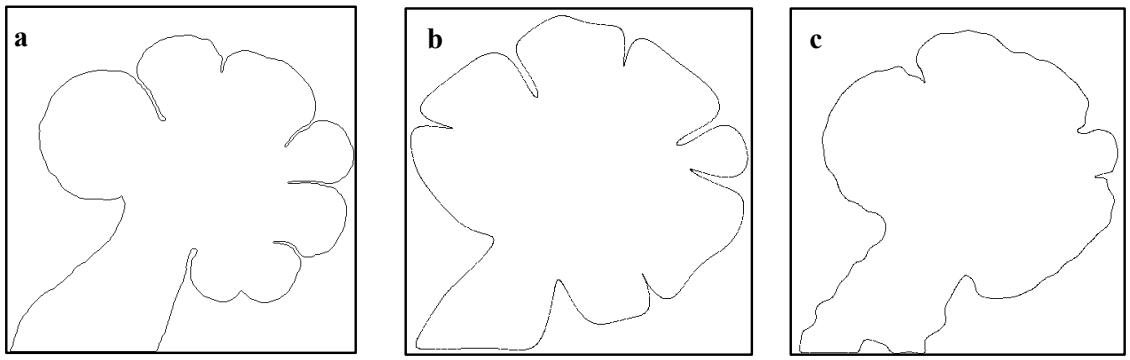


Figure 3.3: Target configuration of the ground truth data set, and the final configurations reached by the simulations for the coupled model and the GGH simulation. The target ground truth configuration is shown in (a), the coupled model's configuration is shown in (b), and the GGH model's configurations is shown in (c). The coupled model creates more clefts than the GGH model and this is one of the primary reasons that it has better quantitative agreement with the ground truth in regards to the morphological features.

3.4 Conclusions

In this paper, we establish a proof of concept for coupling two different models to predict time evolution of a complex biological process. In particular, we consider branching morphogenesis of mouse salivary gland and couple the level set method that captures the topological evolution while remaining agnostic to cellular organization and dynamics, with a discrete model that predicts cellular organization but is agnostic to shape evolution. The coupling is justified by the underlying biology: epithelial-mesenchymal interactions at the gland membrane during development. The level set receives as input the rate and direction of growth from the discrete model, while the shape information obtained from the level set algorithm constrains the growth and organization of the cells. We note that by merging these two models we introduced a multiscale modeling approach that functions at both the tissue and organ levels.

The coupled modeling is compared to another energy minimization-based modeling approach in the Glazier-Graner-Hogeweg model that is used for this particular application domain, and to the ground truth. Our results are promising - the coupled model produces comparable results (for some features, the prediction is closer

to the ground truth) but has a much faster (more than 20 times) execution time on the same computational platform. We believe that such couplings can be extended to other models and application domains; however, merging models should be driven by the underlying biological process.

CHAPTER 4
GRAPH-THEORETIC ANALYSIS OF EPILEPTIC SEIZURES ON SCALP EEG RECORDINGS

4.1 Introduction

Epilepsy is a common chronic neurological disorder that affects the nervous system and is typically characterized by the recurrent seizures caused by sudden bursts of electrical energy in the brain. For a successful treatment and understanding of the seizures, it is critical to characterize the physiological changes caused by the seizures.

Clinical evaluation, identification of epileptic seizures, and localization of epileptic seizures significantly rely on monitoring and analysis of long-term electroencephalographic (EEG) signals and benefit from intensive video-EEG monitoring. Large numbers of multi-channel EEG signals are visually analyzed by neurologists with a goal of understanding when and where the seizures start and how they propagate within the brain. However, there are two main disadvantages of visual analysis of EEG signals: it is time-consuming and prone to subjectivity. Therefore, automation of the detection of the underlying brain dynamics in EEG signals is significant in order to obtain fast and objective EEG analysis.

4.1.1 Related Work

A common approach in seizure recognition/detection and also in prediction is to extract information, e.g., features that can characterize seizure morphologies, from EEG recordings [88, 89, 90, 91, 92]. The procedure for feature extraction from multi-channel EEG data is often as follows: First, an EEG signal from a channel is divided into I time epochs (overlapping or non-overlapping) and then J features are extracted from each epoch. Consequently, a signal from a single channel can be

This chapter previously appeared as: Dhulekar N et al. (2014) Graph-theoretic analysis of epileptic seizures on scalp EEG recordings. *In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 155–163

represented as a matrix of size $I \times J$. A great deal of effort from different disciplines has been invested in exploring the features in order to capture characteristics of an epileptic seizure. These features include statistical complexity measures (e.g., fractal dimension, approximate entropy, Lyapunov exponents, etc.) as well as other features from time (e.g., higher-order statistics of the signal in time domain, Hjorth parameters, etc.) and frequency domains (e.g., spectral skewness, spectral entropy, etc.). A list of features used in characterization of epileptic seizure dynamics can be found in recent studies [90, 92, 91]. In both seizure detection and prediction, the approaches are similar in terms of extracting features, but the goals are different. In *seizure recognition/detection* the goal is to detect the seizure onset as accurately as possible. Therefore, in seizure detection we try to differentiate between ictal and non-ictal time epochs, where ictal period (defined as an EEG event rather than a clinical event) can be identified by neurologists with some level of subjectivity based on visual analysis of EEG signals. On the other hand, *seizure prediction* aims to predict seizures taking into consideration very early signs of an upcoming seizure. We therefore focus on the differentiation of pre-ictal and inter-ictal periods. A *pre-ictal period*, however, has neither strict boundaries nor a definition and many studies in the literature have used a variety of features ([93] and references therein) to define a pre-ictal period. Changes in these features occur from several minutes up to several hours before a seizure onset [94, 95, 96]. Therefore, different prediction horizons are often used to asses the performance of proposed models. In addition to the evaluation of the seizure prediction system using different prediction horizons, a seizure prediction technique should also have a precise temporal resolution since the variation in the prediction horizon may simply make the proposed technique impractical as discussed in [97].

In the literature, studies often use either multiple features from a single channel or a single feature from multiple channels since data construction and data analysis techniques are restricted to two dimensions. For instance, in [90], seizure dynamics are analyzed solely on a specific recording, which represents the characteristics of a seizure well. Then the performance of various features from different domains on that particular signal is analyzed simultaneously. On the other hand, [92] an-

alyzes multi-channel EEG data but assesses the performance of each feature one at a time. Furthermore, different studies extract different features and employ different algorithms to distinguish between seizure and non-seizure periods (e.g., [98] and references therein), which makes it difficult to compare the performance of features. An approach capable of simultaneously analyzing features would enable the performance comparison of the features on the same data using the same classifier. Simultaneous analysis of features is also important because it may consider linear or non-linear combinations of features. While a single feature may not be very effective in discriminating between epileptic periods, combinations of several features may well be [99].

We note that all techniques reviewed above focus on the spikes and sharp waves in each electrode via temporal, frequency, or wavelet domain analysis [100, 101, 102, 103]. However, the underlying mechanism may also be captured via analyzing the signals from different electrodes jointly such as multiway analysis techniques [104]. Multi-modality of the data enables one to include as many features as possible in the analysis, and combines the seizure recognition or prediction power of several features. Furthermore, since we can analyze multiple channels simultaneously, we do not have to make any prior assumption regarding the seizure localization and therefore select channels.

Multiway analysis techniques have been previously applied successfully in neuroscience. In [105], EEG data and data collected through experiments with different doses of a drug are arranged as a six-way array with modes: EEG, patients, doses, conditions, time, and channels. The analysis of the six-way dataset demonstrates that significant information is successfully extracted from a complex drug dataset by a multi-linear model rather than two-way models such as *Principal Component Analysis* (PCA) [106, 107]. Multiway models have become more popular in neuroscience with the idea of decomposing EEG data into space-time-frequency components [108]. The three-way array constructed from multi-channel EEG data in [108] with modes $time\ samples \times frequency \times channels$ is analyzed using a multi-linear component model called *Parallel Factor Analysis* model
(PARAFAC) [109]. Components extracted by PARAFAC are observed to demon-

strate the temporal, spectral and spatial signatures of EEG. In another study [110], PARAFAC models with non-negativity constraints are used on *Event-Related Potentials* (ERP) to find the underlying structure of brain dynamics. In addition to the applications of multi-linear component models, multi-linear regression models have also been previously employed in neuroscience, e.g., in [111] for extracting the connection between EEG and *functional Magnetic Resonance Imaging* (fMRI) recordings.

In [112, 104], multi-channel EEG data is converted to a third order tensor $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ by applying continuous wavelet transformation on the signal recorded at each channel. In this tensor, x_{ijk} represents the square of the absolute value of wavelet coefficient at the ith channel for jth scale and the kth time sample; in other words the power of the wavelet transformed data. Analysis of the tensor by PARAFAC [109] algorithm revealed the seizure localization.

Recently there has emerged a new direction in analyzing EEG recordings by building a *synchronization graph* that enables characterization of the pairwise correlations between electrodes using graph theoretical features over time [113, 114]. In the spatio-temporal EEG graphs, *nodes* (vertices) represent the EEG channels and the *edges* (links) represent the level of neuronal synchronization between the different regions of the brain. This approach has been exploited in the analysis of various neuropsychiatric diseases including schizophrenia, autism, dementia, and epilepsy [114]. Within epilepsy research, evolution of certain graph features over time revealed better understanding of the interactions of the brain regions and the seizures. For instance, Schindler et al. analyzed the change in path lengths and clustering coefficients to highlight the evolution of seizures on epileptic patients [115], Kramer et al. considered the evolution of local graph features including betweenness centrality to explain the coupling of brain signals at seizure onset [116], and Douw et al. recently showed epilepsy in glioma patients was attributed to the theta band activity in the brain [117]. In [118], authors independently suggest a similar approach that combines tensor decompositions with graph theory.

In this paper, we study how to build synchronization graphs in depth and present graph mining results to detect epileptic seizure both in temporal and spatial

domains. The organization of the paper is as follows: in Section 4.2, we describe the epileptic EEG dataset, methodology to construct time-evolving EEG synchronization graphs, and the computation of global and local features on these graphs. In Section 4.3, we present results for seizure detection and seizure localization using the features described in Section 4.2. Finally, we provide an overview, discuss the results, and list extensions to this study, in Section 4.4.

4.2 Methodology

4.2.1 Epileptic EEG Dataset

Our dataset consists of scalp EEG recordings of 34 seizures from 15 patients. All the patients were evaluated with scalp video-EEG monitoring in the international 10-20 system (as described in [119]), magnetic resonance imaging (MRI), fMRI for language localization, and position emission tomography (PET). The data was collected in the Neurophysiology Laboratory of Yeditepe University Hospital in Istanbul, Turkey. All the patients had *Hippocampal Sclerosis* (HS) except one (IY) who suffered from *Cortical Dysplasia* (CD). After selective amygdalohippocampectomy, all the patients were seizure free. The patient information is provided in Table 4.1. For 10 patients, the seizure would onset from the left (L), whereas for 5 patients the seizure would onset from the right (R).

One patient has one 30 minute recording, two patients have two 30 minute recordings, one patient has three 30 minute recordings, three patients have single 60 minute recordings, three patients have two 60 minute recordings, four patients have three 60 minute recordings, and one patient has five 60 minute recordings. The recordings include sufficient pre-ictal and post-ictal periods for the analysis. Two of the electrodes (A_1 and A_2) were unused and C_z electrode was used for referential montage that yielded 18-channel EEG recordings. Dr. Aykut-Bingöl's team diagnosed the initiation and the termination of each seizure and reported these periods as the ground truth for our analysis. An example of such a recording can be found in Fig. 2 in [120]. Seizures were 97 seconds long on average and their standard deviation was 121 seconds.

Table 4.1: Patient Types. Almost all the patients exhibited hippocampal sclerosis (HS). There are two types of lateralizations in HS: left (L) and right (R). One patient (IY) exhibited cortical dysplasia (CD).

Patient	Pathology	Lateralization
Patient_1	CD	R
Patient_2	HS	R
Patient_3	HS	R
Patient_4	HS	R
Patient_5	HS	R
Patient_6	HS	L
Patient_7	HS	L
Patient_8	HS	L
Patient_9	HS	L
Patient_10	HS	L
Patient_11	HS	L
Patient_12	HS	L
Patient_13	HS	L
Patient_14	HS	L
Patient_15	HS	L

4.2.2 Construction of EEG Synchronization Graphs

Let $\mathbf{X}[i, m]$ denote the recorded EEG signal, where $i \in \{1, \ldots, 18\}$ represents the index for the ith electrode and $m \in [1, \ldots, f_s \times M]$ represents the time index, f_s represents the sampling frequency, and M is the duration of the recording in seconds. Sampling frequency, f_s, is either 200 Hz or 400 Hz. We first decompose the signal in each electrode into the five traditional frequency bands δ (0.5–3.5 Hz), θ (3.5–7.5 Hz), α (7.5–12.5 Hz), β (12.5–30 Hz), and γ (>30 Hz) through appropriate digital band-pass filters $H_f(\omega)$, where f represents the index for the frequency band. Next, we constructed epochs of equal lengths with 20% overlap between the preceding and following epochs. The number of epochs N is equal to $1.25M/L$, where L is the duration of the epoch in the same time units.

Given the nature of these spatio-temporal recordings, we consider constructing *time-evolving EEG Synchronization Graphs* on the EEG datasets. A graph is constructed for each frequency band and epoch. The *nodes* represent the EEG elec-

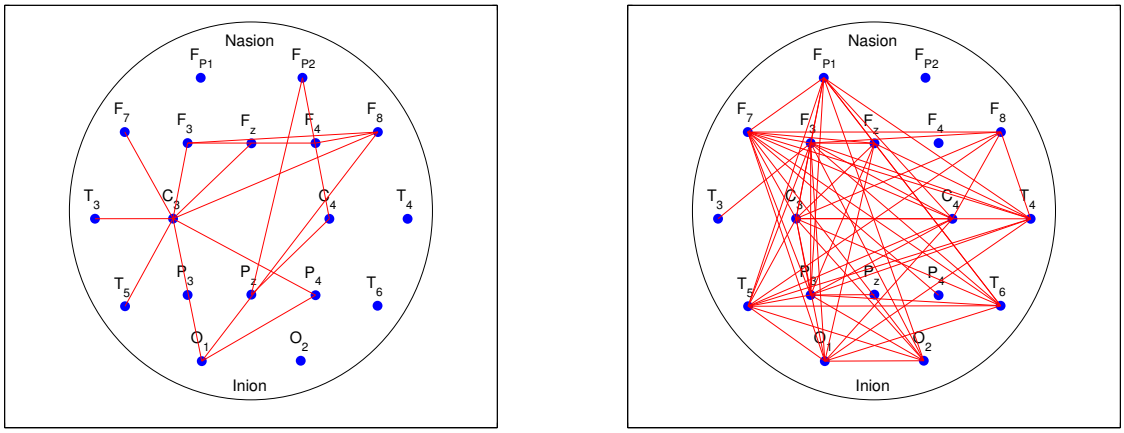
trodes and the *edges* represent a closeness relationship between the nodes in a given epoch. The main goal sought in the time-evolving EEG synchronization graphs is the ability to sense both the spatial and temporal changes in the network, yielding measures of change for the mining.

The selection of epoch length is significant in both the time-domain and frequency-domain analysis techniques.
While longer epochs provide better frequency resolution, far too lengthy epochs may not be stationary [121] and not allow rapid detections of changes in the network [122, 123]. Therefore, shorter epoch lengths are generally preferred. However, extremely short epoch lengths may not allow the generation of meaningful graphs for mining. Therefore, a secondary goal of this research is to analyze and identify the effects of epoch lengths on the time-evolving EEG Synchronization Graphs. Note that if a wavelet-domain based measure is employed for the construction of the time-evolving graphs, these effects are reduced or eliminated. However, as the range of explored frequencies must be arbitrarily defined prior to the wavelet analysis, wavelet domain techniques do not allow the simultaneous exploration of the whole range of the EEG frequency spectrum.

For purposes of illustration, we construct simple time-evolving graphs on the EEG recording shown in Fig. 4.1. The graph edges are established based on the pair-wise relationships between the epoch data. Specifically, an edge between two distinct nodes i and k, where $i, k \in \{1, \ldots, 18\}$, and $i \neq k$, is established if the pair-wise distance for the nth epoch, $d^n_{i,k}$, is less than the threshold τ. A parametric search is performed to find optimal values for the threshold τ. Further information about the search performed on various parameters to determine optimal values is provided in Section 4.3.

Several synchronization measures have been proposed as plausible options for $d^n_{i,k}$. We consider 3 different measures here, namely Correlation Coefficient [124], Phase Lag Index [125], and Phase Locking Value [126]. The three measures are defined as follows:

$$CC_{i,k}(n) = \frac{1}{Lf_s} \frac{\left(\mathbf{x}^n_i - \bar{\mathbf{x}}^n_i\right)\left(\mathbf{x}^n_k - \bar{\mathbf{x}}^n_k\right)^T}{\sigma_{\mathbf{x}^n_i}\sigma_{\mathbf{x}^n_k}} \tag{4.1}$$



(a) Pre-ictal

(b) Ictal

(c) Post-Ictal

Figure 4.1: Sample EEG Synchronization Graphs for pre-ictal, ictal, and post-ictal epochs. It is clearly seen that the ictal period has more coherence between different regions of the brain.

$\bar{\mathbf{x}}_i^n$ and $\bar{\mathbf{x}}_k^n$ are the mean values of $\mathbf{x}_i^n$ and $\mathbf{x}_k^n$, respectively, $\sigma_{\mathbf{x}_i^n}$ is the standard deviation of $\mathbf{x}_i^n$,

$$PLV_{i,k}(n) = \frac{1}{Lf_s}\left|\sum_{m=1}^{Lf_s}\exp\left(j\left(\boldsymbol{\phi}_i^n(m)-\boldsymbol{\phi}_k^n(m)\right)\right)\right| \tag{4.2}$$

$$PLI_{i,k}(n) = \frac{1}{Lf_s}\left|\sum_{m=1}^{Lf_s}\operatorname{sgn}\left(\boldsymbol{\phi}_i^n(m)-\boldsymbol{\phi}_k^n(m)\right)\right| \tag{4.3}$$

where $\boldsymbol{\phi}_i^n = \arctan(\frac{\hat{\mathbf{x}}_i^n}{\mathbf{x}_i^n})$ is the angle of the Hilbert transform $\hat{\mathbf{x}}_i^n$ of the signal $\mathbf{x}_i^n$. For

correlation coefficient (Eq. 4.1), we define pair-wise distance as $d^n_{i,k} = 1 - CC_{i,k}(n)$. Note that smaller threshold values seek higher correlation between the electrodes, therefore yielding sparser graphs. Similarly, higher threshold values would establish an edge even if there is little correlation between the data, thereby yielding denser graphs.

4.2.3 Feature Extraction from EEG Synchronization Graphs

We extract 27 features from the EEG graph for each epoch. These features quantify the compactness, clusteredness, and uniformity of the graph. *Compactness* of a graph is measured by features such as *average eccentricity*, *diameter*, *radius*, and *number of central points* that are based on the path distances between the nodes within the graph. The *clusteredness* of a graph is measured by features such as *average clustering coefficient* and *number of connected components* that are based on how connected are the nodes and their neighbors with respect to each other. In addition, we also extract *spectral graph features* such as *spectral radius* and *spectral gap* using the eigenvalues of the adjacency, Laplacian, and normalized Laplacian matrices, that reveal additional clusteredness and compactness properties about the graphs. A complete list of the features used in this work and their definitions is listed in Tables 4.2 – 4.3.

4.2.4 Mining Global Graph Features for Temporal Clustering

After computing the global graph features for each epoch, we build a feature matrix of dimensionality $N \times 27$, where N is the number of epochs, for the sub-bands of the recording. In order to learn the structural differences within the temporal evolution of the epileptic seizures, we decided to separate the epochs into different clusters using unsupervised learning. We employed k-means clustering due to its successful record in bioinformatics applications [73]. *k-means clustering* is an iterative unsupervised learning method that partitions the samples into k clusters based on their attributes and traits. The method first chooses k random samples from the set and assigns them as the initial cluster centroids. Next, the remaining samples are assigned to the cluster whose centroid is the closest in terms of Euclidean distance. Finally, the cluster centroids are recomputed by averaging over all samples

Table 4.2: Description of EEG global graph features.

Index	Feature Name	Description
1	Average Degree	Average number of edges per node
2	Clustering Coefficient C	Average of the ratio of the links a node's neighbors have in between to the total number that can possibly exist
3	Clustering Coefficient D	Same as feature 2 with node itself added to both numerator and denominator
4	Average Eccentricity	Average of node eccentricities, where the *eccentricity* of a node is the maximum distance from it to any other node in the graph
5	Diameter of graph	Maximum of node eccentricities
6	Radius of graph	Minimum of node eccentricities
7	Average Path Length	Average number of hops along the shortest paths for all possible pairs of nodes
8	Giant Connected Component Ratio	Ratio between the number of nodes in the largest connected component in the graph and total the number of nodes
9	Number of Connected Components	Number of clusters in the graph excluding the isolated nodes
10	Average Connected Component Size	Average number of nodes per connected component
11	% of Isolated Points	% of the isolated nodes in the graph, where an *isolated node* has a degree of 0
12	% of End Points	% of endpoints in the graph, where an *endpoint* has a degree of 1
13	% of Central Points	% of nodes within the graph whose eccentricity is equal to the graph radius
14	Number of Edges	Number of edges between all nodes in the graph

Table 4.3: Description of EEG global spectral features.

Index	Feature Name	Description
15	Spectral Radius	Largest eigenvalue of the adjacency matrix
16	Adjacency Second Largest Eigenvalue	Second largest eigenvalue
17	Adjacency Trace	Sum of the adjacency matrix eigenvalues
18	Adjacency Energy	Sum of the square of adjacency matrix eigenvalues
19	Spectral Gap	Difference between the magnitudes of the two largest eigenvalues
20	Laplacian Trace	Sum of the Laplacian matrix eigenvalues
21	Laplacian Energy	Sum of the square of Laplacian matrix eigenvalues
22	Normalized Laplacian Number of 0's	Number of eigenvalues of the normalized Laplacian matrix that are 0
23	Normalized Laplacian Number of 1's	Number of eigenvalues of the normalized Laplacian matrix that are 1
24	Normalized Laplacian Number of 2's	Number of eigenvalues of the normalized Laplacian matrix that are 2
25	Normalized Laplacian Lower Slope	The sorted slopes of the line for the eigenvalues that are between 0 and 1
26	Normalized Laplacian Upper Slope	The sorted slope of the line for the eigenvalues that are between 1 and 2
27	Normalized Laplacian Trace	Sum of the normalized Laplacian matrix eigenvalues

within the cluster and the process is repeated until the centroids do not change and within-cluster residual sum of squares is minimized. Note that if significantly more recordings for each patient are available, supervised machine learning techniques will be more suitable to capture these differences. For the purpose of this study, we use k-means in an unsupervised setting, only using the ground truth labeling provided by the doctors to validate our clustering results.

4.2.5 Spatial Analysis by Mining Local Graph Features

The analysis of the global graph features indicated a critical change in their values during the transitionary period from pre-ictal to ictal, and from ictal to post-ictal. As a consequence of this outcome, we studied the change in the local graph features for individual nodes (electrodes) to determine whether we could localize the seizure to a particular part of the brain, or more specifically to a particular electrode. The intuition driving this hypothesis is that the electrode closest to the part of the brain in which the seizure onsets will record this change in brain activity before the other electrodes. The local graph features for the particular electrode will then indicate this sudden departure from regular behavior. We can verify our results based on the ground truth. To aid us in this objective, we utilize a Tucker3 tensor decomposition approach [127].

We construct the three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ with modes: time samples, local graph features, and electrodes. We model $\underline{\mathbf{X}}$ using a (P, Q, R)-component Tucker3 model on a three-way array given by

$$x_{ijk} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{pqr}\, a_{ip}\, b_{jq}\, c_{kr} + e_{ijk}\ ,$$

where $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ are the component matrices corresponding to the first, second, and third modes, respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ contains the residuals. The illustration of a Tucker3 model on a three-way array is given in Fig. 4.2.

For each patient, we construct the three-way array $\underline{\mathbf{X}}$ with the modes: time samples within the seizure period, local graph features computed for each electrode, and the number of electrodes. We compute 8 local graph features listed in Table 4.4. We run the Tucker3 decomposition on this tensor.

