Humayra Tasnim

Candidate

Computer Science

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Melanie E. Moses, Chair

Shuang Luan, Member

Judy Cannon, Member

George Matthew Fricke, Member

Soumya Dutta, Member

Insight Into Complexity: Novel Information Theoretic Analysis of Spatiotemporal Interactions

BY

Humayra Tasnim

B.S., Computer Science & Engineering, University of Dhaka, 2013

M.S., Computer Science & Engineering, University of Dhaka, 2015

M.S., Computer Science, The University of New Mexico, 2021

DISSERTATION

Submitted in Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

Computer Science

The University of New Mexico

Albuquerque, New Mexico

July, 2024

DEDICATION

This dissertation is dedicated to my husband Ashraf, my daughter Zunaira

and my parents Faruk and Lily

ACKNOWLEDGMENTS

I am grateful to the remarkable individuals whose support made this dissertation possible. First I want to thank my advisor Professor Melanie Moses for her mentorship, guidance, and support throughout my Ph.D journey. She taught me how to become a researcher and encouraged me to explore and develop my research perspective without any reservations. She believed in me every step of the way and taught me how to communicate my ideas without hesitation. She is an excellent mentor and I am learning from her every day. Thank you, Melanie, for your wisdom and advice, which will be vital throughout my academic career and life.

I want to thank G. Matthew Fricke who has been my collaborator, instructor, advisor for any technical query and a great friend. I began my first project (cell association in lymph node) with Matthew and learned a great deal from him as we co-authored the paper. Matthew taught me the importance of patience in solving research problems. He guided me in using HPC solutions for the SIMCoV project. I am thankful for his support to access resources from the UNM Center for Advanced Research Computing (CARC).

I want to thank Soumya Dutta, who was my mentor during my internship at Los Alamos National Lab (LANL), later became my collaborator on multiple projects, and served on my dissertation committee. He introduced me to data vision and visualization research, which became a significant part of this dissertation. Although we never met in person, Soumya believed in me and continued to mentor and collaborate with me after the internship. I learned from him about national lab research practices, gathering valuable insights and experience that have greatly contributed to my work and career.

I want to thank Steven Hofmeyr from Lawrence Berkeley National Lab (LBL). It was a great pleasure to work and learn from such an incredibly skilled researcher. I am grateful for the opportunity to have interned at LBL twice under his mentorship. The experiences at the lab have been highly beneficial for my career.

I want to thank Professor Judy Cannon who is my collaborator and committee member. She

has significantly enriched my understanding of the immune system. Judy's lab provided the essential biological data for my computational analysis. Despite my initial unfamiliarity with such data, she generously dedicated her time and resources to ensure I grasped it thoroughly. Judy was always available to address my queries.

I am honored to work with Professor Stephanie Forrest on the SIMCoV project. It is a great opportunity to co-author a paper with such a distinguished professor. I have learned a lot from her. I want to thank Professor Shuang Luan for serving on my dissertation committee and providing valuable suggestions, I thank all my dissertation committee members: Melanie Moses, Judy Cannon, Shuang Luan, G. Matthew Fricke, Soumya Dutta for their time in evaluating this work and suggestions for improvement.

I want to thank my fellow peers, students, and members at the Moses Biological Computation Lab. I want to mention Will Vining, Vanessa Surjadudjaja, Abby Pribisova, Akil Andrews, Carter Frost, Antonio Griego whom I have closely worked, and others: John, Bianca, Julie, Jannatul, Jake, Wayne, and Shannon for their support, encouragement, and friendship. As an international student, I was gladly accepted and welcomed into the lab for which I am always grateful.

I can not express my gratitude enough to my parents Fakrul Alam and Khurshid Jahan for shaping me into who I am today. I am indebted to my mom who traveled thousands of miles to take care of me when I needed it. I thank my little brother Nafees Alam for being a motivator and being there for our parents. I am grateful to my extended family—uncles, aunts, and cousins—for taking care of my family in Bangladesh while I completed my PhD. I am thankful to my in-laws for their words of encouragement and support.

I am forever grateful to my husband, Ashraf for his continuous partnership, patience, and support throughout this journey. Without him, I wouldn't have achieved this much. Lastly, I thank my daughter Zunaira for giving my life new meaning and for being the source of encouragement that pushed me to complete my Ph.D journey.

Insight Into Complexity: Novel Information Theoretic Analysis of Spatiotemporal Interactions

by

Humayra Tasnim

B.S., Computer Science & Engineering, University of Dhaka, 2013

M.S., Computer Science & Engineering, University of Dhaka, 2015

M.S., Computer Science, The University of New Mexico, 2021

Ph.D, Computer Science, The University of New Mexico, 2024

ABSTRACT

Complex systems are comprised of different components. Interactions and associations among these components define the functionality of the system. For example, T cells must directly interact with virally infected cells to kill them. This research characterizes the most relevant components of complex systems by analyzing interacting relationships using information theoretic measures. It emphasizes the importance of spatial and temporal dynamics, which occur when components share spatial proximities or temporal sequences. Novel information theoretic analyses are proposed for quantifying the degree of association among system components, which is key to defining the spatiotemporal dynamics. One focus of this work is the application of these measures to biomedical datasets, bridging the gap between computational science and life sciences. Another focus is on the visual representation of such interactions, providing a new scientific lens to understand relevant features of complex systems. The measures are validated against benchmarks to ensure efficacy and applicability across multidisciplinary fields. This work advances the fields of computational biology and scientific visualization by providing novel, robust tools to analyze and interpret complex spatiotemporal interactions.

TABLE OF CONTENTS

List of Figures

List of Tables

Chapter 1

Introduction

Complex systems are characterized by their diverse components and the interactions that occur among these components. Understanding these interactions is crucial because they often govern the holistic behavior and functionality of the system. Complex systems can exhibit unexpected behaviors that emerge from the collective dynamics of their components, rather than from any single element's properties. According to Melanie Mitchell, a ***Complex System is a system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution*** [135].

This complex and emergent behavior is crucial in many multidisciplinary fields including biology, chemistry, ecology, weather, astronomy, economics, technology, and so on. Research in dynamic complex systems often utilizes mathematical models, statistical inferences, and simulations to predict how complex systems respond to changes in interaction patterns, helping to anticipate the functionality of the system through information flow.

The importance of component interactions in complex systems is further highlighted in biomedical research, where the failure or success of cellular functions can rely on these interactions. This study [105] illustrates how systems biology approaches to understanding cellular interactions can lead to breakthroughs in drug discovery and disease treatment. By

mapping the networks of interactions among proteins, genes, immune cells, and other cellular constituents, researchers can identify key nodes whose dysfunction may lead to disease. These interaction maps also help in predicting how the system might respond to specific interventions, allowing for more targeted and effective therapies. Thus, the study of component interactions in complex systems not only enhances theoretical understanding but also drives practical advances in technology and medicine.

Visualizing the interactions within complex systems is very important, as it transforms abstract data into comprehensible insights. Visualization acts as a bridge between raw computational data and human understanding, allowing researchers and practitioners to perceive patterns, anomalies, and critical links that are not readily apparent from numerical data and statistics alone. For example, in network science, visualizations help identify clusters, central nodes, or potential points of failure in systems ranging from social networks to infrastructure grids. As noted in [78], effective visualization tools not only enhance our ability to communicate complex findings but also significantly improve the decision-making process by providing a clear picture of the dynamics and structure of complex systems. This work uses widely used statistical measure: mutual information [173] and its decomposition measures to quantify interactions as well as capture the visual salience of such interactions. The work starts with quantifying spatial cell association in the immune system. Then, move to developing frameworks to extract important features and summarization of such interactions. Next, use an agent-based computational model to simulate and quantify the spatial damage caused by COVID-19 infection in the lung system.

1.1 Quantifying Spatial Association of Cells in Lymph Node

This work focuses on quantifying the spatial association of cells in lymph nodes. The main question of the study is how naïve T cells interact spatially with key cellular and structural elements within lymph nodes, specifically dendritic cells (DCs), fibroblastic reticular cells (FRCs), and blood vessels. This question is important because this association reveals insights about T

cell motility which is a key step in T cell activation and the initiation of the adaptive immune response, which are critical for fighting infections. This research advances the understanding of spatial interactions in the immune system. Traditionally, the focus has been on the interactions between T cells and DCs, but this study broadens the scope to include other structural and cellular components, using advanced quantitative metrics like the Pearson correlation coefficient (PCC) [151] and normalized mutual information (NMI).

The study uses two-photon microscopy (2PM) to observe T cells in the lymph nodes and employs PCC and NMI to measure the extent of spatial association between T cells and DCs, FRCs, and blood vessels. Remarkably, the study finds that naïve T cells are more frequently associated with FRCs than with DCs, the primary antigen-presenting cells. This suggests that while T cells are biologically programmed to respond to DCs, the structural environment within the lymph node, particularly the network formed by FRCs, plays a crucial role in guiding T cell movement and positioning. We find that FRCs could potentially be as important as DCs in regulating T cell behavior, an aspect that has been proposed in previous immunological research [83, 84, 117], but there is no quantitative evidence.

An important aspect of the study is its investigation of the role of the chemokine receptor CCR7 in T cell localization within lymph nodes. CCR7 is known to facilitate the homing of T cells to lymph nodes and their movement within the nodes. Surprisingly, the study shows that CCR7 deficiency does not decrease T cell association with DCs. In fact, CCR7-deficient T cells displayed a slight increase in association with DCs compared to their wild-type counterparts. This counterintuitive result suggests that while CCR7 enhances T cell mobility, its absence does not necessarily impede the T cell's ability to interact with DCs, possibly indicating that T cell motility and their interactions with DCs are modulated by additional factors beyond just chemokine signaling.

The methodological approaches in the work, particularly the use of NMI and PCC, provide a more holistic view of cellular interactions than traditional methods, allowing for a detailed analysis of how T cells coordinate their movements with the lymph node's architecture. The

application of these quantitative tools to immunology opens up new avenues for understanding complex cellular dynamics in a way that was not previously possible.

The findings of this study have significant implications for the development of immunotherapeutic strategies and vaccines. By illustrating the roles of FRCs and the impact of chemokine receptor signaling on T cell behavior, this research could lead to novel approaches that enhance knowledge about immune response. For instance, targeting the interaction between T cells and FRCs or modulating CCR7-dependent pathways could optimize T cell responses against pathogens or tumors.

In conclusion, this study enriches the understanding of T cell dynamics within lymph nodes and highlights the complex interplay between T cells and the lymph node microenvironment. The findings emphasize the necessity of considering multiple factors, including cellular interactions and structural factors, in the effective activation and function of T cells to initiate immune responses. This work is published and referenced as [183].

1.2 Information-Theory Based Analysis of Spatio-Temporal Datasets

In cases of large, multivariate time-varying datasets such as video sequences, weather patterns over time, or dynamic CT imaging, extracting relevant features that capture both spatial and temporal characteristics efficiently is crucial. The complexity and size of these datasets demand sophisticated techniques for feature extraction to enable effective summarization, optimized storage, and insightful analytics. The primary challenge lies in identifying and extracting salient regions from these datasets without exhaustive exploration, which is computationally expensive and time-consuming. Furthermore, summarizing these data dynamically while tracking the flow of information over time is essential for applications requiring real-time analysis and decision-making, such as in surveillance systems or real-time biological cell interaction.

Specific Mutual Information (SMI) offers a promising approach to tackle this challenge. SMI

is an information-theoretic measure derived from mutual information, a concept used to quantify the amount of information obtained about one random variable through another. This measure is particularly suited for spatiotemporal multivariate datasets where understanding individual data values' contribution towards spatial associations or disassociation among multiple variables over time is important. SMI can be utilized to identify areas within the dataset that hold the most 'informative' value — essentially regions where the occurrence of specific features significantly reduces uncertainty in other parts of the dataset. SMI emphasizes important regions in the variables where statistical multivariate properties exist. This measure can automatically highlight regions with interesting relationships (e.g. high surprise regions, high/low predictable regions). This measure also aligns well with the need for dynamic spatiotemporal data summarization, as it allows for the extraction of concise yet informative summaries of the data, facilitating both storage optimization and enhanced understanding of the underlying processes.

Frameworks developed in this work using SMI analysis of multivariate time-varying images aim to achieve the following:

- Automatic identification of salient regions that reduces the cost of exploration in large datasets.
- Dynamic spatiotemporal data summarization using information fusion for storage optimization.
- Tracking information flow to monitor and analyze the evolution of data over time is crucial for tasks that depend on understanding temporal dynamics, such as predictive modeling and anomaly detection.

1.2.1 Information-theoretic Exploration of Multivariate Time-Varying Image Databases

With the use of high-performance computational resources in scientific research, the generation of large multivariate time-varying datasets is common, with applications spanning climate modeling to dynamic medical imaging. As these datasets grow in size and complexity, traditional analysis and storage methods become inadequate due to the inability to efficiently process and extract meaningful information from vast amounts of data. This challenge necessitates the development of advanced techniques that can facilitate the rapid exploration and analysis of such datasets.

One promising approach is to use information theory-based approaches namely Specific Mutual Information (SMI) which is effective for exploring multivariate datasets. It quantifies the shared information between pairs of variables and reveals how specific values within these variables contribute to this shared information. This makes it valuable for detecting interdependencies and dynamic changes within the data, providing insights that are essential for many scientific and engineering applications.

The Cinema project [5] exemplifies an innovative application in managing large-scale scientific datasets. Cinema databases store visualizations of simulation data, allowing researchers to interactively analyze data through image-based techniques. By incorporating SMI-based measures in the Cinema database, the opacity of the images is modulated emphasizing regions of high informational significance. This method effectively reduces the volume of time and resources scientists need to analyze manually, by automatically highlighting areas with strong multivariate relationships.

This work shows that the technique has practical implications in several fields. In weather science, for example, in the case of a hurricane dataset, it can be used to identify and track evolving meteorological phenomena, such as the formation and movement of the hurricane's center. In combustion science, it helps extract regions within a combustion chamber where

chemical reactions are most intense.

An essential component of this approach is the interactive visualization tool, CinemaView, which supports the analysis by providing a user-friendly interface for navigating through time-varying data. Users can compare different time steps and variables side-by-side, adjusting visualization parameters to suit their specific analysis needs.

The integration of the SMI framework into Cinema databases represents a significant advancement in the analysis of multivariate time-varying datasets. It enhances the efficiency of data exploration and improves the accuracy of feature detection, which is critical for domain experts to make informed scientific decisions. This work is published and referenced as [185].

1.2.2 Dynamic Spatiotemporal Data Summarization using Information Based Fusion

With the rise of supercomputing capabilities, the volume of data produced has soared, intensifying storage and I/O overheads that present significant challenges in data management and storage. This work addresses these challenges using a dynamic spatiotemporal data summarization technique. This technique leverages Specific Mutual Information (SMI) to effectively reduce data storage demands while preserving critical information dynamics within datasets. The approach is distinct in retaining both raw and summarized timesteps, ensuring that no critical information is lost in the summarization process.

The core of the method involves the identification of informative and redundant timesteps within time-varying datasets. Informative timesteps are preserved, while redundant ones are fused using SMI-guided fusion techniques. This optimizes storage without sacrificing data integrity. This process streamlines data handling and enhances visualization capabilities, enabling users to track and analyze information change over time more efficiently.

The versatility of the proposed technique is demonstrated through its application to varied datasets, including particle-based flow simulations, security surveillance systems, and biological cell interactions. For example, in security and surveillance, the method allows for the efficient

summarization of lengthy video sequences, highlighting only those periods where significant activity occurs, thereby optimizing storage and improving the manageability of surveillance data. An integral component of the research is the holistic representation of the fused timesteps. This enables minimal data loss allowing for detailed examination of specific data points over compressed intervals.

The proposed summarization technique significantly impacts data management practices across multiple disciplines by reducing the computational and storage overhead associated with large datasets. It is applicable to both *in situ* and *post hoc* data analysis contributing to deeper insights in various scientific and technological fields. This work is submitted for review and archived [184].

1.3 Analyzing Spatial Features of SARS-CoV-2 Infection Spread in Lung using CT Scans and SIMCoV Model

The COVID-19 pandemic has emphasized the critical need for advanced tools to understand and predict the dynamics of viral infections, particularly in the respiratory system. Computed Tomography (CT) scans have been instrumental in diagnosing and assessing the severity of SARS-CoV-2 infections, revealing characteristic patterns of lung damage such as ground glass opacities (GGOs) [15, 51] and consolidations [82]. These imaging features represent the multifocal distribution of lung lesions and the associated tissue damage, typically due to inflammatory responses. We want to understand the underlying properties of lung damage and the cause of variability across patients. We use a simulation framework of the SARS-CoV-2 infection dynamics to explain the observable conditions in the CT scans.

The Spatial Immune Model of Coronavirus (SIMCoV) [137], is an advanced computational framework designed to simulate SARS-CoV-2 infection in the lungs at a cellular level. Unlike

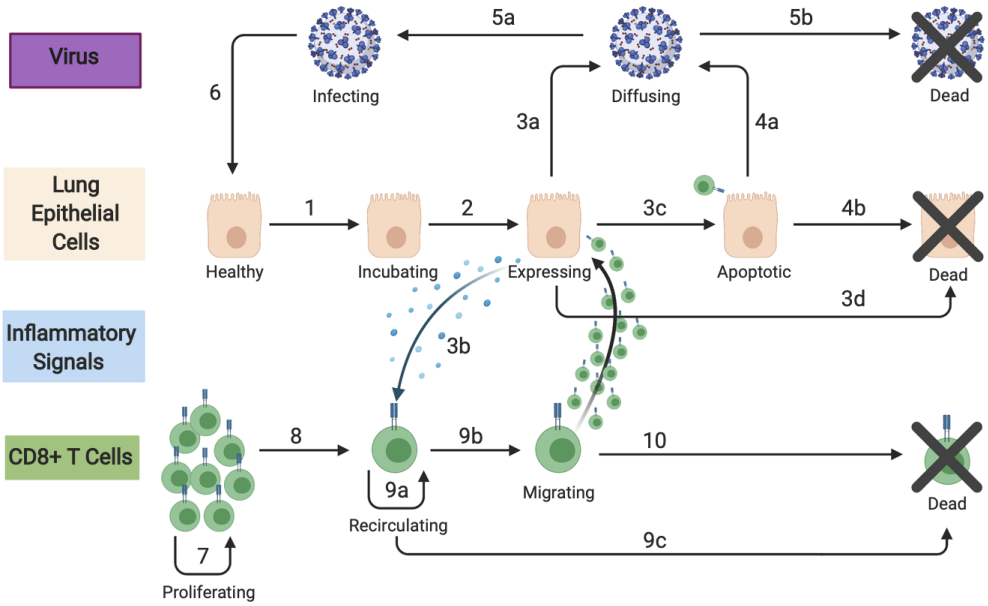


Figure 1.1: SIMCoV model components and their interactions. Epithelial and T cells are represented as agents; virions and inflammatory signals are represented as concentrations. Numbered transitions are described in the Materials and Methods Section of [137]

traditional Ordinary Differential Equation (ODE) models, SIMCoV employs an agent-based modeling approach, allowing for the detailed simulation of viral spread and immune response across hundreds of billions of cells. SIMCoV is the perfect groundwork to analyze and predict the spatial dynamics of COVID-19 in the lung as observed in CT scans.

The SIMCoV model simulates the dynamics of SARS-CoV-2 infections to understand the viral spread dynamics through tissue. It affects lung epithelial cells and examines how the timing and location of immune cells (T cells) influence the spread in the lungs. The components of the model and their interactions are visualized in Figure 1.1 which is referenced from [137]. SIMCoV demonstrates the initial spatial distribution of the virus in the lungs, explains the rates and patterns of viral spread through lung, and analyzes T cell counts and movement patterns, influenced by lung architecture.

Utilizing advanced high-performance computing methods and resources, SIMCoV simulates the viral spread over time using different model configurations. SIMCoV effectively replicates

the viral growth dynamics observed in patients and is the first model to demonstrate how spatially dispersed infections lead to increased viral loads. It also highlights how the timing and strength of the immune response can influence viral dynamics and controls. SIMCoV is an efficient model to understand the within-host dynamics of SARS-CoV-2 infection. The model and simulations suggest that the number of independent infection sites within the lungs is a key driver of peak viral load. The spatial dispersion of the inflammation caused by the virus may be particularly important for SARS-CoV-2 and other lung infections due to the extensive epithelial surface area of the lungs. Therefore, for analyzing spatial features of SARS-CoV-2 infection spread in the lung using CT scans, we are comparing the spatial characteristics of inflammatory signal spread within the alveolar sac structure, simulated by SIMCoV. SIMCoV allows us to investigate why the spread of infection and lung damage vary across patients and appears patchy in many cases.

The multifocal nature of SARS-CoV-2 infection leads to heterogeneous patterns of lung damage, making the disease progression unpredictable in many cases. Current modeling approaches, while useful, often fail to capture the spatial complexities of the infection. SIMCoV's spatially explicit modeling capability presents a unique opportunity to bridge this gap by providing a detailed, scalable platform for studying viral and immune dynamics.

By comparing the spatial features of lung damage observed in CT scans with those generated by SIMCoV, we can enhance our understanding of the initial conditions leading to varying levels of severity in patients. This comparison will also help in refining SIMCoV's parameters to better replicate and eventually predict individual patient outcomes based on early CT scans.

Chapter 2

Background

This chapter discusses the background of some concepts and related works that have been used in the research scope .

2.1 Information Theory Based Approaches and Concepts

Information theory tools [49] have been used extensively for solving problems across computational domains. Information theory provides information content for a variable and can measure similarity. From Shannon's paper [173], it can be stated that information is a defined measurable quantity. According to Claude Shannon in 1948: ***"A basic idea in information theory is that information can be treated very much like a physical quantity, such as mass or energy."***

2.1.1 Entropy

Entropy measures the amount of information in the probability distribution of a random variable [173]. It indicates the uncertainty in the outcome of an event. Entropy can be understood by considering a coin toss. The probability of heads is $p(x) = \frac{1}{2}$ and the probability of tails is $p(y) = \frac{1}{2}$. The entropy H is $-(\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2}))$. Since $\log_2(\frac{1}{2}) = -1$, H=1 bit.

The formula for calculating entropy is:

$$\mathrm{H}(r) = -\sum_{r} \mathrm{p}(r) \log_2 \mathrm{p}(r), \tag{2.1}$$

where H(r) is the entropy of variable r and $p(r)$ is the probability of r occurring. Here we use $\log_2$ so that entropy is measured in bits, the unit of information. The expression is negated because the $\log_2$ of probabilities (which are always less than or equal to 1) is always negative or zero.

2.1.2 Joint Entropy

We use joint entropy to measure the uncertainty in the outcome of two variables:

$$\mathrm{H}(r,g) = -\sum_{r}\sum_{g} \mathrm{p}(r,g) \log_2 \mathrm{p}(r,g) \tag{2.2}$$

where $\mathrm{p}(r,g)$ is the joint probability distribution function of r and g.

2.1.3 Mutual Information (MI)

Mutual Information (MI) is one of the well-known measures to quantify the mutual correlation between two variables. Mutual information quantifies the total amount of information overlap between two variables, i.e., if we observe a certain variable, then MI tells us how much uncertainty has been reduced regarding the information of another variable. Given two random variables X and Y, MI $I(X,Y)$ is formally defined as:

$$\mathrm{I}(X,Y) = \sum_{y \in Y} \sum_{x \in X} \mathrm{p}(x,y) \log \frac{\mathrm{p}(x,y)}{\mathrm{p(x)\,p(y)}} \tag{2.3}$$

where $\mathrm{p}(x)$ and $\mathrm{p}(y)$ are the probabilities of occurrence of values x for X and y for Y respectively and $\mathrm{p}(x,y)$ is the joint probability of occurrence of values x and y together.

MI can also be calculated using entropy one variable and the joint entropy of two variables

using 2.4.

$$\mathrm{MI}(r,g) = \mathrm{H}(r) + \mathrm{H}(g) - \mathrm{H}(r,g) \tag{2.4}$$

Intuitively, this formula calculates MI by subtracting the joint entropy of r and g from the total entropy in both r and g, which leaves the overlap in entropy of r and g.

MI quantifies the total association or disassociation between two variables and provides a single value in bits.

2.1.4 Normalized Mutual Information (NMI)

We normalize MI by dividing by the minimum of the internal entropies, since it provides an upper bound on MI, for a proof see [87].

$$\mathrm{NMI} = \frac{\mathrm{MI}(r,g)}{\min(\mathrm{H}(r), \mathrm{H}(g))} \tag{2.5}$$

The value of NMI is bounded between 0 and 1.

2.1.5 Specific Mutual information

MI can be further decomposed into specific mutual information (SMI) measures to quantify individual data values' contribution towards such association or disassociation. For specific scalar values $x \in X$, SMI computes the information content of x when another variable Y is observed. In this case, X is called the reference variable and Y is called the target variable. Knowledge about the scalar values in the reference variable can increase knowledge about the target variable. This increase in information or decrease in uncertainty helps in identifying important regions in the float-image data. MI can be decomposed in multiple ways to obtain several SMI measures and we focus on two such SMI measures, Surprise and Predictability, [26, 55] for finding different types of multivariate characteristics between variable pairs.

SMI measure Surprise: $I_1(x;Y)$

The *Surprise* measure quantifies the change in the information content in the occurrences of the target variable after observing individual scalar values of the reference variable which has the potential of providing information which would seem improbable otherwise, hence the name surprise [26, 55]. The regions where data values have higher surprise values can be informative. For two random variables X and Y, surprise is denoted as I_1 and presented as:

$$\mathrm{I}_1(x;Y) = \sum_{y \in Y} \mathrm{p}(y|x) \log \frac{\mathrm{p}(y|x)}{\mathrm{p}(\mathrm{y})} \tag{2.6}$$

where $x \in X$ is the reference variable and $y \in Y$ is the target variable. $\mathrm{p}(y)$ is the probabilities of occurrence of values y for Y and $\mathrm{p}(y|x)$ is the conditional probabilities of values y given values x. Surprise is always positive as it is the distance between $\mathrm{p}(y|x)$ and $\mathrm{p}(y)$. A high $I_1(x;Y)$ implies that after observing the reference variable x, some low probability values of $y \in Y$ have become more probable.

SMI measure Predictability: $I_2(x;Y)$

The *Predictability* measure provides us with the amount of increase/decrease in uncertainty about the target variable after observing the reference variable [26,55]. This quantification of the uncertainty change helps to identify statistically significant regions in the images. Predictability is denoted as I_2 and can be computed as:

$$\mathrm{I}_2(x;Y) = -\sum_{y \in Y} p(y) \log p(y) + \sum_{y \in Y} p(y|x) \log p(y|x) \tag{2.7}$$

where $x \in X$ is the reference variable and $y \in Y$ is the target variable. $\mathrm{p}(y)$ is the probabilities of occurrence of values y for Y and $\mathrm{p}(y|x)$ is the conditional probabilities values y given values x. Based on the amount of information increase and decrease, I_2 can be both positive and negative. A high positive $I_2(x;Y)$ value indicates that the uncertainty of target variable Y has decreased

when value x is observed. On the other hand, a high negative $I_2(x;Y)$ value indicates that the uncertainty of target variable Y has actually increased. According to information theory, data values that are less probable or unpredictable contain more information representing salient regions in the data with diverse characteristics that are worth deeper exploration.

2.2 Information Theory for Data Analysis and Visualization

The use of information theoretic measures [49, 191] to solve data analysis and visualization problems is well-known. Mutual information has been used to perform data registration [47, 92, 94, 126, 152], view selection [194], estimation of surface similarities [90], shape complexity [161], and for quantifying information transfer from data to image space [28]. For exploring similarities among level-sets, information theory has also been used [32, 203]. Various decomposition of mutual information, called specific mutual information measures have become recently popular for fusing multi-modal data [27], analyzing isosurface uncertainties between variable pairs [23], and designing transfer functions [25]. Point-wise mutual information is also applied to quantify important data value combinations from time-varying data [66], and for retrieving opposite information from a given variable pair [91]. For a detailed review of information theory applications in data analysis and visualization, interested readers are referred to [42, 43, 166, 198].

2.3 Time Step Selection and Data Summarization

Detection of key time points in a data set is an important problem for time-varying data analysis. Several approaches have been proposed for key time step detection for large time-varying data sets [187, 220]. These techniques assume the availability of all the time steps. When the storage of all time steps is not possible, real-time techniques are required so that they can be applied *in situ*. Myers et al. [142] proposed an *in situ* streaming regression-based strategy for detecting salient time points. To enable adaptive *in situ* workflow during the simulation run, Maher

et al. [165] proposed a trigger-based solution for combustion simulations. These techniques generally allow the detection of key time points and do not offer any data summarization capability.

The computer vision community has developed several techniques for doing spatio-temporal fusion of large data obtained from different sources. These approaches do not necessarily combine time steps based on the key time points. Pulong and Kang proposed a technique for fusing temperature data obtained from MODIS and AMSR-E instruments using a dynamic fused Gaussian process [124]. Nguyen et al. [145] developed a technique for summarizing large spatio-temporal images obtained from remote sensing applications. In a recent work, Shah et al. [171] proposed an algorithm for real-time summarization of data streams for smart grid applications. Compared to the above techniques, the proposed method is different in the sense that our method needs to work *in situ* under strict memory and computational resource constraints and is primarily developed for very large-scale three-dimensional scientific data sets. The proposed method aims at identifying the key time steps based on some user-provided criteria and then generate summaries for the intermediate non-key frames so that the reduced output data can store a holistic view of the entire simulation data allowing flexible *post hoc* analysis and visualization.

Chapter 3

Quantitative Measurement of Naïve T Cell Association with Dendritic Cells, FRCs, and Blood Vessels in Lymph Nodes

3.1 Publication Notes

Citation: Tasnim H, Fricke GM, Byrum JR, Sotiris JO, Cannon JL and Moses ME (2018) Quantitative Measurement of Naïve T Cell Association With Dendritic Cells, FRCs, and Blood Vessels in Lymph Nodes. Front. Immunol. 9:1571. doi: 10.3389/fimmu.2018.01571

Editor: Vitaly V. Ganusov, University of Tennessee, Knoxville, United States

Received: 07 February 2018

Accepted: 25 June 2018

Published: 26 July 2018

Formatting: The original published text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Data and Software Availability: The code used in this paper is publicly available at https://github.com/BCLab-UNM/NMIFrontiers2018.

The data used is available at http://digitalrepository.unm.edu/cs_sp/1/.

Funding: This work was supported by funding from the following: DOD STTR Contract FA8650-18-C-6898 (**JLC** and **MEM**), NIH 1R01AI097202 (**JLC**), the Spatiotemporal Modeling Center (P50 GM085273), the Center for Evolution and Theoretical Immunology 5P20GM103452 (**JLC**), and a James S. McDonnell Foundation grant for the study of Complex Systems (**MEM, GMF**). Thanks to the UNM Cancer Center Fluorescence Microscopy Facility (P30-CA118100) as well as the BRAIN Imaging Center (P30GM103400) for help with two-photon microscopy. **JRB** was supported by T32 NIH 5 T32 AI007538-19 as well as the Ruby Predoctoral Travel Fellowship from the Molecular Genetics and Microbiology Department at UNM HSC. **JLC** is a member of the Center of Biomedical Research Excellence (CoBRE) Autophagy, Inflammation, and Metabolism (AIM) in Disease (P20GM121176). **HT** and **MEM** were supported by an LDRD grant from Sandia National Laboratories.

Ethics Statement: Breeding, maintenance, and use of animals used in this research conform to the principles outlined by the Institutional Animal Care and Use Committee (IACUC) at the University of New Mexico Health Sciences Center. The IACUC at the University of New Mexico approved the protocol for animal studies (protocol number 16-200497-HSC). Anesthesia via ketamine and xylazine was performed during mouse injections, and euthanasia was administered via isofluorane overdose followed by cervical dislocation.

Acknowledgements: We would like to thank Nick Watkins and Sandra Chapman for the useful discussion as well as the very helpful suggestions by reviewers to include analysis of regions of 2PM images, than just pixel-based comparisons.

3.2 Abstract

T cells play a vital role in eliminating pathogenic infections. To activate, naïve T cells search lymph nodes (LNs) for dendritic cells (DCs). Positioning and movement of T cells in LNs is influenced by chemokines including CCL21 as well as multiple cell types and structures in

the LNs. Previous studies have suggested that T cell positioning facilitates DC colocalization leading to T:DC interaction. Despite the influence chemical signals, cells, and structures can have on naïve T cell positioning, relatively few studies have used quantitative measures to directly compare T cell interactions with key cell types. Here we use Pearson correlation coefficient (PCC) and normalized mutual information (NMI) to quantify the extent to which naïve T cells spatially associate with DCs, fibroblastic reticular cells (FRCs), and blood vessels in LNs. We measure spatial associations in physiologically relevant regions. We find that T cells are more spatially associated with FRCs than with their ultimate targets, DCs. We also investigated the role of a key motility chemokine receptor, CCR7, on T cell colocalization with DCs. We find that CCR7 deficiency does not decrease naïve T cells association with DCs, in fact, CCR7$^{-/-}$ T cells show slightly higher DC association compared with wild type T cells. By revealing these associations, we gain insights into factors that drive T cell localization, potentially affecting the timing of productive T:DC interactions and T cell activation.

3.3 Introduction

The adaptive immune response depends on T cell interactions with DCs in the paracortex, or T cell zone, of LNs. The rate at which naïve T cells sample DCs determines how fast the immune system can mount a response to infection [134]. The development of imaging methods such as two-photon microscopy (2PM) and histocytometry have enabled direct observation of cell locations in tissues. Many studies showing the relative location of T cells and DCs suggest that they are both positioned in the LN to maximize the likelihood of T:DC interactions [30, 209]. Despite advances in the ability to image and observe T cells in LNs, few studies make direct quantitative comparisons of how closely T cells associate with multiple other cells types in LNs.

T cells enter the paracortex of the LN from small post-capillary blood vessels termed high endothelial venules (HEVs). T cells, DCs and FRCs occupy this region along with blood vessels

(BVs). T cells move amongst DCs, FRCs, and other T cells to interact with DCs presenting antigen. FRCs are stromal cells that encapsulate a collagen fiber conduit network which allows for transport of lymph fluid carrying soluble antigen and chemokines [12, 88, 148, 176]. FRCs produce the chemokine CCL21, which has an established role in naïve T cell homing into the paracortex from blood vessels [178, 197]. FRCs also provide structural support required for efficient T cell activation [147]. [13] showed the FRC network is closely associated with naïve T cells moving within the paracortex, suggesting that FRCs may provide a network on which T cells migrate.

There are several hypotheses regarding the role of individual cell types in mediating T:DC interactions. HEVs are the entry points for T cells entering the LN. [85] suggests that DCs gather near HEVs to maximize their contact rate with incoming T cells. Others have suggested that DCs may congregate at the intersections of the FRC network, allowing T cells that travel along the edges of the network to encounter T cells at an increased rate [59, 99, 186, 217]. Spatial interactions between T cells and blood vessels, FRCs, and DCs are important if they change how T cells move through the paracortex and the timing of encounters with antigen-presenting DCs, the key step in T cell activation and the initiation of the adaptive immune response.

In addition to structural and cellular cues, chemical mediators, including chemokines, contribute to T cell motion and T:DC contacts in the LN. For example, the signaling molecule LPA produced by FRCs has been shown to mediate rapid T cell motion in LNs [182]. In addition, C-C chemokine receptor type 7 (CCR7), the receptor recognizing CCL21, is important for high speed T cell motility in the LN [10, 112]. While CCR7 increases T cell movement speed in LNs, whether CCR7 impacts T:DC contacts has not been investigated.

Understanding the contribution of cellular and structural LN components to T cell localization requires a quantitative metric that allows direct comparisons of spatial associations of multiple cell types. Several other groups have reported spatial relationships between cells and structures using methods such as visual inspection [85, 131] and comparison of turning angles of T cell movements with structures [13, 138]. However, none of these directly compare

associations between multiple cell types or structures with a consistent quantitative metric.

In this study, we use both the Pearson correlation coefficient (PCC) [3, 18] as well as mutual information (MI) [173] to compare the spatial association of multiple cell types and structures. PCC measures the covariance of homologous pixel intensities, and has been often used to determine colocalization, particularly of fluorescent proteins, in multiple biological systems including the study of T cells [57, 61]. PCC and MI can be calculated without the need to identify individual cell boundaries which can be difficult for 2PM images.

MI is an application of Shannon entropy (which measures the amount of uncertainty about the value of a random variable in bits) originally defined to understand limitations on signal processing and communication [173]. MI quantifies the reduction in uncertainty about one variable when one knows the value of another variable. In analyzing spatial associations, we measure the reduction in uncertainty about the location of one cell type given the location of another cell type. MI has been successfully used in other biomedical image processing applications, particularly in measuring image similarity in X-rays and MRIs for automated image registration [104, 153, 181, 196]. Further, MI and other information theoretic measures are increasingly recognized as powerful tools for analysis of non-linear complex systems, including complex biological systems such as the immune system [120, 155]. In this paper, we use MI to quantify the spatial association of T cells with other cell types (e.g., DC or FRC). We use MI as a measure of spatial association that is independent of specific types of cells or structures. Additionally, MI is theoretically insensitive to coarse graining [54]. Thus, MI can measure the amount of spatial dependence of one fluorescent marker on another while minimizing observational bias. MI, unlike distance measures such as nearest-neighbor analysis, is parsimonious, since it does not require extensive image processing to remove photon noise and determine cell boundaries. Instead, MI can operate on the image directly without the introduction of thresholds. In preliminary work we used MI to quantify the association of T cells and DCs and found less correspondence between T cell and DCs than expected [79].

However, MI is not comparable across images with different sizes and amounts of fluores-

cence. In this study, we use normalized mutual information (NMI) [87, 180] that scales MI to be between 0 and 1, which allows quantitative comparisons of spatial associations between cells fluorescing in one color channel and another cell type fluorescing in a different color channel across experiments [48, 154, 193, 208]. Since PCC and NMI are both pixel based methods that do not correspond to cell sizes, we create regions within the images that match cellular scales and apply PCC and NMI. Analyzing regions as well as pixels allows these methods to capture associations at biologically relevant scales. Both regional PCC and NMI analyses show T cells associate much less with their ultimate targets, DCs, than with FRCs. Our results also show that CCR7 does not increase T cell association with DCs. Our study uses quantitative metrics to directly compare spatial association of T cells with other cell types in lymph nodes, revealing insights into T cell search.

3.4 Methods and Materials

3.4.1 Mice and Reagents

Experiments were performed with C57BL/6 mice (Jackson Laboratories), B6.Ubiquitin-GFP mice (Jackson Laboratories), B6.CCR7$^{-/-}$ mice (Jackson Laboratories) and B6.Cg-Tg(Itgax-Venus)1Mnz/J mice (Jackson Laboratories). Both female and male mice were used between 8-20 weeks of age. Breeding, maintenance, and use of animals used in this research conform to the principles outlined by the Institutional Animal Care and Use Committee (IACUC). The IACUC at the University of New Mexico approved the protocol for animal studies (protocol number 16-200497-HSC). Anesthesia via ketamine and xylazine was performed during mouse injections, and euthanasia was administered via isofluorane overdose followed by cervical dislocation. For blood vessel staining, DyLight 594 labeled *Lycopersicon Esculentum* (tomato) lectin (Vector Laboratories) was used at a dose of 70 µg per mouse. To isolate naïve T cells, Pan T Cell Isolation Kit II (mouse, Miltenyi Biotec, 130-095-130) was used according to manufacturer's instructions. To fluorescently label naïve T cells, CellTracker™Orange (5-(and-

6)-(((4-chloromethyl)benzoyl)amino)tetramethylrhodamine) (CMTMR) Dye (ThermoFisher Scientific, C2927) was incubated with naïve T cells at a final concentration of 5 µM at 37 °C for 30 min before being washed. Labeled naïve T cells were then immediately adoptively transferred into recipient mice.

3.4.2 Mouse procedures

For all images: 10^7 naive T cells were adoptively transferred into mice 14-16 hours prior to LN harvest for imaging by 2PM. For T:DC images: T cells from naïve wild type (WT) mice were labeled with orange vital dye CMTMR and adoptively transferred into naïve CD11c-yellow fluorescent protein (YFP) mice in which all $CD11c^+$ DCs are YFP^+. For T:BV images: T cells from naïve Ubiquitin-green fluorescent protein (GFP) mice were adoptively transferred into naïve C57Bl/6 recipient mice. DyLight 594-labeled *L. Esculentum* (tomato) lectin was injected intravenously into the recipient mice 5 min before harvesting the LNs for imaging. The fluorescent lectin binds to glycoproteins on blood vessel endothelial cells and emits red fluorescence. For T:FRC images: T cells from naïve WT mice were labeled with CMTMR and adoptively transferred into Ubiquitin-GFP recipient mice that were lethally irradiated (10 Gy). The mice were reconstituted with C57Bl/6 bone marrow 4 weeks prior to T cell adoptive transfer. In this chimeric mouse model, the stromal cell populations fluoresce GFP while the hematopoietic cell populations are non-fluorescent.

3.4.3 Two-Photon Microscopy set up

Two-photon microscopy was performed using either a ZEISS LSM510 META/NLO microscope or Prairie Technologies Ultima Multiphoton microscope from Bruker.

Prairie Technologies Ultima Multiphoton microscope from Bruker: A Ti-Sapphire (Spectra Physics) laser was tuned to either 820 nm for excitation of CMTMR or 850 nm for simultaneous excitation of YFP and CMTMR, GFP and DyLight 594, or GFP and CMTMR excitation. The Prairie system was equipped with galvo scanning mirrors and an 801 nm long pass dichroic to

split excitatory and emitted fluorescence. Emitted fluorescence was separated with a 550 nm long-pass dichroic mirror. Fluorescence below 550 nm was split using a 495 nm dichroic and filtered with 460/60 nm and 525/50 nm filters before amplification by photo-multiplier tubes. Fluorescence above 550 nm was split with a 640 nm long-pass dichroic mirror before passing through 590/50 nm and 670/50 nm filters before amplification by GaAsP photo-multiplier tubes. A UMPlanFLN 20x water immersion objective (0.5 numerical aperture) was used. Prairie View 5.4 software (Prairie Technologies) was used to acquire time-lapse z-stacks.

ZEISS LSM510 META/NLO: Chameleon Ti:Sapphire laser tuned to 850 nm (Coherent) was used for excitation of either GFP and CMTMR, YFP and CMTMR, or Dylight 594 and GFP. A 560 nm dichroic mirror and 500-550 nm and 575-640 nm bandpass filters were used for detection of fluorophores. Movies were captured with the ZEN user interface (Zeiss). In both imaging systems, Z-stacks with step size of 4 µm were repeatedly imaged over time to obtain movies of 10-45 min in duration. All analyses were performed on 2D image z stacks captured by 2PM.

3.4.4 Lymph node preparation for live imaging

After euthanasia, LNs from mice were surgically dissected and transferred to a Chamlide AC-B25 imaging chamber (Live Cell Instruments) with a customized coverslip platform to allow flow beneath the LN. The LN was stabilized with a tissue slice harp (Warner Instruments) and superfused with oxygenated Dulbecco's Modified Eagle's Medium (Gibco, 21063-045) and maintained at 37 °C. For experiments in which blood vessels were imaged in conjunction with T cells or DCs, with 70 µg DyLight 594-labeled lectin (from *L. Esculentum*, Vector Laboratories) was intravenously administered by tail vein injection 5 min before euthanasia.

3.4.5 Calculation of Mutual Information

MI measures how much the value of one variable tells us about the value of another variable. In this study, MI is used to quantify how much the locations of DCs, FRCs and blood vessels

reveal about the locations of T cells. We calculate the MI of color intensities resulting from 2PM imaging of two cell types. Each image is composed of a sequence of 2-color 3D images. In these images one cell type is dyed red and another green. We calculate the MI of the red and green channels from every image to determine the association of the corresponding cell types for that image.

The 2PM images contain red, blue and green channels. For every time step, we extract the red and green channels into two separate 3D images r and g.

The 2PM images contain red, blue and green channels. For every time step we extract the red and green channels into two separate 3D images r and g.

The MI calculation procedure can be summarized in the following 3 steps:

1. We calculate the entropy of variables in Xi and Y image r and image g: $\mathrm{H}(r)$ and $\mathrm{H}(g)$. This measures the uncertainty of the color intensity in each image.

2. We calculate the joint entropy $\mathrm{H}(r,g)$ which measures the uncertainty about the color intensities in corresponding positions in both images.

3. We calculate MI as the sum of the entropies of the individual images $\mathrm{H}(r)$ and $\mathrm{H}(g)$ minus the joint entropy of the two images $\mathrm{H}(r,g)$. This reveals how much uncertainty about the color intensity and location of one cell type (i.e., T cells) is reduced when we know the color intensity and locations of the other cell type.

