Fall 11-15-2018

# Multi-Resolution Analysis of Large Molecular Structures and Interactions

Kasra Manavi
*University of New Mexico*

Kasra Manavi

*Candidate*

Computer Science

*Department*

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

*Approved by the Dissertation Committee:*

Dr. Lydia Tapia, *Chairperson*

Dr. Bruna Jacobson

Dr. Shuang Luan

Dr. Darko Stefanovic

Dr. Bridget S. Wilson

# Multi-Resolution Analysis of Large Molecular Structures and Interactions

by

**Kasra Manavi**

B.S., University of New Mexico, USA, 2009
M.S., Texas A&M University, USA, 2012

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

December, 2018

# Dedication

*To Dad, Mom and Sima.*

*"Something Smart." - Someone*

# Acknowledgments

There are a few people I would like to acknowledge for their time and efforts in assisting me in this endeavor. I would like to thank my advisor Dr. Lydia Tapia for her mentorship and guidance during my PhD. I also would like to thank Dr. Shuang Luan, Dr. Darko Stefanovic, Dr. Bridget Wilson and Dr. Bruna Jacobson for being on my committee and their advisement. Thanks to the UNM Department of Computer Science and members of the Tapia Lab for aiding me in completing my degree. Finally, I want to thank Chris Casey, Johnnie Wright and Adam Fidel for their friendship and encouragement throughout my college career.

# Multi-Resolution Analysis of Large Molecular Structures and Interactions

by

**Kasra Manavi**

B.S., University of New Mexico, USA, 2009

M.S., Texas A&M University, USA, 2012

Ph.D., Computer Science, University of New Mexico, 2018

## Abstract

Simulation of large molecular structures and their interactions has become a major component of modern biomolecular research. Methods to simulate these type of molecules span a wide array of resolutions, from all atom molecular dynamics to model interaction energetics to systems of linear equations to evaluate population kinetics. In recent years, there has been an acceleration of molecular structural information production, primarily from x-ray crystallography and electron microscopy. This data has provided modelers the ability to produce better representations of these molecular structures.

The purpose of this research is to take advantage of this information to develop multi-resolution models for the analysis of large molecule motions and interactions. Our methodology focuses on the use of structural models of a given biological system and simulating the molecules using different conditions (number or ratio of molecules being simulated) and constraints (rigid or semi-flexible models). We combine computational geometry and statistical techniques to perform efficient structural modeling and simulation.

Our goal is to utilize our methods to analyze the effect of geometry on molecule interactions, e.g., shape of packed structures or influences of steric hindrance caused by interacting molecules. We focus our work on larger molecular systems, both in size of the molecular structures and number of interacting molecules. The focus of our evaluation is the human allergic immune response. The immune response is triggered by cell surface molecular aggregation of antibodies via an antigen. With our analysis we gain insight into how different allergen geometries affect the size and shape of aggregate structures that form on the cell surface.

We perform a multi-resolution analysis of our structures and model the problem in two ways, a lower resolution rigid body representation which can model the aggregation process, and a higher resolution flexible model which can be used to fit structural experimental data. In the lower resolution work, we develop methods to geometrically model, simulate and analyze antibody aggregation. We show our technique handles the large size and number of molecules involved in aggregation, and we study the impact of model resolution on simulations of geometric structures. In the higher resolution work, we introduce methods to model and fit molecular structures into electron microscopy datasets (20Å - 40Å resolution). We use Gaussian mixture models to describe molecular systems with high flexibility thus enabling the generation of conformations that fit an input tomographic tilt series, a set of 2D images of a 3D molecule taken at a variety of angles. We also apply our method to experimental data, fitting a structure imaged using cryo electron microscopy tomography.

# Contents

# List of Figures

# List of Tables

# List of Algorithms

# Chapter 1

# Introduction

Structural simulation of large molecule interactions has become a major component of modern biomolecular research. One application area of simulations is modeling molecular assembly, the process in which a group of molecules adopt a functional macromolecular arrangement. The assembled structure is composed of a series of molecules bound together to create a functional component of a given biochemical process. The structures of these assemblies are dependent on several features of the individual molecules including size, structure/conformation, and valency (number of binding interfaces). These features determine the size, complexity and functionality of the assembled structure.

In this work, our focus is on the aggregation of antibodies and allergens, an assembly process which triggers the human allergy immune response [100]. This process occurs on the cell surface of key mediators of the allergic reaction, mast cells and basophils, in their response to environmental changes. IgE antibodies, produced by the lymph nodes to identify foreign threats to the body, are found bound to Fc$\epsilon$RI cell surface receptors, priming said cells for activation [100]. When an allergen (antigen) is present, the IgE antibodies bind to the antigen and start to form aggregate structures (crosslink). This aggregation process initiates a signaling cascade that propagates inside the cell, resulting in degranulation. Degranulation is the process where immune mediators are released from the cell; in the case of mast cells this includes histamine, serotonin, and leukotrienes among other molecules. These immune mediators produce an increase in local vascular permeability and induce an inflammatory response [100, 2]. In most cases, the typical allergic response includes a runny nose, irritated skin or itchy eyes, but in a hypersensitive person this immune response is severe, resulting in anaphylaxis and possible death. It is believed that the shape, size and valency of an allergen impact the strength of the response [1]. Figure 1.1 displays a series of allergens, which vary greatly in structure and valency, but still activate the same signaling pathway.

Computational simulations of molecular interactions can provide insights into the behaviors of biomolecules at resolutions not possible from experiment. Traditional approaches to simulate large molecular structures and interactions include include Molecular Dynamics (MD) [103, 102] and Monte Carlo (MC) [37, 22, 105] methods. Large-scale MD simulations (millions of atoms) have been essential in elucidating viral capsid assemblies, ribosome activity, and bioenergy systems [94]. MC methods have been used to study amyloid aggregation [68] and ion channels [16].

Unfortunately, MD and MC simulations have limitations when modeling molecular interactions. MD simulations are often computationally intensive, can only provide simulation times of up to a few microseconds, which is insufficient for processes which take place on longer time scales, and are dependent on force field parameterization [12]. MC simulations produce an ensemble of representative conformations, but do not provide information about time directly due to reduced physical complexity and may not produce a complete representation

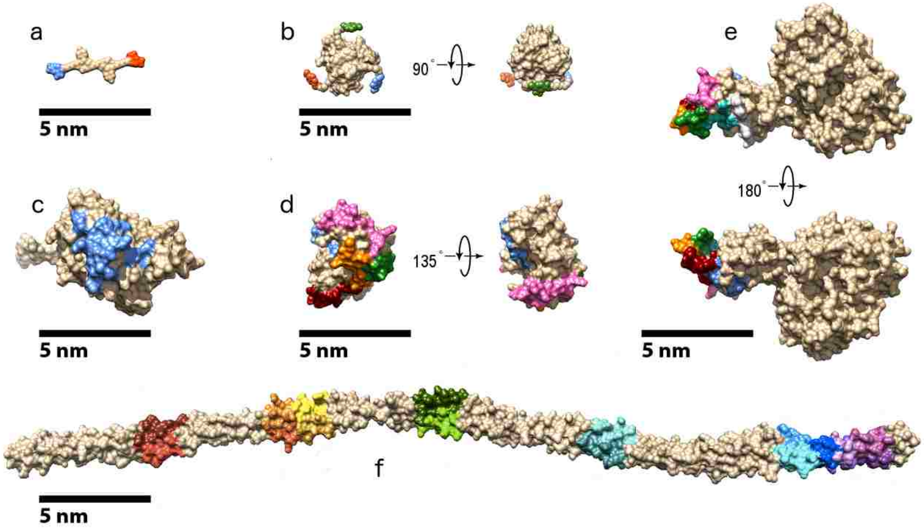


Figure 1.1: Various antigen/allergen structures that effectively signal for the allergic immune response. Note the differences in sizes, valency and binding site topography (colored areas). *a.*) Synthetic, bivalent antigen (Dct)2-cys. *b.*) Synthetic trivalent antigen DF3. *c.*) Birch pollen allergen Bet v 1. *d.*) Cedar pollen (Juniper) allergen Jun a 1. *e.*) The common peanut allergen Ara h 2. *f.*) The common shellfish allergen (shrimp tropomyosin) Pen a 1.

of the conformational space [92]. New methods are being developed which focus on producing alternative ways to simulate these molecular systems to overcome some of these challenges. There are two main efforts to study these systems with more molecules and at longer timescales: "modified" MD methods (including steered MD and coarse-grained MD) [60, 94], and robotic-inspired methods [3]. These methods have been applied to single molecule folding/unfolding [7, 117, 115, 116] and small molecule interactions [112, 11].

The purpose of this research is to extend existing robotics-inspired methods by developing models at multiple resolutions for simulation and analysis of large molecule motion and interaction. Our methodology focuses on the use of structural models of a given biological system and simulating molecules using different conditions (number/ratio of molecules being simulated) and constraints (rigid and semi-flexible models). We combine computational geometry and statistical techniques to generate structural models of biological data for more efficient simulation. We increase both the breath of analysis and computational efficiency by modeling the system at multiple levels of detail. Figure 1.2 describes how our methods are used to model molecular structures.

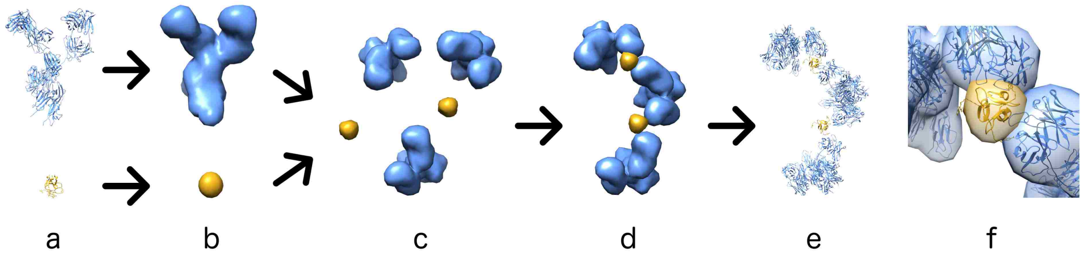


Figure 1.2: Low resolution modeling of our molecular system. *a*.) Modeling begins with all atom structures of molecules involved in the interaction. In this example we model IgE and antigen DF3. *b*.) All atom models are used to construct lower-resolution models. *c*.) We simulate the interaction of the lower-resolution model using a variety of conditions and constraints dependent on the type of simulation. *d*.) Simulation with the low-resolution models results in aggregate structures that can be analyzed. *e*.) All-atom reconstruction is performed to recreate the aggregate structures. *f*.) More detailed analysis of the structures can now be performed on the all-atom reconstruction. (Note: Ribbon format shown representing all-atom model for visual clarity)

To evaluate our methods, we focused on modeling the molecular basis of the allergic response in humans, specifically, how allergen structure and valency impact antibody aggregation. Methods introduced in our preliminary work [84, 81, 82, 80] include model construction techniques and methods for simulation and analysis of large molecular systems. We then extended our work to include articulated body modeling and aggregate conformation fitting [83, 79].

## 1.1 Research Objective

The objective of this research is to develop methods to efficiently model structures and simulate interactions of biomolecular systems. Our goal is to increase the understanding of large molecular systems by developing multi-resolution models to learn about the impact of structure on these systems. Figure 1.3 outlines the three levels of detail, i.e., Degrees Of Freedom (DOF) we choose to model and incorporate in to our methods. The highest level of detail is all atom modeling, where explicit atom position, properties and bonds are maintained by the model. We use flexible modeling for a medium level of detail, where simplified models still capture a majority of the molecular flexibility and occupied volume of a molecule. At the lowest level of detail, we perform rigid body modeling which only captures the occupied volume of a static state.

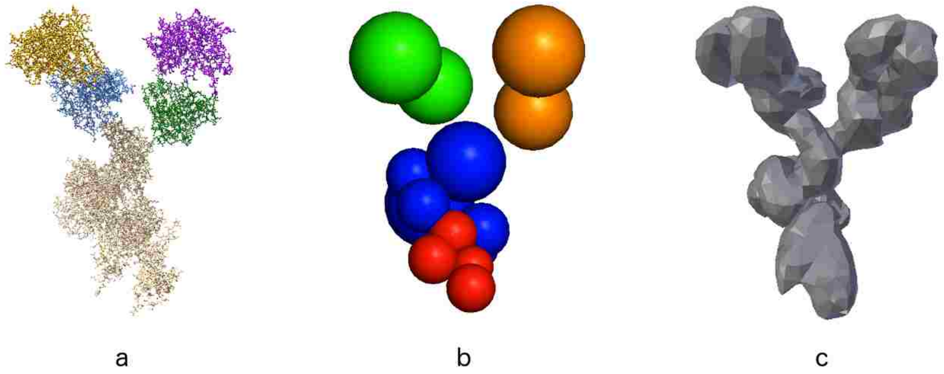


Figure 1.3: Example levels of complexity in our modeling of molecular systems. *a*.) All atom models represent the highest level of complexity available, and they are used as a basis for lower-resolution models and for evaluating biological validity. *b*.) A flexible model represents a mid-level of complexity, enabling the modeling of different conformations of a given molecule. *c*.) At the lowest level of complexity, rigid body models are simple but efficiently model a static molecular structure and occupied volume.

We achieved our objectives by combining computational geometry techniques and statistical models to efficiently simulate the interactions of large molecules. Our methods focus on molecular conformation determination, the generation and evaluation of candidate structures of a given system. To evaluate our methods, we used biological data collected from experiments pertaining to the human allergen immune response. Our work provides insights into how the size/shape of aggregate structures that are formed on the cell surface. In order to derive these insights, we developed two simulations, the first using 3D models of biomolecules to perform complex binding/aggregation, and the second modeling the molecules with flexibility to fit experimentally imaged aggregate structures.

For molecular aggregation, we simulate molecular interaction using a Monte Carlo approach with relaxed constraints and analyze the resulting aggregate structures [84, 81, 82, 45, 80, 46]. This method of simulation allows us to evaluate how aggregation is affected by allergen dose, shape, size and valency. We also developed a novel implementation of Rule-Based Modeling (RBM) that allowed us to analyze steric effects between different allergen models. Results from these simulations provide quantification of the impact of reduced model resolution.

For flexible fitting, we use a Gaussian Mixture Model (GMM) to model the structure of the molecule and fit it to a dataset collected using electron microscopy [83, 79]. This method of simulation allows us to determine actual structures that have been imaged in a microscope, for example aggregate structures.

There is a trade-off between speed and accuracy when comparing all-atom simulations to methods that utilize coarse-graining or reduced-resolution modeling. MD based methods with all-atom representations and detailed atom potentials are accurate, but slow. Our methods, based on molecular geometry and simplified interaction descriptions, are less accurate but more computationally efficient.

## 1.2 Contributions

This research focuses on the development of simulation and analysis tools for modeling the structures and interactions of large molecular systems. We have developed two simulations, the first simulating complex binding/aggregation of rigid body models, and the second modeling a more physical, semi-flexible structures to fit experimental data.

In our preliminary work on molecular aggregation, we focused on developing a method for simulation and analysis of molecular aggregation [84, 81, 82]. With this work we were able to study IgE antibody aggregate formation. Where experimental techniques only report proximity of clustered receptors, our simulations were able to provide unique insights into the aggregation process by classifying the most common geometry associated with receptor aggregates [84]. We have extended this work to evaluate the impact of antigen valency on IgE aggregate size and topography [81].

We have also worked on investigating how model resolution impacts simulation accuracy and efficiency [82]. From that work we were able to quantify and potentially account for the impacts of model simplification. To investigate this connection deeper, we developed a novel implementation of Rule-Based Modeling (RBM) that encodes molecular geometry into the rules [45, 80, 46]. In this work, we began with exploring how RBM can be designed to incorporate geometry into biochemical models [45]. We did this by analyzing RBMs for different antigen geometries and resolutions, and determined how steric effects between allergen binding regions vary with molecular geometry and model resolution. We extended this work with an evaluation of how RBM can complement other computational methods that explicitly represent molecular geometry [46]. This work was a unique integration of geometric rule-based modeling and three-dimensional simulations, showing differences in model resolution/quality of a given method can be quantified using another method. In the capstone of this project, we studied the impact of model resolution on simulations of geometric structures using our Monte Carlo simulation and RBM [80]. In this work we evaluated aggregate clustering and were able to reproduce experimental data. Our study of the aggregation of Pen a 1 (the common shrimp allergen) was considered a novel contribution to the allergen literature [46].

In the second body of work, we focused on developing a method to fit images of molecular aggregates generated using Cryogenic Electron Tomography (Cryo ET) [83]. Cryo ET is a method to capture an ordered set of images of a molecular structure, known as a tilt series, and use those images to reconstruct a 3D model of the molecule. Issues arise fitting existing models to reconstructions due to factors including reconstruction distortions and lack of resolution. To address these issues, in this work we show we can fit tilt series directly, avoiding reconstruction distortion fitting, using a GMM description based off of atomic models and are capable of representing low resolution data. We then extend the work by modifying the optimization method and fitting more complex structures. This work is the first of its kind in that the focus of the method is fitting a tilt series directly as opposed to a volumetric reconstruction of a tilt series.

The body of the research presented in this document is based on the following publications:

- Kasra Manavi, Bridget S. Wilson, and Lydia Tapia. Simulation and analysis of antibody aggregation on cell surfaces using motion planning and graph analysis. In Proc. of the Association for Computing Machinery Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB), August 2012.

- Kasra Manavi, Alan Kuntz, and Lydia Tapia. Geometrical insights into the process of antibody aggregation. In Proc. of the Association for the Advancement of Artificial Intelligence Conference Workshop on Artificial Intelligence and Robotics Methods in Computational Biology (AAAI-AIRMCB), July 2013.

- Kasra Manavi and Lydia Tapia. Influence of model resolution on antibody aggregation simulations. In Proc. of the Robotics Science and Systems Workshop on Robotics Methods for Structural and Dynamic Modeling of Molecular Systems (RSS-RMMS), July 2014.

- Brittany Hoard, Bruna Jacobson, Kasra Manavi and Lydia Tapia. Extending Rule-Based Methods to Model Molecular Geometry, In Proc. of the Institute of Electrical and Electronics Engineers International Conference on Bioinformatics and Biomedicine (IEEE-BIBM), November 2015.

- Kasra Manavi, Bruna Jacobson, Brittany Hoard, and Lydia Tapia. Influence of model resolution on geometric simulations of antibody aggregation. Robotica, May 2016.

- Brittany Hoard, Bruna Jacobson, Kasra Manavi and Lydia Tapia. Extending Rule-Based Methods to Model Molecular Geometry and 3D Model Resolution. BMC Systems Biology, August 2016

- Gaussian Mixture Models for Fitting Tomographic Tilt Series. U.S. Provisional Patent filed. Inventors: Lydia Tapia, Kasra Manavi, Bridget Wilson, and Niels Volkmann, July 2017

- Kasra Manavi, Sahba Tashakkori and Lydia Tapia. Gaussian mixture models with constrained flexibility for fitting tomographic tilt series. In Proc. of the ACM-BCB Computational Structural Biology Workshop (CSBW), August 2017.

- Kasra Manavi *et. al.*, Fitting Tomographic Tilt Series Using Gaussian Mixture Models and Genetic Algorithm Optimization, *In Preparation*.

## 1.3 Outline

A summary of related work in the field is presented in Chapter 2. In Chapter 3 we provide an overview of rigid body modeling of the molecular aggregate problem. We the move on the a discussion of the methods we developed to perform flexible model construction and fitting in Chapter 4. Conclusions and future work are discussed in Chapter 5.

# Chapter 2

# Related Work

In our work, we focus on the structural simulation of molecular aggregation. We do this with two goals, one to understand the aggregation process and the other to determine aggregate structures. This chapter gives necessary background from three diverse fields, reviewing areas of research related to our work. We start with a discussion on methods for simulating molecular motion and aggregation in Section 2.1. From there we move on to a description of human allergy immune response resulting from IgE antibody aggregation in Section 2.2. We move on to Section 2.3 discussing methods for molecular structure determination. We conclude with Section 2.4 introducing the molecules pertinent to our IgE antibody aggregation simulations.

## 2.1 Molecular Motion Simulation

Aggregation simulations pose special challenges due to their large sizes (both large molecules and large number of interacting molecules) and long timescales. Large simulations are often sped up by introducing constraints into the system that remove fast degrees of freedom, e.g., static bond length [103]. Rigid body molecules are ultimately constrained: all torsions, angles and bonds are fixed at their equilibrium distances [103]. This allows for an increase of the basic 1 femtosecond timestep in dynamics simulation [103, 102]. Monte Carlo studies allow even larger simulation steps resulting in significant speedup but can only provide physically relevant dynamics in certain (special) cases [37, 22, 105].

Coarse graining is another way to accelerate simulation by reducing the cost of energy/force calculations. In coarse graining, molecules or molecular subunits can be represented by charged spheres [70, 69, 58] or point masses [89]. These coarse graining approaches work well for reducing the complexity of the molecular representation, allowing for dynamics studies that can reveal the kinetic factors impacting assembly, but can have difficulty providing insights into questions that require more geometric details. Other methods that do not reduce the molecular geometry can yield detailed pictures of kinetic factors, steric hindrance and non-specific binding [43, 52].

In order to preserve structural information, other studies have incorporated polygon-based models of molecular surfaces in order to simulate interactions [76, 20, 21, 40, 77, 120]. These papers highlight the recent push to gain a more detailed understanding of the role of biomolecular surfaces during interaction modeling [90]. Surface representations of molecules have been used to study protein shape and model cavities, e.g., tunnels and clefts [76, 20, 21, 77, 120]. Even the prediction of molecular binding specificity and protein docking has been shown to benefit from the use of polygon-based models [20, 40]. Large scale assembly processes have been shown to benefit from easy and understandable visualization [35]. Another approach to reduce computational cost while maintaining critical structural details has been the advent of multi-scale methods (using a combination of low and high resolution data) [41, 109, 125]. For example, models of membrane-bound molecules are studied with three methods in [125]: molecular dynamics to study the inter-domain flexibility, Monte Carlo simulations to study multi-domain motion and lattice simulations to study clustering.

### 2.1.1 Design/Prediction of Assembled Molecule Structures

Engineering proteins to perform specific tasks include the design of interfaces between proteins and small molecules [66, 6] and designing protein assemblies [62]. Our work is more related to the latter, focusing on molecular aggregation and structure of assemblies. Methods for designing protein-based assemblies come in two forms: stochastic (resulting in irregular structures with probability-derived attributes), and deterministic (producing exactly specified geometric features) [62]. A majority of the computational design methods have focused on interface construction [49, 53], but new methods go further and fully design self-assembling molecules [59]. These methods are similar in that they generally start out performing rigid body docking followed by iterative design/minimization steps to refine the interface. Tools such as Rosetta [67] are becoming increasingly powerful in enabling the design of these molecular structures.

A wide ranging set of fields, from medicine to industrial manufacturing, stand to benefit from the use of computational methods to determine possible geometric structures of assembled molecules. Many methods use lattice models with force fields and focus on the interactions of proteins with both denatured [18] and native [130] conformations. A coarse grained MD-based approach to study polymer-drug aggregation was done in [93].

### 2.1.2 Robotics Methods for Molecular Simulation

Methods which utilize robotic motion planning for molecular simulations have been reviewed in [3]. These methods focus on using the notion of configuration space ($C$-space) [74], the state space of all possible configurations a robot can attain. For molecular simulations, molecular structures are viewed as robots. A configuration, $c$, defines a pose of the robot/molecule that lies in the $C$-space , $C$, the set of all possible configurations the robot can take. There may be restricted areas of the space due to collision or other constraints that are included in the subset $C_{obst}$. For molecules, these restricted areas are also delineated by energy of the configuration. All remaining configurations are in the subset $C_{free}$ which represents feasible poses. From this perspective of $C$-space, the molecular simulation problem becomes the motion planning problem of finding a trajectory that is fully contained in $C_{free}$ that connects start and goal configurations. To leverage $C$-space, sampling-based planners sample a set of configurations from the $C$-space and uses them to provide a characterization of the $C$-space.

Two major motion planning techniques have been derived from the $C$-space formalization, Probabilistic Roadmap Methods (PRMs) [56] and Rapidly-exploring Random Trees (RRTs) [64]. In PRMs, sample configurations are drawn from $C$-space and those in $C_{free}$ are stored in a roadmap (graph) used to find trajectories. In RRTs, a start configuration is used as the root of an iteratively constructed tree that is grown explore $C$-space. Motion planning methods have been used to study protein folding [7, 117, 116], RNA folding [115], and ligand (small molecule) binding [112, 11]. The extensions required for applying motion planning algorithms into molecular simulations include molecular representation, collision detection and energy calculation.

A variant of RRT, Manhattan-Like RRT (ML-RRT) [25], has been proposed for disassembly path planning, the problem of computing motions to disassemble objects. This is done by decoupling the motions of different parts of the system. These methods have been applied to receptor-ligand complexes [24] focusing on the exit problem, finding a trajectory to remove a bound ligand (small molecule) from a receptor and determine which parts of the molecules need to move to find a solution [24]. The combination of decoupled motions and $C$-space exploration was able to reduce computation time compared to RRT and RRT variants.

Other protein simulation methods focus on multiscale modeling or utilizing prior constraints. In [4], Normal Mode Analysis (NMA) of a coarse grained elastic network is used to compute large-scale motions of biomolecules. Normal modes, eigenvectors of the Hessian matrix, are used to predict the low frequency modes of motions of a given molecular structure. RRT is used to explore the linear combinations of modes generated from the sampled configuration. In [98], the authors present a generalized setup for including prior information into RRTs. The method utilizes prior information including atom distances, helix line fit, and secondary structure RMSD to bias the path towards external constraints. It was shown that partial information can improve sampling-based method performance by reducing the $C$-space size.

## 2.2 IgE Antibody Aggregation

The human allergy immune response is initiated when cell surface bound IgE-Fc$\epsilon$RIs crosslink (bind together via antigen), forming clusters which signal the cell for degranulation [100]. Experimental studies using nanoparticles have shown antigen size and valency (number of antigen binding sites) impact degranulation of rat basophilic leukemia cells [50]. Spatiotemporal analysis of IgE aggregation has been done using nanoscale imaging and motion tracking techniques [128, 9, 123]. Methods to analyze clustering of micrograph probes were developed in [128], including Ripley and Hopkins statistic calculations. These calculations use the locations of static gold nanoparticle labeled IgE-Fc$\epsilon$RI which have been imaged using transmission electron microscopy [123]. Spatial clustering analysis of IgE-Fc$\epsilon$RI has been done using methods from [128] as well as hierarchical clustering techniques to quantify the numbers and sizes of clusters [27]. Tracking of quantum dot labeled IgE-Fc$\epsilon$RI has determined temporal information such as diffusion rates [9]. While these experimental methods have been able to measure attributes about receptor dynamics, they do not retain information about the aggregate binding patterns. Because of this, distinguishing linked (bound) molecules from simply proximal receptors is challenging.

Rule-based modeling has been used to model antibody-antigen interactions. A kinetic rule-based version of the Trivalent Antigens and Bivalent Receptors (TLBR) system which accounted for two types of cycles (dimers and heximers) was introduced in [127]. This method was based on a previous equilibrium theory model [36]. This rule-based model was extended to consider aggregate structure constraints [89]. In our prior work, we developed a rule-based model which considers steric constraints, but using a different antigen [46]. This was achieved by rules that considered molecular geometry.

## 2.3 Electron Microscopy for Molecular Structure Determination

Fitting a known structure to experimentally imaged molecules is critical to understanding molecular conformations. In this section we review Cryo ET and the molecular replacement problem (Section 2.3.1). We then review more specific aspects of Cryo ET that we focus on in our work (Section 2.3.2).

### 2.3.1 Cryo Electron Microscopy

Electron Microscopy (EM) has become an essential part of understanding cellular function [33, 65]. EM is performed by preparing a sample and placing it into an electron microscope for imaging. This sample is exposed to an electron beam which is collected via a detector, producing an image of the sample. In this image, the intensity of the pixel is proportional to the density of the 3D object.

There are various methods of EM [33], including Cryo ET [32]. In Cryo ET, preparation includes flash freezing the sample, locking the object of interest in vitreous (disordered) ice. The sample is then rotated about an axis and imaged, producing an ordered set of images from different perspectives called a tilt series. The tilt series is processed using image analysis techniques to perform volume reconstruction of the structure. This volume is then typically fit using structural models to determine the molecular conformation.

#### Sample Preparation and Imaging

Processing a sample begins with the preparation of a sample and its placement on an EM grid [114]. An EM grid is a fine-mesh copper disk with a thin layer of carbon added to the top surface. The disk mesh granularity and carbon layer pattern used are dependent on the type of sample being imaged, e.g., molecular structure size and collection technique. Once the sample is placed on the grid, the sample is then frozen in place. This is done by quickly placing the grid into a cold environment, typically performed by plunging the grid into liquid ethane.

At this point, the sample is prepared for imaging using an electron microscope. There are two main methods for imaging a given sample, untilted and tilted imaging [33, 32]. In untilted imaging, the sample is placed in a microscope and exposed to an electron beam to produce an image. This results in images being collected in an unordered fashion, i.e., the relationship between perspectives needs to be inferred from the data as a post process. In tilted imaging, a sample is placed in the microscope and imaged, then rotated about an axis and imaged again until a series of images is collected. Tilted imaging produces images where the relationship between perspectives is known due to the ordered nature of the collection process, i.e., the transformation matrices between perspectives about an axis of rotation are know.

### Volume Reconstruction and Fitting

The volume reconstruction and fitting process is outlined in Figure 2.1. If the sample is processed using untilted imaging, analysis of images is performed using single particle reconstruction techniques [33]. This process begins with picking a set of particles from the set of images produced from the microscope (the more particles the better). Once this set of particles has been picked, the images are aligned and averaged to produce a clearer view from a given perspective. These different perspectives are then combined to create a 3D volume of the structure using methods such as random canonical tilt or common lines [33].

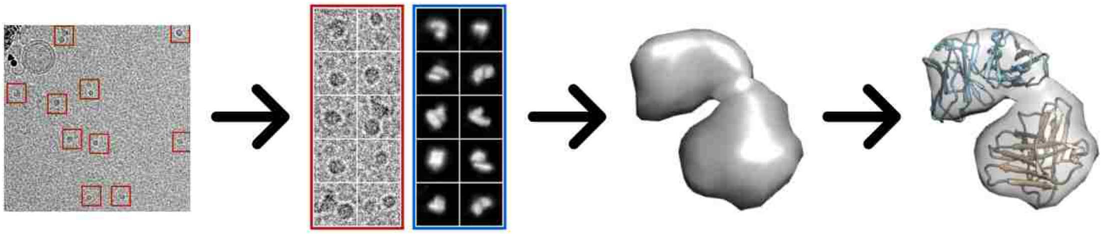


Figure 2.1: A diagram outlining the process of reconstruction and fitting of EM data. (*Far Left*) Particles are picked from the dataset produced by the electron microscope. (*Center Left*) Picked particles (*red/left*) area are aligned / averaged / processed to produce projections for reconstruction (*blue/right*). (*Center Right*) Processed projections are used to reconstruction a 3D volume of the structure. (*Far Right*) Fitting techniques are used to fit atomic model to 3D reconstruction.