We examine the plotting of the residual sum of squares vs. hoteling T-squared values [128] in the electrodes mode which show the outliers in the upper right corner (as shown in Fig. 4.3). These outliers are the electrodes that we consider as progenitors of the seizure, and use their location to determine the side of the brain where the seizure begins. The electrodes present towards the upper left or bottom left

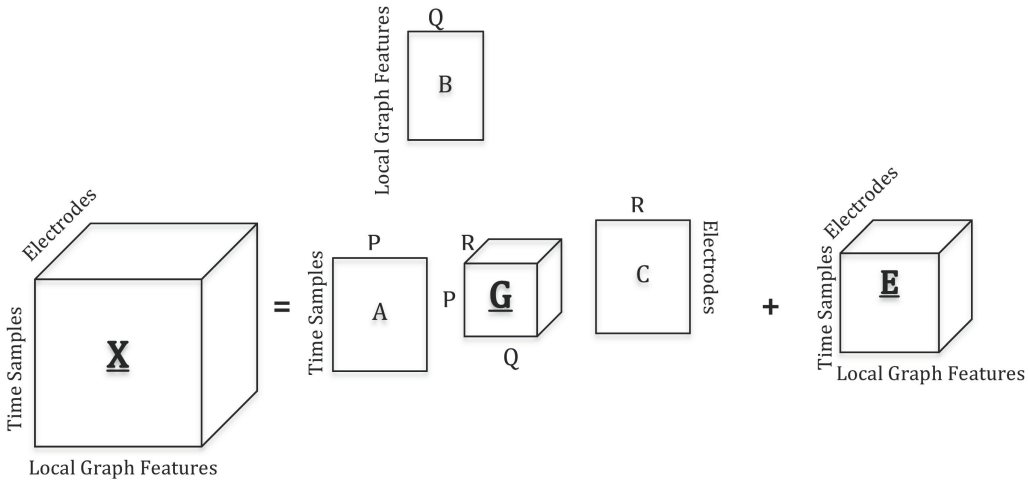


Figure 4.2: Tucker3 tensor decomposition. (P,Q,R)-component Tucker3 model, with a three-way array $\underline{\mathbf{X}} \in \mathbb{R}^{I \times J \times K}$ is modeled with component matrices $\mathbf{A} \in \mathbb{R}^{I \times P}$, $\mathbf{B} \in \mathbb{R}^{J \times Q}$, and $\mathbf{C} \in \mathbb{R}^{K \times R}$ in the time samples (first), local graph features (second), and electrode (third) modes, respectively. $\underline{\mathbf{G}} \in \mathbb{R}^{P \times Q \times R}$ is the core array and $\underline{\mathbf{E}} \in \mathbb{R}^{I \times J \times K}$ contains the residuals.

corners are not significant. In the figure, electrodes are named using the following nomenclature: "[L—C—R][1–8]_identifier", where "L","C", and "R" represent left, center, and right sides of the brain, respectively. The electrodes are numbered from 1–8 and each electrode is given a unique identifier.

4.3 Experimental Results

We conducted a parametric search to determine the effect of parameters to capture the characteristics of the seizures. We tested four different epoch lengths that vary from a relatively short epoch length of 5 seconds to a longer duration of 12.5 seconds with 2.5 second increments. We tested five different threshold (τ) values for each synchronization cost function tested in our methodology. After obtaining the EEG graphs for each epoch in a particular frequency band in the recording, we extracted the graph features and constructed the data matrix to be clustered. In order to reduce the scale differences among the features, the data was normalized so that the features have zero mean and unit variance across the epochs. We tested four different l values ranging between two and five considering pre-ictal, ictal, post-ictal, as well as the transitional periods between these stages.

Table 4.4: Description of EEG local graph features

Index	Feature Name	Description
1	Degree of node	Number of edges incident to the node
2	Clustering Coefficient C	The ratio of the edges between the node's neighbors to the total number that can possibly exist
3	Clustering Coefficient D	The ratio of the edges between the node's neighbors and the node itself to the total number that can possibly exist
4	Clustering Coefficient E	Same as clustering coefficient C but discounting the isolated nodes
5	Eccentricity	Maximum shortest path length from the node to any other node in the graph
6	Eccentricity 90%	Maximum shortest path length from the node to 90% of the reachable nodes in the graph
7	Closeness Centrality	Sum of the distances of the node from all other nodes in the graph determines its importance
8	Betweenness Centrality	Measures how many shortest paths between all pairs of nodes include the node

The result of k-means clustering is typically biased by the initial cluster centroids that are chosen randomly. In order to obtain unbiased results, we repeated the clustering 100 times and choose the clusters that were obtained most frequently as the final result. The resulting clusters are evaluated using F-measure and cluster entropies. *F-measure* is a function of precision (P) and recall (R) of a particular class type after retrieval and is given by $F = 2\frac{PR}{P+R}$. The *entropy* of a cluster is given by $E = -\sum_{k=1}^{K} \Pr(c_k) \log_2 \Pr(c_k)$, where $Pr(c_k)$ is the probability of a time epoch belonging to cluster c_k. For an ideal clustering, we want F-measure to be close to 1 and entropy to be close to 0. Therefore, we combine these two metrics into a single metric as F/E where uniform clusters yield higher values. A total of 75 parameters are tested for each recording and we choose the clusters that yield

the highest value as the representative result for that recording. Figure 4.4 displays clusters for a particular EEG recording. As can be noticed from the figure, the seizure time points are present within a single cluster, namely cluster 5 (marked in red), whereas the pre-ictal and post-ictal time points belong to various clusters, namely clusters 1–4.

4.3.1 Results and Interpretations

In the following 2 sections, we present results for the seizure detection and seizure localization algorithms. Please note that in two patients with identifiers ATU_1 and ATU_2, and ABA_1, we could not detect and localize the seizures well. In these patients, we reviewed the seizure samples again and then we realized that we did not get the exact seizure periods. This result confirmed the veracity of our data analysis.

4.3.1.1 Seizure Detection

We found that our seizure detection algorithm was able to correctly detect the seizure in 30 of the 34 patient recordings for a success rate of 88.24%. If we discount the recordings with identifiers ATU_1 and ATU_2, which had issues even with external video EEG information, our accuracy increases to 93.75%.

We also present the seizure detection results based on 4 criteria:

1. Comparison of multiple frequency bands to a single frequency band consisting of the entire signal.
   We found that considering the top 5 F-score values for each patient recording, multiple frequency bands outperformed single frequency band in 72.35% of the patient recordings.

2. The frequency band with the most consistently high F-score values.
   The θ frequency band had the best F-scores in 34.71% of the patient recordings.

3. The synchronization metric with the most consistently high F-score values.
   The Phase Lag Index synchronization metric had the best F-scores in 51.18% of the patient recordings.

4. The epoch length with the most consistently high F-score values.

   Epoch length of 12.5 seconds had the best F-scores in 45.29% of the patient recordings. As mentioned in Section 4.2.2, determining the optimal epoch length is crucial in constructing meaningful graphs on which mining techniques can be applied.

4.3.1.2 Seizure Localization

Table 4.5 presents the patient pathologies with the results of seizure localization. A "+" sign indicates correctly identified seizure, whereas a "−" sign indicates an incorrectly identified seizure. We were able to correctly localize the seizure in 26 of the 34 recordings leading to a 76.47% success rate. If we discount the recordings with identifiers ATU_1 and ATU_2, which had issues even with additional video EEG information, our accuracy increases to 81.25%. We recognized that the clustering coefficient synchronization metric was responsible for almost all the incorrect localizations.

4.4 Discussion and Conclusion

Synchronization in the firing of neurons is one of the characteristics of epileptic seizure. In this work we modeled this phenomena by building synchronization graphs, as a method to quantify the correspondence among electrodes recording EEG signals. The electrodes form the nodes in the graph and an edge is inserted between these electrodes when their pair-wise synchronization is greater than a threshold. We extracted graph features computed over the entire graph (global features), and for individual nodes (local features), and used these features to analyze the data in the feature space to detect seizure localization in both spatial and temporal domains. We identified the free parameters in this model and ran an extensive parametric search to determine the best possible configurations. When evaluated against the ground truth provided by the medical team, our results were found to be encouraging, but have scope for further improvement. One possible improvement would be to extend the feature set to include more graph features and then run a feature selection algorithm.

Table 4.5: Seizure localization results. Table lists patient identifier, pathologies, and the result of the seizure localization. A + sign indicates a correctly localized seizure, whereas − sign indicates an incorrectly localized seizure.

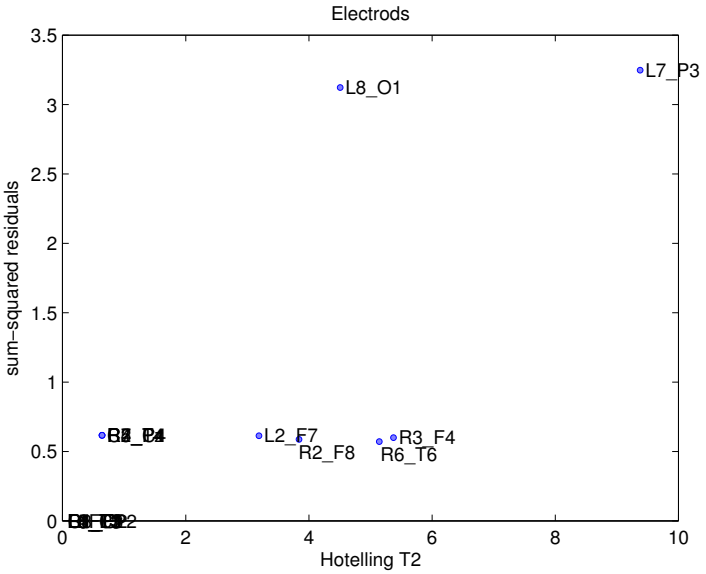
Patients	Pathology and Lateralization	Seizure Localization
Patient_1	CD; R	+ -
Patient_2	HS R	+
Patient_3	HS; R	+ +
Patient_4	HS; R	+ + +
Patient_5	HS; R	+ + + + -
Patient_6	HS; L	+
Patient_7	HS; L	+
Patient_8	HS; L	-
Patient_9	HS; L	+ -
Patient_10	HS; L	- -
Patient_11	HS; L	+ -
Patient_12	HS; L	none + -
Patient_13	HS; L	+ + +
Patient_14	HS; L	+ + +
Patient_15	HS; L	+ + +

As we mentioned in Section 4.2.1, patients considered in this study had undergone surgery thus allowing us to precisely determine the location of the seizure origin. In our future work we will expand our data to include patients without surgical treatment.

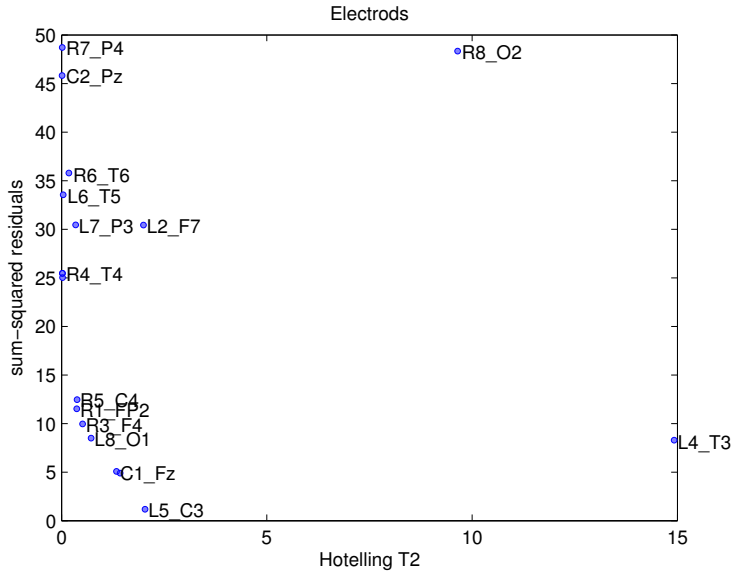
The time-evolving graphs can be enriched in two novel ways. First, given the time evolution of the graphs and uncertainty on the optimum epoch length, the epoch lengths can be chosen stochastically or adaptively to create different time scales. Given that the nodes are constant (i.e. electrodes or sensors remain the same), we can use a node as the common point of reference to integrate the data/knowledge at different scales. In other words we can use node/vertex based augmentation across different time scales. The information/data from different scales can also have different weights (importance). For example, if we are going to use global features to represent the data, weighted averages can be utilized. Assigning optimum weights to scales is another possible research problem that can be addressed as an extension to this work. Second, we can augment the edges with the "similarity vector" between a pair of nodes. This vector contains several "node" features extracted from the "signal" (not from the graphs) and each feature may have a different weight as well.

Acknowledgments

This work was supported in part by National Institutes of Health grants RO1 EB008016 and R01 DE019244. The authors would also like to acknowledge valuable inputs and discussions with Dr. Metin Erturkler and Selim Orhan.



(a) NT_3



(b) FZE_4

Figure 4.3: Temporal seizure localization. Figure shows two examples of seizure localization, for NT_3 the seizure onsets from the left, whereas for FZE_4 the seizure onsets from the right. In the case of NT_3, electrode P3 is chosen as the likely location for seizure onset - this electrode is on the left side of the brain. In the case of FZE_4, electrode O2 is chosen as the likely location for seizure onset - this electrode is on the right side of the brain.

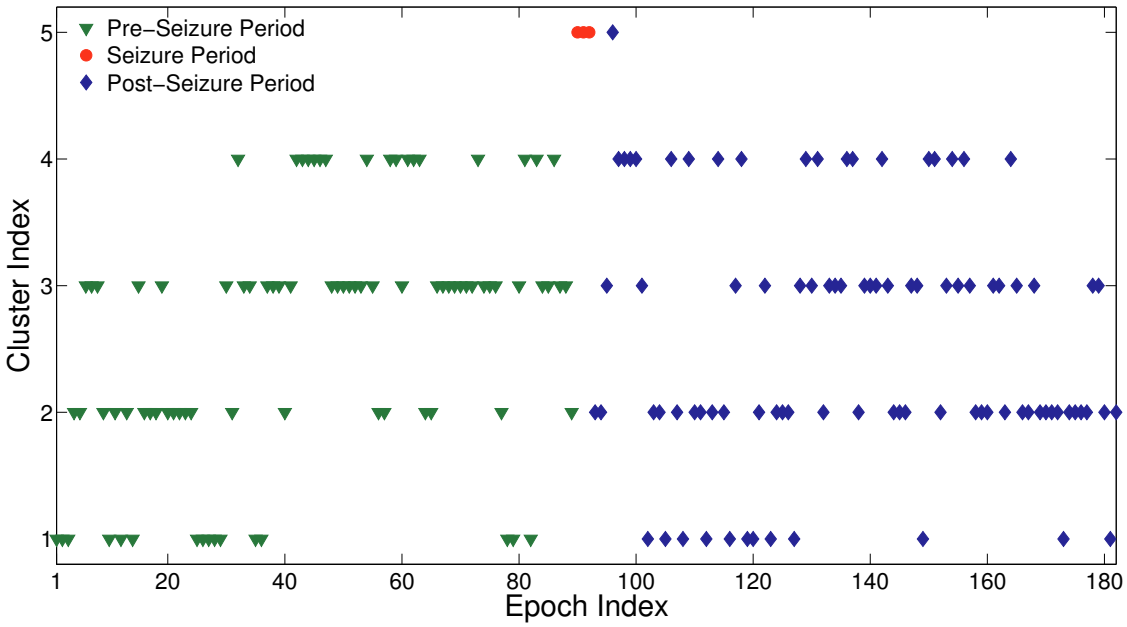


Figure 4.4: Clusters for an EEG recording. The seizure time points are all within one cluster (5) and are marked in red. The pre-ictal time points are marked in green, and the post-ictal time points are marked in blue.

CHAPTER 5
SEIZURE PREDICTION BY GRAPH MINING, TRANSFER LEARNING, AND TRANSFORMATION LEARNING

5.1 Introduction

Epilepsy is one of the most common disorders of the central nervous system characterized by recurring seizures. An epileptic seizure is described by abnormally excessive or synchronous neuronal activity in the brain [129]. Epilepsy patients show no pathological signs of the disease during inter-seizure periods, however, the uncertainty with regards to the onset of the next seizure deeply affects the lives of the patients.

Seizure prediction refers to predicting the onset of epileptic seizures by analyzing electroencephalographic (EEG) recordings without any apriori knowledge of the exact temporal location of the seizure [130]. A method with the capacity to successfully predict the occurrence of an epileptic seizure would make it possible for the patient to be administered therapeutic treatments thereby alleviating the pain [131]. Many approaches have been suggested as possible seizure prediction algorithms, with modest levels of success. The first attempt at a seizure prediction algorithm was made by Viglione and Walsh in 1975 [132] and investigated spectral components and properties of EEG data. This was followed in the 80s decade by different groups attempting to apply linear approaches, in particular autoregressive modeling, to seizure prediction [133, 134, 135]. Moving forward 20 years, Mormann et al. posed the question whether characteristic features can be extracted from the continuous EEG signal that are predictive of an impending seizure [136]. In 2002, the First International Workshop on Seizure Prediction [93] was conducted to bring together experts from a wide range of backgrounds with the common goal of improv-

This chapter is in Press: Dhulekar N, Nambirajan S, Oztan B, and Yener B (2015) Seizure prediction by graph mining, transfer learning, and transformation learning. *In Proceedings of the 11th International Conference on Machine Learning and Data Mining (MLDM).*

ing the understanding of seizures, and thus advancing the current state of seizure prediction algorithms on a joint data set [93].

Most general approaches to the seizure prediction problem share several common steps including (i) processing of multichannel EEG signals, (ii) discretization of the time series into fixed-size overlapping windows called epochs, (iii) extraction of frequency bands to analyze the signal in frequency and/or time domains using techniques such as wavelength transformation [104], (iv) extraction of linear and non-linear features from the signal or its transformations; these features can be univariate, computed on each EEG channel separately, or multivariate, computed between two or more EEG channels, and (v) learning a model of the seizure statistics given the features by using supervised machine learning techniques such as Artificial Neural Networks [137, 138, 139, 140, 141], or Support Vector Machines [142, 143]. A thorough survey of the various linear and non-linear features can be found in [144, 145, 91]. These features are usually calculated over epochs of predetermined time duration (around 20 seconds) via a moving window analysis. It has been found that univariate features, such as Lyapunov exponents, correlation dimension, and Hjorth parameters, calculated from the EEG recordings performed poorly as compared to bivariate and multivariate features [96, 130, 146, 147, 92, 95, 94]. This is understandable given that the seizure spreads to all the electrodes, whereas not all electrical activity in the brain may result in the onset of a seizure - it might be a localized discharge at a certain electrode. Although it has been shown that univariate features are less significant for seizure prediction, the importance of non-linear features over linear features is not quite as straightforward. It has recently been observed that non-linear techniques might not enhance the performance of the seizure prediction algorithm considerably over linear techniques, and also have considerable limitations with respect to computational complexity and description of epileptic events [148, 149, 150, 151, 152].

A phase-locking bivariate measure, which captures brainwave synchronization patterns, has been shown to be important in differentiating interictal from preictal states [153, 94]. In particular, it is suggested that the interictal period is characterized by moderate synchronization at large frequency bands while the pre-

ictal period is marked by a decrease in the beta range synchronization between the epileptic focus and other brain areas, followed by a subsequent hypersynchronization at the time of the onset of the ictal period [154].

Many different approaches have been applied towards determining these features, such as frequency domain tools [155, 156], wavelets [157, 158], Markov processes [159], autoregressive models [160, 161, 142], and artificial neural networks [162]. If it were possible to reliably predict seizure occurrence then preventive clinical strategies would be replaced by patient specific proactive therapy such as resetting the brain by electrical or other methods of stimulation. While clinical studies show early indicators for a pre-seizure state including increased cerebral blood flow, heart rate change, the research in seizure prediction is still not reliable for clinical use.

Recently there has been an increased focus in analyzing multivariate complex systems such as EEG recordings using concepts from *network theory* [163, 164, 165, 166], describing the topology of the multivariate time-series through *interaction networks*. The interaction networks enable characterization of the pair-wise correlations between electrodes using graph theoretical features over time [113, 114]. In the spatio-temporal interaction networks, *nodes* (vertices) represent the EEG channels and the *edges* (links) represent the level of neuronal synchronization between the different regions of the brain. This approach has been exploited in the analysis of various neuropsychiatric diseases including schizophrenia, autism, dementia, and epilepsy [114, 167, 168]. Within epilepsy research, evolution of certain graph features over time revealed better understanding of the interactions of the brain regions and the seizures. For instance, Schindler et al. analyzed the change in path lengths and clustering coefficients to highlight the evolution of seizures on epileptic patients [115], Kramer et al. considered the evolution of local graph features including betweenness centrality to explain the coupling of brain signals at seizure onset [116], and Douw et al. recently showed epilepsy in glioma patients was attributed to the theta band activity in the brain [117]. In [118] authors independently suggest a similar approach that combines tensor decompositions with graph theory. Even with this significant body of research what remained unclear was whether the network-related-approach can adequately identify the inter-ictal to pre-ictal transition [167].

In this paper, we continue studying a form of interaction networks dubbed *synchronization graphs* [169] and introduce new features as the early indicators of a seizure onset, thereby identifying the inter-ictal to pre-ictal transition.

Summarily, the current approaches aim to develop features that are naturally characteristic of seizure activity. While these approaches are both intuitive and instructive, ictal activity is often a small portion of the available data, and statistical learning techniques, which require a large corpus of data for reliable prediction, can be expected to perform poorly as seizure-predictors. However, these techniques seem promising for accurate prediction of non-ictal activity with respect to which ictal activity may be identified as an anomaly. Provided that the only anomalous activity in the data is the seizure, or that other anomalies present with discernible signatures, this provides an equivalent method of predicting seizure. In general we operate under the paradigm that any feature or parameter that distinguishes between ictal and non-ictal activity is a mathematical characteristic of seizure, although it may not be a natural physiological indicator. We rigorously define the notion of what it means to be a good mathematical characteristic of a seizure, rate our seizure-discriminating parameter accordingly, systematically increase how well it discriminates between ictal and non-ictal activity, and qualify our predictions using such a discriminating parameter.

Furthermore, since cortical activity is continuously recorded as EEG signals, it can be represented as a time-series, and analyzed using time-series forecasting methods. The objective of time-series forecasting is to use equally spaced past and current observations to accurately predict future values. *Autoregressive* models (AR) are commonly used tools for time-series prediction, and have been used to capture the spatio-temporal properties of EEG signals [160, 161]. We further improve on the AR model by using *Transfer Learning* [170] to learn the best forecast operator for a particular EEG recording from other EEG recordings. Transfer learning is a general form of learning such that there need not be any similarity in the distributions of the training and testing data. In our context, transfer learning does not require the past values and future values of the output variable to be correlated. In addition, transfer learning is particularly useful when data is only partially available or corrupted by

noise, where such data can be effectively supplemented by clean data from a different experiment. We further improve on the transfer learning by modifying the transfer set into the most similar form to the dataset being investigated by means of a simple transformation (based on the Procrustes problem [171]).

The three main contributions of this work are as follows: (i) Formulating seizure prediction as a problem of anomaly detection and developing a discriminating parameter for the anomaly (ii) bridging the two concepts of AR modeling and interaction graphs by constructing an AR(1) model on the features extracted from the time-evolving EEG synchronization graphs (as well as other features obtained from the EEG signal itself), and (iii) introducing the concepts of transfer and transformation learning to improve the predictions of the AR(1) model.

The organization of the paper is as follows: in Section 5.2, we describe our methodology starting with the epileptic EEG dataset, initial noise removal, and procedure to construct EEG synchronization graphs and extract features from the graph. We then detail the working of a feature selection method based on quadratic programming, build autoregressive models on the selected features, and transfer and transformation learn on these features. Finally, we use an alarm-based detection system to signal the seizure. In Section 5.3 we present and discuss the results for seizure prediction, and comparison to a benchmark. We provide an overview and outline possible extensions to this study, in Section 5.4.

5.2 Methodology

Our seizure-prediction paradigm is centered around discriminating between seizure and non-seizure activity. So we attempt to learn from normal activity maximally and from seizure activity minimally. We then interpret consistent and clear deviations from our understanding of normal activity as seizure. To this end: we initially suppress seizure activity; develop synchronization graphs to describe the seizure-suppressed cortical activity well; select maximally descriptive, minimally redundant features; cross-learn common attributes of seizure-suppressed activites across patients; and measure how well we predict on data from which the seizure has not been suppressed. Conditioned on this sequence of operations performing

well, we reason that any prediction error on the data where the seizure has not been suppressed is due to seizure-related activity. We develop a simple way to determine when such seizure-related activity is reliable enough to declare an imminent seizure, and use it to make such a declaration. The following sections describe the steps involved from taking the EEG signal as input to predicting the seizure. The steps are illustrated in the block diagram shown in Fig. 5.1.

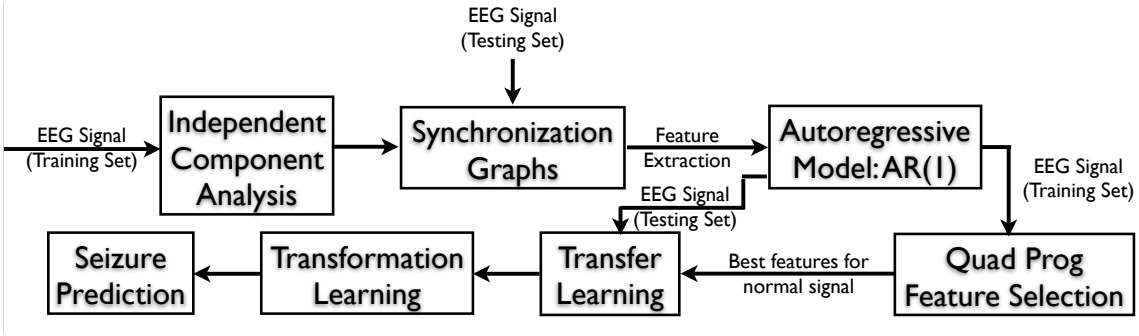


Figure 5.1: Block diagram representing the entire sequence of steps involved in the methodology. We apply Independent Component Analysis (ICA) for artifact removal and noise reduction, which allows us to learn non-ictal activity. This step is carried out only on the training set of EEG recordings, and the testing set of EEG recordings is kept separate. Synchronization graphs are constructed by using Phase Lag Index as explained in Section 5.2.3. These graphs are constructed for both the training and testing sets. Based on the features extracted from the synchronization graphs and the signal itself, an autoregressive model is built 5.2.6, and this model allows us to identify predictive importance of features that are determined via a Quadratic Programming Feature Selection (QPFS) technique 5.2.5. The feature selection technique is applied only on the training set. The important features are then used for transfer and transformation learning 5.2.7 which improves the performance of seizure prediction 5.2.8.