Entropy

Entropy measures the amount of information in the probability distribution of a random variable [173]. It indicates the uncertainty in the outcome of an event. Entropy can be understood by considering a coin toss. The probability of heads is $p(x) = \frac{1}{2}$ and the probability of tails is $p(y) = \frac{1}{2}$. The entropy H is $-(\frac{1}{2} \times \log_2(\frac{1}{2}) + \frac{1}{2} \times \log_2(\frac{1}{2}))$. Since $\log_2(\frac{1}{2}) = -1, \mathrm{H} = 1$ bit.

The formula for calculating entropy is:

$$\mathrm{H}(r) = -\sum_{r} \mathrm{p}(r) \log_2 \mathrm{p}(r), \tag{3.1}$$

where H(r) is the entropy of variable r and $p(r)$ is the probability of r occurring. Here we use $\log_2$ so that entropy is measured in bits, the unit of information. The expression is negated because the $\log_2$ of probabilities (which are always less than or equal to 1) is always negative or zero.

Entropy is maximized for a random event in which the probabilities of all outcomes are equally likely (all N possible outcomes have a probability of occurrence of $\frac{1}{N}$) leading to an entropy of $\log_2(N)$ bits. Entropy is minimized for a completely predictable event in which one outcome has a probability of occurrence equal to 1, and all other outcomes have 0 probability of occurrence, leading to an entropy of zero.

We calculate the entropy of color intensities in the red and green images. Each image has 256 possible color intensities for both the red and green images. Thus the maximum $\mathrm{H}(r)$ and the maximum $\mathrm{H}(g)$ is $\log_2(256) = 8$ bits which would occur if each of 256 color intensities were equally likely.

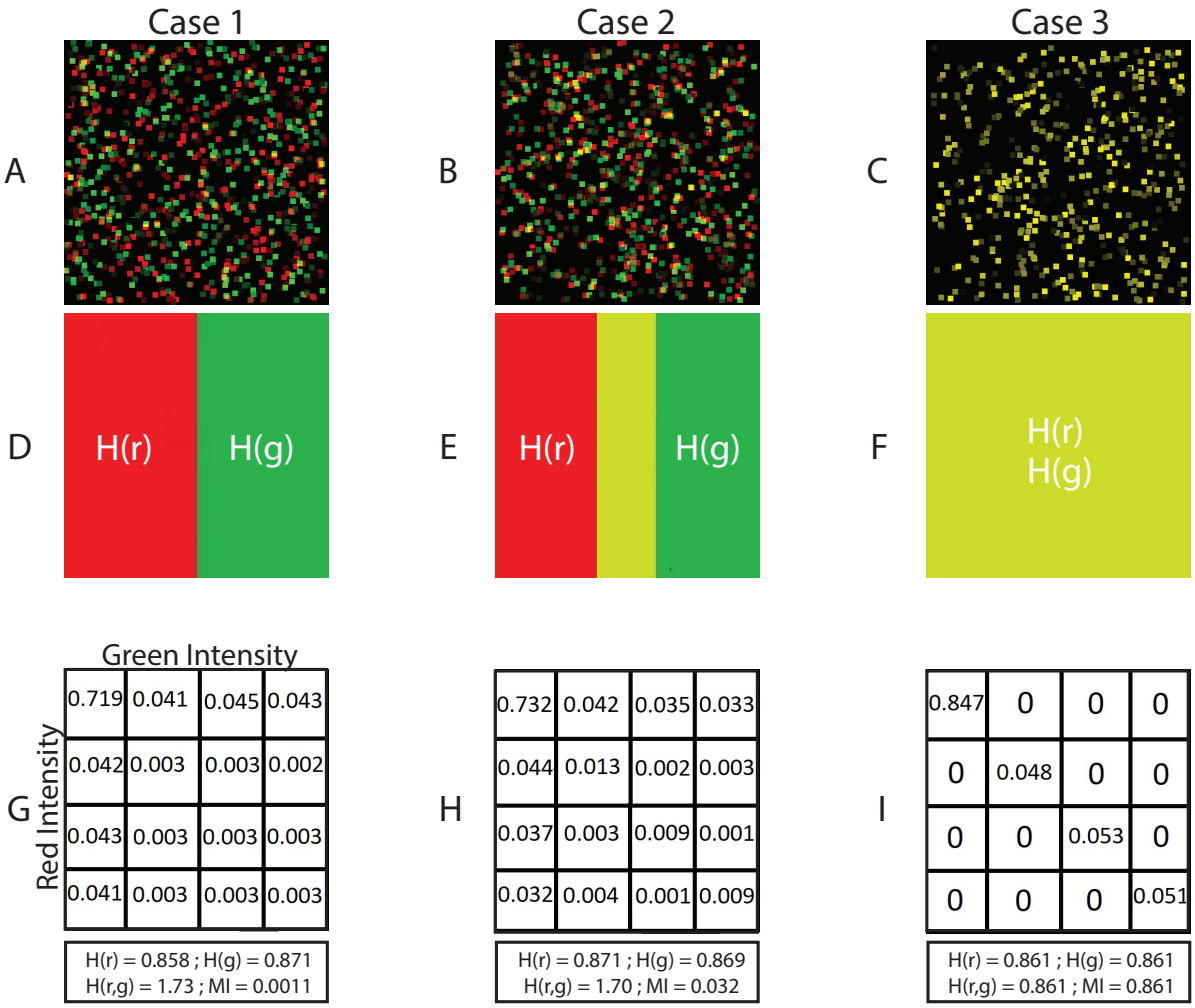

Case 1
Case 2
Case 3
A
B
C
D
H(r)
H(g)
E
H(r)
H(g)
F
H(r)
H(g)
Green Intensity
Red Intensity
G
0.719 0.041 0.045 0.043
0.042 0.003 0.003 0.002
0.043 0.003 0.003 0.003
0.041 0.003 0.003 0.003
H(r) = 0.858 ; H(g) = 0.871
H(r,g) = 1.73 ; MI = 0.0011
H
0.732 0.042 0.035 0.033
0.044 0.013 0.002 0.003
0.037 0.003 0.009 0.001
0.032 0.004 0.001 0.009
H(r) = 0.871 ; H(g) = 0.869
H(r,g) = 1.70 ; MI = 0.032
I
0.847 0 0 0
0 0.048 0 0
0 0 0.053 0
0 0 0 0.051
H(r) = 0.861 ; H(g) = 0.861
H(r,g) = 0.861 ; MI = 0.861

Figure 3.1: Illustration of low, medium and high MI. Simulated images of 500 red and 500 green cells are shown in (A), (B) and (C). Each cell is 11×11 pixels (square shaped) where the red cells are placed some distance from green cells, following a Gaussian distribution with mean 0, and a specified standard deviation, σ. The color intensity of each cell is chosen uniformly at random. However, each pair of green cells and red cells share the same color intensity. In (A), the red and green cell placements are uncorrelated and uniform randomly distributed. In (B), the placements of red and green cells are partially correlated (σ=5). In (C), the location of red and green cells are identical (σ=0). (D-F) are set diagrams indicating the shared information between red and green channels. In (D), the two color channels are independent since cell locations are uncorrelated with each other providing minimum MI. In (E), the two images are partially correlated which increases the MI, shown by the yellow shaded region. In (F), the two images are completely correlated maximizing the MI of the two color channels, resulting in complete intersection of the information in the red and green channels (yellow region). (G), (H), and (I) are joint probability tables for images (A), (B), and (C) where 256 color intensities are binned into 4 color intensities for purposes of illustration, resulting in a 4×4 probability table. In (G), the probability values are low and evenly spread across the table, except for the upper left corner, indicating overlap in the space with no cells (MI = ¡0.001). In panel (H), the probability values are higher along the diagonal than in other parts, indicating partial correlation in the placement of red and green cells (MI = 0.0320). In (I), there are probability values on the diagonal only and the probabilities off the diagonal are 0 since there is complete correlation in the placement of red and green cells (MI = 0.8610). The calculation of entropy $\mathrm{H}(r)$ and $\mathrm{H}(g)$, joint entropy $\mathrm{H}(r,g)$, and MI are shown for each case.

Joint Entropy

We use joint entropy to measure the uncertainty in the outcome of two variables:

$$\mathrm{H}(r,g) = -\sum_{r}\sum_{g} \mathrm{p}(r,g)\log_2 \mathrm{p}(r,g) \tag{3.2}$$

where $\mathrm{p}(r,g)$ is the joint probability distribution function of r and g.

The two variables may be unrelated. For example, the joint entropy in the outcome of tossing a fair coin twice is calculated from the probabilities of four possible events [heads, heads], [heads, tails], [tails, heads] and [tails, tails]. The probability of each event is $\frac{1}{4}$, resulting in a joint entropy of 2 bits. Since the events are independent, the joint entropy is equal to the sum of the entropies of each individual coin toss.

Alternatively, two variables could be related. In the extreme case, two variables could be

completely correlated so that the value of one variable gives perfect information about the value of the other variable. For example, if the second coin toss occurred by picking up the coin and placing it back on the table with the same face up as before, then the probabilities of events [heads, heads] and [tails, tails] are both $\frac{1}{2}$, and the probabilities of [heads, tails] and [tails, heads] are both zero. The joint entropy is 1, and equal to either of the individual entropies.

In our analysis of fluorescent images we are interested in the co-occurrence of red and green colors. That is, we wish to know whether knowing the color intensity of green pixels tells us anything about the color intensity of red ones in the same location. We calculate the probabilities of all possible color intensities (0 to 255) in all corresponding locations of the red and green images. We define the joint probability $\mathrm{p}(r,g)$ as the probability of each pair of color intensities (0 to 255) occurring in the corresponding location in the red and green images. There are 256×256 = 65,536 possible combinations of color intensities. We calculate the number of times every intensity combination occurs in corresponding locations in an image. Then we divide by the total number of locations in the images to turn those occurrences into probabilities. These probabilities are entered in Equation (3.2) to calculate the joint entropy.

The joint entropy is low when color intensities repeatedly co-occur. Note that, joint entropy can be low when either the same color intensities repeatedly overlap, or when different color intensities overlap. For example, if red systematically has lower intensity than green, joint entropy would still be low if a green intensity of, say, 220 was frequently co-located with a red intensity of 180. Joint entropy only depends on the frequency of pairs of values co-occurring in the same locations. Joint entropy is high when there is no association in color intensities between the red and green images. Thus, in Figure 3.1(A) where red and green cells are uniformly randomly distributed, there is minimal co-occurrence of the intensities, and therefore all values in the probability table are low and uniformly distributed. In contrast, when red and green cells co-occur with the same intensities in the same locations (Figure 3.1(C)), the probabilities on the diagonal are high leading to the minimum possible joint entropy. We observe these scenarios in Figure 3.1(G) and Figure 3.1(I) which are the corresponding joint

probability tables for Figure 3.1(A) and Figure 3.1(C). For illustration purposes, the 256 color intensity values are binned into 4 color intensities.

Mutual Information

MI is calculated from the entropy of each image and the joint entropy of the two images using Equation (3.3).

$$\mathrm{MI}(r,g) = \mathrm{H}(r) + \mathrm{H}(g) - \mathrm{H}(r,g) \tag{3.3}$$

Intuitively, this formula calculates MI by subtracting the joint entropy of r and g from the total entropy in both r and g, which leaves the overlap in entropy of r and g.

In Figure 3.1, we illustrate how MI is calculated from a set of 3 simulated images. The first case (Figure 3.1(A)) shows simulated red and green cells placed uniformly in random locations. In most cases, red and green do not overlap as shown in Figure 3.1(D) (although by random chance, there is little co-occurrence of red and green cells that appear yellow). We calculate MI using Equation (3.3). Because there is little or no co-occurrence of red and green pixels in Figure 3.1(A), the joint entropy $\mathrm{H}(r,g) \approx \mathrm{H}(r) + \mathrm{H}(g)$, so MI ≈ 0.

The second case, in Figure 3.1(B), shows red cells placed within in a Gaussian distributed range of the green cells creating partial co-occurrence of red and green pixels. We can observe this region in Figure 3.1(E) (colored in yellow) which is the MI, calculated by summing the entropy of red and green images independently, and then subtracting the joint entropy (Equation (3.2)). The process to calculate the joint entropy of the two images are described in Section 3.4.5 Joint Entropy.

The third case (Figure 3.1(C)) is a special case where the red and green pixels are of same intensity residing in the same location. When separated as two images, red and green cells completely overlap, shown in Figure 3.1(F). In this case, information about the location of red cells provides all the information about the location of green cells. Because there is total correspondence between the intensity of red and intensity of green in the same location, the joint entropy $\mathrm{H}(r,g) = \mathrm{H}(r) = \mathrm{H}(g)$, and the MI therefore equals $\mathrm{H}(r)$ (and also equals $\mathrm{H}(g)$).

3.4.6 Normalized Mutual Information

The MI analysis quantifies in bits the amount information shared by images showing the locations of two different cell types. However, the number of bits is influenced by the dimension of images and the numbers and sizes of cells. It does not provide us with a universal scale with which to compare the association of T cells with other cell types. For this, we define and calculate NMI as:

$$\mathrm{NMI} = \frac{\mathrm{MI}(r,g)}{\min(\mathrm{H}(r),\mathrm{H}(g))} \tag{3.4}$$

We normalize MI by the minimum entropy image. MI depends on both the joint entropy and the internal (marginal) entropies of each color channel. The internal entropies vary across experiments, resulting in MI values that are not directly comparable. We normalize by dividing MI by the minimum of the internal entropies, since it provides an upper bound on MI, for a proof see [87].

The value of NMI is bounded between 0 and 1, where 0 indicates no occurrence of the red and green cells in the same location as in Figure 3.1(A), and 1 indicates complete colocalization of the red and green cells as shown in Figure 3.1(C). NMI allows us to directly compare spatial association of cells, regardless of the cell types, cell sizes, and image dimensions in our experiments.

We validated the NMI metric on simulated data generated as 512×512 RGB images shown in Figure 3.2(A). Each cell is 11×11 pixels (square shaped) with randomly chosen color intensities ranging from 0 to 255. In each image, 500 green cells are placed uniformly at random along with a number of red cells uniformly distributed between 100 and 500. We placed each red cell within a distance determined by a Gaussian distribution from each green cell with standard deviations (σ) ranging from 0 (generating complete correlation of the red and green pixels) to 10 (generating a low probability of overlap of red and green pixels). We treat the image as a torus to avoid edge effects when placing red cells. We also analyzed images in which both green and red cells are placed uniformly at random ($\mathscr{U}$), and therefore with no spatial

association and minimum MI.

NMI is designed to normalize for variations in cell numbers and differences in fluorescence between fields. Normalization makes the method more robust to cell count. To assess the potential effect of cell numbers on NMI, we simulated images in which we varied the cell numbers from 100-500 and calculated NMI for differing cell numbers with complete cell overlap ($\sigma = 0$, increasingly spatially separated $\sigma = 1$ or $\sigma = 3$ or cells placed in uniformly random distribution) Figure 3.3. We also calculated PCC as a comparison. We find that NMI is less sensitive to variations in cell numbers than PCC, particularly in cases in which there is already spatial association.

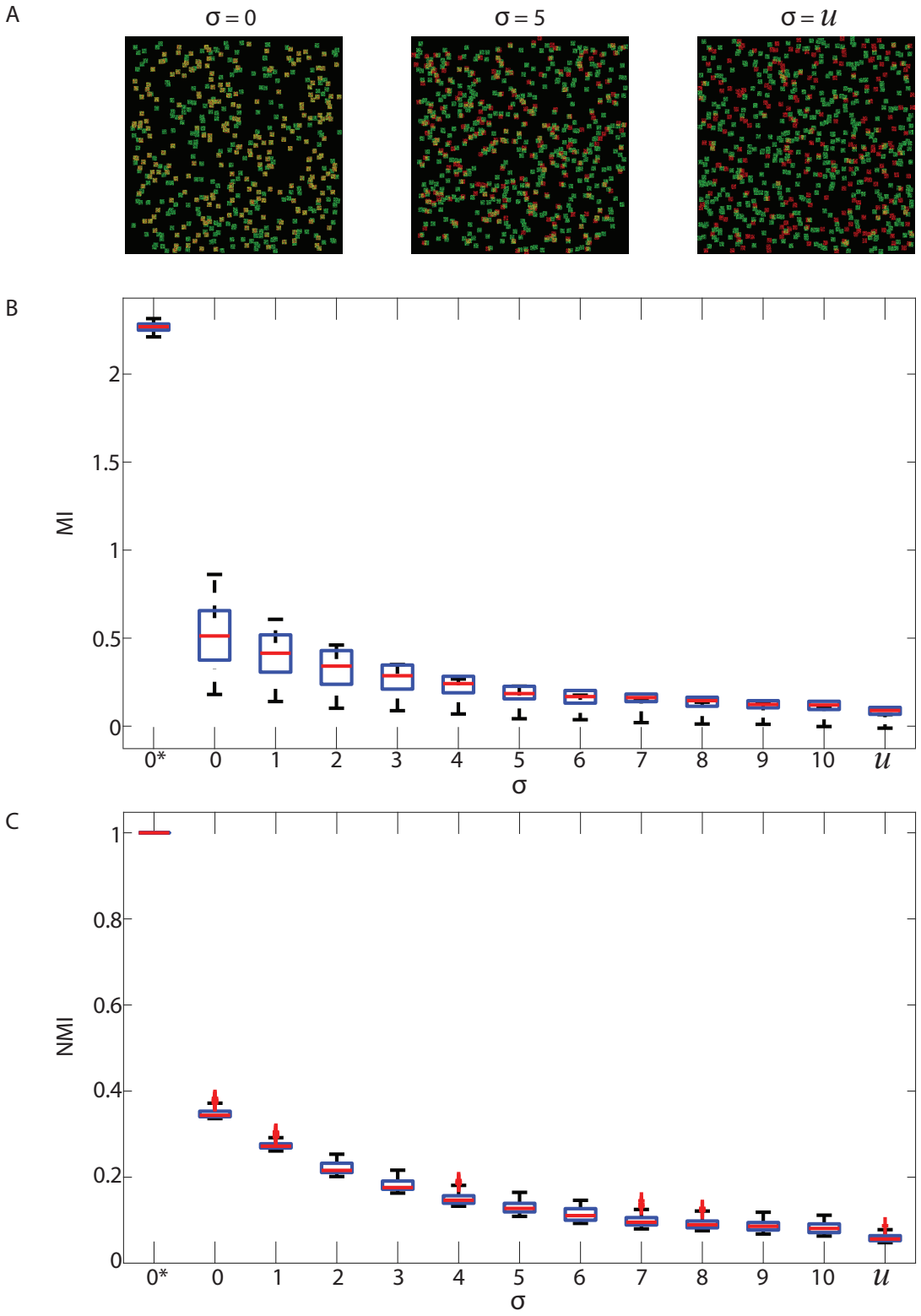

A
σ = 0
σ = 5
σ = u
B
MI
2
1.5
1
0.5
0
0*
0
1
2
3
4
5
6
7
8
9
10
u
σ
C
NMI
1
0.8
0.6
0.4
0.2
0
0*
0
1
2
3
4
5
6
7
8
9
10
u
σ

Figure 3.2: Validation of MI and NMI. Panel A shows 3 samples of simulated 512×512 images that consist of 500 green cells and a number of red cells uniformly distributed between 100 and 500. Each pixel intensity of the red and green cells is randomly assigned and each cell is 11×11 pixels (square shaped). The red cell locations are chosen from a Gaussian distribution centered at the location of green cells with standard deviation (σ) 0 and 5 in the first and second images, and uniformly random in the third image. (B) and (C) consist of multiple boxplots of MI (B) in bits and NMI (C) values for simulated images where the standard deviation (σ) ranges from 0 to 10 and 2 additional special cases: 0* and $\mathscr{U}$. 0* indicates that red and green color intensities are identical in corresponding locations which maximizes both MI and NMI. $\mathscr{U}$ indicates that the cells are placed uniformly at random within the image and with uniform random color intensity, resulting in the lowest MI and NMI. Increasing σ decreases the spatial association of cells and both MI and NMI systematically decrease, demonstrating that they are useful metrics that indicate spatial association between cells.

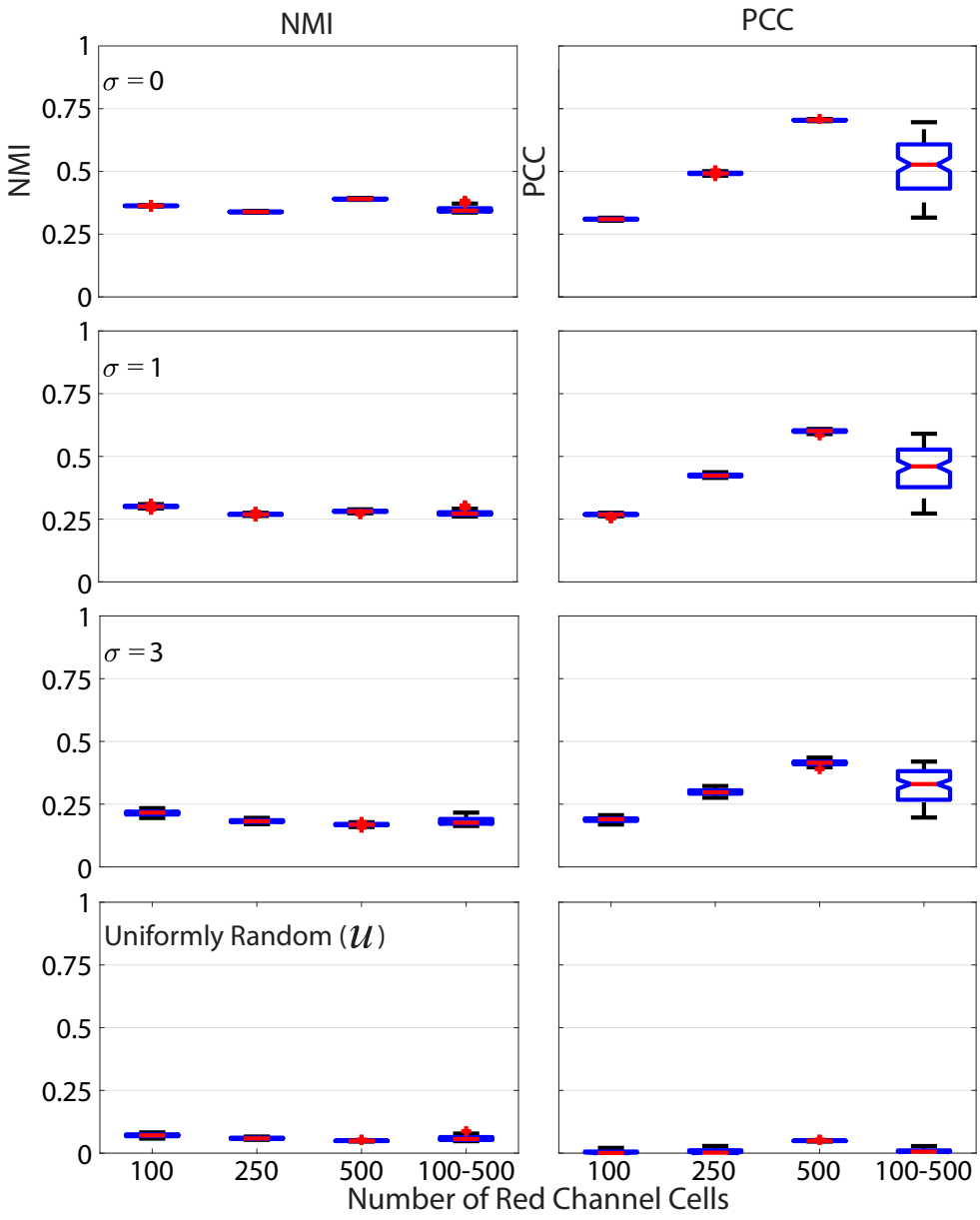

NMI
PCC
NMI
PCC
$\sigma = 0$
$\sigma = 1$
$\sigma = 3$
Uniformly Random ($\mathcal{U}$)
1
0.75
0.5
0.25
0
100
250
500
100-500
Number of Red Channel Cells

Figure 3.3: NMI is more robust than PCC to cell count. Simulated images were generated in which numbers of cells in the green and red channels are varied by number and positions varied as indicated. Apparent association of cell types based purely on the increased chance of two cells being near one another as the number of cells goes up is a concern. The normalization factor in NMI is intended to compensate for this artifact. Insensitivity to variation in cell number while preserving sensitivity to the underlying association between cell types distinguishes NMI from PCC. The number of cells in the green channel is kept constant at 500 while the number of cells in the red channel is varied. NMI results are shown in the left column and PCC in the right column. The spatial association between cell types in the model decreases from $\sigma = 0$ in the top row to uniform random placement in the bottom row.

3.4.7 Regionalization of Images

The NMI method takes into account the intensity and localization of pixels. However, cell sizes consist of multiple pixels. A naïve T cell has a diameter of approximately 5 µm-7 µm whereas the approximate length of a pixel is 1 µm. Therefore, we created regions in the image and call this process "regionalization". In regionalization, we chose a pixel (p) and calculated a region around it with given length, for example in a 5×5 pixel (6 µm × 6 µm) region, p is the middle pixel. We calculated the average intensity of the corresponding region and replaced the value of p with the average intensity value. Then we iterated over all pixels. We discarded the regions along the image boundaries where complete regions could not be formed. This method produced new images where each pixel has the average intensity of its region. We calculated the MI, NMI, and PCC of these regionalized images. We used region sizes: 5×5 pixels (6 µm × 6 µm), 15×15 pixels (18 µm × 18 µm), 25×25 pixels (30 µm × 30 µm). We are most interested in the results of region sizes between 5×5 (6 µm × 6 µm) and 15×15 pixels (18 µm × 18 µm), since these scales are most relevant to our biological data.

We validated both NMI and PCC for regionalized images. For validation, we used 512 × 512 simulated images that are constructed using the same method mentioned in Section 3.4.6 Normalized Mutual Information. Analysis is performed on 500 green cells and 500 red cells. These simulated images are then divided into regions using the regionalization method. The size of the regions are consistent with the ones we used for experimental data. Results from

NMI and PCC analysis on these images are shown in Figure 3.4. NMI and PCC decrease with decreasing spatial association, following a trend similar to that in the validation analysis shown in Figure 3.2, although region size influences PCC more than NMI.

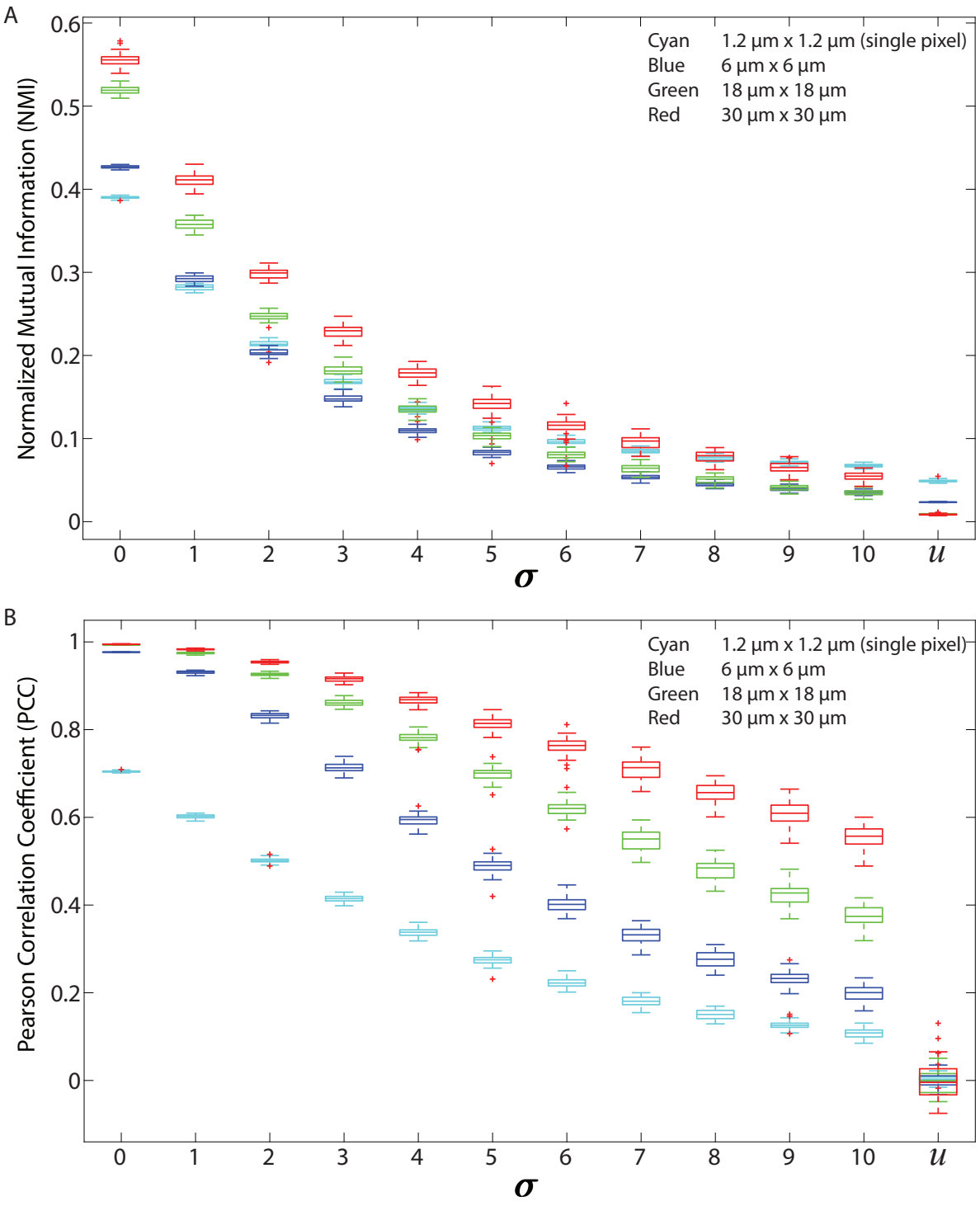

A
Normalized Mutual Information (NMI)
0.6
0.5
0.4
0.3
0.2
0.1
0
Cyan 1.2 µm x 1.2 µm (single pixel)
Blue 6 µm x 6 µm
Green 18 µm x 18 µm
Red 30 µm x 30 µm
0 1 2 3 4 5 6 7 8 9 10 u
σ
B
Pearson Correlation Coefficient (PCC)
1
0.8
0.6
0.4
0.2
0
Cyan 1.2 µm x 1.2 µm (single pixel)
Blue 6 µm x 6 µm
Green 18 µm x 18 µm
Red 30 µm x 30 µm
0 1 2 3 4 5 6 7 8 9 10 u
σ

Figure 3.4: Regionalized PCC and NMI on simulated data. Simulated images are 512×512 pixels with 500 red and 500 green 11×11 pixel square shaped cells. The red cell locations are chosen from a Gaussian distribution centered at the location of green cells with standard deviation (σ), which ranges from 0 to 10 and $\mathscr{U}$. $\mathscr{U}$ indicates that the cells are placed uniformly at random within the images and with uniform random color intensity. (A) NMI of simulated images with regions of 6 µm × 6 µm (blue), 18 µm × 18 µm (green), 30 µm × 30 µm (red), and single pixel (1 µm × 1 µm, cyan). (B) PCC of simulated images using the same regions.

3.5 Results

3.5.1 PCC shows T cells associate more with FRCs than DCs in LN

To ask whether naïve T cells associate with DCs in the LN, we used PCC, a standard colocalization measure. As a comparison, we also calculated the PCC of T cells and FRCs because it has been suggested that T cells use FRCs as a network for migration through the LN [13]. We transferred CMTMR-labeled T cells into CD11c-YFP mice, harvested LNs for 2PM imaging, and calculated PCC of T cells and DCs from multiple images of T cells and DCs. We imaged FRCs as previously described by [13] by irradiating Ubiquitin-GFP animals, reconstituting with whole bone marrow from non-GFP animals for 4-8 weeks, and co-imaged GFP+ FRCs with co-transferred CMTMR labeled T cells. We find the PCC of T:DC microscopy images was low (Figure 3.5(A)) (median = 0.19, results given to two significant figures throughout). In fact, the PCC of T cells to DCs was significantly lower than PCC of T cell with FRCs (T:FRC PCC median = 0.38). In Figure 3.5, we use interquartile-range notched box plots to visualize the statistical relationships between measurements [128]. Non-overlapping notches indicate the measurements were drawn from different distributions at the 95% confidence level. While previous studies have determined association of T cells with FRCs and DC subsets separately, we quantitatively compare the effect of FRCs relative to DCs on T cell positioning. These results suggest that FRCs show much higher correlation with naïve T cell locations in the T cell zone of LNs than the presumed intended targets of DCs.

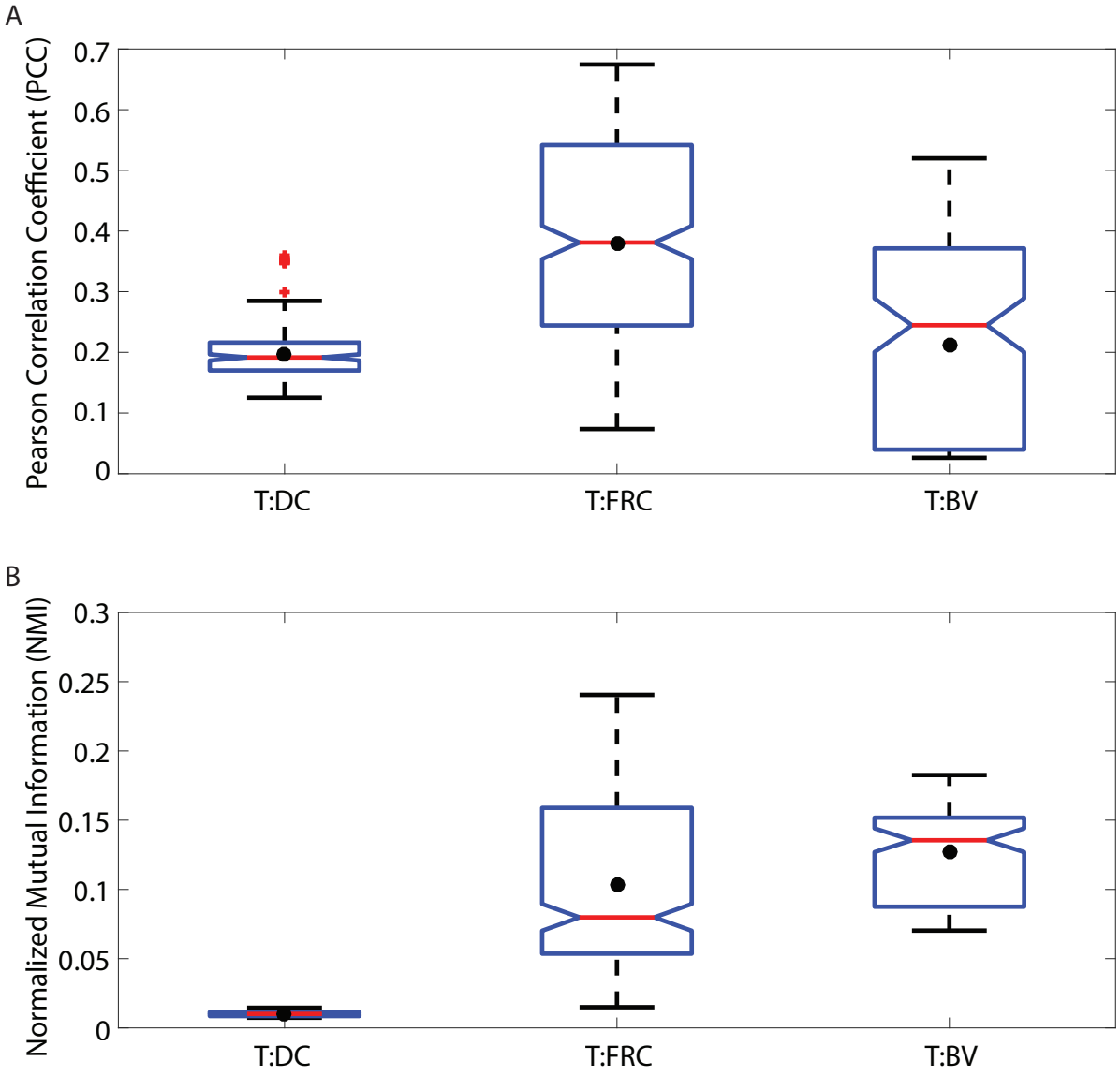


Figure 3.5: Notched boxplots displaying PCC (A) and NMI (B) values for T:DC, T:FRC, and T:BV images. Data include 6 T:DC image z stacks (2 experiments on 2 different days, 2 mice, 4 lymph nodes), 12 T:FRC image z stacks (3 experiments on 3 different days, 6 lymph nodes), 4 T:BV image z stacks (2 mice on 2 different days, 3 lymph nodes). Black dots indicate the mean. Median T:DC PCC value = 0.1922, median T:FRC PCC value = 0.3810, median T:BV PCC value = 0.2447. Mann Whitney p values for T:DC-T:FRC $<$ e-4, T:DC-T:BV = 0.0293, and T:FRC-T:BV $<$ e-4. Median T:DC NMI value = 0.0101, median T:FRC NMI value = 0.0798, median T:BV NMI value = 0.1355. Mann Whitney p values for T:DC-T:FRC, T:DC-T:BV, and T:FRC-T:BV comparisons $<$ e-4.

3.5.2 Application and validation of NMI as a novel method to assess T cell association with cell types in LN

While PCC provides a quantitative metric to assess the correlation among pixels in images, PCC assumes that these correlations are linear [3, 61, 77, 159]. We use NMI (a normalized version of MI) to quantitatively assess spatial relationships between cell types without assuming linearity. The principles of MI are illustrated using simulated images in Figure 3.1. MI has been previously used to understand co-registration of MRI images, but not previously applied to fluorescent images.

We calculated the entropy of fluorescence signals using Equation (3.1) and then calculated the joint entropy using Equation (3.2)(for detail see Methods). We then calculated the MI of the signals using Equation (3.3). To validate our MI calculations, we created simulated images with fields of green and red "cells" in which there is no association (Figure 3.1(A)), partial association (Figure 3.1(B)), and complete association (Figure 3.1(C)) of fluorescent objects with sizes similar to that of cells. The 3 cases can be simplified by observing the images in Figure 3.1(D) (no association) , Figure 3.1(E) (partial association marked as yellow area) and Figure 3.1(F) (complete association marked as yellow area). The joint probability tables (simplified examples in 4×4 color intensities shown in Figure 3.1(G), Figure 3.1(H), Figure 3.1(I)) are used to calculate the joint entropy. If there is no spatial association, the joint probability table shows evenly distributed low values (Figure 3.1(G)). Given the partial spatial association of cells, the joint probability table shows increased values across the diagonal (Figure 3.1(H)). Given completely overlapping signals, the joint probability table shows high values across the diagonal (Figure 3.1(I)). Because MI is calculated from fluorescent images in which different images possess different internal entropies, we normalized the MI values to provide a universal scale (between 0 and 1) with which to compare one image to another. We calculated NMI by normalizing MI with the minimum entropy of the two images, thus enabling quantitative comparisons across fields.

In Figure 3.2(A), we show examples of simulated images created for validating NMI (described in Section 3.4.6 Normalized Mutual Information) in which red cells were placed with standard deviation (σ) of 0 and 5 as well as red cells placed uniformly at random. We expect the MI and NMI values to decrease as the standard deviation increases, as shown in Figures 3.2(B) (MI) and 3.2(C) (NMI). As expected, MI and NMI are maximum in the special case 0* where the intensity, size and location of the cells are all identical; MI and NMI decrease as the spatial association between the cells decreases. While the MI can be greater than 1 bit (Figure 3.2(B)), the NMI metric is normalized to be between 0 and 1 (Figure 3.2(C)), demonstrating that NMI can provide comparisons to account for differing levels of fluorescence across multiple fields on a common scale.

As a further validation, we tested whether NMI calculations on our experimental data range between 0 and 1. Figure 3.6 shows that the NMI of an image with itself is 1 (Matched Red:Red and Matched Green:Green). We calculated NMI of two unrelated images from two different experimental fields (Unmatched Red:Green). For example, the red cell image may be taken from a T:DC experiment and the green cell image from a T:FRC experiment. As expected, NMI in these cases is very close to 0 (Figure 3.6). We then calculated the NMI of T:DC and T:FRC interactions using the same images on which we calculated PCC (Figure 3.5(B)). We find that similar to PCC analyses, NMI shows significantly higher association for T:FRC than T:DC (T:FRC NMI median = 0.08; T:DC NMI median = 0.01).

3.5.3 Regional PCC and NMI analyses

We first calculated both PCC and NMI using pixel-based comparisons (Figure 3.5). We find that PCC and NMI show a significantly higher association of T cells with FRCs than DCs. However, NMI and PCC pixel based metrics can be problematic. Intercellular interactions in 2PM images are challenging to quantify by existing colocalization analyses because individual cells occupy discrete physical space, but pixel-based colocalization methods measure the amount of fluorescence signal overlap in individual pixels. In fact, any actual overlap in cell signal

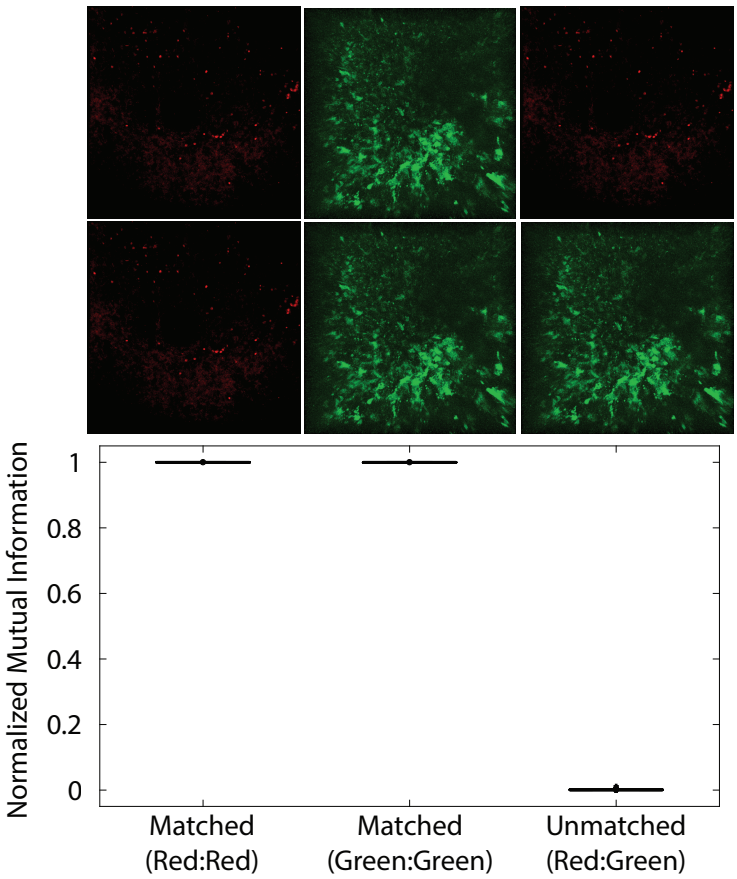


Figure 3.6: Illustration of the highest and lowest NMI that can be generated from the experimental data. The NMI of an image with itself is the maximum value of 1, shown for an example image of red cells and an example image of green cells. To obtain a minimum value, we calculate NMI between two images, one red and one green from two different fields so that the images are unrelated. We calculated NMI from 5,036 pairs of frames (Unmatched Red:Green). For this unmatched scenario, the NMI is very close to 0 (median is 0.008).

as measured by PCC and NMI is likely artefactual in that cells do not physically overlap in space. Also, it is possible that true intercellular contacts would be underestimated due to image resolution and the inability to resolve smaller protrusions such as dendrites of DCs. To account for cell-cell association rather than actual signal overlap based on pixels, we regionalized our images using sliding windows of multiple pixels, the size of which matched approximate sizes of T cells, DCs, and FRCs (estimated 5-7 μm diameter). The regionalized image has the same number of pixels as the original, but each pixel contains information drawn from the region surrounding it. Given that each pixel is approximately 1 μm in length, we created regions of 5×5

pixels (6 µm × 6 µm) and 15×15 pixels (18 µm × 18 µm) to account for potential extensions beyond the cell bodies. We also extended the analysis to larger region sizes. Fluorescence in regions was determined by taking the average fluorescence of all the pixels within the region (for detail see Section 3.4.7 Regionalization of Images). We used this method to generate new regionalized images and performed both PCC and NMI to take into account potential interactions of cells without directly overlapping fluorescent signals.

We first tested the "regionalization" effect by performing PCC and NMI on simulated images (as shown in Figures 3.1 (A), (B), and (C) and 3.2(A)) to determine the effect of cell density, degree of pixel overlap, and regionalization on co-association (Figure 3.4). We created simulated images that approximate the amount of fluorescence in our experimental images. We varied the distance between the simulated cells to model different amounts of spatial association. We applied our regionalization method to these simulated images and calculated NMI and PCC values. We found that larger regions produce higher NMI and PCC values. Compared to NMI, PCC is less sensitive to changes in spatial association but more sensitive to region size (compare Figure 3.4(A) and 3.4(B)). Despite these differences, both NMI and PCC provide a quantitative measure that can be used to detect variation in spatial association among cell types.

