Quantifying The Kinetics of Phase Separation *in silico* and *in vivo*.

Baljyot Parmar

Master of Science

Department of Biology

McGill University

Montreal,Quebec

July, 2019

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Master of Science

DEDICATION

This document is dedicated to my sister.

ACKNOWLEDGEMENTS

I thank my parents. Special thanks to all my mentors: Steph Weber, Rodrigo Reyes-Lamothe, Paul Francois and Anton Zilman. Special thanks to Anne-Marie Ladouceur for introducing me to *E. coli* and including me in her project. She created the strains, and protocols used in the methods of Chapter 2. The idea to incorporate single molecule imaging was initiated by her and Nicolas Soubry from the Reyes lab.

ABSTRACT

Classically cells compartmentalize their constituents into spatial regions bounded by membranes to modulate biochemical reactions. Recently, it has been shown in both eukaryotic and prokaryotic systems that this compartmentalization can occur in lieu of membranes. The theory of phase transitions, the changing of states in a system, has been used to explain this process, but the mechanistic details are not well understood. Namely, the role of certain intrinsically disordered proteins and the properties which allow them to facilitate these transitions, and the material properties they impart on these structures remains unclear. In this thesis we outline two distinct methods by which the role of these proteins can be deciphered. Firstly, we introduce a simple Monte Carlo simulator for a collection of protein systems that can vary in charge profile to understand the propensity of different charge profiles to phase separate in solution. Secondly, we use single molecule imaging to track individual molecules associated with these phase separated states to determine the difference in kinetics relative to other regions in the cell.

ABRÉGÉ

Les cellules classiquement classent leurs constituants dans des régions spatiales délimitées par des membranes pour moduler les réactions biochimiques. Récemment, il a été démontré dans les systémes eucaryotes et procaryotes que cette compartimentation peut se produire á la place des membranes. La théorie des transitions de phase, le changement d'états dans un systme, a été utilisée pour expliquer ce processus, mais les détails mécanistes ne sont pas bien compris. Á savoir le rle de certaines protéines intrinséquement désordonnes et les propriétés qui leur permettent de faciliter ces transitions, ainsi que les propriétés matrielles qu'elles conférent Á ces structures. Dans ce manuscrit, nous décrivons deux méthodes distinctes par lesquelles le rôle de ces protéines peut tre déchiffré. Tout d'abord, nous prsentons un simulateur de Monte Carlo simple pour une collection de systémes protéiques dont le profil de charge peut varier afin de comprendre la propension de différents profils de charge á se séparer en phase dans une solution. Deuxiémement, nous utilisons l?imagerie de molécule unique pour suivre les molécules individuelles associées á ces états séparés de phase afin de déterminer la différence de cinétique par rapport aux autres régions de la cellule.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

1.1 Summary

Membrane bound organelles in eukaryotic systems are involved in the segregation of cellular components and chemical processes which allows the cell to create unique compartments that might not be possible without this level of spatial control. Recently, the occurrence of membrane-less organelles across multiple organisms and in different regions of the cell has prompted questions into the formation and regulation of these compartments [1–6]. These are compartments with a distinct chemical composition with respects to the rest of the cell in which it resides. Examples of such bodies can be found in the nucleoli, P-bodies or stress granules [7]. It was found that many of these membrane-less bodies have liquid properties, namely, they exchange materials with the surrounding, fuse, have a spherical shape and are deformed by external flow [7, 8]. This has led many to assert that these liquid droplets exist as phase separated regions in a liquid environment. The classical example of phase separation is of salad dressing, which is a 2 phase system of oil and water that will de-mix if given sufficient time, hence increasing its phases from 1 to 2. These phase separated droplets occur as a result of the enthalpic contributions to the free energy

of the system exceeding the entropic contributions [1] , which favour mixing. [8, 9] It should be noted that not all of these ordered states are liquid, many exist along a continuum from liquid to solid-like. This is entirely controlled by the enthalpic term in the free energy, which implies interactions between components.

Although the physical concepts of phase transitions have been suggested as a mechanism for the formation of these compartments, there is not sufficient literature to describe the details. In particular, what molecular interactions are responsible for the formation of these bodies and their respective material properties? Also, once these bodies are created, how are they distinguishable kinetically from other regions of the cell? This manuscript aims to answer these two question through *in silico* and *in vivo* methods, namely Monte Carlo simulations and live-cell single molecule experiments.

1.2 Introduction to Phase Separation

Many papers published on the application of phase transitions in biology focus on one subset of these transitions: liquid-liquid. Before we delve into the examples of these transitions and their properties it is worthwhile to define a phase and how this relates to the biological systems we want to investigate. A phase is a region of space in which the properties of that region are shared within the region's confines and differs from other such regions. In this way, a lone iceberg floating on an endless ocean of water is a separate phase from the ocean. This logic is applied to the regions of the cell, in which the nucleoid of prokaryotes can be considered a distinct phase

[1] Responsible for mixing.

compared to the rest of the cell. Just in this distinction we learn almost nothing, however, this changes when one asks about the properties of these phases. How do they behave under various (physiological and not) conditions? How do they form, and what are the essential components required for the assembly?