5.2.1 Epileptic EEG Data Set

Our dataset consists of scalp EEG recordings of 41 seizures from 14 patients. All the patients were evaluated with scalp video-EEG monitoring in the international 10-20 system (as described in [119]), magnetic resonance imaging (MRI), fMRI for language localization, and position emission tomography (PET). All the patients

had *Hippocampal Sclerosis* (HS) except one patient (Patient-1) who suffered from *Cortical Dysplasia* (CD). After selective amygdalohippocampectomy, all the patients were seizure free. The patient information is provided in Table 5.1. For 4 patients, the seizure would onset from the right, whereas for 10 patients the seizure would onset from the left.

The recordings include sufficient pre-ictal and post-ictal periods for the analysis. Two of the electrodes (A_1 and A_2) were unused and C_z electrode was used for referential montage that yielded 18-channel EEG recordings. A team of doctors diagnosed the initiation and the termination of each seizure and reported these periods as the ground truth for our analysis. An example of such a recording can be found in Fig. 2 in [120]. Seizures were 77.12 seconds long on average and their standard deviation was 48.94 seconds. The high standard deviation of the data is an indication of the vast variability in the patient data which further complicated the task of seizure prediction.

5.2.2 Seizure Suppression

In order to suppress seizure-activity, we resort to modeling EEG signal acquisition as follows. We assume that: (1) seizure activity is statistically independent of normal activity and (2) there may be numerous statistically independent cortical activities, both seizure related and otherwise, that combine to provide the signal captured by a single electrode (3) the seizure activity is non-gaussian. Based on these two assumptions, we look to locate and discard the seizure-related activity, thereby suppressing the seizure. Under the assumptions stated above, the problem of extracting seizure-related activity is mathematically equivalent to the cocktail party problem exemplifying blind source separation, which is solved by the state of the art technique of Independent Component Analysis [172, 173, 174], which has thus far been used mainly to remove artifacts from EEG data [175, 176, 177]. Here we use ICA to locate seizure-related activity and remove it in a manner similar to artifact-removal. Formally, given that $\mathbf{X} \in \mathbb{R}^{n,d}$ is a linear mixture of k d-dimensional independent components contained in $\mathbf{S} \in \mathbb{R}^{k,d}$, we may write

$$\mathbf{X} = \mathbf{AS},$$

where $\mathbf{A} \in \mathbb{R}^{n \times k}$ is *the mixing matrix* and $\mathbf{S} \in \mathbb{R}^{k}$. In general, both $\mathbf{A}$ and $\mathbf{S}$ are unknown and we compute the independent components, with respect to an independence maximization measure, as $\mathbf{S} = \mathbf{WX}$, where $\mathbf{W}$ is the inverse of the mixing matrix.

Once $\mathbf{A}$ is computed, we discard seizure related activity by zeroing the columns having the lowest euclidean norms. We reason this as follows: since much of the data is normal function, the independent components corresponding to seizure-related activity do not contribute to most of the data; their contribution is concentrated in time (corresponding to concentration in row-indices of $\mathbf{A}$). Due to the inherent scaling-degeneracy in the problem of blind source separation, we obtain an $\mathbf{A}$ having unit row-norms. This leads to the coefficients corresponding to seizure-related independent components to be tightly controlled, resulting in columns corresponding to the seizure-related independent components being of low euclidean norm. We heuristically zero the lowest two columns of $\mathbf{A}$ to form $\mathbf{A}_o$ and declare

$$\mathbf{X}_o = \mathbf{A}_o\mathbf{S}$$

to be the seizure-suppressed EEG data. It is important to note that the seizure is not completely suppressed, but the independent components retrieved allow us to model the non-ictal activity more precisely.

5.2.3 Construction of EEG Synchronization Graphs

For the signal $\mathbf{X}[i, m]$, we construct epochs of equal lengths with an overlap of 20% between the preceding and following epochs. The number of epochs, n, is equal to $1.25M/L$, where L is the duration of the epoch in same time units. Since the EEG recordings contain both temporal and spatial information, we construct *time-evolving EEG Synchronization Graphs* on the EEG datasets. A synchronization graph is constructed for each epoch, giving an indication of the spatio-temporal correspondence between electrodes - these relationships can then be utilized to obtain changes in the network by identifying descriptive features. The *nodes* represent the EEG electrodes and the *edges* represent a closeness relationship between the nodes in a given epoch. We use an epoch length of 5 seconds.

A sample time-evolving graph on an EEG recording is shown in Fig. 4.1. The pair-wise relationships between the electrodes during an epoch are used to construct the graph edges. If the pair-wise distance between two nodes i and k, where $i, k \in \{1, \ldots, 18\}$, and $i \neq k$, for epoch t, given as $d^n_{i,k}$, is less than a specified threshold, τ, then an edge is inserted into the graph between the two nodes. Note that smaller threshold values seek higher correlation between the electrodes, thereby yielding sparser graphs. Similarly, higher threshold values would establish an edge even if there is small correlation between the data, thereby yielding denser graphs. For our analysis, we performed a parametric search and found the best value of τ to be 1.

Several synchronization measures have been proposed as plausible options for $d^n_{i,k}$ to set up the edges in the graph. Based on earlier results presented in [169], we chose Phase Lag Index (PLI) [125] for $d^n_{i,k}$. PLI is defined as follows:

$$PLI_{i,k}(n) = \frac{1}{Lf_s} \left| \sum_{m=1}^{Lf_s} \operatorname{sgn}\left(\boldsymbol{\phi}^n_i(m) - \boldsymbol{\phi}^n_k(m)\right) \right| \tag{5.1}$$

where $\boldsymbol{\phi}^n_i = \arctan(\frac{\hat{\mathbf{x}}^n_i}{\mathbf{x}^n_i})$ is the angle of the Hilbert transform $\hat{\mathbf{x}}^n_i$ of the signal $\mathbf{x}^n_i$.

5.2.4 Feature Extraction from EEG Synchronization Graphs

We extract 26 features from the EEG synchronization graph for each epoch. These features quantify the compactness, clusteredness, and uniformity of the graph. Apart from these graph-based features, we compute two spectral features - the variance of the stationary distribution on an undirected markov chain on the graph, and the second largest eigenvalue of the Laplacian of the graph. In addition we compute certain natural statistics: the mean jump size between epochs and its variance, to measure the similarity to a Weiner process, and finally the *hinged mean* and *hinged variance*, defined as the mean and variance, respectively, of the signal at the current epoch centered/hinged at the mean of a strictly trailing window. These features arise naturally in change-point-detection and are motivated by the natural belief that, as a stochastic process, the EEG signal undergoes a statistical change when a seizure begins. To this feature set we also added time-domain and

spectral features. The time-domain features include the Hjorth parameters - activity, mobility, and complexity, and the frequency-domain features include skewness of amplitude spectrum and spectral entropy [104]. In all, we calculated 38 features.

In subsequent text, we refer to the feature matrix as $\mathbf{D} \in \mathbb{R}^{n,d}$, with n epochs and d features. We refer to the feature vector at time t (row t of $\mathbf{D}$) as $\mathbf{d}^t$, and the time-series corresponding to feature i (column i of $\mathbf{D}$) as $\mathbf{d}_i$. A complete list of the features used in this work and their definitions is listed in Tables 4.2, 4.3, 5.2. For further information regarding the features, we refer the reader to [62, 104].

5.2.5 Determining The Significance of Features

The computed features were motivated by discussions with the subject matter experts, with the view of casting a meaningful but wide net to capture attributes of an epileptic seizure. However, this doesn't strictly preclude the possibility that certain features may be redundant or low in predictive importance. Furthermore, we wish to select features that are particularly descriptive of the non-ictal activity, of which the data is largely comprised. Therefore, we quantify the predictive significance of the features in a natural but effective way, and score the features to maximize their predictive importance for the entire data, and minimize redundancy, using the method in [178], which we summarize here. The primary advantages of using QPFS are as follows: (i) QPFS is based on efficient quadratic programming technique [179]. The quadratic term quantifies the dependence between each pair of variables, whereas the linear term quantifies the relationship between each feature and the class label. (ii) QPFS provides a considerable time complexity improvement over current methods on very large data sets with high dimensionality.

5.2.5.1 Measuring Redundancy

Our notion of redundancy arises naturally from the interpretation of brain activity as a stochastic process, whence the usual notion of linear dependence is replaced with the notion of statistical correlation. Specifically, suppose the data matrix, $\mathbf{D} \in \mathbb{R}^{n,d}$, spanning n epochs and consisting of d features. We define, the correlation matrix, $\mathbf{Q} \in \mathbb{R}^{d,d}$, element-wise, where $\mathbf{Q}(i,j)$ is the Pearson correlation

coefficient between the feature vectors $\mathbf{d}_i, \mathbf{d}_j \in \mathbb{R}^n$:

$$\mathbf{Q}(i,j) = \frac{\mathbf{d}_i^\top \mathbf{d}_j}{\|\mathbf{d}_i\| \, \|\mathbf{d}_j\|}.$$

The quadratic form $\mathbf{x}^\top \mathbf{Q}\mathbf{x}$ thus has the natural interpretation of yielding the sample-covariance of a compound feature, with coefficients contained in $\mathbf{x}$, which is the notion of redundancy that we wish to minimize.

5.2.5.2 Measuring Predictive Importance

We first recall that the activity of the brain at time t is completely captured by $\mathbf{d}^t$. We define the predictive importance, f_i, of the feature i, as the r.m.s. influence of $\mathbf{d}_i^t$ on $\mathbf{d}_j^{t+1}, 1 \leq j \leq n$, measured by the coefficients in the forecast operator corresponding to i. Formally, let $\mathbf{\Psi} \in \mathbb{R}^{d,d+1}$ be the forecast operator. Then our best prediction of $\mathbf{d}^{t+1}$ is $\tilde{\mathbf{d}}^{t+1}$ where

$$\tilde{\mathbf{d}}^{t+1} = \mathbf{d}^t \mathbf{\Psi}^\top.$$

The influence, $\mathbf{p}_i(j)$, of feature i on j, contained in $\mathbf{p}_i \in \mathbb{R}^d$, may be determined by predicting via $\mathbf{\Psi}$ using its indicator vector, $\mathbf{e}_i$:

$$\mathbf{p}_i = \mathbf{e}_i \mathbf{\Psi}^\top,$$

whence the r.m.s. influence, f_i, of i is simply

$$f_i = \|\mathbf{p}_i\| = \|\mathbf{\Psi}_i\|,$$

the column-norm of the forecast operator corresponding to column i. We define $\mathbf{f} \in \mathbb{R}^d$ such that $_\mathbf{i} = \|\mathbf{\Psi_i}\|$, as the predictive importance vector.

5.2.5.3 Optimizing Redundancy and Predictive Importance

We obtain a significance-distribution over the features that maximizes predictive importance and minimize redundancy by solving

$$\begin{aligned} \mathbf{x}^* = \quad & \arg\min \quad q(\mathbf{x}); \\ & \text{subject to} \quad \mathbf{x} \in \mathbb{R}^d,\ \mathbf{x} \geq \mathbf{0},\ \textstyle\sum_i \mathbf{x}_i = 1, \end{aligned} \tag{5.2}$$

where the constraints arise from forcing the resulting vector to be a distribution, from which we omit an appropriately sized tail, or select just the support if it is small. To make the objective function stable under scaling of the data, we normalize $\mathbf{f}$ to obtain

$$\hat{\mathbf{f}} = \mathbf{f} / \left\|\mathbf{f}\right\|_\infty .$$

To effect a meaningful trade-off between minimizing redundancy and maximizing predictive importance, we take a convex combination of the corresponding terms:

$$q(\mathbf{x}) = (1-\alpha)\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \alpha \hat{\mathbf{f}}^\top \mathbf{x},$$

where α is chosen, as in [178] as

$$\alpha = \frac{\sum_{i,j} \mathbf{Q}(i,j)/n^2}{\sum_{i,j} \mathbf{Q}(i,j)/n^2 + \sum_k \mathbf{f}_k/n}.$$

Since both the predictive importance and the correlation matrix are statistical in nature, they are less affected by the relatively fleeting seizure. So we expect the significant features obtained via QPFS to be features that are highly predictive of non-ictal activity. We use the MATLAB utility quadprog to solve Eq. 5.2.

5.2.6 Autoregressive Modeling on Feature Data

Research has indicated that promising results regarding early detection or prediction of the seizure can be achieved by application of an autoregressive model (AR) to the EEG signal [180, 142]. Also, AR models are linear and as shown in prior research are comparable to non-linear models in their predictive capability [148, 149, 150, 151, 152]. We expand on these earlier results by applying an autoregressive

model to the features extracted both from the graph and the signal itself.

An autoregressive model of order 1, AR(1), is applied to the matrix $\mathbf{D}$, extracted from the time-evolving EEG synchronization graphs. For an AR(1) model the output at time t is only dependent on the values of the time-series at time $t-1$. As a result, the implicit assumption when using AR(1) is that $\mathbf{d}^t$ is a markov chain indexed by t. Formally

$$\mathbf{d}_i^t = \rho_{i0} + \rho_{i1}\mathbf{d}_1^{t-1} + \rho_{i2}\mathbf{d}_2^{t-1} + \ldots + \rho_{im}\mathbf{d}_m^{t-1} + \epsilon_t \tag{5.3}$$

where ρ_{ij} are the linear coefficients computed via autoregression. In matrix form, (5.3) is $[\mathbf{D}]_1^t = \boldsymbol{\Psi} \cdot [1, \mathbf{D}]_0^{t-1} + \boldsymbol{\epsilon}$, where the notation $[\mathbf{A}]_a^b$ denotes a matrix containing all rows from a to b of $\mathbf{A}$, including rows a, b. We compute $\boldsymbol{\Psi}$ to minimize the error $\boldsymbol{\epsilon}$ in euclidean norm,

$$\begin{aligned} \boldsymbol{\Psi} &= \arg\min_{\mathbf{Z}} \left|\left| [\mathbf{D}]_1^t - \mathbf{Z} \cdot [\mathbf{1}, \mathbf{D}]_0^{t-1} \right|\right|_F^2, \\ \boldsymbol{\Psi} &= [\mathbf{D}]_1^t \cdot \left([\mathbf{1}, \mathbf{D}]_0^{t-1}\right)^\dagger. \end{aligned} \tag{5.4}$$

where $\mathbf{A}^\dagger$ denotes the moore-penrose pseudoinverse of $\mathbf{A}$. The role of the operator $\boldsymbol{\Psi}$ is to predict $\mathbf{D}(t)$ as a function of $\mathbf{D}(t-1)$. Any operator that does this will be called subsequently as the *forecast operator*. Thus, using an AR(1) model we arrive at a forecast operator, $\boldsymbol{\Psi}$.

5.2.7 Transfer Learning and Transformation Learning on Autoregressive Model

We critically improve this forecast operator obtained from AR(1) in two directions. First, we improve it under the assumption that the data obtained from different patients are not completely independent of each other; that data obtained from one patient holds some information common to all patients, along with information specific to the patient. Thus, we *transfer* knowledge from one patient to another, motivated by the existing work on Transfer Learning. Specifically, given a feature data set $\mathbf{D}$, the feature transfer set, $\hat{\mathbf{D}}$, and the corresponding forecast operators $\boldsymbol{\Psi}$ and $\hat{\boldsymbol{\Psi}}$ respectively we transfer knowledge from $\hat{\mathbf{D}}$ to $\mathbf{D}$ by regularizing

(5.4) with

$$\lambda||\mathbf{\Psi} - \hat{\mathbf{\Psi}}||_F^2.$$

The parameter λ, playing the familiar role of the Tikhonov Regularizer, is the *transfer coefficient*, governing how much we learn from $\hat{\mathbf{D}}$ onto $\mathbf{D}$. The forecast operator obtained from this transfer learning is simply

$$\bar{\mathbf{\Psi}} = \arg\min \left\{ \left\| [\mathbf{D}]_{t-1}^0 * \mathbf{\Psi} - [\mathbf{D}]_t^1 \right\|_F^2 + \lambda \left\| \mathbf{\Psi} - \hat{\mathbf{\Psi}} \right\|_F^2 \right\} , \tag{5.5}$$

the analytical solution for which is:

$$\mathbf{\Psi} = \hat{\mathbf{\Psi}} + \left({[\mathbf{D}]_{t-1}^0}^\top [\mathbf{D}]_{t-1}^0 + \lambda \mathbf{I} \right)^\dagger \left([\mathbf{D}]_{t-1}^0 \right)^\top \left([\mathbf{D}]_t^1 - {[\mathbf{D}]_{t-1}^0}^\top \hat{\mathbf{\Psi}} \right) \tag{5.6}$$

As more data is obtained for $\mathbf{D}$, the value of λ is reduced because now the core set is getting better at predicting its own future values. To test our estimates, we use the following split between the training and testing data. First, we split $\mathbf{D}$ into training (TR) and testing (TE) sets. Within the training set, we create a further split thereby creating training prime ($\overline{TR}$) and validation (Val) sets. We then train our AR(1) model on $\overline{TR}$, and then use $\hat{\mathbf{\Psi}}$ to improve this estimate by testing on Val. Then, we retrain the model using the learned parameters on the entire training set TR and finally test on TE.

Next, we account for the differences in the collected data that may arise as a result of non-uniformities in the process of acquiring data. We do this under the assumption that the spectral nature of the data is minimally variant with changes across the various setups for acquiring data from multiple patients, and that the flow of time is immutable. To learn from $\hat{\mathbf{D}}$ onto $\mathbf{D}$, we find the object, $\hat{\mathbf{D}}_{\mathbf{D}}$, retaining the spectral nature of $\hat{\mathbf{D}}$, and respecting the directionality of time, that is the closest to $\mathbf{D}$. Formally, we find a rotation $\mathbf{\Gamma}(\hat{\mathbf{D}}, \mathbf{D})$ such that

$$\hat{\mathbf{D}}_{\mathbf{D}} = \hat{\mathbf{D}}\mathbf{\Gamma}(\hat{\mathbf{D}}, \mathbf{D}),$$

$$\mathbf{\Gamma}(\hat{\mathbf{D}}, \mathbf{D}) = \arg \min_{\substack{\mathbf{U} \\ \mathbf{U}^\top \mathbf{U} = \mathbf{I}}} \left\| \mathbf{D} - \hat{\mathbf{D}}\mathbf{U} \right\| .$$

This is the Procrustes problem that has been well-studied [171], and has a closed form solution in terms of the SVD of $\mathbf{D}, \hat{\mathbf{D}}$. Let these SVDs be

$$\mathbf{D} = \mathbf{U_D}\mathbf{\Sigma_D}\mathbf{V_D}^\top, \quad \hat{\mathbf{D}} = \mathbf{U_{\hat{D}}}\mathbf{\Sigma_{\hat{D}}}\mathbf{V_{\hat{D}}}^\top.$$

Then

$$\mathbf{\Gamma}(\hat{\mathbf{D}}, \mathbf{D}) = \mathbf{V_{\hat{D}}}\mathbf{V_D}^\top, \Rightarrow \hat{\mathbf{D}}_\mathbf{D} = \mathbf{U_{\hat{D}}}\mathbf{\Sigma_{\hat{D}}}\mathbf{V_D}^\top.$$

We now transfer-learn using $\hat{\mathbf{D}}_\mathbf{D}$. In summary, we first notice that knowledge can be transferred from other but similar data and then transform such similar data sets into their *most learnable* forms using a simple transformation.

5.2.8 Declaration Of Imminent Seizure

We use the prediction errors incurred by the use of our forecast operator as the eventual *ictal discriminator*. We compute an estimate of the probability of deviation towards seizure using these errors, and declare that a seizure is imminent when this probability is reliably high. We outline how we compute this probability and quantify the sense of reliability we use, in that order.

5.2.8.1 Probability Of Deviation Towards Seizure

Let $\boldsymbol{\epsilon}(t)$ be the prediction error at time, t. Using a moving window of size $\Delta = 30$, we first use a simple statistical thresholding on the errors to determine if an *alarm* has to be thrown, which signifies an outlier to normal function, and a potential seizure. Specifically, let $\kappa(t)$ be the binary variable indicating whether or not an alarm is thrown - 1 when it is thrown and 0 when it is not. Let $\mu(t, \Delta), \sigma(t, \Delta)$ denote the mean and standard deviation of the sequence $\boldsymbol{\epsilon}(t), \boldsymbol{\epsilon}(t+1), \cdots, \boldsymbol{\epsilon}(t+\Delta)$. Then

$$\kappa(t) = \begin{cases} 1 & \text{if } \boldsymbol{\epsilon}(t+\Delta) - \mu(t, \Delta) > \tau^* \sigma(t, \Delta) \\ 0 & \text{otherwise} \end{cases}$$

where τ^* is a tolerance/sensitivity parameter. Clearly, κ is an indicator of the one-sided tail of the distribution from which $\boldsymbol{\epsilon}(t)$ is drawn. Under our assumption that a recorded activity is either normal function or seizure, the measure of the tail of

the distribution of errors during normal function is an appropriate estimate of the probability of seizure. We estimate the size of this *tail of normality* for an interval by the ensemble average of $\kappa(t)$ for the interval. When $\kappa(t)$ indicates seizure repeatedly in a manner highly unlikely to have arisen from random sampling from the tail of normality, we declare an imminent seizure. In practice, we choose $\tau^* = 3$, and declare a seizure when we see 3 consecutive alarms. We justify this choice as follows: in the case where $\kappa(t)$ indicates the result of the high-entropy fair coin toss (i.e. 1 and 0 with equal probability), the probability of obtaining 3 consecutive alarms is 12.5%. In practice our estimate of the size of the tail of normality is significantly below 1/5, resulting in our three-in-a-row rule to be an even rarer occurrence than once in 125 occurrences or 0.8%.

5.3 Results

We present results for the Quadratic Programming Feature Selection algorithm, determining the best forecast operator, and a comparison of the performance of our seizure prediction algorithm on basic autoregression vs. with the addition of transfer and transformation learning.

5.3.1 Quadratic Programming Feature Selection Results

The feature-significance vectors obtained from solving the QPFS problem in (5.2) were found to be highly sparse, and the features that were supported by these vectors were chosen without exception - 9 in all: (i) Average Degree, (ii) Diameter of graph, (iii) Average Path Length, (iv) Giant Connected Component Ratio, (v) Number of Connected Components, (vi) Percentage of Isolated Points, (vii) Number of eigenvalues with value 0 of the normalized Laplacian matrix, (viii) Number of eigenvalues with value 2 of the normalized Laplacian matrix, and (ix) Normalized Laplacian trace.

5.3.2 Baseline SVM results

To establish a baseline to validate the efficacy of our results, we compare our algorithm to the following benchmark algorithm:

Application of Support Vector Machine to feature matrix **D**:
We provide as input the features identified by QPFS from **D** to a two-class Support Vector Machine (SVM). We learn a model of the inter-ictal and ictal states based on their respective feature values. We then classify using the SVM the seizure onset in the pre-ictal region based on the feature values in that region. The intuition being that the initial part of the pre-ictal region will have features similar to the inter-ictal region, whereas the latter part will be more similar to the ictal region in the feature space. We consider the pre-ictal region to start 10 minutes prior to the onset of the seizure.

We found that the benchmark did not predict the seizure in 14 of the 41 analyzed recordings. The median prediction time for the recordings for which seizures were predicted was 1.25 minutes prior to the seizure.

5.3.3 Autoregression vs. Transfer and Transformation Learning

One of the objectives of this study was to improve the basic autoregressive model by the application of transfer learning and transformation learning. We now show that the additional functionality makes the prediction either at least as good as that by the AR(1) model or better for a significant percentage of the dataset. We found that in 60% of the analyzed EEG recordings, transformation learning was able to predict the seizure prior to AR(1), or transfer learning. In 52.5% of the analyzed EEG recordings, transfer learning performed better than the AR(1) model predicting the seizure earlier. Finally, in 67.5% of the cases, either transfer learning or transformation learning was better than the AR(1) model. The median prediction times prior to the occurrence of the seizure for the three methods are 10 min, 10.96 min, and 11.04 min for AR(1), transfer learning, and transformation learning, respectively. Considering only the recordings where transformation learning or transfer learning outperformed AR(1), the median prediction times change to 9.33 min for AR(1), 11.17 min for transfer learning, and 11.92 min for transformation learning.

In Fig. 5.2, the first row ((a)–(c)) consists of an analyzed EEG recording where the AR(1) model was able to predict the seizure before the other two techniques.

The second row ((d)–(f)) consists of an analyzed EEG recording where the AR(1) model with transfer learning and transformation learning was able to predict the seizure before the other two techniques. Finally, the third row ((g)–(i)) consists of an anomalous result where the AR(1) model with transfer learning predicted the seizure before either AR(1) model or the AR(1) model with transfer and transformation learning.

5.4 Conclusions and Future Work

In this study, we outline a seizure prediction algorithm designed for EEG epileptic seizure data by constructing an autoregressive model improved by the addition of transfer learning and transformation learning on features extracted by building synchronization graphs on the independent components of the EEG signal. We use a quadratic programming algorithm called Quadratic Programming Feature Selection (QPFS) to select the features with the highest predictive importance and minimal redundancy.