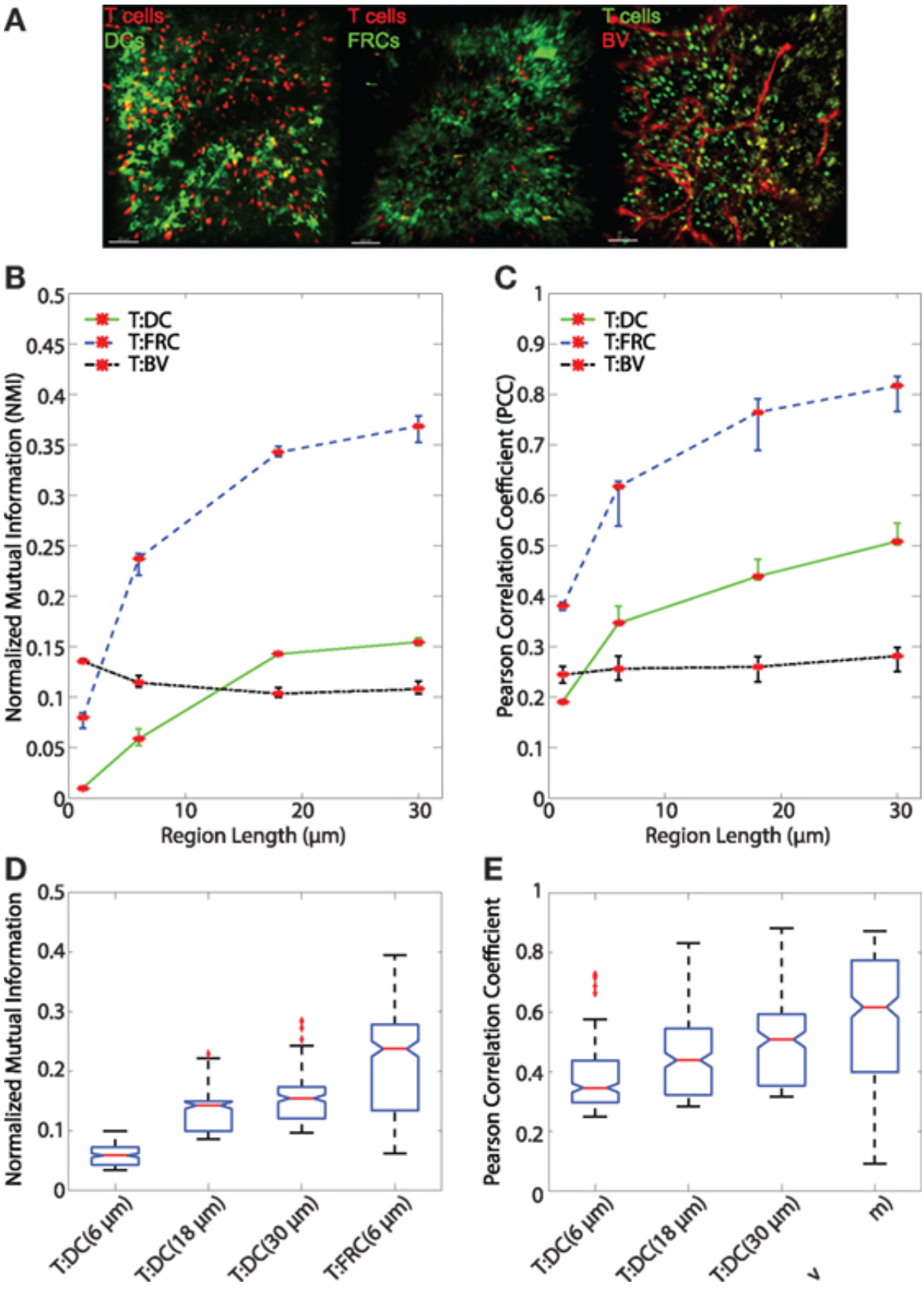

A
T cells
DCs
T cells
FRCs
T cells
BV
B
T:DC
T:FRC
T:BV
Normalized Mutual Information (NMI)
Region Length (μm)
C
T:DC
T:FRC
T:BV
Pearson Correlation Coefficient (PCC)
Region Length (μm)
D
Normalized Mutual Information
T:DC(6 μm)
T:DC(18 μm)
T:DC(30 μm)
T:FRC(6 μm)
E
Pearson Correlation Coefficient
T:DC(6 μm)
T:DC(18 μm)
T:DC(30 μm)

Figure 3.7: (A) Sample images of T:DC (T cells labeled in red and DCs labeled in green), T:FRC (T cells labeled in red and FRCs labeled in green), and T:BV (T cells labeled in green and blood vessels labeled in red). (B, C) Line plots representing the NMI (B) and PCC (C) of T cells and DCs (T:DC, green line), T cells and FRCs (T:FRC, blue dashed line), and T cells and blood vessels (T:BV, black dotted line). NMI and PCC were calculated on pixels (Region Length = 1 µm), or regionalized images of increasing side length (6 µm, 18 µm and 30 µm). Red stars indicate medians for the corresponding region size, and error bars indicate the 95% confidence interval around the median [9]. For NMI, Mann Whitney p values for T:DC-T:FRC, T:DC-T:BV, and T:FRC-T:BV comparisons $<$ e-4 for all region lengths except T:DC-T:BV(region length = 18 µm) p value = 0.0012. For PCC, Mann Whitney p values for T:DC-T:FRC, T:DC-T:BV, and T:FRC-T:BV comparisons $<$ e-4 for all region lengths except T:DC-T:BV (region length = 1 µm) p value = 0.0293. (D, E) Notched box plots comparing the NMI (D) and PCC (E) of T cells and DCs with T cells and FRCs at physiologically relevant region lengths of (6 µm, 18 µm and 30 µm) for T:DC associations and 6 µm for T:FRC associations. Note different scales on the y-axis. Both NMI and PCC are greater for the physiologically relevant region sizes for T:FRC than for T:DC (comparing T:DC at 30 µm to T:FRC at 6 µm $p = 0.0022$; for all other comparisons $p < e-4$). T:DC images were from 6 image z stacks consisting of 4089 frames from 2 mice and 4 lymph nodes. T:FRC images were from 12 image z stacks consisting of 9,468 frames from 3 mice and 6 lymph nodes. T:BV images were from 4 image z stacks consisting of 4,361 frames from 2 mice and 3 lymph nodes.

3.5.4 Regional analyses confirm that T cells are more associated with FRCs than with DCs

After validating both the NMI metric and the regionalization of images, we analyzed regionalized images to quantify spatial association of T cells with DCs and FRCs using both PCC and NMI. Both PCC and NMI show that T cells associate less with DCs than FRCs (Figure 3.7(B) for NMI and Figure 3.7(C) for PCC). T cells are more associated with FRC across all region sizes. In pixel-based comparisons (without regionalizing), the T:DC association was very low (Table 3.1, (Figure 3.7,NMI = 0.0101; PCC = 0.1916) while T:FRC association was significantly higher (NMI = 0.0798; PCC = 0.3810). Both NMI and PCC values for T:DC interactions increased with increasing region sizes, T:FRC association also increased at each region size. Regionalizing PCC into 18 µm $\times$ 18 µm region (15$\times$15 pixels) resulted in the same trend among the compared cell types as NMI (Figure 3.7(B) NMI; T:DC median = 0.1427, T:FRC median = 0.3426; (Figure 3.7(C) PCC T:DC median = 0.4396, T:FRC median = 0.7646, Table 3.1).

Data Type	**Median** NMI	**95% Confidence Interval**	**Median PCC**	**95% Confidence Interval**
Random Control	0.0008	[0.0007, 0.0008]	0.0008	[0.0005, 0.0010]
Same Image control	1	[1, 1]	1	[1, 1]
1 μm × 1 μm (Single Pixel)				
T:DC (WT)	0.0101	[0.0090, 0.0102]	0.1916	[0.1879, 0.1941]
T:DC (CCR7$^{-/-}$)	0.0158	[0.0156, 0.0161]	0.1527	[0.1338, 0.1589]
T:FRC	0.0798	[0.0691, 0.0846]	0.3810	[0.3729, 0.3886]
T:BV	0.1355	[0.1348, 0.1381]	0.2447	[0.2281, 0.2610]
6 μm × 6 μm				
T:DC (WT)	0.0588	[0.0524, 0.0685]	0.3467	[0.3427, 0.3808]
T:DC (CCR7$^{-/-}$)	0.0857	[0.0808, 0.0886]	0.4252	[0.3720, 0.4334]
T:FRC	0.2377	[0.2207, 0.2427]	0.6175	[0.5392, 0.6283]
T:BV	0.1144	[0.1101, 0.1214]	0.2565	[0.2342, 0.2815]
18 μm × 18 μm				
T:DC (WT)	0.1427	[0.1418, 0.1443]	0.4396	[0.4327, 0.4734]
T:DC (CCR7$^{-/-}$)	0.2633	[0.2576, 0.2679]	0.5866	[0.5794, 0.5957]
T:FRC	0.3426	[0.3384, 0.3487]	0.7646	[0.6893, 0.7913]
T:BV	0.1036	[0.1002, 0.1093]	0.2603	[0.2302,0.2805]
30 μm × 30 μm				
T:DC (WT)	0.1547	[0.1509, 0.1589]	0.5089	[0.5020, 0.5448]
T:DC (CCR7$^{-/-}$)	0.3075	[0.2980, 0.3165]	0.6590	[0.6527, 0.6673]
T:FRC	0.3685	[0.3525, 0.3789]	0.8169	[0.7659, 0.8352]
T:BV	0.1080	[0.1034, 0.1159]	0.2816	[0.2514,0.2984]

Table 3.1: Median NMI and PCC values among cell types with 95% confidence interval. Both NMI and PCC values increase with region size except for T:BV.

Figure 3.7(D) and (E) compare physiologically relevant regions that approximate cell sizes and account for potential dendritic extensions with larger regions for DCs at 18 μm and 30 μm than FRCs at 6 μm. Again, T:FRC associations are greater than T:DC associations using both NMI and PCC. Thus, across region sizes, both NMI and PCC analyses show significantly higher T cell association with FRCs compared with DCs. These results suggest that despite the fact that DCs are considered the ultimate targets for T cell search, FRCs a greater determinant of naïve T cell positioning within the LN.

In addition to FRCs and DCs, structures such as blood vessels in the LN can be sources of chemokines [88, 179], and T cells may move along vessels in other tissues [138]. Several studies suggest DCs are biased to localize near blood vessels and efficiently activate antigen-specific T cells [14, 131]. We used NMI and PCC to ask whether vasculature can determine T cell localization in LN. We transferred GFP+ T cells for 16 hours as previously described, then just prior to imaging, we injected animals with DyLight 594-lectin which binds endothelial cells lining blood vessels. We then imaged T cells in conjunction with vasculature in LNs. With the

pixel based PCC (Figure 3.5(A)) and NMI analyses (Figure 3.5(B)), T cell association with blood vessel appears higher than T cell association with DCs, and NMI shows higher T cell association with blood vessels than even FRCs. However, with increasing region size, PCC and NMI analyses of T:BV values stayed consistent while T:DC values increased, for example, in the 18 μm length region, NMI of T:DC was 0.1427 and T:BV was 0.1036. The same trend was seen for PCC (T:DC = 0.4396, T:BV = 0.2603). The consistent value of NMI and PCC analyses of T:BV across regions likely reflects the sharp resolution of the blood vessel fluorescence compared with the more blurred extensions of FRCs and DCs. With increasing region size matching cellular scales, T cells show lower association with BVs (Figure 3.7 (B) and (C)). These results suggest that T cells likely do not use crawling along vessels as a means to migrate within T cell zones of LNs.

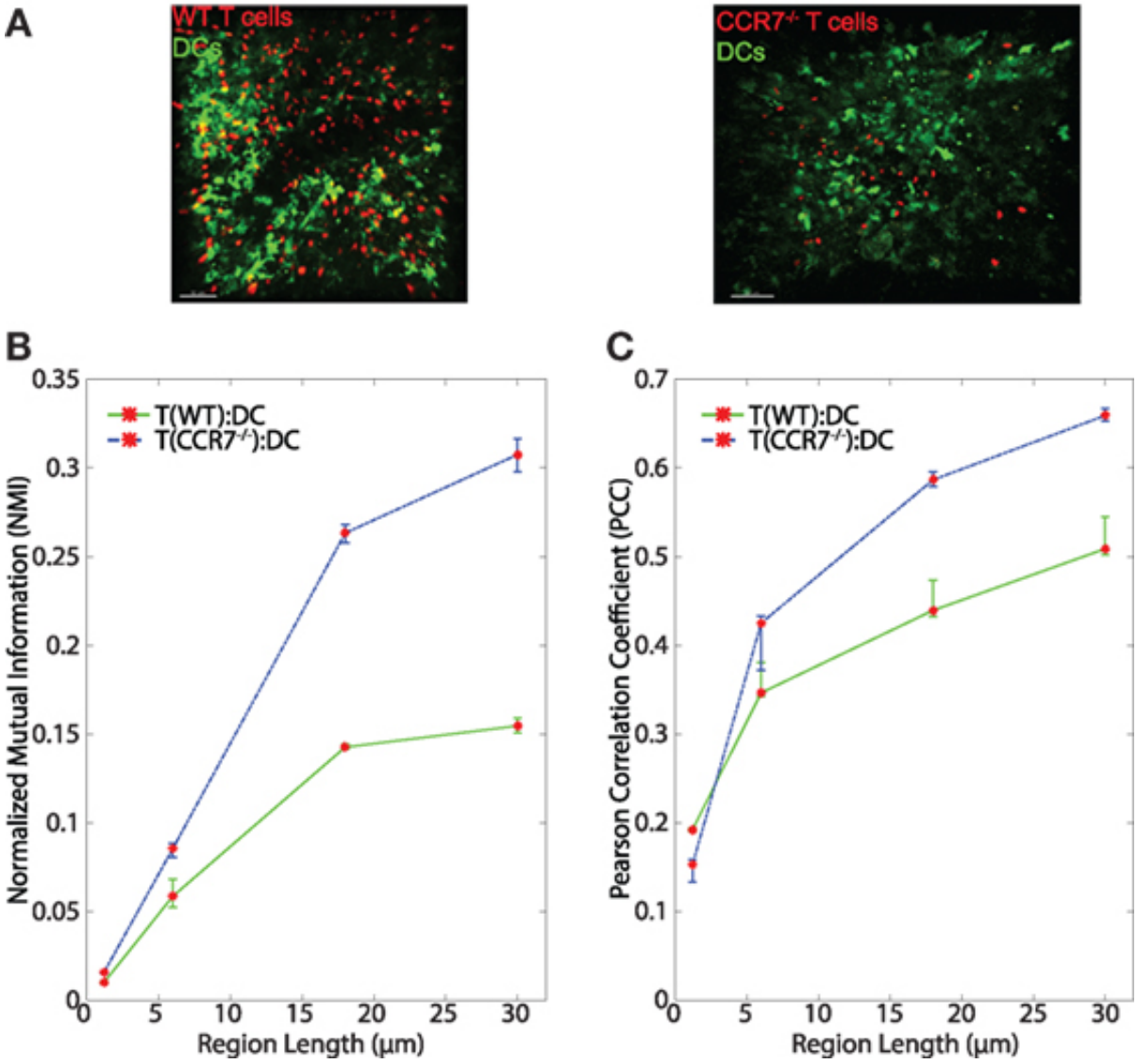

A
WT T cells
DCs
CCR7-/- T cells
DCs
B
T(WT):DC
T(CCR7-/-):DC
Normalized Mutual Information (NMI)
Region Length (μm)
C
T(WT):DC
T(CCR7-/-):DC
Pearson Correlation Coefficient (PCC)
Region Length (μm)

Figure 3.8: (A) Sample images of WT T:DC and CCR7$^{-/-}$ T:DC. T cells are labeled in red and DCs are labeled in green. In WT T:DC, T cells are wildtype naïve T cells and in CCR7$^{-/-}$ T:DC, T cells are from CCR7-deficient animals. (B,C) Line plots representing the NMI (B) and PCC (C) of WT T cells and DCs (T(WT):DC, green line) and CCR7$^{-/-}$ T cells and DCs (T(CCR7$^{-/-}$):DC, blue dashed line). NMI and PCC were calculated on pixels (Region Length = 1 µm), or regionalized images of increasing side length (6 µm, 18 µm and 30 µm). Red stars indicate medians for the corresponding region size, and error bars indicate the 95% confidence interval around the median [9]. For NMI, Mann Whitney p values for T(WT):DC-T(CCR7$^{-/-}$):DC comparisons $<$ e-4 for all region lengths. For PCC Mann Whitney p values for T(WT):DC-T(CCR7$^{-/-}$):DC comparisons for region lengths 1.2, 6, 18, and 30 µm: Region length 1 µm p $<$ e-4, 6 µm p = 0.9152, 18 µm p = 0.0021, 30 µm p $<$ e-4. WT T:DC images were from 6 image z stacks consisting of 4089 frames using 2 mice and 4 lymph nodes. CCR7$^{-/-}$ data are from 12 image z stacks consisting of 11,294 frames using 4 mice and 8 lymph nodes.

3.5.5 CCR7 does not enhance T:DC association

The chemokine CCL21 plays an important role in driving rapid motility of naïve T cells in LNs, and this rapid motility has been suggested to enhance T cell interactions with DCs [209]. We tested whether signaling through CCR7 might provide information to T cells to enable closer T:DC associations. To do this, we transferred CMTMR-labeled CCR7$^{-/-}$ T cells into CD11c-YFP mice, harvested LNs for 2PM imaging, and calculated NMI and PCC of CCR7$^{-/-}$ T cells and DCs. Contrary to our hypothesis, we found that in general, CCR7$^{-/-}$ T cells and DCs showed slightly higher NMI and PCC than WT T:DCs (Figure 3.8(B), NMI WT: 0.0101; CCR7$^{-/-}$: 0.0158 and Table 3.1). WT T cells showed higher co-association with DCs compared with CCR7$^{-/-}$ T cells in only one case, pixel-based PCC analysis, while with increasing region size and in all NMI analyses, CCR7$^{-/-}$ T cells were slightly increased in DC association over WT T cells (Figure 3.8(B) and (C), Table 3.1). Based on both NMI and PCC analyses, these data show that CCR7 does not promote increased T cell localization with DCs. Absence of CCR7 did not increase T:DC association to the level of T:FRCs, as NMI and PCC values of T:FRC remained significantly higher than CCR7$^{-/-}$ T:DC association. These results suggest that high speed motility promoted by CCR7 signaling likely functions to promote T cell exploration of the LN paracortex rather than increase T cell localization close to DCs.

3.6 Discussion

In this work, we analyze 2PM movies to quantitatively compare T cell association with different cell types and structures in the naïve lymph node using both PCC and NMI. To account for the limitations of 2PM to resolve cell structures, we create regions that correspond to physiologically relevant cell sizes. Both PCC and NMI across multiple region sizes show that T cells share more spatial association with FRCs than with DCs. Furthermore, CCR7$^{-/-}$ T cells do not associate less with DCs than WT T cells; in fact, our results suggest that CCR7$^{-/-}$ T cells may associate slightly more with DCs than WT T cells.

Many studies have investigated T cell search for DCs in the naïve LN since DCs are the key cell type that is required to present cognate antigen to T cells leading to the initiation of the adaptive immune response [106, 209]. [206] suggest that cell positioning within the LN maximizes the likelihood of T cell interaction with DCs. Other studies hypothesize that DCs are situated atop the FRC network to facilitate T cell interactions with DCs as the T cells move along the FRCs [83] and that T cells enter the paracortex from HEVs at specific entry points contiguous with the FRC network, enabling T cells to be "received" by a greeting line of DCs positioned on top of the FRCs near the HEV entry points [117]. Further, different subpopulations of DCs have been shown to localize to specific regions in the LN, suggesting that DC positioning relative to T cells may facilitate T cell activation [84]. However, our quantitative analysis using NMI and PCC suggest that T cell association with FRCs does not lead to similarly high association with DCs. The lack of association between T cells and DCs suggests that T cells have no *a priori* knowledge of DC positions and that DCs are unlikely to attract T cells to DC locations prior to infection. While there is evidence that upon DC activation and infection, chemokines are important to mediate T cell repositioning to DCs [39, 89, 115], our data suggests that chemokines CCL19/21 that bind to CCR7 do not play a role in T cell positioning to DCs in the absence of infection. [80] previously demonstrated that T cells move with a lognormal correlated random walk, which aligns with several other studies in the LN [17, 133]. Our results

suggest that random movement, rather than guided movement, may be the strategy that naive T cells use to interact with DCs.

Although T cells and DCs have low NMI and PCC, we find that unexpectedly, lack of CCR7 does not decrease association between T cells and DCs, in fact, CCR7-deficient T cells show slightly increased association with DCs. CCR7 mediates high speed motility in LNs [100]. One possible explanation for our finding is that CCR7 deficiency in T cells results in slower T cells that cannot efficiently move away from DCs once they have made contact. Alternatively, CCR7 signaling might be important for T cells to move along FRCs where they receive chemokinetic and survival signals, including both CCL21 and other cytokines such as IL-7 so that in the absence of CCR7, T cells stay closer to DCs, which are not the primary source of CCL21 [101, 119]. While it is known that CCR7-deficient T cells are less capable of activation, our quantitative analysis suggests that this may not be due to lack of T:DC contacts but rather may be due to CCR7 effects on overall motility or effects on cosignaling with T cell receptors.

We validated both NMI and PCC on simulated data where we directly manipulated the spatial association of cells and showed that both metrics decrease as spatial association decreases and as region size increases (Figure 3.4). We designed NMI to normalize for differences in fluorescence between fields, and NMI can quantify non-linear relationships between variables [177] while PCC is based on correlation coefficients [3, 61]. Additionally, information based measures are theoretically insensitive to coarse graining [54]. Our regional NMI analyses in both simulated and experimental images is consistent with this theoretical prediction in that NMI is less sensitive to region size than PCC (Figure 3.4 and 3.7). We find that NMI is also less sensitive to variations in cell number than PCC, particularly in cases in which there is already spatial association (Figure S2). Further, NMI based on regions avoids problems associated with pixel-distance measures that arise from 2PM images containing transient single pixel noise [149]. Cell-distance measures are also problematic because they require the boundaries of cells, or their centroids, to be well defined. That is usually not the case in 2PM images, especially in

the case of DCs and FRCs. We find there are advantages to our approach over regional mutual information (RMI) [164], in particular RMI fails for region sizes greater than 6 μm in length. For scales where RMI can be applied the results are in line with PCC and NMI.

While both NMI and PCC consistently show that T cells are more spatially associated with FRCs than with DCs, we note several caveats in interpreting these results. We considered that T cells may share the highest NMI or PCC with the most numerous cells or structures that occupy the most volume in the paracortex, simply because they cannot move away from the abundant cell type or structure without encountering another cell or structure of the same kind. However, our simulations (Figure 3.2(C)) validated that NMI is insensitive to variation in cell number, with 5-fold variation in cell number causing much less effect on NMI than changes in spatial association. While the amount of background noise (low-level fluorescence of individual pixels) has some effect on NMI and PCC, that effect does not change the conclusion that NMI and PCC both indicate higher spatial association of T cells with FRC than with DC.

Similar to previous studies, our experimental method uses irradiation to image FRCs showing residual GFP+ hematopoeitic cells (between 5-10%). Thus, it is possible that T:DC can contribute to the T:FRC NMI and PCC. However, because NMI and PCC of T cells with DCs is significantly lower, it is unlikely that the increase in T cell association seen with FRCs is due to residual DC signal. There may also be limitations in the use of two photon imaging as the primary mode of visualizing T cell interactions in the T cell zone as the T cell zone is usually deeper in the LN cortex. Thus, although many publications have used two photon imaging to understand T cell motion in LNs, T cell associations with FRCs and DCs may vary depending on the specific areas that are imaged. Additionally, it is possible that staining specific subsets of T cells or DCs may reveal more or less spatial association than we see with total T cells and all CD11c+ cells.

In summary, our results show that NMI and PCC both provide quantitative methods to analyze the relationship between two sets of objects, validated in simulations. NMI and PCC show significant differences for different cell populations labeled with two different fluorescent

markers, providing quantitative comparisons of fluorescent microscopy images across multiple fields [180]. Thus, both NMI and PCC of physiologically relevant regions are useful tools to quantify the relationship between fluorescent cell types. Since MI is a general method for measuring colocalization of fluorescence microscopy images including 2PM signals, the NMI and regional analyses may be broadly applied to any colocalization study of differentially fluorescent objects in the LN and more generally.

Chapter 4

Information-Theoretic Exploration of Multivariate Time-Varying Image Databases

4.1 Publication Notes

Citation: H. Tasnim, S. Dutta, T. L. Turton, D. H. Rogers, and M. E. Moses, "Information Theoretic Exploration of Multivariate Time-Varying Image Databases," in Computing in Science & Engineering, vol. 24, no. 3, pp. 61-70, 1 May-June 2022, doi: 10.1109/MCSE.2022.3188291.

Editor: Ilkay Altintas, University of California San Diego and Nicola Ferrier, Argonne National Laboratory

Publisher: IEEE

Published: 11 July 2022

Formatting: The original published text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Funding: This work was supported by funding from the following: 10.13039/100000015-U.S. Department of Energy;10.13039/100008902-Los Alamos National Laboratory; 10.13039/100022829-

Triad National Security LLC; 10.13039/100000015-U.S. Department of Energy (Grant Number: 89233218CNA000001 and LA-UR-21-21824)

Acknowledgements: We would like to thank Moses Bio Computational lab for the useful discussion and suggestions.

4.2 Abstract

Modern scientific simulations produce very large datasets, making interactive exploration of such data computationally prohibitive. An increasingly common data reduction technique is to store visualizations and other data extracts in a database. The Cinema project is one such approach, storing visualizations in an image database for post hoc exploration and interactive image-based analysis. This work focuses on developing efficient algorithms that can quantify various types of multivariate dependencies existing within multi-variable datasets. It applies specific mutual information measures for the quantification of salient regions from multivariate image data. Using such information measures, the opacity of the images is modulated so that the salient regions are automatically highlighted and the domain scientists can interactively explore the most relevant regions for scientific discovery.

4.3 Introduction

Image-based data reduction techniques have emerged as one of the viable solutions to minimize the size of the stored data so that it can be analyzed and visualized interactively post hoc by the application scientists [114]. Storing large-scale three-dimensional multivariate simulation datasets in the form of an indexed image database, called a Cinema Database[1] [5], facilitates exploration of the large-scale scientific data efficiently without overwhelming the users. These Cinema databases are ideally generated in situ, i.e., when the simulation is running on the supercomputer and the data is not yet moved to the disks. Instead of keeping the raw data,

[1] https://cinemascience.github.io

Cinema databases are stored onto disk as a proxy for the data, capturing various types of visualizations of the data. Later during offline analysis, the Cinema databases can be explored interactively to analyze the data in the image space. The success of this approach has been shown in many application domains [5, 16].

Even though Cinema databases result in a significant amount of data reduction, such databases still consist of multiple variables, timesteps, visualization parameters, etc. Hence, efficient image-based data analysis and visualization algorithms are necessary to find salient data features automatically so that the domain experts do not have to manually explore them. This problem becomes more challenging when the experts want to analyze features in the multivariate spatiotemporal domain to study their interaction pattern. In many scientific applications, variables collectively show association/dissociation relationships and such properties are often correlated to a physical phenomenon in the data. For example, in hurricane simulation data, low-pressure and low-velocity regions are characterized as the hurricane eye, indicating the strength of the storm. Therefore, multivariate analysis techniques are essential to efficiently detect association/dissociation relationships in image databases. Ideally, these relationships should be visually incorporated to the image database to support further interactive exploration for new scientific discovery.

In this work, we propose an information-theoretic analysis framework that works on multivariate time-varying Cinema databases and performs automatic identification of salient regions given a pair of variables. The technique uses *specific mutual information measures* (SMI) that are a decomposition of traditional mutual information so that the information content of specific data values can be quantified. Each SMI measure captures a unique multivariate property of the data. Using the strength of these SMI measures, the opacity of the images is modulated during visual analysis so that the important spatial regions are highlighted automatically and the users can quickly focus on them while exploring the Cinema databases. The analysis results are presented interactively using a web-based visual-analytics tool, CinemaView[2], which allows

[2]https://github.com/cinemascience/cinema_view

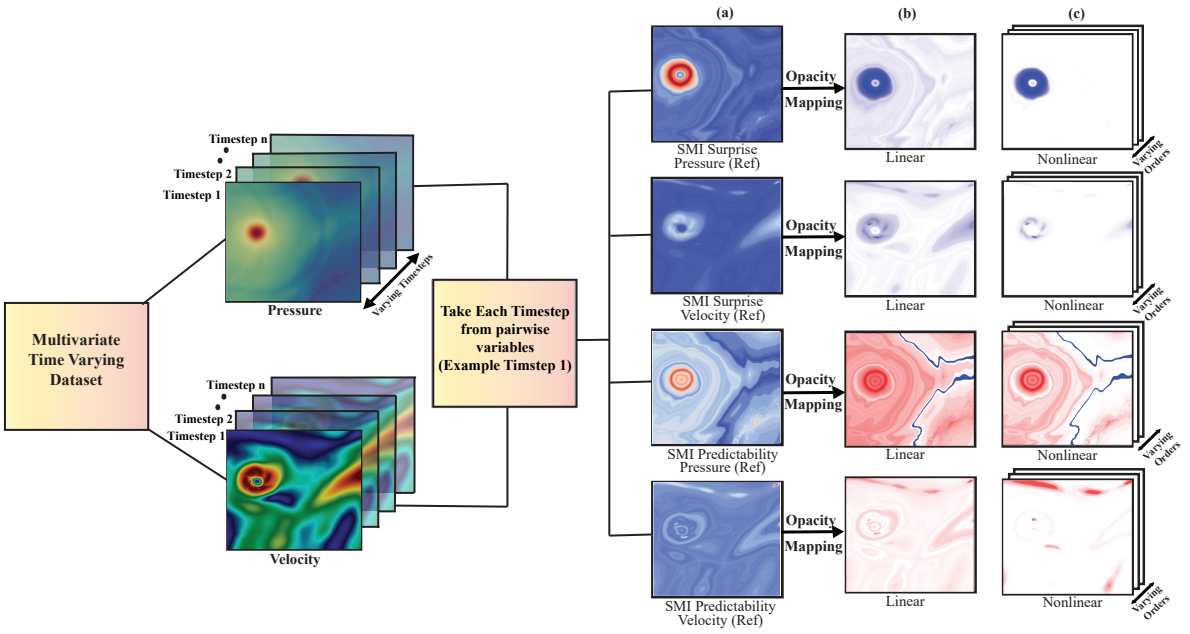


Figure 4.1: An illustrative diagram of our workflow. Here we have chosen the variables pressure and velocity from the Hurricane Isabel dataset to demonstrate the steps in our technique. Specific mutual information (SMI) measures: Surprise and Predictability are applied on the variable pairs and corresponding images are shown in column (a). After modulating the opacity using linear and nonlinear mapping functions, images with salient regions are analyzed as shown in columns (b) and (c) respectively.

side-by-side interactive comparison of analysis results. The efficacy of the proposed framework is demonstrated by applying it to scientific simulation datasets from weather and combustion sciences.

The contributions of our work are twofold:

- We propose a new technique to perform automatic feature analysis in multivariate time-varying scientific data. Our image-based representations of the 3D spatiotemporal data help reduce the overhead of the analysis significantly.

- We propose an information-theoretic opacity mapping technique to highlight the statistically salient regions in the data considering pairs of variables.

4.4 Related Works

In this section, we present a comparative discussion of the existing related works and indicate how our work is different. Information theory [49] have been used successfully for solving problems across many computational domains [40, 183]. Instead of using traditional mutual information, the use of various decomposition of mutual information, called specific mutual information (SMI), have gained significant attraction in recent years. By applying SMI, Bramon et al. showed that multi-modal 3D medical datasets can be fused into a single dataset [26]. In another work, Bramon et al. used mutual information to design color transfer function for medical data [25]. To analyze uncertainty of isosurfaces in scientific 3D data, Biswas et al. [22] used SMI and Dutta et al. extended this work into time-varying domain [67]. In contrast to the above works, in this work, we have focused on 2D image-based databases, generated from multivariate time-varying simulations, where our primary focus is to use SMI to automatically first detect the statistically salient regions considering images from variable pairs and then use the SMI values at each pixel location to define opacity values so that the salient regions are automatically highlighted. These images will be ideally generated during the simulation run, i.e., in situ, and as these simulations can have many variables and hundreds to thousands of time steps, we believe that our approach can significantly accelerate the multivariate analysis for the domain scientists by providing them an image-based time-varying summary of simulation variable interactions where the salient regions are automatically highlighted.

4.5 Proposed Methods

4.5.1 Overview

Our aim is to develop an interactive analysis technique to enable scientists to explore salient regions in time-varying multivariate datasets. The images in the Cinema database are derived from three-dimensional simulation data for each variable over multiple timesteps. To study the

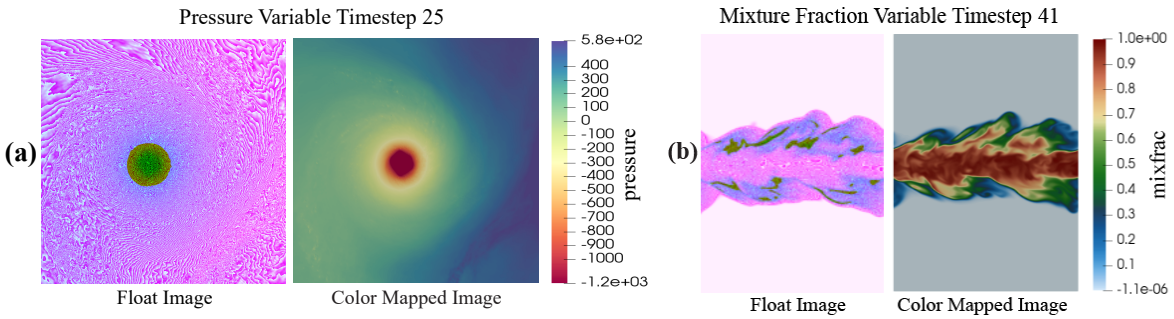


Figure 4.2: Visualization of float and colored images. 4.2(a) presents float image and the corresponding colored image using the colorbar shown on right for the *pressure* variable from the Hurricane Isabel dataset. 4.2(b) presents an example of the *mixture fraction* variable from the Turbulent combustion dataset.

relationship among multiple variables, we use *specific mutual information* (SMI) to provide information about a target variable based on the knowledge of a specific scalar value of another reference variable. We employ two SMI measures to explore multivariate interaction between variable pairs and use the SMI values to design opacity mapping for the images to highlight statistically salient regions automatically. A workflow of the proposed framework is presented in Figure 4.1.

4.5.2 Information-Driven Framework For Multivariate Feature Exploration

Cinema Database and Image Format

To generate the Cinema database images, 2D slice rendering is applied to the 3D scalar valued variables. Instead of applying a transfer function via a colormap and storing the RGB valued images, we use perspective projection on the 2D slice of the 3D data so that each pixel stores the corresponding value of scalar data [16]. Such images are called *float images* and are stored using standard PNG format. This also allows us to compute the SMI measures directly using the raw data values rather than data distorted by an underlying colormap. A colormap can then be applied post hoc. In Figure 4.2, we show examples of the float images and corresponding

color mapped images that are used in this work.

Specific Mutual Information Measures

The key factor in this work is determining the degree of association among the different variables in order to identify and highlight salient regions. Because scientific data often has nonlinear dependencies between variables, any correlation analysis technique must handle nonlinear cases. There are several correlation analysis techniques available for measuring variable relationship. Mutual Information (MI) is one of the well-known measures to quantify the mutual correlation between two variables. MI's ability to capture nonlinear dependency between variables makes it a better choice than a more typical approach such as Pearson's correlation. Mutual information quantifies the total amount of information overlap between two variables, i.e., if we observe a certain variable, then MI tells us how much uncertainty has been reduced regarding the information of another variable. Given two random variables X and Y, MI $I(X,Y)$ is formally defined as:

$$\mathrm{I}(X,Y) = \sum_{y \in Y} \sum_{x \in X} \mathrm{p}(x,y) \log \frac{\mathrm{p}(x,y)}{\mathrm{p(x)\,p(y)}} \tag{4.1}$$

where $\mathrm{p}(x)$ and $\mathrm{p}(y)$ are the probabilities of occurrence of values x for X and y for Y respectively and $\mathrm{p}(x,y)$ is the joint probability of occurrence of values x and y together.

MI quantifies the total association or disassociation between two variables and provides a single value. Since we aim to extract salient regions, we need a measure that can provide us with information related to individual scalar values. Traditional MI can be further decomposed into specific mutual information (SMI) measures to quantify individual data values' contribution towards such association or disassociation. For specific scalar values $x \in X$, SMI computes the information content of x when another variable Y is observed. In this case, X is called the reference variable and Y is called the target variable. Knowledge about the scalar values in the reference variable can increase knowledge about the target variable. This increase in information or decrease in uncertainty helps in identifying important regions in the float-image data. MI can be decomposed in multiple ways to obtain several SMI measures and we focus

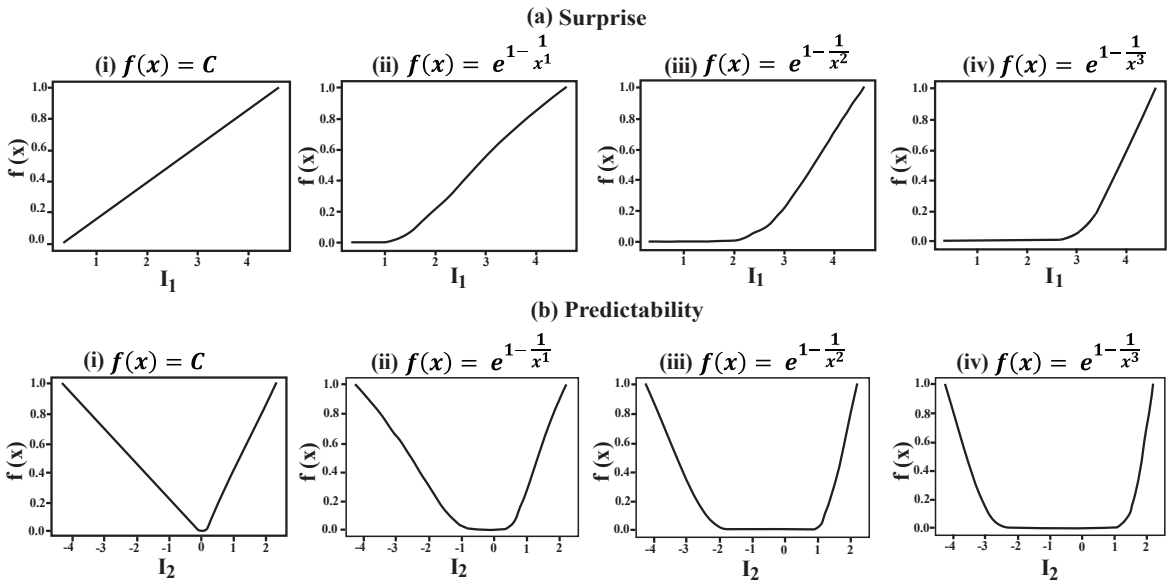


Figure 4.3: Function plots of the opacity mapping for modulating transparency in the images. Upper row 4.3(a), presents plots from SMI measure surprise (I_1) and lower row 4.3(b), presents plots from SMI measure predictability (I_2). Column (i) represents linear mapping and columns (ii), (iii) and (iv) represent increasing order of nonlinear mapping. x-axis of the plots shows the values from the SMI measure and y-axis shows the mapped values from the corresponding functions.

on two such SMI measures, Surprise and Predictability, [26, 55] for finding different types of multivariate characteristics between variable pairs.

SMI measure Surprise: $I_1(x;Y)$

The *Surprise* measure quantifies the change in the information content in the occurrences of the target variable after observing individual scalar values of the reference variable. This measure has the potential of providing information which would seem improbable otherwise, hence the name surprise [26, 55]. The regions where data values have higher surprise values can be informative. For two random variables X and Y, surprise is denoted as I_1 and presented as:

$$\mathrm{I}_1(x;Y) = \sum_{y \in Y} \mathrm{p}(y|x) \log \frac{\mathrm{p}(y|x)}{\mathrm{p}(\mathrm{y})} \tag{4.2}$$

where $x \in X$ is the reference variable and $y \in Y$ is the target variable. $\mathrm{p}(y)$ is the probabilities of occurrence of values y for Y and $\mathrm{p}(y|x)$ is the conditional probabilities of values y given values x. Surprise is always positive as it is the distance between $\mathrm{p}(y|x)$ and $\mathrm{p}(y)$. A high $I_1(x;Y)$ implies that after observing the reference variable x, some low probability values of $y \in Y$ have become more probable. This surprising element is potentially informative for our analysis.

SMI measure Predictability: $I_2(x;Y)$

The *Predictability* measure provides us with the amount of increase/decrease in uncertainty about the target variable after observing the reference variable [26,55]. This quantification of the uncertainty change helps to identify statistically significant regions in the images. Predictability is denoted as I_2 and can be computed as:

$$\mathrm{I}_2(x;Y) = -\sum_{y \in Y} p(y) \log p(y) + \sum_{y \in Y} p(y|x) \log p(y|x) \tag{4.3}$$

where $x \in X$ is the reference variable and $y \in Y$ is the target variable. $\mathrm{p}(y)$ is the probabilities of occurrence of values y for Y and $\mathrm{p}(y|x)$ is the conditional probabilities values y given values x. Based on the amount of information increase and decrease, I_2 can be both positive and negative. A high positive $I_2(x;Y)$ value indicates that the uncertainty of target variable Y has decreased when value x is observed. On the other hand, a high negative $I_2(x;Y)$ value indicates that the uncertainty of target variable Y has actually increased. According to information theory, data values that are less probable or unpredictable contain more information representing salient regions in the data with diverse characteristics that are worth deeper exploration. Therefore, the surprise and predictability measures provide different statistically meaningful results, an important consideration in the workflow.

SMI-driven Opacity Mapping Functions

These two SMI measures can now be applied to the image data to identify and highlight statistically salient regions. Since each pixel in the data has a scalar value, SMI measures can be estimated at every spatial pixel location. Note that high surprise regions and high / low predictable regions indicate salient variable relationships. We want to emphasize such regions where statistically significant multivariate properties exist between the selected variable pair. One of the ways to highlight the regions is by modulating the opacity channel of the image. This suppresses unimportant pixel values while directing focus to important regions. In the following, we show how different types of opacity mapping functions for SMI values can be used to automatically highlight important regions in the images. The design goal of such opacity functions is to make the regions containing high SMI values more opaque so that they are clearly visible and suppress regions with low SMI values by making them transparent. The choice of opacity mapping functions is quite broad and we consider linear and nonlinear mapping functions.

Linear Mapping Strategy of SMI Values

A linear mapping function can be trivially designed. We normalize the values of I_1 and I_2 in the range of $[0,1]$ using the following linear function.

$$f(x) = Constant \tag{4.4}$$

As shown in Figure 4.3, for the surprise measure, I_1, 4.3a(i) shows a linear relationship representing the I_1 values between $[0,1]$ for a pair of variables. Since predictability, I_2, produces both positive and negative values, we model them separately. We normalize positive values between $[0,1]$ and negative values between $[-1,0]$. Combining both at 0, we get a 'V-shaped' plot, as shown in Figure 4.3b(i). By designing linear mapping functions such as these, lower SMI valued or unimportant regions will be transparent and higher SMI valued or informative

regions will become opaque.

Nonlinear Mapping Strategy of SMI Values

The linear mapping strategy computes opacity value as a linear function of SMI values. However, this may not provide sufficient differentiation in the opacity to highlight the most salient regions. In order to design a mapping strategy where the higher SMI valued regions are clearly visible by further suppressing the low valued regions, we introduce nonlinear mapping functions, where the transparency value mapping can be modulated exponentially, giving us more control during analysis. We define the following nonlinear exponential function:

$$f(x) = e^{1-\frac{1}{x^a}}; a >= 1 \tag{4.5}$$

where a is the exponential control parameter. As a increases, higher SMI values are assigned higher exponential weight. a provide a control parameter that a user can use to set a threshold on the measures that are improtant for a specific analysis. Figure 4.3, columns(ii), (iii) and (iv), illustrates how the function changes with increased values of a from 1 to 3. In the case of I_1, as a increases, the plot gets steeper by assigning less weight to lower values and more weight to higher values. For example, in the case of Figure 4.3a(iv), the regions with highest I_1 values will be most opaque making anything below threshold transparent, thus highlighting the significant regions in the images.