Of interest to this manuscript are three distinct phases: liquids, solids, and gels [2] . We define a liquid phase as one that is malleable to its surroundings and exhibits lower molecular organization than the other two. Intuitively we can think to the properties of water at room temperature to understand the nature of liquids in various conditions. If poured into a container liquids orient themselves to take the shape of the confine and retain no memory of their previous shape. This property is on a continuum, in which different liquids pour at different rates given the same impulse, where water has a higher rate while jam or honey respond at lower rates. A key property of liquids is that at the boundary of the phase, the net attractive force felt by the particles is inward (toward the centre of the phase). This creates a tendency for the particles at the boundary to want to move toward the centre, and in doing so they reduce the surface area until the attractive forces balance out at the boundary, giving the phase a spherical shape. [7] Solids are quite the opposite, in that the particles within are oriented in a rigid lattice-like structure. This gives them the properties we normally associate with solid objects, like: solid metals, your laptop, etc. As opposed to liquids, solids tend to retain their previous shape under higher stresses. There is a continuum here of course, but in general, unless an extreme force

[2] Since these are observed in various biological systems

is applied to a solid, it will retain its shape. Unlike liquids, solids can take any sort of boundary configuration, since they are lattice bound. It is best to think of building legos, and all the amorphous shapes which are possible with them. Also, since the particles are bound in a lattice, they have almost no movement [3] , and as such cannot mix with other solids in a way liquids can. [7] Lastly, gels are highly interconnected systems of polymer chains in which a single polymer chain interacts with multiple other such chains. Although the interactions between the polymer chains is not as strong as the lattice-like structures found in solids, the sheer amount of these weaker interactions amount to create a solid-like structure. Glass is an example of a highly jammed gel-like system. [7] Although it seems to be solid, a vertical glass panel is slowly flowing downward due to gravity, but this process is extremely slow and will not affect all the fancy building for a long, long time. Nonetheless, it is important to understand the distinction between solids and gels, namely, the nature of the interactions which give them their characteristic properties. One can think of gels as mutable solids, in which the experimenter/cell can modulate the nature of the interactions to achieve solid-like or liquid-like properties as desired. It should be noted that within these definitions is a continuum of characteristics; viscosity of liquids and the tensile strength of solids can vary drastically.

Having said this, the process of phase separation occurs when a single phase decomposes into two or more distinct phases. The classic example, used in almost

[3] Even at absolute zero the subatomic particles in the lattice have movement, due to Heisenberg's uncertainty principle, but this is a digression.

every paper on the topic, is oil and water mixing and separation in vinaigrettes. [7] If left to time, a well mixed vinaigrette solution will eventually separate into two or more distinct phases [4] . In our previous example of the iceberg and the ocean, we can imagine that system formed by a large iceberg melting some of its mass into water, or vice versa. This process would not be considered phase separation but a transition; the state of the system is just transitioning into another of its phases even though there are two phases in the system in the end. In the case of pure water, the changing of ice to water is a transition because there is only one chemical species in the system (water), while systems of oil and water allow for the separation of these chemical species to occur. This distinction is important when looking at biological examples, both in this manuscript and beyond, to accurately understand the assembly and dynamics of these systems. The change from a single solution to the decomposed phase occurs as a result of multiple factors, but the simplest variable to consider is the concentration of the components. In a supersaturated state the system can separate into two distinct phases: a dilute and dense phase, with respect to the components. This simple process has been known to chemists and the physics communities for a long time, only recently have groups started to observe this behaviour in the living cell. [10–13]

[4] Depending on what you initially put in it. This will follow Gibbs phase rule in the very long time scales.

1.3 Membrane-less Organelles in Cells

Early evidence for the liquid-like properties of membrane-less organelles comes from the seminal paper by the Hyman lab in 2009, which found that P granules [5] behaved as liquids. [14] The paper discusses multiple properties for the liquid nature of the P granules, [6] namely: they are spherical, can fuse and show evidence for free diffusion of their components. They further proposed that this system was a consequence of liquid-liquid phase separation at work in the cell. [14] They argue that it is through this mechanism that the asymmetry in the embryonic development is achieved. [14] As opposed to the simple liquids discussed before, this liquid system is comprised of multiple species of amorphous polymers which interact both with short range van der Waals and long range electrostatic interactions. [15]

In the years following this paper, multiple groups have identified other occurrences of membrane-less organelles occurring across multiple cellular and life domains. Akin to the the P granules in germ cells, P bodies exist as a phase separated system associated with mRNA turnover in the cytoplasm of various types of cells. [16] It has been shown that along with proteins involved in mRNA turnover, the formation of P bodies depends critically on the availability of RNA, suggesting that it plays a stabilizing role in the formation of the condensates. [16, 17] RNA acting as

[5] Bodies rich in RNA and RNA-binding proteins in the *Caenorhabditis elegans* embryo.