If the sample was processed using a tilted imaging solution, tomography can be used to reconstruct a volume [32]. Tomography is a technique that utilizes a series of images to produce a 3D reconstruction of the underlying structure. Images are captured in an ordered fashion (e.g., single/dual axis tilt series) and are combined to produce a 3D density map.

EM and Cryo ET are becoming cornerstones of modern structural biology research, but several challenges remain, including reconstruction evaluation and model fitting [42].

Depending on the type of sample, density map reconstruction can be challenging for both single particle and tomographic methods. This step, required before an atomic model can be fit, can suffer from issues including distortions in image alignment and the missing wedge problem [75]. During single particle reconstruction, images of a dataset are clustered and aligned to produce an averaged image from a particular perspective. The quality of the clustering/alignment results are highly dependent on the size of the input dataset and image quality. Distortions in a given perspective occur when not enough images are gathered to properly describe the sample from the given perspective. When performing tomographic reconstruction of a tilt series, challenges arise due to the missing wedge problem. This problem is due to the nature of the image collection method and how it has the potential to miss features of the molecule due to the limited angle range. A set of perspectives are captured by rotating the sample about an axis, but if the angle range is narrow, portions of the sample are not imaged. This results in a loss of information about the sample and reconstructions are elongated perpendicularly to the axis of rotation.

In addition to the challenge of reconstruction, fitting structural models, typically all-atom structures, to reconstructed density maps can be difficult. Resolution of a density map is important to the type of fitting technique used. At high resolution ($<$ 10 Å), all-atom fitting techniques work well, but lower resolutions (20-40 Å) still pose a challenge to existing methods due to a lack of detail [121]. This is particularly true for larger asymmetric molecular systems that are typically imaged at lower resolutions.

### 2.3.2 Projection and Tomography

In this section we review work related to projection matching, tomographic reconstruction, and antibody structures generated using EM.

#### Projection Matching

Several computational geometry methods have been used to model volumes from projection information and vice versa [13, 101, 88, 124]. For example, the optimal packing/covering problems have been solved using phi-functions developed to evaluate the interaction of geometric objects called phi-objects [13]. Phi-functions take in two phi-object positions as input and returns a value inversely related to the amount overlap between the objects (negative if they overlap, zero if they touch, and positive if they are separated). Another example is shadow art, the idea of occluding light from a source to produce an image. One such method uses light sources and a desired shadow art as input and produces a sculpture that is capable of generating the scene [88]. This work has been expanded to model shadow theater where shadow art is generated by the pose of a single or multiple performance artists [124]. Another similar method turns 2D silhouettes from different perspectives into 3D models [101].

### EM and Structure Determination

Integrating comparative modeling and EM data to produce atomic models is reviewed in [119]. Fitting structures to reconstructed 3D EM data can be broken down into two main methodologies, rigid and flexible [29]. Most six degree of freedom rigid fitting is done using methods like geometric techniques [19], GMMs [57], or Zernike descriptors [28]. Flexible fitting focuses on using molecular simulation methods [108, 72], robotic motion planning techniques [5], as well as statistical techniques [107, 26] to determine candidate conformations. GMMs have been applied to other aspects of EM analysis, including reconstruction of single particle imaging [55] and structural dynamic evaluation [54].

### Antibody Structure Determination

In our work, we focus our analysis on the IgE antibody, responsible for the human allergic immune response. Immunoglobulin (Ig) proteins have been determined to be highly flexible and can form asymmetrical structures [106, 17]. The structures obtained from X-ray diffraction analysis show that IgG, another member of Ig protein family, is composed of three major structural subunits: two identical binding arms (Fab arms) and a membrane bound constant domain [104]. Since antibodies are known to have very flexible and dynamic structures, populations of different conformations have been found to co-exist in images [129, 118]. Therefore, commonly used methods such as X-ray crystallography, which rely on molecular averaging, often do not reflect protein dynamics and flexibility [131]. In contrast, EM can be used to reconstruct unique and independent samples [104].

## 2.4 Molecular Structures

In our work, we focus on the human allergy immune response resulting from the aggregation of IgE antibodies via antigen. An all-atom structure of the IgE-Fc$\epsilon$RI complex was initially described in [78]. The IgE structure, composed of both heavy and light chains, is modeled bound to the $\alpha$-subunit of the cell surface receptor Fc$\epsilon$RI as shown in Figure 2.2. The receptor complex model was constructed using available molecular structures from the Protein Data Bank [14] (PDBs: 1OAU, 2VWE, 1O0V, 1F6A) and is composed of 1,709 amino acids (13,477 atoms). The Y-shaped structure has 3 major regions, the lower region referred to as the constant domain and the upper regions called Fab arms. Antigen bind to the antibody at the HyperVariable (HV) regions of IgE located at the end of the Fab arms.

Much work has been done to study synthetic antigens that initiate an immune response. Synthetic antigens such as the bivalent (Dct)2-cys (DCT2), trivalent DF3, and multivalent DNP-BSA$_n$ have been constructed to study receptor aggregation [97, 111, 126]. The structures of these antigens are well known and documented. All these structures use DNP, a hapten used in molecular biology to bind to DNP-

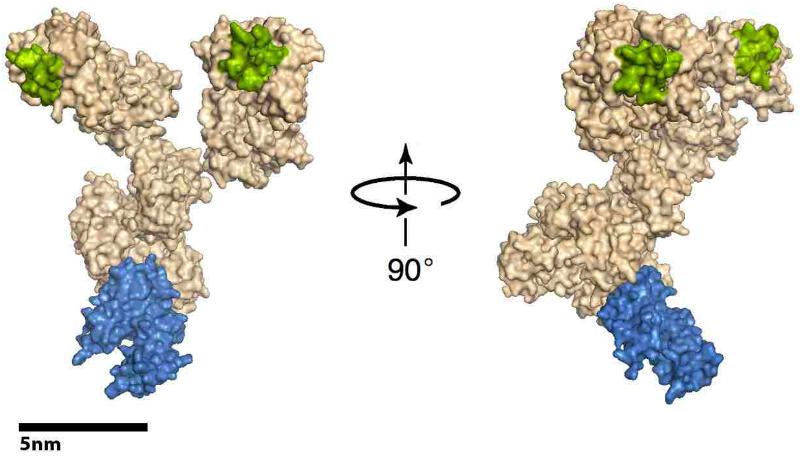


Figure 2.2: The molecular structure of IgE antibody/cell surface receptor complex. The $\alpha$-subunit of the Fc$\epsilon$RI receptor (*blue*) has a transmembrane domain keeping it tethered to the cell surface. The antibody (*tan*) is tightly bound to the extracellular region of the receptor. Two (2) antigen binding sites are located at the ends of the antibody arms (*green*), resulting in a bivalent molecule.

specific IgE antibodies, resulting in model systems with high immunogenicity. Each antigen binding site is comprised of a DNP linker that has been attached to the molecular structure.

In our preliminary experiments we use models of antigen DCT2 and DF3. DCT2 is a synthetic bivalent antigen (Figure 2.3) with 2 binding sites on opposite sides of the molecule. DF3 is another synthetic antigen (Figure 4.11) with 3 binding sites.

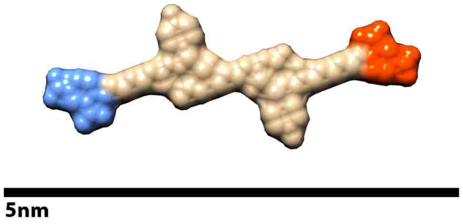


Figure 2.3: The molecular structure of synthetic antigen DCT2 (*tan*). The molecule is symmetric and the DNP linkers (*blue and red*) are attached at both ends, creating a bivalent antigen.

These antigens have reduced valency compared to DNP-BSA$_n$, a synthetic multi-valent ($n$ being the valency) antigen used to study receptor aggregation. The number of DNP linkers bound to BSA$_n$ can vary (2-25 binding sites). However, there is no control of spatial distribution of the binding sites on BSA$_n$, so there is no guarantee of binding site uniformity. On the other hand, DCT2 and DF3 have well defined structures with known binding site locations, making them easier to model.

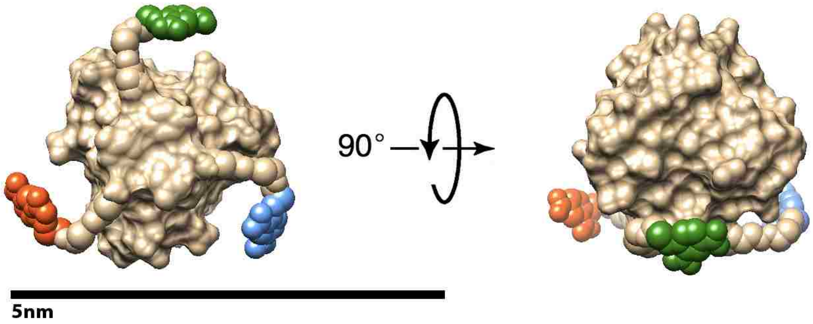


Figure 2.4: The molecular structure of synthetic antigen DF3 (*tan*). The fibritin trimer has 3 DNP linkers (*red, green and blue*) attached to the N-termini of trimer subunit.

The molecular structure of antigen DCT2 was generated from the PubChem open chemistry database (SID:135154086), and is composed of 78 atoms. DF3 was generated by starting with the base fibritin trimer (PDB:1RFO) [38] and adding flexible DNP linkers (about 1 nm in length) to the N-terminus of each of the three subunits. DF3 is comprised of 81 amino acids with 1,365 atoms total.

Synthetic antigen allowed us to probe the system with ideal models, but studies of natural allergens initiating degranulation have more pertinence to human studies. One allergen that has been of particular interest is the common shrimp allergen, Pen a 1. The immune response is triggered by the shrimp tropomyosin molecule, a 40 nm double-stranded coiled coil structure, (Figure 2.5), which crosslinks IgE antibodies. The allergen has been predicted to have 5 binding regions on each of the strands of the coiled coil [10] and a total of 16-18 binding sites [51, 99]. Structural models for shrimp tropomyosin were available in the Protein Data Bank (PDB:1C1G) [122] and in the Structural Database of Allergenic Proteins (SDAP Model: #284) [10]. The Pen a 1 model used was composed of 568 amino acids totaling 4,580 atoms.

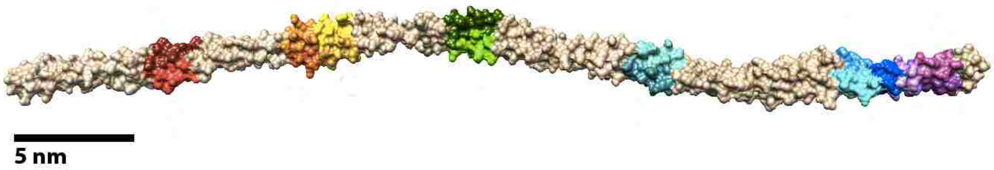


Figure 2.5: The molecular structure of Pen a 1, a common shrimp allergen (tan). A total of 16-18 binding sites (*various colors*) are located in 5 regions on the coiled coil structure.

# Chapter 3

# Rigid Body Molecular Modeling

Our work developing multi-resolution models starts with a focus on methods to simulate and analyze molecular interactions for modeling molecular aggregation. In this chapter we reduce the complexity of the problem by simulating rigid bodies and using a model description based on molecular geometry. This representation allows us to efficiently simulate molecular interactions and gain insights into the aggregation of IgE antibodies when exposed to a given allergen. We begin with an introduction to our methods for model construction in Section 3.1. We then talk about the sampling techniques and methods used for our simulation of molecular motions and aggregation in Section 3.2, and corresponding analysis techniques in Section 3.3. Finally, we present the three sets of results from material published in [84, 81, 82, 80] in Section 3.4 based on the simulation of IgE aggregation.

## 3.1 Rigid Molecular Model Construction

Before simulations can be run, rigid geometric models of the molecular structures described in Section 2.4 need to be created. Here we describe the methods developed to generate reduced resolution rigid molecular models.

### 3.1.1 Resolution Reduction

Since it would be computationally prohibitive to use all-atom models at the molecule counts we simulate, we reduce the complexity by only modeling molecular geometry (Figure 3.1). To construct our geometric models, we begin with an all-atom structure. Using the multi-scale model extension of UCSF Chimera 1.9 [95], we generate isosurface models of the molecules. An isosurface represents points of constant value of a variable in space like an isoline does on planes. In our instance, atomic density described as a volume is used to render isosurface models at specified values of density. The volume described by the surface generated indicates where space is occupied by the molecular model. We generate models at resolutions ranging from 4 Å for the smaller molecules and 6 Å for the larger molecules.

The resulting model of the occupied volume, referred to henceforth as the base model, is considered to be the model with the highest geometric resolution, i.e., the most detailed model (Figure 3.2 *top left*). Due to the nature of isosurface construction, the base models generated are highly detailed (contain many polygons). Unfortunately, this high level of detail in the geometry hinders performance of conformation validity checking because binding site proximity and isosurface collision detection calculations are dependent on the geometric detail. To overcome this obstacle, we evaluate the cost versus benefit of decreasing model resolution.

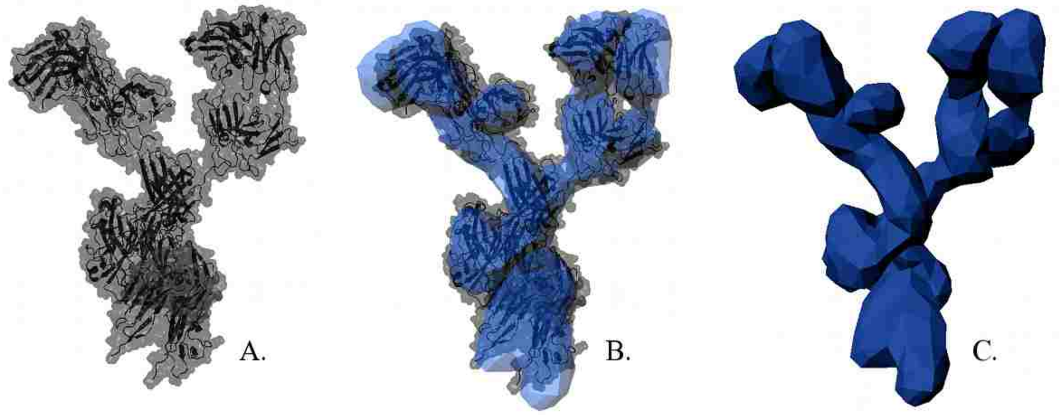


Figure 3.1: Models are polygonal reductions of the iso-surface of the all-atom molecular structure. The relationship between these representations is shown above: A.) High resolution iso-surface of the all-atom model, B.) the polygon model shown overlaying the iso-surface, and C.) final reduced polygon model.

The base model can be reduced in complexity using standard polygon reduction techniques which include controlled vertex/edge/face decimation, vertex clustering and mesh optimization [23]. To do this, we use the polygon reduction feature in Maya [85], a modeling software package, that allows the generation of models with a reduction specified by the percentage of polygons to be retained. Figure 3.2 displays our model construction process, starting with base model generation and the production of reduced models at various resolutions.

### 3.1.2 Binding Region Modeling

To describe binding events with our rigid body models, we use regions to describe binding sites. Binding events are triggered when the regions of two interacting molecules overlap. During the simulation, bound molecule are held rigid relative to each other.

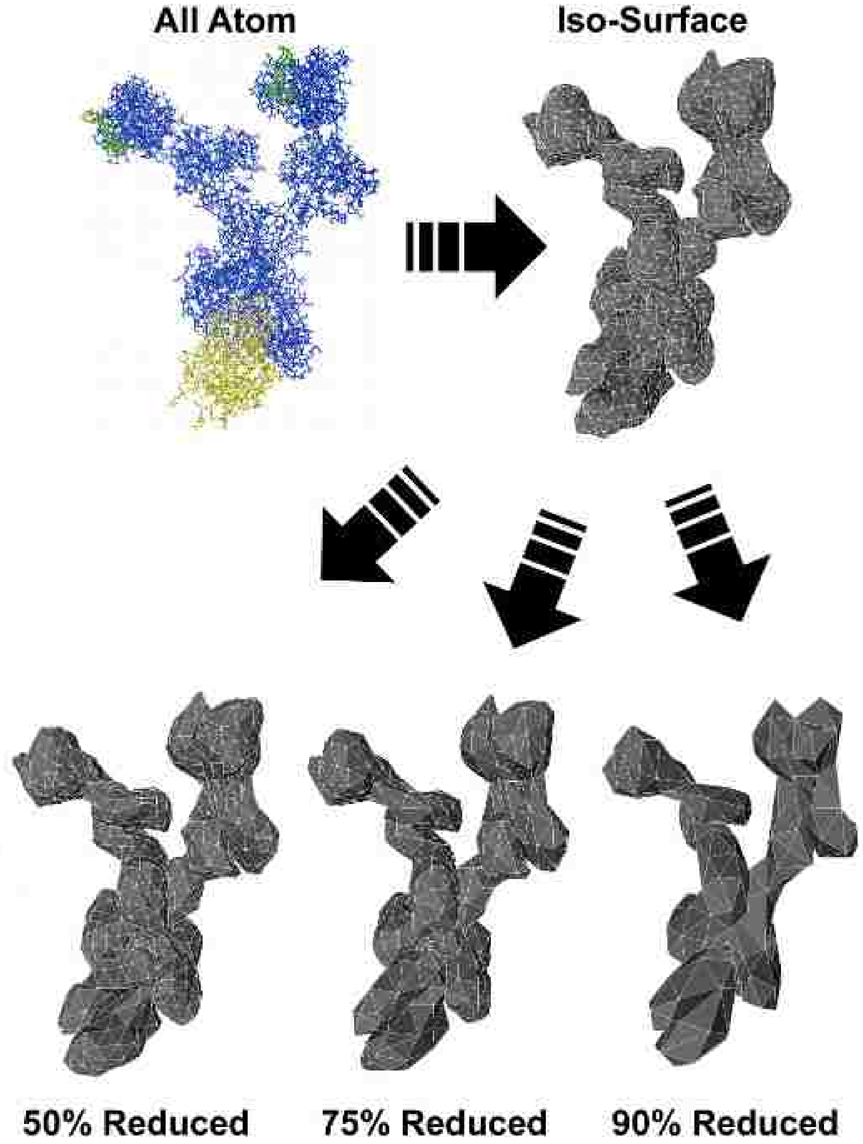


Figure 3.2: The model construction process starting with an all-atom model (IgE-Fc$\epsilon$RI) (*top left*), generating the iso-surface base model (*top right*), and then applying polygon reduction to generate a wide array of models with lower resolutions (*bottom*).

The binding sites on an IgE antibody are assigned to be the HV regions that reside at the end of the Fab arms (Figure 2.2). To model the binding regions, we use spherical description with a vertex describing the region center and a radius for the region boundary. We calculated the HV region's center of mass and found the closest solvent accessible residue in the HV region to define the region center. These vertices representing the binding locations were found using the all-atom model and added to the geometric model. We used a radius of 5 Å from the vertices to describe regions of binding for IgE.

Antigen DCT2 and DF3 have flexible DNP linkers that bind to the antibody binding sites [47, 78]. For these DNP-based antigens, we model a spherical binding volume centered at the DNP linker's center of mass with a radius of half the linker length (7.5 Å).

Pen a 1 binding sites are located on the surface of the molecule. The binding sites of Pen a 1 were identified using an algorithm for linear epitope prediction [51], a method of epitope prediction utilizing protein sequence information. Vertices on surface of Pen a 1 near the center of the amino acids involved in binding were located using the all-atom model and then added to the geometric models. A binding radius of 3 Å was used to describe the binding regions of Pen a 1.

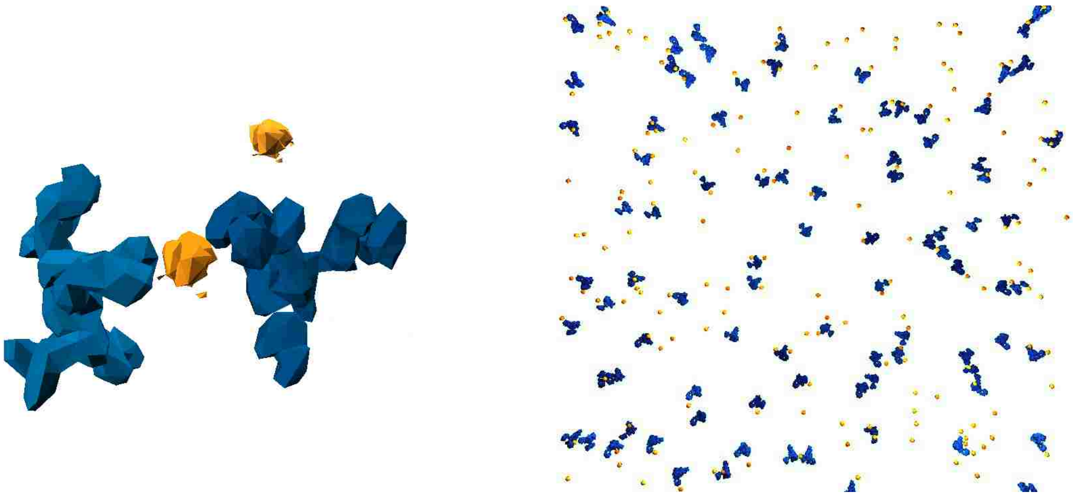
## 3.2 Aggregation Simulation

In this section we discuss our methods for molecular simulation. For the antibody aggregation problem, we focus on simulating the molecular motions that result in the production of aggregate structures [84, 81, 82, 80].

To generate potential aggregate conformations, we simulate antibodies and antigen interaction. We do this by modeling the molecular interactions using a simple Monte Carlo-based simulation and using a graph to maintain connectivity information. Simulations are initialized with randomly placed receptors and antigens in a bounding volume in a collision-free state. The molecules are allowed to move on the XY-plane and rotate about the Z-axis, defining three DOF per molecule. We use three DOF for three reasons: 1.) The planar nature of the cell surface at the resolution we model, 2.) Creating ideal conditions for antibodies and antigen to bind to understand their structure / composition, and 3.) Reducing the computational cost to increase efficiency. The complexity of the simulation (total number of DOF) depends primarily on the number of molecules simulated. For example, 20 molecules requires exploration of a 60 DOF $C_{space}$. Figure 3.3 shows small and large scale examples of our simulation.

Algorithm 3.1 outlines how the receptors and antigens move and how binding events are handled. At each time interval, a Monte Carlo step is taken and all positions of the molecules in the simulation are updated. This step is determined using random sampling, a technique often used to solve high-dimensional motion planning [56] problems. Within our random sampling scheme, biological constraints of the system are considered, e.g., molecule speeds and rotation correlation times, in addition to association and dissociation rates. This means that at each time step, every pair of molecules whose binding regions are overlapping will bind with a probability defined by the association rate. Alternatively, each bound pair of molecules is probabilistically evaluated for bond breakage according to the dissociation rate.

As the molecules move in the bounding volume, receptors and antigens begin to bind and form aggregates. It has been shown larger aggregates have slower diffusion on the cell surface [9]: As the aggregates increase in size, the collection



(a) Small simulation

(b) Large simulation

Figure 3.3: Simulations of different sizes. a.) Small simulation with 2 receptors and 2 antigens. The 2 receptors (*blue*) are bound by an antigen (*yellow*), and there is a second free antigen. b.) Large simulation with 90 receptors and 180 antigens. The state is near the beginning of the experiment, showing a well mixed system with some early binding.

of molecules as a whole slows down and begins to move at a reduced speed. Simulations are run until a stopping criteria is met, e.g., stable graph formation or time step limit reached.

## 3.3 IgE Aggregate Analysis

In order to simplify the analysis of assembled structures, we formulated a graph-based structure to capture molecular interactions. In these graphs, there are two classes of molecules, receptors and antigens, which are represented as vertices in the graph with different labels. If a receptor binds to an antigen, it forms an edge in the graph to represent the bond between the two molecules. Such graph allows us to encode the molecular structure of aggregates in a simple representation, enabling us to efficiently simulate and analyze complex aggregates.

At each timestep, $i$, a graph $G_i$ of all molecules in the simulation and connections made between them is saved. The graph of the final state of the system, $G_{final}$, is saved at the end of the simulation. These graphs can be analyzed using standard graph metric tools. For example, when the simulation reaches steady state, the number of edges in graphs of consecutive timesteps ($G_n$,..., $G_{final}$) should stabilize around an average value. Also, the number and sizes of connected components in $G_{final}$ measures the number and sizes of aggregates formed.

Thus, the graph $G_i$ contains information about free (non-connected vertices) and aggregated molecules (connected vertices) of a given simulation. Since antigen only binds to receptors and vice versa, the graph is bipartite. An individual aggregate structure at time $j$, $a_j$, can be identified as a connected component found in $G_i$. We focus our analysis on the set of all aggregate structures, $A$, found in $G_{final}$. From our preliminary work, we highlight two graph-based analysis that we can run on aggregate structures: aggregate classification and common aggregate substructure. Traditional graph-based techniques can be simply applied. For example, classification is performed using a depth first search traversal, and commonly formed aggregate structures are identified through subgraph isomorphism.

### 3.3.1 DF3 Aggregate Analysis

In order to characterize the aggregate structures generated using antigen DF3, we used a graph traversal algorithm. The characteristics of the aggregate structures allow us to define four major classifications (shown in Figure 3.4):

- **Singleton** (1 receptor with at least 1 antigen bound)
- **Linear Chain** (2 or more receptors forming a chain)
- **Cyclic *n*-mer** (2 or more receptors forming a cycle)
- **Complex Aggregate** (3 or more receptors forming a combination of single bound receptors, linear chains and cyclic *n*-mers)

When an aggregate has two vertices, it must be a Singleton. When an aggregate has three vertices, two of which are antigens, it also will be labeled a Singleton. On the other hand, a graph with three vertices with only one being a antigen is a Linear Chain. Aggregates of four vertices or larger are distinguished as Linear Chains, Cyclic *n*-mers, or Complex Aggregates. If these larger aggregates are traversed and no repeated vertices are seen, it is labeled a Linear Chain. However, if a single cycle exists in the graph, then the structure is labeled a Cyclic *n*-mer where *n* refers to the number of receptors. The final structure category, Complex Aggregate, is identified when finding multiple repeated-molecules or extra molecules beyond those in a Cyclic *n*-mer. This means Complex Aggregates can be any combination of Linear Chains and Cyclic *n*-mers.

These classifications are based on experimental studies of IgE aggregation. The labeling of Fc$\epsilon$RI with nanogold particles produces 2D plots of dark spots [123] which can be used to compare to our receptor positions. Regular structures such as Linear Chains and Cyclic *n*-mers may be identifiable, but the remaining classifications can be difficult, if not impossible, to distinguish.

Classification is performed by traversing the aggregate graph using depth-first search. This produces a search tree where cycles and dead ends can be identified. Each of these is a feature that can be used to detect the four aggregate classifica-

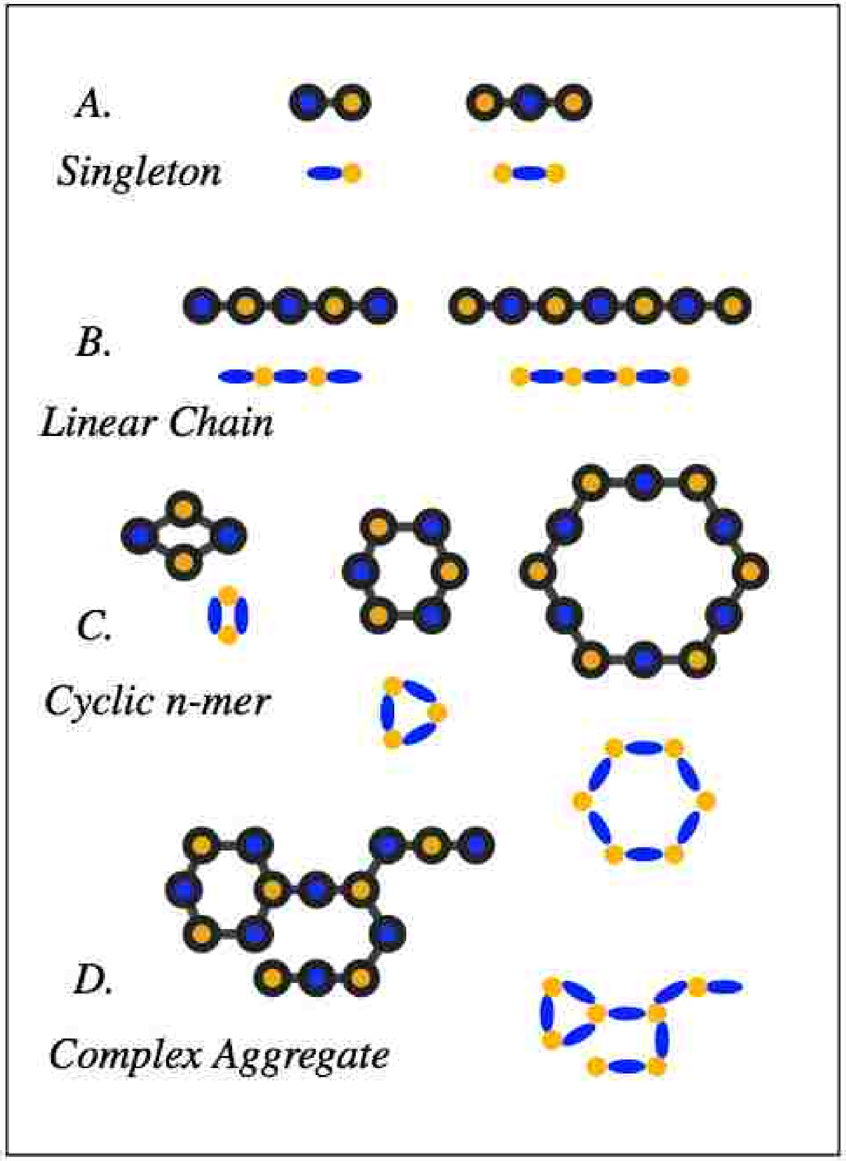


Figure 3.4: Aggregates represented as graphs. The diagrams demonstrate sample receptor *(blue)* and antigen *(yellow)* binding patterns. The aggregates are classified into 4 categories: *A.)* Singletons are just single receptors bound to a antigen or two, *B.)* Linear Chains are two or more receptors forming a sequential chain, *C.)* Cyclic *n*-mers are where two or more receptors form a cycle, and *D.)* Complex Aggregates are made up of combinations of Linear Chains and Cyclic *n*-mers, in this case a Cyclic trimer and two Linear Chains.

tions. For example, Linear Chains have no cycles but do have dead ends. On the other hand, Cyclic *n*-mers have only cycles but no dead ends. Complex Aggregates consist of combinations of cycles and dead ends.