This approach is extended for the I_2 analysis by using the function separately for positive and negative values. As seen in Figures 4.3b(ii), b(iii) and b(iv), with higher orders of a, the V-shape from the linear mapping becomes more 'U-shaped' with steepening curves emphasizing the most significant positive and negative I_2 values.

With the parameter a, the user can set the opacity threshold for results useful to their specific analysis and achieve control over the images they want to visualize for further exploration.

4.6 Results

The results of our work are presented using an interactive visual analytics tool, CinemaView, to study salient regions in image datasets. CinemaView is a browser-based viewer that allows interactive exploration of image databases stored as a Cinema database. Figures 4.4 and 4.5 show the user interface of the CinemaView tool. Figure 4.4(a) shows the color mapped ground truth images of two selected variables, pressure and cloud, followed by the images representing the analysis of the variables using surprise (I_1) and predictability (I_2) as opacity mapping functions. Images containing both linear and nonlinear mapping can be visualized simultaneously using this tool as shown in Figure 4.4(a). In this study, we present results by using order up to 3 for the nonlinear opacity mapping functions. The right panel of the CinemaView interface provides interactive widgets that can be used to adjust image size and to explore the results over time. There is a drop-down menu where the user can select the dataset to view. CinemaView is intuitive and user-friendly and it allows interactive exploration of multiple image databases simultaneously in a side-by-side fashion. Users can easily compare and contrast the relationships among multiple variables and study their evolution over time (supplementary video).

4.6.1 Hurricane Isabel Dataset

Hurricane Isabel data was produced by the Weather Research and Forecast (WRF) model, courtesy of NCAR and the U.S. National Science Foundation (NSF). This dataset consists of 13 variables and 48 timesteps with a spatial resolution of $250 \times 250 \times 50$ for a single timestep. In this work, we show analysis results obtained using the pressure and cloud variables.

Figure 4.4(a) presents analysis results for timestep 7. The pressure is the reference variable and the cloud is the target variable. Thus the specific mutual information measures are calculated for values of pressure. After computing I_1 measures, the results are stored as images for visual analysis. Since each pixel in the raw data has a pressure value and each pressure value has an

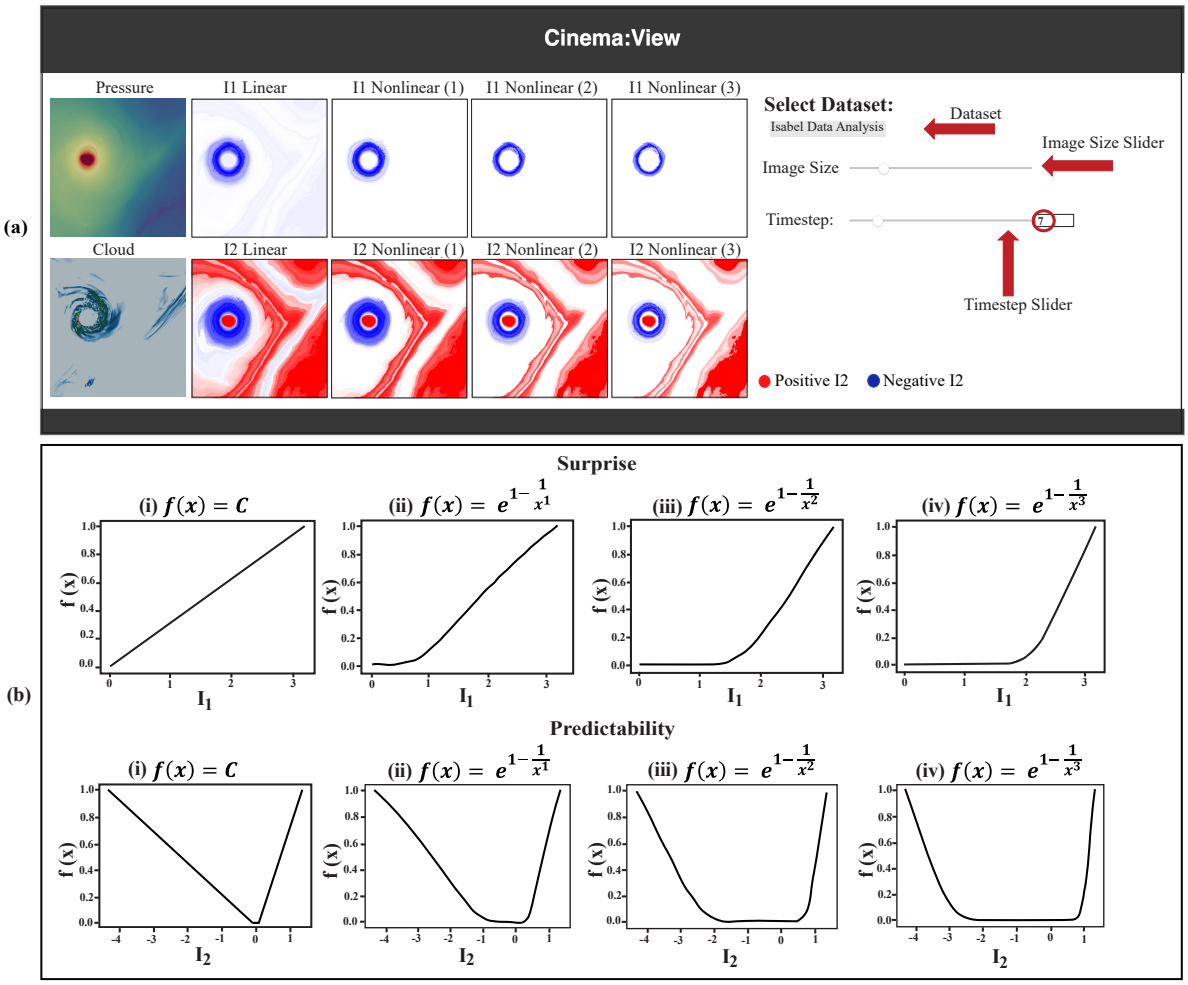


Figure 4.4: (a) presents salient regions between pressure and cloud variable analysis from the Hurricane Isabel dataset at timestep 7 using CinemaView. The first images of each row are the color mapped images of the reference variable pressure and target variable cloud. The first row shows the combined analysis using surprise (I_1) as the opacity mapping function. The blue regions represent detected salient areas. The second row shows combined analysis using predictability (I_2) for the opacity mapping function. The red regions represent positive predictability and the blue regions represent negative predictability. The elements annotated with red arrows and circles show the interactive tools of CinemaView. (b) presents function plots of the opacity mapping for modulating transparency in the corresponding images. The upper row shows surprise (I_1) plots and lower row shows predictability (I_2) plots. Column (i) represents linear mapping and columns (ii), (iii), and (iv) represent increasing order of nonlinear mapping. x-axis of the plots shows the values from the SMI measure and y-axis shows the mapped values from the corresponding functions.

associated surprise (I_1) value, we create a new image where each pixel contains the I_1 value and the opacity at each location is also controlled by a linear/nonlinear mapping function using the associated surprise values. This is then repeated for each timestep. The corresponding opacity mapping functions used to modulate the opacity for timestep 7 are shown in Figure 4.4(b), where the goal is to highlight regions that have high surprise value. As shown in Figure 4.4(b), we modulate the order of the opacity function so that we can emphasize regions with high magnitude of I_1 values.

In Figure 4.4(a), the high I_1 valued regions are presented with different shades of blue where the different shades indicate the opacity modulated regions with darker blue depicting higher surprise values. From the I_1 linear mapping results, we can observe that the areas around the hurricane eye are highlighted as having high I_1 values and indicate that such regions have become more probable after the cloud variable is observed. These regions coincide with the hurricane eyewall – a salient region in the pressure data. It is also observed that, by increasing the ordering of the nonlinear mapping, we can refine the most significant and surprising regions around the hurricane eyewall.

The second row of Figure 4.4(a) (except the first image) presents I_2 analysis results. As the I_2 values can be both positive and negative, for visualization purposes, those regions are highlighted using shades of blue and red. Blue and red indicate negative and positive I_2 values, respectively. From the I_2 analysis results, we see that the hurricane eye region is red (positive I_2) which means it is a highly predictable region when pressure and cloud variables are analyzed. It is known that in the hurricane eye region, pressure values are typically low and cloud values are mostly homogeneous and thus such region is detected as a predictable region. If we focus at the region around the hurricane eye's boundary, we find that a region is identified as uncertain and has negative predictability values and so has blue color. This is also a consistent observation since this region is known as the eyewall and the target/observed variable cloud has high variability and so is less predictable. Finally, moving away from the hurricane eyewall, the cloud values again become less varying and such regions are detected as

more predictable regions (red color) away from the hurricane eye. The white regions in these images indicate regions where both the positive and negative I_2 values are relatively low and so they are transparent. From the predictability plots in Figure 4.4 b(i), b(ii), b(iii) and b(iv), the white areas represent the parts where the 'V-shape' flattens into 'U-shape' as we increase the order of the nonlinear mapping. As the order is increased, stronger predictable and uncertain regions become highlighted as significant regions.

4.6.2 Turbulent Combustion Dataset

The Turbulent combustion simulation data is made available by Dr. Jacqueline Chen at Sandia Laboratories through the US Department of Energy's SciDAC Institute for Ultrascale Visualization. This dataset has 5 scalar variables and 122 timesteps with a spatial resolution of $240 \times 360 \times 60$ for a single timestep. During the combustion process, fuel and oxidizer react and the flame exists where fuel and oxidizer are in stoichiometric proportions [7]. The mixture fraction is an important variable in this dataset that indicates the fraction of mass at the fuel stream origin. So, we have used the mixture fraction (mixfrac) as the reference variable and hydroxyl radical (Y_OH) as the target variable since both of these can be used to study the flame regions of the simulation [7]. By analyzing the interacting relationship of these two variables, important features can be studied and detailed information about the combustion process can be gleaned.

In Figure 4.5, we show results from timesteps 5, 41, and 80 as three different representative timesteps, highlighting three stages of the time-varying simulation. Timestep 5 in Figure 4.5(a) shows the initial state of the combustion variables interacting when the flames just started burning. Timestep 41 in Figure 4.5(b) represents an intermediate time when the combustion process is active and and finally, Figure 4.5(c) presents the result from a later timestep 80 when the flame has expanded. From these three figures, the salient regions clearly change their shape and position over time, indicating how this method is able to capture temporal changes.

The salient regions detected from the I_1 analysis signifies the areas where the combustion

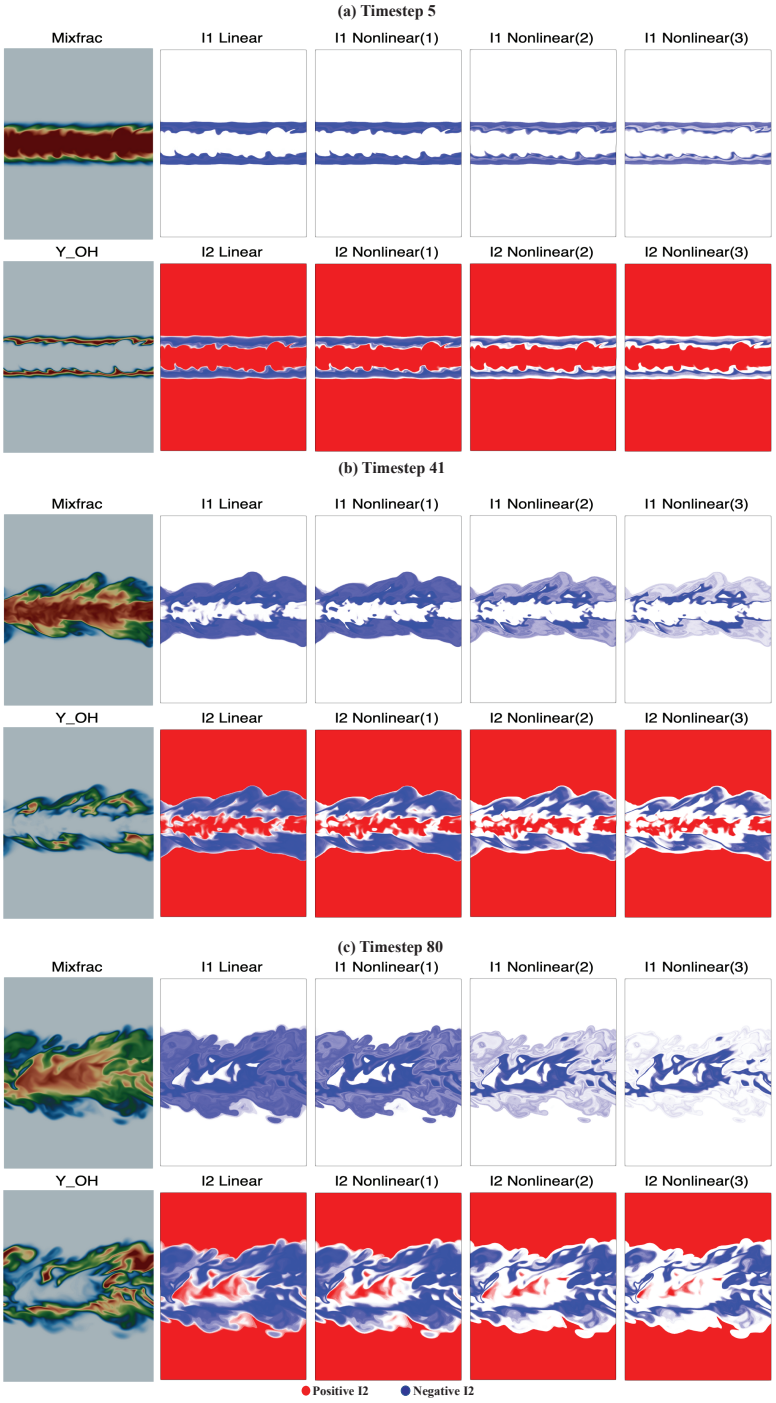


Figure 4.5: Salient regions between reference mixfrac and target Y_OH variable analysis from Turbulent combustion dataset at (a) timestep 5, (b) timestep 41 and (c) timestep 80. The first images of each row are the color mapped images of the reference variable mixfrac and target variable Y_OH. After the color mapped image, the top row from every timestep shows combined analysis of the variables using surprise (I_1). The blue regions represent the salient areas (flames). Similarly, the bottom row shows analysis using predictability (I_2). Red and blue regions represent positive and negative predictability respectively.

process is happening around the flames. I_1 analysis shows blue regions identifying the areas with combustion flames. As we proceed to nonlinear mapping with increased order, higher I_1 valued regions get highlighted with dark blue and lower I_1 valued regions become transparent with lighter shades of blue, displaying the flame regions in a more refined manner.

From the I_2 analysis results, we see two types of regions, blue and red. As before, the blue regions show the locations where the values of the target/observed variable (Y_OH) are not homogeneous when observing the reference variable mixfrac. From all of the three timesteps, we find that the blue regions coincide well with the regions detected by the I_1 analysis, i.e., the regions where the flame is. In this region, the complex chemical reactions take place and so is hard to predict. From our I_2 analysis, such regions are detected as having negative I_2 values which means such regions have higher uncertainty, therefore, less predictable. On the other hand, the red regions in these results show predictable regions of Y_OH when mixfrac is observed. The two outer red regions (the top and the bottom part) are the background regions where the combustion is not happening and hence the data values are mostly homogeneous. As a result, such regions are correctly identified as the highly predictable regions. The red regions in between two blue uncertain regions indicate that at the center of the simulation, there are some places where the variable Y_OH is more predictable and hence has positive I_2 values. It is also observed that as we increase the order of our opacity mapping function for both linear and nonlinear approaches, we can obtain further refined views of these predictable and uncertain regions where the darker (more opaque) regions indicate locations with higher magnitude of I_2 values. From these analysis results, we observe that both I_1 and I_2 analysis on the Turbulent combustion dataset bring out salient regions that the user can further study in more detail for exploring important characteristics of these variables over space and time.

4.7 Conclusions and Future Work

Our work successfully enables scientists to explore and extract salient regions in time-varying multivariate data sets. This technique is generalizable and is not limited to the data sets analyzed in this work. In future work, we plan to accelerate the computation of the information measures by using GPU-based parallel computing. The computation for each timestep can be further parallelized since the computation at each timestep is independent. We also plan to design more sophisticated optimization functions for opacity mapping. Instead of generating different orders for opacity modulation, an optimization-based approach could generate regions that are most useful to the domain scientists.

4.8 Funding

This work was supported by the U.S. Department of Energy through the Los Alamos National Laboratory. Los Alamos National Laboratory is operated by Triad National Security, LLC, for the National Nuclear Security Administration of U.S. Department of Energy (Contract No. 89233218CNA000001). This research is released under LA-UR-21-21824.

Chapter 5

In Situ Adaptive Spatiotemporal Data Summarization

5.1 Publication Notes

Citation: S. Dutta, H. Tasnim, T. L. Turton and J. Ahrens, "In Situ Adaptive Spatio-Temporal Data Summarization," 2021 IEEE International Conference on Big Data (Big Data), Orlando, FL, USA, 2021, pp. 315-321, doi: 10.1109/BigData52589.2021.9671581.

Publisher: 2021 IEEE International Conference on Big Data (Big Data)

Date of Conference: 15-18 December 2021

Published: 13 January 2022

Formatting: The original published text has been preserved as much as possible while still adhering to the formatting requirements of this dissertation.

Funding: This work was supported by funding from the following: 10.13039/100006168-National Nuclear Security Administration; 10.13039/100006235-Lawrence Berkeley National Laboratory

Acknowledgements: This research was supported by the Exascale Computing Project (17-SC-20-SC), a collaborative effort of the U.S. Department of Energy Office of Science and

the National Nuclear Security Administration and is released under LA-UR-21-28689 Ver. 2. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231. We thank our ECP collaborators, especially Jordan Musser and Ann S. Almgren.

5.2 Abstract

Scientists nowadays use data sets generated from large-scale scientific computational simulations to understand the intricate details of various physical phenomena. These simula- tions produce large volumes of data at a rapid pace, containing thousands of time steps so that the spatiotemporal dynamics of the modeled phenomenon and its associated features can be captured with sufficient detail. Storing all the time steps into disks to perform traditional offline analysis will soon become prohibitive as the gap between the data generation speed and disk I/O speed continues to increase. In situ analysis, i.e., in- place analysis of data when it is being produced, has emerged as a solution to this problem. In this work, we present an information-theoretic approach for in situ reduction of large- scale time-varying data sets via a combination of key and fused time steps. We show that this approach can greatly minimize the output data storage footprint while preserving the temporal evolution of data features. A detailed in situ application study is carried out to demonstrate the in situ viability of our technique for efficiently summarizing thousands of time steps generated from a large-scale real-life computational simulation code.

5.3 Introduction

With the increase in computing capabilities, large-scale scientific simulations now produce very large data sets containing thousands of time steps. These computer simulations help scientists in understanding the intricate nature of various phenomena, e.g., the evolution of hurricanes

and tornadoes, formation and dynamics of bubbles in a gas-solid mixing process, assessing the consequences of potential asteroid impacts etc. All of these phenomena are time-varying in nature and their simulations produce time-varying data sets that can take terabytes (TBs) to petabytes (PBs) of disk storage. Soon we will have exascale supercomputers [72], enabling scientists to generate exabytes (EBs) of data. Storing all such data will be prohibitive since the data generation velocity will outpace the rate at which it can be stored into persistent disks [44, 63]. The bottleneck of slow disk I/O and extreme data volume will entail novel data triage strategies that can work real-time with the simulation, i.e., *in situ*, and produce informative data summaries, significantly smaller than the raw simulation output, enabling flexible *post hoc* analysis.

Currently, to manage the output data size, simulation scientists often skip regular intervals of time steps and store every n^{th} (n typically varies between $50 \sim 100$) time step. By doing so, the scientists remain oblivious of the events that take place in those skipped time steps. A better strategy is to detect the key time steps and store only the key time steps so that the important events can be preserved. In this case, even though the key time steps are stored, a comprehensive summary of all the time steps will still be missing. Another complicating factor is that many existing key time step detection techniques for scientific data sets assume the availability of all the time steps [187, 199]. For an *in situ* approach, where data becomes available in a streaming fashion, one time step at a time, such algorithms (a) may not be readily applicable, (b) could be computationally expensive. In recent years, researchers have focused on developing *in situ* techniques that allow identification of important time points during the simulation [110, 142, 165]. However, such techniques typically do not offer any integrated data summarization strategy. Therefore, new automatic time-varying data summarization techniques are needed that will work *in situ* and scale with the data generation velocity while producing informative and comprehensive data summaries with minimal storage footprints.

In this work, we propose a spatiotemporal data summarization technique that uses information-theoretic measures to quantify data value importance between consecutive time steps and

summarizes data from a sequence of time steps into a single fused data set. As the simulation runs for long hours in supercomputers to produce scientifically meaningful data, the proposed technique analyzes data from thousands of time steps *in situ*, i.e., when the data is being generated, identifies key time steps based on an user provided criterion, and summarizes the data between every two consecutive key time steps into a single summarized data set that captures a comprehensive view of the features for the time window. Our work can leverage the existing *in situ* key time step detection approaches [110, 142, 165] and produce data summaries for the intermediate time steps. The proposed method stores raw simulation data for each key time step along with time-varying data summaries for time steps between every two key time steps. We show that the output data size for our method is significantly smaller compared to the raw simulation data size and that the summary data can be effectively analyzed and visualized interactively during *post hoc* exploration. To show the efficacy of the proposed technique, we apply our method to several time-varying data sets and conduct a detailed *in situ* application study with a large-scale simulation to demonstrate the *in situ* applicability and performance of our technique. Therefore, our contributions to this work are twofold:

- We propose an information-theoretic adaptive spatiotemporal data summarization technique for large-scale time-varying data sets that produces summary data as a combination of key and fused time steps to preserve (a) the important events and (b) a comprehensive view of the whole simulation data.
- We study the effectiveness of the proposed algorithm *in situ* with a large-scale simulation and demonstrate its practical applicability and *in situ* viability.

5.4 Related Works

5.4.1 In Situ Analysis

With modern supercomputers producing large-scale data sets, in situ analysis has emerged as a promising solution and several in situ analysis frameworks such as Ascent [109], ParaView Catalyst [75], and VisIt libSIM [207] have been developed. Further, a significant amount of research has been done to develop data reduction techniques for producing reduced data summaries that can be stored and used as a proxy for the raw data. Cinema [4] is such an in situ image-based data reduction and visualization approach. Among other in situ techniques, compression [111, 116, 118], sub-sampling [21, 202, 210], and distribution-based summaries [63, 64, 214] are popular. In this work, we advocate a hybrid approach where we store the raw data for important key time steps and summarize the intermediate time steps to achieve sufficient data reduction.

Detection of key time points in a data set is an important problem for time-varying data analysis. Several approaches have been proposed for key time step detection for large time-varying data sets [187, 220]. These techniques generally allow the detection of key time points and do not offer any data summarization capability. The computer vision community has developed several techniques for doing spatiotemporal fusion of large data obtained from different sources. Pulong and Kang proposed a technique for data fusion [124]. Nguyen et al. [145] developed a technique for summarizing large spatio- temporal images. In a recent work, Shah et al. [171] proposed an algorithm for real-time summarization of data streams for smart grid applications.

The use of information-theoretic measures [49, 191] to solve data analysis and visualization problems is well-known. Mutual information has been used to perform data registration [47, 92, 94, 126, 152], view selection [194], and for quantifying information transfer from data to image space [28]. For exploring similarities among level-sets, information theory has also

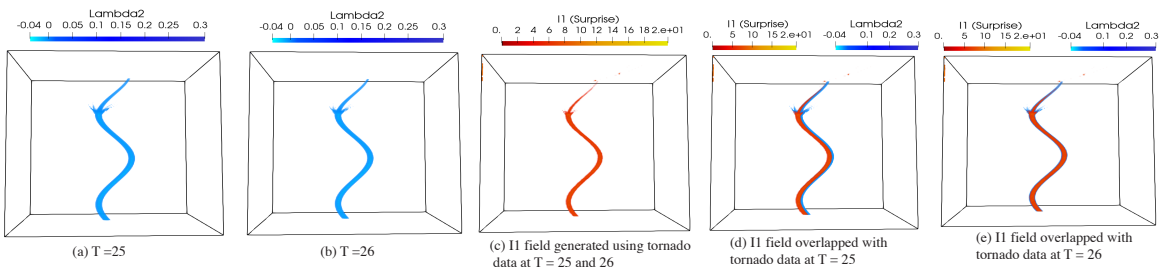
been used [32, 203].Various decomposition of mutual information, called specific mutual information and pointwise mutual information measures have become recently popular for fusing multi-modal data [27] and multivariate sampling [66] for data reduction. For a more detailed review of information theory applications in data analysis and visualization, interested readers are referred to [42, 43, 166, 198].

5.5 Methods

In this work, we propose a new technique for summarizing a sequence of time-varying scalar fields into a single scalar field that captures the dynamic temporal evolution of the data features. The users can study the summary fields to obtain a comprehensive view of the time-varying nature of the features without needing to go over each time step individually. This approach achieves significant data reduction for the *post hoc* analysis while preserving the important feature dynamics of a sequence of time steps so that analysis time is reduced and scientific discovery is accelerated. In the following section, we first introduce the concepts of the information theory measure that we use to quantify *informativeness* of specific data values over time and then present the technique for producing time-varying data summaries for a sequence of time steps. Note that we develop this algorithm for *in situ* use cases, where we run our algorithm online when the simulation is running and access the time step data one-by-one in a streaming fashion as they are produced.

5.5.1 Data Value Informativeness Quantification

Since the goal is to combine data from a sequence of time steps, it is important to quantify the informativeness of each data point so that we can prioritize one data point over others during the summarization process. In information theory [49], mutual information (MI) is a well-known measure that estimates the amount of information overlap between two random variables and



(a) T =25 (b) T =26 (c) I1 field generated using tornado data at T = 25 and 26 (d) I1 field overlapped with tornado data at T = 25 (e) I1 field overlapped with tornado data at T = 26

Figure 5.1: Visualization of $I_1 field$ generated using two consecutive time steps of the analytical Tornado data set. Volume rendering technique is used to generate the visualization results. (a) and (b) show the vortex region of the Tornado data and (c) shows the corresponding $I_1 field$. In this illustrative example, data from T=25 is observed and so the high I_1 valued region overlap accurately with the vortex region at T=26 as shown in (e).

can be formally computed following Equation 5.1:

$$I(Y;X) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \tag{5.1}$$

In Equation 5.1, $I(Y;X)$ is the MI between two random variables Y and X, $y \in Y$ represents a specific value of Y and $x \in X$ is a value of X. The joint probability between x and y is written as $p(x,y)$ and the marginal probabilities of x and y are $p(x)$ and $p(y)$ respectively. MI for two random variables computes to a single number reflecting the total shared information between X and Y. Since we need information content of each data value so that we can perform spatiotemporal data summarization, we focus on a decomposition of MI that can estimate the information content of each data value of one variable, while observing values from another variable. Such decomposition of MI is called *specific information* measures.

Specific information measure was first introduced by DeWeese and Meister [56] and can be formally derived from Equation 5.1 as shown in Equation 5.2 and 5.3. The specific information, called *surprise*, denoted as $I_1(y;X)$ in Equation 5.3, represents the informativeness of a data value y when the whole variable X is observed. Here, $p(x|y)$ represents the conditional

probability of value x given y.

$$\begin{aligned} I(Y;X) &= \sum_{y \in Y} p(x) \sum_{x \in X} p(x|y) \log \frac{p(x|y)}{p(x)} \\ &= \sum_{x \in X} p(x) I_1(y;X), \end{aligned} \tag{5.2}$$

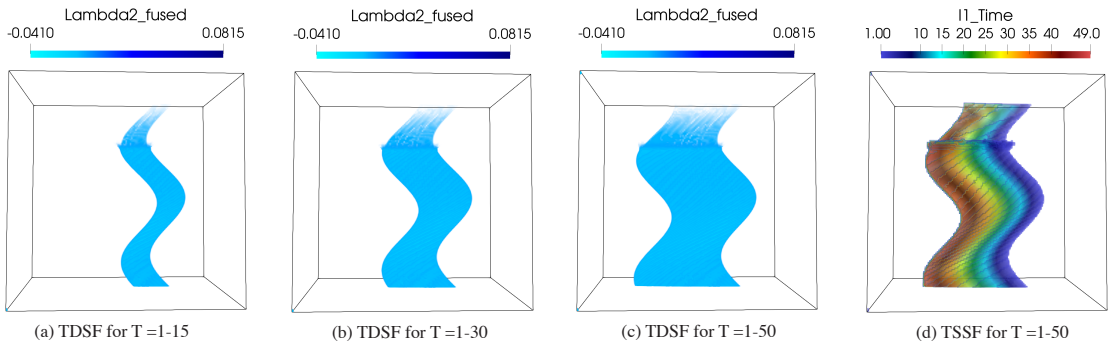
$$I_1(y;X) = \sum_{x \in X} p(x|y) \log \frac{p(x|y)}{p(x)} \tag{5.3}$$

For a data value y, a high value of $I_1(y;X)$ indicates that some infrequent occurrences of $x \in X$ have become more probable after observing the value y from Y, amounting to a surprising result, hence the name surprise. The value of surprise ($I_1(y;X)$) is always positive, i.e., $I_1(y;X) \geq 0 \ \forall \ y \in Y$ since it represents the KL-divergence between the distributions $p(X)$ and $p(X|y)$ [56].

In our work, we use surprise as the measure to estimate the informativeness of a data value when data values from another time step are observed. More specifically, if we assume that X and Y represent the same data variable from time step t and $t+1$, then we can estimate the informativeness of each data value at time step $t+1$ as $I_1(y;X)$, by observing the same variable from the previous time step t. This gives us a way of finding the highly surprising regions in the data when we compare it with a previous time step. These surprising regions (i.e., regions with high $I_1(y;X)$ values) can indicate the regions where the data features exist and need to be captured in the temporal summary field.

5.5.2 Information Fields

Our primary target application is time-varying 3-D scalar fields with the goal to summarize a sequence of 3-D scalar fields into a single scalar field that can provide a comprehensive summary of the data features for the selected time sequence. Such summaries can indicate how the data features of interest have evolved within the time window and can also reveal their tracking information. Equation 5.3 shows how surprise can be estimated for each data value in variable Y. In practice, computation of such information theory measures is done by first establishing a



(a) TDSF for T =1-15 (b) TDSF for T =1-30 (c) TDSF for T =1-50 (d) TSSF for T =1-50

Figure 5.2: Demonstration of the proposed spatiotemporal data summarization scheme using a sequence of time steps from the Tornado data set. (a), (b), and (c) show the TDSFs of the Tornado data when time steps between 1-15, 1-30, and 1-50 have been summarized using the proposed algorithm. In (d) we present the TSSF for the Tornado data corresponding to the TDSF shown in (c). The colors in (d) shows the temporal evolution of the vortex region over the time window and how it moves gradually from right to left.

communication channel $Y \rightarrow X$ between the variables X and Y as discussed in [27] and then computing the surprise using the communication channel. Normalized histograms can be used to estimate probability distributions while computing the values of $I_1(y;X)$. After the surprise (I_1) values are computed, we create a new scalar field where at each spatial grid point (with data value $y \in Y$), we put the corresponding value of $I_1(y;X)$. Since such a scalar field contains information values at each grid point, it can be called an *information field* or $I_1 field$. The $I_1 field$ computed between two time steps can be visualized directly and regions with high I_1 values can be explored as salient regions.

Figure 5.1 shows an example of an $I_1 field$ constructed using two time steps of an analytical Tornado data set. This data set of dimension $128 \times 128 \times 128$, contains velocity vectors and is generated by an analytical function [52]. The data set has 50 time steps and simulates a tornado-like vortex structure. For this study, we have modified the analytical equation so that the center of the tornado changes position with time, creating a moving vortex in the spatial domain. Tracking and visualizing this vortex is of interest in this data. To detect the vortex region, we have used the lambda2 (λ_2) vortex criterion [96]. The visualizations shown here are generated using the Ray-casting-based Volume Rendering technique [60] from ParaView [11]

that allows interactive visualization of 3-D scalar field data sets. Figure 5.1(a) and 5.1(b) show the vortex at T=25 and T=26 respectively. Even though they look very similar, the vortex at T=26 has moved slightly toward the left from its position at T=25. Figure 5.1(c), presents the $I_1 field$ computed at T=26 when the data at T=25 is observed. We refer to the time step that is the observed variable as the reference time step. We find that the $I_1 field$ at T=26 captures the location of the vortex region accurately. In Figure 5.1(d) and 5.1(e), we superimpose the estimated $I_1 field$ with the λ_2 vortex fields from T=25 and T=26 respectively. Figure 5.1(d) shows that the $I_1 field$ at T=26 captures the slight shift on the vortex structure and only partially overlaps with the vortex at T=25, whereas, in Figure 5.1(e), a complete overlap of the $I_1 field$ with the underlying vortex is seen at T=26.

5.5.3 Time-varying Feature-based Data Summarization using Information Fields

The insights obtained from Figure 5.1 allow us to develop the idea of time-varying data summarization using $I_1 field$s from a sequence of consecutive time steps. One can imagine that if we compute the $I_1 field$s for every consecutive pair of time steps, each $I_1 field$ will assign high values to the statistically salient regions of the data. Then, if we create a new fused summary field where at each spatial location, we assign the data value from the time step where the I_1 value is the highest over the chosen time window, we can combine all the high I_1 valued regions from a time window into a single field. Hence, for each spatial location p, the assigned value is calculated as:

$$Val(p) = max(I_1^t(p)), \forall t = t_{start}, .., t_{end} \tag{5.4}$$

where t_{start} and t_{end} represent start and end time steps, $I_1^t(p)$ is the value of I_1 at point p in time t. Conceptually, this technique will maximize the spatiotemporal information in the combined field by selecting data points that have maximum I_1 values over the time window. This combined field will capture the time-varying nature of the data by focusing on the salient regions with

high I_1 values.

Since domain scientists primarily want to study the important features in their data, we devise our summarization strategy for the feature regions when a domain-specific feature descriptor is available. This methodology allows the user to provide a feature descriptor, such as a threshold, and while performing the temporal summary, we check if the current data point is a feature and then only summarize such points. For all the non-feature points in the data, we assign a constant value to them so that when the summary fields are analyzed and visualized, the non-feature points can be emphasized less using volume rendering techniques so that the users can focus on the evolution of the features without any occlusion from non-featured regions.

Figure 5.2 demonstrates this spatiotemporal data summarization scheme using the analytical Tornado data. Figure 5.2(a), 5.2(b), and 5.2(c) show the volume rendering of the summary fields when 15, 30, and 50 time steps of Tornado data are summarized into a single field. These summary fields are denoted as the *temporal data summary field* (TDSF). It is seen that these TDSFs can capture the evolution of the vortex in Tornado data as the vortex moves from right to left. To capture how the TDSFs are generated and associate each part of the TDSF with its relevant time step, we also generate another field, the *time step summary field* (TSSF). For each spatial location, the TSSF assigns the time step number from which the data (with the highest I_1 value) is selected. Figure 5.2(d) shows the TSSF for Tornado data that corresponds to TDSF at Figure 5.2(c). The colors in Figure 5.2(d) reflect the time steps and, using a colormap that naturally delineates bands, we can see that the vortex moves from right to left over time as the color changes from blue to red.

By exploring the TDSF and TSSF together, users can get a comprehensive view of the evolution of the vortex in the Tornado data without needing to inspect each time step individually. Disk storage can be significantly reduced by retaining the Tornado data at T=1 (initial time step) and T=50 (final time step), while keeping the TDSF and TSSF fields as a replacement for all the 48 intermediate time steps. We observe that the storage for the raw Tornado data is 489MB, whereas the proposed technique will only take 40MB disk space, achieving approximately 92%

storage reduction. Using this technique, we can generate temporal summary fields (TDSFs and TSSFs) for sequences of time steps, retaining raw simulation data for the start and end time steps of each sequence along with the corresponding TDSFs and TSSFs for achieving sufficient data reduction.

5.6 In Situ Application Study

5.6.1 Application Background

In this section, we apply our algorithm *in situ* to a data set generated from a large-scale computational fluid dynamics–discrete element model (CFD-DEM) code, MFIX-Exa [1, 140], which is being developed at the National Energy Technology Laboratory (NETL) to study multiphase flows. MFIX-Exa generates particle-based data to study the working principles of chemical looping reactors (CLR). Such reactors contain fluidized beds where particles interact and, under certain physical conditions, bubbles (void regions) are formed as shown in the left image of Figure 5.3.. The study of the dynamics and interaction of such bubbles is critical since the formation of large, fast-moving bubbles in fluidized beds can cause poor gas/solid mixing, lowering the conversion efficiency and stability of the reactor.

Data produced from a single MFIX-Exa run can contain tens of millions of particles per time step and can have thousands of time steps, needing terabytes to petabytes of storage. A full-scale simulation of MFIX-Exa is set to achieve exascale performance [73, 140] in the upcoming years as part of US DOE's Exascale Computing Project (ECP) [72]. As a consequence, storing all the raw particle data for a *post hoc* analysis will be prohibitive. Therefore, new algorithms are required that can detect the bubbles *in situ* and summarize their temporal dynamics so that the output data size is significantly reduced and scalable bubble dynamics analysis will be possible. To address this need, we have deployed our proposed algorithm *in situ*, i.e., as the data is being produced, and generate bubble-based summarization fields so that the raw particle data are not required to be stored at each time step, thereby significantly reducing the overall storage needs.

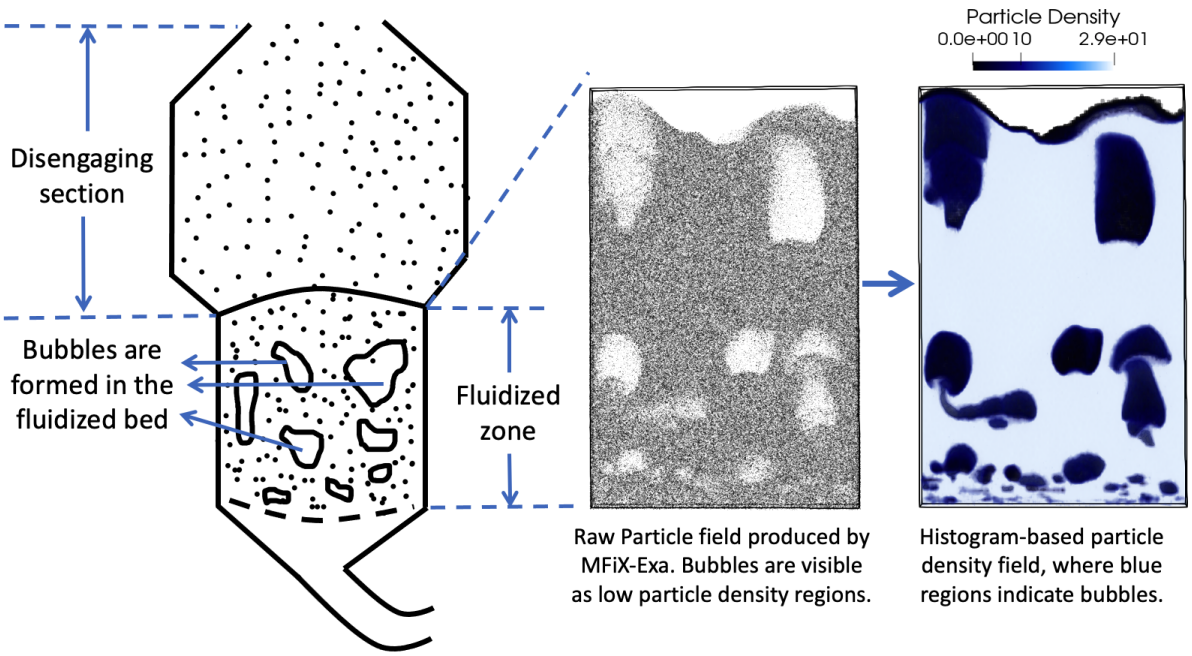


Figure 5.3: The left image shows a schematic diagram of a CLR and the fluidized bed region is highlighted where bubbles are formed. The middle image shows the raw particle visualization from a time step and the empty low particle density regions can be observed. The right image shows the estimated particle density scalar field for this data where the bubble regions are seen as low-density regions (dark blue regions).

To perform *in situ* analysis using MFIX-Exa, custom code is added to both MFIX-Exa and AMReX code bases. Our *in situ* code is also developed in C++ and uses the VTK [169] library for data processing. We have developed an *in situ* adapter function that directly accesses the raw particle data from AMReX particle containers and converts it to a VTK data set which is used in our algorithm. As the simulation code and our *in situ* algorithm run on the same memory and computing resources, this *in situ* integration works in synchronous mode, tightly coupled with the simulation code.

5.6.2 In Situ Algorithm for Streaming Environment

Since MFIX-Exa produces unstructured particle fields, we first convert it to a scalar particle density field. To estimate the particle density, we create a spatial 3-D histogram using particle locations. Note that the particles are distributed into multiple computing nodes and so first

we create a local partial histogram using global particle bounds at each MPI process and then merge all the local histograms to construct the global histogram using MPI reduction. As the bin frequencies of this 3-D histogram reflect the number of particles in a local region of the domain, we convert this 3-D spatial histogram into regularly structured grid data where the number of histogram bins translates to the spatial dimensions of the structured grid and the bin frequency values are interpreted as particle density at each grid point. An example of this histogram-based density field is shown in Figure 5.3. The center image shows the raw particle field where the void regions are the bubbles. The right image shows the structured particle density scalar field estimated using the spatial histogram. The dark blue regions in this image show regions that correspond to low particle density regions and are considered bubbles.

A threshold on these particle density fields can be used to segment the bubbles. These bubbles follow complex time-varying dynamics where they are formed at the bottom of the fluidized bed, and over time rise and merge with other bubbles or split into multiple bubbles before reaching the top boundary. The MFIX-Exa domain scientists want to understand these complex bubble interactions while evaluating their computational model. The interesting time points for this simulation are when relatively larger bubbles undergo a merge/split event. However, since the simulation data gradually evolves over time and such events do not happen at each time step, this is an ideal use case for our approach. In this case, the sequence of time steps between merge/split events can be summarized into a fused field. To preserve the raw particle data at the key time steps when a merge/split event happens, we first segment the density field and count the number of segments where each segment indicates a bubble. For the next time step, if the number of segments remains the same as the previous time step – indicating no merge/split has happened – we apply our summarization algorithm to fuse all such intermediate time steps. When the count of the bubbles changes, the algorithm outputs the summarized TDSF and TSSF at that time step and also stores the raw particle data, re-initializes the TDSF and TSSF, and continues the process from the next time step.

The algorithm uses a threshold value to segment and detect the bubble regions (regions

$\leq$ TH) while generating the summarized fields. In the *in situ* environment, we only have access to one time step at a time, requiring modifications to the methodology. Since the size of the estimated particle density field is quite small, we keep the particle density field from the previous time step in memory. The joint histogram computation needed to compute the surprise (I_1) values requires two sequential time steps. We also initialize TDSF and TTSF as global data objects. At each new time step, for every spatial location, if the value of I_1 is higher than the current value, we update the data value at that location with the data value from the current time step and also update the time step number with the current time step number for the same spatial location in the TSSF. This process incrementally constructs the TDSF and TSSF for a sequence of time steps in the *in situ* setting. Once the bubble count changes, we output the current TDFS, TSSF, and the particle raw data and reinitialize the TDSF and TSSF using the values from the current time step. This algorithm can run continuously with the simulation and produce TDSFs and TSSFs for sequences of time steps when the bubble count remains the same. Hence, this method adaptively stores key time steps from the MFIX-Exa simulation and summarizes the intermediate time steps, achieving a significant data reduction while preserving the details of the bubble dynamics.

5.6.3 Analysis Results

We have tested the effectiveness of our method by running it *in situ* with the MFIX-Exa simulation. The simulation test case represents a scenario where a constant density, constant viscosity gas is used to fluidize spherical particles of uniform radius. The fluidized bed has a constant velocity gas inlet at the bottom of the bed and the simulation contains $\approx$ 4.1 million particles. As the simulation progresses and reaches a steady-state, bubbles start to form inside the fluidized bed. We have run this simulation for 6000 time steps starting from a previously stored checkpoint file at T=25000 to reach the point when bubbles are already forming.

In Figure 5.4, we show the results of *in situ* data summarization for one of the time windows, with start time step 25090 and end time step 25340. Figure 5.4(a) and 5.4(c) show the estimated

	No. of processors	Simulation (mins)	In situ processing (mins)	Simulation raw I/O (mins)	In situ I/O (mins)
MFIX-Exa Case (∼4.1M particles, 6000 time steps)	1024	553.05	24.37	72.7	2.26

Table 5.1: Computational performance for the in situ application study using MFIX-Exa simulation.