[6] Which are now used in many papers to support their own systems as a phase separated one.

a stabilizing variable in the formation of membrane-less droplets is a reoccurring theme, which hints at the molecular underpinning of these systems (discussed later).

Similarly, stress granules are liquid-liquid phase separated bodies in the cytoplasm of many cell types, enriched in mRNA and proteins associated with mRNA translation. These bodies form dynamically under signals of stress to the cell and dissociate as quickly under the loss of the stress signal. [18] They display the same fusing and wetting properties of the P granules observed in the *Caenorhabditis elegans* embryo while also being sensitive to the concentration of RNA in the cell. [17,18] Kroschwald et al. found that yeast and mammalian cells differ in the material properties of stress granules by treatment with 1-6 hexanediol [7] . [19] While their mammalian counterparts show a liquid-like granule, the yeast ones are more solid, but the underlying interactions of both bodies are the same. [19] These interactions are not prion-like protein aggregates, rather the material properties are conferred through promiscuous interactions between prion-like proteins and RNAs in the system; this allows the cell to react in a modular way to counteract various levels of stress. [19] A functional connection between P bodies and stress granules has been proposed based on the observation that the latter fails to form without the former. [17]

The ability of the cell to sense small changes in stress signals is not fully understood. An interesting case involving poly(A)-binding protein, shows that this highly conserved RNA-binding protein phase separates in response to high temperatures, in yeast and humans. [20] RNA-binding domains and modular hydrophobic

[7] A drug associated in disrupting weak interactions between polymer species.

domains along the protein are critical to the formation of stress granules and the phase separating ability of this protein, both *in vivo* and *in vitro*. [20] Riback et al. further showed that the assembly of the poly(A)-binding protein is modulated by the temperature and pH the cell experiences and forms dynamically at the onset of stress. [20] Without definitively pinpointing the function, the group showed that stopping poly(A)-binding protein mediated phase separation results in atypical growth patterns, suggesting a role in translation. [20, 21]

The nucleolus is a tri-compartmentalized structure [8] in eukaryotic cells that functions as the centre for ribosome biogenesis, although the structure changes during different cell cycle stages. [22] The distinct compartments are shown to be liquid-like and it has been proposed that they form through liquid-liquid phase separation. [22] The different compartments act as an assembly line for rRNA processing and its movement through the regions is hypothesized to occur by enzymatic modifications. [17] Feric at al. show *in vitro* that the two proteins, FIB1 and NPM1 [9] phase separate into the distinct ringed structure observed *in vivo*, suggesting that multiphase systems can easily be created by simple polymer species. [23]

Another RNA, protein rich body in the nucleus is the paraspeckle, which consists of core paraspeckle proteins (PSF, NONO and PSPC1) along with long non-coding RNA *NEAT1*. [17] They are generally found in the interchromatin space in

[8] Consisting of the fibrillar centre (FC), dense fibrillar centre (DFC) and the granular component (GC)

[9] Which localize to DFC and GC regions of the nucleolus respectively.

the nucleus, which is distinct to the nuclear speckles (discussed later). The body is associated with RNA transcription through the activity of RNA Pol II. [17] If the RNA Pol II activity is inhibited the bodies fuse with the nucleolus but does not fully integrate into the nucleolar matrix, suggesting different material properties of these two phases. [17] Nuclear speckles localize adjacent to the interchromatin space and are enriched in mRNA splicing factors and serine/arginine-rich proteins. Like most of these nuclear bodies the exact function of the nuclear speckles is not known, but the occurrence of the RNA splicing factors suggests their role in regulating gene expression. [17]

It is clear that phase separated systems are ubiquitous in various eukaryotic cells and different cellular regions, even if a clear function cannot be associated to them. The size and structure of these systems is varied but there is a reoccurring theme of RNA and RNA binding proteins that are enriched and many times essential to the formation and stability of these systems. Two questions arise, what are the molecular underpinnings of these systems, and do all forms of life use phase separation? The latter is the motivation to look at RNA Polymerase clusters in *E. coli*, in Chapter 3, while the former is the motivation for the molecular simulations in Chapter 2. The following section attempts to consolidate the knowledge on the nature of the molecular composition of these systems and their significance in the polymer physics of phase separation.

1.4 Molecular Players of Phase Separation in the Cell

Many studies on the protein constituents of phase separated systems have shown that their formation is dependent on the occurrence of multiple interacting domains

along a protein chain. [15] As stated in Boeynaems et al. there are three distinct ways this sort of multivalency can occur: the interaction between folded proteins and from oligomeric systems, "sticky" folded domains which interact 1-1 with other such domains and are linked by mobile linkers, and intrinsically disordered regions (IDRs) which can facilitate promiscuous short ranged weak interactions between and within a polymer. [15] The extent to which the system uses each of these mechanism and the respective interaction strengths can play a role in modulating the material properties of the phase separated system.