The ability to identify common aggregate substructure could provide insight into likely and common aggregate formations. We use subgraph isomorphism in order to identify the largest and most frequently occurring aggregate formations. McGregor's common subgraph algorithm [86] can extract these substructures from the aggregate graphs produced by the simulation.

**Spatial Clustering Analysis**

Due to motion and association, the antigen/receptor positions will become more clustered as aggregates form, eventually reaching a state that initiates the signaling cascade for degranulation. These clusters can be observed experimentally [9, 123], and theoretical studies of clustering, for example [36, 127, 89] can be an instrument to compare our model with experiment. To quantify clustering in our models, we use a geometry-based statistical analysis of clustering tendency, the Hopkins statistic [48]. The Hopkins statistic is a measure of spatial randomness which utilizes nearest-neighbor distance of randomly sampled points and randomly selected probes (known molecule locations). We use receptor molecule positions for our cluster analysis in line with previous work [128, 27]. We calculate the Hopkins statistic in a similar fashion as in [128], and for nearest neighbor calculations, we use the Euclidean distance between two points. The values calculated for the Hopkins statistic range from [0,1]. The closer the Hopkins statistic value is to 0.5, the more randomly spaced the points are, whereas the closer the value is to 1.0, the more clustered the data is.

### 3.3.2 Pen a 1 Aggregate Analysis

In addition to the 3-D Monte Carlo method simulating IgE binding to Pen a 1, we developed a supplemental simulation utilizing Rule-Based Modeling (RBM). RBM is an approach to modeling where a set of rules are used to indirectly specify a mathematical model that can be evaluated using Markov chain or differential equation based method. Our RMB incorporates geometric information in both the rules for aggregate formation and rate constants. The RBM is implemented with RuleBender [113] using the BioNetGen language [15]. This method automates the generation of the coupled differential equations associated with the creation of new molecule aggregates as IgE binds to the available binding sites of Pen a 1.

**Steric Analysis Using Rule-Based Modeling**

Steric effects of receptors bound to an antigen with multiple binding sites in relation to binding site exclusion has been analyzed in [44]. However, they only investigated low dimensional (1D and 2D) shapes with specific geometries (surface or array) and either ordered or uniformly-random binding site distributions. This information is critical to understanding what structural characteristics are shared among antigen effective at eliciting an immune response. To gain insights into properties of the volumes of our aggregate structures, we take our modeled aggregates and generate all-atom structures. With these all-atom models, we can take measurements of the aggregate structure and analyze features of the aggregate such as steric hindrance and measure internal distances. We can also quantify the model construction quality. An example Pen a 1 all-atom aggregate is shown in Figure 3.5.

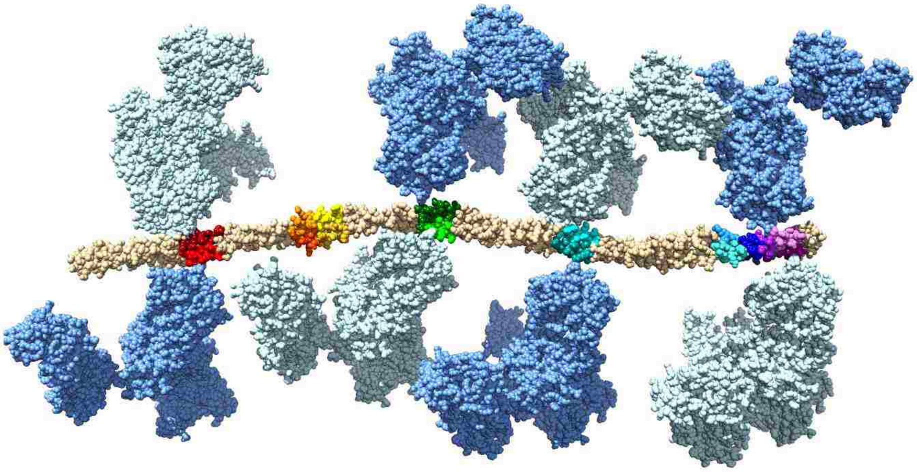


Figure 3.5: An aggregate structure generated using our method. The eight IgE-Fc$\epsilon$RI (light/medium blue) are bound to the Pen a 1 antigen (tan) at various binding sites on the antigen (various colors).

To gain further insights into how antigen conformation plays a role in steric hinderance, we use a novel application of RBM to evaluate the effect. We analyze how the conformation of the antigen affects steric constraints of the system. Due to the size of IgE and the distances between binding sites on Pen a 1, neighboring binding site occupancy is important. A description for dependency on neighbor occupancy is shown in Figure 3.6. Steric hindrance induced by neighbor occupation can be broken down into three categories. First, IgE can easily bind to a region if neighboring sites are free (Figure 3.6 (a)). Second, on the strand with negative curvature around a region, occupation of nearest-neighboring regions can reduce accessibility of IgE to this region, effectively reducing the binding rate constant (Figure 3.6 (b)). Finally, on the opposite side with positive curvature, IgE can still bind to a region even if its nearest neighbors are occupied (Figure 3.6 (c)). The binding rules (listed in the appendix) are written with explicit neighboring site dependency.

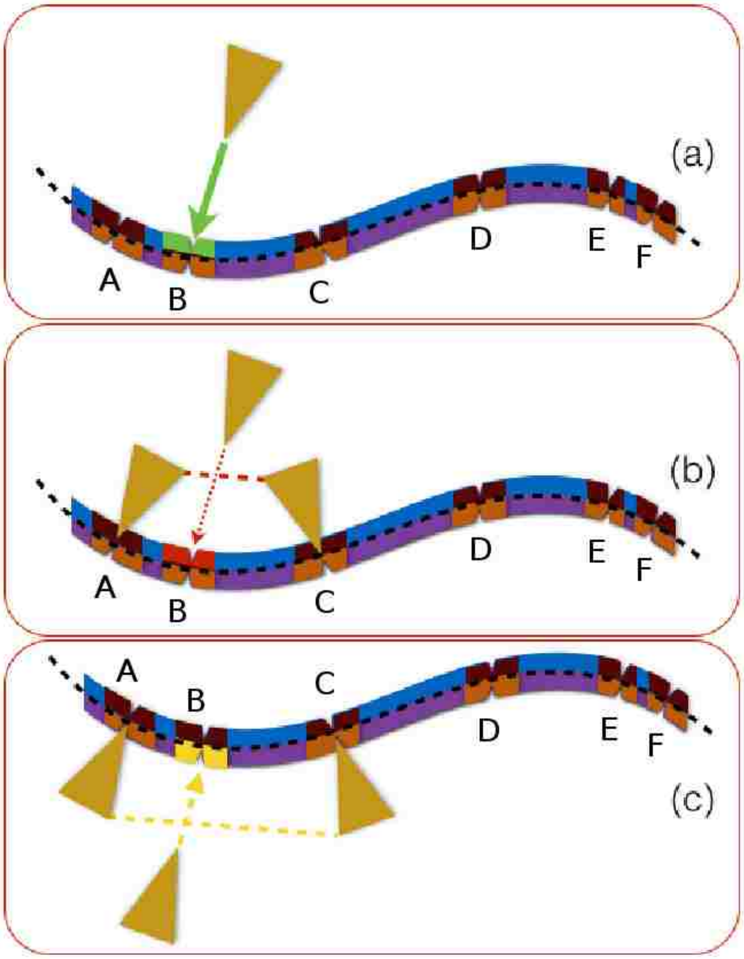


Figure 3.6: Steric hindrance induced by neighbor occupation. The six binding regions are labeled $A, B, C, D, E, F$. (a) No neighbors: receptors are free to bind, (b) Negative curvature reducing binding rate constant, and (c) Positive curvature with possible effect on binding rate.

### Model Construction and Calculations

From this geometric analysis, we know that a negative curvature in the coiled coil may introduce hindrances to the accessible surface area for IgE binding and potentially brings binding sites closer. This makes the accessibility of a receptor to a particular binding site dependent on whether its neighboring sites are bound to IgE or not (namely their occupation states). However, because there are 16-18 available binding sites in the antigen, the introduction of geometric effects may lead to the number of rules becoming too overwhelming to implement, as a dependency on the occupation state of neighboring binding sites has to be explicitly added to some of the rules of binding events. Therefore, we make a few simplifying assumptions to generate the rules and associated ODEs:

- We assume that IgE binds to a single binding site in Pen a 1, i.e., binding events in which IgE binds to two sites on the same Pen a 1 are forbidden.
- To compare with the Monte Carlo simulations, which were carried out for a single Pen a 1 molecule, we do not allow crosslinking through IgE binding to two or more different Pen a 1 molecules.
- We significantly reduce the number of rules (and ODEs) by assuming that each IgE binds to a region on Pen a 1 known to have one or more binding sites. This is a reasonable assumption as binding sites in the same region are close ($<$ 5 nm); In the event of IgE binding to one binding site in a particular region, the other binding site(s) in the same region may be automatically blocked.
- Because the binding region on the tail of Pen a 1 is longer than the others (see Figure 2.5), a more physical representation is made by splitting this longer region into two independent ones in our RBM, resulting in each strand of the coiled-coil structure having six binding regions.
- We can further decrease the number of rules by considering that each strand in Pen a 1 binds IgE independently of the other, i.e., the occupation state of any binding site on one strand of the coiled coil is independent of the occupation state of any binding site on the other strand. Since each strand has six regions, the maximum number of conformations of IgE binding for each

> strand is $2^6 = 64$. The aggregate sizes of IgE-Pen a 1 with 12 binding regions is now given by the combined independent probability of the aggregate formation in each strand of the coiled coil.

The probability of finding aggregates of size zero to twelve is calculated by simulating each strand of the coiled coil separately, with different rules depending on the positive or negative curvature of the strand. The probability $P(n)$ to form an aggregate of size $n$ is given by:

$$\begin{aligned} P(n \leq 6) &= \sum_{m=0}^{n} P_I(m) P_{II}(n-m), \\ P(n > 6) &= \sum_{m=n-6}^{6} P_I(m) P_{II}(n-m), \end{aligned} \tag{3.1}$$

where $P_{I(II)}(n)$ is the independent probability of forming an aggregate of size $n$ in strand $I(II)$.

**Model Rate Constants**

In order to analyze the influence of both rules and rate constants on our rule-based modeling results, we create a set of rules for Pen a 1. The *General* rule set in Tables A1 and A2 takes into account neighboring binding site interactions and employs hierarchical binding rate constants. This means that when IgE binds to any site $i$ on Pen a 1, the associated rate constant depends on the occupation of its nearest (first-order) and next-nearest (second-order) neighbors. We define four hierarchies of binding, thus the binding rate constants are four independent parameters. The unbinding rate constants are equal to 0.01 $s^{-1}$ for all aggregate formation rules. Since the actual rate of Pen a 1 is unknown, we use the rate of DF3 as a substitute. Neighbors are defined by geometry: as the Pen a 1 molecule has a slight S-shaped curvature, binding sites on the concave (negative curvature) sides of the molecule are closer than sites on the convex (positive curvature) side of the molecule.

Resolution changes can be simulated in two ways: by fixing the binding rates and changing the rules for each resolution or by keeping rules fixed and changing the binding rates associated with each rule. The former requires a very careful analysis of the geometry of the receptor/antigen system at each resolution. The latter only needs a set of rules that roughly represents the site interactions given the curvatures of the protein complex, and rates can be changed as binding to particular sites is allowed or disallowed. Therefore, we focus on an implementation of the latter by generating a set of rules which capture the geometric constraints of the molecules and define rate constants that are dependent on the occupation of neighboring sites.

The advantage of using independent rate constants as parameters related to neighbor occupation is that we can simulate high and low resolution studies by turning rules *on* or *off*. The rate constant of a given rule determines whether or not the rule is *on* (non-zero rate constant) or *off* (rate constant set to zero). Rules that favor the formation of large aggregates are turned *on*.As our Monte Carlo results indicate, the loss of detail leads to a reduced volume of Pen a 1 and receptors, thus exposing possible binding sites. At high resolution, the volume of receptors is larger and the extra detail can reduce binding site availability if a number of sites are already occupied. This indicates that the binding rate constants for rules associated with the formation of large aggregates should be turned *on* (for more binding events), and for formation of small aggregates some of these rate constants should be *off*, thus allowing fewer binding events.

The four binding rate constants are assigned to the rules as follows: $k_{f1}$ is assigned to rules that specify that none of the neighboring binding regions are occupied, $k_{f2}$ is assigned to rules that specify that one nearest-neighbor region is occupied, $k_{f3}$ is assigned to rules that specify that one next-nearest neighbor is occupied, and $k_{f4}$ is assigned to rules that specify that two nearest neighbors are occupied. As the tail region in Pen a 1 was split into two independent regions for our model, (regions $E$ and $F$ in the rule set, as seen in Figure 3.6) we treat these regions as a special case and assign the rate $k_{f4}$ to the E and F binding rules that specify that nearest-neighbor E or F is occupied.

We demonstrate how the rules were selected using the rules associated with binding region A in Figure 3.6 as an example. Region B is a nearest neighbor to region A, so we specify region B as a region that affects the binding rate of region A in the rules for A binding for both strands. On strand I, the next nearest neighbor to region A, region C, is located in a region of positive curvature (see Figure 3.6(c) for an illustration of positive curvature). Therefore, in the A binding rules for strand I, we do not specify region C as a neighbor that affects the binding rate of region A. On strand II, region C is located in a region of negative curvature (see Figure 3.6(b) for an illustration of negative curvature) along with region A. Therefore, in the A binding rules for strand II, we specify region C as a neighbor that affects the binding rate of region A.

We use the *General* rule set (Appendix Tables A1 and A2) to illustrate how the influence of neighbor occupancy hierarchy on binding site probability favors particular aggregate sizes. Our results show that if all rules are *on* with identical binding rates $k = 1.0$ molecule$^{-1}$s$^{-1}$, the distribution of aggregate size is skewed to larger aggregates. However, as we turn rules *off* by setting their associated binding rates to zero in hierarchic order (larger to smaller aggregates), we see the progression shown in Figure 3.7, until we obtain a single peak at aggregate size 4 if all nearest and next-nearest neighbors to site $i$ need to be empty for a binding event to occur. Note that binding affinities are not known for the binding sites of Pen a 1, so we use values known for DF3. Since our system is finite in size, the association rate unit of molecule$^{-1}$s$^{-1}$ is used (further discussed in Section 3.4).

The analysis shown in Figure 3.7 shows how turning on and off different binding rates affects the distribution of aggregate sizes. The results indicate that it is possible to fine tune a particular aggregate distribution by choosing the rule set wisely (based on geometric input) and by setting binding rate constants appropriately. Finding a proper rule set is one of the main difficulties of this method, but PDB structures and feedback from Monte Carlo 3-D rigid body simulations can give crucial input on this step. Binding rate constants can be varied as well. Our Monte Carlo simulations assume a constant binding rate of $k = 1.0$ molecule$^{-1}$s$^{-1}$ for all sites. However, to mimic loss of accessible volume to a particular binding site on Pen a 1, the rates can be varied to improve fits to data. These rates are considered free parameters of the simulation and they represent physical binding rates qualitatively.

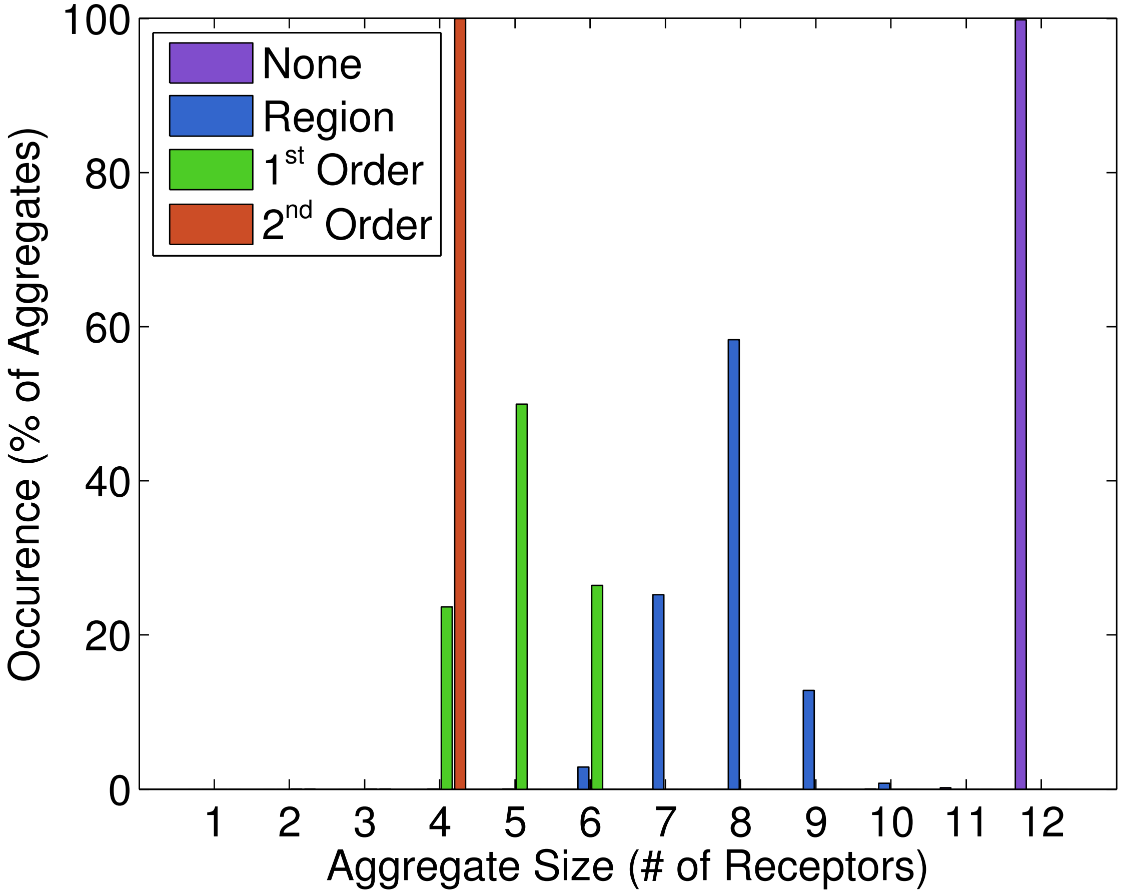


Figure 3.7: For the same *General* sets of rules, we selectively turn hierarchic rates *on* and *off*. *On* binding rates have their value fixed at 1.0 molecule$^{-1}$s$^{-1}$ and *off* binding rates are set to zero. As we make the rules more restricted, smaller aggregates are formed. The purple peak at 12 is labeled "None" for having no restrictions on binding due to neighbors ($k_{f1}$, $k_{f2}$, $k_{f3}$, and $k_{f4}$ are *on*). This is why the largest possible aggregates are formed almost 100% of the time. The data labeled "Region" (blue) allows for binding to a site even if nearest and next-nearest neighbor sites are bound ($k_{f1}$, $k_{f2}$, and $k_{f3}$ are *on*; $k_{f4}$ is *off*). First order interactions (nearest-neighbors, green) are not allowed for this data set ($k_{f1}$ and $k_{f3}$ are *on*; $k_{f2}$ and $k_{f4}$ are *off*). The peak in red (second order) does not allow binding to sites if both their nearest and next-nearest neighbors are occupied ($k_{f1}$ is *on*; $k_{f2}$, $k_{f3}$, and $k_{f4}$ are *off*).

**Algorithm 3.1** IgE Aggregation Simulation Algorithm.

**Input:** Receptors $R$, antigens $A$ and graph $G$.

**Output:** A set $S$ of the resulting aggregates

```
Initialize($R$,$A$,$G$)
for $timestep$ = 0:MAX_TIME do
    for each molecule $m \in R \cup A$ do
        $m$.DetermineMotion($G$)
        moleculeList $old$ = $m$.KnownBoundSites()
        moleculeList $new$ = $m$.PotentialBindingSites()
        for each $t \in old$ do
            $S$.TryRemoveLink($G$,$m$,$t$,D_RATE)
        end for
        for each $t \in new$ do
            $S$.TryAddLink($G$,$m$,$t$,A_RATE)
        end for
    end for
    if $G$.StabilityReached() then
        break()
    else
        $G$.StoreConnectionCount()
    end if
end for
set $S$ = $G$.DetermineAggregates()
return $S$
```

### Model Comparison

In order to quantify the difference between the Monte Carlo and rule-based modeling aggregate sizes for each resolution, the residual sum-of-squares (RSS) normalized by the number of possible aggregate sizes (13) was calculated for each resolution. The equation used to calculate the normalized RSS is:

$$RSS = \frac{\sum_{i=1}^{N}(P_{MC}^{i} - P_{RBM}^{i})^2}{N}$$

where $N$ is the total number of possible aggregate sizes in a histogram (each histogram has the same number of possible aggregate sizes), $P_{MC}^{i}$ is the occurrence probability of the $i$th aggregate size of the Monte Carlo data, and $P_{RBM}^{i}$ is the occurrence probability of the $i$th aggregate size of the rule-based modeling data.

Since the data points used in this calculation are probabilities, the maximum possible normalized RSS is one, and the minimum possible normalized RSS (corresponding to two identical histograms) is zero.

## 3.4 IgE Antibody Aggregation Results

For our rigid body modeling of aggregation, we present three sets of results. In the first set of results we focus our preliminary analysis of the method [84] (Section 3.4.1). We then move on to an analysis of the impact of valency on our simulations [81] (Section 3.4.2). We finish our analysis of the method by quantifying the impact of resolution on our simulation [82, 80] (Section 3.4.3).

### 3.4.1 Method Evaluation

We begin with an analysis of the base methods used to simulate aggregation of antibodies published in [84]. We used the antigen DF3 for this experiment. We use 90 receptors in our experiment and specified antigen counts at 30, 45, 60, 90, 135, and 180. These molecular counts were chosen to match experimental conditions which keep receptor concentration consistent and vary antigen concentration [9]. The bounding area used was 400 nm x 400 nm (160,000 nm$^2$) and is fixed over the course of all experiments. Simulating 90 receptors on a patch this size results in a density of $\sim$600 receptors/$\mu$m$^2$. In all simulations we apply reflecting boundary conditions, ensuring that the number of molecules is kept constant. Thus, molecules are not permitted to exit the area representing the membrane patch and reflect off boundaries when reached. A fixed time interval of 13.2 ms was used for all simulations and each simulation was run for a total of 36,000 time steps, with a total time sufficient to reach steady state. Association and dissociation rates of 1.0 molecule$^{-1}$s$^{-1}$ and 0.025 s$^{-1}$, respectively, were used. Since our systems are finite in size, the association rate unit molecule$^{-1}$s$^{-1}$ are calculated from the original units of M$^{-1}$s$^{-1}$, M being molar concentration, following calculations from [89].

The base speed, $s$, for molecules is set to be 0.09$\mu$m$^2$/s from [9]. Recent experimental evidence suggests that unbound molecules move at faster rates than bound molecules [9]. Since this reduction in speed has not been fully quantified, we use a theoretical basis for modeling the slow down in our system [34, 61]. While there are many slow down models that could be employed, we decided to use a model where receptor count is used as a surrogate for size measurement to determine relative slow down. The geometric analysis of this work emphasizes packing structure rather than aggregation kinetics, so the choice in slow down method should not impact aggregate structure packing at steady state. The slow down is incorporated into the simulation by diffusing aggregates inversely proportional to their size, i.e., the diffusion coefficient of an aggregate linking 3 receptors is 1/3 of the original coefficient. We note that our slow down scheme does not account for physical barriers that exist in the cell membrane that may induce slowing/immobilization (like those in [91]).

Aggregates become more stable as they get larger [9]. The rotation of an aggregate is dependent on multiple receptor contact points with the cell membrane. The more contact points and aggregate has, the more restricted the aggregate dynamics become. We model rotation of an aggregate by choosing a direction and angle value based on receptor contacts points. For any given aggregate rotation, the angle is constrained by receptor displacement, i.e., how much the receptors move given a rotation operation. The receptor furthest from the center of the aggregate would be the most displaced by a rotation operation. Thus, the aggregate rotation is limited by the diffusion constant of the receptor furthest from the aggregate center.

Simulations were created using PMPL, a motion planning library developed at Texas A&M University and graph analysis was performed using elements of Boost Graph Library [110]. Experiments were run in a Linux environment on a single processor of an Intel i7 quad-core with 8G of RAM. Multiple (10) runs were done for each experiment.

### Equilibrium of Aggregate Formation

Aggregates should be the most complex after the simulation is allowed to run to a steady state. To evaluate this, we quantified the stability of the graph $G$ in terms of the number of edges. This is due to the fact that the addition and removal of edges indicates a change in aggregate structure. As the number of edges in the graph starts to level out, we can conclude we have reached a steady state.

The average number of edges in $G$ for different ratios of antigen DF3 is shown in Figure 3.8. In all of the simulations, the number of edges quickly grows no matter the receptor to antigen ratio. Starting near 9,000 time steps, all the curves begin to level off with much smaller growth in the number of edges, indicating that at this point the aggregates were mostly formed and slowly reaching towards a steady state. This result is consistent with observations in changes of Fc$\epsilon$RI mobility [9], which are associated with changes in aggregate size. Abrupt slowing of IgE-receptor aggregates can be observed with the addition of polyvalent antigen and are typically complete within 60-90 secs [9].

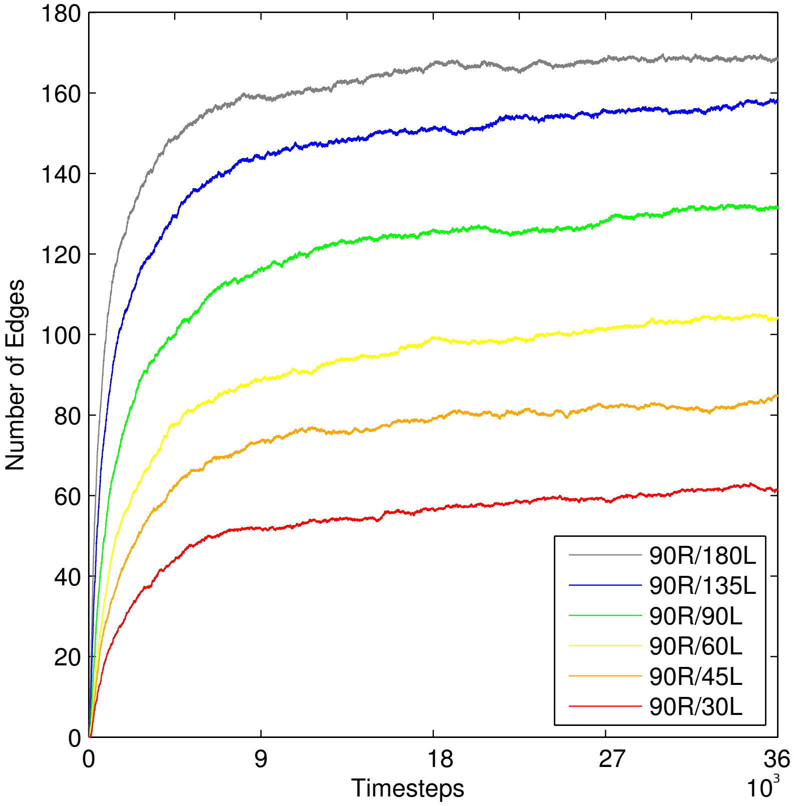


Figure 3.8: The number of edges in $G$ over time used to estimate simulation stable state. The $x$-axis is simulation time and the $y$-axis is the number of connections. Different receptor to antigen ratios are shown, where R and L represent receptor and antigen, respectively.

### Aggregate Size

Electron microscopy has shown that large Fc$\epsilon$RI "signaling patches" form within 1-2 minutes of addition of polyvalent antigen [9]. One limitation of this technique is that, while these patches may contain tens to hundreds of IgE-Fc$\epsilon$RI and antigen, it is not possible to estimate the range of aggregate sizes within these signaling patches. This is due to the fact the resolution of the data cannot provide any connectivity information of the aggregates, thus it is near impossible to distinguish between actually bound and simply proximal receptors. Since we have connectivity information about our graphs, we can report aggregate sizes.

Results from simulations of antigen DF3 are shown in Figure 3.9, where aggregate size was measured for every subgraph in G of three vertices or more. The number of vertices in a subgraph distinguish the size of an individual aggregate. Experimental studies have focused on reporting temporal information about receptors, so we measure aggregate size as the subset of vertices labeled as receptor, i.e., aggregate size is the number of receptors. After the simulations were run, aggregate sizes were collected and averaged.

In Figure 3.9 there are clear aggregate size differences depending on the receptor to antigen ratio. With a high ratio of receptor to antigen (red bar in Figure 3.9), there are fewer antigens so receptors can not easily find unbound antigen. Since there is a low number of antigens, aggregates do not get very large. Looking at the highest ratio experiment, aggregates had only up to seven receptors. With a low ratio (purple bar in Figure 3.9), a saturation of antigen and lack of receptors is observed, resulting in small aggregate structures. This is because the binding sites of the receptors are quickly filled with unbound antigen keeping aggregate sizes small, resulting in aggregates of at most size six.