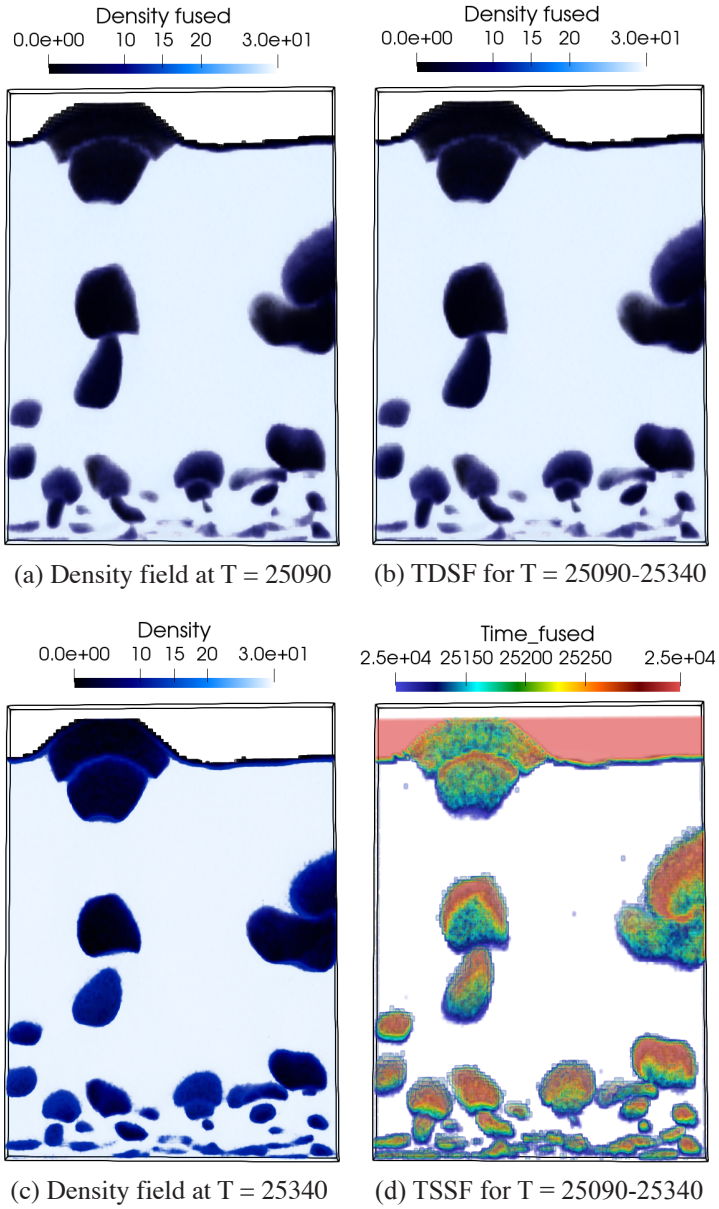
density fields at T=25090 and T=25340 respectively. We observe that the bubbles (dark regions with low particle density) have evolved and have moved upward. To focus the analysis only on the bubble regions, we have used the density threshold=12 for segmenting the bubbles. Also, since the state of the bubbles changes very slowly between consecutive time steps, we call the *in situ* routine at every 10^{th} time step. The *in situ* processing frequency is an input parameter and the users can set it to the desired value based on how frequently *in situ* processing is needed. Furthermore, since the domain experts are more interested in the evolution of larger bubbles, in this study, we only count the number of bubbles containing more than 750 voxels. In Figure 5.4(b) and Figure 5.4(d), we present the TDSF and TSSF for this time window. These fields highlight the evolution of the bubbles for the intermediate time steps. Note that at T=25340, two bubbles merge (the bubbles at the center-left of Figure 5.4(a)) and as a result, the number of bubbles changes. To preserve this time step as one of the key time steps, our technique outputs the raw particle data along with the summarized TDSF and TSSF for the time window T=25090-25340. For the entire *in situ* run of 6000 time steps, our method identified 54 key time points, summarizing the intermediate data for each pair of consecutive key time steps between key time points. These results demonstrate the usefulness of the proposed method for analyzing and summarizing large-data sets *in situ* where we can access the simulation data at a much higher temporal frequency, bypassing the expensive disk I/O, which would be prohibitive for traditional *post hoc* analysis.

5.6.4 Storage Savings and Computational Performance

The *in situ* studies are done using the supercomputer Cori at the National Energy Research Scientific Computing Center (NERSC). NERSC is one of the primary high-performance scientific computing facilities for the Office of Science in the U.S. Department of Energy (DOE). Cori is a Cray *XC*40 system, capable of achieving a peak performance of about 30 petaflops.

For these studies, the raw simulation particle data is stored using PLOTFILE format and contains particle ids, particle locations, and their velocities. We ran 6000 time steps of the simulation. The proposed method stored 54 key time steps with the TDSFs and TSSFs. The spatial dimension of the generated TDSFs and TSSFs are $128 \times 16 \times 128$ and are stored in VTK format. We find that the proposed method needs 16.03 GB storage, while if we store all the raw data for every 10th time step, then we would require 170 GB storage. Hence the proposed method is able to reduce $\approx 91\%$ disk storage.

In Table 5.1, we provide the *in situ* computational performance of our technique to demonstrate its *in situ* viability. Typically, when an *in situ* analysis is performed with a simulation, it is desirable that the *in situ* processing will take only a small fraction of the simulation time. Our study is run using 1024 processors and it is observed that the in situ processing time is significantly smaller compared to the simulation time. Also, from the fifth and sixth column of Table 5.1, we observe that the in situ I/O, which includes timings for storing the raw data for key time steps and the TDSFs and TSSFs, is significantly smaller compared to the raw data I/O if we store the particle data at every 10th time step to conduct similar analysis offline. In addition, we also measure the timings if our algorithm is executed post hoc and found that the post hoc disk I/O takes 246.27 minutes, which is significantly higher compared to the *in situ* I/O. However, by processing the data *in situ*, we are able to bypass this slow *post hoc* I/O. Therefore, by performing *in situ* analysis, the proposed method saves both storage and computational time while enabling flexible *post hoc* analysis.



(a) Density field at T = 25090

(b) TDSF for T = 25090-25340

(c) Density field at T = 25340

(d) TSSF for T = 25090-25340

Figure 5.4: *In situ* application study results of the proposed method when run with the MFIX-Exa simulation. (a) and (c) show the particle density field from T=25090 and 25340 respectively. The bubble features are observed with dark blue regions. The TDSF generated for the intermediate time steps (T=25090-25340) is provided in (b) and the corresponding TSSF is shown in (d). We find that the TSSF is able to provide a comprehensive summary of all the bubbles within this time window and as two bubbles (two bubbles at the center left of (a) merge at T=25340, the bubble count changes at T=25340 and the proposed technique outputs summary results.

5.7 Conclusion

In conclusion, we have presented an *in situ* technique for summarizing large-scale spatiotemporal data sets to reduce the size of the output data significantly while preserving the important state of the features. The proposed method detects key time steps based on a suitable user-provided criterion and fuses data between every pair of key time steps into a summarized data set. Finally, the summary data sets are stored along with the raw data from the key time steps so that they can be analyzed and visualized during *post hoc* exploration. We verify the efficacy of our method by conducting an in situ study with a large-scale simulation.

In the future, we plan to develop criteria for detecting key time steps that will not need any domain knowledge so that key time steps can be detected in a purely data-driven way which will make the algorithm applicable across a wide range of scientific data sets. We also wish to run a GPU implementation of this technique with a larger case of MFIX-Exa to study the computational performance further.

Chapter 6

Dynamic Spatiotemporal Data Summarization using Information Based Fusion

6.1 Abstract

In the era of burgeoning data generation, managing and storing large-scale time-varying datasets poses significant challenges. With the rise of computing capabilities, the volume of data produced has soared, intensifying storage and I/O overheads. To address this issue, we propose a dynamic spatiotemporal data summarization technique that identifies informative features in key timesteps and fuses less informative ones. This approach minimizes storage requirements while preserving data feature dynamics. Unlike existing methods, our method retains both raw and summarized timesteps, ensuring a comprehensive view of information changes over time. We utilize information-theoretic measures to devise the fusion process, resulting in a visual representation that captures essential data patterns. We demonstrate the versatility of our proposed technique across datasets from diverse application domains. Our research contributes to the realm of data management and analysis, introducing enhanced efficiency and deeper

insights across diverse multidisciplinary domains. We provide a streamlined approach for analyzing large-scale datasets that can be applied to both post hoc and streaming analysis. This not only addresses escalating data storage challenges but also accelerates informed decision-making. Our method empowers researchers to explore salient temporal dynamics with minimal storage, enhancing a more intuitive understanding of complex data. This not only addresses the escalating challenges of data storage and I/O overheads but also unlocks the potential for accelerated informed decision-making. Our method empowers researchers and experts to explore salient temporal dynamics while minimizing storage requirements, thereby fostering a more effective and intuitive understanding of complex data.

6.2 Introduction

In today's data-driven world, the exponential growth in data generation has brought forth significant challenges for storage and associated I/O overheads. Modern computing capabilities have enabled the creation of massive datasets at an accelerated pace [45, 158]. Many of these datasets exhibit a dynamic temporal nature, spanning thousands of timesteps and needing large storage. Analyzing such a large number of timesteps poses significant challenges. One popular approach is to summarize the data by identifying the key timesteps. However, while key timestep-based approaches preserve the important events, automatic detection of key timesteps is non-trivial and the data dynamics for the intermediate non-key timesteps are completely ignored. Therefore, to preserve the temporal dynamics, only key timestep-based solutions may not be desired. We need novel data summarization methods that preserve both key events and overall temporal dynamics of the data in a storage-efficient compact format enabling accelerated analytics on large time-varying data.

To address the aforementioned need, we propose a data summarization technique that aims to minimize the storage overhead while preserving the salient temporal dynamics. We also emphasize visualizing these dynamics by tracking changes over time. Our approach involves a

dynamic spatiotemporal summarization (DSTS) technique, which adaptively identifies both key and redundant timesteps. We store the key timesteps and summarize redundant timesteps into a single timestep, highlighting the salient temporal characteristics of the features. The summarization technique ensures storage reduction with minimal information loss.

To achieve this, we use information-theoretic measures namely the Specific Mutual Information to guide the data fusion for the summary generation. The core idea of the summarization is to identify informative temporal features within the redundant (non-key) timesteps and fuse them using principles from information theory. By selecting the most relevant features from the redundant timesteps and summarizing through information-guided fusion, we retain the temporal dynamics. This approach optimizes storage requirements and facilitates the visualization and tracking of information change over time, providing valuable insights about data patterns.

In this paper, we present the details of the dynamic spatiotemporal analytics framework and the information-guided fusion process for summarization. We demonstrate the DSTS technique's versatility by applying it to different datasets, including scalar data from particle-based simulations and image sequences from surveillance video and biological cell interactions. Our results show significant storage reduction without compromising critical insights, emphasizing the effectiveness of our approach for efficient data management and visual exploration.

Our research aim in this work is to develop a method that efficiently handles large-scale time-varying datasets across various domains. Our solution seeks to bridge the data reduction landscape by providing approaches to effectively manage large and intricate temporal datasets. By leveraging the power of information theory, we aim to contribute to a more efficient data management strategy, especially in the context of dynamic and multifaceted datasets.

The contributions of the paper are:

- Develop a dynamic spatiotemporal summarization (DSTS) technique for large-scale time-varying datasets. The summary provides three features: key timesteps, fused timesteps, and holistic visual representation of information change.

- Propose several information-theoretic fusion strategies and comprehensively compare,

contrast, and evaluate their characteristics and applicability in summarizing the datasets.

- Demonstrate the flexibility and effectiveness of the proposed DSTS technique through application to diverse types of time-varying datasets including scientific flow simulations, surveillance video, and cell interactions in the immune system.

- Explore the impact of the proposed technique in optimizing data storage with minimal data loss.

6.3 Related Works

In this work, we focus on identifying and storing informative key timesteps while summarizing less informative (non-key) ones by fusion. The summaries reduce storage overhead while the key and fused timesteps pre serve data dynamics. Various data compression and reduction techniques have been explored. Among such methods, Cinema [5] is an image-based in situ data reduction and visualization approach. Lossless and lossy compression methods are also applied for data reduction [37, 192]. Among other techniques, statistical methods have been applied to perform data reduction and summarization [65, 211, 215]. These approaches store the reduced timesteps only. In our work, we aim to retain information from both raw and reduced timesteps to comprehensively capture information changes over time while preserving data dynamics.

Identifying key timesteps is imperative in analyzing time-varying data. There are numerous approaches [143, 188, 221]] that have been proposed for identifying key timesteps. These studies focus on only capturing key timesteps without the summarization capacity. Other studies focused solely on data reduction: in [6] similar timesteps are grouped and one is selected, in [200] salient timesteps were selected by comparing dissimilarity with previous timesteps. Unlike these studies, our work combines summarization with key timestep selection and data reduction.

Data fusion techniques [38] for large-scale spatiotemporal datasets has been a popular

field across various domains like remote sensing [113, 146], geoscience [125, 212], network architectures [97, 172]], computer vision [71, 201, 213], and time-varying scientific data [69]. In computer vision, various data summarization strategies have been explored, including Gaussian entropy fusion [81] and probabilistic skimlets fusion [218]]. Additionally, deep learning methods have also been applied for summarization [219]. Unlike some existing techniques, our approach doesn't need training and can be readily applied to large-scale datasets. Moreover, it is computationally efficient, rendering it applicable for both streaming and offline data analysis. The feasibility of in situ application is showcased in our preliminary research [69].

Information theory [49, 173, 190] has been employed to measure the relationships between variables in data across multiple computational domains [136, 167, 183]]. Mutual information (MI) is extensively applied for feature selection, exploration, extraction, and tracking [22, 185]. Image registration is another popular application [35, 93, 127]. MI, as well as its decomposition measures like specific mutual information and pointwise mutual information, have been widely used in multi-modal data fusion [26], data analysis, and visualization [8, 24, 41, 68, 95, 200]. Other use cases include view selection [195], feature similarity [33], and transfer function and design [29, 163]].

6.4 Information-Driven Framework for Feature-Based Temporal Data Summaries

6.4.1 Framework Workflow

In this section, we provide a comprehensive step-by-step description of our proposed technique's mechanism. We intend to construct an efficient, generic, and fast data summarization workflow with minimal customization to adapt to a variety of application domains. Figure 6.1 illustrates the schematic of this proposed workflow. To demonstrate each step of this workflow, we will refer to Figure 6.2, which serves as an illustrative application of our method using a synthetic

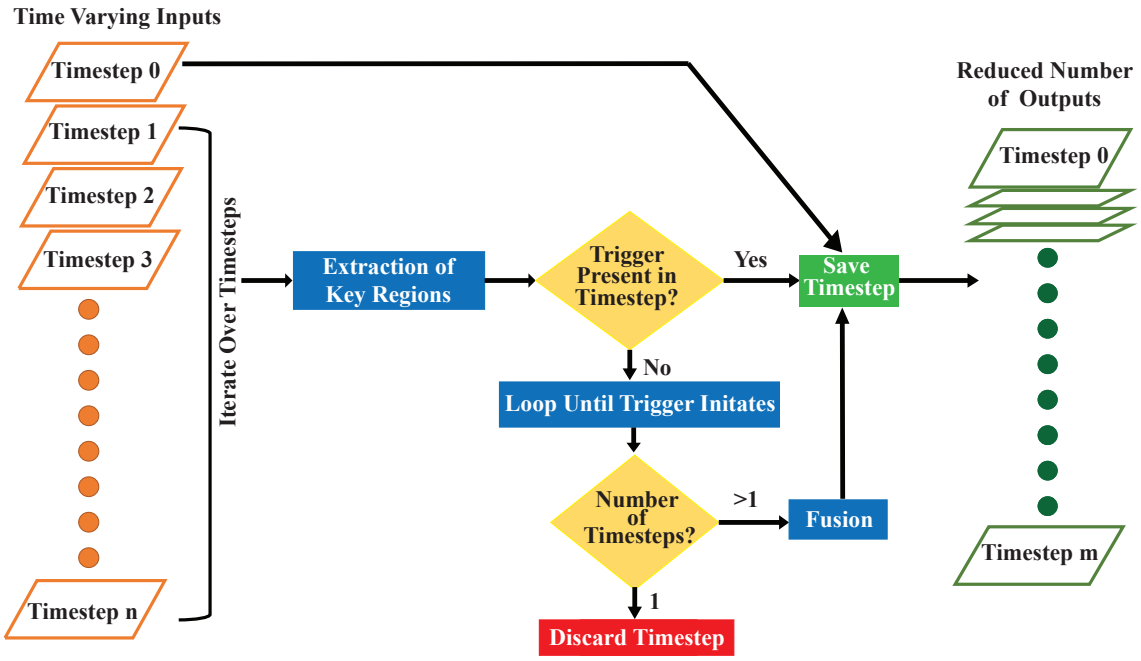


Figure 6.1: Schematic diagram of our workflow. Standard computational flowchart [175] symbols are used for representations: input/output, process, decision, and arrows indicating relationships between symbols.

data set. In this application, we simulate a rolling ball moving from left to right at each timestep until it exits the view area.

Our proposed method is designed for time-varying data containing various types of salient features. In Figure 6.2, the simulated rolling ball application consisting of 19 timesteps (T0 - T18), shows the ball's positional change over time. Each timestep is represented as 800×400 pixel 2D RGB image. At timestep T0, the frame is empty; the ball has not yet entered the view area. The ball enters at T1 and changes position until T17; finally exiting the view area at T18. The proposed method iteratively processes the sequence of input timesteps. After the first timestep, for every subsequent timesteps, the key regions are extracted using a segmentation method proposed in [108]. Criteria for extraction of such key regions is determined by the domain knowledge. In this case, the key feature is the presence of the ball and its location. So we segment the region containing the ball and create binary masked images shown in Figure 6.2 (masked row). These images contain only two data values: 0 (no ball) and 255 (ball).

After extracting the key region, we check if a certain property is present in that timestep. We denote this property as *trigger* which is a change in the key region. The change can be in terms of count, size, shape, connectivity, space, or association. In this rolling ball demonstration, the triggers are the first appearance and final exit of the ball from the view area. When the ball enters and then exits the area, it is considered as salient information. But the time the ball remains in the area, the only novel information is its change of position. If a trigger is present in the current timestep, then it is considered as a key timestep. Hence our method saves it as it is. If the trigger is not present, we proceed to the next timestep, do a similar check, and continue the process until a trigger is encountered. These intermediate sequential timesteps that did not have the trigger are chosen to be fused into a single timestep as the amount of novel information within such a sequence is low. If the number of timesteps to be fused is one then we can discard it as the previous timestep has the necessary information. If the number is greater than one then we perform pairwise information-guided fusion on these timesteps and convert them as one single timestep to be saved as a temporal summary. Referencing the demonstration in Figure 6.2, T0 is saved. Subsequently, for T1 through T17, no triggers are identified, and again, T18 is saved. Therefore, T1 through T17 are fused as shown in figure 6.2(a) and (b). Note that, since our method processes one timestep at a time incrementally as they appear, it can be applied to applications where data is streamed for real-time processing. The following section describes our use of the information theory-guided fusion method.

6.4.2 Characterization of Samplewise Information for Fusion

In this work, our aim is to track and summarize the information change in fused timesteps. Therefore, we need a quantification of information content for each data point in the timesteps. We use the term ”sample” to refer to individual data points. Each timestep contains multiple samples representing the values of the data. In the case of images like the rolling ball, these samples range from 0 to 255, while for other types of data variables, they may be scalar values. Quantifying information for these samples will help identify important spatial features for the

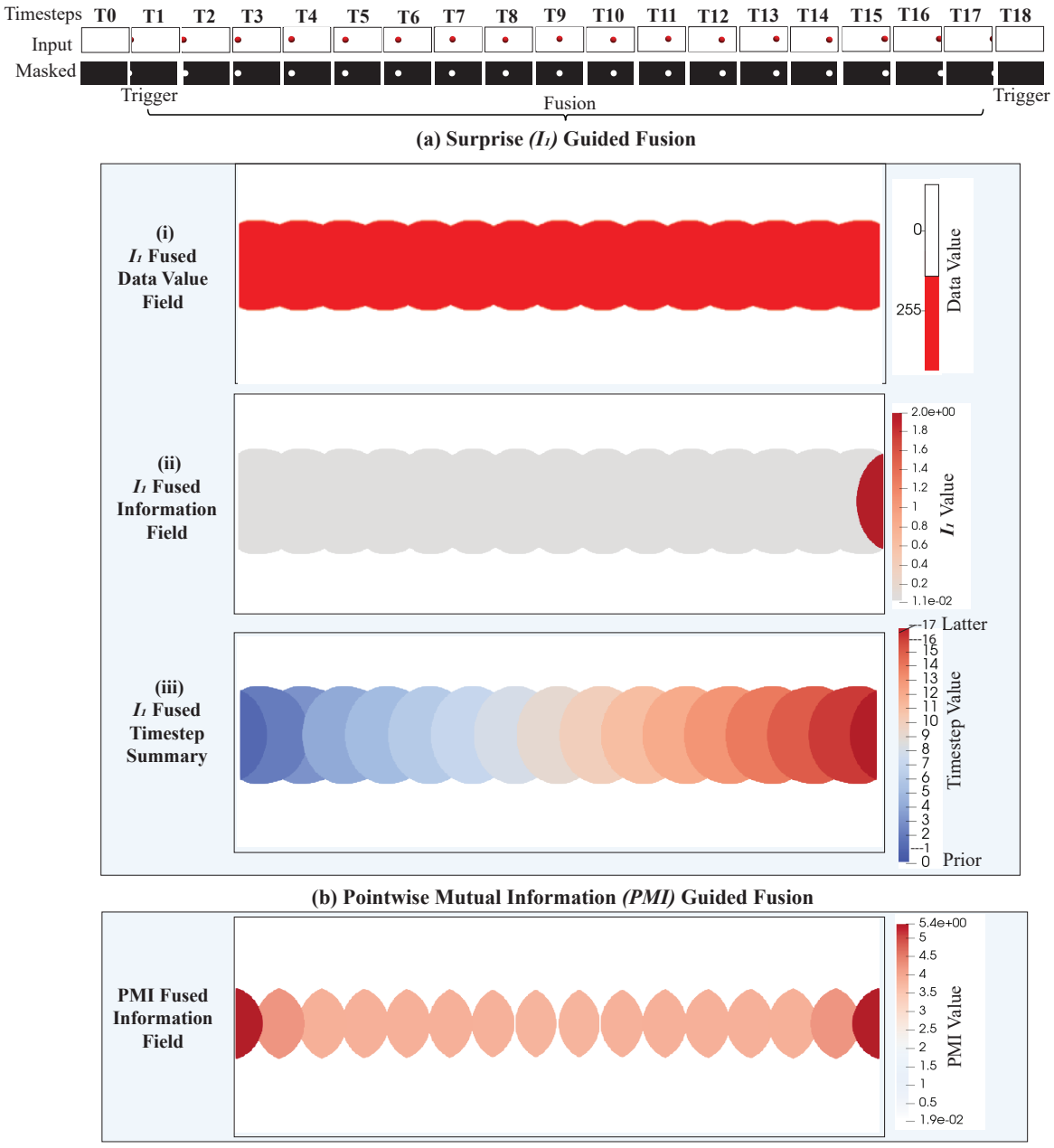


Figure 6.2: Illustration using a simulated rolling ball with 19 timesteps (T0 -T18). At each timestep, the ball moves 0.5 units to the right. Timesteps are 2D RGB images with 800×400 dimensions with data (pixel) values ranging between 0 to 255 (input row). The masked row presents binary images with values 0 (no ball) and 255 (ball). T1 to T17 are fused using (a) Surprise (I_1) guided fusion and (b) PMI guided fusion. a(i) shows the I_1 fused data value field (0 white and 255 red). a(ii) shows I_1 fused information value field. a(iii) displays I_1 fused timestep summary with numbered color labels for each timestep. The numbers indicate spatial information changes over time. Surprise effectively captures spatiotemporal properties, whereas alternative PMI measure does not perform well. (b) shows the scenario with the information field using PMI values.

timestep.

Mutual Inforamtion

In information theory, Mutual Information (MI) [173] is a prominent measure that estimates the total amount of shared information between two random variables. Given two random variables X and Y, MI $I(X;Y)$ is formally defined as:

$$\mathrm{I}(X;Y) = \sum_{y \in Y} \sum_{x \in X} \mathrm{p}(x,y) \log \frac{\mathrm{p}(x,y)}{\mathrm{p(x)\,p(y)}} \tag{6.1}$$

where $\mathrm{p}(x)$ and $\mathrm{p}(y)$ are the probabilities of occurrence of values x for X and y for Y respectively. $\mathrm{p}(x,y)$ is the joint probability of occurrence of values x and y together. MI assesses the degree of association or disassociation between two random variables and gives a single value. Since we aim to extract feature-based data summaries, we need samplewise spatial and temporal information characterization. Therefore, we leverage the decomposition of MI which quantifies each data value's contribution toward the association or dissociation. The decomposition of MI is termed as Specific Mutual Information or SMI [55]. SMI measures the information content of the individual scalar values of one variable (reference) when another variable (target) is observed. There are multiple methods for MI decomposition [34, 55]. For apprehending the fusion criteria essential for summarizing the data, the properties of the SMI measure, *Surprise* holds the most potential.

SMI Measure Surprise

The Surprise measure denoted as I_1 was first introduced by [55]. Surprise quantifies the information change of the target variable after observing the individual scalar values of the reference variable. The derivation of Surprise from MI is as follows.

By definition, the conditional probability of x given y is:

$$\mathrm{p}(x|y) = \frac{\mathrm{p}(x,y)}{\mathrm{p(y)}} or \, \mathrm{p}(x,y) = \mathrm{p}(x|y)\,\mathrm{p(y)} \tag{6.2}$$

Replacing the joint probability in Equation 6.1, we get,

$$\begin{aligned} \mathrm{I}(X;Y) &= \sum_{y \in Y} \mathrm{p(y)} \sum_{\mathrm{x} \in \mathrm{X}} \mathrm{p(x|y)} \log \frac{\mathrm{p(x|y)}}{\mathrm{p(x)}} \\ &= \sum_{y \in Y} \mathrm{p(y)}\,\mathrm{I}_1(\mathrm{y};\mathrm{X}) \end{aligned} \tag{6.3}$$

where,

$$\mathrm{I}_1(y;X) = \sum_{x \in X} \mathrm{p}(x|y) \log \frac{\mathrm{p}(x|y)}{\mathrm{p(x)}} \tag{6.4}$$

Equation 6.4 represents the surprise measure of data value y from Y after observing all the values of X. A high value for $I_1(y;X)$ means after observing y, some previously low probable values of $x \in X$ have become highly probable. This likelihood increase is the element of surprise and a salient finding for further analysis. Surprise is also the only positive decomposition of MI since it is the Kullback-Leibler distance between $\mathrm{p}(x|y)$ and $\mathrm{p(x)}$ [107]. In the process of the data summarization, the data samples with high surprise values stand out and are identified as important features.

6.4.3 Surprise (I_1) Guided Fusion Technique

When the low informative timesteps are chosen, the fusion initiates for summarization. The fusion is done on pairwise timesteps. For every pair of data samples from the timesteps, we store samples with high I_1 values. This fusion strategy was introduced by [26] for fusing different datasets to gain the most informative combination. The condition to compute the fused value using I_1-fusion is:

For every data sample pair with (x, y), the fused value, f is,

$$\mathrm{f} = \begin{cases} \mathrm{x}, & \text{if } \mathrm{I_1(x;Y)} > \mathrm{I_1(y;X)} \\ \mathrm{y}, & \text{otherwise} \end{cases} \tag{6.5}$$

Here x and y are individual data values from two data sets X and Y. Our fusion criteria is based on the idea of Equation 6.5, however, instead of different datasets we are using two subsequent timesteps from the same dataset. To fuse multiple timesteps, we begin by creating a fused timestep using the first two timesteps. Then, we repeat the fusion process by comparing the fused timestep with the next timestep and continue until all desired timesteps have been fused. Our strategy involves updating the fused timestep during each iteration and selecting the spatial and temporal values with the highest information content. By the end of the process, the resulting fused timestep will represent a summary capturing their most informative properties with direction. The fusion process is described in detail in Algorithm 1. After each fusion process, the algorithm provides 3 fused fields as shown in 6.2(a). I_1 fused data value field contains the values of the data samples with high surprise measure. I_1 fused information value field contains the I_1 values for the same sample positions. In the timestep summary fields, the same data samples are labeled with their originating timestep numbers.

Applying the fusion process in Algorithm 1 on T1 - T17 of the simulated rolling ball, the I_1 fused data value field is generated highlighting the path of the ball with values 255 as shown in Figure 6.2 a(i). The regions without the ball are valued 0. Figure 6.2 a(ii) represents the fused information fields with I_1 values. From the color bar's gradient, we observe that the surprise values exhibit limited variation, spanning approximately from 0 to 2. A minimal data value range results in minimal surprise variation. Figure 6.2 a(iii) presents the timestep summary where the numbers of originating timesteps are labeled for the salient samples. Here, we employed distinct colors to label the timesteps, enabling clear visualization and differentiation of each timestep. This color-coded representation shows the flow of information, facilitating

the tracking of information changes over time. Here, the confidence threshold is employed to downplay the non-important regions. In this particular case, the threshold value is set to 255. Any value below 255, representing the absence of the ball, is assigned as timestep 0. In Figure 6.2 a(iii), these regions are depicted as white or transparent (steps 19 - 24 in Algorithm 1). The method reduces the number of output timesteps from 19 to 3 in the simulated rolling ball case, achieving substantial data reduction with minimal loss. The fused timestep effectively visualizes the information changes over time, serving as a summary of the original data dynamics.

6.4.4 Alternative Fusion Approaches

The measure *Surprise* effectively apprehends the spatiotemporal features for the data summarization. However, we also explore other potential information measures to devise alternative techniques for generating data summaries. These information-theoretic measures include Pointwise Mutual Information (PMI) [46], SMI measures (1) Predictability (I_2) [55] and (2) *Stimulus* Specific Information (I_3) [34]. These measures were investigated because they are the decomposition of MI and they possess the ability to analyze the contributions of individual data values in quantifying the information content of spatiotemporal data.

PMI Guided Fusion

PMI [46] quantifies the degree of association (or disassociation) between individual data points given two variables. If X and Y are two variables, then each data point can be represented by the value pair (x,y) where $x \in X$ and $y \in Y$. The statistical association between these two points can be measured by their PMI value:

$$\mathrm{PMI}(x,y) = \log \frac{\mathrm{p}(x,y)}{\mathrm{p(x)\,p(y)}} \tag{6.6}$$

where $\mathrm{p}(x)$ and $\mathrm{p}(y)$ are the probabilities of occurrence of values $x \in X$ and $y \in Y$. $\mathrm{p}(x,y)$ is the joint probability of occurrence of values x and y together. Comparing Equations 6.1

Algorithm 1 Fusion Process

Require: data1: Array of data values from fused timestep. Initialized with the first timestep.
data2: Array of data values from the subsequent timestep.
Ifield1: Array of I_1 values for $I_1(x;Y)\ \forall x \in X$
Ifield2: Array of I_1 values for $I_1(y;X)\ \forall y \in Y$
timestep_fuse: Array of timestep values. Starts with 0
time: Current timestep value
conf_th: Confidence threshold for the key regions

Ensure: fused_field_data: Array of the fused data values
fused_field_I1: Array of the fused I_1 values
timestep_fuse: Array of the fused timestep values.

0: **procedure** CREATEFUSIONFIELDS(List of Input)
0: *fused_field_data* ← array of zeros with shape *data1*
0: *fused_field_I1* ← array of zeros with shape *data1*
0: **for** $i \leftarrow 0$ to $data1.shape[0]-1$ **do**
0: **for** $j \leftarrow 0$ to $data1.shape[1]-1$ **do**
0: **if** $Ifield1[i][j] > Ifield2[i][j]$ **then**
0: $fused\_field\_data[i][j] \leftarrow data1[i][j]$
0: $fused\_field\_I1[i][j] \leftarrow Ifield1[i][j]$
0: **if** $time = 1$ **then**
0: $timestep\_fuse[i][j] \leftarrow time$
0: **end if**
0: **else**
0: $fused\_field\_data[i][j] \leftarrow data2[i][j]$
0: $fused\_field\_I1[i][j] \leftarrow Ifield2[i][j]$
0: $timestep\_fuse[i][j] \leftarrow time+1$
0: **end if**
0: **end for**
0: **end for**
0: **for** $i \leftarrow 0$ to $data1.shape[0]-1$ **do**
0: **for** $j \leftarrow 0$ to $data1.shape[1]-1$ **do**
0: **if** $fused\_field\_data[i][j] < conf\_th$ **then**
0: $timestep\_fuse[i][j] \leftarrow 0$
0: **end if**
0: **end for**
0: **end for**
0: **return** $fused\_field\_data, fused\_field\_I1, timestep\_fuse$
0: **end procedure**=0

and 6.6, we can infer that the expected PMI values over all occurrences of variables X and Y correspond to the MI value $I(X;Y)$. PMI is a symmetric measure that can generate values ranging from negative to positive, depending on whether the distributions are complementary or overlapping. If the information overlap is high ($\mathrm{p}(x,y) > \mathrm{p}(x)\,\mathrm{p}(y)$), then $\mathrm{PMI}(x,y) > 0$. The low association is indicated by $\mathrm{p}(x,y) < \mathrm{p}(x)\,\mathrm{p}(y)$, resulting in $\mathrm{PMI}(x,y) < 0$. If x and y are statistically independent then $\mathrm{p}(x,y) = \mathrm{p}(x)\,\mathrm{p}(y)$ and $\mathrm{PMI}(x,y) = 0$.

Given the PMI measure, we can devise a fusion strategy similar to I_1 where I_1 values are substituted with PMI values in Algorithm 1. The resulting fused information field on the simulated rolling ball is shown in Figure 6.2(b). We observe that PMI fails to capture the spatial characteristics of the key regions and only captures the overlapped regions indicated by high positive PMI values. The non-overlapping regions are transparent showing low information overlap. The PMI values here are 0 meaning data distribution is complementary and statistically independent. As the spatial position of sample pairs plays a critical role in PMI calculation, the fused field properties can exhibit significant variation depending on the degree of overlap between the key features.

I_2 Guided Fusion

Predictability (I_2) is another decomposition of MI introduced by [55]. This SMI measure quantifies the change in the uncertainty of one variable (X) after observing the individual value of another variable $(y \in Y)$ and is computed as:

$$\mathrm{I}_2(y;X) = -\sum_{x\in X} p(x)\log p(x) + \sum_{x\in X} p(x|y)\log p(x|y) \tag{6.7}$$

where $y \in Y$ is the reference variable and $x \in X$ is the target variable. $\mathrm{p}(x)$ is the probabilities of occurrence of values x for X and $\mathrm{p}(x|y)$ is the conditional probabilities values of x given y. Upon observing the variable y, the uncertainty of variable X can either increase or decrease, leading to the possibility of both positive and negative values for the I_2 measure. In some cases,

the increased uncertainty can reveal significant information about the relationship between the variables. However, when we use the I_2 measure instead of the I_1 measure in the fusion process, the resulting fused information does not offer a meaningful summary over time. Our hypothesis is that this measure is more suitable for feature extraction and uncertainty quantification across various datasets rather than for analyzing consecutive timesteps within a single dataset.

I_3 Guided Fusion

There is another measure that is derived from the decomposition of MI and was introduced for measuring the association of stimulus and response in certain neural systems [34]. It is termed as Stimulus Specific Information (SSI), denoted by I_3:

$$\mathrm{I}_3(y;X) = -\sum_{x \in X} p(x|y)\,\mathrm{I}_2(x;Y) \tag{6.8}$$

The response and stimulus are the two variables X and Y. This measure focuses on establishing the association between the two variables to extract the maximum amount of information from their relationship. It emphasizes that the most informative data values from the first variable are related to the most informative data values of the second variable [34]. In some cases, I_1 can be an alternate measure for I_3, but the interpretation is different based on the data [34]. When I_3 is used instead of I_1 on the simulated rolling ball dataset, it captured very similar properties shown in Figure 6.2(a). However, when applied to a more complex dataset, it failed to capture the spatial properties of the features in the summarization. This is explained in detail in Section 6.5.1 and shown in Figure 6.3(b).

From the various methods mentioned, it's clear that different approaches emphasize distinct data properties. After a thorough investigation, we found that the *Surprise* measure aligns best with our objectives of highlighting features and summarizing spatiotemporal data while tracking information flow. In the next section, we apply this method to more complex datasets to demonstrate its versatility across different data scenarios.

6.5 Applications

In this section, we assess the versatility of our proposed DSTS method across multidisciplinary applications, demonstrating its effectiveness in handling complex and diverse datasets. We select three applications across multiple domains:

- A scalar dataset obtained from a particle-based multiphase flow simulation MFIX-Exa.
- A surveillance video dataset with extended timesteps.
- An image-based dataset consisting of complex immune cell interactions.

6.5.1 MFIX-Exa Flow Simulation

MFIX-Exa [141] is a multiphase flow simulation developed by the National Energy Technology Laboratory (NETL), USA. Using MFIX-Exa, particle-based data is generated for studying the operational principles of chemical looping reactors. In such simulations, formation of void regions, known as bubbles, is an important phenomenon. Understanding the temporal evolution of these bubbles holds significant importance for domain experts. MFIX-Exa generates data with millions of particles and thousands of timesteps. This extensive raw data presents significant challenges in transferring to storage due to limited I/O bandwidth. Hence, experts seek solutions to extract bubble-specific information while reducing output storage and preserving temporal bubble dynamics. Our proposed DSTS method can be used to provide this solution.

Data Context and Features

For analyzing bubble dynamics, typically the raw particle data is first converted to a scalar density field. Then bubbles can be segmented as the connected regions with low particle density. For more details about this pre-processing, please refer to [70].

In this work, we assume that the scalar density fields are already available and we use 2D slices extracted from the density fields. These slices contain scalar values representing particle

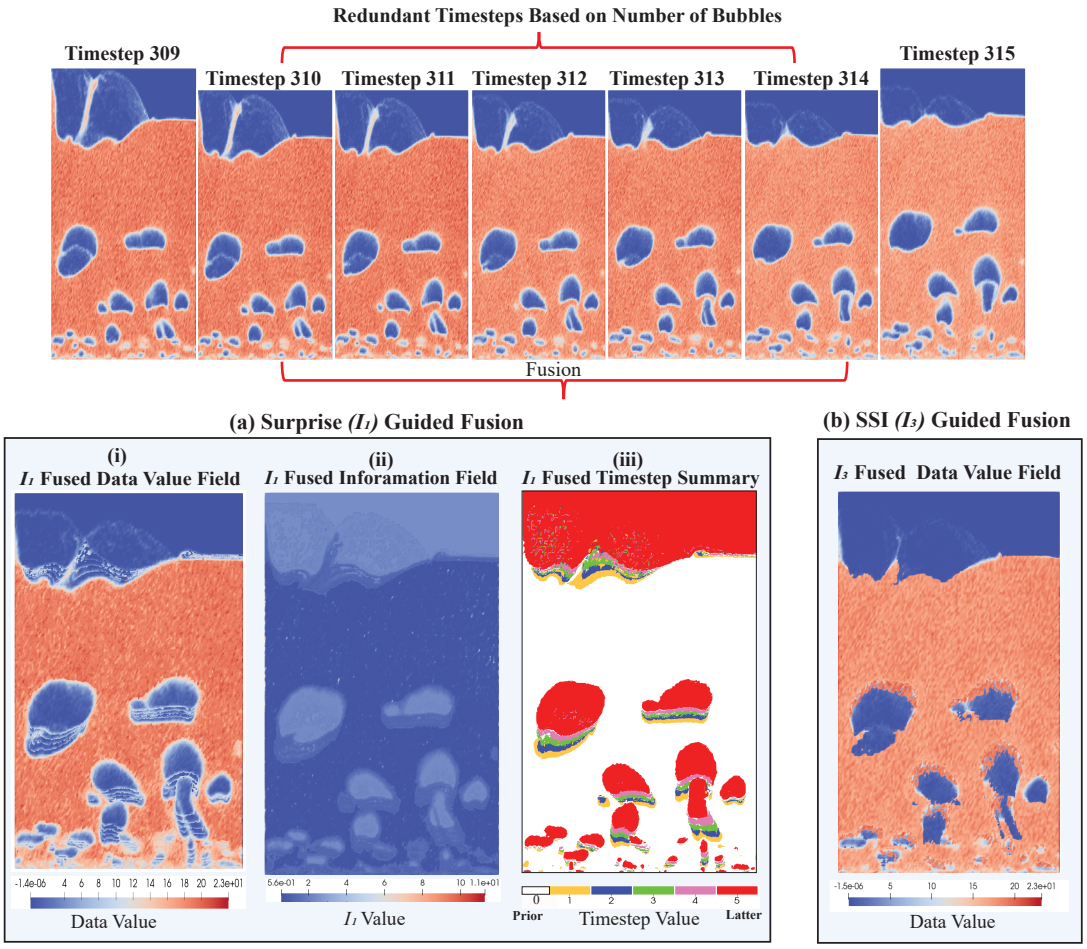


Figure 6.3: Analysis of DSTS method for MFIX-Exa simulation. The first row represents a window of timesteps (309 - 315) where the bubbles are highlighted as the blue regions. Timesteps are raw images with 488×842 dimensions. 310 to 314 are fused using (a) Surprise (I_1) guided fusion (i-iii) and alternative (b) Stimulus Specific Information (SSI) or I_3 guided fusion. a(i) shows the I_1 fused data value field of the timesteps, a(ii) shows I_1 fused information value field reflecting I_1 values and a(iii) shows I_1 fused timestep summary with 5 timesteps. The color bar labels 5 different colors summarizing the direction of the bubbles in one timestep. The alternative I_3 is unable to capture the path of bubbles as reflected in (b) I_3 fused data field.

density. Our evaluation dataset consists of multiple timesteps (count 332), and each timestep corresponds to 2D data samples with dimensions of 488×842. The sample values fall within the range of [-1×10^{-6}, 29.08]. As mentioned earlier, the key features are the bubbles with low particle density. In [70], the detection, segmentation, and characterization of the bubbles are studied in an extensive manner. In our work, we use the VTK [170] library to extract the connected components and then use a low scalar density threshold value to filter the bubbles. Over time, the bubbles undergo phases like creation, merge, split, and dissolve into air. Domain experts want to comprehend the evolution of bubbles and explore the relationships between various bubble characteristics such as their size, shape, number of bubbles, etc. [70]. Important events in this simulation can be characterized by the creation of a bubble, the merging of two or more bubbles, or the dissolving of a bubble. Note that for all of these events, the total number of bubbles will change. Hence, a timestep with the bubble number changed from a previous timestep, can be considered as a ”trigger”. Here, we ignore counting changes in very small bubbles since the domain experts are more concerned about the bubbles when they grow in size. Key timesteps are saved when the trigger occurs, and intermediate steps between two triggers are fused using Algorithm 1 for summrization. This process continues for all timesteps.

Results for Data Summarization

Figure 6.3 shows the analysis of the DSTS method for MFIX-Exa simulation. Timesteps 309 - 315 are shown in the first row where bubbles are the blue regions. Timesteps 310 to 314, during which the number of bubbles remains unchanged, are summarized through the fusion process. Figures 6.3 (a) represent results from the I_1 guided fusion. The I_1 fused data value field a(i) shows the scalar values ranging [-1×10^{-6}, 23] for the fused timesteps. Here the change in bubble movement is very prominent. Figure 6.3 a(ii) presents the I_1 values ranging [6×10^{-1}, 11] for the fused timesteps. The range of I_1 values is smaller, making it less sensitive to bubble movement compared to particle-density values. However, it effectively highlights the main bubbles and their temporal dynamics. The timestep summary field in Figure 6.3 a(iii)

represents the timestep values from which the bubbles originate. Here 5 timesteps are distinctly color-coded to highlight the flow of the information between interacting bubbles. The color map reflects the direction of the bubbles from the start to the end position. This summary field emphasizes key features and visually indicates the spatial information flow over time. The white background (labeled 0) filters all the density values that are of low importance for this dataset.

We have also implemented the alternative SSI (I_3) guided fusion technique as mentioned in Section 6.4.4 for MFIX-Exa. Figure 6.3(b), represents the I_3 fused data value field. Here the bubbles are only partially highlighted and the change in the bubbles' movement is also hard to interpret. While I_3 captures some spatial features of the bubbles, the edges are blurred. Thus, the *Surprise* fusion method proves to be better than the SSI measure.

Based on the outcomes in both the rolling ball and MFIX-Exa simulation applications, it is evident that the timestep summary field stands out as the most informative visualization for summarization. This representation encapsulates both spatial and temporal dynamics. Consequently, for our analysis of the next two applications, we only show the surprise fused timestep summary fields.