To study the role of this multivalency, Rosen and colleagues looked at *in vitro* multi-domain polymers of PRM and SH3 connected with flexible linkers in varying domain concentration. [24] They show that the valency of the polymer chains play a role in driving phase separation. [24] Using the natural nephrin-NCK-N-WASP system [10] as a three-domain extension of the SH3-PRM system, they showed a similar phase profile to the *in vitro* system. [17, 24] As a functional element, the phase separated N-WASP allows the nucleation of actin filaments through their Arp2/3 complex. [17, 24] This example is rich in both the effectors underlying phase separation while also displaying a clear functional role for the system. The critical role for the phosphorylation of the tyrosine sites on nephrin, enabling the interactions to occur, also gives a good introduction to the importance of post-translational modifications on the dynamics of phase separated systems.

[10] Nephrin contains 3 tyrosine phosphorylation sites which bind the SH2 domains on NCK, which in turn has 3 SH3 domains that bind the 6 PRM domains in N-WASP.

1.4.1 Intrinsically Disordered Proteins

IDRs as mentioned before are these disordered domains, namely without a strict 3D structure, that have been shown throughout the literature to be dominant in facilitating phase separation through *in vitro* studies. [15, 25] The heterogeneity in regards to the conformations of these proteins allows them to partake in multiple dynamic interactions between other such proteins or with structured domains. [25] Many times an interplay between folded domain interactions and those facilitated through IDRs can lower the barrier to phase separation. The nephrin system best represents this interplay with the folded SH3, PRM domains linked with regions of amino acids that are disordered, allowing the distances between the domains to be irregular. Molecular simulations of PRM-SH3 interacting polymers show that the length and rigidity of these linker regions modulates the critical concentration needed for a sol-gel transition to occur. [26] A consequence of this heterogeneity in conformation is that the structures formed through these interacting regions can have varying properties. As with the SH3-PRM system, Harmon et al. show via simulations that at a critical concentration and interaction strength the system can undergo liquid-liquid phase separation and then gelation. [25, 26] Many times the structured domains form a rigid backbone for the phase separated structure and other weakly interacting species are recruited to the region but have no role in facilitating the formation [11] of these bodies. This sort of model for phase separation is referred

[11] This is a generalization since there exists many permutations on how these phase separated structures form and are stabilized with respect to the constituents.

to as the scaffold-client model, where one or more species are responsible for actually creating a scaffold phase and the others are recruited through their ability to interact with parts of the scaffold. [15, 25, 26]

The promiscuous nature alone, of IDR interactions can lead to liquid-like phases and even gels. The nuclear pore complex (NPC) provides a great example on the ability of these weak multivalent interactions to create gel-like systems in the cell. [9, 27] The overall structure of the NPC exists as a dense gel-like phase, acting as a diffusion barrier for large molecules but the core structure is composed from nucleoporins (Nups) that contain disordered FG motifs. Deletions of FG domains in *S. cerevisiae* resulted in a significant loss in the barrier of NPC, along with the fact that purified FG domains alone were able to create NPC-like systems *in vitro* suggest that these disordered FG motifs are essential. [27]

Apart from lacking a distinct stable configuration these intrinsically disordered proteins (IDP) tend to include multiple low complexity domains (LCD), which as the name suggests are rich only in a select few amino acids. [9,15] These amino acids with uncharged polar side chains, aromatic rings, and charged properties cluster into short linear interaction motifs (SLiMs) which form the interaction basis for these systems. [7, 9, 28] There is further organization within the repeats of the SLiMs which combined with the charged nature of the residues involved creates unique charge patterns that are important in the protein's phase separation. [9, 28]

In vertebrate oocytes the protein Xvelo (Velo1) localizes to Balbiani bodies and with its prion-like domains (PLD) with disordered linkers, self-assembles into an amyloid-like matrix with the ability to bind RNA. [15, 25, 29] PLDs are an example

of these LCDs which are enriched in polar, uncharged amino acids [12] and interact to create amyloid-like fibres in these phase separated systems. [15, 25] Xvelo exists in a soluble form in mature oocytes where the PLD does not interact with any other proteins, presenting a curious case in which amyloid structures can be reversible *in vivo*. [29] Mutations in the stress granule proteins of the hnRNPA family showed the dependence of both a prion-like domain and a RNA binding region for successful phase separation. [9, 15, 25] Similar studies with FUS [13] , an ALS associated protein thought to have an impact on transcription, shows its ability to phase separate into liquid, gel and solid prion-like states through its IDRs, prion-like and RNA binding domains. [15, 30, 31] These examples show the interplay between IDRs, PLDs and RNA binding regions in allowing these systems to confer various material states. Furthermore, deletions of FUS' PLDs and RNA binding regions show a loss of ability in forming phase separated systems, suggesting that the promiscuous interactions these domains confer is critical to the formation of these systems. [9, 32]