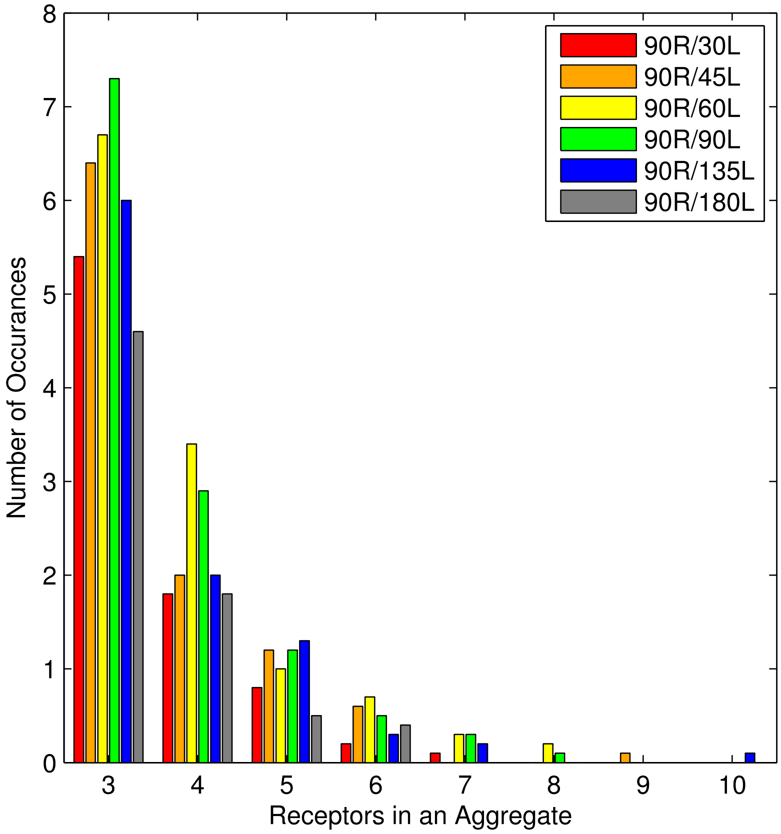
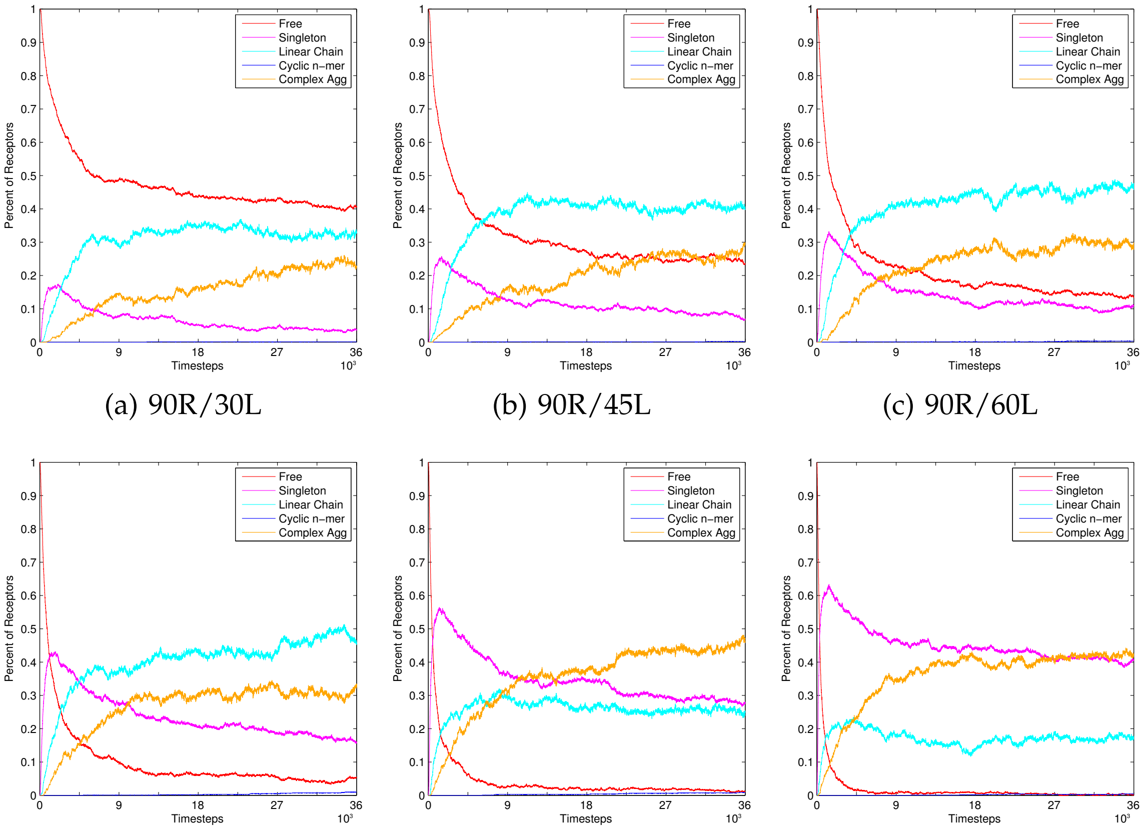


Figure 3.9: Histograms of aggregate size (as defined by number of receptors). Aggregate sizes were collected at the end of runs and averaged.

The ratios which lie in the middle (orange, yellow, green and blue bars in Figure 3.9) show different characteristics than at the ratios at the extremes. As the ratio of receptor to antigen increases, the largest aggregate size and number of aggregates initially increase, but eventually peak and decline. We see this non-monotonic dependence clearly in Figure 3.9, finding higher counts of small aggregate structures and larger aggregate sizes for ratios near the middle compared to those at the extremes.

### Aggregate Populations

Population kinetics describe the population of a class of objects over time. We classify aggregate structures using the four classes defined in Figure 3.4, singleton, cyclic *n*-mer, linear chain and complex aggregate. Given these four classes and an additional unbound receptor class, referred to as free, we can measure how the population of each class changes over the course of the simulation. For each ratio experiment, the class of receptors are plotted against simulation time in Figure 3.10. Values are averaged over all runs.



(a) 90R/30L

(b) 90R/45L

(c) 90R/60L

(d) 90R/90L

(e) 90R/135L

(f) 90R/180L

Figure 3.10: Population kinetics of the simulations of different receptor (R) to antigen (L) ratios for antigen DF3.

We notice that all six plots show more gradual change in class after 9,000 time steps. This relates to Figure 3.8 where the number of edges begins to converge. Therefore, Figure 3.10 confirms that the majority of aggregates are fully formed when steady state in edge count is reached.

One notable trait in Figure 3.10 is that the percentage of unbound (free) receptors decreases quickly in all six ratios. When there is a high receptor to antigen ratio as in Figure 3.10(a), there are always some remaining free receptors. As this ratio decreases, the number of free receptors go down since there are more antigens to bind to the free receptors. We also see that right after the antigen start interacting with the antibodies (near time step 1,000), the percentage of free receptors in the highest receptor to antigen ratio (Figure 3.10(a)) is 70%, where in the lowest ratio it is 10% Figure 3.10(f). This matches experimental results showing antigen concentration is related to immobility (and aggregate formation) [9].

Recall that singleton refers to a single receptor bound to either one or two antigens. In the run where there is a low receptor to antigen ratio (Figure 3.10(a)), we see a low percentage of singletons (magenta line), but we see this percentage increase as the receptor to antigen ratio decreases. This is intuitive since unbound receptors should have an easier time finding one or two unbound antigens in these low ratio (high antigen count) cases.

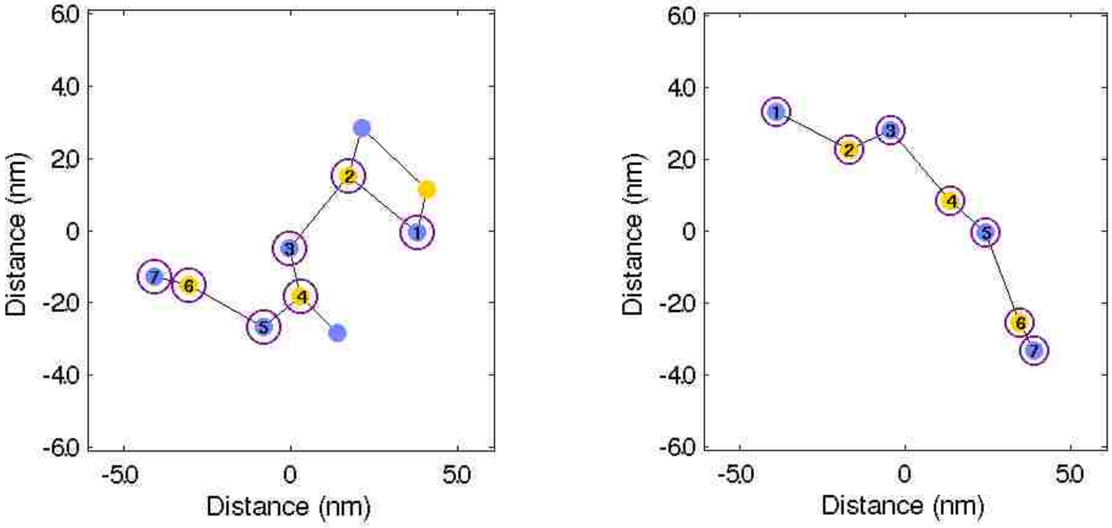
Another results that Figure 3.10 indicates is that cyclic *n*-mers are not very common in any case due to the rigid-body nature of our simulation, this type of structure is considered very constrained. It requires that antigens and receptors to bind at angles that are optimized in order to form cycles. Another requirement is that nothing else can bind beyond those receptors needed for the cycle. Meeting these requirements simultaneously is unlikely given the accessibility of the three binding sites of DF3. Recall that if other structures extend off the Cyclic n-mer, it will be classified as a Complex Aggregate.

### Aggregate Substructure Analysis

To evaluate the the aggregates generated, we need to be able to compare their structures. Common substructures seen in the graph topology of compared aggregates can provide insights into the aggregate structure composition and conditions necessary to generate large aggregates. In this study, the two largest aggregates from each of the ten runs of the six ratios (120 total) were collected for analysis. McGregor's common subgraph was performed between all pairs of selected aggregates to produce a set of common substructures. To describe these common structures, classification was run using the categories described before in Section 3.3.1.

Two resulting aggregates and their common subgraph is shown in Figure 3.11. Even though the two original aggregates had two different classes (Complex Aggregate and Linear Chain for a and b, respectively), their common subgraph was a Linear Chain of three receptors and three antigens. The vertices that are part of the subgraph are circled with their corresponding vertices labeled numerically for easier comparison. The vertices are in positions that relate to their center of mass of the molecules. However, the vertices are not scaled for the size of the molecule.

Table 3.1 outlines the classifications of the subgraphs identified from pairwise comparison. In these results, there appears to be a relationship between common subgraph classification and the ratio of receptors to antigens. With few antigens (three higher receptor to antigen ratios), there was an even ratio of Linear Chains and Complex Aggregates found as the subgraph isomorphism. However, there is increase in Linear Chains when the saturation of antigens increases. In the three lowest receptor to antigen ratio experiments, we see a much higher ratio of Linear Chains versus Complex Aggregates as the common substructure. We note that in the population kinetics results that at the lowest receptor to antigen ratios, Complex Aggregates overtake Linear Chains in classification. Linear Chain geometry appears to be better suited for generating large aggregates in high ratio experiments.



(a) Complex Aggregate

(b) Linear Chain

Figure 3.11: Subgraph Isomorphism between two DF3 antigen aggregates from the 90R/90L ratio run. The first (a) is a Complex Aggregate and the second (b) is a Linear Chain. Purple rings outline the most common subgraph structure and the numbers label the correspondence between the graphs.

| | Receptor Ratio | | | | | |
|---|---|---|---|---|---|---|
| | High | ⟶ | | ⟶ | ⟶ | Low |
| Classification | 30L | 45L | 60L | 90L | 135L | 180L |
| Linear Chain | 48% | 36% | 45% | 90% | 86% | 89% |
| Cyclic n-mer | 0% | 0% | 0% | 0% | 0% | 0% |
| Complex Agg | 52% | 64% | 55% | 10% | 14% | 11% |

Table 3.1: Classification of the most common subgraphs extracted from pairwise aggregate comparison. Aggregates were generated using antigen DF3. Ninety (90) receptors were used for all runs.

### 3.4.2 Antigen Valency Study

Results on antigen valency from [81] are presented in this section. In this experiment, we wanted to gain insights into how valency impacts aggregate structures. To do this we compare DCT2 and DF3, two DNP based antigens that are similar in size but differ in valency (bivalent vs trivalent, respectively). For each antigen, we simulate a variety of antigen to receptor ratios. Receptor count is kept consistent at 90, and we varied the antigen count, in order to match experimental analysis. We ran counts of 30, 45, 90, and 180 for both antigens. We simulated the molecules on a 400 nm x 400 nm (160,000 nm$^2$) membrane patch. A fixed time interval of 13.2 ms was used for all simulations and 36,000 steps were taken. The association and dissociation rates used were 1.0 molecule$^{-1}$s$^{-1}$ and 0.025 s$^{-1}$, respectively. The speed for all molecules, $s$, is 0.09 $\mu$m$^2$/s and the speed of the aggregates are reduced to $s/|a|$, where $a$ is the number of molecules in the aggregate. Multiple (10) runs of each experiment were performed. Experiments were run on single cores with Intel Xeon E5645 processors and 4GB RAM per processor.

#### Equilibrium of Aggregate Formation

To analyze the stability of the system, we look at the number of edges in the graph $G$. The average number of total edges in the entire simulation of graph $G$ over the course of the experiments are shown in Figure 3.12. As can be seen, the number of edges initially grows quickly in all of the experiments. Near 9,000 timesteps, the rate at which edges are being added to $G$ starts to slow down, as evidenced by a leveling off of the edge count. We infer this as the aggregates becoming stable and fully formed. This result is found to be consistent with observations seen in IgE-Fc$\epsilon$RI where it shown that changes in mobility are associated with aggregation [9]. Also in Figure 3.12, DF3 produces more edges than DCT2 for the same ratio of antigen to receptor. This is expected due to the valency of the antigen. DF3 has 50% more binding sites accessible per molecule.

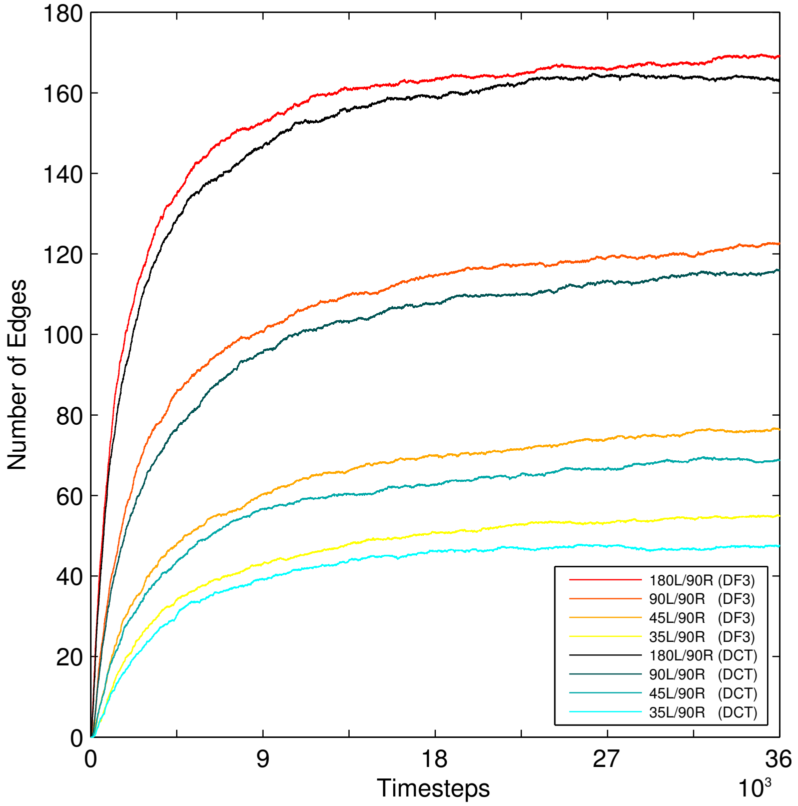


Figure 3.12: The number of connections vs. time step. The system's steady state is indicated by the flattening of the curves. The $x$-axis is simulation timestep and the $y$-axis is the number of edges in $G$.

### Aggregate Size

Aggregate size was measured for every connected component in $G$ produced over the course of the simulations, and the results is shown in Figure 3.13. Like the experiment prior, we measured aggregate size by receptor count. After the simulations were run, aggregate size counts were collected and averaged for each antigen, and aggregate sizes of two or larger were reported.

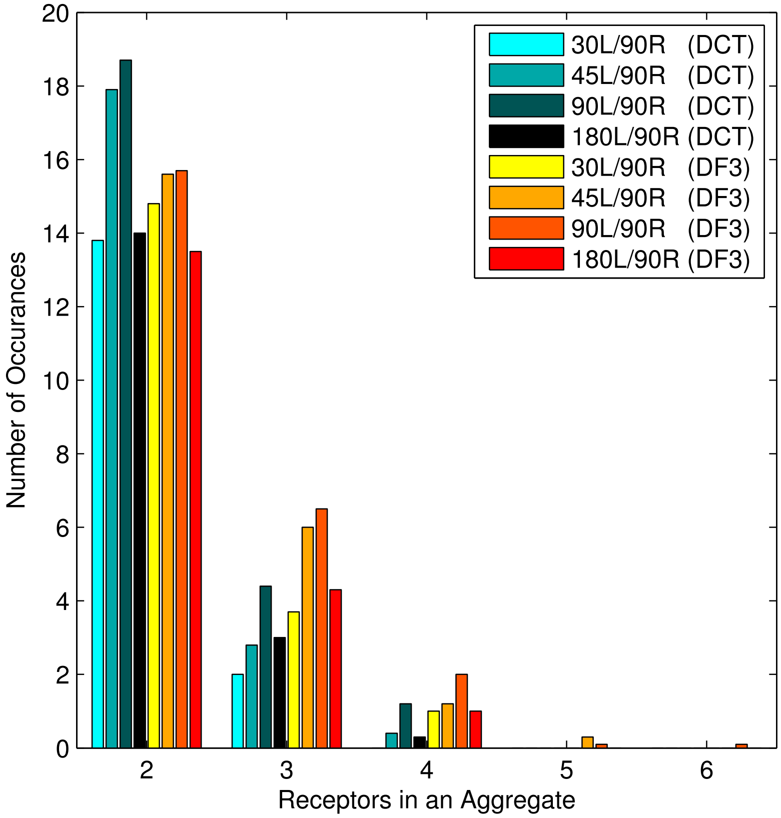


Figure 3.13: Aggregate sizes and number of occurrences. Aggregate sizes (number of receptors) were enumerated at the end of run and averaged. The $x$-axis is the size of aggregates and $y$-axis is the average number of aggregates of that size.

We see two trends in Figure 3.13. First, the different ratios for both antigens, the middle ratios (45 and 90 antigen counts) produce more aggregates of any given size relative to the extreme ratios (30 and 180 antigen counts). Second, these median ratios produce larger aggregates compared to the extreme ratios.

In Figure 3.13, there are clear aggregate size differences depending on the receptor to antigen ratio. In low antigen count simulations, antigens have a hard time binding to already bound receptors since there are so few. Because of this, aggregates tend to stay small. Looking at high antigen count simulations, we see that a saturation of antigens produces smaller aggregates. This is attributed to receptor binding sites quickly filling up with unbound antigens reducing their ability to crosslink. More moderate ratios produce more aggregates and larger aggregates.

Starting from receptor saturated experiments (high receptor to antigen ratio) and increasing the antigen count, aggregate counts of all sizes increase. This trend continues until antigen saturation when we see a decline in the number of aggregates. This trend is seen in the bell shape curve in the number of occurrences of any given aggregate size and molecule counts (Figure 3.13). These results are consistent across both antigens.

We also see in Figure 3.13 that the 2 antigens simulated produce aggregates of different sizes. Bivalent DCT2 produces slightly more small aggregates, but trivalent DF3 overall produces larger aggregates. This can be caused by the valency difference, trivalent DF3 has similarly accessible binding sites but can produce more complex structures (cycles, chains and trees) than bivalent DCT2 (cycles and chains).

### Resulting Aggregates

Antigens with different valencies can produce different aggregate formations. A sample of the resulting aggregates constructed during our simulation is shown in Figure 3.14. We see that trivalent antigen are capable of generating aggregates that cannot be made using bivalent antigen, i.e., trees.

Details of aggregate structures (Figure 3.14) are interesting because aggregate binding patterns are difficult to see using experimental imaging techniques. For example, notice the compactness of Figure 3.14 A. Even though the DCT2 antigen is only able to produce simple structures, the receptor positions are compact, similar to the DF3 aggregate (Figure 3.14 B). These reconstructions enable the extrac-

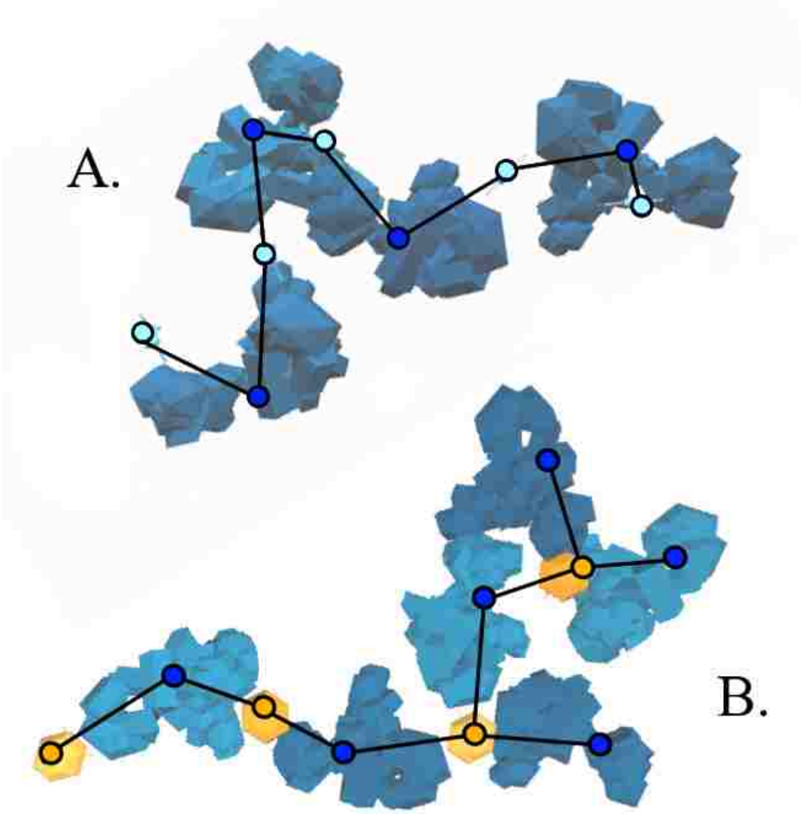


Figure 3.14: Aggregates produced during our simulations. (A.) DCT2 Aggregate (Size 4) (B.) DF3 Aggregate (Size 6). Note aggregate size is dependent on receptor (*blue*) count, antigens DCT2 (*cyan*) and DF3 (*orange*) are disregarded.

tion of more information about the aggregates besides their connectivity graphs. For example, metrics like the distribution of distances between receptors can be measures and compared across aggregates made with different antigens.

### 3.4.3 Model Resolution Study

In this set of results, we are focused on determining the impact that model resolution has on our simulation. These results were published in [82, 80]. For these simulations, we simulate a discrete patch of membrane 200 nm x 200 nm (40,000 nm$^2$). We simulate 24 receptors for all experiments, resulting in a density of $\sim$600 receptors/$\mu$m$^2$. In two different experiments, we simulate twelve DF3 and one Pen a 1 antigen molecules at four distinct resolutions, reducing the models of both antigen and receptor by 0%, 50%, 75% and 90%. Due to the presence of multiple DF3 molecules, it is possible to observe crosslinking in the DF3 simulations. However, no antigen-mediated crosslinking can be observed in the Pen a 1 simulations as only one molecule of Pen a 1 is simulated in each experiment.

We use the diffusion coefficient 0.09 $\mu$m$^2$/s of IgE-Fc$\epsilon$RI found in [9] for all molecules. We use a time step of 10 $\mu$s and run experiments for 500,000 time steps, long enough for the simulations to reach a steady state. Association and dissociation rates of 1.0 molecule$^{-1}$s$^{-1}$ and 0.01 s$^{-1}$, respectively, were used for both antigens. Simulations were run on single cores of Intel Xeon E5645 processors with 4 GB of RAM per processor. Thirty (30) runs of each experiment were performed.

### Volume and Timing

We begin with an analysis of polygon reduction and the impact it has on model volume. Table 3.4.3 shows the number of polygons and volume for each model. The polygon reduction algorithm works by specifying a percentage of the polygons to reduce, leading to the close correspondence between the reduction percentage and the number of polygons. We find that volumes decrease as the number of polygons is reduced. Such decrease is expected, and can be quite dramatic (nearly 50% for 90% reduced Pen a 1). We note that volume reduction does not necessarily mean less realistic results; "*soft docking*" approaches [39] allow a certain degree of inter-protein penetration to approximate flexibility given rigid structures.

| Molecule Name | Model Property | Model Percent Reduction | | | |
|---|---|---|---|---|---|
| | | 0% | 50% | 75% | 90% |
| Receptor | Polygons | 4876 | 2438 | 1216 | 490 |
| | Volume ($nm^3$) | 234.98 | 227.90 | 208.73 | 162.31 |
| | Volume (%) | 100.00 | 96.99 | 88.83 | 69.07 |
| DF3 | Polygons | 1208 | 604 | 302 | 120 |
| | Volume ($nm^3$) | 15.83 | 14.90 | 13.16 | 9.74 |
| | Volume (%) | 100.00 | 94.13 | 83.13 | 61.53 |
| Pen a 1 | Polygons | 2328 | 1164 | 582 | 234 |
| | Volume ($nm^3$) | 51.60 | 49.95 | 44.86 | 28.80 |
| | Volume (%) | 100.00 | 96.80 | 86.94 | 55.81 |

Table 3.2: Model reduction statistics including polygon counts and volumes of the molecular models generated at a variety of resolutions.

As seen in Figure 3.15, the reduction in polygons has a clear effect on runtime. We see a linear increase in runtime versus model polygon count. This is due in part to the nature of the rigid body modeling; collision detection is a major factor in computation time and is highly dependent on model complexity [63]. We attribute the calculation of binding site interactions, whose costs depend on valency and molecule size, to the difference in slope between DF3 and Pen a 1 runtimes.

#### Impact of Resolution on Quality of Results

Simulations are run until a steady state is reached. To ensure the system is stable, we count the number of bonds between molecules, i.e., the number of edges in $G$. Figure 3.16 (a) shows the number of edges in $G$ for DF3 in blue. We see that for DF3, all of the reductions generate similar numbers of connections. The more reduced models produce slightly more connections but all of the average lines are very close. The mean of each reduction is contained in the overlap of the standard deviation of all reductions.

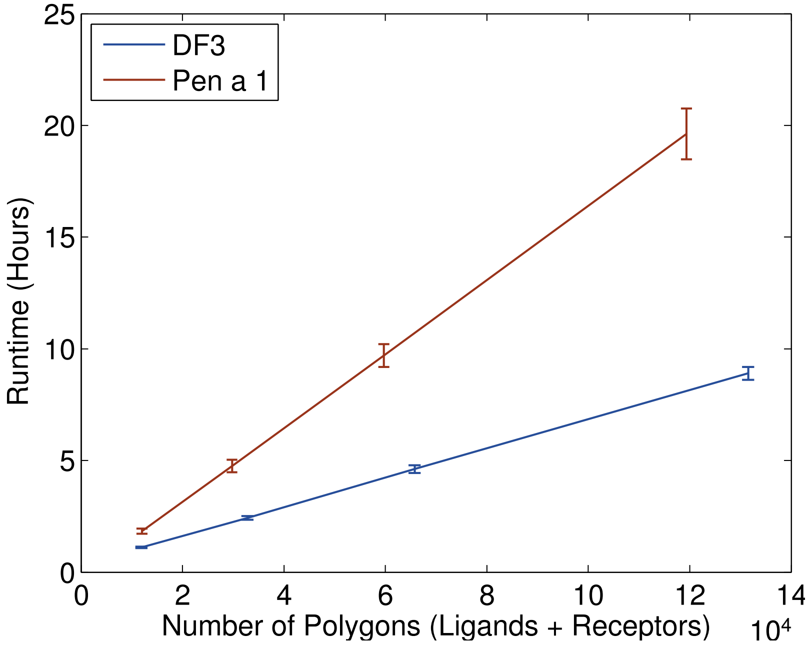
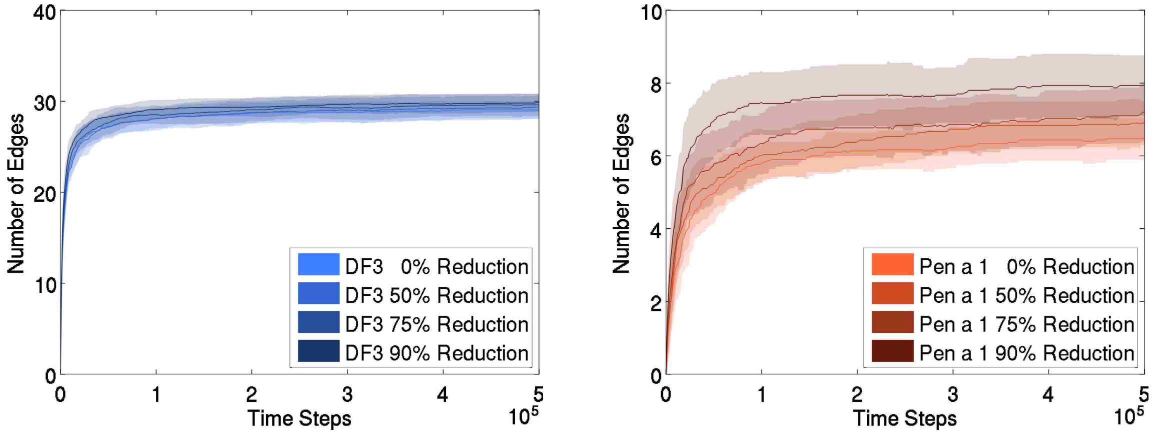


Figure 3.15: We compare the runtime of the different resolutions of the same model. The $x$-axis is the number of polygons used to describe the models and the $y$-axis is the time took to run the experiment in hours.

In Figure 3.16 (b) we see that Pen a 1 model resolution has a higher impact on the number of connections that are made. The 90% reduced model made on average nearly two more connections than the 0% reduced model for a single antigen. This is one of the side effects of reducing model volume. With the reduction, there is more open volume around the binding sites, reducing steric hindrance of receptors trying to bind to sites in the same or adjacent regions.

To further analyze the implications of model reduction, we plot histograms of aggregate size versus percentage of occurrence (Figure 3.17). We see in Figure 3.17 (a) that there is minimal impact on aggregate size distributions for the DF3 experiment. The distribution for each model reduction seem to be the same.



(a) DF3

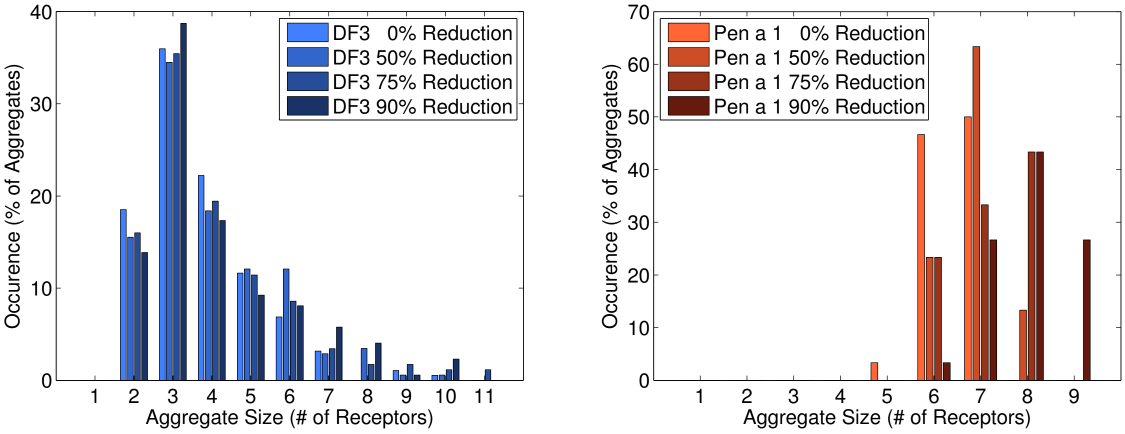
(b) Pen a 1

Figure 3.16: Influence of model resolution on the number of connections made during a given simulation. The $x$-axis is simulation time step and the $y$-axis is the number of edges in $G$. For DF3 (a) connections do not seem to be affected much by model resolution. Pen a 1 binding is affected by the use of different resolutions (b); lower resolutions generate more connections than higher resolutions.