One of the primary concerns with the seizure prediction area is the definition of a Seizure Prediction Horizon (SPH). In the literature prediction horizons have varied from several minutes to a few hours [181]. We would like to come up with a more rigorous theoretical basis for assigning prediction horizons. Another future direction is with respect to the various thresholds used in the study. Although, well-motivated and justified from the literature, we would like to obtain these thresholds from first principles. Examples of these thresholds include epoch lengths for the synchronization graphs, sensitivity parameters for raising an alarm, and number of columns to zero out from the mixing matrix in Independent Component Analysis (ICA). Yet another important future direction is analyzing the partial contribution of each module in the pipeline to determine the effect of individual modules in improving the basic prediction. Specifically, we would like to examine the influence of ICA vs. transformation learning to determine which of the two is better used for the initial surgery to suppress seizure - to ensure that we don't use a gas-engine for the short haul and a horse for the long one. Furthermore, we would like to qualify the use of transfer learning based on the similarity of the data sets being learned

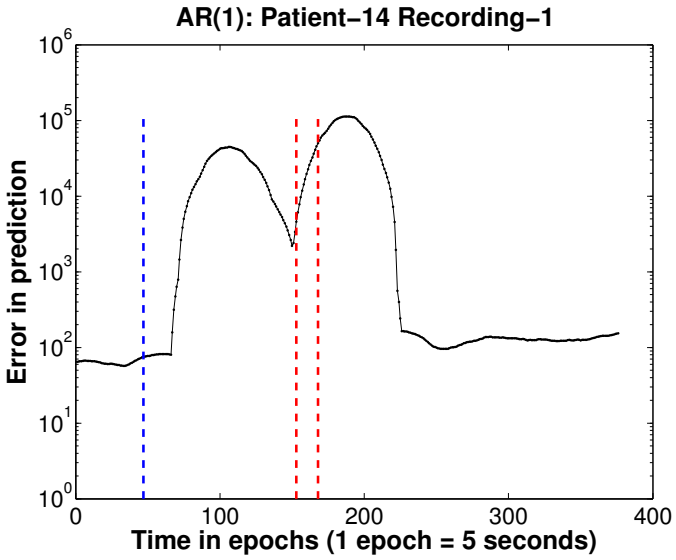
across: establish a metric of closeness of data sets/learnability across patients and recordings. Finally, the problem of seizure prediction is accompanied by the problem of localizing seizure, which, apart from requiring new methods, also sets a higher standard for understanding seizure. We hope to contribute to this problem in the future as well.

Table 5.1: Patient Types. Almost all the patients (except one patient) exhibited hippocampal sclerosis (HS). There are two types of lateralizations in HS: left (L) and right (R). One patient (Patient-1) exhibited cortical dysplasia (CD).

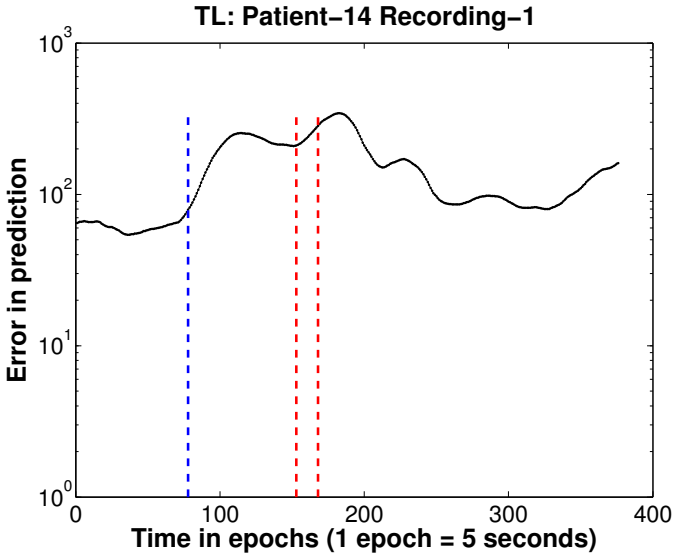
Patient	Pathology	Lateralization	Number of Recordings	Length of Recordings (in min)
Patient-1	CD	R	2	30
Patient-2	HS	R	2	30
Patient-3	HS	R	3	60
Patient-4	HS	R	5	60
Patient-5	HS	L	1	60
Patient-6	HS	L	1	30
Patient-7	HS	L	2	60
Patient-8	HS	L	2	60
Patient-9	HS	L	3	60
Patient-10	HS	L	3	30
Patient-11	HS	L	2	60
Patient-12	HS	L	5	41
Patient-13	HS	L	5	35
Patient-14	HS	L	5	35

Table 5.2: Names and description of EEG global graph features. Features 1–26 are computed on the synchronization graph, features 27 and 28 are signal-based. Features 29–31 are representative of change-point detection, and features 32 and 33 are spectral features. Features 34–36 and 37–38 are time-domain and frequency-domain features, respectively.

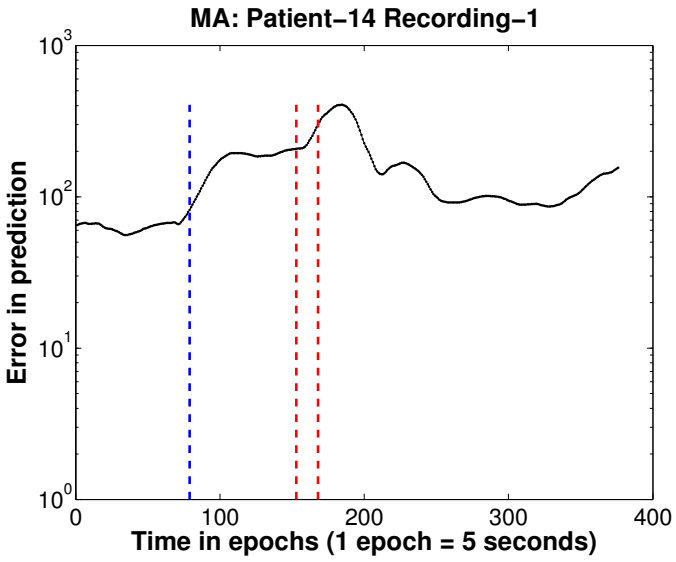
Index	Feature Name	Description
27	Mean of EEG recording	Mean of EEG signal for each electrode and epoch
28	Variance of EEG recording	Variance of EEG signal for each electrode and epoch
29,30	Change-based Features	Mean and variance of jump size in EEG signal for each electrode and epoch
31	Change-based Feature 3	Variance of EEG signal for particular electrode in given epoch after subtracting the mean of up to 3 previous windows
32	Spectral Feature 1	Variance of eigenvector of the product of the adjacency matrix and the inverse of the degree matrix
33	Spectral Feature 2	Second largest eigenvalue of the Laplacian matrix
34,35,36	Hjorth parameters (time-domain)	Activity, Mobility, and Complexity
37,38	Frequency-domain features	Skewness of amplitude spectrum and Spectral entropy



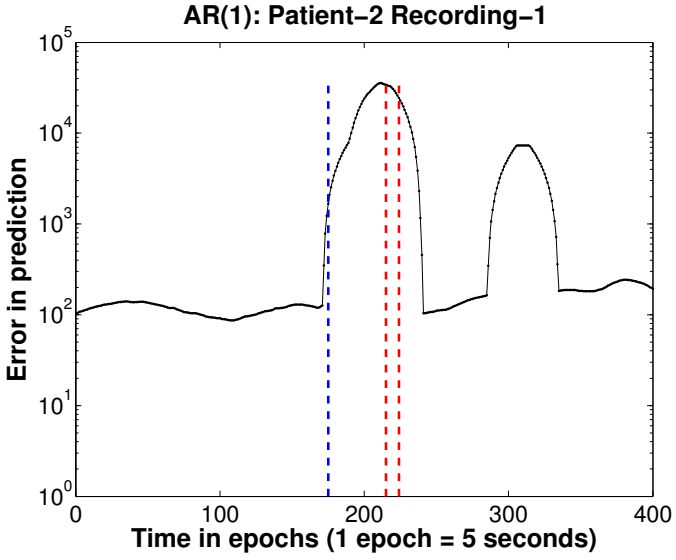
(a) AR(1) model



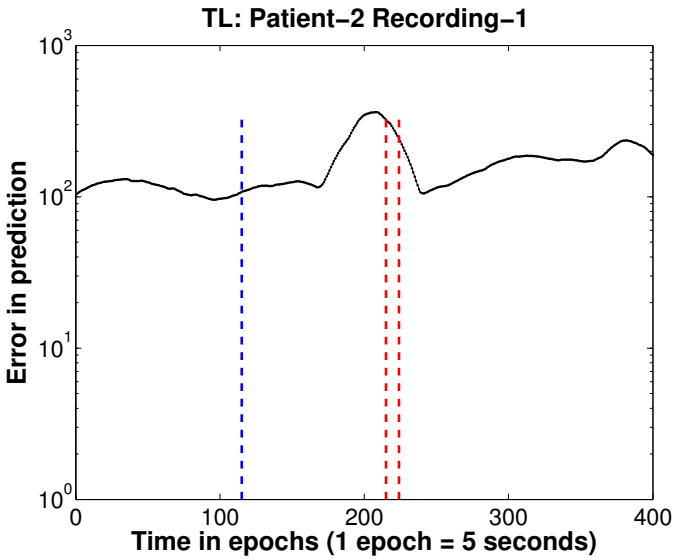
(b) AR(1) model with transfer learning



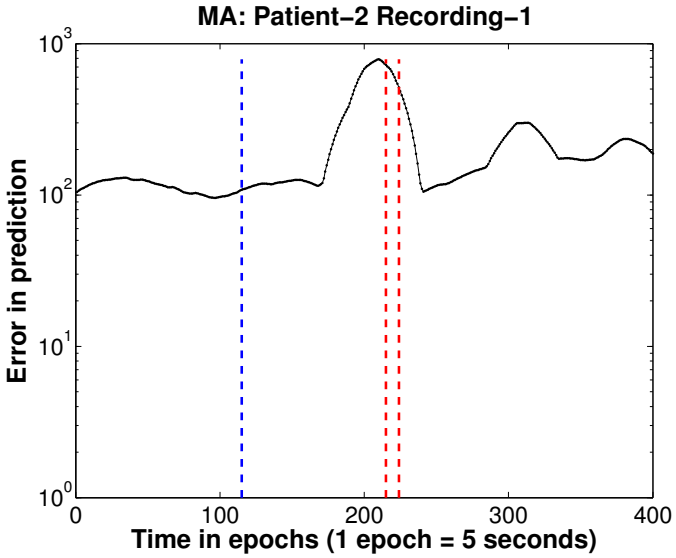
(c) AR(1) model with transfer and transformation learning



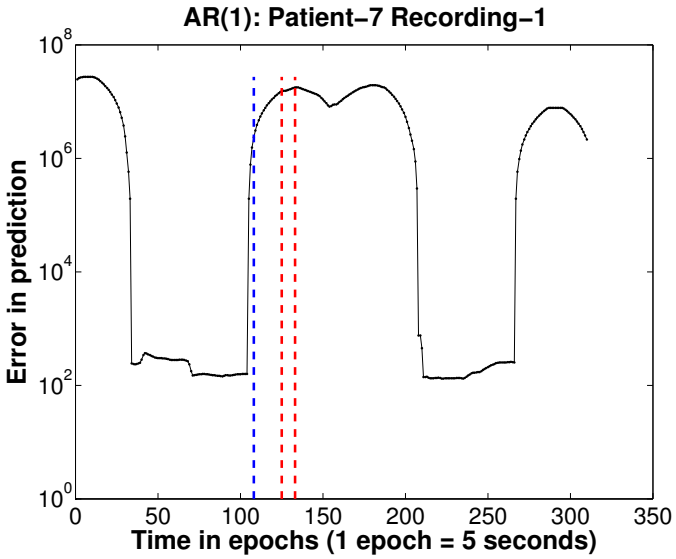
(d) AR(1) model



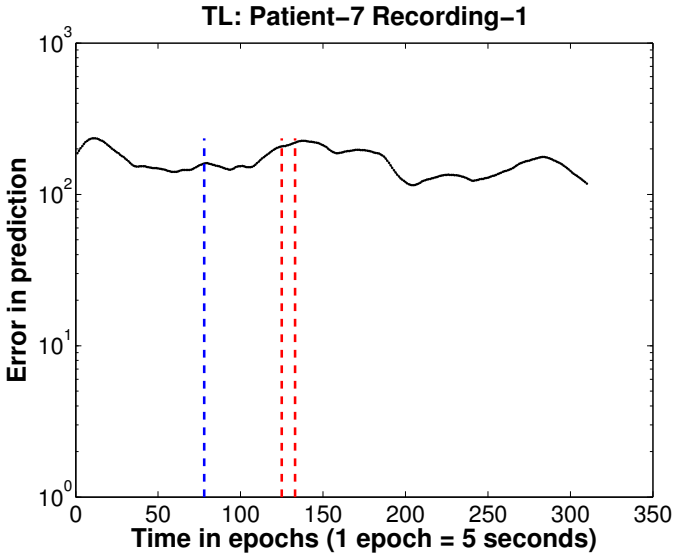
(e) AR(1) model with transfer learning



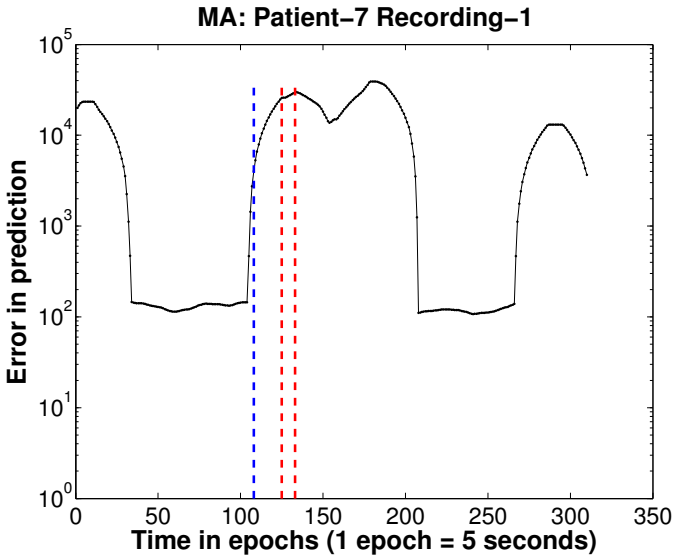
(f) AR(1) model with transfer and transformation learning



(g) AR(1) model



(h) AR(1) model with transfer learning



(i) AR(1) model with transfer and transformation learning

Figure 5.2: Comparison of the AR(1) model ((a), (d), and (g)) with the AR(1) model improved by transfer learning ((b), (e), and (h)) and transformation learning ((c), (f), and (i)). The epoch at which the seizure is detected is shown in blue, the start and end of the seizure region are marked in red. The first row is an example of where the AR model does better than both transfer learning and transformation learning. The second row is a typical example of the AR(1) model with transfer and transformation learning outperforming both the AR(1) model and the AR(1) model enhanced by transfer learning. The third row is an example of an anomaly where the AR(1) model with transfer learning performs much better than the AR(1) model and the AR(1) model with transfer and transformation learning.

CHAPTER 6
LUNG NECROSIS AND NEUTROPHILS REFLECT COMMON PATHWAYS OF SUSCEPTIBILITY TO *MYCOBACTERIUM TUBERCULOSIS* IN GENETICALLY DIVERSE, IMMUNE COMPETENT MICE

6.1 Introduction

Outbred mice have been used in a variety of research settings because they may better reflect human diversity than inbred mice [182, 183, 184, 185, 186, 187]. Using a diverse experimental population may be beneficial to reduce erroneous interpretations due to strain-specific effects and to detect phenotypes that may be absent from inbred mice because laboratory strains are quite related [188, 189]. Here, we use Diversity Outbred (DO) mice because the population is highly heterogeneous due to its 8 unique founder strains [186, 187]. DO mice have not been fully characterized as a model for *M. tuberculosis*, and little is known regarding their immune or inflammatory responses. Although 3 of the founder strains (C57BL/6, 129, and A/J) have identified requirements for immunological resistance to *M. tuberculosis* and some correlates of susceptibility [190, 191, 192, 193], the other 5 founder strains have not been investigated. To our knowledge, DO mice have been included in only one infection study [194] and one aging study [195]. Thus, additional work is needed. Here, we infect DO mice with *M. tuberculosis* by aerosol, and evaluate morbidity, bacterial burden, and a set of immunological and inflammatory molecules in lungs, blood, and plasma. Multiple methods (simple and complex statistical analyses, and machine learning) identified and ranked features, which were used to generate and test models capable of classifying relative susceptibility.

Decades of research show that the host response to *M. tuberculosis* is complex, involving many cell types, molecules and signaling pathways. Resistance to mycobacterial infection in humans, and to *M. tuberculosis* in mice, requires TH1

polarized, cell-mediated immunity. As no studies have reported immune responses in *M. tuberculosis* infected DO mice, we focused on cytokines that significantly alter survival of inbred mice when these molecules are absent or blocked, including: IFN-γ, IL-12, TNF, IL-2, and IL-10 [196, 197, 198, 199, 200, 201, 202]. This is not an exhaustive list, but reflects immune deficiencies that increase the risk of mycobacterial disease in humans [203, 204, 205]. Interestingly, these defects are not typical of active TB [206, 207, 208] which manifests as granuloma necrosis and neutrophilic inflammation. The underlying mechanisms are not completely known. One mechanism of necrosis in mice [191, 209, 210, 211, 212] and some pulmonary TB patients [213] is genetic polymorphism in Ipr1 (mouse) and SP110 (human). However, the murine polymorphism is rare (existing naturally in 1 inbred substrain), and not all TB patients have polymorphisms in SP110 [214] or polymorphisms have no association with TB [215]. Therefore, it is worth pursuing additional models to identify more mechanisms of necrosis. Neutrophils are also negative indicators of pulmonary TB, recognized more than two decades ago [216], and interest in neutrophils has increased following blood profiling studies in humans [217]. Many [194, 218, 219, 220, 221, 222] but not all [223, 224] studies implicate neutrophils as causal mediators of lung damage. Therefore, we included necrosis, neutrophils, and neutrophil chemokines CXCL1, CXCL2, and CXCL5.

Following aerosol infection, DO mice vary in survival, morbidity, bacterial burden, granuloma composition, immunological cytokines, and neutrophil chemokines. Nearly half of the DO mice developed morbidity which was associated with necrosis of lung tissue and inflammatory cells, and neutrophil infiltration. Strong and significant disease correlates were lung neutrophil chemokines (CXCL1, CXCL2, CXCL5), TNF, and dead cells. In contrast, cytokines (IL-12, IL-2, IFN-γ, and IL-10) were weak correlates. Peripheral (blood/plasma) cytokine/chemokine levels were not strong correlates of disease or of lung responses, with the exception of plasma CXCL1.

We generated models to classify *M. tuberculosis* susceptibility and tested the models on independent data, using features consistently identified as important by statistical analyses and machine learning. Lung TNF, CXCL1, CXCL2, CXCL5,

IFN-γ, IL-12 and blood IL-2 and TNF were identified as important. The most accurate models used only CXCL1, CXCL2 and CXCL5 to classify Supersusceptible, Susceptible, Resistant and Non-infected mice. TNF was an important feature, but models with TNF were less accurate because they confused Resistant mice with Supersusceptible mice. Cytokines discriminated non-infected from infected mice, but could not distinguish susceptibility.

In summary, of the molecules that we measured, neutrophil chemokines CXCL1, CXCL2 and CXCL5 best discriminated highly susceptible DO mice while low/no immune cytokines identified Non-infected individuals. Similar to humans [194] neutrophils are features of disease and CXCL1 may be a good peripheral biomarker. Overall, these results suggest that future studies with DO mice may identify specific genetic and molecular features of susceptibility, in particular those associated with neutrophils and necrosis.

6.2 Materials and Methods

6.2.1 Ethics Statement

Experiments were approved by Tufts University IACUC protocols G2012-53 and G2012-151. Biosafety Level 3 (BSL3) work was approved by IBC registration (GRIA04).

6.2.2 Infection with *Mycobacterium tuberculosis* and Quantification of *Lung Bacilli*

Female J:DO (009376) and C57BL/6J (000664) mice (The Jackson Laboratory, Bar Harbor, ME) were maintained under BSL3 conditions with sterile food, bedding, and water in the New England Regional Biosafety Laboratory (North Grafton, MA). At 8 weeks of age, J:DO and C57BL/6J mice were infected by aerosol exposure using a CH Technologies, Inc. machine resulting in 127 ± 68 *M. tuberculosis* strain Erdman bacilli to the lungs to generate training data, and 97 ± 61 bacilli to the lungs to generate test data. Lung bacillary burden was determined following homogenization in sterile PBS using gMACS M-tubes (Miltenyi Biotech), plating serial dilutions on OADC-supplemented 7H11 agar, and counting CFUs after 3 weeks at 37°C.

6.2.3 Identification of Susceptibility Classes

Mice were euthanized when morbidity developed, or at day 35 of infection, whichever came first. Mice euthanized prior to day 35 were classified as Super-susceptible. Susceptible mice had no outward signs of illness on day 35, but had lost some weight determined by retrospective analysis. Resistant mice had stable weight or gained weight throughout. Age- and sex-matched non-infected control mice, housed identically, were euthanized on day 35. These classes were used to organize the microscopy results after blinded evaluation, and for testing models.

6.2.4 Cytokine Measurements

ELISAs for TNF, IFN-γ, IL-12, IL-2, and IL-10 were performed using antibody pairs and standards or OptEIA kits (BD Biosciences) on serially diluted homogenized lung or from antigen-stimulated whole blood as described [192] with the exception that blood was diluted 1:5. Samples at or below the IFN-γ level of detection were repeated using the eBioscience Ready-Set-Go ELISPOT kit [225]. CXCL1, CXCL2, and CXCL5 were quantified using R&D Systems ELISA kits on diluted homogenized lung and plasma.

6.2.5 Enumeration of Dead Cells

Dead cells were enumerated following collagenase/DNase digestion of the lungs and trypan blue staining on single cell suspensions [226].

6.2.6 Light Microscopy

Lung lobes were inflated and fixed with 10% neutral buffered formalin, processed, embedded in paraffin, cut at 5μm, and stained with hematoxylin and eosin at the CSVMHistology Laboratory. Two serial sections, 100μm apart, were examined by a board certified veterinary pathologist (GB) without knowledge of the groups.

6.2.7 Statistical Analyses

GraphPad Prism 6.0 was used for correlations. Data had non-normal distributions, so Spearman correlation coefficients were calculated, and identified as very weak $(0 - 0.19)$, weak $(0.20 - 0.39)$, moderate $(0.40 - 0.59)$, strong $(0.60 - 0.79)$, or

very strong $(0.80-1.0)$ and considered statistically significant if $p < 0.05$. MATLAB was used for analysis of variance (ANOVA) followed by Tukeys multiple comparison tests to identify which features had significantly different means between any two classes, defined as $p < 0.05$.

6.2.8 Machine Learning

Machine learning can help identify relationships in complex data by exploring data in multidimensional space [227, 228]. We applied standard machine learning methods for model training and testing, using mice with all parameters measured ($N = 78$ mice for training and $N = 60$ mice for testing). Five supervised (Classification tree [229, 230, 231], Relief Attribute Evaluation [232], Consistency Subset [233], Spectral Feature Selection [234], a custom Principal Components Analysis), and two unsupervised (Brute Force and standard Principal Component Analysis [235]) methods were used. Of all the methods, Classification trees [229, 230, 231] and Relief Attribute Evaluation [232] were the most informative because these methods ranked molecular features by importance. Classification trees were pursued for validation and testing because they provided the most accurate models.

6.3 Results

6.3.1 DO Mice Have a Spectrum of Responses to Aerosolized *M. tuberculosis*

All mice gained weight for at least two weeks after infection. Afterwards, nearly half of the DO mice developed morbidity before 35 days, resulting in significantly reduced survival compared to the founder C57BL/6J strain (Fig. 6.1A). Neither starting nor peak body weight protected DO mice, as there was no correlation between survival and absolute weight (not shown). Instead, reduced survival strongly reflected the rate of weight loss (not shown), as observed in other heterogeneous mice [195]. Relative weight loss and lung *M. tuberculosis* burden had a strong, significant inverse correlation (Fig. 6.1B), as expected.