6.5.2 Surveillance Data Analysis and Optimization

Data summarization techniques can significantly benefit the security camera footage analysis. Security camera systems generate vast amounts of data, and reviewing the continuous stream of videos can be time-consuming and cumbersome. DSTS offers an efficient solution by allowing the optimization of camera footage. Suspicious activities can be detected by choosing an appropriate "trigger" and the generated summary fields provide experts with a comprehensive visual representation of the key timesteps. The most significant impact of our proposed method in this application is on archiving the storage optimization.

To demonstrate DSTS in security camera footage analysis, we used the publicly available SBM-RGBD Dataset [36, 168]. This dataset was originally created for the Workshop on

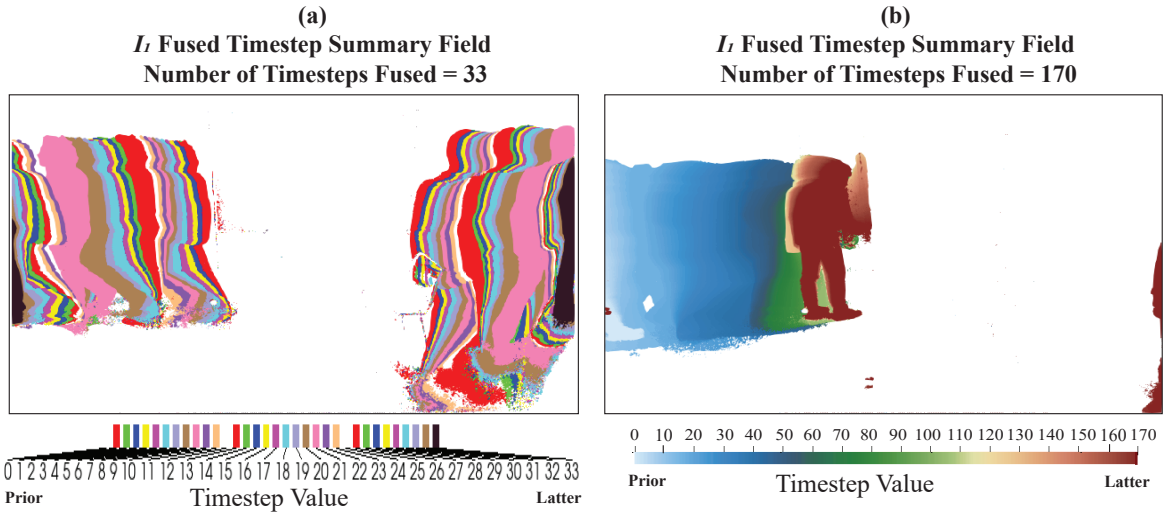


Figure 6.4: Results of the DSTS method for SBM-RGBD dataset. Here the emphasis is on representing the extensive number of timestep summaries. (a) shows a summarization of 33 fused timesteps using a discrete color bar. (b) showcases a summarization of 170 fused timesteps, employing a continuous color bar to depict information changes over a longer period. The color bars point out the spatial direction of the information flow by denoting the prior and latter states.

Background Learning for Detection and Tracking from RGBD Videos [160]. The dataset comprises 33 RGBD videos, totaling 15033 timesteps, recorded indoors using a Microsoft Kinect sensor [36]. The dataset contains videos capturing moving objects at intervals, which aligns with our data requirements. Here, we used one of the videos titled ”Multipeople2” which shows four individuals walking in and out of the view area, engaging in discussions, and writing on a whiteboard. Our method shows an effective demonstration of summarizing the movement patterns of the individuals.

Data Context and features

The videos have 640×480 resolution and the length is 1400 timesteps. The dataset comes with PNG files for each timestep of the input video. The key regions (features) are the individuals and their movement. The whiteboard and a chair are stationary in the background. To extract key regions, we have used a background subtraction algorithm, called ViBe [19]. The algorithm aims to identify moving objects within consecutive images or videos by distinguishing between

the foreground (moving objects) and the background (stationary elements). The ViBe algorithm is adaptive and computationally lightweight, making it suitable for real-time applications like object tracking, surveillance, and motion detection. Its straightforward pseudocode in [19] facilitates easy implementation. The ViBe algorithm converts the RGB images into masked binary images with segmented individuals.

In scenarios with multiple individuals, we adopt a concept similar to that used for counting bubbles in the MFIX-Exa (Section 6.5.1). We apply the concept to count the number of individuals in each timestep by analyzing the largest connected regions in the masked images. Since it is a binary image, the data samples have two values: 0 (no individual) and 255 (individual). By setting a size threshold, we can accurately count the number of individuals in each timestep. Given that individuals move in and out of the view area, the number of individuals can be used as a trigger for our application. Whenever a person enters or exits, that is a key timestep. The consecutive timesteps between two triggers are then fused using Algorithm 1. This fusion process effectively summarizes the movement patterns of individuals within one timestep. The combination of the key and summary timesteps, provides an intuitive visual representation of the significant moments in the video, making the analysis of surveillance scenarios more informative.

Results for Data Summarization

Figure 6.4 showcases the summarization fields for two separate fused timesteps in the dataset. Both fields show I_1 fused timestep summaries representing spatial features and movement directions of individuals over time. In this application, our main aim is to showcase how effectively the method can fuse longer timesteps while ensuring that both spatial and temporal dynamics remain just as noticeable as in shorter timesteps.

In Figure 6.4(a), the summarization field depicts a fusion of 33 timesteps, where two individuals walk out of the view area. The leftmost person exits first, initiating the trigger and stopping the fusion process. Each timestep is represented by a discrete color, highlighting the

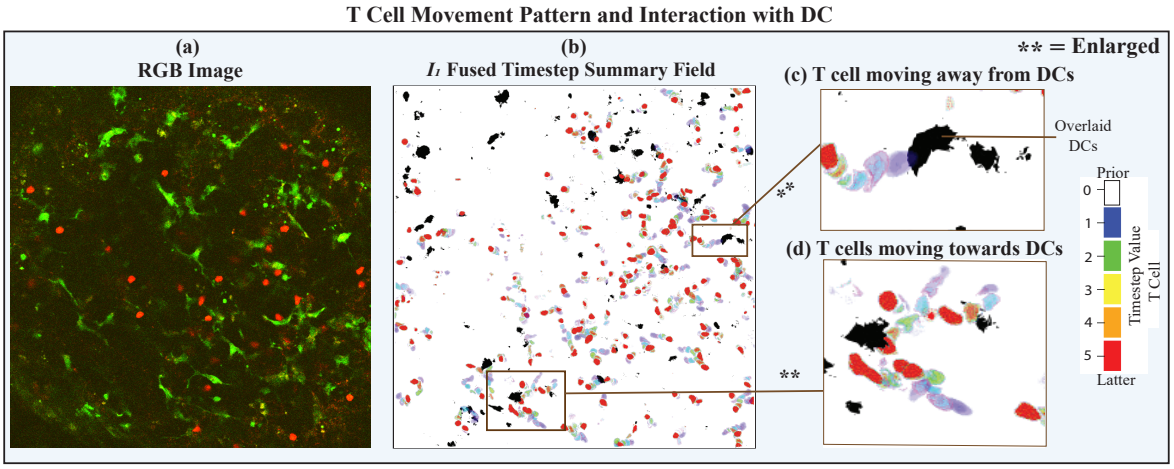


Figure 6.5: Results of the DSTS method for cell interaction in Lymph Node. The results emphasize T cell movement patterns and interaction with DCs in LN. (a) is an illustrative timestep from the dataset. Here red indicates the T cells and green indicates the DCs. (b) is the surprise fused timestep summary fields of T cell for 5 timesteps. The black cells in the field are overlaid DCs to highlight cell contact. (c) and (d) are 2 enlarged positions from the (b) field to emphasize T cell movement. (c) shows that a T cell is moving away from the DCs. (d) shows multiple T cells moving toward the DCs. The corresponding timestep values are provided in the color bars to highlight the direction.

changes in movement over 33 timesteps. Sample values below 255 are set to 0, representing the white background, as the data value of individuals is 255.

Figure 6.4(b) shows a timestep summary for a longer period of 170 timesteps. The summarization field captures an individual walking into the view area and writing on the board, while another person has just stepped in, initiating the trigger. Continuous colors are used to display the movement changes due to the length of the fused timesteps.

Expectedly, key and summarized timesteps in this dataset result in a significant reduction from 1400 timesteps to only 49 timesteps. The highest number of timesteps being fused is 262. This notable optimization ensures that only relevant information is stored, reducing storage requirements without compromising critical insights into the movement patterns of individuals. By efficiently identifying and storing key moments, our method enhances the analysis of crowded environments, enabling rapid detection of suspicious activities, and thereby serving as a valuable tool for surveillance.

6.5.3 Tracking Cell Interactions in Lymph Nodes

This dataset contains consecutive images of cellular interactions within the lymph node (LN). LNs are essential for immune function, playing a crucial role in initiating immune response and facilitating immune cell communication [134]. In the LN micro-environment, naïve T cells are activated by interactions with different cell types. Understanding these interactions provides valuable insights into immune activation [31]. We reanalyze data from [183], where information theory-based approaches were used to identify and quantify the spatial relationships between naïve T cells and three target cellular components: dendritic cells (DCs), fibroblastic reticular cells (FRCs), and blood vessels (BVs). These interactions are critical in T cell movement and the timing of encounters with antigen-presenting DCs. This process is a key step in T cell activation and the initiation of the adaptive immune response.

The data for the study was gathered using two-photon microscopy (2PM) [162] to acquire 3D image stacks of LN tissue samples from mice. The imaging process captured dynamic movies lasting 10 to 45 minutes, resulting in a sequence of 3D images. This dataset is well-suited for the application of our method. Next section demonstrates that our approach offers both quantitative analysis and visualization of cell movement and communication. Additionally, this dataset showcases our technique's applicability to 3D images and movies, highlighting its suitability for complex spatial interactions in biological systems.

Data Context and Features

Figure 6.5(a) shows an RGB image with T cells dyed red and DCs dyed green. Each voxel contains the color intensities of the dye in the red, blue, and green channels. For every time step, we extract the red and green channels into two separate images. We focus on the red channel in order to analyze T-cell motility.

Because these images contain a lot of noise, we implement a pre-processing step using the median filter [86], to reduce noise while preserving the edges of the cells for improved

visualization. Since the red channels specifically represent the T cells, no segmentation is required.

Our goal is to visualize how T cells move and interact in these movies. In [183], MI and normalized mutual information (NMI) were used to quantify associations between cells. Here for each timestep, we use the MI value between two cell types as a "trigger". If the MI value for a specific timestep exceeds a specified threshold, we save that as a key timestep. If the MI value falls below the threshold, we find the next timestep in which the MI value exceeds the trigger threshold and fuse the intermediate ones. This allows us to efficiently capture and represent significant interactions while fusing less informative time steps.

Results of Data Summarization

Figure 6.5 focuses on the T cell movement and interaction with DCs in the summarized timesteps. Figure 6.5(a) is a sample timestep of the T:DC dataset with $512 \times 512 \times 22$ dimensions. This dataset has a total of 51 timesteps. Figure 6.5(b) displays the I_1 fused timestep summary field, representing five fused timesteps from this dataset. The black cells in the summarization represent the DCs' value field overlaid on the fused summary field, visually illustrating the physical interactions between T cells and DCs. Given that there are multiple interactions captured in each timestep, we highlight two specific interactions by enlarging the locations in Figures 6.5(c) and (d). In Figure 6.5(c), we observe a T cell moving away from the DCs. The color bar on the right indicates the first (blue) and last (red) timesteps in the summary field, clearly indicating the movement direction. In Figure 6.5(d), we see multiple T cells moving toward the DCs making explicit contact. This visualization allows for a comprehensive understanding of the dynamic interactions between T cells and DCs, providing valuable insights into the temporal dynamics of immune cell communication.

Our method, successfully visualizes physical contact between cells and tracks movement over time. Other studies [62, 139] using similar datasets have presented the statistical quantification of association. We believe that the addition of the visualization capability has the

potential to unveil new insights for experts to analyze these associations further. This could contribute notably to the study of T cell motility, a crucial aspect for understanding immune response dynamics.

6.6 Discussion

The proposed DSTS technique has demonstrated its effectiveness and flexibility across several applications. Starting with a synthetic simulation of a rolling ball to analyzing complex cellular interactions within lymph nodes, the method effectively showcased its robustness. The combination of the key and fused timestep resulting from the method provides a compact yet comprehensive data summarization. The intuitive visual representation is a plus in highlighting an ideal fusion of data while preserving salient information changes over time. We evaluate the feasibility of multiple information theory measures and establish that SMI measure *Surprise* performs the best to capture the complex spatiotemporal features effectively.

We select applications from multiple domains to shed light on different aspects of the DSTS method. This technique offers a practical solution to downsize and analyze the features in the MFIX-Exa simulation. This application analysis establishes that the method can handle raw scalar data as well as image-based data that incorporates the rest of the applications.

The RGBD tracking dataset is introduced to show the method's ability to summarize and highlight important movement patterns of individuals in a video sequence. The results from this dataset reflect that longer fused timesteps are equally apprehensible as the shorter ones. This has promising implications for surveillance and security applications.

The cellular interaction in LN is a more complex dataset. T cells which are the key regions (features) are ample in number and the interactions with DCs are sporadic in nature. Our method is able to track multiple cell interactions. The summarization highlights immune cell communication by providing a comprehensive visualization of the T cell movement. This visualization potentially introduces new possibilities for immunological research.

All the applications in this work present post hoc data analysis. Since the method is not computationally expensive it can be easily combined to analyzing data in a streaming framework. Through the integration of this method in any in situ streaming framework, the resulting data will be summarized in real-time ensuring optimal storage reduction.

6.7 Conclusion and Future Work

We acknowledge that challenges may arise in selecting appropriate triggers and threshold values, especially in complex datasets with multiple key features, interactions, and noise. However, the flexibility of the technique allows for the adjustment of parameters to tailor the summarization process to different applications. Additionally, future research could explore combining different information-theoretic measures to further enhance the summarization capabilities to multivariate time-varying datasets.

In summary, our proposed technique is a powerful tool for visualizing and analyzing large-scale time-varying datasets, demonstrating adaptability to offer valuable insights into data dynamics across various domains. The approach holds promise for novel discoveries and applications, ultimately enhancing our understanding of complex data systems.

Chapter 7

Analyzing the Spatial Spread of SARS-CoV-2 in Lung CT Scans using SIMCoV

7.1 Abstract

The Spatial Immune Model of Coronavirus or SIMCoV is a computational model developed to study SARS-CoV-2 infection by analyzing the spatial distribution of infected cells in the lungs and the immune response in patients. Running efficiently on HPC systems, SIMCoV can simulate hundreds of millions of cells (both lung and immune) and is able to show how the virus spreads and then declines after initiating the immune response. In this project, we propose to compare SIMCoV results and Computed Tomography (CT) scans from COVID patients. Based on the CT scan analysis of infection spread, lung damage, and severity of the disease, the idea is to generate simulation scenarios where SIMCoV can show similar spatial features in the result. This comparison would let us work out an explanation for the initial conditions of the viral dynamics for which we can see different levels of lung damage in the patients. The goal is to identify the initial conditions and viral and immune dynamics to parameterize SIMCoV to

match the lung damage in CT scans. If we successfully establish this relationship, SIMCoV can be extended to predict lung damage and severity in patients based on the initial analysis from their CT scans.

7.2 Introduction

Computed tomography (CT) scans of COVID-19 patients have been valuable tools for assessing and studying SARS-CoV-2 infection since the beginning of the COVID-19 pandemic. CT scans of SARS-CoV-2 infection are characterized by multi-focal distribution of lesions, particularly Ground Glass Opacities (GGOs) [15, 51] and consolidations [82], both of which are likely to indicate tissue damage caused by inflammatory cell infiltration. Although SARS-CoV-2 infection is multi-focal, there has been little work on how spatial relationships impact SARS-CoV-2 infection in the lung. The spatial Immune Model of Coronavirus, or SIMCoV [137], is an agent-based, computational model of SARS-CoV-2 infection in the lung that compares favorably to existing ODE models of SARS-CoV-2 infection. In this work, we further explore SIMCoV as a tool to understand SARS-CoV-2 viral and immune dynamics by developing a methodology for characterizing spatial relationships in SARS-CoV-2 infection and comparing predictions from SIMCoV to CT scans of COVID-19-infected patients which show spatial heterogeneity in disease distribution in the lung. Using the spatially explicit SIMCoV model will allow us to use CT scans from patients to test the role of spatial spread of disease in COVID-19.

We study the complex dynamics of lung lesions, particularly Ground Glass Opacities (GGO), as observed in the CT scans shown in figure 7.1. We focus on the detailed analysis of growth rates of lung lesions throughout the disease using sequential CT scans. This analysis is facilitated by the SIMCoV [137] simulation that represents the spatial and temporal progression of lesions. Tracking changes over time provides a comprehensive understanding of lesion dynamics, important for predicting disease progression and studying treatment strategies.

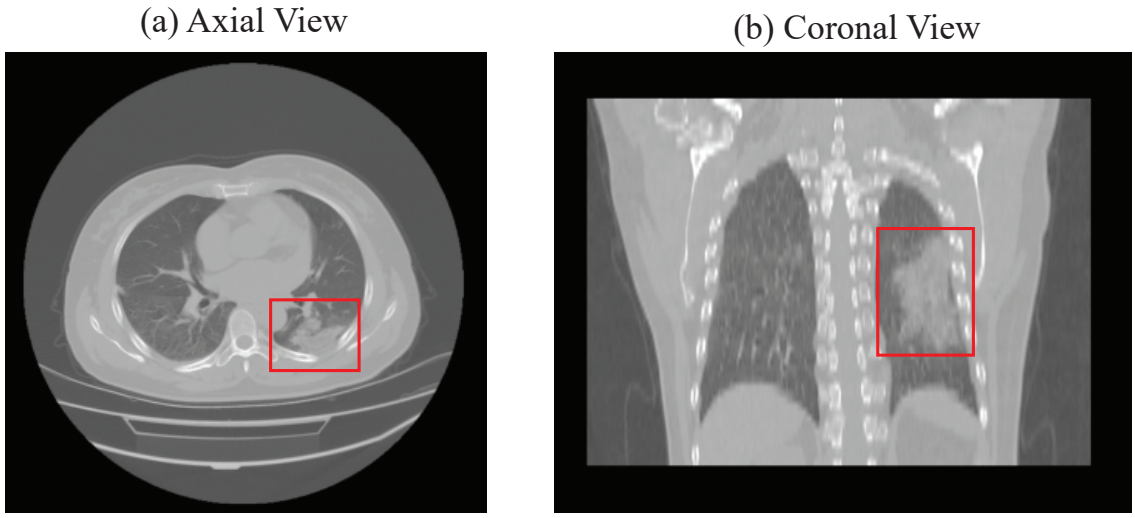


Figure 7.1: Chest CT scan from the dataset [98] showing Ground Glass Opacities (GGO) indicated by the red rectangle. GGOs appear as diffuse, foggy regions caused by the partial collapse of the alveolar sacs and partial filling with fluid [15, 51]. The complete filling of fluid in the alveolar spaces is termed consolidation [82]. (a) shows the axial view and (b) shows the coronal view of the scan.

This work proposes a novel simulation model called **"MultiSac"** using properties of SIMCoV [137] that represent the spatial structure of alveolar sacs within the lung, incorporating various cell types such as air, epithelial cells, and interstitial space. By assuming specific diffusion parameters for each cell type, we demonstrate how the lung's spatial architecture influences the spread and extent of damage. This modeling allows for a more fine-tuned understanding of disease mechanisms at the cellular level.

The work conducts a comparative analysis of lesion growth in patients with the spread of inflammatory signals of the SIMCoV simulations. This comparison highlights the correlation between physical lung damage and the underlying inflammatory processes. By aligning the growth patterns of lesions with those of inflammatory signals in SIMCoV, insights can be gathered about the infection and underlying immune response by analyzing the parameters.

The findings from the work have meaningful clinical and research implications. Understanding the growth patterns of GGOs [15, 51] and the factors influencing them can improve diagnostic accuracy, inform treatment decisions, and drive the development of therapeutic inter-

ventions. Our MultiSac model enables the exploration of various hypotheses about the effect of lung structure on the underlying immune response to infectious diseases like COVID-19. The next part of the work focuses on identifying the set of critical parameters that replicate both the viral and inflammatory dynamics observed in patients. These parameters are important in simulating the spatial growth of lung damage accurately. By establishing these parameters, the work enhances the predictive feature of the SIMCoV model given the early examples of CT scans from patients, making it a useful tool for both research and clinical applications. Overall these detailed contributions form the foundation of our research, providing the comprehensive framework for simulating infection in lung structure for understanding the progression of lesions and the impact of underlying cellular and inflammatory processes.

7.3 Patient Data Analysis

In this work, we aim to analyze and track the growth rates of lung lesions caused by SARS-CoV-2 throughout the disease. Therefore, we required the datasets containing patients who went through serial CT scan exams. We use the serial CT scan dataset from the research work [98] that explores the temporal relationships between chest CT scans and laboratory measurements in COVID-19 patients to understand disease progression and severity. The research involves 739 patients with confirmed COVID-19, of whom 29 underwent serial CT and laboratory tests over an average period of 50 days. The study employs both manual and AI-based segmentation to quantify lung opacities, including ground-glass opacities (GGO) [15,51] and consolidation [82]. The study provides important insights into the temporal onset of opacities; i.e. according to their data, lung opacities appeared approximately 3.4 days before symptom onset, with a peak occurring around the day of symptom onset [98]. The work also correlates laboratory results with CT findings. The study concludes that sequential CT and lab measurements provide valuable insights into the disease course of COVID-19 and may aid in early detection, prognostication, and clinical trial design. In our work, we second this conclusion while we also

take a step forward in hypothesizing the key insights about the disease and validating the facts with an agent-based simulation model implemented by factoring in the structure of the lung.

7.3.1 Lung and Lesion Identification and Volume Calculation

After CoVID-19, there have been lots of studies [74,103,132,157] about COVID detection using lung CT scans and chest X-Rays. Most of these works propose automated AI-based approaches for lung and lesion detection. Therefore, public datasets of CT scans [50, 123, 129] are available with segmented lungs and lesions. However, in most cases, these are not sequential scans which we require and also developing methods for efficient lung and lesion detection is out of scope for this study. However, we have used these two datasets [50, 129] with pre-segmented lungs and lesions to develop the volume calculation and visualization method (detailed analysis are shows in Figures A.1 and A.2). We use the sequential CT scan dataset of 29 patients from [98] for multiple days before and after symptom onset. The information about the patient demographics is detailed in the methods section.

From this dataset, we obtained the raw CT scans. One point to be noted here is that the days in which CTs were taken were not consistent among patients. Also, data is not available for every day post infection (DPI). For example: one patient has 2,4, 8 DPI CT scans while another one has 5, 24, 47 DPI. The results from [98], provide us with the total percentage of opacity (due to GGOs and consolidation) in the lung. They are referred to as "lung opacities". Since we want to analyze the spatial properties of the lesion in the lung structure and use a spatial model, we are interested in the total and individual volumes of the lesions observed in the patients. Therefore, the steps that we followed for our analysis are:

- **Lung Segmentation and Volume calculation**: The CT scans are available in the NIFTI format. We use the tool Slicer (Version 4.11) [2, 76] that has a lung analyzer module that segments the lung area and calculates the lung volumes from CT scans. The step-by-step tutorial can be found in [122].

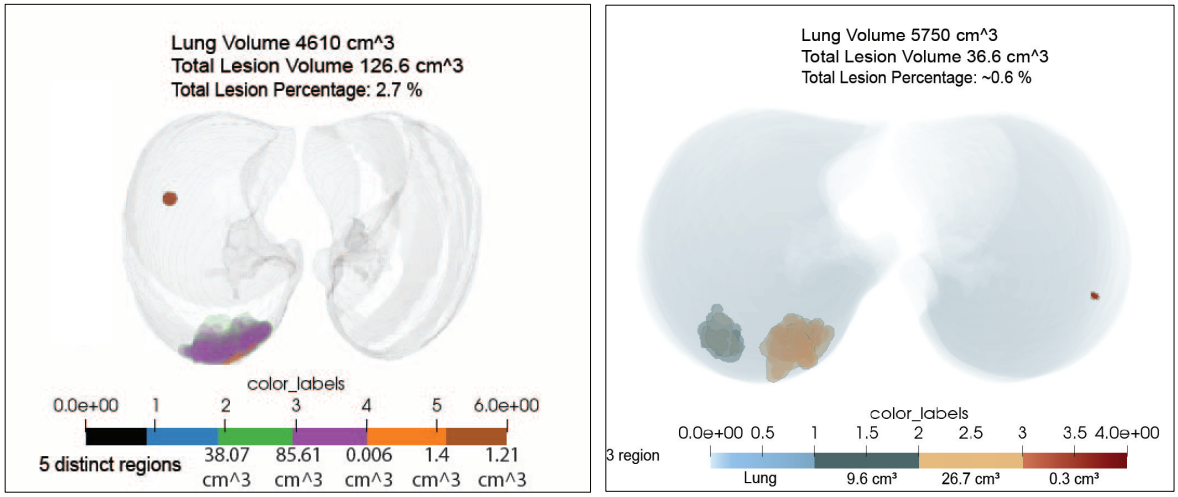


Figure 7.2: 3D spatial visualization of the infected lung with lesions from sample COVID patients from datasets [129] and [50]

- **Lesion Identification and volume calculation**: We have collaborated with radiologists from the UNM Radiology department for the CT scan interpretation. They helped us with the manual annotation of the lesions from the scans. The labeled scans were later segmented using computer vision-based approaches [102]. From the total lesion segmentation, we extracted the individual lesion using Python Library Connected Components 3D [156]. At the end of this step, we have individual lesion analyses from every patient at every timepoint.

- **Visualization**: After the extraction of lung and lesion segments, we overlay the lesions on the lungs using VTK library [170] and visualize them using Paraview [11]. All the calculation and visualization is done in 3D.

Figure 7.2 shows the visualization of COVID-infected lungs and lesions from pre-segmented datasets [50, 129]. Each lesion is labeled with a different color for identification and volume calculation. The first one or two color labels are identified as the lung.

Next, We follow this method for the sequential CT scans from [98]. As mentioned the CT scans average over 50 days pre and post-symptom onset. The patient-wise details about lung opacity are presented in spreadsheets in [174]. Since we are looking at the properties when

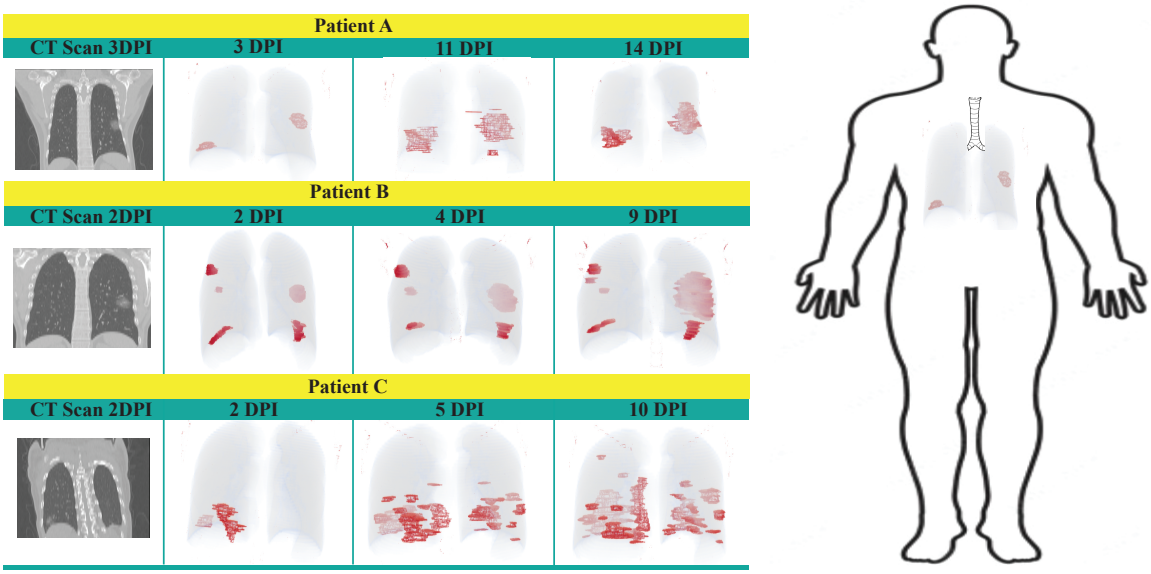


Figure 7.3: Left most panels show 2D slices of CT images. The other panels show regions of tissue damage in three example patients from dataset [98] at the estimated days post-infection/ DPI over the course of their disease. The lung is shown in gray and the lesions/lung damage is shown in red (Coronal View). The damages quantified demonstrate different growth rates of infection in areas of opacity. The human image is used to show the positioning of the lung in the coronal view of the lung CT scans.

COVID lesions are growing in the lung, we analyze the CT scans with a positive lesion growth rate. We have used 19 patients with 2 or 3 time points where positive growth of the lesions is observed. With our calculation and analysis, we had the following information about the 19 patients for each time point:

- The total volume of the lung.
- The total and individual volumes of the lesions.
- Percentage of the lung with lesions.
- Growth rate of the lesions from the previous time point.
- 3D Visualization of all the scans

The detailed results in spreadsheets and visualization are documented in [174].

Figure 7.3 illustrates an analysis of lung tissue damage in three patients over the course of their disease, using CT images. In the leftmost panels, 2D slices of the CT images are displayed. These slices provide a comprehensive view of the internal lung structures. The subsequent panels depict regions of tissue damage highlighted in red, against the lung tissue shown in gray, which represent the progression of the disease at various days post-infection (DPI). The visual representation emphasizes the temporal changes in the lungs' opacity, indicating areas of infection and damage. The figure also includes a coronal view of the human body to contextualize the positioning of the lungs within the chest cavity, aiding in the spatial understanding of the CT scans.

Like the sample three patients, each of the 19 patients in the dataset have different extents and patterns of lung damage observable over time (detailed analysis are available in [53]). This comparison highlights the heterogeneity in disease progression among individuals. The proposed visualization approach provides valuable insights into the temporal and spatial progression of lung damage in COVID-19 patients.

7.3.2 Lesion Growth Rate Analysis

The referenced work [98] provides lung opacity values for every CT scan, providing the percentage of total lesion volume divided by total lung volume. We calculated growth rates of the total lesion volume from this lung opacity data. We plotted the growth rates from our method, compared to the growth rate of the total percentage of lession in the original paper in a notched box plot shown in Figure 7.4(a). The box plot on the left represents the growth rate data from the reference study [98]. The median is below 50%, with a wider spread of data points and several outliers above 200%. The box plot on the right represents the growth rate data from our method that estimates the individual volume of each lesion at each time point. The median is similar to the reference method but slightly lower, and the data points are more tightly clustered, indicating less variability. There are fewer outliers compared to the reference. We conducted a comparative analysis of the two approaches to confirm that, despite

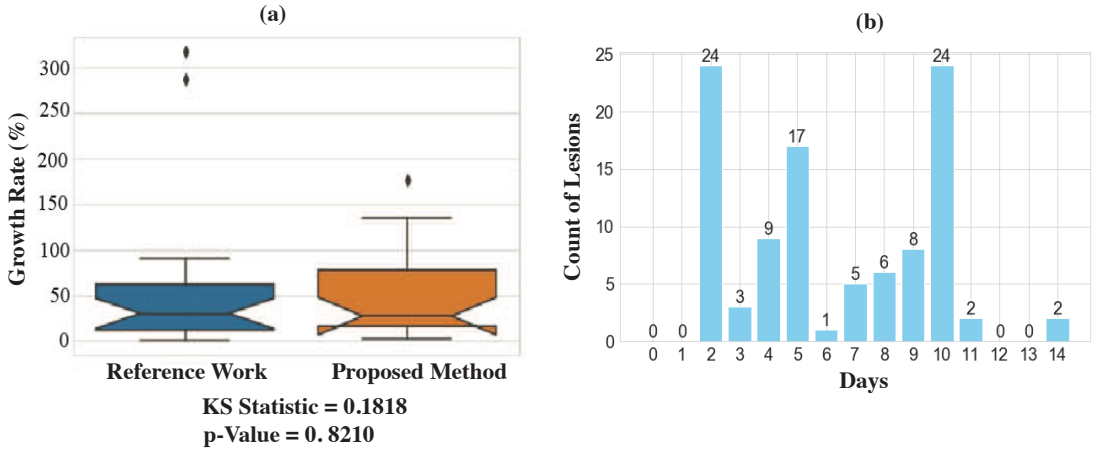


Figure 7.4: (a) shows notched box plots comparing the growth rates of lung tissue lesions between the reference work [98] and the proposed method. The y-axis represents the growth rate percentage, with the central notched line in each box indicating the median growth rate. The interquartile range (IQR) is represented by the width of each box, and the whiskers indicate the range of the data excluding outliers. Outliers are displayed as individual points. The Kolmogorov-Smirnov (KS) [58] test was applied to assess the difference between the two methods, resulting in a KS statistic of 0.1818 and a p-value of 0.8210, indicating no significant difference between the distributions of the two methods. (b) shows individual lesion counts over days for the 19 patients used in the proposed method.

employing different methodologies, our proposed method yielded results consistent with those of the reference study.

To prove that statistically the results come from the same distribution we performed the Kolmogorov-Smirnov (KS) test [58]. KS test is a non-parametric statistical test used to compare two samples or a sample with a reference probability distribution. The test evaluates the null hypothesis that the two samples are drawn from the same continuous distribution. The KS statistic is the maximum distance between the empirical cumulative distribution functions of the two samples. The p-value determines the significance of the observed difference. For the provided box plot comparing the growth rates between the reference and the proposed study, the KS statistic is calculated as 0.1818 with a p-value of 0.8210. This is a low KS statistic value. A higher KS statistic would indicate a larger difference between the distributions. The p-value is significantly higher than the common significance level of 0.05, therefore we fail to reject the null hypothesis. The KS test results support the visual observations from the box plot. The

small KS statistic and high p-value suggest that there is no significant difference between the growth rate distributions of the two approaches.

For this study, we selected 19 patients, each with 2 or 3 time points showing a positive total growth rate of lesions. Initially, we conducted a patient-wise analysis, but due to the inconsistency of time points across patients, comparative analysis proved challenging. Additionally, there were missing days without any lesion information or examples for a range of days from only a single patient. Our objective is to calculate a daily growth rate of the lesions in CT scans and compare them to SIMCoV simulations. Therefore, we changed our approach to focus on characterizing individual lesions over multiple days across patients, considering only those lesions that remained distinct and whose growth rates could be accurately tracked. A total of 46 lesions met these criteria, and the frequency of these lesions over days is summarized in Figure 7.4(b). Furthermore, our current simulation capabilities do not extend to simulating the entire lung volume, which ranges from 4000 to 6000 cm^3 (4 to 6 liters) [216]. We can simulate only a small portion of the lung ($\approx$ 4 cm³) containing 1 or 2 individual lesions. Consequently, a lesion-wise analysis is more appropriate for our comparative study.

From Figure 7.4(b), we observe that there is no lesion data available for days 0, 1, 12, and 13. To address this, we estimated the volume for these missing days using interpolation and Gaussian smoothing. We opted to use the median volume rather than the mean volume from the patient data for days with multiple volume instances, as the median is less affected by outliers and skewed data, providing a more robust measure of central tendency. After interpolation, we smoothed the data as it helps to reduce noise and create a more continuous and realistic representation of the volume data over time. There are some assumptions that we proposed for interpreting and analyzing the data and simulation scenarios.

- There is data from only 1 patient with 2 lesions for days 11 and 14 (Patient lesion analysis shown in Figure 7.3 Patient A). We will be simulating this case separately and not including it in this analysis.

- Day 6 data is also from 1 patient and it reflects a peak in the lesion volume which skews

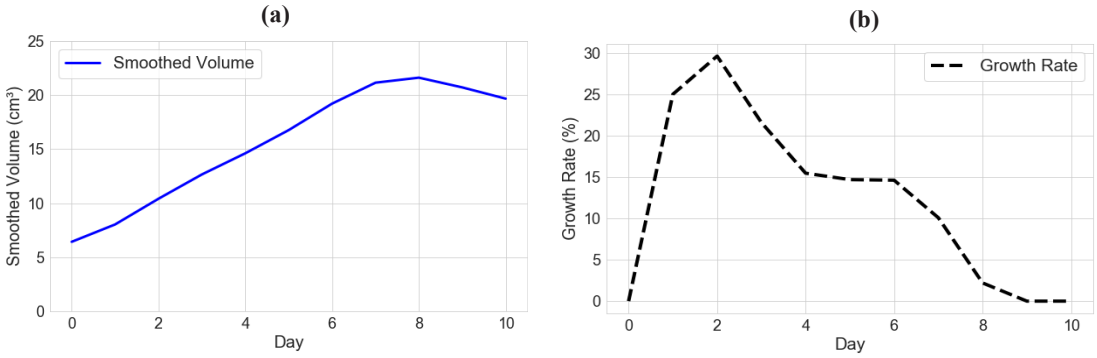


Figure 7.5: (a) Line plot showing the smoothed volume of lung lesions over days post-infection. The x-axis represents the days, while the y-axis shows the smoothed volume in cubic centimeters (cm^3), starting from 0. The smoothing was achieved through Gaussian smoothing applied to interpolated median volumes, providing a continuous representation of lesion volume changes over time. (b) Line plot showing the growth rate percentage of lung lesions over days post-infection calculated from the volume plot. The x-axis is the days, and the y-axis represents the growth rate percentage. The growth rate shows the dynamic progression of lesion growth during the infection period.

the data and growth rate. This can be either an artifact or patient-specific scenario. We will be simulating this case separately as well.

- Several studies have documented that the peak viral load of SARS-CoV-2 occurs approximately 7-10 days post-infection. For instance, research has shown that the highest viral load in the respiratory tract is typically reached within the first week of symptom onset and gradually declines thereafter [130]. This pattern indicates that patients have a positive lesion growth rate during this early period, aligning with similar findings across various studies on viral load dynamics and transmission potential. Therefore, we are including results until day 10 and interpolating volumes at days 0, 1, and 6.

The results from this analysis are shown in Figure 7.5. (a) shows the smoothed volume of lung lesions over days post-infection across all patients. Gaussian smoothing was applied to the interpolated median volumes from all patients to reduce noise and provide a clearer trend of volume changes over time. The volume starts at a lower value and increases steadily, peaking around day 8 before slightly declining by day 10. This suggests a general increase in

lesion size over time, with a reduction towards the later days when the body's immune response kicks in. (b) shows the line plot showing the growth rate percentage of lung lesions over days post-infection calculated from the smoothed volume. The growth rate starts at 0%, rapidly increases to a peak around day 2 at approximately 30%, and then gradually decreases. This indicates that the rate of lesion growth is highest early in the infection and slows down over time. The plots together provide a comprehensive view of the lesion dynamics during the infection period across the patients. The initial rapid growth phase followed by stabilization or reduction in volume suggests that the lesions expand quickly initially but then start to decrease as the infection spreads through lung and later the body's immune response becomes effective. The growth rate plot of the patients will be used in the next sections to compare and explain with SIMCoV simulations.

7.4 MultiSac Model in SIMCoV Simulation

Spatial Immune Model of Coronavirus Infection (SIMCoV) [137] is a computer-simulated agent-based model developed to study SARS-CoV-2 infection by analyzing the spatial distribution of infected cells and immune response in patients. The details about SIMCoV is discussed in detail in the introduction section 1.3.

In the original model, SIMCoV is initialized with the SARS-CoV-2 virus that can infect lung epithelial cells. In the experiments, epithelial cells are modeled as a 2D grid or 3D grid, where each grid point represents a $5 \times 5 \times 5\mu m^3$ volume. Space is represented by a discrete Cartesian grid. Grid points are spaced five microns apart (roughly the diameter of a T cell), and components of the model occur only at these discrete locations. The model is run as a discrete-time simulation, where each time step represents one minute, approximately the time it takes for a T cell to move five microns (one grid point) [121, 137, 139]. There are four main components of SIMCoV: epithelial cells, CD8+ T cells, virions, and inflammatory signals. The 2D or 3D layer of epithelial cells is susceptible to infection with virions. After a while, T

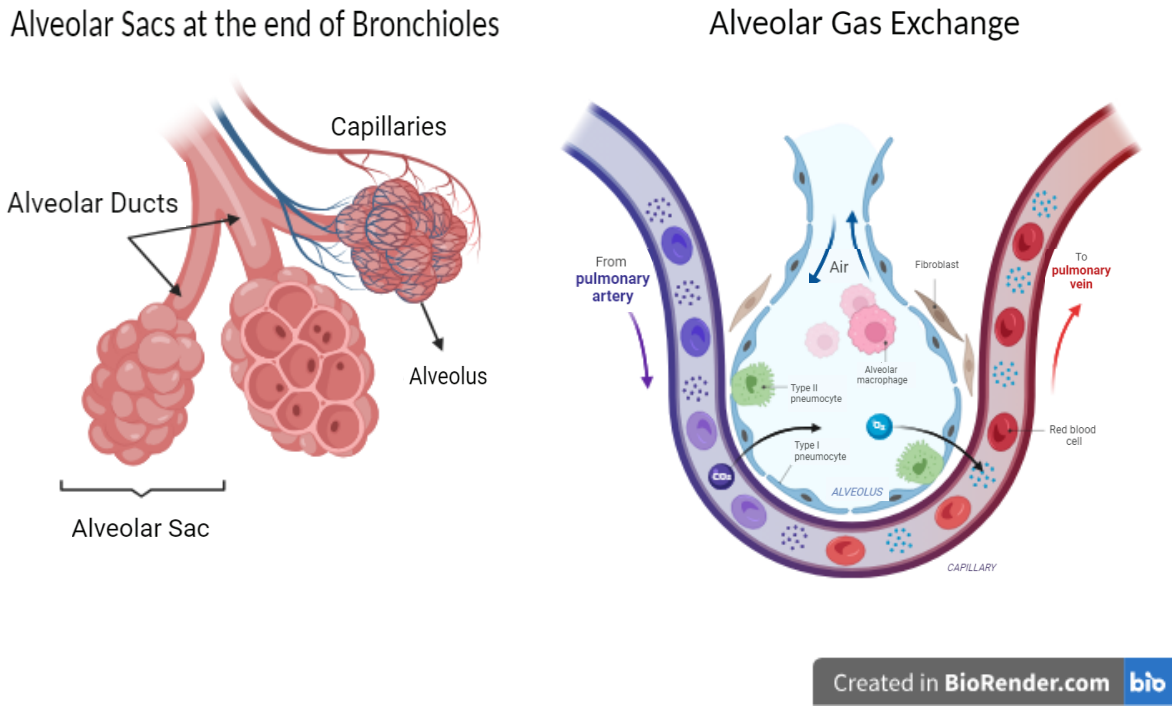


Figure 7.6: The Left figure shows the structure of alveolar sacs at the end of bronchioles. Alveolar sacs are composed of multiple grape-like alveoli, each surrounded by a network of capillaries. The alveolar ducts connect the alveoli to the bronchioles, enabling airflow into the alveoli. The right figure shows the process of alveolar gas exchange. Oxygen from inhaled air diffuses through the alveolar walls into the blood in the capillaries, while carbon dioxide from the blood diffuses into the alveoli to be exhaled. Different alveolar cells, including type I pneumocytes and type II pneumocytes are highlighted. The figure illustrates the role of alveolar sacs in respiratory function. This figure is created using [20]

cells arrive to clear the infection reciprocating the body's immune response. When infection takes place, inflammatory signals are produced which correspond to cytokines that cause T cell extravasation into the lung tissue. In our analysis, we are comparing the lesion inflammation from the CT scans with the component inflammatory signal in SIMCov.

To enhance the biological relevance of the existing SIMCoV model, we propose including the spatial structure of alveolar sacs. This integration aims to provide a more accurate representation of lung anatomy and physiology, which would improve the model's effectiveness in simulating spatial disease dynamics.

The alveolar sacs are important components of the respiratory system, playing a central role in gas exchange [150]. These tiny air sacs are located at the end of the bronchioles in the lungs and are clustered together in structures known as alveolar sacs. Figure 7.6 created using [20] illustrates the detailed anatomy and function of these structures. Each alveolar sac consists of multiple alveoli (single alveolus), which are small, balloon-like structures filled with air [204, 205]. The alveoli are interconnected and are surrounded by a network of capillaries. This proximity facilitates efficient gas exchange between the air in the alveoli and the blood in the capillaries. Alveolar ducts lead into the alveolar sacs, acting as channels that allow air to flow into the alveoli [150, 204]. A dense network of capillaries surrounds each alveolus. The thin walls of the capillaries and alveoli allow for the rapid exchange of gases. As shown in Figure 7.6, oxygen from inhaled air diffuses through the walls of the alveoli and into the blood in the capillaries. Simultaneously, carbon dioxide from the blood diffuses into the alveoli to be exhaled [189].