This is not the case for Pen a 1, seen in Figure 3.17 (b). The distribution has the two least reduced models peaking near aggregates of size seven whereas the two most reduced models peak near aggregates of size eight. This is attributed to the volume reduction of the model which in effect relaxes the steric constraints. With a smaller volume, more free space is available for a molecule to pack into a tight space within a given aggregate.

### Clustering Analysis of DF3

To quantitatively analyze the clustering of the system, we measure the Hopkins statistic of the receptors over the course of the simulations. We focus on the analysis of clustering for DF3 due to the availability of experimental data [78]. Unfortunately, a similar analysis for Pen a 1 does not result in significant data as there is only one Pen a 1 allergen in each simulation run. To evaluate cluster-



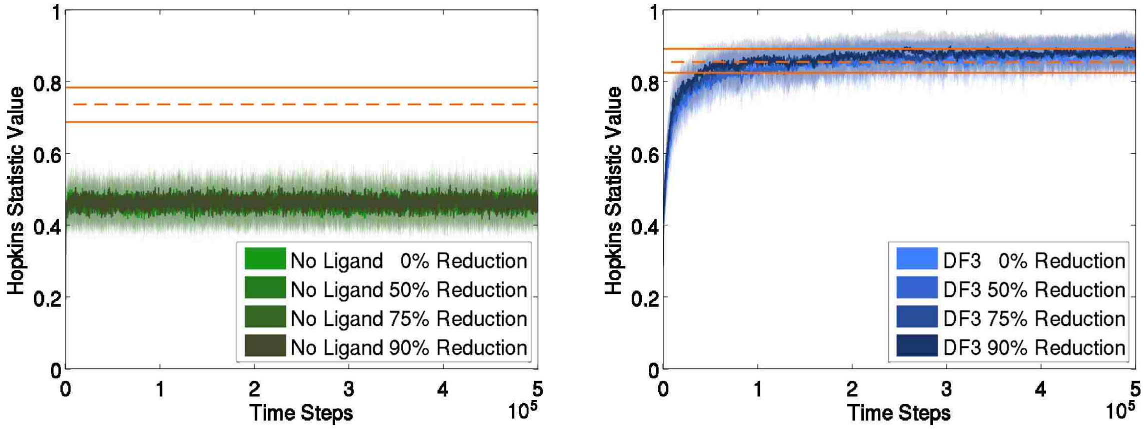
(a) DF3

(b) Pen a 1

Figure 3.17: Influence of model resolution on the size of aggregates generated during a given simulation. Histograms show aggregate size vs. percentage of aggregates of that size. For DF3 (a), the different resolutions do not affect the distribution of aggregate sizes. For Pen a 1 (b), aggregate size seems to be dependent on model resolution, with lower resolution Pen a 1 models producing larger aggregates.

ing, the Hopkins statistic values were calculated over the course of the simulation and the results are an average of the simulations for each experiment as seen in Figure 3.18. These values are then plotted and compared to the values obtained experimentally in [78], as shown in Figure 3.18.

For a baseline, we performed a Hopkins statistic calculation for a simulation with only receptors and no antigen and produced the plot in Figure 3.18 (a). We find that for the *no antigen* simulation, the value does not change and is hovering at around 0.5, indicating that the receptors are essentially randomly distributed. We compare our *no antigen* simulation with the 0 nM experiment in [78] (Figure 3.18 (a), red dashed line). Our value (mean 0.5) differs from what is seen experimentally (mean 0.74); however, this difference can be attributed to the fact that cell membranes present topological inhomogeneities that result in natural receptor organization [71, 96, 8, 73]. These features are not incorporated in our sim-



(a) No Antigen

(b) DF3

Figure 3.18: Influence of model resolution on the clustering of receptors during a given simulation. Hopkins Statistic values (measure of clustering) are plotted over the course of the simulation without antigen (a) and with DF3 (b). All resolutions converge to the same results. Comparison with experiments in [78] are shown as the red lines (mean-dashed, variance-solid).

ulations for simplicity. Therefore, we observe no clustering instead of the slightly clustered distribution observed in experiments.

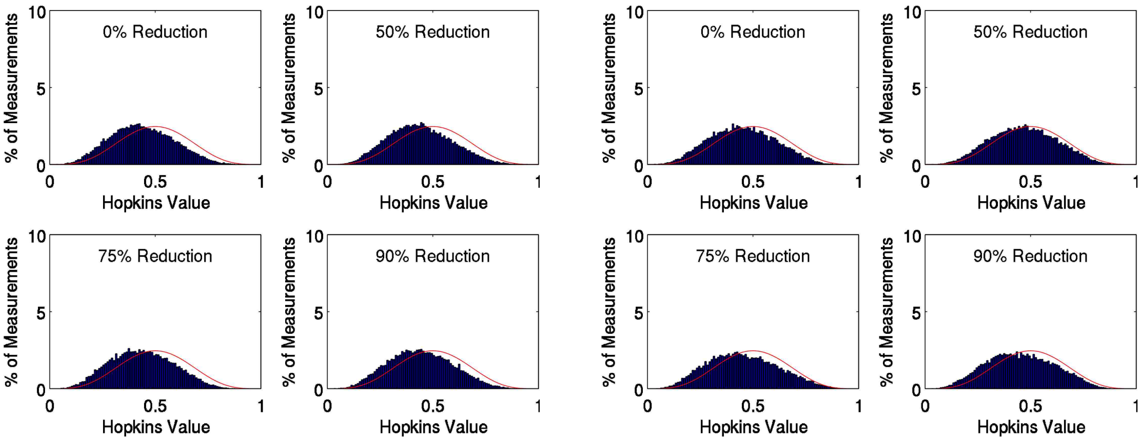
Based on the conclusion of [78], where the authors imply that the antigen-induced aggregate state can influence cell signaling, we assume that final antigen-receptor aggregate size can be associated with cellular degranulation. Thus, to analyze clustering in the presence of antigen, we compare the Hopkins statistic results from the experiment in [78] that resulted in optimal histamine secretion (10 nM of DF3) to the Hopkins statistic obtained from our Monte Carlo simulations. We find that the Hopkins statistic value at equilibrium from experiment has a mean value of 0.85 with error bars between 0.82 and 0.89 (Figure 3.18 (b), red dashed line), while our simulations give a mean of 0.88±0.05. This overlap indicates similar clustering observed in both our Monte Carlo simulation and experimentally derived results. In addition, we observe that model resolution does not impact the amount of clustering that occurs for DF3, as all of the values converge to the same result, as seen in Figure 3.18.

To verify whether the averaged values are representative of the underlying clustering, we plot the histograms of the Hopkins statistic values calculated at the beginning and end of each experiment. These histograms are plotted against a normal distribution representing uniformly random distributed data to provide an intuition of the amount of clustering. For each experiment, we performed 30 runs, each contributing 1000 calculations, resulting in a total of 30000 measurements per histogram. The beginning histograms are taken 1000 steps into the simulation to ensure that the molecules move away from their initial grid state, which brings bias into the calculation. We see in Figure 3.19 that in the *no antigen* experiments, there is no change between the histograms at the beginning of the experiments (Figure 3.19 (a)) to the histograms at the end of the experiments (Figure 3.19 (b)), and both histograms are very close to the normal distribution (red line) indicating no clustering.

However, for DF3, we see (Figure 3.20) that there is a significant shift in the histogram from start to end. The beginning of the experiment starts off as a random distribution (Figure 3.20 (a)). By the end of the experiment, we see a shift in the histogram away from the red normal distribution line (Figure 3.20 (b)), indicating clustering in the simulations. We note this shift is consistent for all resolution models of DF3.

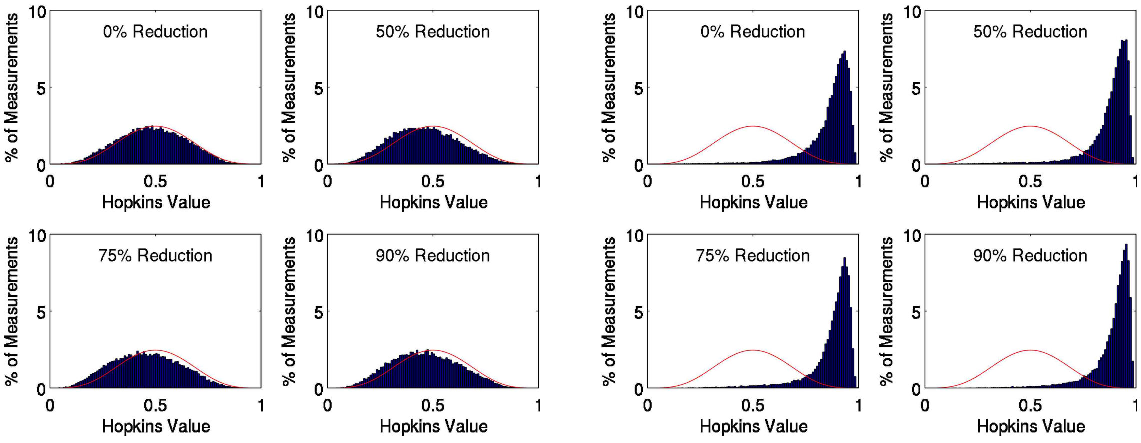
### Analysis of Model Quality

We also investigate the impact of model reduction on all-atom aggregate structures. In order to evaluate this, after aggregates are constructed with low-resolution polygon models, we construct the corresponding all-atom structure.



(a) No Antigen (b) DF3

Figure 3.19: Hopkins statistic for an experiment with *no antigen*. Histograms of Hopkins values for the beginning of the experiment (panel (a)) and end (panel (b)) of the experiment on bottom. Red curve indicates a normal distribution, i.e., no clustering.



(a) No Antigen (b) DF3

Figure 3.20: Hopkins statistic for an experiment with antigen DF3. Histograms of Hopkins values for the beginning of the experiment (panel (a)) and end (panel (b)) of the experiment on bottom. Red curve indicates a normal distribution, i.e., no clustering.

However, since the polygon models are much simpler than the all-atom structures, there may be unintended interactions. For example, when non-bonded atoms are too close, repulsion may occur due to van der Waals interactions. In order to evaluate the possible effects of transitioning from polygon to all-atom models, we counted the number of $C_\alpha$ atoms and DNP linker carbon rings within 7 Å (lower than the shortest interaction radius of 8Å) for IgE-Fc$\epsilon$RI and DF3. For Pen a 1, distances were calculated between $C_\alpha$ atoms for the aggregated molecules. In order to indicate these proximal non-binding residues, we refer to them as *potential collisions*. Also, antigen binding sites are not included in the enumeration.

Table 3.3: Percentage (%) of residues that exhibit a *potential collision*. Antigen residues involved in binding are not included.

| Antigen Simulated | Model Percent Reduction | | | |
|---|---|---|---|---|
| | 0% | 50% | 75% | 90% |
| DF3 | 0.0074% | 0.0122% | 0.0215% | 0.1016% |
| Pen a 1 | 0.0158% | 0.0350% | 0.0816% | 0.2238% |

We can see from the results in Table 3.3 that model resolution has an impact on the number of *potential collisions* that exist in aggregate structures. *Potential collision* residues increase as resolution decreases. We see that DF3 is not significantly impacted by model resolution up to 75%. However, at 90%, there is an order of magnitude increase in *potential collisions*. We see that Pen a 1 model reduction generally has a higher percentage of residues in *potential collision* compared to DF3. This is attributed to the flexibility of the DF3 binding site. The DNP linker has a large, relatively open volume that can be bound, while the binding sites of Pen a 1 are smaller in volume since they are on the molecular surface and are partially occupied by the molecular volume. Therefore, the antibodies have to be closer to the allergen surface in Pen a 1.

We note that overall, the number of residues in *potential collision* is minimal. Even at 90% reduction for the Pen a 1, aggregates generated have about 0.2% of residues in *potential collision*. These interactions could be addressed through locally evaluated energetics and perturbations.

### Rule-Based Modeling Results

So far, we have presented the results of our Monte Carlo simulations for different resolutions, which explicitly include geometric effects in aggregate formation. Rule-based models of aggregate formation, on the other hand, need to encode all geometric information in the rules of antibody-antigen binding and their binding/unbinding rate constants.

Our approach is to vary the binding rates to reproduce the aggregate size distribution at a particular resolution for the *General* rule set. One way to achieve this is by doing a multi-parameter optimization of the four binding rate constants. This is especially useful when fitting the model to experimental data. In Figure 3.21, we compare the aggregate size distribution for the rule-based model with the values found from the Monte Carlo simulation.

Most of the results shown in Figure 3.21 were obtained by analyzing the rates via single and two-parameter scans. For the 0% and 50% resolutions, we fixed $k_{f1}$ to unity in all runs and performed scans of $k_{f3}$ from 0.0 to 1.0 molecule$^{-1}$s$^{-1}$ and $k_{f2}$ from 0.0 molecule$^{-1}$s$^{-1}$ to $k_{f3}$, both at 0.01 intervals. For the 90% resolution, we fixed $k_{f1} = k_{f2} = k_{f3} = 1.0$ molecule$^{-1}$s$^{-1}$ and varied $k_{f4}$ from 0.0 to 0.05 molecule$^{-1}$s$^{-1}$ by 0.001 increments. In all cases, the value chosen for the variable parameter was the one that resulted in the smallest RSS from the Monte Carlo data. However, it is important to highlight that even though the RSS from the Monte Carlo data is a reasonable measure of comparison between the two meth-

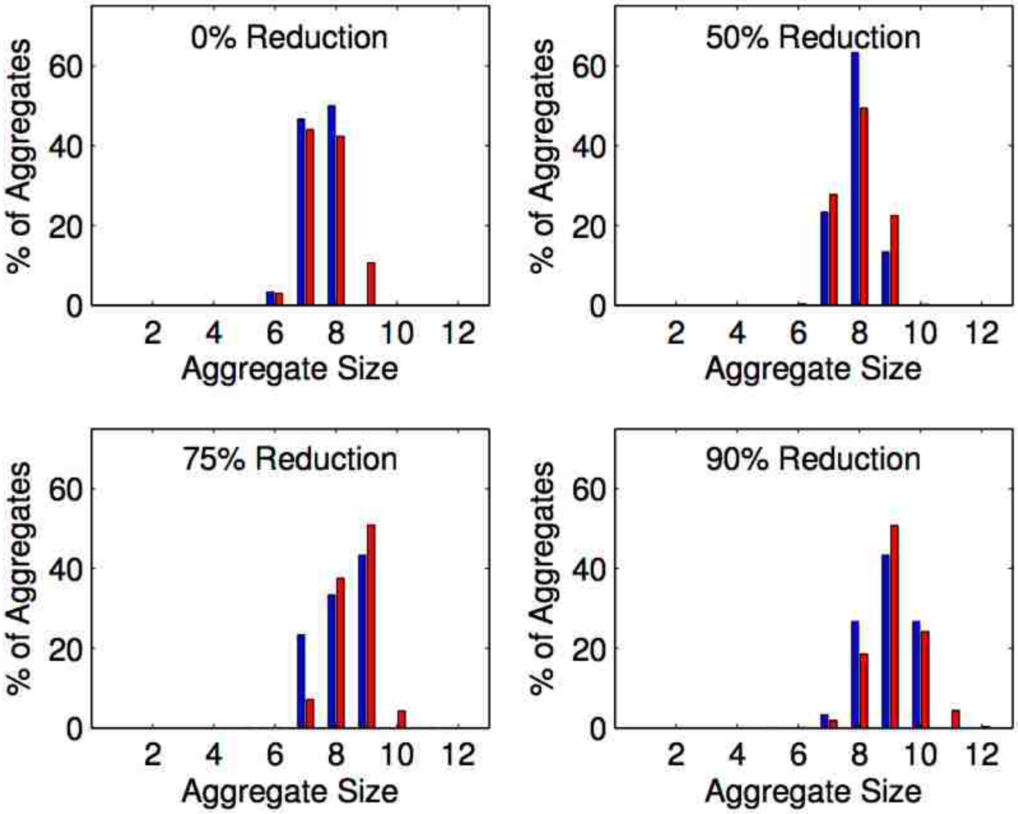


Figure 3.21: Comparison of Monte Carlo (Blue) and rule-based model (Red) results for the *General* rule set with variable rates for different resolutions. Rates and RSS values are shown in Table 3.4.

ods, our Monte Carlo data was obtained from 30 independent runs, and additional runs could change the overall distribution of the histograms, thus changing the RSS significantly. For this reason, we use this number mostly as a guide and avoid fitting rates to exactly reproduce the Monte Carlo data.

The rates listed in Table 3.4 for the 75% resolution were fitted from the trends of $k_{f1} = k_{f3} = 1.0$ molecule$^{-1}$s$^{-1}$ and increasing $k_{f2}$ and $k_{f4}$ for lower resolutions. They show a sufficiently close result to the Monte Carlo data (Figure 3.21). The trend of increasing rate values of $k_{f2}$ (binding with one nearest neighbor occupied) and $k_{f4}$ (binding with two nearest neighbors occupied) reinforce the intuition that the volume loss due to resolution reduction impacts the binding of neighboring regions.

At the 75% resolution, a two-parameter scan for $k_{f2}$ and $k_{f3}$ yields a slightly better RSS for $k_{f2} = k_{f3} = 0.12$ molecule$^{-1}$s$^{-1}$ at fixed $k_{f1} = 1.0$ molecule$^{-1}$s$^{-1}$ and $k_{f4} = 0$, which deviates from the parameter trends. This particular result also seems misleading because at higher resolutions one would expect that $k_{f1} \leq k_{f3}$, due to restrictions on neighbor interactions. However, it is expected that multi-parameter scans may lead to numerous minima of RSS, thus the best-fit solution may not be unique.

Table 3.4: Binding and unbinding rate constants and RSS differences for the rule-based model to capture the aggregate size distributions of different resolutions.

| Rate | Model Percent Reduction | | | |
|---|---|---|---|---|
| Value | 0% | 50% | 75% | 90% |
| $k_{f1}$ (molecule$^{-1}$s$^{-1}$) | 1.00 | 1.00 | 1.00 | 1.00 |
| $k_{f2}$ (molecule$^{-1}$s$^{-1}$) | 0.07 | 0.12 | 0.50 | 1.00 |
| $k_{f3}$ (molecule$^{-1}$s$^{-1}$) | 1.00 | 1.00 | 1.00 | 1.00 |
| $k_{f4}$ (molecule$^{-1}$s$^{-1}$) | 0.00 | 0.00 | 0.00 | 0.006 |
| $k_r$ (s$^{-1}$) | 0.01 | 0.01 | 0.01 | 0.01 |
| RSS | 0.001367 | 0.002283 | 0.002731 | 0.001135 |

# Chapter 4

# Flexible Molecular Modeling

The second body of work we present focuses on our efforts modeling molecular aggregate structures described in [83, 79]. In this chapter we utilize semi-flexible, reduced-resolution models that describe molecular density. This is different from our previous model which is fully rigid and only represents the occupied volume of the structure. Here we discuss our methods for semi-flexible, reduced-resolution modeling of molecules and how we ascertain molecular conformations from Cryo EM tilt series. We outline our methods for model construction in Section 4.1. We then detail our methods for projection construction and tilt series fitting in Sections 4.2 and 4.3, respectively. We then present results from fitting simulated and experimental data in Section 4.4.

## 4.1 Flexible Molecular Model Construction

Generation of a semi-flexible reduced-resolution model of a molecular system begins with its all-atom molecular structure. This structure is first decomposed into rigid subunits, which are areas of the molecular structure where atoms remain static relative to each other throughout conformation changes. Then, Principal Component Analysis (PCA) is applied to the subunits of this decomposed model to simplify the description of the subunits and generate a GMM. Once this is done, flexibility between rigid regions is added, and the subunits are refined to complete the model construction. An example of this process applied to the IgE-Fc$\epsilon$RI complex (described in Section 2.4) is outlined in Figure 4.1. The resulting model is then used to fit tomographic tilt series using genetic algorithm (GA) optimization.

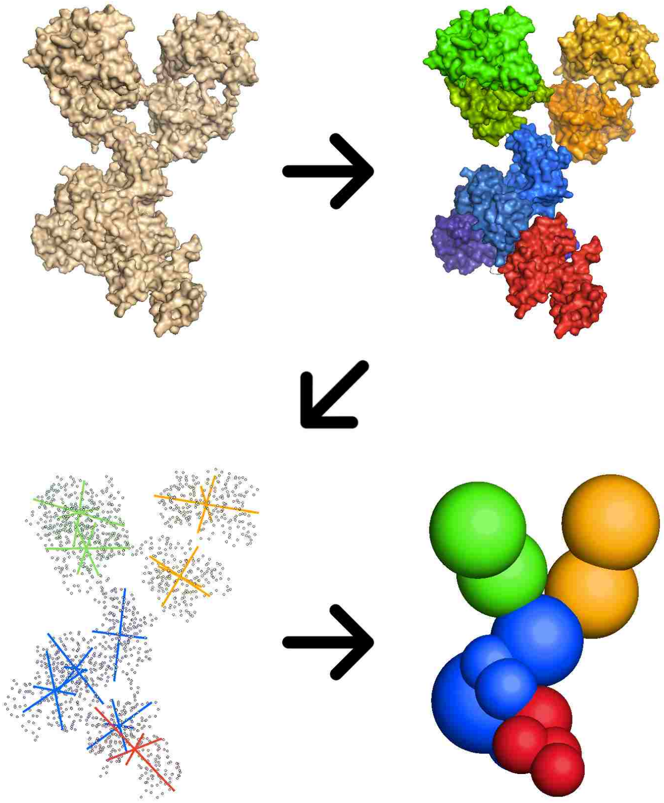


Figure 4.1: Process of generating a GMM from an all-atom structure. First, the all-atom model (*top left*) is decomposed into rigid subunits (*top right*). The rigid subunits are then processed using PCA (*bottom left*). The results of PCA are used to construct a GMM (*bottom right*).

Flexible model construction begins with identification of the rigid subunits of the all-atom model through KINARI-Web rigidity analysis [31]. This web server uses inter-atomic connectivity and interaction information to predict which groups of atoms are likely to move together in a coordinated fashion. If a contiguous portion of the protein is calculated to be flexible, the sequence of residues between the rigid subunit termini are identified as a flexible region. If both termini are associated with the same subunit, the sequence is considered part of that subunit and reclassified as rigid. If the ends are from different subunits, i.e., a flexible linker, the $\alpha$-carbons of the amino acids on both termini of the flexible regions are stored with their associated GMM subunit. For the IgE-Fc$\epsilon$RI structure, ten rigid subunits were classified into five regions outlined in Table 4.1.

| Structural Subunit | Chain & Residues | Region |
|---|---|---|
| Fc$\epsilon$RI $\alpha$-subunit 1 | A 5-84 | 1 |
| Fc$\epsilon$RI $\alpha$-subunit 2 | A 88-169 | 1 |
| Constant 1 | H 256-355, I 256-355 | 1 |
| Constant 2a | H 364-462 | 1 |
| Constant 2b | I 364-462 | 1 |
| Constant 3 | H 469-571, I 469-571 | 1 |
| Fab 1 - C terminal | H 151-247, L 140-234 | 2 |
| Fab 1 - N terminal | H 21-144, L 26-127 | 3 |
| Fab 2 - C terminal | I 151-247, M 140-234 | 4 |
| Fab 2 - N terminal | I 21-144, M 26-127 | 5 |

Table 4.1: IgE-Fc$\epsilon$RI subunits calculated from rigidity analysis. A flexible region is considered part of a rigid region if both ends of the region are associated with the same structural subunit.

After rigid subunits have been identified, the decomposed subunits are converted into a GMM representation by performing PCA on the atom positions of each identified subunit structure. PCA analyzes the positions of the atoms in a subunit and returns eigenvectors, or principal components, that describe the spread of the atoms in 3D space. The eigenvalues associated with the components represent the amount of variance on a particular axis (Figure 4.1, *bottom left*). A rigid subunit is divided into smaller subunits if the ratio between the lowest and highest eigenvalues of the PCA is high, i.e., if the distribution is more elliptical than spherical, the subunit is divided. The model used a ratio of 1.8× or greater as a cutoff for this division. From this criteria subunits 1 and 2 of Fc$\epsilon$RI $\alpha$-subunit and constant subunits 2a and 2b of the IgE complex were split along their principle component axis.

Next, the molecular subunits are converted to a GMM description with Gaussian functions centered at the subunit's atoms mean position $\mu$ with standard deviation $\sigma$. The standard deviation is set to the square root of the largest subunit eigenvalue (as computed by PCA). This construction method is applied to all subunits of the IgE complex, resulting in a model composed of fourteen Gaussian functions (Figure 4.1, *bottom right*).

Distance constraints are incorporated into the GMM to represent flexibility that exists between rigid subunits. Specifically, GMM subunits that are linked together via flexible portions of the protein backbone have a distance constraint. The distance between adjacent $\alpha$-carbon positions of the protein backbone in flexible regions are averaged and stored as a distance, $d$. To determine the length of a distance constraint associated with a flexible region, we multiply the number of amino acids in the flexible region, $n$, times the average distance $d$. During simulation, the distance between constrained rigid subunits' $\alpha$-carbon positions must be less than $n \cdot d$.

There are three major areas of the IgE-Fc$\epsilon$RI structure, the constant domain and the two Fab arms. These three areas are broken down into five rigid regions and six flexible regions. Fab arms are the structures composed of two subunits connected by two flexible regions, producing two distance constraints per arm. Each Fab arm is connected to the constant domain, producing two distance constraints. This results in six distance constraints total, outlined in Table 4.2.

| Class | Subunits | Chain & Residues | Length |
|---|---|---|---:|
| Arm Linker | Con 1, Fab 1C | H 248-255 | 8 |
| Arm Linker | Con 1, Fab 2C | I 248-255 | 8 |
| Fab Short | Fab 1C, Fab 1N | H 145-150 | 6 |
| Fab Long | Fab 1C, Fab 1N | L 128-139 | 12 |
| Fab Short | Fab 2C, Fab 2N | I 145-150 | 6 |
| Fab Long | Fab 2C, Fab 2N | M 128-139 | 12 |

Table 4.2: Flexible regions between rigid antibody subunits. Con 1 represents the Constant subunit and Fab subunits are labeled “Fab $XY$”, $X$ being the arm index (1 or 2) and $Y$ being the termini (N or C). Length is in number of residues in the flexible region.

## 4.2 Projection Generation

The main data structures utilized by our method are 2D single-channel density images from the processed tilt series or generated from projections of the GMM. Those processed tilt series can originate from experimental data or be simulated from molecular models. GMM projections are generated by rendering the set of Gaussian functions from a series of perspectives associated with the tilt series.

Various levels of image- and post-processing are applied to experimentally derived tilt series to prepare the data for fitting. Processing begins with the selection of molecular structures of interest from a preliminary evaluation of the tilt series captured by the electron microscope. These selected structures are used to extract images from the raw tilt series for generation of a 3D reconstruction. The reconstructed model is filtered using non-local means filtering and truncated at upper and lower bounds. The remaining density is discretized into bins using histogram equalization and used to create a set of images from the same perspectives as the originating tilt series. These new images are processed using an iterative means filter and collected together to form the final processed tilt series for fitting.

Simulated tilt series derived from molecular models begin with the creation of a density map of a given atomic model using Chimera [95]. Since our method is focused on low resolution fitting, our density maps typically range from 20Å - 40Å. This density map is equivalent to the reconstruction described when processing experimental tilt series. Non-local means filtering, truncation, and histogram equalization are applied to the density map as previously described, and projections are generated to create a processed tilt series of the molecular model.

The projections of the GMM conformations are generated with a simple rendering technique for Gaussian functions. Since the density space was discretized, $\sigma$ is scaled to render different density values. This technique produces orthographic images of the conformation, rendering circles dependent on the camera perspective, subunit Gaussian with $\mu$ and $\sigma$, and $\sigma$-scale values at which to render densities. Note that the discretization of the density is not necessary for the method to work; rather this step is due in part to the processing method of the lower resolution experimental data.

## 4.3 Tilt Series Fitting

Conformations, or poses, of the GMM are generated and optimized using a GA to identify a best fit to the provided tilt series (Figure 4.2). The process begins with creation of an initial set of GMM conformations, also referred to as the initial population. The individuals of the population have a set of properties, referred to as their genotype, that can be mutated and altered in an evolutionary fashion.

In this study, each element in the genotype represents the amount of translation {x,y,z} or rotation {$\alpha$,$\beta$,$\gamma$} applied to a particular IgE-Fc$\epsilon$RI subunit. In this case, the conformation of IgE-Fc$\epsilon$RI is encoded in a vector of floating point values to facilitate the evolutionary operations. The initial population is generated by randomly perturbing the starting conformation of the GMM.