It is common in human TB studies to stratify by clinical symptoms or disease severity, and we applied this approach to DO mice. Three susceptibility classes were

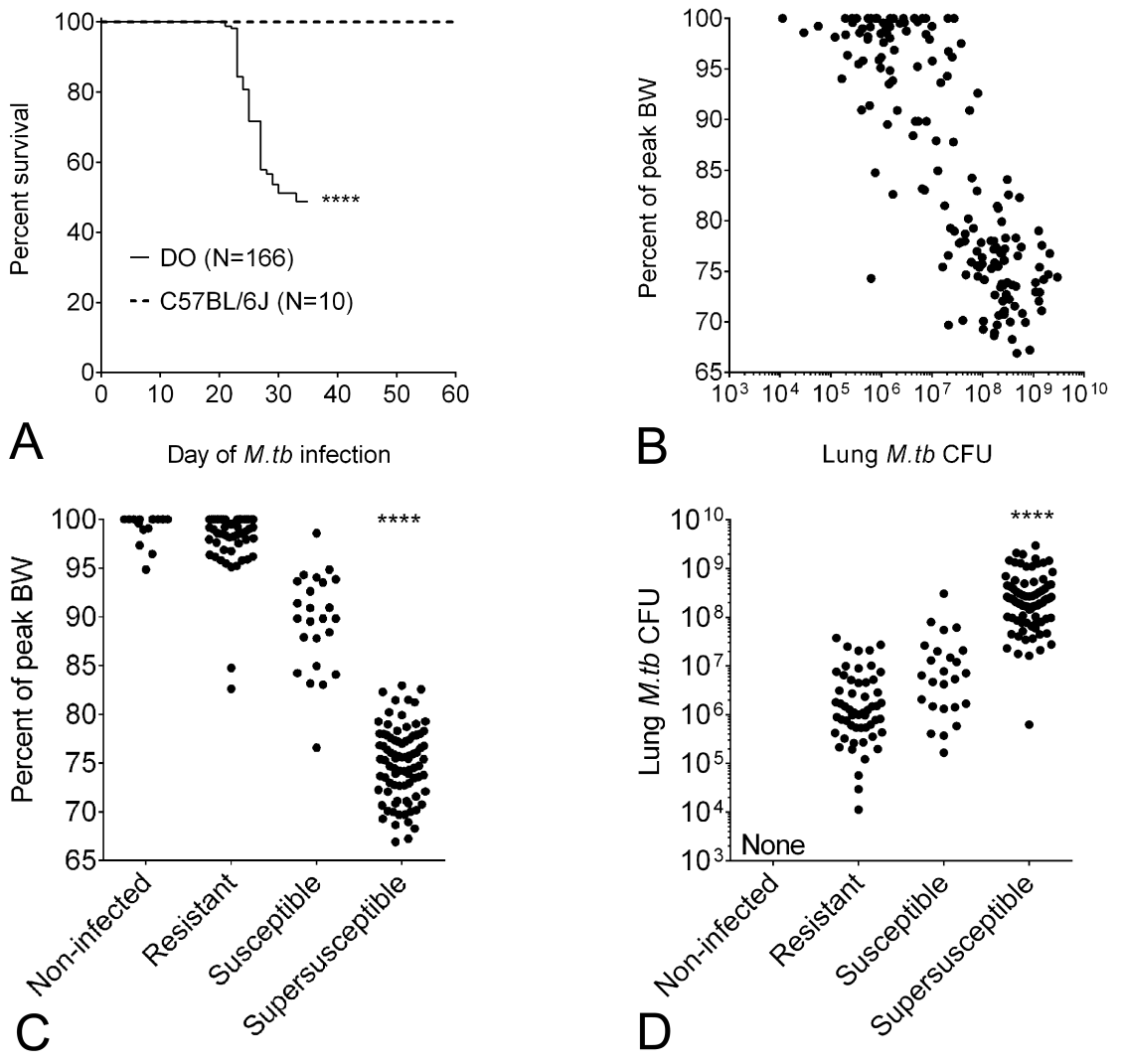


Figure 6.1: ***M. tuberculosis*** **infection and TB disease in DO mice. 8-week-old, female, non-sibling, DO mice** $(N = 166)$ **and C57BL/6J mice** $(N = 10)$ **were infected with** 100 **M.tb bacilli by aerosol. Survival was determined by euthanasia due to morbidity or at day** 35 **of infection, whichever came first. Survival of DO mice compared to the parental C57BL/6J strain was analyzed by Log rank test** $^{****}p < 0.001$ **(A). Body weight and lung** ***M. tuberculosis*** **burden, were strong disease indicators and inversely related to each other, Spearman** $r - 0.79, p < 0.001, 95\%$ **CI** -0.84 **to** -0.72 **(B). Weight (C) and lung** ***M. tuberculosis*** **burden (D) were analyzed by ANOVA with Tukeys post-test,** $^{****}p < 0.0001$**. Data are combined from 2 independent experiments.**

detected: Supersusceptible, Susceptible, and Resistant. Figures 6.1C and 6.1D show that Supersusceptible mice lost the most body weight and had the highest lung bacterial burdens; Susceptible mice were intermediate for both; and Resistant mice lowest for both. Figure 6.2 shows common lung lesions for each susceptibility class with overlap as the lesions represent a biological spectrum. The lungs of Supersusceptible DO mice typically contained regions of lung tissue and macrophage necrosis, capillary thrombosis, neutrophils, and few lymphocytes. The lungs of Resistant mice typically contained small granulomas with macrophages and lymphocytes, but little necrosis or neutrophils, which resembles the C57BL/6 founder strain [236, 237]. Susceptible DO mice had mixed patterns. These findings link susceptibility of DO mice to similar lesions (necrosis, neutrophils) in human patients with active TB [220, 221, 206, 238, 239, 240, 241] and in other susceptible inbred mice [219, 242, 243, 244], none of which are DO founder strains. Together these results indicate that lung necrosis and neutrophilic influx are common disease pathways, and likely reflect many different genetic, molecular, and cellular mechanisms.

6.3.2 Correlates of Disease in *M. tuberculosis* Infected DO Mice

Figure 6.3 depicts the cytokine/chemokine data. Correlations were performed between disease (survival, relative weight loss, bacterial burden) and lung; disease and blood/plasma; and then lung and blood/plasma. Correlates of survival were identical to weight and bacterial burden (not shown). Lung and body weight correlations are shown in Table 6.1. Lung and *M. tuberculosis* CFU correlations are shown in Table 6.2. Dead cells in the lungs, lung CXCL1, CXCL2, CXCL5, and TNF were very strong or strong disease correlates, supporting prior work [194, 221, 245]. Lung cytokines (IFN-γ, IL-2, IL-12, IL-10) had weak or very weak correlations with disease and variable statistical significance.

Table 6.3 and Table 6.4 show correlations between blood/plasma cytokines/chemokines and disease (weight, *M. tuberculosis* lung burden). Only plasma CXCL1 had moderately strong, statistically significant correlations. Plasma CXCL5 had no correlation with weight or *M. tuberculosis* lung CFU, likely because some Resistant mice have high levels (Figure 3). CXCL2 was not detectable in plasma. All cy-

Table 6.1: Correlations of Lung Features and Body Weight

	Correlation (r)	Strength of r	95% CI	Sample size	P value of r
CXCL1	−0.75	Strong	−0.81 to −0.66	146	< 0.0001
CXCL2	−0.74	Strong	−0.81 to −0.66	146	< 0.0001
CXCL5	−0.56	Strong	−0.82 to −0.50	146	< 0.0001
TNF	−0.62	Strong	−0.71 to −0.51	166	< 0.0001
Dead cells	−0.60	Strong	−0.71 to −0.43	80	< 0.0001
IFN-γ	−0.30	Weak	−0.44 to −0.15	163	< 0.0001
IL-2	−0.24	Weak	−0.42 to −0.04	97	0.0183
IL-12	0.36	Weak	0.21 to 0.48	161	0.02
IL-10	0.04	Very weak	−0.12 to 0.20	150	0.61
IFN-γ/TNF	0.41	Moderate	0.27 to 0.53	161	< 0.0001
IFN-γ/IL-10	−0.26	Weak	−0.41 to −0.09	144	0.0016
IL-2/IFN-γ	0.15	Very weak	−0.006 to 0.30	94	0.90

tokines and cytokine ratios (IFN-γ/TNF, IFN-γ/IL-10, and IL-2/IFN-γ) had weak or very weak correlations, some with statistical significance. These findings in DO mice reflect the human condition, where statistically significant but weak correlations with *M. tuberculosis* burden or disease severity are reported using these same cytokines and ratios [246, 247, 248, 249, 250].

In our data set, only plasma CXCL1 correlated with lung CXCL1 levels, and as expected from Fig 6.3, Supersusceptible mice had significantly higher amounts in lung and plasma than Susceptible, Resistant, and Non-infected mice ($p < 0.001$, ANOVA with Tukeys post-test). No other significant correlations between plasma

Table 6.2: Correlations of Lung Features and *M. tuberculosis* Lung CFU

	Correlation (r)	Strength of r	95% CI	Sample size	P value of r
CXCL1	0.83	Very strong	0.77 to 0.87	146	< 0.0001
CXCL2	0.81	Very strong	0.74 to 0.86	146	< 0.0001
CXCL5	0.62	Strong	0.50 to 0.71	146	< 0.0001
TNF	0.66	Strong	0.56 to 0.74	166	< 0.0001
Dead cells	0.59	Moderate	0.41 to 0.70	80	< 0.0001
IFN-γ	0.32	Moderate	0.17 to 0.45	163	< 0.0001
IL-2	0.19	Very weak	-0.17 to 0.38	97	0.06
IL-12	-0.30	Weak	-0.44 to -0.15	161	< 0.0001
IL-10	0.009	Very weak	-0.15 to 0.17	150	0.91
IFN-γ/TNF	-0.43	Moderate	0.55 to -0.29	163	< 0.0001
IFN-γ/IL-10	0.25	Weak	0.08 to 0.40	144	0.0031
IL-2/IFN-γ	-0.06	Very weak	-0.21 to 0.10	163	0.45

and lung levels were present including CXCL5, cytokines, and cytokine ratios (not shown). Multidimensional analyses (2D-correlation coefficients [251] and quadratic discriminant analysis [252, 253]) also failed to detect significant relationships between lung and blood/plasma profiles.

Together, these results show that lung neutrophil chemokines and necrosis are TB disease correlates but immune cytokines are not. Of the molecules that we measured, plasma CXCL1 may be a good biomarker for neutrophil-associated lung damage, which has been suggested previously [194]. CXCL5 was a lung correlate of disease, but not a peripheral correlate, because some Resistant mice had high plasma levels.

Table 6.3: Correlations of Plasma/Blood Features and Body Weight

	Correlation (r)	Strength of r	95% CI	Sample size	P value of r
CXCL1	−0.50	Moderate	−0.61 to −0.37	159	< 0.0001
CXCL5	−0.15	Very weak	−0.01 to 0.29	159	0.065
TNF	0.31	Weak	0.15 to 0.45	160	< 0.0001
IFN-γ	0.25	Weak	0.09 to 0.30	161	0.0014
IL-2	0.30	Weak	0.15 to 0.44	161	< 0.0001
IL-12	0.15	Very weak	−0.008 to 0.30	160	0.0557
IL-10	−0.07	Very weak	−0.07 to 0.23	161	0.34
IFN-γ/TNF	0.21	Weak	0.05 to 0.38	137	0.0106
IFN-γ/IL-10	0.24	Weak	0.06 to 0.40	126	0.0075
IL-2/IFN-γ	0.32	Weak	−0.40 to −0.09	150	0.0022

Note for Tables 6.1-6.4: 8-week-old, female, non-sibling, DO mice ($N = 166$) were infected with 100 M.tb bacilli by aerosol. Cytokines and chemokines in lungs were directly quantified in homogenized tissue. Blood cytokines (IL-2, IFN-γ, IL-12, TNF, IL-10) were quantified after antigen-stimulation of heparinized whole blood. Neutrophil chemokines (CXCL5, CXCL1) were directly quantified in plasma. Since data were not normally distributed, Spearman correlation coefficients (r) were calculated and interpreted as very weak ($0 - 0.19$), weak ($0.20 - 0.39$), moderate ($0.40 - 0.59$), strong ($0.60 - 0.79$), or very strong ($0.80 - 1.0$). The 95% confidence interval, sample size, and p-values are reported.

Table 6.4: Correlations of Plasma/Blood Features and *M. tuberculosis* Lung CFU

	Correlation (r)	Strength of r	95% CI	Sample size	P value of r
CXCL1	0.53	Moderate	0.40 to 0.63	159	< 0.0001
CXCL5	-0.10	Very weak	-0.25 to 0.06	159	0.23
TNF	-0.27	Weak	-0.41 to -0.12	160	0.0005
IFN-γ	-0.24	Weak	-0.38 to -0.08	161	0.0023
IL-2	-0.22	Weak	-0.36 to -0.06	161	0.0060
IL-12	-0.07	Very weak	-0.23 to 0.086	160	0.35
IL-10	-0.09	Very weak	-0.24 to 0.07	161	0.27
IFN-γ/TNF	-0.18	Very weak	-0.35 to -0.01	137	0.0313
IFN-γ/IL-10	-0.29	Weak	-0.45 to -0.12	126	0.0009
IL-2/IFN-γ	-0.25	Weak	-0.40 to -0.09	150	0.0022

6.3.3 Neutrophil Chemokines Can Classify Relative Susceptibility to *M. tuberculosis*

Statistical and machine learning methods identified important features from our 17 variables. ANOVA extracted features with statistical differences followed by Tukeys multiple comparison tests to identify pairs of classes whose averages were significantly different (Table 6.5). These statistical results support observations (Fig. 6.3) and correlations (Tables 6.1-6.4) such that Supersusceptible DO mice had high lung TNF, CXCL1, CXCL2, CXCL5 and high plasma CXCL1. Non-infected mice had low to no IFN-γ, IL-12, IL-2 or IL-10. No features appeared unique for Resistant or Susceptible mice. Table 6.5 further reveals that none of these features alone can discriminate the four different classes and that some statistical

results do not make biological sense. For example, using *M. tuberculosis* CFUs, Non-infected mice could not be identified by statistical methods because the confidence interval for CFUs in Resistant mice extended below zero. Therefore, a combination of features may better parse the data into biologically relevant classes. This led us to use machine learning methods, which to our knowledge have not been applied to experimental *M. tuberculosis in vivo*.

Machine learning identified the same molecular features that were strong disease correlates and/or had some statistical ability to distinguish classes: lung CXCL1, CXCL2, CXCL5, TNF, IFN-γ, IL-12; and two blood features IL-2 and TNF. Classification trees produced the most accurate models. To validate the trees, we utilized leave-one-outcross-validation on a set of 78 mice from the first experiment that had complete data for all parameters. Cross-validation allows some generalizability for model development and testing, but we chose a more stringent approach by testing models against completely independent data. Trees using the strongest disease correlates, CXCL1, CXCL2, CXCL5 (Fig. 6.4) and TNF, CXCL1, CXCL2, and CXCL5 (not shown) had the highest accuracy rates of 77% and 58%, respectively. Trees with TNF confused Resistant with Supersusceptible mice, which may reflect different roles of TNF in resistance and disease ([245, 254, 255, 256, 257, 258]).

Many additional trees were generated and tested but accuracy did not improve, ranging between $38-61\%$ (Table 6.6). Models with accuracy $> 50\%$ reflected TNF as a dominant feature. Organizing data by anatomic site or by function (IL-2, IL-12, IL-10, IFN-γ) with or without TNF, and with or without CXCL1, CXCL2, and CXCL5 did not improve accuracy, either. Neither did generating models using cytokine ratios (not shown).

Overall, the classification tree based on neutrophil chemokines CXCL1, CXCL2, and CXCL5 performed well and was particularly good at discriminating Supersusceptible DO mice from all other categories. This model also confirms inferences from inspecting raw data (Fig. 6.4) and statistical analyses (correlations and ANOVA/MCT) but has the benefit of being easily testable using independent data. The model was not perfect. It misclassified Resistant and Susceptible mice, and misclassified Non-infected mice as Resistant in testing. We next investigated how

Table 6.5: Ability of single features to distinguish each susceptibility class, as using analysis followed by multiple comparison tests. The rows correspond to features while the columns represent the classes. Cells within each row show the features which separate pairs of classes with statistical confidence by ANOVA followed by Tukeys post-test ($p < 0.05$). For example, the feature body weight for the Supersusceptible (SS) class is significantly different from all other classes, as is body weight for the Susceptible (S) class. Body weight for Resistant and Non-infected classes are also significantly different from SS and S, but not from each other. In other words, body weight alone cannot discriminate R from N. Features with cells containing an X could not separate any classes from each other due to overlapping confidence intervals.

Feature	Supersusceptible (SS)	Susceptible (S)	Resistant (R)	Non-infected (N)
Body weight	All	All	SS, S	SS, S
M. tb CFU	All	SS	SS	SS
Lung CXCL1	All	SS	SS	SS
Lung CXCL2	All	SS	SS	SS
Lung CXCL5	R, N	x	SS	SS
Lung TNF	All	SS	SS	SS
Lung IFN-γ	N	N	N	All
Lung IL-12	N	N	N	All
Lung IL-10	N	N	N	All
Lung	Insufficient data for text experiment			
Plasma CXCLI	x	x	x	x
Plasma CXCL5	x	x	x	x
Blood TNF	All	SS	SS	SS
Blood IFN-γ	x	x	x	x
Blood IL-12	S	SS	x	x
Blood IL-10	x	x	x	x
Blood IL-2	x	N	x	S

Table 6.6: Data were acquired from an experiment of *M. tuberculosis* infected and non-infected DO mice, using 78 mice that had all parameters quantified. Trees were validated by the leave-one-out method, and then tested using data from 60 mice in an independent experiment. Results are summarized from classification trees and confusion matrices that included combinations of neutrophil chemokines (CXCL1, CXCL2, CXCL5), TNF, and immune molecules (IFN-γ, IL-12, IL-10, IL-2). Accuracy is reported as overall accuracy for training and testing, separated by a semi-colon. Comments summarize the strengths and weakness of the models generated.

Source	Features	Accuracy (%)	Comments
Lung, plasma	CXCL1, CXCL2, CXCL5	63; 77	This is Fig. 6.4. Best at discriminating SS and N. Sometimes confuses S and R.
Lung, blood, plasma	CXCL1, CXCL2, CXCL5, TNF	71; 58	Good at discriminating SS. Confuses R with SS.
Lung	CXCL1, CXCL2, CXCL5, TNF	55; 53	Confuses SS, S, and R. Good at discriminating N.
Lung, blood, plasma	CXCL1, CXCL2, CXCL5, TNF, IFN-γ, IL-12, IL-10	61; 55	Good at discriminating SS and N. Confuses S and R.
Lung	CXCL1, CXCL2, CXCL5, TNF, IFN-γ, IL-12, IL-10	50; 58	Confuses SS, S, and R. Good at discriminating N.
Blood, plasma	CXCL1, CXCL2, CXCL5, TNF, IFN-γ, IL-12, IL-10	57; 60	Good at discriminating SS. Confuses S, R, and N.
Lung, blood, plasma	CXCL1, CXCL2, CXCL5, TNF, IFN-γ, IL-12, IL-10, IL-2	61; 55	Good at discriminating SS and N. Confuses S and R.
Lung, blood	FN-γ, IL-12, IL-10, IL-2	38; 47	Confuses SS, S, and R. Good at discriminating N.
Lung	FN-γ, IL-12, IL-10	53; 45	Confuses SS, S, and R. Good at discriminating N.

mice were misclassified.

6.3.4 Some DO Mice Misclassified by Neutrophil Chemokines Have Unique Phenotypes

To determine whether the magnitude of cytokine/chemokine responses influenced feature selection, data were normalized. However, normalization had no effect (not shown). The classification tree developed in training can only assess misclas-

sifications in testing. As the numbers of misclassifications were too small for data mining, all results were manually inspected to identify recording errors (none were found) and to seek biologically relevant explanations. Thirteen of 60 test mice were misclassified by Fig. 6.4. Three were Non-infected mice misclassified as Resistant, reflecting higher values in the test (average 0.8 ng/ml) than the training (average 0.4 ng/ml) data. Thus, these Non-infected mice were erroneously classified as Resistant because of inter-experiment variability. Additional training data from Non-infected mice would improve these misclassifications.

Seven Resistant mice were misclassified. Two were misclassified because of high CXCL1 and CXCL5, possibly due macrophage production without substantial neutrophil recruitment, evident by microscopy. Of note, these are not the two Resistant mice lower body weights (Fig. 6.1C). Five misclassified Resistant mice had CXCL1 less than the 1st node (2.91 ng/ml), and were misclassified at various other nodes, despite having typical lung lesions. Three misclassified Susceptible mice had lung CXCL1 less than the 1st node (2.91 ng/ml) and these mice had a spectrum of lung lesions. Only one Supersusceptible mouse was misclassified. This mouse was clearly different by molecular features (CXCL1 less than 2.91 ng/ml) and by microscopy. Instead of neutrophil infiltration and necrosis, morbidity reflected massive lung infiltration by non-necrotic macrophages.

Examination of misclassified mice was helpful. Misclassified Non-infected mice reflected inter-experiment variability, which could be improved with additional training data. Two misclassified Resistant mice better fit the Susceptible class, showing that the model may better classify ambiguous cases. It was unclear how Susceptible mice were misclassified because the numbers were tiny. The one misclassified Supersusceptible mouse had a unique phenotype: marked infiltration by non-necrotic macrophages without neutrophils.

6.4 Discussion

For the first time, we investigate survival, morbidity, bacterial burden, and a set of immune and inflammatory molecules in *M. tuberculosis* infected DO mice. Cytokines/chemokines were chosen because their absence/blockade alters immunity

or disease (IFN-γ, IL-12, TNF, IL-2, IL-10) [196, 197, 198, 199, 200, 201, 202] or because of recent discoveries as drivers of detrimental inflammation or potential biomarkers (CXCL1, CXCL2, CXCL5) [194, 221]. We did not perform extensive profiling, but our results do provide a rational basis for pursuing such studies. Given the remarkable susceptibility of some DO mice, reducing the infectious dose may be important to slow lung damage, and improve detection of individual granulomas and their features before morbidity develops. This approach has been successful with the highly susceptible C3HeB/FeJ substrain to identify genes and loci that contribute to susceptibility [209, 210, 259, 260].

Approximately half of the DO population had significantly reduced survival as compared the C57BL/6J founder strain (Fig. 6.1A), which typically survives > 1 year given the same dose [190, 193, 261, 262, 263]. This suggests that other DO founder strains contribute genetic material that enhances susceptibility to *M. tuberculosis*. To our knowledge, no studies have identified genes or pathways that contribute to susceptibility using DO mice, and further research is needed. Mechanisms of DO susceptibility are unlikely to directly involve the private Ipr1 polymorphism responsible for lung granuloma necrosis in the C3HeB/FeJ substrain [191, 210, 213] as the polymorphism is absent from DO founder strains (analyses performed by Dr Daniel Gatti in Dr Gary Churchills laboratory, The Jackson Laboratory, Bar Harbor, ME using the Mouse Phylogeny Viewer and the Sanger Mouse Genome Database accessed on October 6th 2014 [264], `http://www.sanger.ac.uk/resources/mouse/genomes/`). Therefore, we expect that future studies in DO mice will identify new genes/loci that contribute to macrophage and lung granuloma necrosis. We anticipate this will be relevant to some human cases, because not all TB patients have polymorphisms in SP110 (the homologue to mouse Ipr1) [214], and some populations have SP110 polymorphisms that are not associated with TB [215, 265].

Neutrophils in the blood and lungs are now well-recognized characteristics of humans with active TB [194, 216, 239] and some inbred mice during early [218] and chronic [219] infection. Since neutrophils have characteristic morphologies, we used light microscopy to identify mature neutrophils in lung sections of DO mice. Clearly, microscopy is limited to structural analyses, and additional assays are needed to de-

fine functional significance, and to investigate other immune cells (subsets of T cells, B cells, macrophages, dendritic cells etc) that participate in the host response to *M. tuberculosis*. We did not attempt to discriminate immature neutrophils (bands) because the nuclear morphology likely overlaps with immature myeloid-derived suppressor cells recently described in lethal murine tuberculosis [223].

Consequences of neutrophil influx in response to *M. tuberculosis* infection have not all been proven, but growing evidence suggests that enzymes, oxidants, and inflammatory cytokines/chemokines cause damage and drive cycles of detrimental inflammation [194, 220, 221, 266, 267, 268, 269]. Given that some DO mice recruit abundant neutrophils to the lungs and have high levels of some neutrophil chemokines, DO mice will also be a good model to identify genetic control of signals that attract neutrophils, consequences of neutrophil recruitment, and possibly identify and test targets for preventative or therapeutic intervention. We have not investigated mechanisms, but the pathways likely involve CXCL1, CXCL2, and CXCL5 through the CXCR2 receptor, as well as Type I interferon signaling, IL-17 and S100A8/A9 [194, 220, 221, 270, 271] and undiscovered pathways. In DO mice, lung levels of CXCL1, CXCL2, and CXCL5 were strong disease correlates, but only CXCL1 was a good peripheral biomarker that reflected lung disease and lung levels, supporting the use of CXCL1 as a biomarker [194]. We suggest CXCL1 could also be explored as a predictive marker of disease progression in longitudinal studies using DO mice infected with a lower dose of *M. tuberculosis*.

Previous computational approaches in *M. tuberculosis* infection have focused on immunologic cytokines, effects of aging on immune resistance, and roles for TNF [254, 272, 273, 274]. To our knowledge, we are the first to apply machine learning to build models for classifying relative susceptibility to *M. tuberculosis* in a genetically diverse population *in vivo*, and then to test model performance using independent data. Here, we used observational, bimodal correlation analyses, ANOVA followed by MCT, and multiple types of machine learning algorithms analysis to confirm that the important molecular features were consistent. Models that included disease indicators (survival, weight, bacterial burden) were highly accurate ($> 95\%$) in discriminating all 4 classes (Supersusceptible, Susceptible, Resistant,

and Non-infected) but we did not show or discuss them because they provide less biological insight. We focused instead on molecular features from the lungs and blood/plasma. Multiple unsupervised and supervised methods were used, but we pursued the Classification tree method, a supervised approach, because it yielded most accurate models in training and testing. In fact, the most accurate tree used only three neutrophil chemokines: CXCL1, CXCL2, and CXCL5 to discriminate the 4 classes (Supersusceptible, Susceptible, Resistant and Non-infected) with 77% accuracy on test data. This tree best discriminated Supersusceptible mice from all other classes, but performed less well for Susceptible, Resistant, and Non-infected mice. Further investigation revealed that Non-infected mice were misclassified because CXCL2 levels were slightly higher in the test data than the training data. Misclassifications for Susceptible and Resistant mice were not always clear, indicating that none of the molecules we measured were discriminatory for these classes. Thus, identification of other features is needed to improve accuracy.

We recognize that computational approaches could have improved accuracy such as collapsing the classes into 2 (Supersusceptible and other); internal replication of data to yield groups of equivalent numbers; or combining independent experiments to generate a larger training data set and then testing accuracy by randomly leaving out a portion of the data. Although these changes would have increased performance, we chose not to pursue them because they may not reflect the biological reality.