During an infection, the alveoli in the lung become inflamed and filled with fluid or pus, resulting in the white shadows observed on CT scans. Therefore, we aim to model the alveolar sac to capture this pathological condition in the simulation. The structures and cells that we want to model in the simulation are:

- Air: Each alveolus is filled with air.
- Alveolar Epithelium: This forms the outer layer of the sac. There are 2 types of cells here:
  - Type I Pneumocytes: They form 95% of the alveolar surface and are responsible for the gas exchange [150, 204]. These cells are non-infectable.
  - Type II Pneumocytes: 5% of the surface are formed with these infectable cells.
- Inside of the sac has multiple alveolar structures. Each alveoli is like a grape with air inside and has alveolar epithelium outside [150].

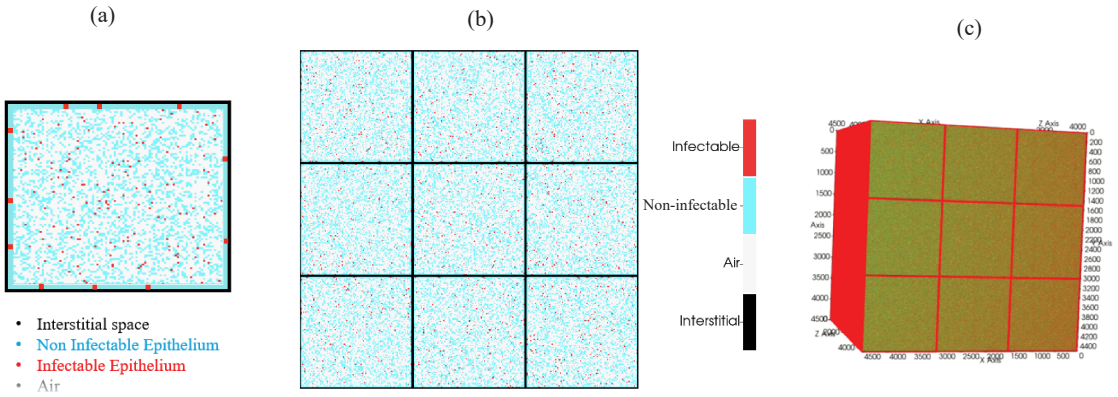


Figure 7.7: (a) shows the structure of one alveolar sac which 4 different cell types: Interstitial space, infectable epithelium, non-infectable epithelium, and air. (b) and (c) illustrate the representation of alveolar sacs in a simulation grid of $300 \times 300 \times 300$ voxels ($4500\mu m \times 4500\mu m \times 4500\mu m$), equating to approximately $0.1 cm^3$ in volume. (b) is a 2D slice of the simulation showing the spatial distribution of cells within the simulation. (c) is a 3D representation of the simulation presenting 27 alveolar sacs.

- The space between the alveolar sacs is called interstitial space in this work.

7.4.1 Structure of the Proposed Multisac Model

Integrating the spatial structure of alveolar sacs into the SIMCoV model enhances the biological accuracy and relevance of these simulations. Given that SIMCoV operates on a grid model, we assume alveolar sacs to be cube-like rather than sphere-like for the ease of implementation. Figure 7.7 (a) illustrates the 2D structure of an alveolar sac. From the outermost to the innermost layers, the first layer, shown in black, represents the interstitial space. This is followed by the alveolar epithelium layer, which consists of 95% non-infectable Type I Pneumocytes (blue) and 5% infectable Type II Pneumocytes (red). The innermost layer contains multiple alveoli within a sac, each resembling a grape with air inside and an alveolar epithelium outside. Modeling the fine details of this inner space is complex because some alveoli may be squished together without distinct layers. Therefore, we use the air, infectable, and non-infectable epithelium distribution in this layer based on the number of alveoli, rather than attempting to represent each individual alveolus precisely.

For the Multisac model and simulation, we updated the cell types from the generic SIMCoV epithelial cells with a diameter of 5μm to alveolar cells with a diameter of 15μm. This change effectively reduced the spatial and temporal resolution by a factor of three while increasing the simulation dynamics by three-fold compared to the previous version of SIMCoV [137]. Given this single-cell-to-single-cell conversion, we assume that our biological parameters remain valid when applied to the larger cell size, allowing each grid point to represent 15μm. Each alveolus is modeled with a diameter of 300μm [150, 216], and we assume each side of the alveolar sac is 1500μm, resulting in 100 grid points across each side of the sac in our model. Consequently, each sac (3D) contains approximately 125 alveoli. Figure 7.7 (b) and (c) illustrate the representation of alveolar sacs in a simulation grid of $300 \times 300 \times 300$ voxels ($4500\mu m \times 4500\mu m \times 4500\mu m$), equating to approximately $0.1 cm^3$ in volume. Figure 7.7 (b) shows a 2D slice of the simulation, where each sac is separated by layers of interstitial space depicted in black. (c) presents a 3D representation of 27 alveolar sacs, highlighting the spatial organization and distribution within the simulation.

7.4.2 Effects of the Proposed Multisac Model

In the previous version of SIMCoV, there were four main components: epithelial cells, T cells, virions, and inflammatory signals. In our proposed MultiSac model, the epithelial cells have been expanded into three distinct types: air, epithelial, and interstitial. This necessitates the definition of new virion diffusion parameters for each specific cell type. Virion diffusion refers to the fraction of virions that diffuse into all neighboring grid points per minute [137]. Given the change in cell size for each grid point from 5μm to 15μm, the parameter dynamics of the components have been adjusted three-fold based on the properties of the corresponding component. Specifically, probabilistic parameters have increased three-fold due to the larger cell size, while spatial and temporal parameters have decreased three-fold to reflect the faster dynamics within the larger volume.

Based on the default COVID-19 parameter values from SIMCoV [137] Table 2, the cor-

responding changes in key parameters for the MultiSac model are summarized in Table 7.1. Additionally, new virion diffusion values for air and interstitial cell types are introduced in the table and highlighted in yellow to indicate their specific adjustments for the new cell types. For instance, virions are expected to diffuse faster in the air and slower in the interstitial space. The air diffusion parameter is set to 1.0, representing the maximum possible fraction, and is 20 times higher than the default epithelial diffusion value of 0.05. Conversely, interstitial diffusion is reduced by a factor of five, set at 0.01, reflecting the slower diffusion rate in this denser tissue type. These values were selected based on the different experimental runs and comparison with the patient scenario. This approach ensures that the MultiSac model accurately represents the biological processes and dynamics at the new scale, maintaining the integrity and relevance of the simulation outcomes.

Parameters	Occurrence	default SIMCoV	MultiSac Model	Calculation
Incubation Period	num ts	480	160	x(1/3)
Apoptosis Period	num ts	180	60	x(1/3)
Expressing Period	num ts	900	300	x(1/3)
Infectivity	per ts	0.001	0.003	Probability t.f. x(3)
Virion Production	per ts	1.1	3.3	Additive t.f. x(3)
Virion Clearance	per ts	0.004	0.01195	$1-(1-x)^3$
Virion Diffusion Epithelial	per ts	0.15	0.05	Einstein-Smoluchowski with x=15 and t=180
Virion Diffusion Air	per ts	N/A	1.0	N/A
Virion Diffusion Interstitial	per ts	N/A	0.01	N/A
Inflammatory Signal Production	per ts	1.0	1.0	No change (max cell inflammation is 1.0)
Inflammatory Signal Decay	per ts	0.01	0.0297	$1-(1-x)^3$
Inflammatory Signal Diffusion	per ts	1.0	1.0	No change (max cell inflammation is 1.0)
Antibody Period	num ts	5760	1920	x(1/3)
Tcell Generation Rate	per ts	105000	315000	Additive t.f. x(3)
Tcell Initial Delay	num ts	10080	3360	x(1/3)
Tcell Vascular Period	num ts	5760	1920	x(1/3)
Tcell Tissue Period	num ts	1440	480	x(1/3)
Tcell Binding Period	num ts	10	3	x(1/3)

Table 7.1: Changes in the key parameters for MultiSac Model

Comparative Analysis with Default SIMCoV

To observe the effects of the MultiSac SIMCoV Model compared to the default SIMCoV (2D and 3D), we analyzed and compared the inflammatory signal count and growth rate across the three models. The results of these comparisons are presented in Figure 7.8. This analysis allows us to understand the impact of introducing the spatial structure of alveolar sacs on

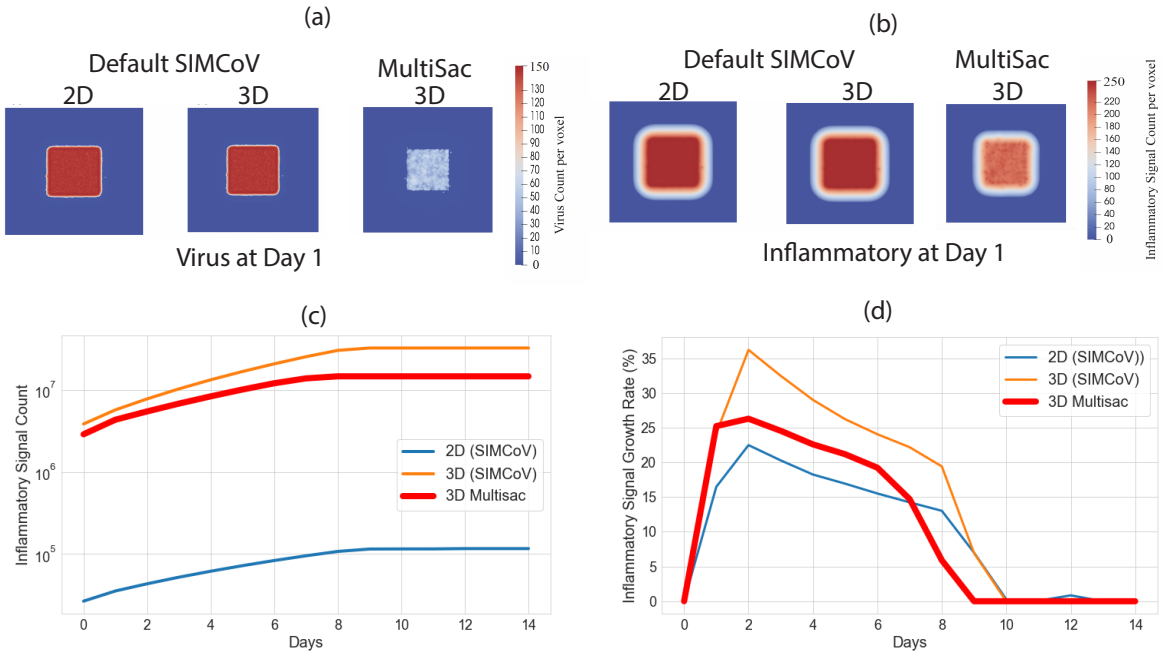


Figure 7.8: (a) and (b) shows the spatial position of the virus and inflammatory signal respectively on day 1 of the simulations (sliced as 2D for presentation purposes) for three different models: 2D default SIMCoV, 3D default SIMCoV, and the 3D MultiSac SIMCoV model.(c) presents the inflammatory signal count and (d) shows the corresponding growth rates over 14 days across three different models: 2D default SIMCoV, 3D default SIMCoV, and the 3D MultiSac SIMCoV model

the dynamics of inflammatory signals, which we use as a proxy for the lesions in CT scans. The simulations were run for $1000 \times 1000 \times 1000$ voxels ($15000\mu m \times 15000\mu m \times 15000\mu m$), equating to approximately $3.3cm^3$ in volume. The simulations were initialized by infecting the volume of one alveolar sac. In the default SIMCoV simulation, all cells are infectable, resulting in the infection of all voxels. This scenario is observed in Figure 7.8 (a), which shows the spatial position of the virus on day 1 of the simulations (sliced as 2D for presentation purposes).

In the MultiSac case, only the infectable epithelial cells, equivalent to type II pneumocytes, are infected. Figure 7.8 (b) shows the spatial positioning of the inflammatory signal on day 1. We calculated the counts of the inflammatory signals, which are compared as the equivalent of inflammation or lesions observed in CT scans. This comparison highlights the differences in infection patterns and inflammatory responses between the default and MultiSac models.

Figure 7.8 (c) presents the inflammatory signal count over 14 days across three different models: 2D default SIMCoV, 3D default SIMCoV, and the 3D MultiSac SIMCoV model. For the 2D default model, the inflammatory signal count starts low and increases gradually over time, reaching a plateau around day 10. The overall count is significantly lower compared to the 3D models. This is likely due to the limited spatial representation in 2D, which restricts the simulation of the diffusion processes. In case of the 3D default model, the inflammatory signal count rises sharply, with a slight plateauing towards the end. The count is substantially higher than in the 2D model, reflecting the enhanced spatial interaction dynamics that 3D simulations can capture. In the 3D MultiSac model, the count is close to that of the 3D SIMCoV model but slightly lower. This is attributed to the presentation of alveolar structures, where only specific cell types (type II pneumocytes) are infectable, leading to the containment of infection spread.

Figure 7.8 (d) shows the inflammatory signal growth based on the count in (c) across the three models. The trend is similar to what we have discussed for (c). All the models peak around day 2 and then gradually decrease leveling off near zero around day 9/10 as the T cells arrive at Day 7. As observed in (c), the 3D default model has the highest growth rate, the 3D multisac is slightly lower but follows a similar trend. The 2D model has the lowest which is expected.

Comparing these models reflects that, the Multisac model restricts infection to specific cell types. We can propose that the MultiSac model provides a more accurate representation of how inflammatory signals propagate and stabilize over time. In the CT scans, lesions appear as discrete patches rather than a diffuse pattern. The MultiSac model simulates localized infection and varying diffusion rates mirror the clinical reality of patchiness from respiratory infections.

Distribution of Cells in the MultiSac Model

We have conducted experiments to determine and test the distribution of cells in the model that is most biologically relevant. The structure and cell types of the MultiSac Model are shown in Figure 7.9 (a). We also experimented to see the effect on inflammatory signal growth rates if

the sac structure was not present in the model. In this case, infectable cells were distributed throughout the entire simulation with the same number and density as in the sac structure. This scenario is visualized in Figure 7.9 (b), presenting two cases:

- Distributed Structure (epithelial): Infectable cells are surrounded by non-infectable epithelial cells. The virion diffusion rate for epithelial cells is 0.05.
- Distributed Structure (air): Infectable cells are surrounded by air. The virion diffusion rate for air is 1.0.

In both cases, the diffusion rate for infectable cells remains 0.05. These variations help us understand the impact of different surrounding environments on the diffusion and spread of infectable cells, providing insights into how the presence or absence of the sac structure with interstitial space affects inflammatory signal growth rates. For this experiment also, the simulations were run for $1000 \times 1000 \times 1000$ voxels (15000µm $\times$ 15000µm $\times$ 15000µm), equating to approximately 3.3cm^3 in volume. The simulations were initialized by infecting the volume of one alveolar sac.

Figure 7.9 (c) and (d) consists of two line plots comparing the inflammatory signal count and growth rate over 14 days across three different scenarios: Distributed (Epithelial), Distributed (Air), and MultiSac Structure.

In the case of the distributed structure (epithelial), the inflammatory signal count and growth rise steadily but at a slower rate compared to the other two scenarios. This scenario represents infectable cells distributed throughout the simulation space surrounded by non-infectable epithelial cells. The slower increase in inflammatory signals can be attributed to the same diffusion rate of virions in both infectable and non-infectable epithelial cells (0.05), which restricts the spread of infection. There is no presence of air as well.

In the case of the distributed structure (air), the inflammatory signal count and growth increase rapidly in the initial days, and then stabilizing due to T cell arrival. In this case, infectable cells are surrounded by air, which has a higher virion diffusion rate (1.0). This

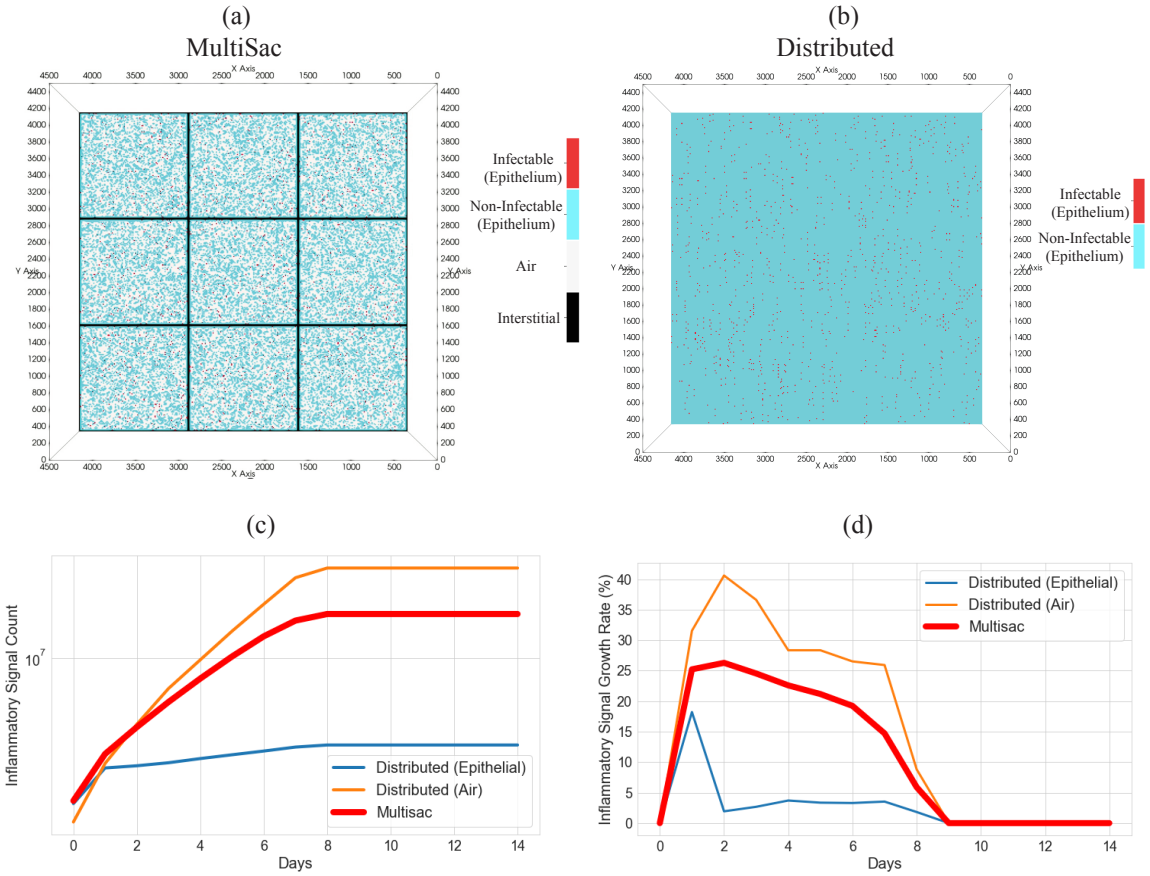


Figure 7.9: (a) shows the MultiSac structure and (b) shows the distributed cell structure. (c) presents a line plot of the inflammatory signal count and (d) plot shows the corresponding growth rates over 14 days across three different models: MultiSac structure, distributed cell structure with epithelial cells, and distributed cell structure with air.

facilitates faster spread of virions, leading to a quicker rise in inflammatory signals. However, this diffuses quickly and there is a steep decline in the growth rate later meaning the virus and inflammation also clear out too quickly.

In the Multisac structure, the distribution of infectable, non infectable epithelial cells and air balances the spread in the sac. Also, the interstitial space confines the spread of infection to specific regions. This results in a rapid initial increase in inflammatory signals, followed by stabilization as the infection becomes contained within the sacs. This is a controlled structure that captures and explains both rapid initial spread and effective containment, aligning with observed patterns of lesion distribution in clinical CT scans. The MultiSac model provides a better understanding of why lesions are not seen uniformly across the lung in CT scans.

The experiments presented in this section highlight the importance of spatial structure in the simulation model to track disease progression and inflammatory responses.

7.5 Comparing MultiSac SIMCoV with Patient Analysis

Understanding the dynamics of inflammation or lesion growth in lung infections is crucial for developing effective treatments and interventions. The MultiSac SIMCoV model, which includes a detailed representation of alveolar sacs, provides a sophisticated tool for simulating these dynamics. To validate the accuracy of the model, we compare the inflammatory signal growth rate generated by the MultiSac SIMCoV model with the lesion growth rate from the patients (Figure 7.5 (b)) discussed in section 7.3. Figure 7.10 presents this comparison over 10 days.

The MultiSac SIMCoV model extends the default SIMCoV [137] framework by incorporating the spatial structure of alveolar sacs for enhanced biological relevance. In this model, the epithelial cells are differentiated into four types: air, epithelial (infectable and non-infectable), and interstitial space cells. Only specific cell types (type II pneumocytes) are infectable, reflecting more accurate biological processes. Virion diffusion rates are adjusted based on different

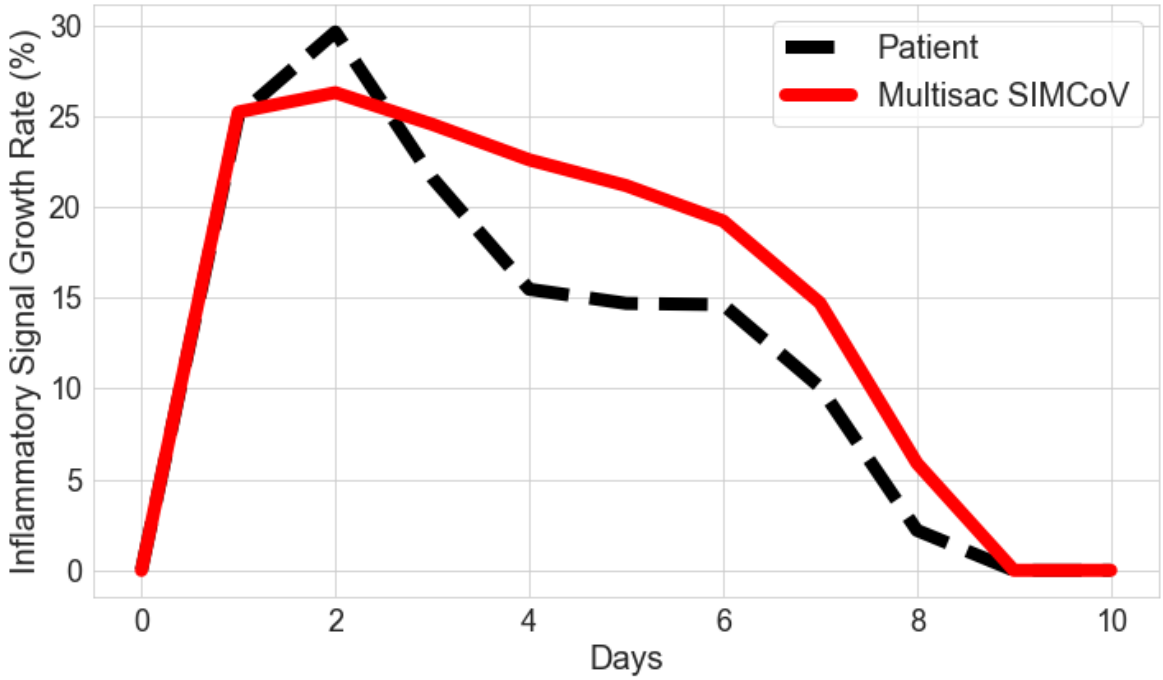


Figure 7.10: Comparison between the inflammatory signal growth rate of the MultiSac SIMCoV model (red line) and SARS-CoV-2 Patient data (black dashed line)

cell types and environments, with faster diffusion in air and slower in epithelial and interstitial spaces. The simulation used in this analysis uses the parameters mentioned in Table 7.1.

The patient data used for comparison reflects the inflammatory lesion growth rate observed in CT scans in clinical settings. This data provides a real-world benchmark to assess the model's performance and accuracy. The growth rate in patients typically shows a rapid initial increase followed by a gradual decline as the infection spreads through space which is the lung tissue and finally zeros after body's immune response kicks in.

Figure 7.10 compares the inflammatory signal growth rate between the MultiSac SIMCoV model (red line) and COVID-19 patient data (black dashed line). Both the MultiSac model and patient data show a sharp increase in the inflammatory signal growth rate, peaking around day 2. The initial growth rate in the MultiSac model closely follows the patient data, indicating that the model accurately captures the early infection dynamics. After peaking, both curves exhibit a decline in the growth rate . The MultiSac model shows a slightly smoother decline compared

to the patient data. This highlights the model's ability to replicate the peak infection period and the subsequent containment of the spread of infection because of the structure. The real patient data shows more variability, possibly due to individual patient differences in immune response and disease progression. The model can also simulate the scenarios if parameters especially the virion diffusion rates are varied. The decrease in lesion growth after day 2 indicates the containment of infection and the reduction of new inflammatory signals because of the multiple sac structure until the body's immune response kicks in.

In the MultiSac Model, T cells arrive at day 7 which is reflected in the declining growth after day 7 and zeroing around day 9. We see a similar trend in the patient data as well. The close alignment between the model and patient data in this phase further validates the model's effectiveness in handling immune response dynamics.

The comparative analysis demonstrates that the MultiSac SIMCoV model effectively simulates the key phases of inflammation growth observed in patients:

- Accuracy in early infection: The model's close relevance with patient data during the initial growth phase suggests that it accurately captures the early spread of infection.
- Realistic peak and infection dynamics: The model's ability to get closer to the peak and subsequent decline in growth rate reflects its sophisticated handling of infection dynamics. The growth decline phase between the model and patient data indicates that the MultiSac model can effectively simulate the containment and control of inflammation,
- Immune response: The timing of T cell arrival in the MultiSac model can handle the immune response initiation scenario in actual patients.

Using only the default parameters originally fit to viral load data from a different set of patients [137], the MultiSac Model can successfully replicate the average median case of lesion growth across a new set of patients. The slight differences observed in the comparative analysis can be handled by further refinements, such as incorporating more variability in diffusion and immune response parameters to reflect individual patient scenarios. The MultiSac SIMCoV

model provides a biologically plausible explanation of key factors that govern lesion growth in the lung caused by SARS-CoV-2 infection. By comparing the growth of inflammation to the growth of lesions in patient CT scans, we validate that SIMCoV captures key features of spatial structure and immune dynamics that control the spread of damage in the lung. This increased understanding of how damage spreads in the lung has the potential to aid in the development of future treatments and interventions. By capturing the underlying dynamics of SARS-CoV-2 infection and immune response, the MultiSac model represents a significant advance in the simulation and analysis of COVID-19 infections.

7.6 Discussion and Next Steps

The MultiSac SIMCoV model introduces a detailed representation of alveolar structures, which significantly influences the rate of spread and growth of lung damage caused by SARS-CoV-2. The structure of alveolar sacs introduced in this model is validated to be efficient in simulating how infection spreads within lung tissue. The growth rate of lung damage observed in the MultiSac SIMCoV model closely mirrors the growth rates seen in patient CT scans, emphasizing the model's biological relevance and accuracy.

The structured alveolar model in MultiSac SIMCoV restricts the spread of infection to specific cell types (type II pneumocytes), resulting in a controlled and realistic simulation of lung damage progression. The similarity of growth rates between the MultiSac model and patient CT scans indicates that the model successfully replicates the dynamics of lung infection and the inflammatory response observed in clinical settings. The analysis suggests that the inherent structure of lung tissue, particularly the distribution and density of alveolar sacs, plays an important role in controlling the spread and growth of lung damage in COVID-19 patients. This insight highlights the importance of accurately modeling and incorporating lung anatomy in simualtions to understand disease progression better.

To further refine and validate the MultiSac SIMCoV model, sensitivity analysis to parameters

using calibration tools like Calipro [144] will be essential. This analysis will help identify key parameters that significantly impact the model's output and allow for fine-tuning to enhance accuracy and predictive capability.

Creating individual patient scenarios based on specific patient analysis will enable personalized modeling of disease progression. We already have the individual analysis of 19 patients [53]. This approach will help capture the variability in immune responses and infection dynamics among different patients, providing tailored treatment strategies.

Observations from the patient analysis suggest that lesions often appear on the periphery of the lung. It raises several hypotheses that we want to explore:

- Immune response slow at the periphery: It is possible that the immune response is slower or less effective at the lung periphery, allowing the virus to proliferate more readily in these regions.

- Virus spreads fast on the periphery: Alternatively, the virus might spread more rapidly along the periphery, potentially due to structural factors or differences in tissue composition that facilitate quicker viral movement.

- Structural influence: Another hypothesis is that certain structural elements, such as a higher density of alveoli or the presence of impermeable boundaries, may influence the spread pattern. Further investigation into the lung's microanatomy could reveal whether these structural factors play a significant role in peripheral lesion formation.

By addressing these future work areas, we aim to enhance the MultiSac SIMCoV model's predictive ability and biological relevance, contributing to more effective and personalized treatments for infections caused by SARS-CoV-2.

7.7 Acknowledgements

I would like to thank and acknowledge George Matthew Fricke and his team at the UNM Center for Advanced Research Computing (CARC) for their support and assistance in setting up the SIMCoV framework and providing resources for large-scale simulation. I thank our collaborators Professor Melanie Moses, Professor Judy Cannon, Akil Andrews from UNM, Professor Stephanie Forrest and Kirtus Leyba from Arizona State University (ASU), Steven Hofmeyr from Lawrence Berkeley National Lab (LBL), and Ronak Etemadpour and Hossein Mehdikhani Karimabad from the UNM radiology department. I would like to thank LBL for partial funding and for providing HPC resources for this work.

Chapter 8

Pieces to Patterns: Discussions and Future Work

Information is a difference that makes a difference

– Gregory Bateson

The study and analysis of complex systems is a multidisciplinary field that encircles various domains such as physics, biology, economics, and engineering. One powerful approach to understanding these systems is through the use of information-theoretic measures. These measures provide a framework for quantifying the information content and the relationships between different components of the system over space and time. In this dissertation work, we have established how these measures can be utilized to uncover the most relevant components of complex systems by analyzing their spatial and temporal relationships. We have also established that regardless of application fields the information theory approaches can be used for analysis.

This work proposes a novel development and application of the measure Normalized Mutual information (NMI) to quantify how different cells, in lymph nodes interact spatially and temporally within the immune system. It plays a significant role in understanding the process of immune response and T cell activation in the body. NMI, along with Specific

Mutual Information (SMI), provides robust tools for capturing complex interactions within biological datasets. These methods quantify the amount of shared information between variables, making them particularly useful in identifying and analyzing spatial and temporal patterns in biomedical data. The ability to quantify such interactions is essential in systems biology, where understanding the network of interactions among genes, proteins, and cells can lead to insights into disease mechanisms and potential therapeutic targets.

This work demonstrates that SMI measures are useful for understanding, highlighting, and tracking complex interactions across multiple domains, including the interaction of various atmospheric variables to understand the formation and progression of weather phenomena like hurricanes. These measures can explore the spatial and temporal dynamics of chemical reactions, particularly in combustion processes. Automatically tracking the cell-to-cell interaction in the biological system with minimal computation cost is a particular application for SMI. It can also be used in security and video surveillance to detect anomalies. All these use cases and applications share a common benefit: they help reduce the cost of analyzing and storing vast amounts of data.

The dissertation also proposes MultiSac SIMCoV model that includes lung structures like alveolar sacs in the simulation model. It emphasizes the fact that these simulations are biologically relevant and the analysis of the spread of SARS-CoV-2 in the lungs is more accurate. This model provides a detailed understanding of how the virus affects lung epithelial cells and the immune response, highlighting the spatial dynamics of the infection. MultiSac SIMCoV's ability to simulate the distribution and progression of the inflammation caused by the virus offers a valuable tool for predicting patient outcomes and proposing medical interventions. The use of CT scans in conjunction with MultiSac SIMCoV enables a comparison of simulated infection patterns with actual patient data. This approach not only validates the model but also enhances our understanding of the variability in disease severity among patients. By identifying key factors that influence the spread and impact of the virus, SIMCoV contributes to the development of more effective treatment strategies.

The domains and methods proposed in this dissertation have vast potential for future work toward several key areas:

- Enhanced Bio-Computational Models: Improving the computational efficiency of NMI and SMI calculations, possibly through GPU acceleration, can allow for real-time analysis of large biomedical datasets. This will facilitate the broader application of these methods in clinical datasets and health records.

- Integration with Advanced Imaging Techniques: Combining NMI and SMI with advanced imaging techniques such as two-photon microscopy and high-resolution CT scans can provide more detailed spatial maps of cellular interactions and feature extraction. One possible application would be using NMI to quantify the spatial association of lesions with the periphery of the lung. It is already mentioned that most lung lesions are observed on the periphery of the lung. (See Section 7.6 and figures in Appendix A) This integration can lead to better diagnostic tools and more precise tracking of disease progression.

- Personalized Medicine: Using information-theoretic measures to analyze patient-specific data can lead to personalized treatment plans. By understanding the unique spatial and temporal patterns of disease in each patient, treatments can be tailored to target specific pathways and interactions, improving efficacy and reducing side effects.

- Broader Biological Applications: Agent-based models can be adapted to any model based on agent interactions and properties. Beyond lung infections, SIMCoV can be developed to analyze other areas of biomedical research, such as cancer biology, neurodegenerative diseases, and immune response studies. The versatility of this model makes it a valuable tool for a wide range of biological investigations. Moreover, SIMCoV's computational capacity can be significantly enhanced to simulate larger systems, such as the entire lung.

- Data Summarization and Visualization: Developing more sophisticated algorithms using information theory measures to summarize and visualize multivariate time-varying data

can aid researchers in interpreting complex datasets. For example, these algorithms can be used to identify significant patterns and relationships in large ecological systems like fish foraging behavior in coral reefs, predator-prey dynamics in savannas, and plant-pollinator interactions in rainforests.

APPENDICES

Appendix A

Analyzing the Spatial Spread of SARS-CoV-2 in Lung CT Scans using SIMCoV

A.1 Experiments

A.1.1 Lung and Lesion Visualization

We have analyzed two other datasets using our method mentioned in Section 7.3 to present the lung and lesion overlay visualization. [129] dataset has 9 patients and [50] dataset has 20 patients. These analyses will be further used to study the lung-lesion association and causes for lesions mostly appearing on the periphery of the lung.

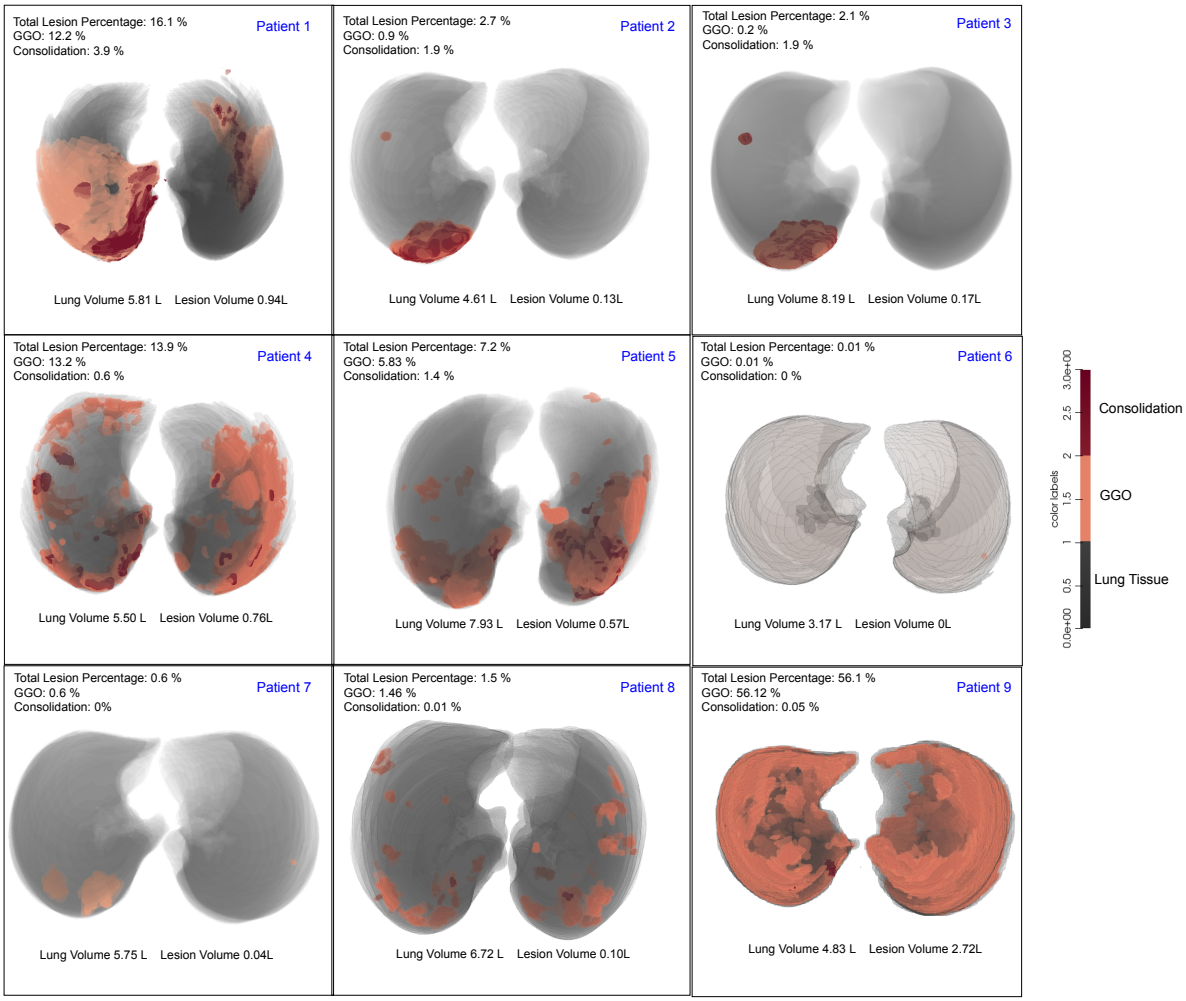


Figure A.1: Visualization of Lung and Lesion from COVID-19 dataset [129]. The GGOs and consolidations are indicated separately.

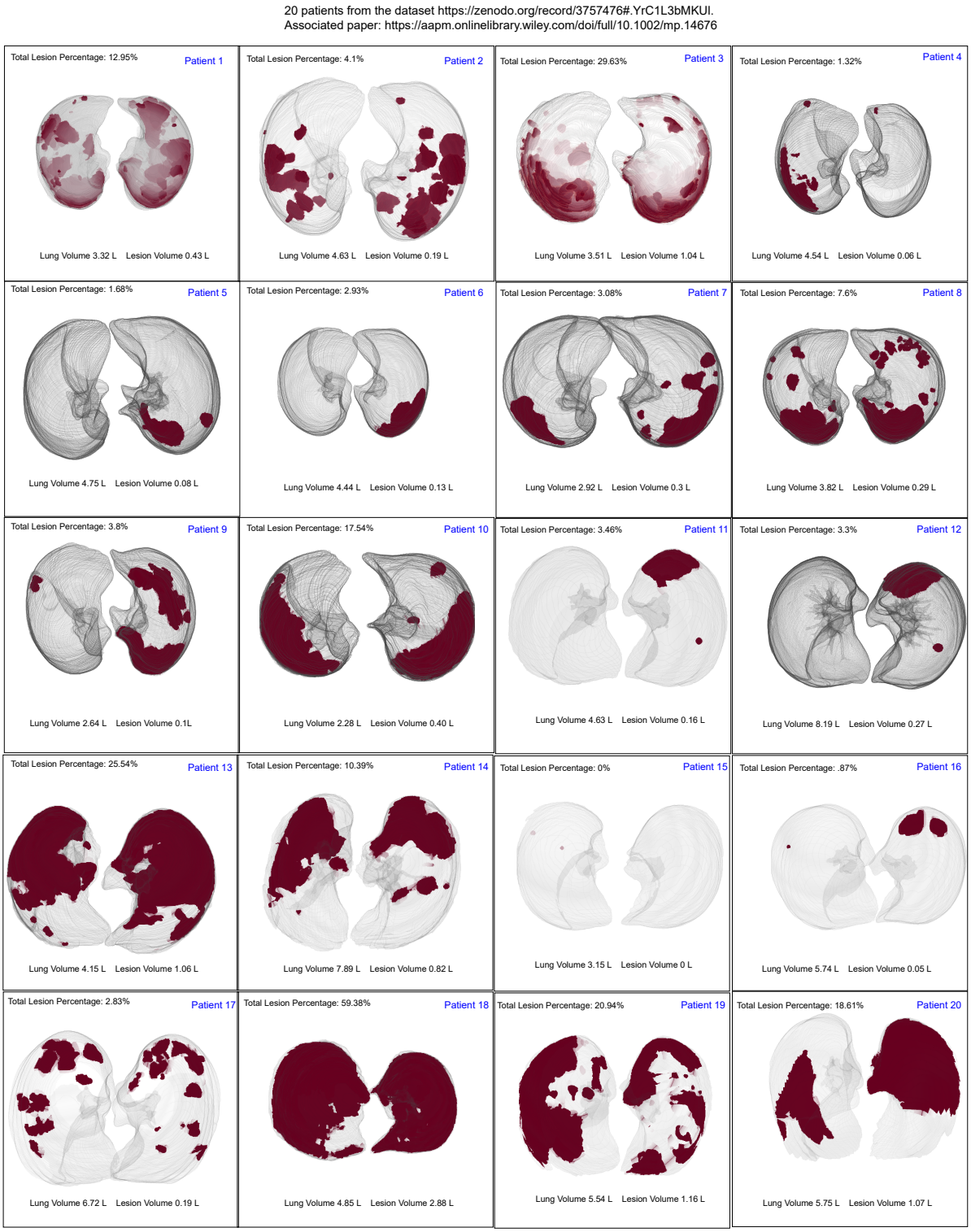


Figure A.2: Visualization of Lung and Lesion from COVID-19 dataset [50]. The GGOs are presented in red.

Bibliography

[1] MFIX-Exa. https://amrex-codes.github.io/MFIX-Exa/docs_html/, 2021 (accessed August 25, 2021).

[2] 3d slicer. https://www.slicer.org/. Accessed: 2024-07-12.

[3] Jeremy Adler and Ingela Parmryd. Quantifying colocalization by correlation: the Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytometry Part A*, 77(8):733–742, 2010.

[4] J. Ahrens, S. Jourdain, P. OLeary, J. Patchett, D. H. Rogers, and M. Petersen. An image-based approach to extreme scale in situ visualization and analysis. In *SC14: International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 424–434, 2014.

[5] James Ahrens, Sébastien Jourdain, Patrick O'Leary, John Patchett, David H. Rogers, and Mark Petersen. An Image-Based Approach to Extreme Scale in Situ Visualization and Analysis. *International Conference for High Performance Computing, Networking, Storage and Analysis, SC*, 2015(January):424–434, 2014.

[6] Hiroshi Akiba, Nathaniel Fout, and Kwan-Liu Ma. Simultaneous classification of time-varying volume data based on the time histogram. In *EuroVis*, volume 6, pages 1–8, 2006.

[7] Hiroshi Akiba, Kwan-Liu Ma, Jacqueline H. Chen, and Evatt R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Comput. Sci. Eng.*, 9(2):76–83, 2007.

[8] Hiroshi Akiba, Kwan Liu Ma, Jacqueline H. Chen, and Evatt R. Hawkes. Visualizing multivariate volume data from turbulent combustion simulations. *Computing in Science and Engineering*, 9(2):76–83, 2007.

[9] Douglas Altman, David Machin, Trevor Bryant, and Martin Gardner. *Statistics with confidence: confidence intervals and statistical guidelines*. John Wiley & Sons, 2013.

[10] François Asperti-Boursin, Eliana Real, Georges Bismuth, Alain Trautmann, and Emmanuel Donnadieu. CCR7 ligands control basal T cell motility within lymph node slices in a phosphoinositide 3–kinase–independent manner. *The Journal of experimental medicine*, 204(5):1167–1179, 2007.

[11] Utkarsh Ayachit. *The ParaView Guide: A Parallel Visualization Application*. Kitware Inc., 4.3 edition, 2015. ISBN 978-1-930934-30-6.

[12] Espen S Baekkevold, Takeshi Yamanaka, Roger T Palframan, Hege S Carlsen, Finn P Reinholt, Ulrich H von Andrian, Per Brandtzaeg, and Guttorm Haraldsen. The CCR7 ligand elc (CCL19) is transcytosed in high endothelial venules and mediates T cell recruitment. *Journal of Experimental Medicine*, 193(9):1105–1112, 2001.

[13] Marc Bajénoff, Jackson G Egen, Lily Y Koo, Jean Pierre Laugier, Frédéric Brau, Nicolas Glaichenhaus, and Ronald N Germain. Stromal cell networks regulate lymphocyte entry, migration, and territoriality in lymph nodes. *Immunity*, 25(6):989–1001, 2006.

[14] Marc Bajénoff, Samuel Granjeaud, and Sylvie Guerder. The strategy of T cell antigen-presenting cell encounter in antigen-draining lymph nodes revealed by imaging of initial T cell activation. *The Journal of experimental medicine*, 198(5):715–724, 2003.