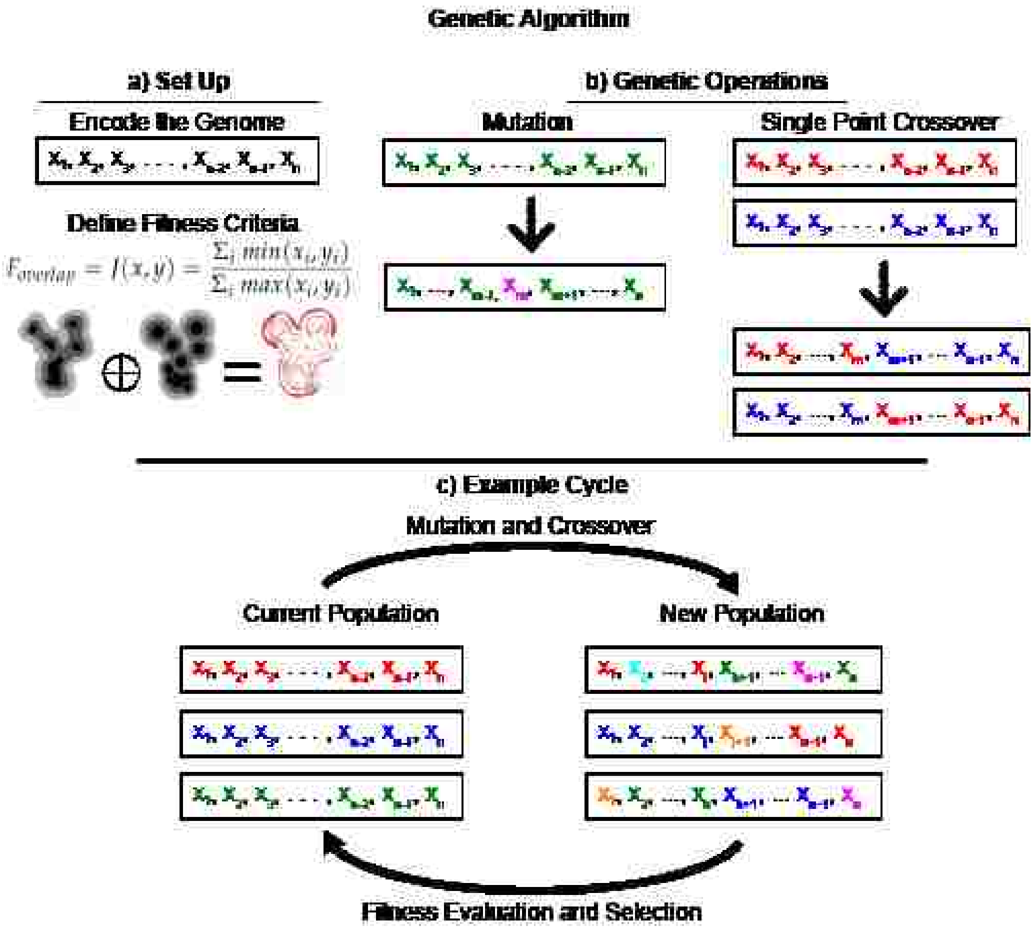


Figure 4.2: Using genetic algorithms to optimize overlap between GMMs and tilt series. a) *Setup:* "genes" are stored in a bit string "genome". Fitness is the overlap measured between the individual and the tilt series. b) *Genetic Operations:* Evolutionary operators include single point crossover (*red* and *blue*) and mutation (*magenta*). c) *Example Iteration:* The current population is subjected to crossover (*red*, *blue*, *green*) and mutation (*magenta*, *cyan*, *orange*) to create a new population. The fittest individual(s) of the new population is passed to the next iteration.

The GA proceeds by selecting individuals from the current population to evolve for the next generation of conformations. The GA here uses tournament selection [87], a method where random subsets of the population are put into ″tournaments″ based on fitness, and the winning individual(s) are subject to crossover and mutation. Evolution begins with crossover, which creates new conformations by swapping genes at random points in the genotypes of a pair of individuals. Crossover is particularly effective when it combines the most optimized parts of two individuals′ genomes. Evolution continues with mutation occurring at a uniform random selection rate. If a gene of an individual is selected for mutation, a new value for the gene is drawn from a Gaussian distribution. In both mutation and crossover operations, the validity of the conformation is evaluated by checking for collisions and distant constraint violations before setting the genotype of a given individual.

Quality of a member of the population is determined by a fitness function that compares projections of the GMM conformation to the tilt series and finds the correspondence between the two sets. Since our main data structures is a 2D single-channel density image, we compare the pixel values of the projections. Projections of the GMM are generated from corresponding perspectives in the tilt series as previously described and are compared as seen in Figure 4.3. The overlap between GMM projections and tilt series images are evaluated by aligning the images and determining the Jaccard similarity coefficient, a generalized form of intersection over union evaluation, for the corresponding pixels:

$$J(x,y) = \frac{\sum_i min(x_i, y_i)}{\sum_i max(x_i, y_i)} \tag{4.1}$$

For each pair of corresponding pixels $(x_i, y_i)$, the minimal pixel value is divided by the maximal pixel value, returning a value between 0.0 and 1.0, with higher values representing higher similarity. The average of these quotients across all images in the tilt series is returned as the measure of overlap between the two projections.

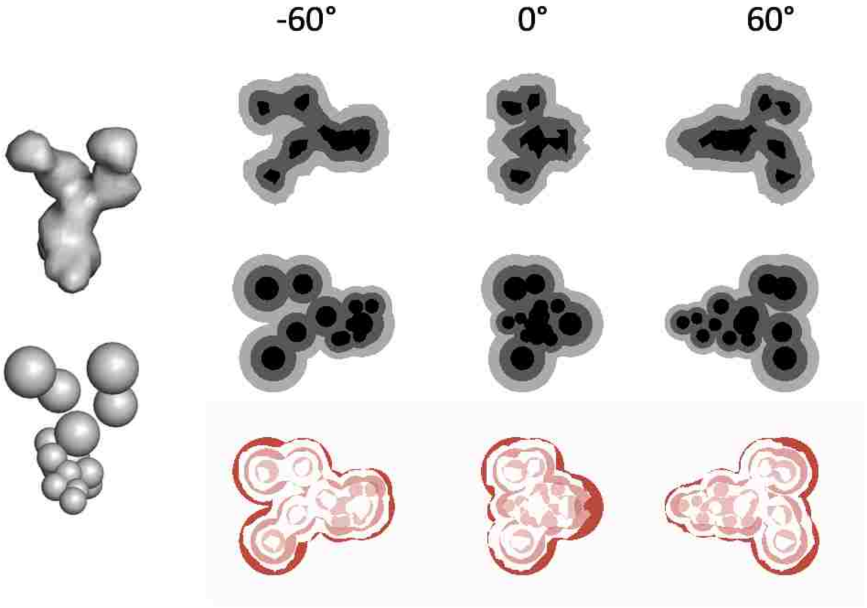


Figure 4.3: Comparison of tilt series to evaluate overlap score. Molecular data (*top left*) and GMM (*bottom left*) orthogonally projected at angles {-60,0,60} (*top/middle rows, right*). Projections from the same perspective are overlaid and evaluated for overlap (*bottom row, right*).

The population is monitored over the course of the optimization to prevent the GA from converging to a local minima. This is detected when the population fitness does not show improvement over consecutive iterations. If a local minima is detected, the worst performing subunit of the model is targeted for improvement. During these subsequent iterations, the evolutionary operations are performed only on the genes of worst performing subunit to force the GA to explore alternative conformations.

## 4.4 Experiments

### 4.4.1 GA Fitting Analysis on Simulated Single Molecules

The performance of the GA was first evaluated on a single IgE-Fc$\epsilon$RI molecule by fitting randomly generated GMM conformations to simulated tilt series of the native conformation. Experiments used single axis projection sets with rotation performed about the Y-axis. These sets were generated from [-45°, 45°], [-60°, 60°] and [-90°, 90°] angle ranges, respectively representing 90°, 120° and 180° of total range. We chose the two narrower ranges due to their experimental relevancy and the widest range was used to represent the theoretical limit of capturing a single axis tilt series. Projections were generated every 3° based on typical experimental conditions. The molecule was oriented either parallel or perpendicular to the axis of rotation as seen in Figure 4.4.

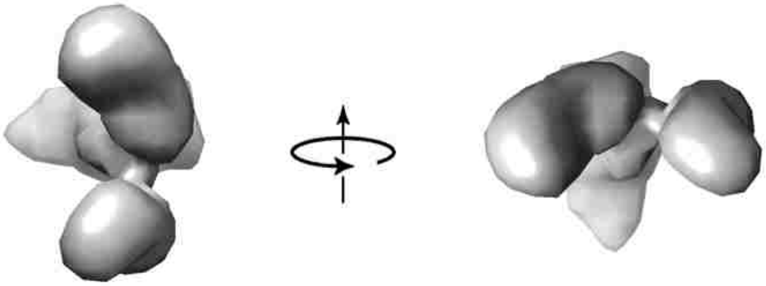
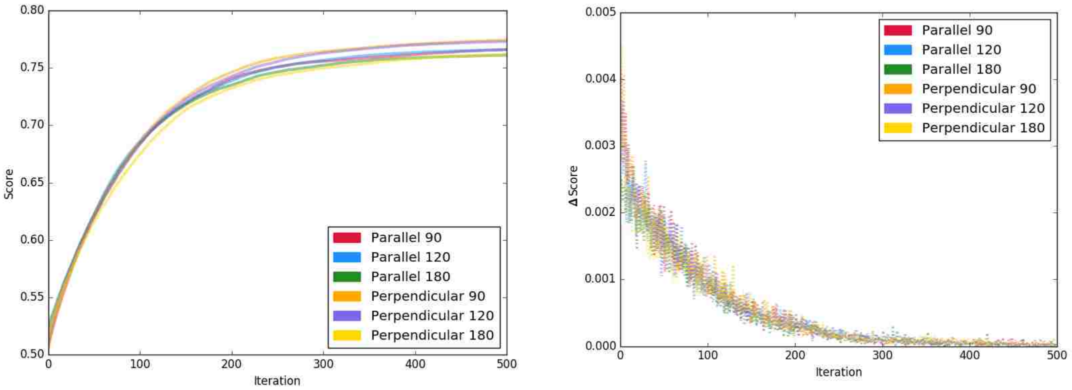


Figure 4.4: Simulated tilt series conformations parallel (*left*) and perpendicular (*right*) relative to the tilt series axis.

A 30 Å density map of the default antibody conformation was processed using the technique previously described to prepare tilt series for fitting (Section 4.2). The density map was truncated at density values 0.781 and 0.004 (value selection based on a normalized density map), discretized into 3 bins using histogram equalization, and then had a Gaussian blur applied with $\sigma = 5$ based on Å physical units. Collision was detected using a $\sigma$-scale value of 1.5 and a distance-per-residue $d$ of 3.5 Å was used for distance constraints. Variation in the initial conformations of the GMM was modeled by randomly sampling new Fab arm positions. One hundred experiments of each angle range in both parallel and perpendicular oriented data sets were performed. Each experiment ran for five hundred iterations, as shown in Figure 4.5.



(a) Scores

(b) Change in Score

Figure 4.5: Evaluation of convergence for the GA optimization. The average of and change in overlap score was evaluated for six simulated single molecule tilt series. All data sets had similar maximum overlap scores and converged at approximately the same time step.

After new populations were created using selection and crossover, all of the individuals were evaluated for mutation. During mutation, every gene of an individual, i.e., $\{x, y, z\}$ and $\{\alpha, \beta, \gamma\}$, had a 10% probability its value would be perturbed using a Gaussian distribution with $\mu = 0$ and $\sigma = 2.0$.

Overlap scores were monitored over each iteration using an improvement criteria and iteration interval. The population was considered to be in local minima if scores did not improve by at least $\epsilon$ after $\Delta$ iterations, where $\epsilon = 0.005$ and

$\Delta = 10$. To escape the local minima, the genes which corresponded to the worst performing subunit were identified and targeted for refinement until the score improved. These worst performing subunit was identified as the subunit which had the lowest individual overlap score. During these iterations, the mutation rate was increased 4-fold to account for the smaller number of genes in the targeted region.

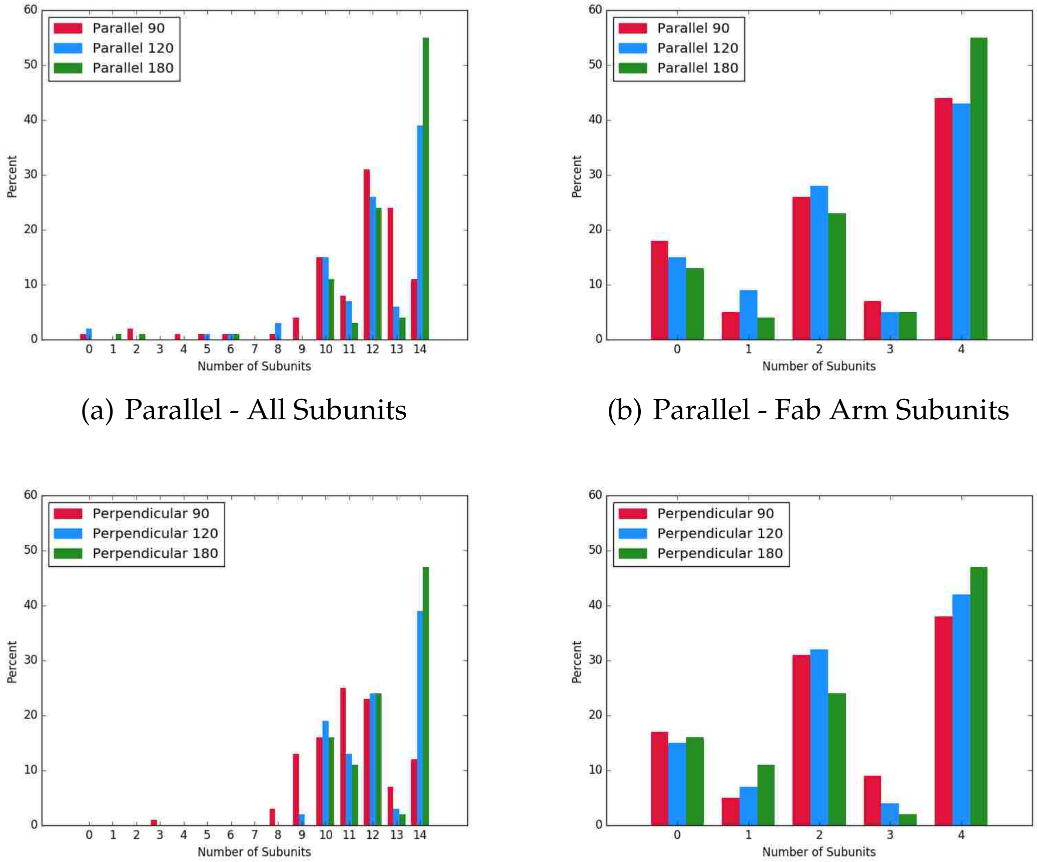
Final conformations of the single molecule fittings were evaluated for quality by measuring the distance relative to the known native conformation via subunit Root Mean Square Deviation (RMSD). In this instance, RMSD was evaluated as a distance between the GMM positions. The RMSD provides a measure of quality external to the GA by comparing the relative distances of corresponding subunits in the final and native conformations. A cutoff distance of 1.0 nm (determined from subunit RMSD distributions) was used to determine if a given subunit was properly placed. Average RMSD values are reported in Table 4.3, and the number of properly located subunits per tilt series is shown in Figure 4.6.

| | Subunit RMSD (Å) | |
|---|---|---|
| Tilt Series | Parallel | Perpendicular |
| 90° | 17.23 ± 3.02 | 15.76 ± 2.34 |
| 120° | 16.28 ± 3.03 | 14.85 ± 2.56 |
| 180° | 14.50 ± 2.84 | 13.64 ± 2.58 |

Table 4.3: Average RMSD ( Å ) with standard deviations of the fitted GMM conformations to the native GMM conformation across varied angle ranges of the tilt series imaged parallel and perpendicular to the axis of rotation.

In Table 4.3, the average values of RMSD decrease as total angle range increases. While the standard deviation indicates an overlap between angle ranges, the overall change in value supports the idea that fitting quality is potentially being enhanced by greater visibility provided by a larger range of angles. This trend is reinforced in Figures 4.6(a) and 4.6(c), where the larger angle ranges have the most correctly placed subunits (14 total) in the entire IgE molecule. Figures 4.6(b) and 4.6(d) show little variation in the number of correctly placed subunits between angle ranges, implying that the left skew in 4.6(a) is due to misplaced subunits of the constant domain. This implication is reinforced in Figure 4.6(c), where there are very few fits with less than 10 correctly placed subunits, i.e., the constant domain not being properly placed. This figure shows that the perpendicular tilt series of the IgE molecule generally had more correctly placed subunits and thus higher quality fits than the parallel tilt series. This is attributed to the fact that although the parallel pose had intuitively better conditions to model the arm differences, the perpendicular pose was able to produce better fits due to a stronger description of the constant domain.

These results show that high overlap scores from the fittest individuals may still result in misplaced subunits. These misplacements are due to the design of the GA, which is only concerned with maximizing the value of the fitness function, i.e., the overlap between projections and the tilt series. For example, initial conformation sampling may result in two subunits starting off with their positions swapped, particularly in the Fab arms. As the GA evolves, the positions of the subunits are refined until the upper limit of the distance constraints are reached. This example accounts for many cases in Figure 4.6 (c) and (d) where none or only two of the Fab arm subunits were placed correctly. If only two are correctly placed, typically a Fab arm is flipped. If none are correctly placed, the Fab arms may have completely swapped positions or both arms are flipped.



(a) Parallel - All Subunits

(b) Parallel - Fab Arm Subunits

(c) Perpendicular - All Subunits

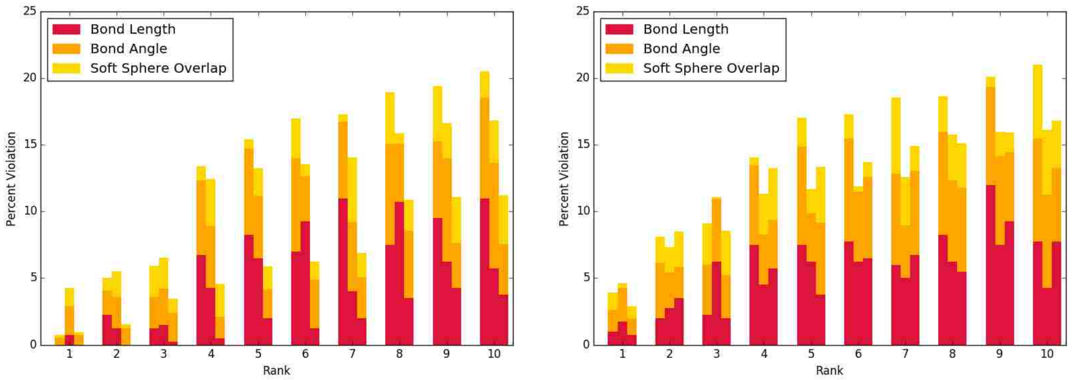
(d) Perpendicular - Fab Arm Subunits

Figure 4.6: Histograms of the percentage of subunits placed correctly. Figures a and b represent the parallel tilt series, while c and d represent the perpendicular tilt series.

### 4.4.2 Single Molecule Reconstruction Analysis

All-atom model reconstruction was performed on the final conformations of the single molecule tilt series fits to measure the effectiveness of the method at creating feasible all-atom structures. GMM conformations were reconstructed by placing the rigid molecular subunits modeled by a GMM into their fitted positions and orientations. Modeller [30] was then used to generate the flexible loop structures between GMM subunits. The reconstructions were evaluated using violations reported by Modeller in major constraints such as bond length, bond angle, and soft sphere overlap. Similar to the overlap score reported by the GA, the violation metrics are not dependent on knowledge of native conformations or reconstructed density maps.

Candidate structures were found by analyzing the percentages of reported constraint violations in the reconstructions. Figure 4.7 reports the ten conformations with the lowest average number of violations for each tilt series and angle range. Conformations with lower ratios of violations to constraints are considered higher quality. This figure shows that within these top ranked conformations, the ratios of constraint violations in each ranked conformation typically decrease as the angle range increases, in our case more so for the parallel conformation than the perpendicular. This is consistent with previous observations that greater visibility as provided by larger ranges of angles leads to higher quality conformations.



(a) Parallel Tilt Series

(b) Perpendicular Tilt Series

Figure 4.7: The *red*, *orange*, and *yellow* segments are the percentage of violations of all constraints for bond lengths, bond angles and soft sphere overlaps, respectively. The leftmost bar in each rank cluster is the 90° angle range, 120° is the middle bar, and 180° is the rightmost bar.

The results of the violation analysis were investigated by comparing the overlap scores, density map cross correlation, and subunit placement of the top ten conformations and the rest of the population. Figures 4.8 and 4.9 visualize the performance of the top ten conformations in these metrics in the parallel and perpendicular single molecule tilt series, respectively. The values in the overlap score histograms differ only slightly from the line representing average score of each distribution, but are typically above. The top ten conformations are scattered throughout and do not show any clustering in any angle range or tilt series. This indicates there is some association between overlap score and conformation quality, but it is more so a measure of GA progress. Cross correlation scores in both tilt series exhibit the same improvement as the angle range increases, and the top ten conformations typically have cross correlation scores above average, similar to what was seen in overlap score. Nearly all top ten conformation in each tilt series have one hundred percent correct subunit placement, and that amount improves as the angle range increases.

From the top ten candidate conformations for each tilt series and angle range, one candidate was selected to be the representative sample based on the calculated scores. These selected candidates were reconstructed and displayed in Figure 4.10. Overall, the top conformation for each tilt series and angle range have high cross correlation, similar overlap score, low violation percentage and high subunit placement rankings. Overlap score and cross correlation metrics emphasize quality of fit with the data, whereas violation percentage emphasizes conformational validity. These results showed that the conformations with the lowest percentage of Modeller violations match the true conformation of the molecule in the tilt series, especially when larger angle ranges are used, and certain conformation orientations are better than others for tilt series fitting. Note, subunit placement cannot be applied to experimental data since the true conformation of the imaged structure being fit is not known.

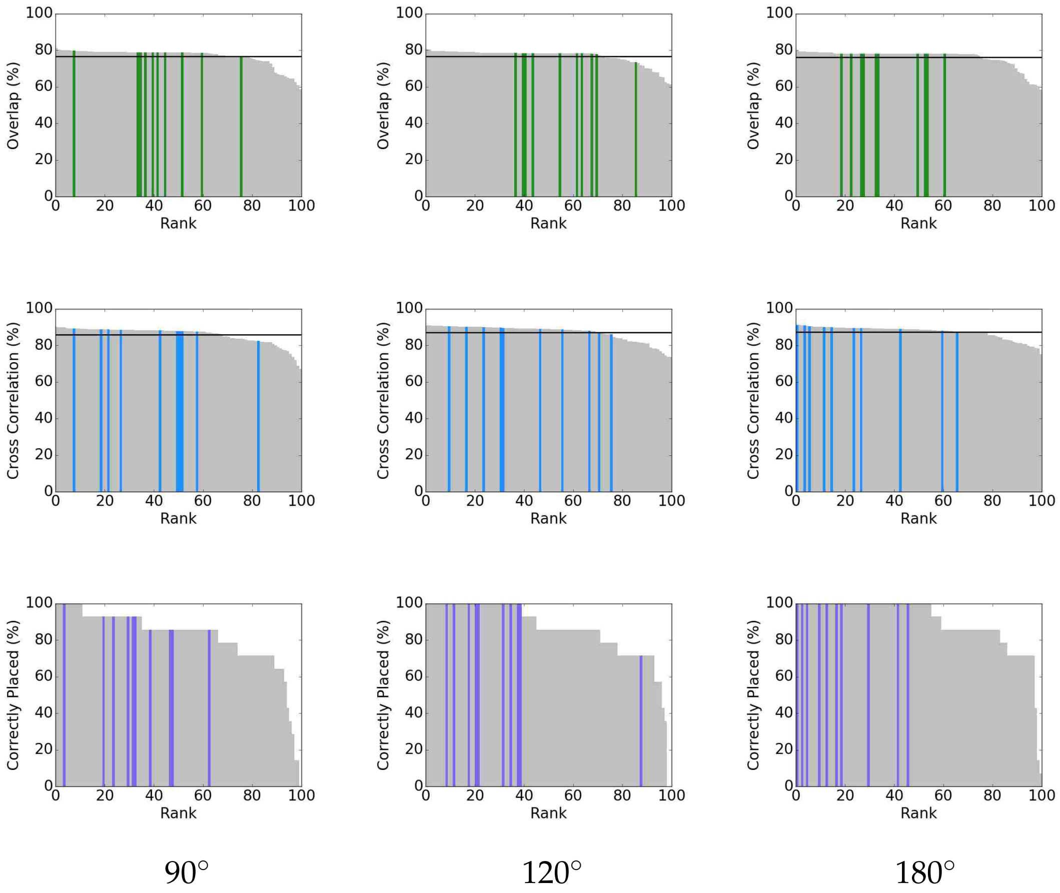


Figure 4.8: Performance of the ten conformations with the least percentages of Modeller violations in the **parallel** single molecule tilt series. From left to right are the 90°, 120°, and 180° angle ranges. From top to bottom are overlap score (green), cross correlation with density model (blue), and percentage of correctly placed subunits (purple). Top performing conformations are visualized as the lighter colored bars in each plot. The average overlap scores and average cross correlations are represented as horizontal lines in the first two rows of plots.

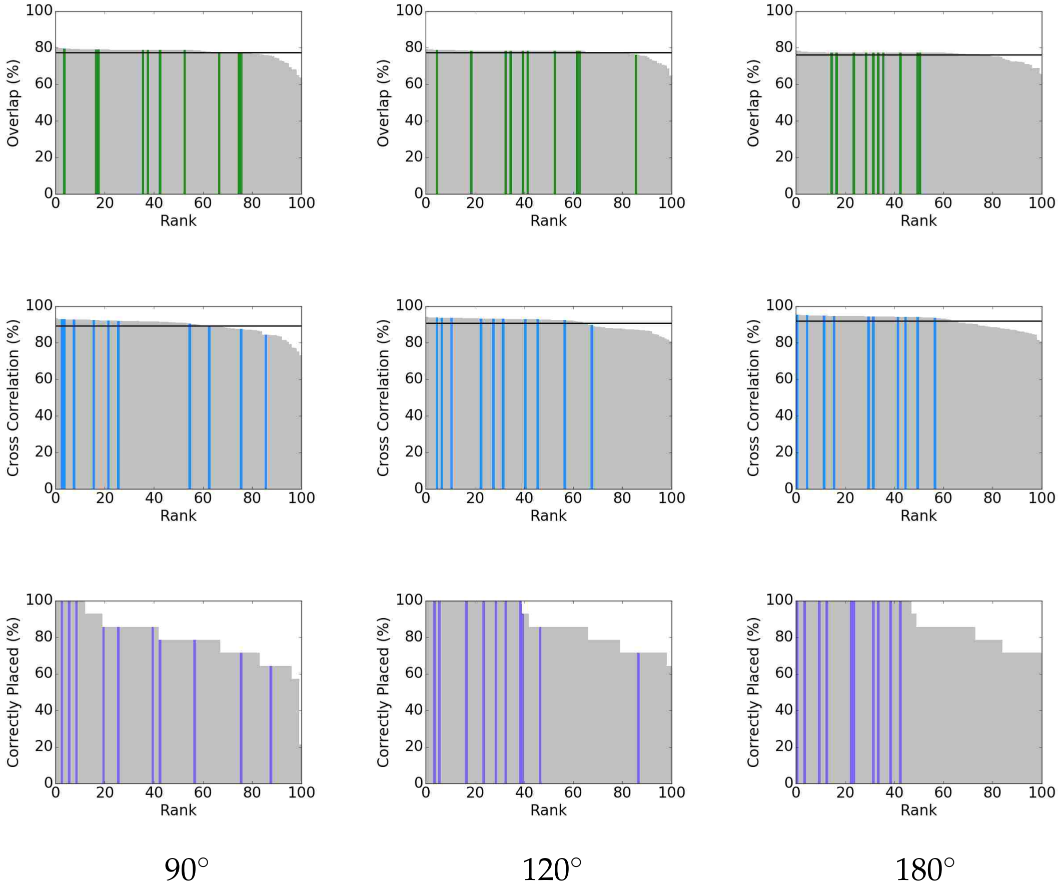
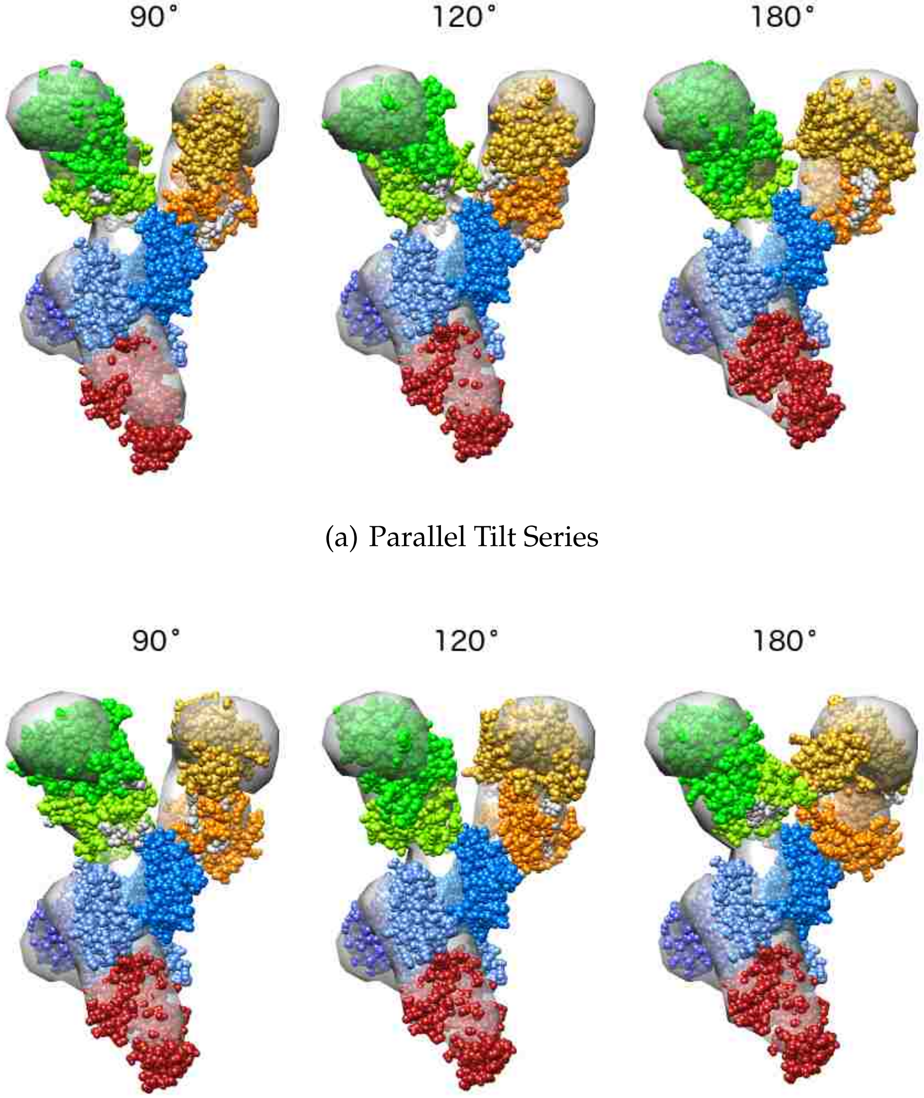


Figure 4.9: Performance of the ten conformations with the least percentages of Modeller violations in the **perpendicular** single molecule tilt series. From left to right are the 90°, 120°, and 180° angle ranges. From top to bottom are overlap (green), cross correlation with density model (blue), and percentage of correctly placed subunits (purple). Top performing conformations are visualized as the lighter colored bars in each plot. The average overlap scores and average cross correlations are represented as horizontal lines in the first two rows of plots.



(a) Parallel Tilt Series

(b) Perpendicular Tilt Series

Figure 4.10: All-atom reconstructions of the top candidates. Parallel tilt series experiments are shown in (a) and perpendicular in (b). In both (a) and (b), top candidates displayed left to right are 90°, 120°, and 180° angle ranges, respectively.