We also investigated four cytokines important for resistance to *M. tuberculosis* (TNF, IFN-γ, IL-12, IL-2), and one that contributes to susceptibility (IL-10) [196, 197, 198, 199, 200, 201, 202]. Although DO mice produced these cytokines, relative susceptibility did not reflect a reduction of proinflammatory cytokines nor an over production of the immune suppressive cytokine IL-10. With the exception of TNF (which was a strong disease correlate in the lung and weak in the periphery), cytokines and cytokine ratios were poor correlates of disease. Machine learning models based on cytokines could not classify susceptibility either, but were very good at discriminating Non-infected mice. These findings reflect recent reviews [206, 207, 208] and meta-analyses [275], which conclude that antigen-specific

TH1 responses are not good correlates of protective immunity or TB disease in humans and more sophisticated investigations are needed. Our results here suggest that the DO population may be a good experimental model to help resolve these questions.

We recognize that a mouse model cannot recapitulate all characteristics of human pulmonary TB. However, DO mice provide an immunologically intact, genetically diverse experimental population that shares some similarities with active TB (necrosis, neutrophils), and, we suspect that reducing the *M. tuberculosis* dose will allow discovery of more relevant human-like phenotypes. Seventeen variables across approximately 150 mice may be considered a small data set, but the concept of applying machine learning to *in vivo M. tuberculosis* infection is new and important, as machine learning provides an accelerated platform to extract information from large and complex data. Thus, machine learning will be particularly useful for large genetic, gene expression, protein array, and computer-aided image analysis studies using DO mice, which can be used to create models capable of predicting outcome before TB disease develops.

Acknowledgements

This paper is the responsibility of the authors and does not represent views of the National Cancer Institute, the National Institute of Allergy and Infectious Diseases, or the National Institutes of Health. Support was provided by Tufts University, the Cummings School of Veterinary Medicine, and the Department of Infectious Disease and Global Health (GB); in part by NIH T35 OD010963-04 (PI: Anwer to RC), by R01CA134451 (PIs: Gurcan, Lozanski) from the National Cancer Institute, by R56 AI111823 (PIs: Campos-Neto, Beamer), and by 5R01HL059836 (PI: Kramnik) from the National Institute of Allergy and Infectious Diseases.

We thank Dr Daniel Gatti in Dr Gary Churchills group (The Jackson Laboratory) for bioinformatics analyses and Dr Joanne Turner (The Ohio State University, Columbus, OH) for virulent *M. tuberculosis* Erdman. We thank Ms Melanie Harwood, Mr Curtis Rich, Mr Donald Girouard, and Dr Donna Akiyoshi at the NERBL. We thank Ms Frances Brown and the histology staff at the Cummings School

of Veterinary Medicine, Tufts University. BEI Resources, NIAID, NIH provided the plasmid pMRLB.7, containing Gene Rv3875 (Protein ESAT-6) (NR-13280) and ESAT-6 Recombinant Protein Reference Standard (NR-14868). The Comparative Pathology & Mouse Phenotyping Shared Resource, Department of Veterinary Biosciences and the Comprehensive Cancer Center, The Ohio State University, Columbus, OH, supported in part by grant P30 CA016058, National Cancer Institute, Bethesda, MD is acknowledged for digital slide scanning using Aperio ScanScope.

Specific Contributions

IK, BY, MG, and GB conceived experiments, analyzed data, interpreted results and prepared the manuscript; GB also performed technical assays and figure preparation; MK performed machine learning, statistical analyses, and manuscript preparation; ND: performed machine learning; RC performed data analysis; DS, and SM performed technical assays. All authors had final approval of the manuscript.

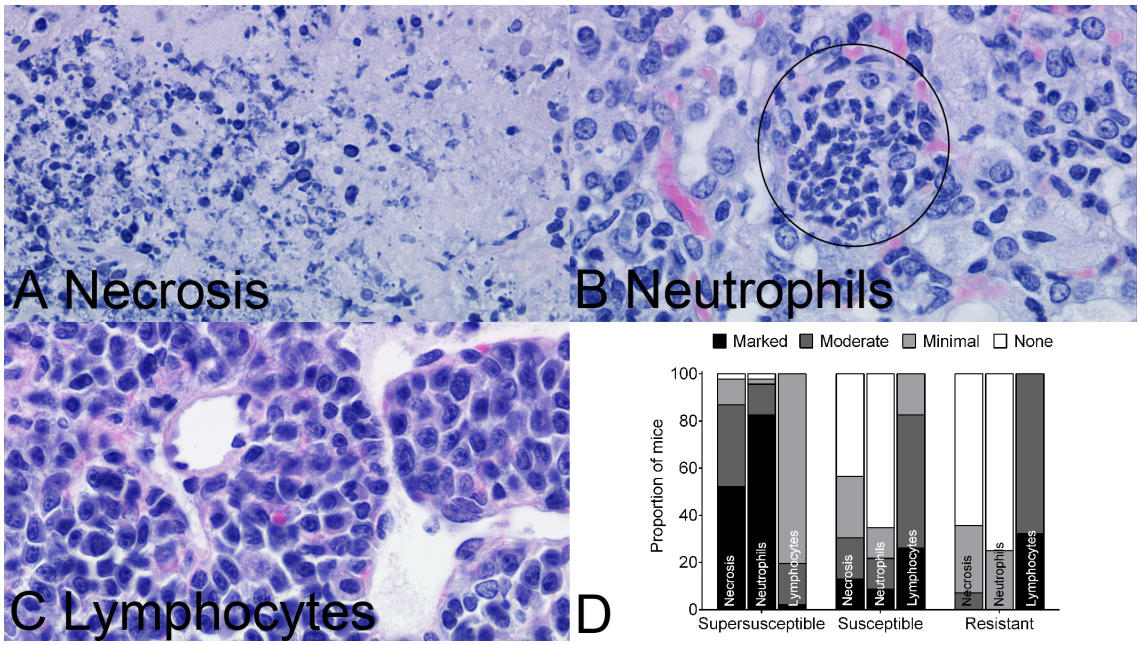


Figure 6.2: Lung lesions in *M. tuberculosis* infected DO mice. 8-week-old, female, non-sibling, DO mice ($N = 97$) infected with 100 M.tb bacilli by aerosol. Hematoxylin and eosin stained lung sections were evaluated Lung from a Supersusceptible mouse is shown magnified 10x (A), with (B) showing lung tissue and inflammatory cell necrosis magnified 200x. Lung from a Resistant mouse is shown magnified 10x (C) showing a non-necrotic lesion with a relative abundance of perivascular lymphocytes magnified 200x (D). Black circles show examples of mature neutrophils in an alveolar space (E) and amongst macrophages within a granuloma (F) magnified 400x. Two lung lobes from each mouse were scored for relative severity of each lesion type by a board certified veterinary pathologist (GB) without knowledge of the groups, and data compiled (G).

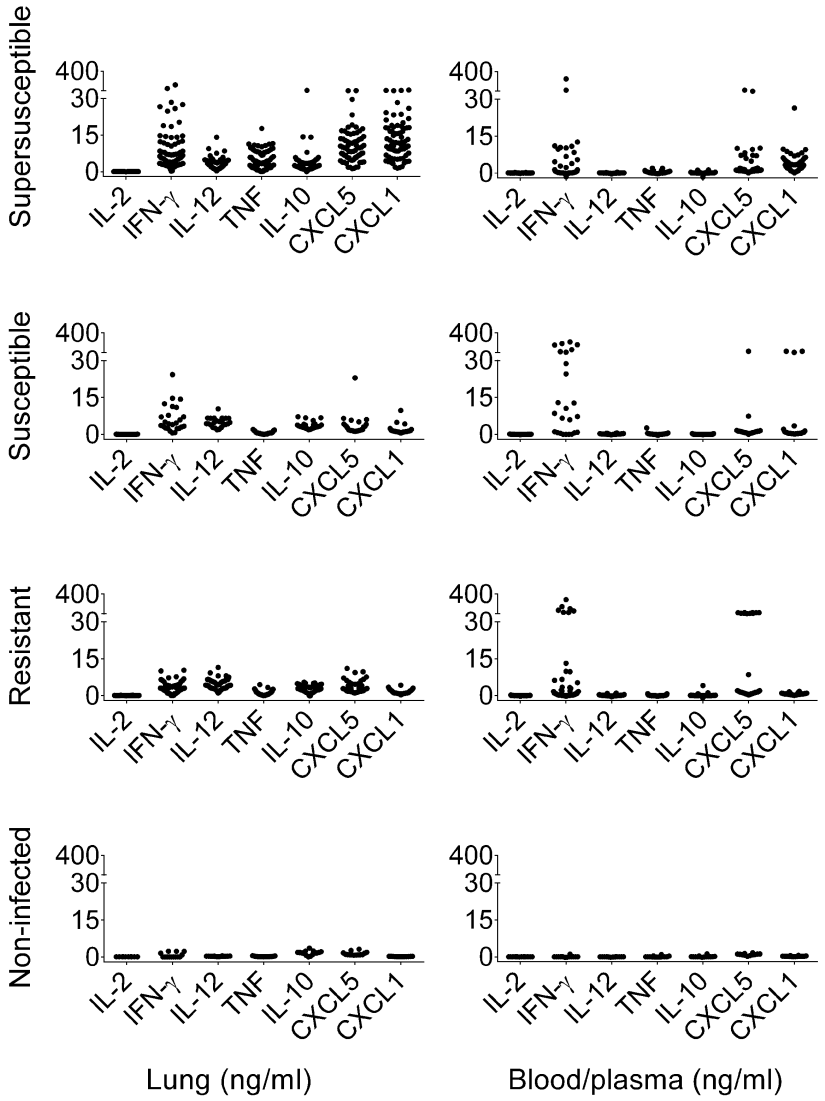
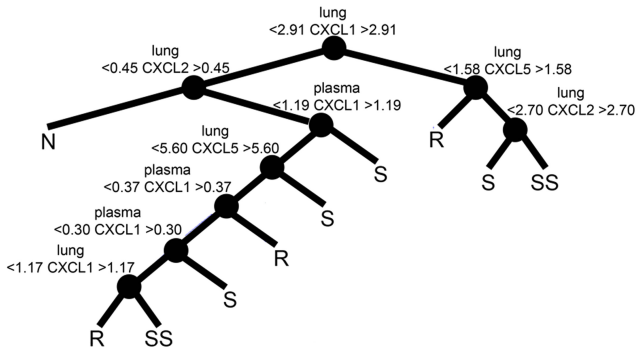


Figure 6.3: Molecular profiles of lung and blood/plasma in DO mice. 8-week-old, female, non-sibling, DO mice $(N = 166)$ **were infected with** 100 **M.tb bacilli by aerosol. Cytokines and chemokines in lungs were quantified in homogenized tissue. Blood cytokines (IL-2, IFN-γ, IL-12, TNF, IL-10) were quantified after antigen-stimulation. Neutrophil chemokines (CXCL1, CXCL5) were directly quantified in plasma but CXCL2 was not detectable. Features are grouped as follows: T cell responses (IL-2, IFN-γ), macrophages (IL-12, TNF, IL-10) and neutrophils (CXCL1, CXCL5, CXCL2) from left to right. Each dot represents the average of duplicate or triplicate samples from one mouse. Data are combined from two independent experiments.**

A



B

Actual Class	Predicted Class Training data (63% accurate)			
	SS	**S**	**R**	**N**
SS	22	3	4	0
S	3	9	7	0
R	6	4	11	1
N	0	0	1	7

C

Actual Class	Predicted Class Test data (77% accurate)			
	SS	**S**	**R**	**N**
SS	29	1	0	0
S	1	0	2	0
R	1	6	17	0
N	0	0	3	0

Figure 6.4: Classification tree and confusion matrices based on neutrophil chemokines. Training data was acquired from *M. tuberculosis* infected and non-infected DO mice, using 78 mice with all parameters measured. Trees were validated by the leave-one-out method, and then tested using data from an independent experiment with 60 mice will all parameters measured. The tree based on lung and plasma CXCL1, CXCL2, and CXCL5 is shown (A) with confusion matrices for training (B) and testing (C).

CHAPTER 7
CONCLUSIONS AND FUTURE WORK

We investigated complex biological systems that can be represented as time-series with varying modalities. These modalities included image-driven, signal-based, and pseudo-time-series data. For each of these different data sets, we developed an appropriate model to accurately represent the evolution of the underlying system. For this purpose we utilized techniques such as dynamic graphs, level sets, and autoregressive processes. We showed that coupling of techniques can significantly improve the modeling ability of the algorithm.

We started with the problem of modeling cleft formation in the first round of branching morphogenesis in the mouse submandibular salivary gland 2. We developed a dynamic-graph-based-descriptive-model that could accurately depict cleft elongation, bud outgrowth, and *de novo* cleft formation. We devised new region of interest features to quantify the growth of the tissue, and compare it to the biological ground truth time-lapse video data. We compared our model with an on-lattice Monte Carlo model, and demonstrated the improved tissue topology, as well as considerable savings in simulation time. We enhanced the descriptive model with the addition of regression-based growth trends, and thus, reduced the requirement for expensive *in vivo* experiments.

We further improved on our dynamic-graph-based-model by coupling it with level sets 3. We demonstrated that this coupling benefits the graph model by creating a more realistic gland topology, while adding cellular constraints to the level sets. This coupling is an early attempt at combining the physics-based-continuous-models with discrete cellular models. This model further outperformed the on-lattice Monte Carlo model, with a further reduction in computational complexity.

We next moved to problems related to epileptic seizures, namely, seizure localization and early seizure prediction 4. Statistics show that about 1% of the world's population suffers from epilepsy, making it a very serious concern. Using the synchronous firing of neurons in the brain as an indicator of seizure, we built graphs to

capture this synchronization event. The scalp EEG electrodes measuring neuronal activity act as the vertices of this synchronization graph, and edges are established between vertices when they record similarly high spiking activity. Computing measurements on these time-evolving synchronization graphs that exhibit significant change during seizures, we were able to localize the seizures temporally. We also used a tensor-based approach using measurements on individual EEG electrodes to spatially localize the seizure to a particular side of the brain.

Although seizure localization is an advancement for the field of epilepsy, the "holy grail" is early seizure detection, or seizure prediction. The objective here is to identify characteristic signatures of epileptic seizures prior to the seizure event. This objective gets more complicated when additional constraints are added including predicting the seizure only a reasonable period before the event, and not missing any seizure events. This early detection period is referred to as the seizure prediction horizon, and cannot extend indefinitely as this would negatively impact the patient's life. Also, every seizure-like event must be identified, i.e. false positives are acceptable within reason. Missing a seizure because it was classified as a false positive would make the system unusable in a real-world situation. On the other hand, predicting a seizure for every seizure-like event would again make it very difficult for the patient to lead a normal life.

Given these constraints, we developed a system that couples measurements from synchronization graphs with an autoregressive process (5 and [276]. This allowed us to describe the system's current state in terms of its previous state (i.e. making use of its Markov property). The novelty in our approach lies in the fact that we considered the seizure to be an anomaly, similar to eye-blinks or patient motion. Thus, we learnt on the non-ictal behavior profiles of the EEG recordings. We improved on this learning by utilizing the concepts of transfer learning and manifold alignment, where we learnt from other recordings of the same patient, as well as recordings of other patients. We then used a one-dimensional error profile based on our prediction of the state of the system vs. the actual state of the system. We found that our algorithm could predict a seizure a reasonable period of time prior to its occurrence.

The last problem we attempted was to investigate the susceptibility of heterogeneous mice populations to *Mycobacterium tuberculosis* 6. We were able to identify important features apart from the obvious indicators that helped with this supervised classification problem. We presented results on the sets of features that could differentiate between the various susceptibility profiles.

In the following sections, we present a few future directions for the work presented in this thesis.

7.1 Identification of Seismic Events in Seismograph Recordings

A key objective of the San Andreas Fault Observatory at Depth (SAFOD) project is to drill through the rupture patch of a small repeating earthquake. The size of the rupture patch is on the order of tens of meters; hence knowing its location to a comparable degree of accuracy is vital to SAFOD's success. Professor Steven W. Roecker has been involved with the analysis of data from the deployment of dense, temporary networks of seismic stations (PASO-UNO and PASO-DOS) in the vicinity of the drill site. There are 35 seismic stations in the PASO-DOS network continuously recording seismic activity. One of the most reliable measurements taken by these stations is the vertical component of motion. Given this network of stations and their recording, our goal is to build a graph with the stations as vertices, establishing edges between stations when they record similar seismic activity. This is along the same vein as the seizure localization problem discussed in Chapter 4. There are multiple objectives for this project:

1. Automated cataloging of seismic events – Given the vast amounts of data recorded by seismographs, it is extremely difficult for seismologists to reliably analyze all the data. By developing an earthquake detection algorithm, we can attempt to assuage this problem, immensely reducing the human hours required to comb through all the seismograph recordings.

2. Identify the epicenter of the earthquake – The second objective is to identify the origin region (or epicenter) of the earthquake. We would like to recognize

which station(s) recorded the earthquake at the earliest time. This would allow seismologists to focus their efforts on performing risk estimation in that particular region.

3. Create a real-time (or near-real time) algorithm – The "holy grail" for earthquake detection would be an algorithm that can perform seismic event detection in real-time (or near-real time). This objective has massive real-world applications such as monitoring volcanoes, and mining and drilling facilities.

We have preliminary results illustrating that changes in the features calculated on the seismograph recordings indicate presence of seismic events.

7.2 Seizure Prediction for Longer Continuous Recordings and its Implications

Our current analysis on seizure prediction was limited to data sets with a single seizure, and only an hour of EEG recording. We would like to test the veracity of the algorithm on longer data sets (up to 4 days) with multiple seizures. We have acquired the data through a new collaboration with the EEG Lab in the Developmental Disability Center at Mt Sinai-Roosevelt Hospital and the Mount Sinai Epilepsy Center at the Mount Sinai School of Medicine.

The larger corpus of data would allow us to tune the parameters of the algorithm making it more robust. Also, our current analysis is limited to temporal lobe epilepsy; there are multiple types of epilepsy including childhood absence, juvenile myoclonic, occipital lobe, frontal lobe, Lennox-Gastaut syndrome, and focal epilepsy. We would like to test whether our algorithm is generalized enough to adapt to any type of epilepsy. Further future directions for this project include a more rigorous understanding of the impact of the various components. More specifically, we'd like to investigate the influence of Independent Component Analysis (ICA) vs. transfer learning. Another open question is with regards to the similarity among data sets from which the transfer is taking place - establishing a methodology to determine how much can be learnt from the transfer sets without a reduction in performance.

LITERATURE CITED

[1] Banerjee J, Chan YC, Sen CK (2011) MicroRNAs in skin and wound healing. *Physiol Genomics* 43(10):543–556.

[2] St-Supery V et al. (2011) Wound healing assessment: does the ideal methodology for a research setting exist? *Ann Plast Surg* 67(2):193–200.

[3] Ma'ayan A (2009) Insights into the organization of biochemical regulatory networks using graph theory analyses. *J Biol Chem* 284(9):5451–5455.

[4] Davies JA (2007) *Branching Morphogenesis*. (Springer Science & Business Media, Berlin).

[5] Affolter M et al. (2003) Tube or not tube: Remodeling epithelial tissues by branching morphogenesis. *Dev Cell* 4(1):11–18.

[6] Denny PC, Ball WD, Redman RS (1997) Salivary glands: A paradigm for diversity of gland development. *Crit Rev Oral Biol Med* 8(1):51–75.

[7] Patel VN, Rebustini IT, Hoffman MP (2006) Salivary gland branching morphogenesis. *Differentiation* 74(1):349–364.

[8] Ray S, Fanti JA, Macedo DP, Larsen M (2014) LIM kinase regulation of cytoskeletal dynamics is required for salivary gland branching morphogenesis. *Mol Biol Cell* 25(16):2393–2407.

[9] Steinberg Z et al. (2005) FGFR2b signaling regulates *ex vivo* submandibular gland epithelial cell proliferation and branching morphogenesis. *Development* 132(6):1223–1234.

[10] Nogawa H, Takahashi Y (1991) Substitution for mesenchyme by basement-membrane-like substratum and epidermal growth factor in inducing branching morphogenesis of mouse salivary epithelium. *Development* 112(3):855–861.

[11] Koyama N et al. (2008) Signaling pathways activated by epidermal growth factor receptor or fibroblast growth factor receptor differentially regulate branching morphogenesis in fetal mouse submandibular glands. *Dev Growth Differ* 50(7):565–576.

[12] Miettinen PJ et al. (1997) Impaired lung branching morphogenesis in the absence of functional EGF receptor. *Dev Biol* 186(2):224–236.

[13] Weller A, Sorokin L, Illgen E, Ekblom P (1991) Development and growth of mouse embryonic kidney in organ culture and modulation of development by soluble growth factor. *Dev Biol* 144(2):248–261.

[14] Luetteke NC et al. (1999) Targeted inactivation of the EGF and amphiregulin genes reveals distinct roles for EGF receptor ligands in mouse mammary gland development. *Development* 126(12):2739–2750.

[15] Miettinen PJ, Huotari MA, Koivisto T, Ustinov J, Palgi, J. ea (2000) Impaired migration and delayed differentiation of pancreatic islet cells in mice lacking egf-receptors. *Development* 127(12):2617–2627.

[16] Kashimata M, Gresik EW (1997) Epidermal growth factor system is a physiological regulator of development of the mouse fetal submandibular gland and regulates expression of the alpha6-integrin subunit. *Dev Dyn* 208(2):149–161.

[17] Morita K, Nogawa H (1999) EGF-dependent lobule formation and FGF7-dependent stalk elongation in branching morphogenesis of mouse salivary epithelium in-vitro. *Dev Dyn* 215(2):148–154.

[18] Häärä O, Koivisto T, Miettinen PJ (2009) EGF-receptor regulates salivary gland branching morphogenesis by supporting proliferation and maturation of epithelial cells and survival of mesenchymal cells. *Differentiation* 77(3):298–306.

[19] Larsen M, Wei C, Yamada KM (2006) Cell and fibronectin dynamics during branching morphogenesis. *J Cell Sci* 119(16):3376–3384.

[20] Sakai T, Larsen M, Yamada KM (2003) Fibronectin requirement in branching morphogenesis. *Nature* 423(6942):876–881.

[21] Onodera T et al. (2010) Btbd7 regulates epithelial cell dynamics and branching morphogenesis. *Science* 329(5991):562–565.

[22] Zaczek A, Brandt B, Bielawski KP (2005) The diverse signaling network of EGFR, HER2, HER3 and HER tyrosine kinase receptors and the consequences for therapeutic approaches. *Histol Histopathol* 20(3):1005–1015.

[23] Kashimata M, Sakagami HW, Gresik EW (2000) Intracellular signalling cascades activated by the EGF receptor and/or by integrins, with potential relevance for branching morphogenesis of the fetal mouse submandibular gland. *Eur J Morphol* 38(4):269–275.

[24] Larsen M et al. (2003) Role of PI 3-kinase and PIP3 in submandibular gland branching morphogenesis. *Dev Biol* 255(1):178–191.

[25] Koyama N et al. (2012) Extracellular regulated kinase5 is expressed in fetal mouse submandibular glands and is phosphorylated in response to epidermal growth factor and other ligands of the Erbb family of receptors. *Dev Growth Differ* 54(9):801–808.

[26] Bogdan S, Klämbt C (2001) Epidermal growth factor receptor signaling. *Curr Biol* 11(8):292–295.

[27] Setty Y, Dalfó D, Korta DZ, Hubbard EJA, Kugler H (2012) A model of stem cell population dynamics: in-silico analysis and in-vivo validation. *Development* 139(1):47–56.

[28] Turing AM (1952) The chemical basis of morphogenesis. *Philos Trans R Soc Lond B Biol Sci* 237(641):37–72.

[29] Eden M (1961) A two-dimensional growth process. *Dyn Fract Surf* 4(1):223–239.

[30] Steinberg MS (1963) Reconstruction of tissues by dissociated cells. Some morphogenetic tissue movements and the sorting out of embryonic cells may have a common explanation. *Science* 141(3579):401–408.

[31] Oster GF, Murray JD, Harris AK (1983) Mechanical aspects of mesenchymal morphogenesis. *J Embryol Exp Morphol* 78(1):83–125.

[32] Murray JD, Oster GF (1984) Generation of biological pattern and form. *Math Med Biol* 1(1):51–75.

[33] Rejniak KA (2007) An immersed boundary framework for modeling the growth of individual cells: An application to early tumour development. *J Theor Biol* 247(1):186–204.

[34] Rejniak KA et al. (2010) Linking changes in epithelial morphogenesis to cancer mutations using computational modeling. *PLoS Comput Biol* 6(8):e1000900.

[35] Metzger R, Klein O, Martin G, Krasnow M (2008) The branching programme of mouse lung development. *Nature* 453(7196):745–750.

[36] Andrew DJ, Ewald AJ (2010) Morphogenesis of epithelial tubes: Insights into tube formation, elongation, and elaboration. *Dev Biol* 341(1):34–55.

[37] Hartmann D, Miura T (2006) Modelling in vitro lung branching morphogenesis during development. *J Theor Biol* 242(4):862–872.

[38] Srivathsan A, Menshykau D, Michos O, Iber D (2013) Dynamic Image-Based Modelling of Kidney Branching Morphogenesis, *Lecture Notes in Computer Science: Computational Methods in Systems Biology.* (Springer, Berlin), pp. 106–119.

[39] Lubkin SR, Li Z (2002) Force and deformation on branching rudiments: Cleaving between hypotheses. *Biomech Model Mechanobiol* 1(1):5–16.

[40] Pouille PA, Farge E (2008) Hydrodynamic simulation of multicellular embryo invagination. *Phys Biol* 5(1):015005.

[41] Fleury V (2011) A change in the boundary conditions induces a discontinuity of tissue flow in chicken embryos and the formation of the cephalic fold. *Eur Phys J E Soft Matter* 34(7):1–13.

[42] Ashkin J, Teller E (1943) Statistics of two-dimensional lattices with four components. *Phys Rev Lett* 64(5):178–184.