[15] A.K. Banarjee. *Radiology Made Easy*. Cambridge University Press, Cambridge, UK, 2009.

[16] Divya Banesh, Joseph A. Schoonover, James P. Ahrens, and Bernd Hamann. Extracting, Visualizing and Tracking Mesoscale Ocean Eddies in Two-dimensional Image Sequences Using Contours and Moments. In Karsten Rink, Ariane Middel, Dirk Zeckzer, and Roxana Bujack, editors, *Workshop on Visualisation in Environmental Sciences (EnvirVis)*. The Eurographics Association, 2017.

[17] Edward J Banigan, Tajie H Harris, David A Christian, Christopher A Hunter, Andrea J Liu, and Becca Asquith. Heterogeneous CD8+ T Cell Migration in the Lymph Node in the Absence of Inflammation Revealed by Quantitative Migration Analysis. *PLoS computational biology*, 11(2):e1004058–e1004058, 2015.

[18] Andrew L Barlow, Alasdair MacLeod, Samuel Noppen, Jeremy Sanderson, and Christopher J Guérin. Colocalization analysis in fluorescence micrographs: verification of a more accurate calculation of pearson's correlation coefficient. *Microscopy and Microanalysis*, 16(6):710–724, 2010.

[19] Olivier Barnich and Marc Van Droogenbroeck. Vibe: A universal background subtraction algorithm for video sequences. *IEEE Transactions on Image processing*, 20(6):1709–1724, 2010.

[20] Biorender. https://www.biorender.com/. Accessed: 2024-07-12.

[21] Ayan Biswas, Soumya Dutta, Jesus Pulido, and James Ahrens. In situ data-driven adaptive sampling for large-scale simulation data summarization. In *Proceedings of the Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, ISAV '18, page 13–18, New York, NY, USA, 2018. Association for Computing Machinery.

[22] Ayan Biswas, Soumya Dutta, Han Wei Shen, and Jonathan Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2683–2692, 2013.

[23] Ayan Biswas, Soumya Dutta, Han-Wei Shen, and Jonathan Woodring. An information-aware framework for exploring multivariate data sets. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2683–2692, 2013.

[24] R. Bramon, M. Ruiz, A. Bardera, I. Boada, M. Feixas, and M. Sbert. An information-theoretic observation channel for volume visualization. *Computer Graphics Forum*, 32(3 PART4):411–420, 2013.

[25] R. Bramon, M. Ruiz, A. Bardera, I. Boada, M. Feixas, and M. Sbert. Information theory-based automatic multimodal transfer function design. *IEEE Journal of Biomedical and Health Informatics*, 17(4):870–880, 2013.

[26] Roger Bramon, Imma Boada, Anton Bardera, Joaquim Rodríguez, Miquel Feixas, Josep Puig, and Mateu Sbert. Multimodal data fusion based on mutual information. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1574–1587, 2012.

[27] Roger Bramon, Imma Boada, Anton Bardera, Joaquim Rodriguez, Miquel Feixas, Josep Puig, and Mateu Sbert. Multimodal data fusion based on mutual information. *Visualization and Computer Graphics, IEEE Transactions on*, 18(9):1574 –1587, sept. 2012.

[28] Roger Bramon, Marc Ruiz, Anton Bardera, Imma Boada, Miquel Feixas, and Mateu Sbert. An information-theoretic observation channel for volume visualization. *Comput. Graph. Forum*, 32(3):411–420, 2013.

[29] Roger Bramon, Marc Ruiz, Anton Bardera, Imma Boada, Miquel Feixas, and Mateu Sbert. Information theory-based automatic multimodal transfer function design. *IEEE Journal of Biomedical and Health Informatics*, 17(4):870–880, 2013.

[30] Anna Brewitz, Sarah Eickhoff, Sabrina Dähling, Thomas Quast, Sammy Bedoui, Richard A Kroczek, Christian Kurts, Natalio Garbi, Winfried Barchet, and Matteo Iannacone. CD8+ T cells orchestrate pDC-XCR1+ dendritic cell spatial and functional cooperativity to optimize priming. *Immunity*, 46(2):205–219, 2017.

[31] Anna Brewitz, Sarah Eickhoff, Sabrina Dähling, Thomas Quast, Sammy Bedoui, Richard A Kroczek, Christian Kurts, Natalio Garbi, Winfried Barchet, Matteo Iannacone, et al. Cd8+ t cells orchestrate pdc-xcr1+ dendritic cell spatial and functional cooperativity to optimize priming. *Immunity*, 46(2):205–219, 2017.

[32] Stefan Bruckner and Torsten Möller. Isosurface similarity maps. *Computer Graphics Forum*, 29:773–782, 2010.

[33] Stefan Bruckner and Torsten Möller. Isosurface similarity maps. In *Computer Graphics Forum*, volume 29, pages 773–782. Wiley Online Library, 2010.

[34] Daniel A. Butts. How much information is associated with a particular stimulus? *Network: Computation in Neural Systems*, 14(2):177–187, 2003.

[35] Nathan D. Cahill. Normalized measures of mutual information with general definitions of entropy for multimodal image registration. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6204 LNCS(8):258–268, 2010.

[36] Massimo Camplani, Lucia Maddalena, Gabriel Moyá Alcover, Alfredo Petrosino, and Luis Salgado. A Benchmarking Framework for Background Subtraction in RGBD Videos. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 10590 LNCS:219–229, 2017.

[37] Franck Cappello, Sheng Di, Sihuan Li, Xin Liang, Ali Murat Gok, Dingwen Tao, Chun Hong Yoon, Xin-Chuan Wu, Yuri Alexeev, and Frederic T Chong. Use cases

of lossy compression for floating-point data in scientific data sets. *The International Journal of High Performance Computing Applications*, 33(6):1201–1220, 2019.

[38] Federico Castanedo et al. A review of data fusion techniques. *The scientific world journal*, 2013, 2013.

[39] Flora Castellino, Alex Y Huang, Grégoire Altan-Bonnet, Sabine Stoll, Clemens Scheinecker, and Ronald N Germain. Chemokines enhance immunity by guiding naive {CD}8+ {T} cells to sites of {CD}4+ {T} cell-dendritic cell interaction. *Nature*, 440(7086):890–895, 4 2006.

[40] Min Chen, Miquel Feixas, Ivan Viola, Anton Bardera, Han-Wei Shen, and Mateu Sbert. *Information Theory Tools for Visualization*. A. K. Peters, Ltd., USA, 2016.

[41] Min Chen, Miquel Feixas, Ivan Viola, Anton Bardera, Han-Wei Shen, and Mateu Sbert. *Information theory tools for visualization*. CRC Press, 2016.

[42] Min Chen, Miquel Feixas, Ivan Viola, Anton Bardera, Han-Wei Shen, and Mateu Sbert. *Information Theory Tools for Visualization*. A K Peters/CRC Press, August 25, 2016.

[43] Min Chen and Heike Jänicke. An information-theoretic framework for visualization. *Visualization and Computer Graphics, IEEE Transactions on*, 16(6):1206–1215, 2010.

[44] Hank Childs. Data exploration at the exascale. *Supercomputing frontiers and innovations*, 2(3), 2015.

[45] Hank Childs. Data exploration at the exascale. *Supercomputing frontiers and innovations*, 2(3):5–13, 2015.

[46] Kenneth Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

[47] André Collignon, Frederik Maes, Dominique Delaere, Dirk Vandermeulen, Paul Suetens, and Guy Marchal. Automated multi-modality image registration based on information theory. In *Information processing in medical imaging*, volume 3, pages 263–274, 1995.

[48] Clyde Hamilton Coombs, Robyn M Dawes, and Amos Tversky. Mathematical psychology: An elementary introduction. *Prentice-Hall*, 1970.

[49] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition*. Wiley-Interscience, 2006.

[50] Covid-19 ct lung and infection segmentation dataset. https://zenodo.org/records/3757476#.Xpz8OcgzZPY. Accessed: 2024-07-12.

[51] Diletta Cozzi, Edoardo Cavigli, Chiara Moroni, Olga Smorchkova, Giulia Zantonelli, Silvia Pradella, and Vittorio Miele. Ground-glass opacity (ggo): a review of the differential diagnosis in the era of covid-19. *Japanese journal of radiology*, 39(8):721–732, 2021.

[52] R.A. Crawfis and N. Max. Texture splats for 3d scalar and vector field visualization. In *Visualization, 1993. Visualization '93, Proceedings., IEEE Conference on*, pages 261–266, Oct 1993.

[53] Ct all patient analysis. https://unmm-my.sharepoint.com/:u:/g/personal/htasnim30_unm_edu/ESa24skeJcxMkp92MLzlxnIBe6WZ1eJGSWekpYDPXeb4Gg?e=4tE7Wn. Accessed: 2024-07-12.

[54] Simon DeDeo, Robert X D Hawkins, Sara Klingenstein, and Tim Hitchcock. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013.

[55] Michael R Deweese and Markus Meister. How to Measure the Information Gained From One Symbol. *Network*, 10(2):123–32, 1999.

[56] Michael R. DeWeese and Markus Meister. How to measure the information gained from one symbol. *Network: Computation in Neural Systems*, 4:325–340, nov 1999.

[57] Jelena Dinic, Astrid Riehl, Jeremy Adler, and Ingela Parmryd. The T cell receptor resides in ordered plasma membrane nanodomains that aggregate upon patching of the receptor. *Scientific reports*, 5:10082, 2015.

[58] Yadolah Dodge. *Kolmogorov–Smirnov Test*, pages 283–287. Springer New York, New York, NY, 2008.

[59] Graham M Donovan and Grant Lythe. T cell and reticular network co-dependence in HIV infection. *Journal of theoretical biology*, 2016.

[60] Robert A. Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. In *Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '88, page 65–74, New York, NY, USA, 1988. Association for Computing Machinery.

[61] Kenneth W Dunn, Malgorzata M Kamocka, and John H McDonald. A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology*, 300(4):C723–C742, 2011.

[62] Kenneth W Dunn, Malgorzata M Kamocka, and John H McDonald. A practical guide to evaluating colocalization in biological microscopy. *American Journal of Physiology-Cell Physiology*, 300(4):C723–C742, 2011.

[63] S. Dutta, C. Chen, G. Heinlein, Han-Wei Shen, and J. Chen. In situ distribution guided analysis and visualization of transonic jet engine simulations. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):811–820, 2017.

[64] S. Dutta, J. Woodring, Han-Wei Shen, J. Chen, and J. Ahrens. Homogeneity guided probabilistic data summaries for analysis and visualization of large-scale data sets. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 111–120, 2017.

[65] Soumya Dutta, Chun-Ming Chen, Gregory Heinlein, Han-Wei Shen, and Jen-Ping Chen. In situ distribution guided analysis and visualization of transonic jet engine simulations. *IEEE transactions on visualization and computer graphics*, 23(1):811–820, 2016.

[66] Soumya Dutta, Xiaotong Liu, Ayan Biswas, Han-Wei Shen, and Jen-Ping Chen. Pointwise information guided visual analysis of time-varying multi-fields. In *SIGGRAPH Asia 2017 Symposium on Visualization*, SA '17, New York, NY, USA, 2017. Association for Computing Machinery.

[67] Soumya Dutta, Xiaotong Liu, Ayan Biswas, Han-Wei Shen, and Jen-Ping Chen. Pointwise information guided visual analysis of time-varying multi-fields. In *SIGGRAPH Asia 2017 Symposium on Visualization*. ACM, 2017.

[68] Soumya Dutta, Xiaotong Liu, Ayan Biswas, Han Wei Shen, and Jen Ping Chen. Pointwise information guided visual analysis of time-varying multi-fields. *SIGGRAPH Asia 2017 Symposium on Visualization, SA 2017*, 2017.

[69] Soumya Dutta, Humayra Tasnim, Terece L. Turton, and James Ahrens. In Situ Adaptive Spatio-Temporal Data Summarization. *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, pages 315–321, 2021.

[70] Soumya Dutta, Terece Turton, David Rogers, Jordan M. Musser, James Ahrens, and Ann S. Almgren. In situ feature analysis for large-scale multiphase flow simulations. *Journal of Computational Science*, 63(July):101773, 2022.

[71] Arda Duzceker, Silvano Galliani, Christoph Vogel, Pablo Speciale, Mihai Dusmanu, and Marc Pollefeys. Deepvideomvs: Multi-view stereo on video with recurrent spatio-

temporal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15324–15333, 2021.

[72] ECP: Exascale computing project. https://www.exascaleproject.org/, (accessed August 16, 2021).

[73] Optimizing a new technology to reduce power plant carbon dioxide emissions. https://www.exascaleproject.org/optimizing-a-new-technology-to-reduce-power-plant-carbon-dioxide-emissions/, 2021 (accessed Aug 30, 2021).

[74] Mohamed Abd Elaziz, Khalid M Hosny, Ahmad Salah, Mohamed M Darwish, Songfeng Lu, and Ahmed T Sahlol. New machine learning method for image-based diagnosis of covid-19. *Plos one*, 15(6):e0235187, 2020.

[75] N. Fabian, K. Moreland, D. Thompson, A. C. Bauer, P. Marion, B. Gevecik, M. Rasquin, and K. E. Jansen. The ParaView coprocessing library: A scalable, general purpose in situ visualization library. In *2011 IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 89–96, 2011.

[76] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

[77] Patrick A Fletcher, David R L Scriven, Meredith N Schulson, and Edwin D W Moore. Multi-image colocalization and its statistical significance. *Biophysical journal*, 99(6):1996–2005, 2010.

[78] Linton C Freeman. Social network visualization, methods of., 2009.

[79] G. Matthew Fricke. *Search in T cell and Robot Swarms: Balancing Extent and Intensity*. PhD thesis, University of New Mexico, 2017.

[80] G Matthew Fricke, Kenneth A Letendre, Melanie E Moses, and Judy L Cannon. Persistence and Adaptation in Immunity: T Cells Balance the Extent and Thoroughness of Search. *PLoS Computational Biology*, 12(3):e1004818, 3 2016.

[81] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010.

[82] Michele Gaeta, Santi Volta, Emanuele Scribano, Giuseppe Loria, Antonino Vallone, and Ignazio Pandolfo. Air-space pattern in lung metastasis from adenocarcinoma of the gi tract. *Journal of Computer Assisted Tomography*, 20(2):300–304, 1996.

[83] Georg Gasteiger, Marco Ataide, and Wolfgang Kastenmüller. Lymph node–an organ for T-cell activation and pathogen defense. *Immunological reviews*, 271(1):200–220, 2016.

[84] Michael Y Gerner, Parizad Torabi-Parizi, and Ronald N Germain. Strategically Localized Dendritic Cells Promote Rapid T Cell Responses to Lymph-Borne Particulate Antigens. *Immunity*, 42(1):172–185, 2015.

[85] Jean-Philippe Girard, Christine Moussion, and Reinhold Förster. HEVs, lymphatics and homeostatic immune cell trafficking in lymph nodes. *Nature Reviews Immunology*, 12(11):762–773, 2012.

[86] Rafael C Gonzales and Paul Wintz. *Digital image processing*. Addison-Wesley Longman Publishing Co., Inc., 1987.

[87] Robert M Gray. *Entropy and information theory*. Springer Science & Business Media, 2011.

[88] J Elizabeth Gretz, Christopher C Norbury, Arthur O Anderson, Amanda E I Proudfoot, and Stephen Shaw. Lymph-borne chemokines and other low molecular weight molecules reach high endothelial venules via specialized conduits while a functional barrier limits access to the lymphocyte microenvironments in lymph node cortex. *Journal of Experimental Medicine*, 192(10):1425–1440, 2000.

[89] Joanna R Groom, Jillian Richmond, Thomas T Murooka, Elizabeth W Sorensen, Jung Hwan Sung, Katherine Bankert, Ulrich H von Andrian, James J Moon, Thorsten R Mempel, and Andrew D Luster. CXCR3 chemokine receptor-ligand interactions in the lymph node optimize CD4+ T helper 1 cell differentiation. *Immunity*, 37(6):1091–1103, 2012.

[90] Martin Haidacher, Stefan Bruckner, and Meister Eduard Gröller. Volume analysis using multimodal surface similarity. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1969–1978, oct 2011.

[91] Martin Haidacher, Stefan Bruckner, Armin Kanitsar, and Meister Eduard Gröller. Information-based transfer functions for multimodal visualization. In *VCBM*, pages 101–108. Eurographics Association, October 2008.

[92] Derek L. G. Hill, Philipp G. Batchelor, Mark Holden, and David J. Hawkes. Medical image registration. *Physics in Medicine and Biology*, 46(3):R1, 2001.

[93] Derek LG Hill, Philipp G Batchelor, Mark Holden, and David J Hawkes. Medical image registration. *Physics in medicine & biology*, 46(3):R1, 2001.

[94] William M. Wells III, Paul Viola, Hideki Atsumi, Shin Nakajima, and Ron Kikinis. Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35 – 51, 1996.

[95] Phillip Isola, Daniel Zoran, Dilip Krishnan, and Edward H. Adelson. Crisp boundary detection using pointwise mutual information. *Lecture Notes in Computer Science*

(including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 8691 LNCS(PART 3):799–814, 2014.

[96] Jinhee Jeong and Fazle Hussain. On the identification of a vortex. *Journal of Fluid Mechanics*, 285:69–94, 1995.

[97] Shafiza Ariffin Kashinath, Salama A Mostafa, Aida Mustapha, Hairulnizam Mahdin, David Lim, Moamin A Mahmoud, Mazin Abed Mohammed, Bander Ali Saleh Al-Rimy, Mohd Farhan Md Fudzee, and Tan Jhon Yang. Review of data fusion methods for real-time and multi-sensor traffic flow analysis. *IEEE Access*, 9:51258–51276, 2021.

[98] Michael T Kassin, Nicole Varble, Maxime Blain, Sheng Xu, Evrim B Turkbey, Stephanie Harmon, Dong Yang, Ziyue Xu, Holger Roth, Daguang Xu, et al. Generalized chest ct and lab curves throughout the course of covid-19. *Scientific reports*, 11(1):6940, 2021.

[99] Tomoya Katakai, Takahiro Hara, Jong-Hwan Lee, Hiroyuki Gonda, Manabu Sugai, and Akira Shimizu. A novel reticular stromal structure in lymph node cortex: an immunoplatform for interactions among dendritic cells, T cells and B cells. *International immunology*, 16(8):1133–1142, 2004.

[100] Tomoya Katakai and Tatsuo Kinashi. Microenvironmental control of High-speed interstitial t cell Migration in the Lymph Node. *Frontiers in immunology*, 7:194, 2016.

[101] Tomoya Katakai, Hidenori Suto, Manabu Sugai, Hiroyuki Gonda, Atsushi Togawa, Sachiko Suematsu, Yukihiko Ebisuno, Koko Katagiri, Tatsuo Kinashi, and Akira Shimizu. Organizer-like reticular stromal cell layer common to adult secondary lymphoid organs. *The Journal of Immunology*, 181(9):6189–6200, 2008.

[102] Dilpreet Kaur and Yadwinder Kaur. Various image segmentation techniques: a review. *International Journal of Computer Science and Mobile Computing*, 3(5):809–814, 2014.

[103] Faten F Kharbat, Tarik A Elamsy, and Nuha H Hamada. Diagnosing covid-19 in x-ray images using hog image feature and artificial intelligence classifiers. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 1–5, 2020.

[104] Junhwan Kim. Visual correspondence using energy minimization and mutual information. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1033–1040. IEEE, 2003.

[105] Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

[106] Matthew F Krummel, Frederic Bartumeus, and Audrey Gérard. T cell migration, search strategies and mechanisms. *Nature Reviews Immunology*, 16(3):193, 2016.

[107] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[108] Jiss Kuruvilla, Dhanya Sukumaran, Anjali Sankar, and Siji P Joy. A review on image processing and image segmentation. In *2016 International Conference on Data Mining and Advanced Computing (SAPIENCE)*, pages 198–203, 2016.

[109] Matthew Larsen, James Ahrens, Utkarsh Ayachit, Eric Brugger, Hank Childs, Berk Geveci, and Cyrus Harrison. The alpine in situ infrastructure: Ascending from the ashes of strawman. In *Proceedings of the In Situ Infrastructures on Enabling Extreme-Scale Analysis and Visualization*, ISAV'17, page 42–46, New York, NY, USA, 2017. Association for Computing Machinery.

[110] Matthew Larsen, Amy Woods, Nicole Marsaglia, Ayan Biswas, Soumya Dutta, Cyrus Harrison, and Hank Childs. A flexible system for in situ triggers. In *Proceedings of the Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, ISAV'18, page 1–6, New York, NY, USA, 2018. Association for Computing Machinery.

[111] H. Lehmann and B. Jung. In-situ multi-resolution and temporal data compression for visual exploration of large-scale scientific simulations. In *IEEE 4th Symposium on Large Data Analysis and Visualization (LDAV), 2014*, pages 51–58, 2014.

[112] Kenneth Letendre, François Asperti-Boursin, Emmanuel Donnadieu, Melanie E Moses, and Judy L Cannon. Bringing Statistics Up To Speed With Data in Analysis of Lymphocyte Motility. *PloS one*, 2015.

[113] Jun Li, Yunfei Li, Lin He, Jin Chen, and Antonio Plaza. Spatio-temporal fusion for remote sensing data: An overview and new benchmark. *Science China Information Sciences*, 63:1–17, 2020.

[114] Shaomeng Li, Nicole Marsaglia, Christoph Garth, Jonathan Woodring, John Clyne, and Hank Childs. Data reduction techniques for simulation, visualization and data analysis. *Computer Graphics Forum*, 37(6):422–447, September 2018.

[115] Jeffrey Lian and Andrew D Luster. Chemokine-guided cell positioning in the lymph node orchestrates the generation of adaptive immune responses. *Current opinion in cell biology*, 36:1–6, 2015.

[116] Xin Liang, Sheng Di, Dingwen Tao, Zizhong Chen, and Franck Cappello. An efficient transformation scheme for lossy data compression with point-wise relative error bound. In *2018 IEEE International Conference on Cluster Computing (CLUSTER)*, pages 179–189, 2018.

[117] Randall L Lindquist, Guy Shakhar, Diana Dudziak, Hedda Wardemann, Thomas Eisenreich, Michael L Dustin, and Michel C Nussenzweig. Visualizing dendritic cell networks in vivo. *Nature immunology*, 5(12):1243–1250, 2004.

[118] P. Lindstrom. Fixed-rate compressed floating-point arrays. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):2674–2683, Dec 2014.

[119] Alexander Link, Tobias K Vogt, Stéphanie Favre, Mirjam R Britschgi, Hans Acha-Orbea, Boris Hinz, Jason G Cyster, and Sanjiv A Luther. Fibroblastic reticular cells in lymph nodes regulate the homeostasis of naive T cells. *Nature immunology*, 8(11):1255, 2007.

[120] Joseph T Lizier. JIDT: An information-theoretic toolkit for studying the dynamics of complex systems. *arXiv preprint arXiv:1408.3270*, 2014.

[121] Mark R Looney, Emily E Thornton, Debasish Sen, Wayne J Lamm, Robb W Glenny, and Matthew F Krummel. Stabilized imaging of immune surveillance in the mouse lung. *Nature methods*, 8(1):91–96, 2011.

[122] Lung analyzer. https://www.youtube.com/watch?v=v1-L_niLZxQ. Accessed: 2024-07-12.

[123] Jun Ma, Yixin Wang, Xingle An, Cheng Ge, Ziqi Yu, Jianan Chen, Qiongjie Zhu, Guoqiang Dong, Jian He, Zhiqiang He, et al. Toward data-efficient learning: A benchmark for covid-19 ct lung and infection segmentation. *Medical physics*, 48(3):1197–1210, 2021.

[124] Pulong Ma and Emily L. Kang. Spatio-temporal data fusion for massive sea surface temperature data from modis and amsr-e instruments. *Environmetrics*, 31(2):e2594, 2020.

[125] Pulong Ma and Emily L Kang. Spatio-temporal data fusion for massive sea surface temperature data from modis and amsr-e instruments. *Environmetrics*, 31(2):e2594, 2020.

[126] Frederik Maes, André Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *Medical Imaging, IEEE Transactions on*, 16(2):187–198, April 1997.

[127] Frederik Maes, Andre Collignon, Dirk Vandermeulen, Guy Marchal, and Paul Suetens. Multimodality image registration by maximization of mutual information. *IEEE Transactions on Medical Imaging*, 16(2):187–198, 1997.

[128] Robert McGill, John W Tukey, and Wayne A Larsen. Variations of box plots. *The American Statistician*, 32(1):12–16, 1978.

[129] Medseg dataset. https://www.medseg.ai/. Accessed: 2024-07-12.

[130] Ada Melo-Vallès, Clara Ballesté-Delpierre, and Jordi Vila. Review of the microbiological diagnostic approaches of covid-19. *Frontiers in Public Health*, 9:592500, 2021.

[131] Thorsten R Mempel, Sarah E Henrickson, and Ulrich H Von Andrian. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature*, 427(6970):154–159, 2004.

[132] Leslie Mertz. Ai-driven covid-19 tools to interpret, quantify lung images. *IEEE pulse*, 11(4):2–7, 2020.

[133] Mark J Miller, Sindy H Wei, Michael D Cahalan, and Ian Parker. Autonomous T cell trafficking examined in vivo with intravital two-photon microscopy. *Proceedings of the National Academy of Sciences*, 100(5):2604–2609, 2003.

[134] Henry P Mirsky, Mark J Miller, Jennifer J Linderman, and Denise E Kirschner. Systems biology approaches for understanding cellular mechanisms of immunity in lymph nodes during infection. *Journal of theoretical biology*, 287:160–170, 2011.

[135] Melanie Mitchell. *Complexity: A guided tour*. Oxford university press, 2009.

[136] Douglas G. Moore, Gabriele Valentini, Sara I. Walker, and Michael Levin. Inform: Efficient information-theoretic analysis of collective behaviors. *Frontiers Robotics AI*, 5(JUN):1–14, 2018.

[137] Melanie E. Moses, Steven Hofmeyr, Judy L. Cannon, Akil Andrews, Rebekah Gridley, Monica Hinga, Kirtus Leyba, Abigail Pribisova, Vanessa Surjadidjaja, Humayra Tasnim, and Stephanie Forrest. Spatially distributed infection increases viral load in a computational model of sars-cov-2 lung infection. *PLOS Computational Biology*, 17(12):1–24, 12 2021.

[138] Paulus Mrass, Sreenivasa Rao Oruganti, G. Matthew Fricke, Justyna Tafoya, Janie R. Byrum, Lihua Yang, Samantha L. Hamilton, Mark J. Miller, Melanie E. Moses, and Judy L. Cannon. ROCK regulates the intermittent mode of interstitial T cell migration in inflamed lungs. *Nature Communications*, 8(1), 2017.

[139] Paulus Mrass, Sreenivasa Rao Oruganti, G Matthew Fricke, Justyna Tafoya, Janie R Byrum, Lihua Yang, Samantha L Hamilton, Mark J Miller, Melanie E Moses, and Judy L Cannon. Rock regulates the intermittent mode of interstitial t cell migration in inflamed lungs. *Nature communications*, 8(1):1010, 2017.

[140] Jordan Musser, Ann S Almgren, William D Fullmer, Oscar Antepara, John B Bell, Johannes Blaschke, Kevin Gott, Andrew Myers, Roberto Porcu, Deepak Rangarajan, Michele Rosso, Weiqun Zhang, and Madhava Syamlal. MFIX-Exa: A path toward exascale CFD-DEM simulations. *International Journal of High Performance Computing Applications*, 2021.

[141] Jordan Musser, Ann S. Almgren, William D. Fullmer, Oscar Antepara, John B. Bell, Johannes Blaschke, Kevin Gott, Andrew Myers, Roberto Porcu, Deepak Rangarajan, Michele Rosso, Weiqun Zhang, and Madhava Syamlal. MFIX-Exa: A path toward exascale CFD-DEM simulations. *International Journal of High Performance Computing Applications*, 36(1):40–58, 2022.

[142] Kary Myers, Earl Lawrence, Michael Fugate, Claire McKay Bowen, Lawrence Ticknor, Jon Woodring, Joanne Wendelberger, and Jim Ahrens. Partitioning a large simulation as

it runs. *Technometrics*, 58(3):329–340, 2016.

[143] Kary Myers, Earl Lawrence, Michael Fugate, Claire McKay Bowen, Lawrence Ticknor, Jon Woodring, Joanne Wendelberger, and Jim Ahrens. Partitioning a large simulation as it runs. *Technometrics*, 58(3):329–340, 2016.

[144] Pariksheet Nanda and Denise E. Kirschner. Calibration methods to fit parameters within complex biological models. *Frontiers in Applied Mathematics and Statistics*, 9, 2023.

[145] Hai Nguyen, Matthias Katzfuss, Noel Cressie, and Amy Braverman. Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185, 2014.

[146] Hai Nguyen, Matthias Katzfuss, Noel Cressie, and Amy Braverman. Spatio-temporal data fusion for very large remote sensing datasets. *Technometrics*, 56(2):174–185, 2014.

[147] Mario Novkovic, Lucas Onder, Jovana Cupovic, Jun Abe, David Bomze, Viviana Cremasco, Elke Scandella, Jens V Stein, Gennady Bocharov, and Shannon J Turley. Topological small-world organization of the fibroblastic reticular cell network determines lymph node functionality. *PLoS biology*, 14(7):e1002515, 2016.

[148] Roger T Palframan, Steffen Jung, Guiying Cheng, Wolfgang Weninger, Yi Luo, Martin Dorf, Dan R Littman, Barrett J Rollins, Hans Zweerink, and Antal Rot. Inflammatory chemokine transport and presentation in HEV. *Journal of Experimental Medicine*, 194(9):1361–1374, 2001.

[149] James B Pawley and Barry R Masters. Handbook of biological confocal microscopy. *Journal of biomedical optics*, 13(2):9902, 2008.

[150] W. Pawlina and M.H. Ross. *Histology: A Text and Atlas: With Correlated Cell and Molecular Biology*. Wolters Kluwer Health, 2018.

[151] Karl Pearson. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London. A*, 186:343–414, 1895.

[152] Josien P. W. Pluim, J. B. Antoine Maintz, and Max A. Viergever. Mutual-information-based registration of medical images: a survey. *IEEE Transcations on Medical Imaging*, pages 986–1004, 2003.

[153] Josien P W Pluim, J B Antoine Maintz, and Max A Viergever. Mutual-information-based registration of medical images: a survey. *IEEE transactions on medical imaging*, 22(8):986–1004, 2003.

[154] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007.

[155] Mikhail Prokopenko, Fabio Boschetti, and Alex J Ryan. An information-theoretic primer on complexity, self-organization, and emergence. *Complexity*, 15(1):11–28, 2009.

[156] Python connected component 3d. https://pypi.org/project/connected-components-3d/. Accessed: 2024-07-12.

[157] Hanae Ramdani, Khadija Benelhosni, Nabil Moatassim Billah, and Ittimade Nassar. Chest ct in covid-19 pneumonia's follow-up: A 30 patients case series. *Annals of Medicine and Surgery*, 84, 2022.

[158] Daniel A Reed and Jack Dongarra. Exascale computing and big data. *Communications of the ACM*, 58(7):56–68, 2015.

[159] David N Reshef, Yakir A Reshef, Hilary K Finucane, Sharon R Grossman, Gilean McVean, Peter J Turnbaugh, Eric S Lander, Michael Mitzenmacher, and Pardis C Sabeti. Detecting novel associations in large data sets. *science*, 334(6062):1518–1524, 2011.

[160] Background learning for detection and tracking from rgbd videos. https://rgbd2017.na.icar.cnr.it/. Accessed: 2023-09-22.

[161] Jaume Rigau, Miquel Feixas, and Mateu Sbert. Shape complexity based on mutual information. In *Shape Modeling and Applications, 2005 International Conference*, pages 355–360, June 2005.

[162] Michael Rubart. Two-photon microscopy of cells and tissue. *Circulation research*, 95(12):1154–1166, 2004.

[163] Marc Ruiz, Anton Bardera, Imma Boada, Ivan Viola, Miquel Feixas, and Mateu Sbert. Automatic transfer functions based on informational divergence. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1932–1941, 2011.

[164] Daniel B Russakoff, Carlo Tomasi, Torsten Rohlfing, and Calvin R Maurer. Image similarity using mutual information of regions. In *European Conference on Computer Vision*, pages 596–607. Springer, 2004.

[165] Maher Salloum, Janine C. Bennett, Ali Pinar, Ankit Bhagatwala, and Jacqueline H. Chen. Enabling adaptive scientific workflows via trigger detection. In *Proceedings of the First Workshop on In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization*, ISAV2015, page 41–45, New York, NY, USA, 2015. Association for Computing Machinery.

[166] Mateu Sbert, Miquel Feixas, Jaume Rigau, Miguel Chover, and Ivan Viola. *Information Theory Tools for Computer Graphics*. Synthesis Lectures on Computer Graphics and Animation. Morgan and Claypool Publishers Colorado, 2009.

[167] Mateu Sbert, Miquel Feixas, Jaume Rigau, Miguel Chover, and Ivan Viola. *Information theory tools for computer graphics*. Springer Nature, 2022.

[168] Sbm-rgbd dataset. https://rgbd2017.na.icar.cnr.it/SBM-RGBDdataset.html. Accessed: 2023-09-22.

[169] W. Schroeder, K. Martin, and B. Lorensen. *The Visualization Toolkit: An Object Oriented Approach to 3D Graphics*. Kitware Inc., fourth edition, 2004. ISBN 1-930934-19-X.

[170] Will Schroeder, Kenneth M Martin, and William E Lorensen. *The visualization toolkit an object-oriented approach to 3D graphics*. Prentice-Hall, Inc., 1998.

[171] Z. Shah, A. Anwar, A. Mahmood, Z. Tari, and A. Y. Zomaya. A spatiotemporal data summarization approach for real-time operation of smart grid. *IEEE Transactions on Big Data*, 6(04):624–637, oct 2020.

[172] Zubair Shah, Adnan Anwar, Abdun Naser Mahmood, Zahir Tari, and Albert Y Zomaya. A spatiotemporal data summarization approach for real-time operation of smart grid. *IEEE Transactions on Big Data*, 6(4):624–637, 2017.

[173] Claude E. Shannon. Claude E. Shannon. *Bell Systems Technical Journal*, 27(3):379–423, 1948.

[174] Simcov ct data and resources. https://unmm-my.sharepoint.com/:f:/g/personal/htasnim30_unm_edu/EkwRbmqTdcpJlrRbGigQBysBdSPKVYm_0Pl6cuLsRDYCqg?e=NICMq8. Accessed: 2024-07-12.

[175] Michael Sipser. *Introduction to the Theory of Computation*. Course Technology, Boston, MA, third edition, 2013.

[176] Michael Sixt, Nobuo Kanazawa, Manuel Selg, Thomas Samson, Gunnel Roos, Dieter P Reinhardt, Reinhard Pabst, Manfred B Lutz, and Lydia Sorokin. The conduit system transports soluble antigens from the afferent lymph to resident dendritic cells in the T cell area of the lymph node. *Immunity*, 22(1):19–29, 2005.

[177] Reginald Smith. A mutual information approach to calculating nonlinearity. *Stat*, 4(1):291–303, 2015.

[178] Jens V Stein and César Nombela-Arrieta. Chemokine control of lymphocyte trafficking: a general overview. *Immunology*, 116(1):1–12, 2005.

[179] Jens V Stein, Antal Rot, Yi Luo, Manjunath Narasimhaswamy, Hideki Nakano, Michael D Gunn, Akio Matsuzawa, Elizabeth J Quackenbush, Martin E Dorf, and Ulrich H von Andrian. The CC chemokine thymus-derived chemotactic agent 4 (TCA-4, secondary lymphoid tissue chemokine, 6Ckine, exodus-2) triggers lymphocyte function–associated antigen 1–mediated arrest of rolling T lymphocytes in peripheral lymph node high endothelial venules. *Journal of Experimental Medicine*, 191(1):61–76, 2000.

[180] Alexander Strehl and Joydeep Ghosh. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec):589, 2002.

[181] Colin Studholme, Derek L G Hill, and David J Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern recognition*, 32(1):71–86, 1999.

[182] Akira Takeda, Daichi Kobayashi, Keita Aoi, Naoko Sasaki, Yuki Sugiura, Hidemitsu Igarashi, Kazuo Tohya, Asuka Inoue, Erina Hata, and Noriyuki Akahoshi. Fibroblastic reticular cell-derived lysophosphatidic acid regulates confined intranodal T-cell motility. *Elife*, 5:e10561, 2016.

[183] H. Tasnim, G.M. Fricke, J.R. Byrum, J.O. Sotiris, J.L. Cannon, and M.E. Moses. Quantitative measurement of naïve T cell association with dendritic cells, FRCs, and blood vessels in lymph nodes. *Frontiers in Immunology*, 9(JUL), 2018.

[184] Humayra Tasnim, Soumya Dutta, and Melanie Moses. Dynamic spatio-temporal summarization using information based fusion, 2023.

[185] Humayra Tasnim, Soumya Dutta, Terece L. Turton, David H. Rogers, and Melanie E. Moses. Information-Theoretic Exploration of Multivariate Time-Varying Image Databases. *Computing in Science and Engineering*, 24(3):61–70, 2022.

[186] Johannes Textor, Judith N Mandl, and Rob J de Boer. The reticular cell network: a robust backbone for immune responses. *PLoS biology*, 14(10):e2000827, 2016.

[187] Xin Tong, Teng-Yok Lee, and Han-Wei Shen. Salient time steps selection from large scale time-varying data sets with dynamic time warping. In *IEEE Symposium on Large Data Analysis and Visualization (LDAV)*, pages 49–56, 2012.

[188] Xin Tong, Teng-Yok Lee, and Han-Wei Shen. Salient time steps selection from large scale time-varying data sets with dynamic time warping. In *IEEE symposium on large data analysis and visualization (LDAV)*, pages 49–56. IEEE, 2012.

[189] G.J. Tortora and B.H. Derrickson. *Principles of Anatomy and Physiology*. Wiley, 2018.

[190] S. Verdu. Fifty years of shannon theory. *IEEE Transactions on Information Theory*, 44(6):2057–2078, 1998.

[191] Sergio Verdú. Fifty years of shannon theory. *Information Theory, IEEE Transactions on*, 44(6):2057–2078, Oct 1998.

[192] K Vidhya, G Karthikeyan, P Divakar, and S Ezhumalai. A review of lossless and lossy image compression techniques. *Int. Res. J. Eng. Technol.(IRJET)*, 3(4):616–7, 2016.

[193] Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[194] Ivan Viola, Miquel Feixas, Mateu Sbert, and Meister Eduard Gröller. Importance-driven focus of attention. *Visualization and Computer Graphics, IEEE Transactions on*, 12(5):933–940, 2006.

[195] Ivan Viola, Miquel Feixas, Mateu Sbert, and Meister Eduard Groller. Importance-driven focus of attention. *IEEE transactions on visualization and computer graphics*, 12(5):933–940, 2006.

[196] Paul Viola and William M Wells III. Alignment by maximization of mutual information. *International journal of computer vision*, 24(2):137–154, 1997.

[197] Ulrich H von Andrian and Charles R Mackay. T Cell Function and Migration. *New England Journal of Medicine*, 2000.

[198] Chaoli Wang and Han-Wei Shen. Information theory in scientific visualization. *Entropy*, 13(1):254–273, 2011.

[199] Chaoli Wang, Hongfeng Yu, and Kwan-Liu Ma. Importance-driven time-varying data visualization. *IEEE Trans. on Vis. and Comp. Graphics*, 14(6):1547–1554, 2008.

[200] Chaoli Wang, Hongfeng Yu, and Kwan-Liu Ma. Importance-driven time-varying data visualization. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1547–1554, 2008.

[201] Qunming Wang, Yijie Tang, Xiaohua Tong, and Peter M Atkinson. Virtual image pair-based spatio-temporal fusion. *Remote Sensing of Environment*, 249:112009, 2020.

[202] T. Wei, S. Dutta, and Han-Wei Shen. Information guided data sampling and recovery using bitmap indexing. In *2018 IEEE Pacific Visualization Symposium (PacificVis)*, pages 56–65, 2018.

[203] Tzu-Hsuan Wei, Teng-Yok Lee, and Han-Wei Shen. Evaluating isosurfaces with level-set-based information maps. In *Proceedings of the 15th Eurographics Conference on Visualization*, EuroVis '13, pages 1–10, Aire-la-Ville, Switzerland, Switzerland, 2013. Eurographics Association.

[204] E.R. Weibel. *Morphometry of the Human Lung*. Elsevier Science, 2013.

[205] J.B. West. *Respiratory Physiology: The Essentials*. Point (Lippincott Williams and Wilkins) Series. Wolters Kluwer Health/Lippincott Williams & Wilkins, 2008.

[206] Jürgen Westermann, Ulrike Bode, Andrea Sahle, Uwe Speck, Nathan Karin, Eric B Bell, Kathrin Kalies, and Andreas Gebert. Naive, effector, and memory T lymphocytes efficiently scan dendritic cells in vivo: contact frequency in T cell zones of secondary lymphoid organs does not depend on LFA-1 expression and facilitates survival of effector T cells. *The Journal of Immunology*, 174(5):2517–2524, 2005.

[207] Brad Whitlock, Jean M. Favre, and Jeremy S. Meredith. Parallel in situ coupling of simulation with a fully featured visualization system. In *Proceedings of the 11th Eurographics Conference on Parallel Graphics and Visualization*, EGPGV '11, pages 101–109. Eurographics Association, 2011.

[208] Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[209] Harikesh S Wong and Ronald N Germain. Robust control of the adaptive immune system. In *Seminars in immunology*. Elsevier, 2017.

[210] J. Woodring, J. Ahrens, J. Figg, J. Wendelberger, S. Habib, and K. Heitmann. In-situ sampling of a large-scale particle simulation for interactive visualization and analysis. In *Proceedings of the 13th Eurographics / IEEE - VGTC Conference on Visualization*, pages 1151–1160. Eurographics Association, 2011.

[211] Jonathan Woodring, James Ahrens, J Figg, Joanne Wendelberger, Salman Habib, and Katrin Heitmann. In-situ sampling of a large-scale particle simulation for interactive visualization and analysis. In *Computer Graphics Forum*, volume 30, pages 1151–1160. Wiley Online Library, 2011.

[212] Penghai Wu, Zhixiang Yin, Chao Zeng, Si-Bo Duan, Frank-Michael Göttsche, Xiaoshuang Ma, Xinghua Li, Hui Yang, and Huanfeng Shen. Spatially continuous and

high-resolution land surface temperature product generation: A review of reconstruction and spatiotemporal fusion techniques. *IEEE Geoscience and Remote Sensing Magazine*, 9(3):112–137, 2021.

[213] Jie Xue, Yee Leung, and Tung Fung. A bayesian data fusion approach to spatio-temporal fusion of remotely sensed images. *Remote Sensing*, 9(12):1310, 2017.

[214] Yucong Chris Ye, Tyson Neuroth, Franz Sauer, Kwan-Liu Ma, Giulio Borghesi, Aditya Konduri, Hemanth Kolla, and Jacqueline Chen. In situ generated probability distribution functions for interactive post hoc visualization and analysis. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pages 65–74, 2016.

[215] Yucong Chris Ye, Tyson Neuroth, Franz Sauer, Kwan-Liu Ma, Giulio Borghesi, Aditya Konduri, Hemanth Kolla, and Jacqueline Chen. In situ generated probability distribution functions for interactive post hoc visualization and analysis. In *2016 IEEE 6th Symposium on Large Data Analysis and Visualization (LDAV)*, pages 65–74. IEEE, 2016.

[216] Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View That Includes Motifs, Discords and Shapelets. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*, pages 1317–1322. IEEE, 12 2016.

[217] Ming Zeng, Peter J Southern, Cavan S Reilly, Greg J Beilman, Jeffrey G Chipman, Timothy W Schacker, and Ashley T Haase. Lymphoid tissue damage in HIV-1 infection depletes naïve T cells and limits T cell reconstitution after antiretroviral therapy. *PLoS pathogens*, 8(1):e1002437, 2012.

[218] Luming Zhang, Yue Gao, Richang Hong, Yuxing Hu, Rongrong Ji, and Qionghai Dai. Probabilistic skimlets fusion for summarizing multiple consumer landmark videos. *IEEE Transactions on Multimedia*, 17(1):40–49, 2014.

[219] Sheng-hua Zhong, Jiaxin Wu, and Jianmin Jiang. Video summarization via spatio-temporal deep architecture. *Neurocomputing*, 332:224–235, 2019.

[220] Bo Zhou and Yi-Jen Chiang. Key time steps selection for large-scale time-varying volume datasets using an information-theoretic storyboard. *Computer Graphics Forum*, 37(3):37–49, 2018.

[221] Bo Zhou and Yi-Jen Chiang. Key time steps selection for large-scale time-varying volume datasets using an information-theoretic storyboard. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018.