### 4.4.3 Simulated Aggregate Tilt Series Fitting

The performance of the GA was next evaluated by fitting an antibody-antigen aggregate generated using a Monte Carlo-based method [84]. The aggregate structure contains three IgE-Fc$\epsilon$RI molecules and two DF3 molecules. Performing GMM construction on DF3 resulted in a single subunit structure.

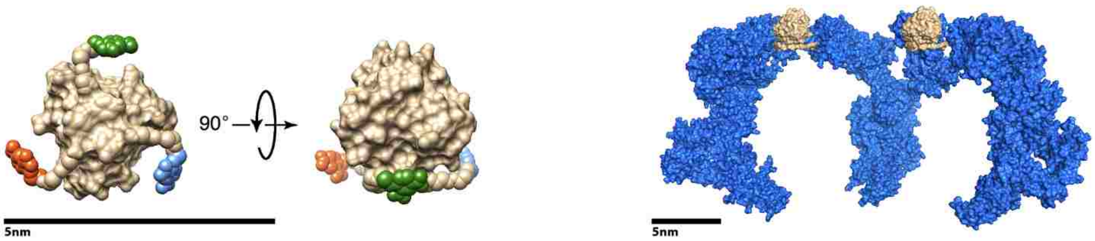
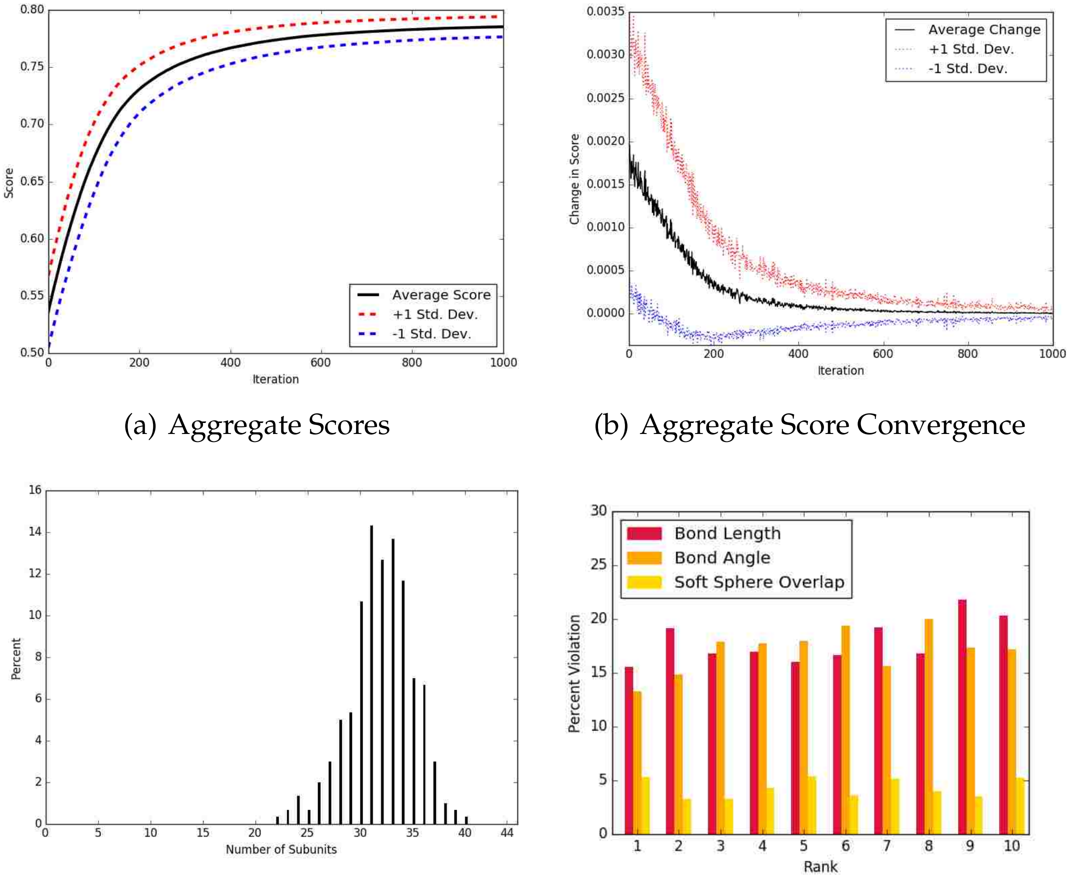


Figure 4.11: (*left*) The molecular structure of synthetic antigen DF3 (*tan*). The fibritin trimer has 3 DNP linkers (*red*, *green* and *blue*) attached to the N-termini of trimer subunits. (*right*) Aggregate structure used to evaluate tilt series fitting. The aggregate is composed of 2 DF3 antigen (*tan*) and 3 IgE antibodies (*blue*)

Randomly generated conformations of the aggregate structure were sampled and then fit to a [-45°, 45°] angle range tilt series with an increment of 3° of its native aggregate density map, representing 90° of total range. This angle range was chosen for its experimental relevancy, as tilt series with smaller ranges are less costly and are likely to make up the majority of experimental data sets. IgE were sampled using the same scheme described in Section 4.4.1, where as DF3 had their position randomly sampled within a 1 nm cube of the default conformation. Figures 4.12(a) and 4.12(b) show convergence within one thousand iterations for each of 300 runs of aggregate fitting.



(a) Aggregate Scores

(b) Aggregate Score Convergence

(c) Correctly Placed Subunits

(d) Modeller Violation Percentages

Figure 4.12: a.) Evaluation of scores for GA optimization of aggregate containing 3 IgE-Fc$\epsilon$RI and 3 DF3 molecules. 300 runs were analyzed. b.) Convergence analysis of the aggregate fitting scores. c.) Correct subunit placement historgram for 300 aggregate fittings. d.) The *red*, *orange*, and *yellow* segments are the ratios of violations to constraints for bond lengths, bond angles and soft sphere overlaps, respectively, for the top 10 scoring conformations.

Final aggregate GMM conformations were evaluated for quality by measuring the RMSD between the GMM and the native conformation. The average RMSD of the 300 aggregate fittings was 20.41 Å ± 4.05 Å. The same cutoff distance of 1.0 nm was used to determine if a given subunit was properly placed. The number of properly located subunits in the whole aggregate is shown in Figure 4.12(c).

The IgE-Fc$\epsilon$RI and DF3 aggregate had a total of 44 subunits - 2 DF3, 30 IgE base (10x3), and 12 IgE Fab Arm (4x3). Figure 4.12(c) shows that majority of the fittings placed between 30 and 35 out of 44 subunits correctly, but no fitting achieved over 40 subunits. Fits of this quality were expected based on the complexity of the fitting problem. This complexity arises from the number of subunits being fit, the amount of subunit overlap in the tilt series, and the potential to swap subunit locations during initial conformation sampling as described in the end of Section 4.4.1. All-atom reconstruction was performed on each of the final fitted aggregate conformations as described in Section 4.4.2. Violation percentages from the top 10 best fit aggregate conformations are shown in Figure 4.12(d). Percentages of violations of reconstructed candidates are higher for this data set than in either of the single molecule tilt series, but this was expected due to the number of molecules in the aggregate and complexity of the fitting problem.

Performance of the aggregate fittings in overlap score, cross correlation with density model, and correctly placed subunits were similar to those shown in Figures 4.8 and 4.9. Higher cross correlation scores and correctly placed subunits are associated with conformations with fewer violations. The selected aggregate shown in Figure 4.13 is the best fit based on violation percentage. This aggregate candidate provides a good fit with only 4 of the 44 subunits out of place. The highest RMSD of a single subunit found was 15.0 Å. This result further confirms the hypothesis that reconstruction violations are a valid final indicator of conformational quality.

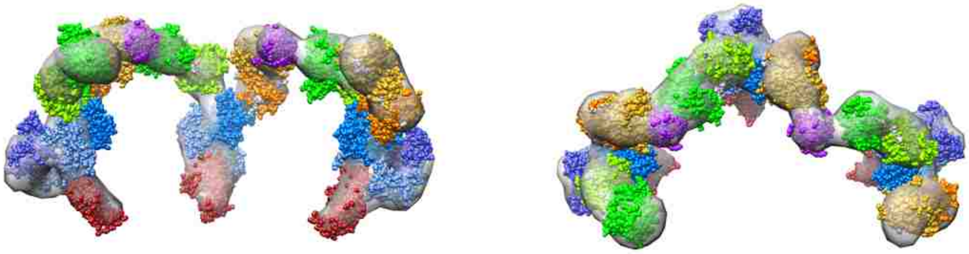


Figure 4.13: Top selected candidate from aggregate fittings. Front perspective (left) and top perspective (right) show that a vast majority of the subunits are correctly placed. Only 4 of the 44 subunits were found to have an RMSD $>$ 10.0 Å. The maximum RMSD of a single subunit was 15.0 Å.

### 4.4.4 Experimental Data Fitting

Experimental tilt series of unbound IgE-Fc$\epsilon$RI were captured using a Titan Krios by our collaborators at Sanford Burnham Prebys Medical Discovery Institute. Data was collected from an angle range of [-45°, 45°] in 3° intervals, resulting in a tilt series of 31 images. The tilt series was processed using the methods described in Section 4.2. Initial and processed density maps are shown in Figure 4.14

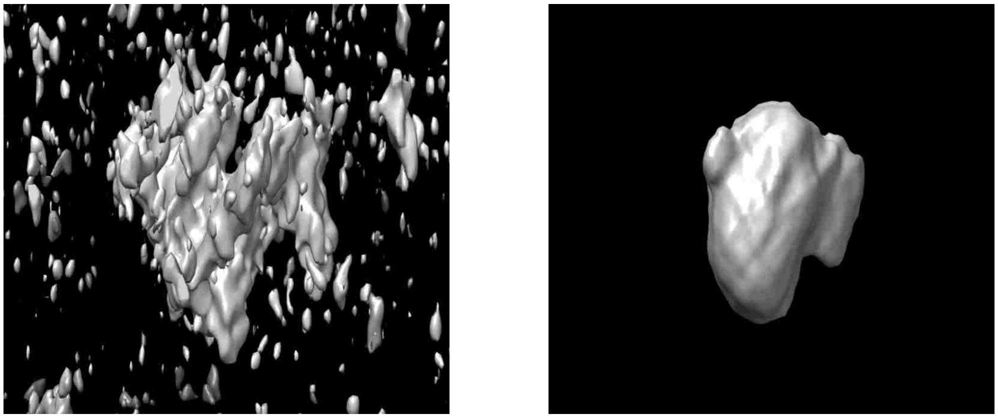
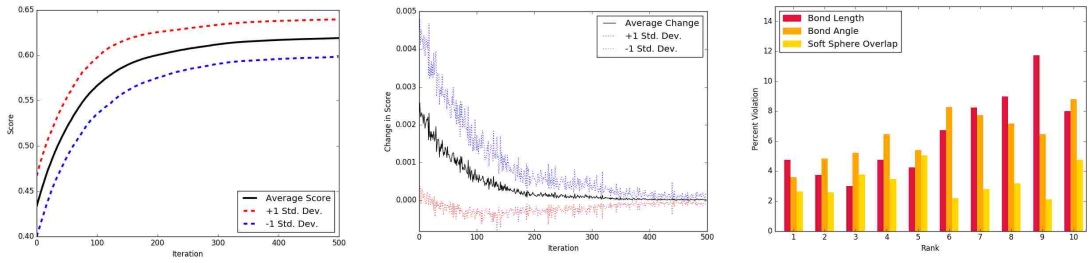


Figure 4.14: Experimental conformation of a single IgE captured by our collaborators (*left*) and the conformation after post processing (*right*).

Fitting of the processed tilt series began with a rigid body fit of the all-atom model to the reconstruction to find a default conformation for fitting using Chimera [95]. This default conformation was subjected to the sampling scheme described in Section 4.4.1 to create initial conformations for fitting. The same antibody model, GA setup, local minima detection parameters, and experimental volume and length also defined in Section 4.4.1 were used. All-atom model reconstruction was performed on the final conformations (as described in Section 4.4.2). Convergence and reconstruction violation results of these runs are shown in Figure 4.15.



(a) Scores (b) Convergence (c) Violation Percentage

Figure 4.15: a.) Evaluation of scores for GA optimization of an experimental tilt series. 100 runs were analyzed. b.) Convergence analysis of the experimental tilt series fitting scores. c.) The *red*, *orange*, and *yellow* segments are the ratios of violations to constraints for bond lengths, bond angles and soft sphere overlaps, respectively for the top 10 scoring conformations of the experimental tilt series.

Since RMSD is not available for experimentally imaged structures, candidates for the conformation with a best fit were found by analyzing a combination of metrics including overlap score, value of cross correlation with the density map, and reconstruction violations. Figure 4.16 displays three conformations that highlight differences between conformations selected with an emphasis on individual metrics.

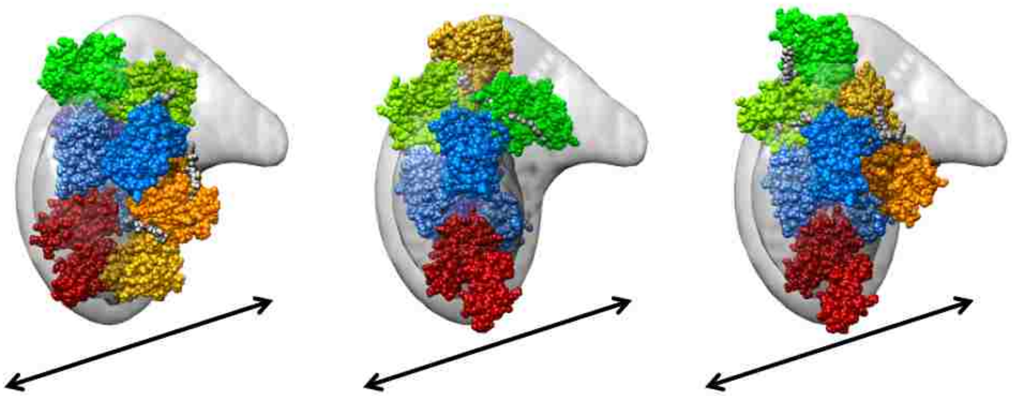


Figure 4.16: Top selected candidates from experimental fitting. The different reconstructions represent an emphasis on a particular metric for evaluation including overlap score (*left*), cross correlation value (*center*) and reconstruction violations (*right*). The axis of rotation is displayed below the conformations.

When an emphasis is placed solely on overlap score, conformations tend to overfit the tilt series and generate conformations with low regard for reconstruction violations (Figure 4.16, *left*). This is reinforced when looking at the all atom reconstruction from the perspectives of the tilt series. The best conformation based on overlap score is shown from different tilt series perspectives in the top row of Figure 4.17. At the perspective at -45° (*left column*), it is clearly seen that the other two conformations (*middle* and *bottom rows*) allow for the $\alpha$-subunit of Fc$\epsilon$RI (*red atoms*) to protrude out of the density description. This makes sense since overlap score emphasizes the highest possible overlap in all perspectives rather than conformational validity.

Conformations fit the density map better when focusing on cross correlation (Figure 4.16, *center*). The impact of the emphasis on cross correlation is seen looking at the candidate selected (Figure 4.17, *middle row*). This candidate has a worse fit at the -45° perspective than the overlap score candidate (*top row*), but a better fit in the 45° perspective. However, there is no guarantee that the density map is a completely accurate representation of the molecule due to the previously described issues with reconstruction evaluation in Section 2.3.1, e.g., stretching due to the missing wedge problem.

Reconstruction violations are a great indicator for structural validity and, unlike the previous two metrics, this metric is not dependent on conformational fitness. It is much easier to filter preliminary conformations using reconstruction violations as the primary metric (Figure 4.16, *right*) because of these reasons. In the bottom row of Figure 4.17 at -45° ($left\ column$) there is a similarity in the Fc$\epsilon$RI $\alpha$-subunit conformation to the cross correlation candidate (*red atoms*, *middle* and *bottom rows*), but the constraint violations candidate had better overlap. All candidates look similar at 0° (Figure 4.17, *center column*), but at 45° the one based on constraint violations appears to have the lowest overlap (*bottom right*). It follows that overlap is lower when considering violation constraints because those conformations are not as extreme, e.g., linkers are not stretched to full length, subunits are not in close proximity, etc. For candidates of the other two metrics, conformational validity is only based on the GMM, which is a more relaxed description of the linker than an all-atom reconstruction.

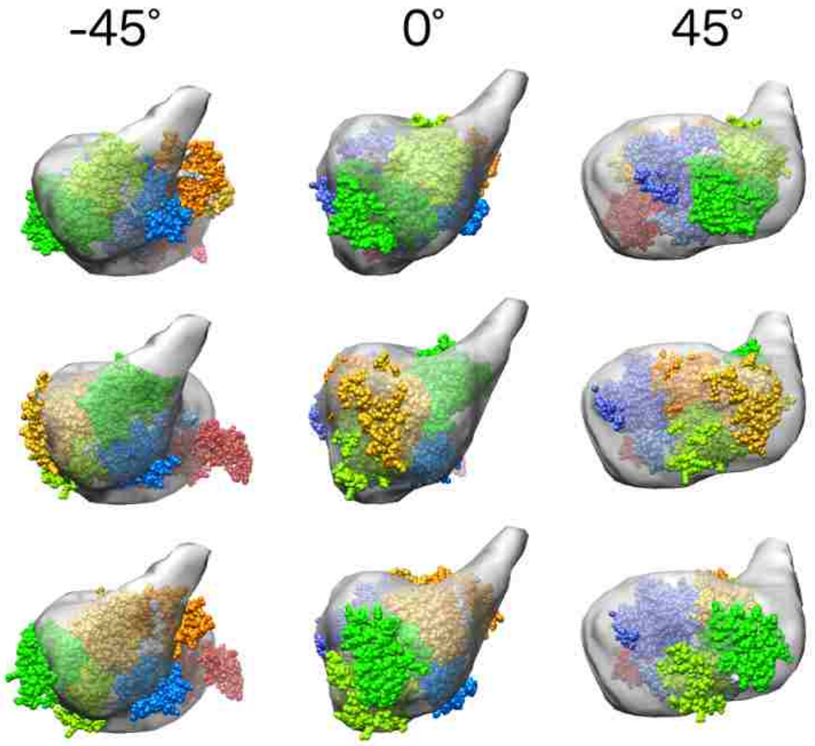


Figure 4.17: All-atom reconstructions viewed at tilt series perspectives. Best candidates selected by overlap score, cross correlation value and constraint violations are shown (*top*, *middle* and *bottom rows*, respectively). Tilt series perspectives from −45° are displayed in the left column, from 0° in the middle, and from 45° in the right.

# Chapter 5

# Conclusions and Future Work

The research presented here describes multi-resolution simulation and analysis techniques that enable the modeling of large molecular structures and their interactions. The focus of the work is on the IgE antibody aggregation problem, an assembly process associated with the human allergy immune response. We combined both computational geometry and statistical techniques to generate molecular models that were more efficient to simulate and increased the breath of analysis modeling the system at multiple levels of detail. We use our methods to understand how allergen structure and valency impact antibody aggregation. In our effort to study geometric packing of large protein complexes, we developed a rigid body aggregation model and investigated the impact of model resolution on aggregate formation and clustering. In a second body of work, efforts to study the allergic mechanisms were pursued developing flexible molecular models to fit experimentally collected aggregate structures.

In our rigid body modeling work, we focused on developing methodologies for simulating and analyzing aggregate formation. In developing simplified models based on experimentally derived data, we were able to study aggregate formation and packing structures under biologically-relevant conditions. The method was used to study how ligand valency impacts aggregation by comparing 2 similarly sized antigen, bivalent DCT2 and trivalent DF3. In the interest of improving computational performance while preserving packing structures, we examined simulations with lower resolution models. We evaluated the impact that model reduction had on our simulation, both in terms of evaluation time and model quality. Our analysis showed that time always decreased with lowered levels of resolution and in certain circumstances, a loss in resolution did not affect the results. We performed a clustering analysis of the DF3 aggregation at different resolutions to compare with experimental data and were able to reproduce the experimental Hopkins statistic metric of cluster formation at all resolutions.

To further our investigation of rigid body model resolution reduction, we utilized RBM to quantify the impact of model resolution on simulation quality. We built RBMs that reflect steric constraints due to molecular conformation allowing us to examine differences in aggregate formation across resolutions. It was shown that model resolution has negligible impact on DF3, and minimal impact on Pen a 1. This was attributed to the geometric nature of the proteins being affected by volume reduction (globular vs rod-like shapes). These results informed us how to tailor the amount of reduction applied to the molecular structures being modeled. Surface/volumetric analysis is necessary to ensure our model construction/reduction appropriately capture molecular topology when dealing with more complex molecular structures.

In our flexible molecular modeling work, we presented a method for fitting of GMM models to projections from tomographic tilt series. The intended goal of this work was to enable the fitting of models to experimentally collected aggregate structures. GMM representations of models of the IgE-Fc$\epsilon$RI complex and antigen DF3 were used to directly fit tilt series projections and we evaluated the method's ability to perform flexible fitting using GA optimization. We performed three evaluations, fitting tilt series from a simulated single antibody model, an aggregate model and an experimental data set. We were able to conclude that the method was able to reconstruct all-atom models directly from tilt series. While visual obstruction can prove challenging, fitting tilt series with flexible structures is possible even with small angle ranges. Different metrics were shown to provide different fits, therefore implementing additional metrics should provide alternative candidates. These measurements, which may be more expensive, e.g., conformational free energy, can provide alternative metrics for candidate evaluation. We showed that our method is capable of fitting low resolution (20Å - 40Å) tilt series directly. This is an improvement over standard methods which would have to fit a low resolution reconstructed density map with a highly detailed atomic model. Standard methods are great with high resolution data, but without enough information about the molecular conformation these methods run into problems with overfitting and structural distortions.

A variety of extension to these works are viable areas for development. When considering rigid body modeling, aggregation dictates how the immune systems responds to a given threat. Extension of this work could be used as a prediction tool for allergen response severity given an the allergens size/valency. Another extension could focus on antigen structure modification that could be used to treat hypersensitivity. Antigen structural modification along with aggregation simulation analysis can be used to understand how to effectively neutralize the response of a given allergen e.g., the introduction of binding competitors or design of hyper-allergenic products.

For future work in our flexible GMM model, we believe an elastic constraint could be incorporated into the existing distance constraint for loops connecting subunits. Even though loop constraint evaluation during simulation would increase with this change, extreme distances would be relaxed over time. This would help in instances where multiple loops connect subunits; for example, the loops connecting the Fab arm termini in our antibody model. Alternative GMM density models, e.g., full model density map vs subunit density map normalization, can also be used to fine tune the fitting process. Another feature that would make the method much more efficient would be to adapt the solution to GPU (graphics processing unit) based computation. The problem of image fitting is naturally positioned to take advantage of the specialty GPU hardware and could enable the generation of significantly more candidate structures given the same wall time.

# Appendices

Table A1: Rule Set for Strand I ($T_I$) of Pen a 1 in pseudo BioNetGen language format. Letters in parentheses represent free binding sites. 'IgE' in parentheses represent occupied binding sites with the subscript indicating which site is occupied. Omitted letters represent binding sites not included in the rule (can be free or occupied). Dissociations are addressed with complementary rules (not shown), with rates k$_r$ = 0.01 s$^{-1}$.

| Binding Site | Reaction Rule | Binding Rate |
|---|---|---|
| A | $T_I$(A,B) + IgE → $T_I$(IgE$_A$,B) | k$_{f1}$ |
| | $T_I$(A,IgE$_B$) + IgE → $T_I$(IgE$_A$,IgE$_B$) | k$_{f2}$ |
| B | $T_I$(A,B,C,D) + IgE → $T_I$(A,IgE$_B$,C,D) | k$_{f1}$ |
| | $T_I$(IgE$_A$,B,C) + IgE → $T_I$(IgE$_A$,IgE$_B$,C) | k$_{f2}$ |
| | $T_I$(A,B,IgE$_C$) + IgE → $T_I$(A,IgE$_B$,IgE$_C$) | k$_{f2}$ |
| | $T_I$(A,B,C,IgE$_D$) + IgE → $T_I$(A,IgE$_B$,C,IgE$_D$) | k$_{f3}$ |
| | $T_I$(IgE$_A$,B,IgE$_C$) + IgE → $T_I$(IgE$_A$,IgE$_B$,IgE$_C$) | k$_{f4}$ |
| C | $T_I$(B,C,D,E,F) + IgE → $T_I$(B,IgE$_C$,D,E,F) | k$_{f1}$ |
| | $T_I$(IgE$_B$,C,D) + IgE → $T_I$(IgE$_B$,IgE$_C$,D) | k$_{f2}$ |
| | $T_I$(B,C,IgE$_D$) + IgE → $T_I$(B,IgE$_C$,IgE$_D$) | k$_{f2}$ |
| | $T_I$(B,C,D,IgE$_E$) + IgE → $T_I$(B,IgE$_C$,D,IgE$_E$) | k$_{f3}$ |
| | $T_I$(B,C,D,E,IgE$_F$) + IgE → $T_I$(B,IgE$_C$,D,E,IgE$_F$) | k$_{f3}$ |
| | $T_I$(IgE$_B$,C,IgE$_D$) + IgE → $T_I$(IgE$_B$,IgE$_C$,IgE$_D$) | k$_{f4}$ |
| D | $T_I$(B,C,D,E,F) + IgE → $T_I$(B,C,IgE$_D$,E,F) | k$_{f1}$ |
| | $T_I$(IgE$_C$,D,E,F) + IgE → $T_I$(IgE$_C$,IgE$_D$,E,F) | k$_{f2}$ |
| | $T_I$(C,D,IgE$_E$) + IgE → $T_I$(C,IgE$_D$,IgE$_E$) | k$_{f2}$ |
| | $T_I$(C,D,E,IgE$_F$) + IgE → $T_I$(C,IgE$_D$,E,IgE$_F$) | k$_{f2}$ |
| | $T_I$(IgE$_B$,C,D,E,F) + IgE → $T_I$(IgE$_B$,C,IgE$_D$,E,F) | k$_{f3}$ |
| | $T_I$(IgE$_C$,D,IgE$_E$) + IgE → $T_I$(IgE$_C$,IgE$_D$,IgE$_E$) | k$_{f4}$ |
| | $T_I$(IgE$_C$,D,E,IgE$_F$) + IgE → $T_I$(IgE$_C$,IgE$_D$,E,IgE$_F$) | k$_{f4}$ |
| E | $T_I$(C,D,E,F) + IgE → $T_I$(C,D,IgE$_E$,F) | k$_{f1}$ |
| | $T_I$(IgE$_D$,E,F) + IgE → $T_I$(IgE$_D$,IgE$_E$,F) | k$_{f2}$ |
| | $T_I$(IgE$_C$,D,E,F) + IgE → $T_I$(IgE$_C$,D,IgE$_E$,F) | k$_{f3}$ |
| | $T_I$(E,IgE$_F$) + IgE → $T_I$(IgE$_E$,IgE$_F$) | k$_{f4}$ |
| F | $T_I$(C,D,E,F) + IgE → $T_I$(C,D,E,IgE$_F$) | k$_{f1}$ |
| | $T_I$(IgE$_D$,E,F) + IgE → $T_I$(IgE$_D$,E,IgE$_F$) | k$_{f2}$ |
| | $T_I$(IgE$_C$,D,E,F) + IgE → $T_I$(IgE$_C$,D,E,IgE$_F$) | k$_{f3}$ |
| | $T_I$(IgE$_E$,F) + IgE → $T_I$(IgE$_E$,IgE$_F$) | k$_{f4}$ |

Table A2: Rule Set for Strand II ($T_{II}$) of Pen a 1 in pseudo BioNetGen language format. Letters in parentheses represent free binding sites. 'IgE' in parentheses represent occupied binding sites with the subscript indicating which site it occupies. Omitted letters represent binding sites not included in the rule (can be free or occupied). Dissociations are addressed with complementary rules (not shown), with rates $k_r = 0.01\ s^{-1}$.

| Binding Site | Reaction Rule | Binding Rate |
|---|---|---|
| A | $T_{II}$(A,B,C) + IgE → $T_{II}$(IgE$_A$,B,C) | k$_{f1}$ |
| | $T_{II}$(A,IgE$_B$) + IgE → $T_{II}$(IgE$_A$,IgE$_B$) | k$_{f2}$ |
| | $T_{II}$(A,B,IgE$_C$) + IgE → $T_{II}$(IgE$_A$,B,IgE$_C$) | k$_{f3}$ |
| B | $T_{II}$(A,B,C,D) + IgE → $T_{II}$(A,IgE$_B$,C,D) | k$_{f1}$ |
| | $T_{II}$(IgE$_A$,B,C) + IgE → $T_{II}$(IgE$_A$,IgE$_B$,C) | k$_{f2}$ |
| | $T_{II}$(A,B,IgE$_C$) + IgE → $T_{II}$(A,IgE$_B$,IgE$_C$) | k$_{f2}$ |
| | $T_{II}$(A,B,C,IgE$_D$) + IgE → $T_{II}$(A,IgE$_B$,C,IgE$_D$) | k$_{f3}$ |
| | $T_{II}$(IgE$_A$,B,IgE$_C$) + IgE → $T_{II}$(IgE$_A$,IgE$_B$,IgE$_C$) | k$_{f4}$ |
| C | $T_{II}$(A,B,C,D) + IgE → $T_{II}$(A,B,IgE$_C$,D) | k$_{f1}$ |
| | $T_{II}$(IgE$_B$,C,D) + IgE → $T_{II}$(IgE$_B$,IgE$_C$,D) | k$_{f2}$ |
| | $T_{II}$(B,C,IgE$_D$) + IgE → $T_{II}$(B,IgE$_C$,IgE$_D$) | k$_{f2}$ |
| | $T_{II}$(IgE$_A$,B,C,D) + IgE → $T_{II}$(IgE$_A$,B,IgE$_C$,D) | k$_{f3}$ |
| | $T_{II}$(IgE$_B$,C,IgE$_D$) + IgE → $T_{II}$(IgE$_B$,IgE$_C$,IgE$_D$) | k$_{f4}$ |
| D | $T_{II}$(B,C,D,E,F) + IgE → $T_{II}$(B,C,IgE$_D$,E,F) | k$_{f1}$ |
| | $T_{II}$(IgE$_C$,D,E,F) + IgE → $T_{II}$(IgE$_C$,IgE$_D$,E,F) | k$_{f2}$ |
| | $T_{II}$(C,D,IgE$_E$) + IgE → $T_{II}$(C,IgE$_D$,IgE$_E$) | k$_{f2}$ |
| | $T_{II}$(C,D,E,IgE$_F$) + IgE → $T_{II}$(C,IgE$_D$,E,IgE$_F$) | k$_{f2}$ |
| | $T_{II}$(IgE$_B$,C,D,E,F) + IgE → $T_{II}$(IgE$_B$,C,IgE$_D$,E,F) | k$_{f3}$ |
| | $T_{II}$(IgE$_C$,D,IgE$_E$) + IgE → $T_{II}$(IgE$_C$,IgE$_D$,IgE$_E$) | k$_{f4}$ |
| | $T_{II}$(IgE$_C$,D,E,IgE$_F$) + IgE → $T_{II}$(IgE$_C$,IgE$_D$,E,IgE$_F$) | k$_{f4}$ |
| E | $T_{II}$(D,E,F) + IgE → $T_{II}$(D,IgE$_E$,F) | k$_{f1}$ |
| | $T_{II}$(IgE$_D$,E,F) + IgE → $T_{II}$(IgE$_D$,IgE$_E$,F) | k$_{f2}$ |
| | $T_{II}$(E,IgE$_F$) + IgE → $T_{II}$(IgE$_E$,IgE$_F$) | k$_{f4}$ |
| F | $T_{II}$(D,E,F) + IgE → $T_{II}$(D,E,IgE$_F$) | k$_{f1}$ |
| | $T_{II}$(IgE$_D$,E,F) + IgE → $T_{II}$(IgE$_D$,E,IgE$_F$) | k$_{f2}$ |
| | $T_{II}$(IgE$_E$,F) + IgE → $T_{II}$(IgE$_E$,IgE$_F$) | k$_{f4}$ |

# References

[1] Rob C. Aalberse. Structural biology of allergens. *Journal of Allergy and Clinical Immunology*, 106(2):228 – 238, 2000.