[43] Ray S et al. (2013) Cell-based multi-parametric model of cleft progression during submandibular salivary gland branching morphogenesis. *PLoS Comput Biol* 9(11):e1003319.

[44] Haghighat A, Wagner JC (2003) Monte carlo variance reduction with deterministic importance functions. *Prog Nucl Energ* 42(1):25–53.

[45] Diestel R (2000) *Graduate Texts in Mathematics: Graph Theory.* (Springer, Berlin).

[46] Mason O, Verwoerd M (2006) Graph theory and networks in biology. *IET Syst Biol* 1(2):89–119.

[47] Jeong H, Mason SP, Barabási AL, Oltvai ZN (2001) Lethality and centrality in protein networks. *Nature* 411(6833):41–42.

[48] Wagner A (2001) The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18(7):1283–1292.

[49] Barabási AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. *Nat Rev Genet* 5(2):101–113.

[50] Samanta M, Liang S (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *Proc Natl Acad Sci U S A* 100(22):12579–12583.

[51] Jeong H, Oltvai ZN, Barabási AL (2003) Prediction of protein essentiality based on genomic data. *ComPlexUs* 1(1):19–28.

[52] Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* 407(6804):651–654.

[53] Overbeek R et al. (2000) WOT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research* 28(1):123–125.

[54] Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL (2002) Hierarchical organization of modularity in metabolic networks. *Science* 297(5586):1551–1555.

[55] Wagner A, Fell D (2001) The small world inside large metabolic networks. *Proc Biol Sci* 268(1478):1803–1810.

[56] Tong AH et al. (2004) Global mapping of yeast genetic interaction network. *Science* 303(5659):808–813.

[57] Lee T et al. (2002) Transcriptional regulatory networks in saccharomyces cerevisiae. *Science* 298(5594):799–804.

[58] Barabási AL (2007) Network medicine – from obesity to the "diseasome". *N Engl J Med* 357(4):404–407.

[59] Eubank S et al. (2004) Modelling disease outbreaks in realistic urban social networks. *Nature* 429(6988):180–184.

[60] Watts DJ, Strogatz SH (1998) Collective dynamics of "small-world" networks. *Nature* 393(6684):440–442.

[61] Bilgin C, Lund AW, Can A, Plopper GE, Yener B (2010) Quantification of three-dimensional cell-mediated collagen remodeling using graph theory. *PLoS ONE* 5(9):e12783.

[62] Bilgin CC et al. (2012) Multiscale feature analysis of salivary gland branching morphogenesis. *PLoS ONE* 7(3):e32906.

[63] Bilgin C, Bullough P, Plopper GE, Yener B (2010) ECM-aware cell-graph mining for bone tissue modeling and classification. *Data Min Knowl Discov* 20(3):416–438.

[64] Bilgin C et al. (2010) Cell-graph modeling of salivary gland morphology. *In Proceedings of IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pp. 1427–1430.

[65] Dhulekar N et al. (2012) A novel dynamic graph-based computational model for predicting salivary gland branching morphogenesis. *In Proceedings of 2012 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 1–8.

[66] Pizer SM et al. (1987) Adaptive histogram equalization and its variations. *Lect Notes Comput Sc* 39(3):355–368.

[67] Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cyber* 9(1):62–66.

[68] Abramoff MD, Magalhaes PJ, Ram SJ (2004) Image processing with ImageJ. *Biol Med Phys Biomed* 11(7):36–42.

[69] Bunks C, Kylander K (2000) *Grokking the GIMP.* (New Riders Publishing, San Francisco).

[70] Koprnicky M, Maher A, Mohamed K (2004) Contour Description Through Set Operations on Dynamic Reference Shapes, *Lecture Notes in Computer Science: Image Analysis and Recognition.* (Springer, Berlin), pp. 400–407.

[71] Chakroborty S, Saha G (2010) Feature selection using singular value decomposition and qr factorization with column pivoting for text-independent speaker identification. *Speech Commun* 52(9):693–709.

[72] Gunduz C, Yener B, Gultekin SH (2004) The cell-graphs of cancer. *Bioinformatics* 20(1):145–151.

[73] Hartigan JA (1975) *Clustering Algorithms.* (Wiley, New York).

[74] Manning CD, Raghavan P, Schütze H (2008) *Introduction to Information Retrieval.* (Cambridge University Press, Cambridge).

[75] Hogeweg P (2000) Evolving mechanisms of morphogenesis: on the interplay between differential adhesion and cell differentiation. *J Theor Biol* 203(4):317–333.

[76] Merks RM, Perryn ED, Shrinifard A, Glazier JA (2008) Contact-inhibited chemotaxis in de novo and sprouting blood-vessel growth. *PLoS Comput Biol* 4(9):e1000163.

[77] Ising E (1925) Beitrag zur theorie des ferromagnetismus. *Zeitschrift fur Physik* 31(1):253–258.

[78] Rejniak KA, Anderson ARA (2011) Hybrid models of tumor growth. *Wiley Interdiscip Rev Syst Biol Med.* 3(1):115–125.

[79] Maini P, Deutsch A, Dormann S (2007) *Cellular Automaton Modeling of Biological Pattern Formation: Characterization, Applications, and Analysis.* (Birkhäuser, Berlin).

[80] Stott E, Britton L, Glazier JA, Zajac M (1999) Stochastic simulation of benign avascular tumour growth using the Potts model. *Math Comput Model* 30(5):183–198.

[81] Shirinifard A et al. (2009) 3D multi-cell simulation of tumor growth and angiogenesis. *PLoS ONE* 4(10):e7190.

[82] Beyer T, Meyer-Hermann M (2009) Multiscale modeling of cell mechanics and tissue organization. *IEEE Eng Med Biol Mag* 28(2):38–45.

[83] Galle J, Loeffler M, Drasdo D (2005) Modeling the effect of deregulated proliferation and apoptosis on the growth dynamics of epithelial cell populations in vitro. *Biophys J* 88(1):62–75.

[84] Klipp E et al. (2009) *Systems Biology: A Textbook*. (Wiley-Blackwell, Hoboken).

[85] Osher S, Sethian JA (1988) Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulation. *J Comput Phys* 79(1):12–49.

[86] Courant R, Isaacson E, Rees M (1952) On the solution of nonlinear hyperbolic differential equations by finite differences. *Commun Pur Appl Math* 5(3):243–255.

[87] Fedkiw R, Sapiro G, Shu CW (2001) Shock capturing, level sets and PDE based methods in computer vision and image processing: A review of Osher's contributions. *J Comput Phys* 185(1):309–341.

[88] Shoeb A et al. (2004) Patient-specific seizure onset detection. *In Proceedings of 26th International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 419–422.

[89] Saab ME, Gotman J (2005) A system to detect the onset of epileptic seizures in scalp EEG. *Clin Neurophysiol* 116(2):427–442.

[90] Päivinen N et al. (2005) Epileptic seizure detection: A nonlinear viewpoint. *Comput Methods Programs Biomed* 79(2):151–159.

[91] van Putten MJ, Kind T, Visser F, Lagerburg V (2005) Detecting temporal lobe seizures from scalp EEG recordings: A comparison of various features. *Clin Neurophysiol* 116(10):2480–2489.

[92] Mormann F et al. (2005) On the predictability of epileptic seizures. *Clin Neurophysiol* 116(3):569–587.

[93] Lehnertz K, Litt B (2005) The first international collaborative workshop on seizure prediction: Summary and data description. *Clin Neurophysiol* 116(3):493–505.

[94] Le Van QM et al. (2005) Preictal state identification by synchronization changes in long-term intracranial EEG recordings. *Clin Neurophysiol* 116(3):559–568.

[95] Iasemidis LD et al. (2005) Long-term prospective on-line real-time seizure prediction. *Clin Neurophysiol* 116(3):532–544.

[96] D'Alessandro M et al. (2005) A multi-feature and multi-channel univariate selection process for seizure prediction. *Clin Neurophysiol* 116(3):506–516.

[97] Schelter B et al. (2007) Seizure prediction: The impact of long prediction horizons. *Epilepsy Res* 73(2):213–217.

[98] Mohseni HR, Maghsoudi A, Shamsollahi MB (2006) Seizure detection in EEG signals: A comparison of different approaches. *In Proceedings of 28th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 6724–6727.

[99] Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. *J Mach Learn Res* 3(1):1157–1182.

[100] Van Gils M et al. (1997) Signal processing in prolonged EEG recordings during intensive care. *IEEE Eng Med Biol* 16(6):56–63.

[101] Hjorth B (1970) EEG analysis based on time domain properties. *Electroen Clin Neuro* 29(3):306–310.

[102] Akin M (2002) Comparison of wavelet transform and FFT methods in the analysis of EEG signals. *J Med Syst* 26(3):241–247.

[103] Rosso OA, Martin MT, Plastino A (2002) Brain electrical activity analysis using wavelet-based informational tools. *Physica A* 313(3-4):587–608.

[104] Acar E, Aykut-Bingol C, Bingol H, Bro R, Yener B (2007) Multiway analysis of epilepsy tensors. *Bioinformatics* 23(13):i10–i18.

[105] Estienne F, Matthijs N, Massart DL, Ricoux P, D L (2001) Multiway modeling of high-dimensionality electroencephalographic data. *Chemometr Intell Lab* 58(1):59–72.

[106] Jackson JE (1980) Principal components and factor analysis: Part I - principal components. *J Qual Technol* 12(4):201–213.

[107] Jackson JE (1981) Principal components and factor analysis: Part II - additional topics related to principal components. *J. Qual Technol* 13(1):46–58.

[108] Miwakeichi F et al. (2004) Decomposing EEG data into space-time-frequency components using parallel factor analysis. *NeuroImage* 22(3):1035–1045.

[109] Harshman RA (1970) Foundations of the PARAFAC procedure: Models and conditions for an explanatory multi-modal factor analysis. *UCLA Phone* 16(1):1–84.

[110] Mørup M, Hansen LK, Herrmann CS, Parnas J, Arnfred SM (2006) Parallel factor analysis as an exploratory tool for wavelet transformed event-related EEG. *NeuroImage* 29(3):938–947.

[111] Martinez-Montes E, Valdes-Sosa PA, Miwakeichi F, Goldman RI, Cohen MS (2004) Concurrent EEG/FMRI analysis by multiway partial least squares. *NeuroImage* 22(3):1023–1034.

[112] Acar E, Bingol CA, Bingol H, Yener B (2006) Computational analysis of epileptic focus localization. *In Proceedings of Fourth IASTED International Conference on Biomedical Engineering*, pp. 317–322.

[113] Bullmore E, Sporns O (2009) Complex brain networks: graph theoretical analysis of structural and functional systems. *Nat Rev Neurosci* 10(3):186–198.

[114] Stam C, van Straaten E (2012) The organization of physiological brain networks. *Clin Neurophysiol* 123(6):1067–1087.

[115] Schindler KA, Bialonski S, Horstmann MT, Elger CE, Lehnertz K (2008) Evolving functional network properties and synchronizability during human epileptic seizures. *Chaos* 18(3):033119.

[116] Kramer MA, Kolaczyk ED, Kirsch HE (2008) Emergent network topology at seizure onset in humans. *Epilepsy Res* 79(2):173–186.

[117] Douw L et al. (2010) Epilepsy is related to theta band brain connectivity and network topology in brain tumor patients. *BMC Neurosci* 11(1):103.

[118] Mahyari A, Aviyente S (2014) Identification of dynamic functional brain network states through tensor decomposition. *In Proceedings of 39th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2014)*.

[119] Jasper HH (1999) The ten-twenty electrode system of the International Federation. *Electroencephalogr Clin Neurophysiol Suppl* 52(1):3–6.

[120] Smith SJM (2005) EEG in the diagnosis, classification, and management of patients with epilepsy. *J Neurol Neurosurg Psychiatry* 76(2):ii2–ii7.

[121] van de Velde M, Ghosh IR, Cluitmans PJM (1999) Context related artefact detection in prolonged EEG recordings. *Comput Methods Programs Biomed* 60(3):183–196.

[122] Levy WJ (1987) The effect of epoch length on the identification of changes in EEG power spectra. *Anesthesiology* 66(4):489–495.

[123] Levy WJ (1986) The effect of epoch length on the identification of changes in EEG power spectra. *Anesthesiology* 65(3A):A539.

[124] Nunez PL et al. (1997) EEG coherency: I: statistics, reference electrode, volume conduction, laplacians, cortical imaging, and interpretation at multiple scales. *Electroencephalogr Clin Neurophysiol* 103(5):499–515.

[125] Stam CJ, Nolte G, Daffertshofer A (2007) Phase lag index: Assessment of functional connectivity from multi channel EEG and MEG with diminished bias from common sources. *Hum Brain Mapp* 28(11):1178–1193.

[126] Lachaux JP, Rodriguez E, Martinerie J, J. VF (1999) Measuring phase synchrony in brain signals. *Hum Brain Mapp* 8(4):194–208.

[127] Tucker LR (1964) The extension of factor analysis to three-dimensional matrices. *Rec Res Psy* 1(1):109–127.

[128] Hotelling H (1931) The generalization of student's ratio. *Ann Math Stat* 2(3):360–378.

[129] Fisher RS et al. (2005) Epileptic seizures and epilepsy: Definitions proposed by the International League Against Epilepsy (ILAE) and the International Bureau for Epilepsy (IBE). *Epilepsia* 46(4):470–472.

[130] Esteller R et al. (2005) Continuous energy variation during the seizure cycle: Towards an on-line accumulated energy. *Clin Neurophysiol* 116(3):517–526.

[131] Elger CE (2001) Future trends in epileptology. *Curr Opin Neurol* 14(2):185–186.

[132] Viglione SS, Walsh GO (1975) Epileptic seizure prediction. *Electroencephalogr Clin Neurophysiol* 39(4):435–436.

[133] Rogowski, Z. GI, Bental E (1981) On the prediction of epileptic seizures. *Biol Cybern* 42(1):9–15.

[134] Salant Y, Gath I, Henriksen O (1998) Prediction of epileptic seizures from two-channel EEG. *Med Biol Eng Comput* 36(5):549–56.

[135] Siegel A, Grady CL, Mirsky AF (1982) Prediction of spike-wave bursts in absence epilepsy by EEG power-spectrum signals. *Epilepsia* 23(1):47–60.

[136] Mormann F, Andrzejak RG, Elger CE, Lehnertz K (2007) Seizure prediction: The long and winding road. *Brain* 130(2):314–333.

[137] Tzallas AT, Tsipouras MG, Fotiadis DI (2007) The use of time-frequency distributions for epileptic seizure detection in EEG recordings. *In Proceedings of 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pp. 3–6.

[138] Tzallas AT, Tsipouras MG, Fotiadis DI (2009) Epileptic seizure detection in EEGs using time-frequency analysis. *IEEE Trans Inf Technol Biomed* 13(5):703–710.

[139] Güler NF, Übeyli ED, Güler I (2005) Recurrent neural networks employing lyapunov exponents for EEG signals classification. *Expert Syst Appl* 29(3):506–514.

[140] Kannathal N, Choo ML, R. AU, Sadasivan PK (2005) Entropies for detection of epilepsy in EEG. *Comput Methods Programs Biomed* 80(3):187–194.

[141] Srinivasan V, Eswaran C, Sriraam N (2005) Artificial neural network based epileptic detection using time-domain and frequency-domain features. *J Med Syst* 29(6):647–660.

[142] Chisci L et al. (2010) Real-time epileptic seizure prediction using AR models and support vector machines. *IEEE Trans Biomed Eng* 57(5):1124–1132.

[143] Chandaka S, Chatterjee A, Munshi S (2009) Cross-correlation aided support vector machine classifier for classification of EEG signals. *Expert Syst Appl* 36(2):1329–1336.

[144] Fisher N, Talathi SS, Carney PR, Ditto WL (2008) Epilepsy detection and monitoring. *Quant EEG Anal Meth Appl* 1(1):157–183.

[145] Giannakakis G, Sakkalis V, Pediaditis M, Tsiknakis M (2015) Methods for Seizure Detection and Prediction: An Overview, *Neuromethods: Modern Electroencephalographic Assessment Techniques*. (Springer, Berlin), pp. 131–157.

[146] Harrison MA, Frei MG, Osorio I (2005) Accumulated energy revisited. *Clin Neurophysiol* 116(3):527–531.

[147] Jouny CC, Franaszczuk PJ, Bergey GK (2005) Signal complexity and synchrony of epileptic seizures: Is there an identifiable preictal period? *Clin Neurophysiol* 116(3):552–558.

[148] Mormann F, Lehnertz K, David P, Elger CE (2000) Mean phase coherence as measure for phase synchronization and its application to the EEG of epilepsy patients. *Physica D* 144(3):358–369.

[149] Litt B et al. (2001) Epileptic seizures may begin hours in advance of clinical onset: A report of five patients. *Neuron* 30(1):51–64.

[150] Cranstoun SD et al. (2002) Time-frequency spectral estimation of multichannel EEG using the auto-slex method. *IEEE Trans Biomed Eng* 49(9):988–996.

[151] Mormann F et al. (2003) Epileptic seizures are preceded by a decrease in synchronization. *Epilepsy Res* 53(3):173–185.

[152] Corsini J, Shoker L, Sanei S, Alarcon G (2006) Epileptic seizure predictability from scalp EEG incorporating constrained blind source separation. *IEEE Trans Biomed Eng* 53(5):790–799.

[153] Le Van QM, Navarro V, Martinerie J, Baulac M, Varela FJ (2003) Toward a neurodynamical understanding of ictogenesis. *Epilepsia* 44(12):30–43.

[154] Mirowski P, Madhavan D, LeCun Y, Kuzniecky R (2009) Classification of patterns of EEG synchronization for seizure prediction. *Clin Neurophysiol* 120(11):1927–1940.

[155] Chu CC, Bronzino JD (1995) *The Biomedical Engineering Handbook*. (CRC Press, Boca Raton).

[156] Muthuswamy J, Thakor NV (1998) Spectral analysis methods for neurological signals. *J. Neurosci. Methods* 83(1):1–14.

[157] Hazarika N, Chen JZ, Tsoi AC, Sergejew A (1997) Classification of EEG signals using the wavelet transform. *In Proceedings of the 13th International Conference on Digital Signal Processing Proceedings*, pp. 89–92.

[158] Murali S, Kulish VV (2007) Modeling of evoked potentials of electroencephalograms: An Overview. *Digit. Signal Process.* 17(3):665–674.

[159] Lytton WW (2008) Computer modeling of Epilepsy. *Nat. Rev. Neurosci.* 9(8):626–637.

[160] Anderson NR et al. (2009) An offline evaluation of the autoregressive spectrum for electrocorticography. *IEEE Trans Biomed Eng* 56(3):913–916.

[161] Subasi A, Alkan A, Koklukaya E, Kiymik MK (2005) Wavelet neural network classification of EEG signals by using AR models with MLE processing. *Neural Netw* 18(7):985–997.

[162] Liu HS, Zhang T, Yang FS (2002) A multistage, multimethod approach for automatic detection and classification of epileptiform EEG. *IEEE Trans Biomed Eng* 49(12):1557–1566.

[163] Strogatz SH (2001) Exploring complex networks. *Nature* 410(6825):268–276.

[164] Newman MEJ (2003) The structure and function of complex networks. *SIAM Rev* 45(2):167–256.

[165] Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU (2006) Complex networks: Structure and dynamics. *Phys Rep* 424(4–5):175 – 308.

[166] Barrat A, Barthelemy M, Vespignani A (2008) *Dynamical Processes on Complex Networks*. (Cambridge University Press, Cambridge).

[167] Kuhnert MT, Elger CE, Lehnertz K (2010) Long-term variability of global statistical properties of epileptic brain networks. *Chaos* 20(4):043126.

[168] Wu H, Li X, Guan X (2006) Networking Property During Epileptic Seizure with Multi-channel EEG Recordings, *Advances in Neural Networks - ISNN 2006*. (Springer, Berlin), pp. 573–578.

[169] Dhulekar N et al. (2014) Graph-theoretic analysis of epileptic seizures on scalp EEG recordings. *In Proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 155–163.

[170] Pan SJ, Yang Q (2010) A survey on transfer learning. *IEEE Trans Knowl Data Engg* 22(10):1345–1359.

[171] Wang C, Mahadevan S (2008) Manifold alignment using procrustes analysis. *In Proceedings of the 25th International Conference on Machine Learning*, pp. 1120–1127.

[172] Comon P (1994) Independent component analysis – A new concept. *Signal Process.* 36(3):287–314.

[173] Comon P, Jutten C (2010) *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. (Academic Press, Oxford).

[174] Jutten C, Herault J (1991) Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. *Signal Process* 24(1):1–10.

[175] Delorme A, Sejnowski TJ, Makeig S (2007) Enhanced detection of artifacts in EEG data using higher-order statistics and independent component analysis. *Neuroimage* 34(4):1443–1449.

[176] Jung TP et al. (2000) Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37(2):163–178.

[177] Lee TW, Girolami M, Sejnowski TJ (1999) Independent component analysis using an extended infomax algorithm for mixed sub-gaussian and super-gaussian sources. *Neural Comput* 11(2):417–441.

[178] Rodriguez-Lujan I, Huerta R, Elkan C, Cruz CS (2010) Quadratic programming feature selection. *J. Mach. Learn. Res.* 11(1):1491–1516.

[179] Bertsekas DP (1999) *Nonlinear Programming*. (Athena Scientific, Belmont).

[180] Alkan A, Koklukaya E, Subasi A (2005) Automatic seizure detection in EEG using logistic regression and artificial neural network. *J Neurosci Methods* 148(2):167–176.

[181] Osorio I, Zaveri H, Frei M, Arthurs S (2011) *Epilepsy: The Intersection of Neurosciences, Biology, Mathematics, Engineering, and Physics*. (Taylor & Francis, New York).

[182] Harrison DE et al. (2009) Rapamycin fed late in life extends lifespan in genetically heterogeneous mice. *Nature* 460(7253):392–395.

[183] Ferris MT et al. (2013) Modeling host genetic regulation of influenza pathogenesis in the collaborative cross. *PLoS Pathog* 9(2):e1003196.

[184] Logan RW et al. (2013) High-precision genetic mapping of behavioral traits in the diversity outbred mouse population. *Genes Brain Behav* 12(4):424–437.

[185] Recla JM et al. (2014) Precise genetic mapping and integrative bioinformatics in Diversity Outbred mice reveals Hydin as a novel pain gene. *Mamm Genome* 25(5-6):211–222.

[186] Churchill GA, Gatti DM, Munger SC, Svenson KL (2012) The Diversity Outbred mouse population. *Mamm Genome* 23(9-10):713–718.

[187] Svenson KL et al. (2012) High-resolution genetic mapping using the Mouse Diversity outbred population. *Genetics* 190(2):437–447.

[188] Yang H, Bell TA, Churchill GA, de Villena FPM (2007) On the subspecific origin of the laboratory mouse. *Nat Genet* 39(9):1100–1107.

[189] Yang H et al. (2011) Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet* 43(7):648–655.

[190] Medina E, North RJ (1998) Resistance ranking of some common inbred mouse strains to *Mycobacterium tuberculosis* and relationship to major histocompatibility complex haplotype and Nramp1 genotype. *Immunology* 93(2):270–274.

[191] Pichugin AV, Yan BS, Sloutsky A, Kobzik L, Kramnik I (2009) Dominant role of the sst1 locus in pathogenesis of necrotizing lung granulomas during chronic tuberculosis infection and reactivation in genetically resistant hosts. *Am J Pathol* 174(6):2190–2201.

[192] Beamer GL, Flaherty DK, Vesosky B, Turner J (2008) Peripheral blood gamma interferon release assays predict lung responses and *Mycobacterium tuberculosis* disease outcome in mice. *Clin Vaccine Immunol* 15(3):474–483.

[193] Jagannath C et al. (2000) Hypersusceptibility of A/J mice to tuberculosis is in part due to a deficiency of the fifth complement component (c5). *Scand J Immunol* 52(4):369–379.

[194] Gopal R et al. (2013) S100A8/A9 proteins mediate neutrophilic inflammation and lung pathology during tuberculosis. *Am J Respir Crit Care Med* 188(9):1137–1146.

[195] Harrison DE, Astle CM, Niazi MKK, Major S, Beamer GL (2014) Genetically diverse mice are novel and valuable models of age-associated susceptibility to mycobacterium tuberculosis. *Immun Ageing* 11(1):24.

[196] Flynn JL, Chan J (2001) Immunology of tuberculosis. *Annu Rev Immunol* 19(1):93–129.

[197] Cooper AM (2009) Cell mediated immune responses in tuberculosis. *Annu Rev Immunol* 27(1):393.

[198] Gong JH et al. (1996) Interleukin-10 downregulates Mycobacterium tuberculosis-induced Th1 responses and CTLA-4 expression. *Infect Immun* 64(3):913–918.

[199] Zhang J et al. (2011) Interleukin-10 polymorphisms and tuberculosis susceptibility: A meta-analysis. *Int J Tuberc Lung Dis* 15(5):594–601.

[200] Beamer GL et al. (2008) Interleukin-10 promotes Mycobacterium tuberculosis disease progression in CBA/J mice. *J Immunol* 181(8):5545–5550.

[201] Turner J et al. (2002) In vivo IL-10 production reactivates chronic pulmonary tuberculosis in C57BL/6 mice. *J Immunol* 169(11):6343–6351.

[202] Higgins DM et al. (2009) Lack of IL-10 alters inflammatory and immune responses during pulmonary Mycobacterium tuberculosis infection. *Tuberculosis* 89(2):149–157.

[203] Casanova JL, Abel L (2002) Genetic dissection of immunity to mycobacteria: The human model. *Annu Rev Immunol* 20(1):581–620.