The abundance of charge patterns that arise in these LCDs confer different electrostatic interactions which can range from π-π, cation-π to charge-charge. [9, 15, 25, 32, 33] Delocalized π electrons on the side chains of certain aromatic and charged residues interact with each other to create stacking interactions which when realized in large numbers form a critical role in liquid-liquid phase separation. [9, 33, 34] Similar interactions occur between these delocalized π electrons and charged side

[12] Namely Q/N-rich LCDs.

[13] A disordered protein involved in stress granule phase separation

chains of certain amino acids, which work in a similar way to facilitate the multivalent interactions in phase separation. [9,32,33] Disrupting key residues in PLDs and these LCDs is enough to lose these interactions and disrupt the phase separating behaviours of FUS and the Nups discussed before. [9, 15, 25, 32–34] Mutations of F to aromatic residues of Nsp1p [14] , but not charged residues, preserved the formation of the gel-like structure of the nuclear pore complex *in vitro*, consistent with other point mutation studies. [15, 33, 34] Apart from the specific residue changes, the position at which they are found along the chain also impart an influence on the phase separation of these proteins. [9, 15]

These concepts of equilibrium phase separation apply to closed systems, but as we know the cell is inherently dynamic and out of equilibrium, which requires it to respond to environmental and cellular cues at various time scales. The importance of charge interactions in the multivalent nature of these IDPs has been established, hence an ideal target for the cell to change its phase separating systems is by modulating the charge patterns. It has been extensively shown that through the use of post-translational modifications the cell is able to respond to the dynamic changes in the environment. [9, 15, 32] Phosphorylation, methylation and SUMOylation (addition of small ubiquitin-like modifiers) are some of the ways in which the cell has been shown to modulate the binding dynamics of the constituent proteins by changing the charge profiles. [9] The nephrin system discussed before shows a simple example of the importance of tyrosine phosphorylation events in facilitating the phase separation

[14] A nucleoporin in the formation of the nuclear pore complex.

of nephrin and nucleation of actin filaments. [24] The three phosphorylation sites on nephrin bind the SH2 domains on NCK which then recruits other proteins through its SH3 domains, effectively building the multivalent nature of the phase separation. [24] FUS is similar, in which the phosphorylation of serines reduced the gel-like nature of *in vitro* FUS droplets and arginine methylation reduced the potential for it to phase separate. [9,31] The targets of these post translational modifications are usually found in the LCDs of these phase separating proteins. At the sequence-level these regions are not well conserved, however, their length and charge composition is. [35,36] This leads to strong functional conservation of these disordered regions in regard to their phase separation. [35,36] The fact that many of the kinases required for these modifications to occur localize to the phase separated systems they control hints at the ubiquitous nature of these modifications to modulate phase separating properties. [15]

1.5 Phase Separation and Disease

The fact that many of the proteins discussed above are associated with neurodegenerative diseases and tend to aggregate in various species including humans, shows a need for a deeper understanding of the nature of these systems. [9] α-synuclein in Parkinson's, β-amyloid in Alzheimer's and RNA binding proteins associated in ALS and FTD show some of the overlap between these types of diseases and aggregation proteins. [9] Stress granule associated proteins like FUS, hnRNPA1, and TDP-43 have been observed to go through a phase transition, similar to the aggregates found in ALS and FTD, through time in equilibrium conditions. [9,30,37] Mutation studies of FUS and TDP-43 show that ALS-like aggregation can be accelerated through the

mutation of key residues in the prion-like domains of these proteins, indicating that only under certain changes introduced within the cell do these systems form, and are kept in a non-aggregate form in other cases. [9, 30, 37] Hence, the appearance of the aggregates in these diseases can be seen as the cell's loss of ability in controlling the aggregating nature of these proteins. Although no direct role between phase separated systems like the stress granule and neurodegenerative diseases has been established, the similarities in the underlying mechanisms may provide further insight into the dynamics of both.

CHAPTER 2
Creating A Non-Lattice Monte Carlo Simulator For Polymer Chains

2.1 Introduction

Macromolecules form the fundamental basis for much of biology. Polymers are a subset of these macromolecules which exhibit a large variation in properties and functions in cell biology. Two subclasses of polymers, genetic polymers (DNA, RNA) and proteins have been extensively studied and will serve as a model for the proposed simulations below [28].