[2] Soman N. Abraham and Ashley L. St John. Mast cell-orchestrated immunity to pathogens. *Nature reviews. Immunology*, 10(6):440 – 452, 2010.

[3] Ibrahim Al-Bluwi, Thierry Siméon, and Juan Cortés. Motion planning algorithms for molecular simulations: A survey. *Computer Science Review*, 6(4):125 – 143, 2012.

[4] Ibrahim Al-Bluwi, Marc Vaisset, Thierry Siméon, and Juan Cortés. Modeling protein conformational transitions by a combination of coarse-grained normal mode analysis and robotics-inspired methods. *BMC Structural Biology*, 13(1), 2013.

[5] Kamal Al Nasr and Jing He. An effective convergence independent loop closure method using forward backward cyclic coordinate descent. *Int. J. Data Min. Bioinformatics*, 3(3):346–361, June 2009.

[6] Brittany Allison, Steven Combs, Sam DeLuca, Gordon Lemmon, Laura Mizoue, and Jens Meiler. Computational design of protein-small molecule interfaces. *Journal of structural biology*, 185(2):193–202, 2014.

[7] N. M. Amato, Ken A. Dill, and G. Song. Using motion planning to map protein folding landscapes and analyze folding kinetics of known native structures. *J. Comput. Biol.*, 10(3-4):239–255, 2003.

[8] Nicholas L. Andrews, Keith A. Lidke, Janet R. Pfeiffer, Alan R. Burns, Bridget S. Wilson, Janet M. Oliver, and Diane S. Lidke. Actin restricts Fc$\epsilon$RI diffusion and facilitates antigen-induced receptor immobilisation. *Nature Cell Biology*, 10(8):955–963, 2008.

[9] Nicholas L. Andrews, Janet R. Pfeiffer, A. Marina Martinez, David M. Haaland, Ryan W. Davis, Toshiaki Kawakami, Janet M. Oliver, Bridget S. Wilson, and Diane S. Lidke. Small, mobile Fc$\epsilon$RI receptor aggregates are signaling competent. *Immunity*, 31(3):469–479, 2009.

[10] R. Ayuso, S. B. Lehrer, and G. Reese. Identification of continuous, allergenic regions of the major shrimp allergen Pen a 1 (tropomyosin). *Int Arch Allergy and Immunology*, 127:27–37, 2002.

[11] O. B. Bayazit, G. Song, and N. M. Amato. Ligand binding with OBPRM and haptic user input: Enhancing automatic motion planning with virtual touch. In *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pages 954–959, 2001.

[12] Kyle A Beauchamp, Yu-Shan Lin, Rhiju Das, and Vijay S Pande. Are protein force fields getting better? a systematic benchmark on 524 diverse nmr measurements. *Journal of Chemical Theory and Computation*, 8(4):1409–1414, 2012.

[13] J. Bennell, G. Scheithauer, Y. Stoyan, and T. Romanova. Tools of mathematical modeling of arbitrary object packing problems. *Annals of Operations Research*, 179(1):343–368, 2010.

[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne. The protein data bank. *Nucleic Acids Research*, 28(1):235–242, 2000.

[15] Michael L. Blinov, James R. Faeder, Byron Goldstein, and William S. Hlavacek. BioNetGen: Software for rule-based modeling of signal transduction based on the interactions of molecular domains. *Bioinformatics*, 20(17):3289–3291, 2004.

[16] Dezsö Boda. Monte carlo simulation of electrolyte solutions in biology: in and out of equilibrium. In Ralph A. Wheeler, editor, *Annual Reports in Computational Chemistry, Volume 10*, chapter 5, pages 127–158. Elsevier B.V., Amsterdam, 2014.

[17] L Bongini, D Fanelli, F Piazza, P De Los Rios, S Sandin, and U Skoglund. Dynamics of antibodies from cryo-electron tomography. *Biophysical Chemistry*, 115(2):235–240, 2005.

[18] Dusan Bratko, Troy Cellmer, John M. Prausnitz, and Harvey W. Blanch. Molecular simulation of protein aggregation. *Biotech. and Bioeng.*, 96(1):1–8, 2007.

[19] Pablo Chacón and Willy Wriggers. Multi-resolution contour-based fitting of macromolecular structures. *Journal of Molecular Biology*, 317(3):375 – 384, 2002.

[20] Brian Y. Chen and Barry Honig. VASP: A volumetric analysis of surface properties yields insights into protein-ligand binding specificity. *PLoS Comput Biol*, 6(8), August 2010.

[21] B.Y. Chen and S. Bandyopadhyay. VASP-S: A volumetric analysis and statistical model for predicting steric influences on protein-ligand binding specificity. In *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 22–29, 2011.

[22] Jim C. Chen and Albert S. Kim. Brownian dynamics, molecular dynamics, and Monte Carlo modeling of colloidal systems. *Advances in Colloid and Interface Science*, 112(13):159 – 173, 2004.

[23] P. Cignoni, C. Montani, and R. Scopigno. A comparison of mesh simplification algorithms. *Computers & Graphics*, 22(1):37–54, 1998.

[24] Juan Cortés, Léonard Jaillet, and Thierry Siméon. Molecular disassembly with rrt-like algorithms. In *2007 IEEE International Conference on Robotics and Automation*, pages 3301–3306, April 2007.

[25] Juan Cortés, Léonard Jaillet, and Thierry Siméon. Disassembly path planning for complex articulated objects. *IEEE Trans. Robot.*, 24(2):475–481, 2008.

[26] Pilar Cossio and Gerhard Hummer. Bayesian analysis of individual electron microscopy images: Towards structures of dynamic and heterogeneous biomolecular assemblies. *Journal of Structural Biology*, 184(3):427 – 437, 2013.

[27] Flor A. Espinoza, Janet M. Oliver, Bridget S. Wilson, and Stanly L. Steinberg. Using hierarchical clustering and dendrograms to quantify the clustering of membrane proteins. *Bulletin of Mathematical Biology*, 74(1):190–211, 2012.

[28] Juan Esquivel-Rodríguez and Daisuke Kihara. Fitting multimeric protein complexes into electron microscopy maps using 3d zernike descriptors. *The Journal of Physical Chemistry B*, 116(23):6854–6861, 2012.

[29] Juan Esquivel-Rodríguez and Daisuke Kihara. Computational methods for constructing protein structure models from 3d electron microscopy maps. *Journal of Structural Biology*, 184(1):93 – 102, 2013.

[30] Narayanan Eswar, Ben Webb, Marc A. Marti-Renom, M.S Madhusudhan, David Eramian, Min-yi Shen, Ursula Pieper, and Andrej Sali. Comparative protein structure modeling using modeller. *Current Protocols in Bioinformatics*, 15(1):5.6.1–5.6.30, 09 2006.

[31] Naomi Fox, Filip Jagodzinski, Yang Li, and Ileana Streinu. Kinari-web: a server for protein rigidity analysis. *Nucleic Acids Research*, 39(suppl 2):W177–W183, 2011.

[32] Joachim Frank. *Introduction: Principles of Electron Tomography*. Springer New York, New York, NY, 2006.

[33] Joachim Frank. *Three-Dimensional Electron Microscopy of Macromolecular Assemblies: Visualization of Biological Molecules in Their Native State*. Oxford University Press, Oxford, 2006.

[34] Y Gambin, R Lopez-Esparza, M Reffay, E Sierecki, NS Gov, MHRS Genest, RS Hodges, and W Urbach. Lateral mobility of proteins in liquid membranes revisited. *Proceedings of the National Academy of Sciences of the United States of America*, 103(7):2098–2102, 2006.

[35] Thomas D. Goddard, Conrad C. Huang, and Thomas E. Ferrin. Software extensions to UCSF chimera for interactive visualization of large molecular assemblies. *Structure*, 13(3):473–482, 2005.

[36] B. Goldstein and A.S. Perelson. Equilibrium theory for the clustering of bivalent cell surface receptors by trivalent ligands. Application to histamine release from basophils. *Biophysical Journal*, 45(6):1109–1123, 1984.

[37] Frank Guarnieri. Theory and algorithms for mixed Monte Carlo-stochastic dynamics simulations. *Journal of Mathematical Chemistry*, 18(1):25–35, 1995.

[38] Sarah Güthe, Larisa Kapinos, Andreas Möglich, Sebastian Meier, Stephan Grzesiek, and Thomas Kiefhaber. Very fast folding and association of a trimerization domain from bacteriophage t4 fibritin. *Journal of Molecular Biology*, 337(4):905 – 915, 2004.

[39] Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443, 2002.

[40] I. Hashmi, B. Akbal-Delibas, N. Haspel, and A. Shehu. Guiding protein docking with geometric and evolutionary information. *J Bioinf and Comp Biol*, 10(3):1242002, 2012.

[41] Allison P. Heath, Lydia E. Kavraki, and Cecilia Clementi. From coarse-grain to all-atom: Toward multiscale analysis of protein landscapes. *Proteins: Structure, Function, and Bioinformatics*, 68(3):646–661, 2007.

[42] Richard Henderson, Andrej Sali, Matthew L Baker, Bridget Carragher, Batsal Devkota, Kenneth H Downing, Edward H Egelman, Zukang Feng, Joachim Frank, Nikolaus Grigorieff, et al. Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2):205–214, 2012.

[43] Anne M. Herian and Steve L. Taylor. Non-specific binding of anti-human IgE peroxidase-linked conjugates to legume proteins in immunoblots. *Journal of Immunological Methods*, 140(2):153 – 158, 1991.

[44] William S. Hlavacek, Richard G. Posner, and Alan S. Perelson. Steric effects on multivalent ligand-receptor binding: Exclusion of ligand sites by bound cell surface receptors. *Biophysical Journal*, 76(6):3031–3043, 1999.

[45] Brittany Hoard, Bruna Jacobson, Kasra Manavi, and Lydia Tapia. Extending rule-based methods to model molecular geometry. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1–8, 2015.

[46] Brittany Hoard, Bruna Jacobson, Kasra Manavi, and Lydia Tapia. Extending rule-based methods to model molecular geometry and 3d model resolution. *BMC Systems Biology*, 2016.

[47] D. Holowka, D. Sil, C. Torigoe, and B. Baird. Insights into immunoglobulin E receptor signaling from structurally defined ligands. *Immunol. Rev.*, 217:269–279, 2007.

[48] Brian Hopkins and J. G. Skellam. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18(70):213–227, 1954.

[49] Po-Ssu Huang, John J. Love, and Stephen L. Mayo. A de novo designed protein-protein interface. *Protein Science*, 16(12):2770–2774, 2007.

[50] Yu-Fen Huang, Haipeng Liu, Xiangling Xiong, Yan Chen, and Weihong Tan. Nanoparticle–mediated IgE–receptor aggregation and signaling in RBL mast cells. *Journal of the American Chemical Society*, 131(47):17328–17334, 2009.

[51] Ovidiu Ivanciuc, Catherine H. Schein, and Werner Braun. SDAP: Database and computational tools for allergenic proteins. *Nucleic Acids Research*, 31(1):359–362, 2003.

[52] E. Jensen-Jarolim, M. Vogel, V. Zavazal, and B. M. Stadler. Nonspecific binding of IgE to allergens. *Allergy*, 52(8):844–852, 1997.

[53] Ramesh K. Jha, Andrew Leaver-Fay, Shuangye Yin, Yibing Wu, Glenn L. Butterfoss, Thomas Szyperski, Nikolay V. Dokholyan, and Brian Kuhlman. Computational design of a PAK1 binding protein. *Journal of Molecular Biology*, 400(2):257–270, 2010.

[54] Slavica Jonić and Carlos Óscar Sánchez Sorzano. Coarse-graining of volumes for modeling of structure and dynamics in electron microscopy: Algorithm to automatically control accuracy of approximation. *IEEE Journal of Selected Topics in Signal Processing*, 10(1):161–173, 2016.

[55] Paul Joubert and Michael Habeck. Bayesian inference of initial models in cryo-electron microscopy using pseudo-atoms. *Biophysical Journal*, 108(5):1165–1175, 2015.

[56] L. E. Kavraki, P. Švestka, J. C. Latombe, and M. H. Overmars. Probabilistic roadmaps for path planning in high-dimensional configuration spaces. *IEEE Trans. Robot. Automat.*, 12(4):566–580, August 1996.

[57] Takeshi Kawabata. Multiple subunit fitting into a low-resolution density map of a macromolecular complex using a gaussian mixture model. *Biophysical Journal*, 95(10):4643 – 4658, 2008.

[58] Young C. Kim and Gerhard Hummer. Coarse-grained models for simulations of multiprotein complexes: Application to ubiquitin binding. *J. Mol. Biol.*, 375(5):1416 – 1433, 2008.

[59] Neil P. King, William Sheffler, Michael R. Sawaya, Breanna S. Vollmar, John P. Sumida, Ingemar André, Tamir Gonen, Todd O. Yeates, and David Baker. Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science*, 336(6085):1171–1174, 2012.

[60] Michael L Klein and Wataru Shinoda. Large-scale molecular dynamics simulations of self-assembling systems. *Science*, 321(5890):798–800, 2008.

[61] Jefferson D. Knight, Michael G. Lerner, Joan G. Marcano-Velázquez, Richard W. Pastor, and Joseph J. Falke. Single molecule diffusion of membrane-bound proteins: Window into lipid contacts and bilayer dynamics. *Biophysical Journal*, 99(9):2879 – 2887, 2010.

[62] Yen-Ting Lai, Neil P. King, and Todd O. Yeates. Principles for designing ordered protein assemblies. *Trends in Cell Biology*, 22(12):653–661, 2012. Special Issue  Synthetic Cell Biology.

[63] S. M. LaValle. *Planning Algorithms*. Cambridge University Press, Cambridge, U.K., 2006. Available at http://planning.cs.uiuc.edu/.

[64] S. M. LaValle and J. J. Kuffner. Randomized kinodynamic planning. *Int. J. Robot. Res.*, 20(5):378–400, May 2001.

[65] Catherine L. Lawson, Matthew L. Baker, Christoph Best, Chunxiao Bi, Matthew Dougherty, Powei Feng, Glen van Ginkel, Batsal Devkota, Ingvar Lagerstedt, Steven J. Ludtke, Richard H. Newman, Tom J. Oldfield, Ian Rees, Gaurav Sahni, Raul Sala, Sameer Velankar, Joe Warren, John D. Westbrook, Kim Henrick, Gerard J. Kleywegt, Helen M. Berman, and Wah Chiu. Emdatabank.org: unified data resource for cryoem. *Nucleic Acids Research*, 39:D456 – D464, 2011.

[66] Benjamin Leader, Quentin J Baca, and David E Golan. Protein therapeutics: a summary and pharmacological classification. *Nature reviews Drug discovery*, 7(1):21, 2008.

[67] Andrew Leaver-Fay, Michael Tyka, Steven M Lewis, Oliver F Lange, James Thompson, Ron Jacak, Kristian W Kaufman, P Douglas Renfrew, Colin A Smith, Will Sheffler, et al. Rosetta3: an object-oriented software suite for the simulation and design of macromolecules. In *Methods in enzymology*, volume 487, pages 545–574. Elsevier, 2011.

[68] Da-Wei Li, Sandipan Mohanty, Anders Irbäck, and Shuanghong Huo. Formation and growth of oligomers: A monte carlo study of an amyloid tau fragment. *PLoS Comput Biol*, 4(12):1–12, 12 2008.

[69] Yunqi Li and Qingrong Huang. Influence of protein self-association on complex coacervation with polysaccharide: A Monte Carlo study. *J. Phys. Chem. B*, 117(9):2615–2624, 2013.

[70] Yunqi Li, Tongfei Shi, Lijia An, and Qingrong Huang. Monte Carlo simulation on complex formation of proteins and polysaccharides. *J. Phys. Chem. B*, 116(10):3045–3053, 2012.

[71] Björn F. Lillemeier, Janet R. Pfeiffer, Zurab Surviladze, Bridget S. Wilson, and Mark M. Davis. Plasma membrane-associated proteins are clustered into islands attached to the cytoskeleton. *Proceedings of the National Academy of Sciences*, 103(50):18992–18997, 2006.

[72] Steffen Lindert, Nathan Alexander, Nils Wötzel, Mert Karakaş, Phoebe L. Stewart, and Jens Meiler. Em-fold: De novo atomic-detail protein structure

determination from medium-resolution density maps. *Structure*, 20(3):464 – 478, 2012.

[73] Daniel Lingwood and Kai Simons. Lipid rafts as a membrane-organizing principle. *Science*, 327(5961):46–50, 2010.

[74] T. Lozano-Pérez. Spatial planning: A configuration space approach. *IEEE Trans. Comput.*, C-32:108–120, 1983.

[75] Vladan Lučić, Friedrich Förster, and Wolfgang Baumeister. Structural studies by electron tomography: from cells to molecules. *Annu. Rev. Biochem.*, 74:833–865, 2005.

[76] Paul Mach and Patrice Koehl. Geometric measures of large biomolecules: Surface, volume, and pockets. *Journal of Computational Chemistry*, 32(14):3023–3038, 2011.

[77] Paul Mach and Patrice Koehl. An analytical method for computing atomic contact areas in biomolecules. *Journal of Computational Chemistry*, 34(2):105–120, 2013.

[78] Avanika Mahajan, Dipak Barua, Patrick Cutler, Diane S. Lidke, Flor A. Espinoza, Carolyn Pehlke, Rachel Grattan, Yuko Kawakami, Chang-Shung Tung, Andrew R. M. Bradbury, William S. Hlavacek, and Bridget S. Wilson. Optimal aggregation of Fc$\epsilon$RI with a structurally defined trivalent ligand overrides negative regulation driven by phosphatases. *ACS Chemical Biology*, 9(7):1508–1519, 2014.

[79] Kasra Manavi and et. al. Fitting tomographic tilt series using gaussian mixture models and genetic algorithm optimization. In *In Preparation.*, 2018.

[80] Kasra Manavi, Bruna Jacobson, Brittany Hoard, and Lydia Tapia. Influence of model resolution on geometric simulations of antibody aggregation. *Robotica*, 34(08):1754–1776, 2016.

[81] Kasra Manavi, Alan Kuntz, and Lydia Tapia. Geometrical insights into the process of antibody aggregation. In *Proc. AAAI Conf. Workshop Art. Int. and Rob. Meth. Comp. Bio. (AIRMCB)*, 2013.

[82] Kasra Manavi and Lydia Tapia. Influence of model resolution on antibody aggregation simulations. In *Proc. Robotics Methods for Structural and Dynamic Modeling of Molecular Systems Workshop at Robotics Science and Systems Conference*, 2014.

[83] Kasra Manavi, Sahba Tashakkori, and Lydia Tapia. Gaussian mixture models with constrained flexibility for fitting tomographic tilt series. In *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics*, 2017.

[84] Kasra Manavi, Bridget S. Wilson, and Lydia Tapia. Simulation and analysis of antibody aggregation on cell surfaces using motion planning and graph analysis. In *Proc. ACM Conference on Bioinformatics, Computational Biology and Biomedicine (ACM-BCB)*, pages 458–465, 2012.

[85] Autodesk Maya, 2014.

[86] James J. McGregor. Backtrack search algorithms and the maximal common subgraph problem. *Software: Practice and Experience*, 12(1):23–34, 1982.

[87] Brad L. Miller and David E. Goldberg. Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, 9:193–212, 1995.

[88] Niloy J. Mitra and Mark Pauly. Shadow art. *ACM Trans. Graph.*, 28(5):156:1–156:7, December 2009.

[89] Michael I. Monine, Richard G. Posner, Paul B. Savage, James R. Faeder, and William S. Hlavacek. Modeling multivalent ligand-receptor interactions with steric constraints on configurations of cell-surface receptor aggregates. *Biophysical Journal*, 98(1):48–56, 2010.

[90] Vijay Natarajan, Patrice Koehl, Yusu Wang, and Bernd Hamann. Visual analysis of biomolecular surfaces. In *Visualization in Medicine and Life Sciences*, pages 237–255. Springer Berlin Heidelberg, 2008.

[91] Dan V Nicolau, John F Hancock, and Kevin Burrage. Sources of anomalous diffusion on cell membranes: A monte carlo study. *Biophysical Journal*, 92(6):1975–1987, 2006.

[92] Eric Paquet and Herna L Viktor. Molecular dynamics, monte carlo simulations, and langevin dynamics: a computational review. *BioMed research international*, 2015, 2015.

[93] Lili X. Peng, Lei Yu, Stephen B. Howell, and David A. Gough. Aggregation properties of a polymeric anticancer therapeutic: A coarse-grained modeling study. *Journal of Chemical Information and Modeling*, 51(12):3030–3035, 2011.

[94] Juan R Perilla, Boon Chong Goh, C Keith Cassidy, Bo Liu, Rafael C Bernardi, Till Rudack, Hang Yu, Zhe Wu, and Klaus Schulten. Molecular dynamics simulations of large macromolecular complexes. *Current opinion in structural biology*, 31:64–74, 2015.

[95] Eric F. Pettersen, Thomas D. Goddard, Conrad C. Huang, Gregory S. Couch, Daniel M. Greenblatt, Elaine C. Meng, and Thomas E. Ferrin. Ucsf chimera–a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, 25(13):1605–1612, 2004.

[96] Linda J. Pike. Rafts defined: a report on the keystone symposium on lipid rafts and cell function. *Journal of Lipid Research*, 47(7):1597–1598, 2006.

[97] Richard G. Posner, Kala Subramanian, Byron Goldstein, James Thomas, Toni Feder, David Holowka, and Barbara Baird. Simultaneous cross-linking by two nontriggering bivalent ligands causes synergistic signaling of IgE Fc$\epsilon$RI complexes. *J. of Immunology*, 155(7):3601–3609, 1995.

[98] Barak Raveh, Angela Enosh, Ora Schueler-Furman, and Dan Halperin. Rapid sampling of molecular motions with prior information constraints. *PLoS Comput Biol*, 5(2):e1000295, February 2009.

[99] Gerald Reese, Julia Viebranz, Susan M. Leong-Kee, Matthew Plante, Iris Lauer, Stefanie Randow, Mar San-Miguel Moncin, Rosalia Ayuso, Samuel B. Lehrer, and Stefan Vieths. Reduced allergenic potency of VR9-1, a mutant of the major shrimp allergen Pen a 1 (tropomyosin). *The Journal of Immunology*, 175(12):8354–8364, 2005.

[100] Juan Rivera and Alasdair M. Gilfillan. Molecular regulation of mast cell activation. *Journal of Allergy and Clinical Immunology*, 117(6):1214–1225, 2006.

[101] Alec Rivers, Frédo Durand, and Takeo Igarashi. 3d modeling with silhouettes. *ACM Trans. Graph.*, 29(4):109:1–109:8, 2010.

[102] A. Peter Ruymgaart and Ron Elber. Revisiting molecular dynamics on a CPU/GPU system: Water kernel and SHAKE parallelization. *Journal of Chemical Theory and Computation*, 8(11):4624–4636, November 2012.

[103] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman Berendsen. Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *Journal of Computational Physics*, 23(3):327–341, March 1977.

[104] Sara Sandin, Lars-Göran Öfverstedt, Ann-Charlotte Wikström, Örjan Wrange, and Ulf Skoglund. Structure and flexibility of individual immunoglobulin g molecules in solution. *Structure*, 12(3):409 – 415, 2004.

[105] E. Sanz and D. Marenduzzo. Dynamic Monte Carlo versus Brownian dynamics: A comparison for self-diffusion and crystallization in colloidal fluids. *The Journal of Chemical Physics*, 132(19):194102, 2010.

[106] Erica Ollmann Saphire, Robyn L. Stanfield, M.D. Max Crispin, Paul W.H.I. Parren, Pauline M. Rudd, Raymond A. Dwek, Dennis R. Burton, and Ian A. Wilson. Contrasting igg structures reveal extreme asymmetry and flexibility. *Journal of Molecular Biology*, 319(1):9 – 18, 2002.

[107] Sjors H.W. Scheres. Relion: Implementation of a bayesian approach to cryo-em structure determination. *Journal of Structural Biology*, 180(3):519 – 530, 2012.

[108] Gunnar F. Schröder, Axel T. Brunger, and Michael Levitt. Combining efficient conformational sampling with a deformable elastic network model facilitates structure refinement at low resolution. *Structure*, 15(12):1630 – 1641, 2007.

[109] A. Shehu, L. E. Kavraki, and C. Clementi. Multiscale characterization of protein conformational ensembles. *Proteins: Structure, Function, and Bioinformatics*, 76(4):837–851, 2009.

[110] Jeremy Siek, Lie-Quan Lee, and Andrew Lumsdaine. *The Boost Graph Library: User Guide and Reference Manual*. Addison-Wesley, 2002.

[111] Dwaipayan Sil, Jong Bum Lee, Dan Luo, David Holowka, and Barbara Baird. Trivalent ligands with rigid DNA spacers reveal structural requirements for IgE receptor signaling in RBL mast cells. *ACS Chemical Biology*, 2(10):674–684, 2007.

[112] Amit P. Singh, Jean-Claude Latombe, and Douglas L. Brutlag. A motion planning approach to flexible ligand binding. In *Int. Conf. on Intelligent Systems for Molecular Biology (ISMB)*, pages 252–261, 1999.

[113] Adam M. Smith, Wen Xu, Yao Sun, James R. Faeder, and G. Elisabeta Marai. RuleBender: Integrated modeling, simulation and visualization for rule-based intracellular biochemistry. *BMC Bioinformatics*, 13(8), 2012.

[114] Murray Stewart. *Electron Microscopy of Biological Macromolecules*, pages 9–39. Springer US, Boston, MA, 1990.

[115] Xinyu Tang, Shawna Thomas, Lydia Tapia, David P. Giedroc, and Nancy M. Amato. Simulating RNA folding kinetics on approximated energy landscapes. *J. Mol. Biol.*, 381:1055–1067, 2008.

[116] Lydia Tapia, Shawna Thomas, and Nancy M. Amato. A motion planning approach to studying molecular motions. *Communications in Information and Systems*, 10(1):53–68, 2010.

[117] Shawna Thomas, Xinyu Tang, Lydia Tapia, and Nancy M. Amato. Simulating protein motions with rigidity analysis. *J. Comput. Biol.*, 14(6):839–855, 2007.

[118] Huimin Tong, Lei Zhang, Allan Kaspar, Matthew J Rames, Liqing Huang, Gary Woodnutt, and Gang Ren. Peptide-conjugation induced conformational changes in human igg1 observed by optimized negative-staining and individual-particle electron tomography. *Scientific Reports*, 3:1089:1–1089:9, 2013.

[119] Maya Topf and Andrej Sali. Combining electron microscopy and comparative protein structure modeling. *Current Opinion in Structural Biology*, 15(5):578 – 585, 2005.

[120] Vishwesh Venkatraman, Lee Sael, and Daisuke Kihara. Potential for protein surface shape analysis using spherical harmonics and 3d zernike descriptors. *Cell Biochemistry and Biophysics*, 54(1-3):23–32, 2009.

[121] Elizabeth Villa and Keren Lasker. Finding the right fit: chiseling structures out of cryo-electron microscopy maps. *Current Opinion in Structural Biology*, 25:118–125, 2014.

[122] F.G. Whitby and G.N. Phillips Jr. Crystal structure of tropomyosin at 7 angstrom resolution. *Proteins*, 38(1):49 – 59, 2000.

[123] Bridget S. Wilson, Janet M. Oliver, and Diane S. Lidke. Spatio-temporal signaling in mast cells. *Advances in Experimental Medicine and Biology*, 716:91–106, 2011.

[124] Jungdam Won and Jehee Lee. Shadow theatre: Discovering human motion from a sequence of silhouettes. *ACM Trans. Graph.*, 35(4):147:1–147:12, July 2016.

[125] Yinghao Wu, Jeremie Vendome, Lawrence Shapiro, Avinoam Ben-Shaul, and Barry Honig. Transforming binding affinities from three dimensions to two with application to cadherin clustering. *Nature*, 475:510–513, 2011.

[126] Keli Xu, Byron Goldstein, David Holowka, and Barbara Baird. Kinetics of multivalent antigen DNP-BSA binding to IgE-Fc$\epsilon$RI in relationship to the stimulated tyrosine phosphorylation of Fc$\epsilon$RI. *J. Immunol.*, 160(7):3225–3235, 1998.

[127] Jin Yang, Michael I. Monine, James R. Faeder, and William S. Hlavacek. Kinetic Monte Carlo method for rule-based modeling of biochemical networks. *Phys. Rev. E*, 78(3):031910, 2008.

[128] Jun Zhang, Karin Leiderman, Janet R. Pfeiffer, Bridget S. Wilson, Janet M. Oliver, and Stanly L. Steinberg. Characterizing the topography of membrane receptors and signaling molecules from spatial patterns obtained using nanometer-scale electron-dense probes and electron microscopy. *Micron*, 37(1):14–34, 2006.

[129] Lei Zhang and Gang Ren. Ipet and fetr: Experimental approach for studying molecular structure dynamics by cryo-electron tomography of a single-molecule structure. *PLOS ONE*, 7(1):1–19, 2012.

[130] Lin Zhang, Diannan Lu, and Zheng Liu. How native proteins aggregate in solution: A dynamic Monte Carlo simulation. *Biophys. Chem.*, 133:71–80, 2008.

[131] Xing Zhang, Lei Zhang, Huimin Tong, Bo Peng, Matthew J Rames, Shengli Zhang, and Gang Ren. 3d structural fluctuation of igg1 antibody revealed by individual particle electron tomography. *Scientific Reports*, 5:9803:1–9803:13, 2015.