[204] Redford PS, Murray PJ, O'Garra A (2011) The role of IL-10 in immune regulation during M. tuberculosis infection. *Mucosal Immunol* 4(3):261–270.

[205] Liang B, Guo Y, Li Y, et al. (2014) Association between IL-10 gene polymorphisms and susceptibility of tuberculosis: evidence based on a meta-analysis. *PloS One* 9(2):e88448.

[206] Nunes-Alves C et al. (2014) In search of a new paradigm for protective immunity to TB. *Nat Rev Microbiol* 12(4):289–299.

[207] Andersen P, Woodworth JS (2014) Tuberculosis vaccines–rethinking the current paradigm. *Trends Immunol* 35(8):387–395.

[208] Modlin RL, Bloom BR (2013) TB or not TB: That is no longer the question. *Sci Transl Med* 5(213):213sr6–213sr6.

[209] Kramnik I, Demant P, Bloom BB (1998) Susceptibility to tuberculosis as a complex genetic trait: analysis using recombinant congenic strains of mice. *Novartis Found Symp* 217(1):120–137.

[210] Pan H et al. (2005) Ipr1 gene mediates innate immunity to tuberculosis. *Nature* 434(7034):767–772.

[211] Kramnik I (2008) Genetic Dissection of Host Resistance to Mycobacterium tuberculosis: the SST1 locus and the LPR1 Gene. *Immunology, Phenotype First: How Mutations Have Established New Principles and Pathways in Immunology* (Springer, Berlin), pp. 123–148.

[212] Sissons J et al. (2009) Multigenic control of tuberculosis resistance: analysis of a QTL on mouse chromosome 7 and its synergism with sst1. *Genes Immun* 10(1):37–46.

[213] Tosh K et al. (2006) Variants in the SP110 gene are associated with genetic susceptibility to tuberculosis in West Africa. *Proc Natl Acad Sci U S A* 103(27):10364–10368.

[214] Png E et al. (2012) Polymorphisms in SP110 are not associated with pulmonary tuberculosis in Indonesians. *Infect Gene Evol* 12(6):1319–1323.

[215] Thye T et al. (2006) No associations of human pulmonary tuberculosis with SP110 variants. *J Med Genet* 43(7):e32.

[216] Barnes PF et al. (1988) Predictors of short-term prognosis in patients with pulmonary tuberculosis. *J Infect Dis* 158(2):366–371.

[217] Berry MPR et al. (2010) An interferon-inducible neutrophil-driven blood transcriptional signature in human tuberculosis. *Nature* 466(7309):973–977.

[218] Keller C et al. (2006) Genetically determined susceptibility to tuberculosis in mice causally involves accelerated and enhanced recruitment of granulocytes. *Infect Immun* 74(7):4295–4309.

[219] Major S, Turner J, Beamer G (2013) Tuberculosis in CBA/J mice. *Vet Pathol* 50(6):1016–1021.

[220] Dorhoi A et al. (2014) Type I IFN signaling triggers immunopathology in tuberculosis-susceptible mice by modulating lung phagocyte dynamics. *Eur J Immunol* 44(8):2380–2393.

[221] Nouailles G et al. (2014) CXCL5-secreting pulmonary epithelial cells drive destructive neutrophilic inflammation in tuberculosis. *J Clin Invest* 124(3):1268–1282.

[222] Nandi B, Behar SM (2011) Regulation of neutrophils by interferon-γ limits lung inflammation during tuberculosis infection. *J Exp Med* 208(11):2251–2262.

[223] Tsiganov EN et al. (2014) Gr-1dimCD11b+ immature myeloid-derived suppressor cells but not neutrophils are markers of lethal tuberculosis infection in mice. *J Immunol* 192(10):4718–4727.

[224] Yang CT et al. (2012) Neutrophils exert protection in the early tuberculous granuloma by oxidative killing of mycobacteria phagocytosed from infected macrophages. *Cell Host Microbe* 12(3):301–312.

[225] Beamer GL, Cyktor J, Carruthers B, Turner J (2011) H-2 alleles contribute to antigen 85-specific interferon-gamma responses during mycobacterium tuberculosis infection. *Cell Immunol* 271(1):53–61.

[226] Beamer GL et al. (2012) CBA/J mice generate protective immunity to soluble Ag85 but fail to respond efficiently to Ag85 during natural *Mycobacterium tuberculosis* infection. *Eur J Immunol* 42(4):870–879.

[227] Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J Roy Stat Soc B Met* 73(3):273–282.

[228] Kornaropoulos EN, Niazi M, Lozanski G, Gurcan MN (2014) Histopathological image analysis for centroblasts classification through dimensionality reduction approaches. *Cytometry A* 85(3):242–255.

[229] Gelfand SB, Ravishankar C, Delp EJ (1989) An iterative growing and pruning algorithm for classification tree design. *In Proceedings of 1989 IEEE International Conference on Systems, Man and Cybernetics*, pp. 818–823.

[230] Rokach L, Maimon O (2005) Top-down induction of decision trees classifiers – A survey. *IEEE Trans Syst Man Cyber C* 35(4):476–487.

[231] Coppersmith D, Hong SJ, Hosking JR (1999) Partitioning nominal attributes in decision trees. *Data Min Knowl Discov* 3(2):197–217.

[232] Kira K, Rendell LA (1992) The feature selection problem: Traditional methods and a new algorithm. *In Proceedings of Association for the Advancement of Artificial Intelligence (AAAI)*, pp. 129–134.

[233] Liu H, Setiono R, et al. (1996) A probabilistic approach to feature selection-a filter solution. *In Proceedings of Thirteenth International Conference on Machine Learning (ICML)*, pp. 319–327.

[234] Zhao Z, Liu H (2007) Spectral feature selection for supervised and unsupervised learning. *In Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 1151–1157.

[235] Boutsidis C, Mahoney MW, Drineas P (2008) Unsupervised feature selection for principal components analysis. *In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 61–69.

[236] Niazi M, Beamer G, Gurcan MN (2014) Detecting and characterizing cellular responses to mycobacterium tuberculosis from histology slides. *Cytometry A* 85(2):151–161.

[237] Vesosky B, Rottinghaus EK, Stromberg P, Turner J, Beamer G (2010) CCL5 participates in early protection against Mycobacterium tuberculosis. *J Leukoc Biol* 87(6):1153–1165.

[238] Leong FJ, Dartois V, Dick T (2010) *A Color Atlas of Comparative Pathology of Pulmonary Tuberculosis*. (CRC Press, Boca Raton).

[239] Eum SY et al. (2010) Neutrophils are the predominant infected phagocytic cells in the airways of patients with active pulmonary tb. *Chest* 137(1):122–128.

[240] Hunter RL, Actor JK, Hwang SA, Karev V, Jagannath C (2014) Pathogenesis of post primary tuberculosis: Immunity and hypersensitivity in the development of cavities. *Ann Clin Lab Sci* 44(4):365–387.

[241] Hunter RL (2011) Pathology of post primary tuberculosis of the lung: An illustrated critical review. *Tuberculosis* 91(6):497–509.

[242] Harper J et al. (2011) Mouse model of necrotic tuberculosis granulomas develops hypoxic lesions. *J Infect Dis* 205(4):595–602.

[243] Lyadova IV et al. (2010) In mice, tuberculosis progression is associated with intensive inflammatory response and the accumulation of Gr-1 cells in the lungs. *PLoS One* 5(5):e10469.

[244] Eruslanov EB et al. (2005) Neutrophil responses to Mycobacterium tuberculosis infection in genetically susceptible and resistant mice. *Infect Immun* 73(3):1744–1753.

[245] Bekker LG et al. (2000) Immunopathologic effects of tumor necrosis factor alpha in murine mycobacterial infection are dose dependent. *Infect Immun* 68(12):6954–6961.

[246] Jamil B et al. (2007) Interferon gamma/IL10 ratio defines the disease severity in pulmonary and extra pulmonary tuberculosis. *Tuberculosis* 87(4):279–287.

[247] Hur YG et al. (2013) Combination of cytokine responses indicative of latent TB and active TB in Malawian adults. *PLoS ONE* 8(11):e79742.

[248] Sahiratmadja E et al. (2007) Dynamic changes in pro- and anti-inflammatory cytokine profiles and gamma interferon receptor signaling integrity correlate with tuberculosis disease activity and response to curative treatment. *Infect Immun* 75(2):820–829.

[249] Millington KA et al. (2007) Dynamic relationship between IFN-gamma and IL-2 profile of Mycobacterium tuberculosis-specific T cells and antigen load. *J Immunol* 178(8):5217–5226.

[250] Suter-Riniker F et al. (2011) Clinical significance of interleukin-2/gamma interferon ratios in Mycobacterium tuberculosis-specific T-cell signatures. *Clin Vaccine Immunol* 18(8):1395–1396.

[251] Noda I (1993) Generalized two-dimensional correlation method applicable to infrared, raman, and other types of spectroscopy. *Appl Spectrosc* 47(9):1329–1336.

[252] Scholkopft B, Mullert KR (1999) Fisher discriminant analysis with kernels. *In Proceedings of the Nineth IEEE Signal Processing Society Workshop on Neural Networks*, pp. 23–25.

[253] Lachenbruch PA, Goldstein M (1979) Discriminant analysis. *Biometrics* 35(1):69–85.

[254] Ray JCJ, Flynn JL, Kirschner DE (2009) Synergy between individual tnf-dependent functions determines granuloma performance for controlling mycobacterium tuberculosis infection. *J Immunol* 182(6):3706–3717.

[255] Fallahi-Sichani M, Schaller MA, Kirschner DE, Kunkel SL, Linderman JJ (2010) Identification of key processes that control tumor necrosis factor availability in a tuberculosis granuloma. *PLoS Comput Biol* 6(5):e1000778.

[256] Law K et al. (1996) Increased release of interleukin-1 beta, interleukin-6, and tumor necrosis factor-alpha by bronchoalveolar cells lavaged from involved sites in pulmonary tuberculosis. *Am J Respir Crit Care Med* 153(2):799–804.

[257] Bekker LG, Maartens G, Steyn L, Kaplan G (1998) Selective increase in plasma tumor necrosis factor-alpha and concomitant clinical deterioration after initiating therapy in patients with severe tuberculosis. *J Infect Dis* 178(2):580–584.

[258] Juffermans NP et al. (1998) Tumor necrosis factor and interleukin-1 inhibitors as markers of disease activity of tuberculosis. *Am J Respir Crit Care Med* 157(4):1328–1331.

[259] Kramnik I, Dietrich WF, Demant P, Bloom BR (2000) Genetic control of resistance to experimental infection with virulent Mycobacterium tuberculosis. *Proc Natl Acad Sci U S A* 97(15):8560–8565.

[260] Yan B, Kirby A, Shebzukhov Y, Daly M, Kramnik I (2006) Genetic architecture of tuberculosis resistance in a mouse model of infection. *Genes Immun* 7(3):201–210.

[261] Kondratieva E, Logunova N, Majorov K, Averbakh Jr M, Apt A (2010) Host genetics in granuloma formation: human-like lung pathology in mice with reciprocal genetic susceptibility to M. tuberculosis and M. avium. *PloS One* 5(5):e10515.

[262] Rhoades ER, Frank AA, Orme IM (1997) Progression of chronic pulmonary tuberculosis in mice aerogenically infected with virulent Mycobacterium tuberculosis. *Tuber Lung Dis* 78(1):57–66.

[263] Turner J et al. (2001) Immunological basis for reactivation of tuberculosis in mice. *Infect Immun* 69(5):3264–3270.

[264] Wang JR, de Villena FP, McMillan L (2012) Comparative analysis and visualization of multiple collinear genomes. *BMC Bioinformatics* 13(3):S13.

[265] Liang L et al. (2011) Association of SP110 gene polymorphisms with susceptibility to tuberculosis in a Chinese population. *Infect Genet Evol* 11(5):934–939.

[266] Amulic B, Cazalet C, Hayes GL, Metzler KD, Zychlinsky A (2012) Neutrophil function: From mechanisms to disease. *Annu Rev Immunol* 30(1):459–489.

[267] Wilkinson TS et al. (2012) Ventilator-associated pneumonia is characterized by excessive release of neutrophil proteases in the lung. *CHEST* 142(6):1425–1432.

[268] Elkington PT, Ugarte-Gil CA, Friedland JS (2011) Matrix metalloproteinases in tuberculosis. *Eur Respir J* 38(2):456–464.

[269] Palanisamy GS et al. (2011) Evidence for oxidative stress and defective antioxidant response in guinea pigs with tuberculosis. *PLoS One* 6(10):e26254.

[270] Tecchio C, Micheletti A, Cassatella MA (2014) Neutrophil-derived cytokines: Facts beyond expression. *Front Immunol* 5(1):1–7.

[271] Gunzer M (2014) Traps and hyper inflammation - new ways that neutrophils promote or hinder survival. *Br J Haematol* 164(2):189–199.

[272] Friedman A, Turner J, Szomolay B (2008) A model on the influence of age on immunity to infection with mycobacterium tuberculosis. *Exp Gerontol* 43(4):275–285.

[273] Kirschner D, Marino S (2005) Mycobacterium tuberculosis as viewed through a computer. *Trends Microbiol* 13(5):206–211.

[274] Guzzetta G, Kirschner D (2013) The roles of immune memory and aging in protective immunity and endogenous reactivation of tuberculosis. *PloS One* 8(4):e60425.

[275] Rangaka MX et al. (2012) Predictive value of interferon-γ release assays for incident active tuberculosis: A systematic review and meta-analysis. *Lancet Infect Dis* 12(1):45–55.

[276] Dhulekar N, Nambirajan S, Oztan B, Yener B (2015) Seizure prediction by graph mining, transfer learning, and transformation learning *In Proceedings of the 11th International Conference on Machine Learning and Data Mining (MLDM)*.

[277] Hogeweg P, Takeuchi N (2003) Multilevel selection in models of prebiotic evolution: compartments and spatial self-organization. *Orig Life Evol Biosph* 33(4-5):375–403.

[278] Hogeweg P (2002) Computing an organism: On the interface between informatics and dynamic processes. *Biosystems* 64(1-3):97–109.

[279] Marée AFM, Hogeweg P (2001) How amoebids self-organize into a fruiting body: multicellular coordination in *Dictyostelium discoideum*. *Proc Natl Acad Sci U S A* 98(7):3879–3883.

[280] Marée AFM, Hogeweg P (2002) Modeling *Dictyostelium discoideum* morphogenesis: the culmination. *Bull Math Biol* 64(2):327–353.

[281] Savill NJ, Hogeweg P (1997) Modeling morphogenesis: From single cells to crawling slugs. *J Theor Biol* 184(3):229–235.

[282] Larson DE et al. (2010) Computer simulation of cellular patterning within the drosophila pupal eye. *PLoS Comput Biol* 6(7):e1000841.

[283] Shrinifard A et al. (2012) Adhesion failures determine the pattern of choroidal neovascularization in the eye: a computer simulation study. *PLoS Comput Biol* 8(5):e1002440.

[284] Hester SD, Belmonte JM, Gens JS, Clendenon SG, Glazier JA (2011) A multi-cell, multi-scale model of vertebrate segmentation and somite formation. *PLoS Comput Biol* 7(10):e1002155.

APPENDIX A
Glazier-Graner-Hogeweg Model

The Glazier-Graner-Hogeweg (GGH) Model [75, 277, 278, 279, 280, 281, 282, 76, 283, 284] is built upon the energy minimization-based Ising model [77], using imposed fluctuations via a Monte Carlo approach. The simulation space is divided into a lattice, and cells are represented by groups of adjacent lattice points. We utilize a two-dimensional lattice, with lattice sites corresponding to pixels, but it is possible to use three dimensions or a different lattice to pixel/voxel correspondence. The model has an effective energy, which is assigned by interactions between lattice sites as well as between cells. Energy assignment is governed by rules based on biological constraints and behaviors. The model then tries to minimize the effective energy using a Monte Carlo simulation process where lattice sites attempt to copy themselves into a neighboring site, growing the cell (and possibly shrinking the neighboring cell).

The effective energy is written as a Hamiltonian equation, where each term is the contribution from a particular interaction or condition. These energy functions are the mathematical representations of the biological behavior or constraint. The energy functions we use in our model are as follows:

Contact energy represents differential adhesion between cells of different types. It is a lattice-lattice interaction that assigns an energy penalty to adjacent lattice points belonging to different cells. It requires that each possible pair of cells (τ_a, τ_b), including self-pairs, be enumerated and given an energy value $J(\tau_a, \tau_b)$. Cell types that adhere to each other are assigned a lower penalty. In the energy function, the cell type τ of a particular lattice point i is given by $\tau_{\sigma(i)}$, where $\sigma(i)$ is the cell ID. The contact energy penalty assigned to a pair of lattice points (i, j) is therefore given as $J(\tau_{\sigma(i)}, \tau_{\sigma(j)})$. To prevent lattice points within the same cell from being assigned a contact energy penalty; this is multiplied by $(1 - \delta_{\sigma_i,\sigma_j})$, where δ is the Kronecker delta. The term for contact energy in the Hamiltonian equation across all pairs of lattice points (i, j) is therefore given as $\sum_{ij} J(\tau_a, \tau_b)(1 - \delta_{\sigma_i,\sigma_j})$, where i

and j are neighboring sites.

Focal point plasticity represents longer-distance adhesive and repulsive interactions between cells not adequately captured by contact energy. It is a cell-cell interaction that creates a link between adjacent cells of the specified types. Pairs of linked cells are given a target distance between cell centroids L, and deviation from this distance is assigned an energy penalty $\lambda_{FPP}(L_{current} - L_{target})^2$, where λ_{FPP} is a factor determining the strength of the penalty. In our model, FPP is used to simulate the changes that occur in the cleft region, where links are set based on the location of cells within the cleft. Deeper cells are assigned shorter target distances to guide the cleft shape, and the strength of the interaction is manipulated dynamically to model changing adhesion conditions within the cleft. The energy contribution from all FPP links within the model can be written as the summation $\sum_{(\sigma,\sigma')} \lambda_{FPP}(L_{current} - L_{target})^2$ for each pair of linked cells (σ, σ').

Area is a cell constraint that penalizes cells for deviating from a target size, simulating the biological tendency for cells to grow and maintain a certain size. The model requires two parameters, the target area A for each cell type τ and a factor λ_{Area}, which determines how strongly cells adhere to this constraint. The energy term of area is given as $\sum_{\sigma} \lambda_{Area}(a(\sigma) - A(\tau_a))^2$ for each cell σ and cell type τ.

Perimeter simulates the limited amount of plasma membrane available to a cell, and can either encourage or discourage the cell from adopting a highly irregular shape. Like area, it assigns an energy penalty for deviating from a target perimeter value P for each cell type τ, and its strength is determined by a multiplier $\lambda_{Perimeter}$, resulting in the energy term $\sum_{\sigma} \lambda_{Perimeter}(p(\sigma) - P(\tau_\sigma))^2$ for each cell σ and cell type τ. The full Hamiltonian equation for our simulation is thus given as

$$
\begin{aligned}
H = \sum_{ij} J(\tau_a, \tau_b)(1 - \delta_{\sigma_i,\sigma_j}) \ldots \\
+ \sum_{(\sigma,\sigma')} \lambda_{FPP}(L_{current} - L_{target})^2 \ldots \\
+ \sum_{\sigma} \lambda_{Area}(a(\sigma) - A(\tau_a))^2 \ldots \\
+ \sum_{\sigma} \lambda_{Perimeter}(p(\sigma) - P(\tau_\sigma))^2
\end{aligned}
$$

Energy minimization is carried out via a Monte Carlo simulation (MCS), as depicted in the overview in Fig. A.1. A sample consists of choosing a lattice point, and attempting to copy its cell ID into a randomly chosen neighbor. If both points belong to the same cell no change occurs; otherwise the outcome of the attempt is determined by the change in effective energy ΔH. Energy-lowering changes ($\Delta H < 0$) are always accepted; energy-increasing changes ($\Delta H > 0$) are accepted with a probability $e^{\frac{-\Delta H}{T}}$ (a Boltzmann acceptance function). In the context of the GGH model, the constant T controls the intrinsic motility of the cell. The use of an acceptance function allows some energy raising copy attempts to succeed, which is important as it prevents the model from becoming stuck at local energy minima. A single MCS step in the GGH model consists of N lattice copy attempts, where N is the total number of lattice sites in the simulation space.

The GGH model also incorporates mitosis as the division of an existing cells lattice sites into two equal parts through its centroid. Since mitosis cannot be described by a lattice-copy event, it is implemented outside of the energy framework. To simulate cell growth and proliferation, we specify a percentage of cells every 100 MCS steps, which then grow to twice their original size and divide, mimicking the growth and mitosis of biological cells. These percentages were calculated from the *ex-vivo* data.

We constructed our model using Compucell 3D, an open-source implementation of GGH (`http://www.compucell3d.org/`). Simulations were initialized from the starting image in each time-lapse image set by overlaying a 6×6 grid and

assigning pixels within each square grid to GGH cells. Seven cell types were defined for the simulations. Border cells (B), representing outer columnar cells at the boundary of the gland. A subset defined as opposite cleft walls (C1, C2) were given additional properties to model the unique behavior of cells within the cleft region. Regular cells (R), representing inner polymorphic cells. Mitotic cells (M) were represented as a subset of these cells, including those derived from border cells. Stem or duct (S), represents differentiated cells in the ductal region of the gland. These cells are included only to provide a structural framework to capture the morphology of the gland, and are not included in our main analysis. Medium (Med) is a simplified representation of the mesenchymal cells and extracellular matrix. It is a special cell that fills all otherwise unoccupied lattice sites, and is not subject to area and perimeter constraints (`http://www.compucell3d.org/Manual`). The different types of cells, their biological equivalents, and the possible differing contact energy values are specified in Table A.1.

Table A.1: Cell Types and Contact Energy Values Used in CompuCell 3D Implementation of GGH Model. Unless otherwise noted, contact energy between two cell types is 5.

Cell Type	Biological Cell Type	Contact Energy Values Differing from Baseline
Med	Mesenchyme	Med-R: 10
R	Inner polymorphic	R-S: 3
M	Mitotic	
S	Duct	S-S: 3
B	Outer columnar	B-S:3
C1	Cleft wall	C1-C1: 1, C1-Med: 1
C2	Cleft wall	C2-C2: 1, C2-Med: 1

As mentioned earlier, mitosis does not readily fit into the energy-minimization Hamiltonian framework, and thus we use a different approach to implement mitosis. A number of cells equal to 0.5% of the total cells in the model were selected for mitosis every 100 MCS steps by designating them as M cells. Of this number, 75% are randomly chosen from R and B cells near the boundary, and 25% are chosen from the R cells in the interior of the gland. These cells attempt to double their size in the

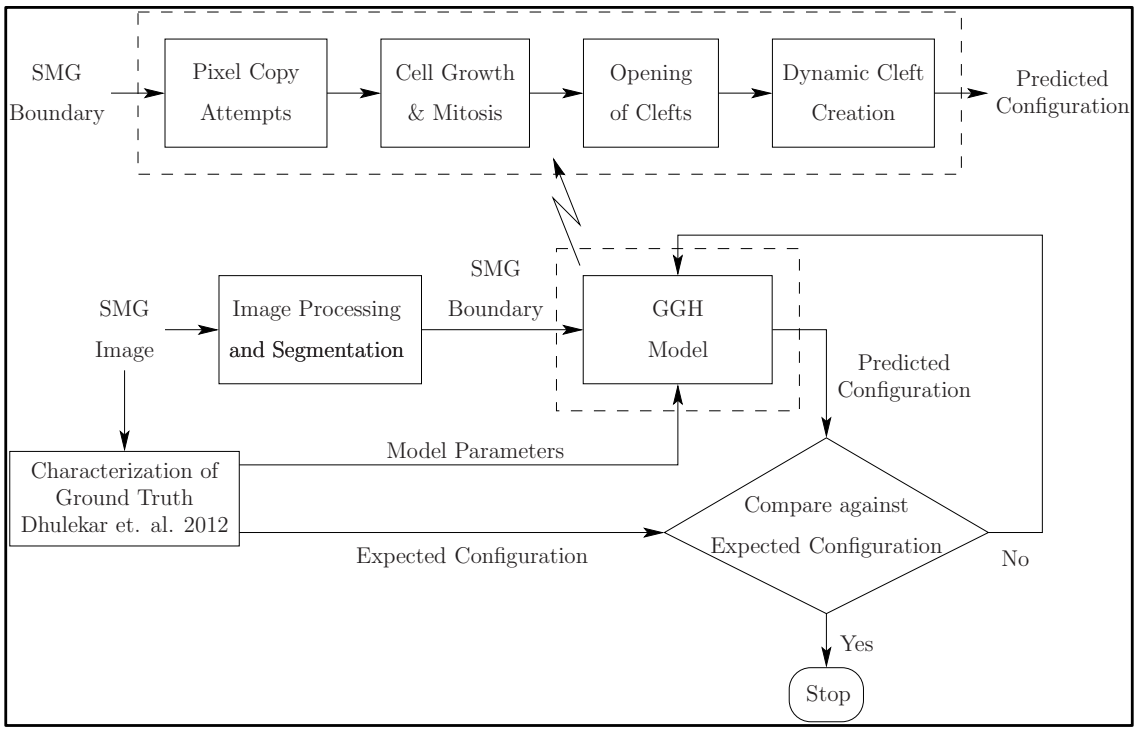


Figure A.1: After initialization of the model from an SMG image, MCS steps are run until the model reaches termination. Each MCS step consists of many pixel copy attempts. The model grows and divides cells at set intervals of MCS steps.

next 70 MCS steps, and are partitioned into two R cells. Supplementary Fig. A.1 presents an overview of the steps involved in our CompuCell 3D implementation for simulating SMG branching morphogenesis.