Recently, it has been documented that this subclass of macromolecules is responsible for the formation of membrane-less organelles *in vivo* under certain conditions [1–6]. Namely, the cell can regulate the formation of these compartments according to its needs or in response to extracellular cues. Although a critical review is found below, it is sufficient to state that there is a consensus on the mechanism by which these polymers form into organelles, namely by phase separation [4, 7, 22, 28, 38, 39]. This concept is well understood in many physical systems, namely in the idealized polymer-solute models such as those introduced by Paul Flory and Maurice Huggins [11–13, 40, 41]. The Flory-Huggins theory makes many simplifying assumptions on the nature of interaction between polymer subunits and the solvent, while disregarding stochasticity (thermal noise derived), but still displays the general behaviour of phase separating systems [11–13, 41]. Recently, there

have been attempts to account for these simplifying assumptions but there is insufficient literature to simulate these polymer systems computationally using only basic assumptions of charge-charge interactions. One of the aims of this work is to bridge this gap in the literature and determine if these simulations are able to describe experimental results and also predict the propensity of untested sequences to phase separate.

2.2 Background

2.2.1 Identification of Membrane-less Organelles in Biology

It has been suggested that proteins which are observed to undergo phase separation are generally intrinsically disordered, allowing them to take varied configurations in space [2,28]. It is this property along with distinct charged domains in the sequence that reportedly allow the proteins to aggregate into these liquid like droplets [2, 28]. The nature of these interactions have been debated, but many groups have shown that weak interactions pertaining to $\pi - \pi$, cation-π, and charged interactions play a role in stabilizing these ordered droplets. [9, 42] It should be noted that interaction domains for RNA and DNA along these proteins stabilize the formation of these systems; this is different from prion-like aggregations which tends to create solid-like systems through oligomerization. [43].

It has been shown that point mutations at key regions or across the whole sequence can also affect the material properties of these bodies by interfering with charge-charge interactions [8,42]. These changes to the charge profile of the sequence can determine the nature of the phase transition and adopted state. For example, rearranging the charge pattern of the N terminus disordered tail of Ddx4, a DEAD-box RNA helicase, disrupts its ability to natively form these liquid droplets. [42] These weak interactions are universal to many intrinsically disordered proteins and as such provide a unified framework with which to study the properties of phase separation in biology.

Berry et al. showed that nuclear bodies enriched in FIB-1, a nucleolar localizing protein in *C. elegans* embryos, are formed via the classical mechanisms of phase separation and stabilized by RNA transcription. [3] FIB-1 is an intrinsically disordered protein with a blocky charge distribution akin to the N-terminal tail of Ddx4, further supporting the universality of the charge-charge interactions playing an important role in phase transitions. [3] Surprisingly, another nucleolar protein, DAO-5, behaves exactly like FIB-1 but has a rapidly alternating charge profile. [3] A charge profile of Fib-1 and Dao-5 is shown in Figure 2-1. It is not known if the alternating charge profile is a detriment for phase separation or another mechanism rescues the activity in this case. This chapter attempts to ask: is the charge distribution enough to explain the behaviour of these proteins *in vivo*?

2.2.2 Mean Field Approaches in Polymer Physics

Phase transitions have been studied in classical physics for a long time. Only recently biologists have begun to apply the theories to systems in the cell. Early work in the 1950s by Paul Flory and Maurice Huggins (separately) introduced lattice approximations of polymer-polymer interactions which showed phase separation of the mixed polymer-solvent solution into high and low density polymer regions depending on an interaction parameter χ, which took into account the interaction energies between neighbouring lattice sites [2, 11–13, 22].

The following will be a brief introduction to the Flory-Huggins (FH) model and Landau theory as it relates to the topic at hand. [1]

The objective of FH is to create an expression for the free energy of a polymer system and minimize it. To create this formulation, imagine lattice sites on a 2D grid [2] Each site on the lattice can be occupied by a monomer of a polymer or by a solvent (let's say water). [11] If there are N monomers in each polymer and a total of n_p polymers with total lattice sites Ω, then the number of solute molecules is $n_s = \Omega - n_p N$. [11] To simplify matters one can reduce these to a single variable $\phi = \frac{n_p N}{\Omega}$ as the volume fraction of the polymers in this lattice system. [11] The monomers can interact with other monomers, nearest neighbors, and also the solute with a certain energy. [11] The total interaction energy of the system then can be written as:

$$U_i = \epsilon_{pp} N_i^{pp} + \epsilon_{ss} N_i^{ss} + \epsilon_{sp} N_i^{sp}$$

[11] where U_i is the energy of the ith configuration, and the ϵ are the interaction energies of pairs of solutes (ss), monomers (pp) and solute-monomers (sp), with the number of those pairs occurring in the system. [11]

The partition function for the system can be written as:

---

[1] For the avid reader, the seminal book on polymer physics: Introduction to Polymer Physics by Masao Doi is highly recommended to expand the current discussion, and much of the formulations is identical to the book. [11]

[2] 2D is used for simplicity.

$$Z = \sum_i e^{-\frac{U_i}{k_b T}}$$

where i runs through all possible configurations of the system, T is the temperature, and k_b is Boltzmann's constant. [11] For most systems there is no simple analytical form but under the simple assumption that each interaction reduces to the average interaction in the system, the expression can be simplified. This assumption is known as mean field theory. [11] Effectively this reduces the partition function from a complicated sum over configurations to the average configuration.

$$Z = e^{-\frac{\langle U \rangle}{k_b T}}$$

[11] The free energy then can be solved for using the entropy relation

$$F = -k_b T log Z$$

which is left to the reader to follow using [11]. The final form of the mean field free energy is as follows:

$$F_{mean_field} = \Omega k_b T f_{mean_field}(\phi)$$

$$f_{mean_field} = \frac{1}{N}\phi ln(\phi) + (1-\phi) ln(1-\phi) + \chi\phi(1-\phi)$$

$$\chi = \frac{z}{2k_b T}(\epsilon_{pp} + \epsilon_{ss} - 2\epsilon ps)$$

where z is the lattice coordination number, indicating the number of interactions a point on the lattice can have. [11]

Without delving further into the algebra, the key point in this formulation is that the fluctuations inherent in the energy calculations are not included and only the average interactions on each unit in the lattice is considered.

This mean field approximation is also seen in Landau theory, which is a study of the order parameter (measure of the degree of order) assuming the free energy can be analytically expressed as a function of said order parameter [10–13].

Landau made an observation that different physical systems can exhibit similar properties at a critical point. The classical example is of the phase transition of water from liquid to gaseous. [11] Both of these systems are homogenous, which implies that at the critical point of transition between the two states this property is continuous. The order parameter is a quantity used to identify order in a system, and its precise expression depends on the system. [11] For example in the FH formulation above, the binary mixture order parameter is the volume fraction of polymer units. [11]

At the critical point there are large correlations in space, namely, two distant points can become highly correlated and the system becomes fractal in nature. [11] This implies that the properties of the system are not dependent on the microscopic details which usually complicate these analytical theories, rather it is the behaviour of global variables and their scaling. It turns out that these global variables and their scaling are shared across many systems. [11] Landau theory clumps these systems into universality classes which share the same scaling, usually in a power law manner. [11]

Using the above fact and the symmetry of the free energy of the FH model with respect to the volume fraction one can propose an analytical form for the free energy of the system at a critical point as:

$$f(\phi) = f_0 + f_2\phi^2 + f_4\phi^4$$

where f_4 is a positive variable, but f_2 can change signs. [11] It might seem that this equation comes out of thin air, but it is actually educated guesswork, for a simple system like this. It turns out that one can recover the same properties of the FH formulation from this simple guessed system.

Although the theory is only valid for temperatures near the critical points, it is still useful to extract critical exponents which are universal quantities independent of the physical system. These exponents are useful not only to find dynamic scaling at different temperature regimes of a system but also to validate with experimental data [10–13]. It should be noted that most mean field critical exponents calculations in dimensions above $d = 4$ are valid while for lower dimensions the exponents do not match experimental data. [10, 11, 13]

Both the FH and Landau formulations try to reduce the interactions in the system to simplify the mathematics but overlook information regarding the microscopic fluctuations. This assumption is almost never true in biological systems [3] , where a study of the total interaction terms needs to be made. However, the problem with most full field theories, which take fluctuations and microscopic configurations into account, is that there is no analytical solution to the free energies of these systems.

[3] Because species can interact in a variety of ways with other species and themselves.

Rather, computational and numerical methods need to be implemented to approximate the free energy landscape. These computations can become impossible to run at large enough scales and suffer from fine tuning by the researcher. In this chapter we choose to explore the partition function with a full field approach because we want to capture the dynamics and intermediate steps of these transitions. Our approach will allow us to explore this system through time and also away from equilibrium, which cannot be done for the previous formulations, giving us a full picture of these systems.

2.2.3 Efforts to Predict Phase Separation Propensity of Proteins

Using the thermodynamic formulation of the Flory-Huggins theory it is easy to identify the occurrence of these transitions. The downside of these theories is that they do not take charge distribution into account when considering the propensity of sequences to phase separate such that a blocky or alternating pattern with the same overall charge would have the same predicted phase behaviour. In this regard many groups have sought ways to incorporate the sequence in their simulations. The Pappu group utilizes a lattice model which simulates two types of polymers chains each containing a series of interacting domains that complement those on the other chain [26]. These modules interact 1:1 with the opposite domains while being separated by intrinsically disorder linear regions. Their simulation found that depending on the propensity of the linkers to contract or extend specific type of transition from disordered to order changed [26]. Namely, the system would undergo either a liquid-liquid phase separation and gelation or gelation without phase separation depending on the distance between the interacting domains.

Recently, Chan's group at the University of Toronto was able to simulate polymers in a lattice system based on actual sequence data, with each unit in the simulation taking the average charge of the respective amino acid [44]. They tested the propensity of multiple sequences of the same length but with varying charge distributions and found that those displaying blocked charges were able to show evidence of phase separation or aggregation at a lower simulation temperature than those without [44]. Namely, the phase boundaries (between ordered and disordered) are determined by the blockiness of the charge in the primary sequences. This leads directly to the idea that the cell can regulate the formation of these bodies by modifications to the primary sequence of these proteins which allows reversibility of this phase transition [44]. Although this approach is in the right direction to explicitly model sequences of charge along a protein, these models are on lattice simulations. By this virtue, the configurations in space the polymers can take is necessarily finite, which does not model the folding nature of these proteins *in vivo*. In this regard a new simulator needs to be created which is non-lattice and also uses a robust metric for classifying local densities of condensed materials.

2.3 Methods and Simulator Design

2.3.1 Creating a Simple and Extendable Model for Polymer Chains

To create this simulator one needs to decide on the way objects interact within the system and how these objects relate to the "real" system in question. In order to ensure that the simulator does not mimic a pipeline for the question at hand, namely, we recover what we put in, the simulator needs to be based on simple interactions. It also must be computationally simple due to the scale of the system we are trying to simulate, but not trivial. As such, this is not a pure molecular dynamics simulation, in which objects carry acceleration and velocities, rather it should explore the entire state space randomly and determine which state is most favourable.

Before this is defined, a coarse grain for the system at hand is required. It is currently unfeasible to simulate a sufficiently sized system of actual proteins or polymers that can be run at large timescales. To combat this curse of scalability and to also keep the simulator simple and easy to adapt to various systems, a different scale must be used. This simulator is essentially a chain and ball scheme, in which each ball is a point charge that can be connected to another point charge through a distance oriented bond. The nature of this bond will be discussed later, but the general idea of the system is to create a collection of these point charges that can be modelled as explicit amino acids in a protein or a clump along a protein sequence. This sort of abstraction allows the user to model any sort of behaviour in a polymer system, for example, a single point charge can be modelled as an interaction domain along an intrinsically disordered protein. Figure 2-2 (top right) shows the proposed model for a base object in the simulation, from which all other objects of higher order

are created. This base object (monomer) is allowed to have 2 properties: position in 3D (x, y, z) and charge, which can take on any real number value.

To define a favourable state, the creation of a cost function is needed. In the context of charged connected monomers the easiest function is the Coulomb's interaction between each monomer species. This function:

$$E_{Coul} = k\frac{q_1 q_2}{r}$$

models the energy of two point charges, q_1, q_2, at a distance r apart. [4]

Since there is no explicit bonds between the point charges the Coulomb's interaction will cause opposite charges to attract until they are superimposed atop one another. To counteract this force, a Lennard-Jones potential is added to the function.

$$E_{LJ} = 4\epsilon\left(\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right).$$

This sort of potential gives a repulsion term when interacting monomers get too close to one another, the extent of this can be modulated by σ term which models the distance r at which the interaction energy is 0. The attraction term is modelling the long range interactions between particles due to van der Waals forces.

Now, another term is needed to implicitly model the behaviour of the intramolecular bonds, namely, between monomers belonging to the same polymer. Classically

[4] It should be noted that the simulator contains periodic boundary conditions and the distances r or displacements x discussed will be according to this fact.

this is modelled using a hookian spring potential in which the energy in the bond is proportional to the square of the net displacement from its equilibrium position x.

$$E_{Bond} = \frac{1}{2} k_{spring} x^2$$

As such the final Hamiltonian of the system is a combination of these potentials.

$$H = E_{Coul} + E_{LJ} + E_{Bond}$$

It should be noted that the user can define any Hamiltonian desired, in accordance to how the objects in the simulation interact. The basic principle of the simulations is to take a sequence of connected point charges and determine the behaviour in a solution of other point charge chains. We decided to approach this with a simple Monte Carlo scheme using a modified Metropolis Hastings algorithm (MCMH). The MCMH scheme is shown as a flow chart in figure 2-3.

An in-depth mathematical explanation of the scheme is not shown, for brevity, but can be readily accessed.[5] The simulator creates a random initial state for the user-selected system and randomly proposes a move be made. If this move lowers the energy of the cost function defined above, the move is accepted. If the energy is not lower, then the Boltzmann probability ratio between the initial and proposed final state is calculated (to satisfy detailed balance).

$$P = e^{-\frac{(H_{Proposed} - H_{Initial})}{k_b T}}$$

[5] Please see the seminal paper from Wilfred Hastings. [45]

Where k_b is the Boltzmann constant and T is the temperature of the system. A uniform random variable on [0,1] is compared to this probability to accept or reject the proposed move. This scheme is repeated numerous times where each iteration can be thought of as the base time scale unit. A move in this context is actually numerous moves along each polymer which are tested as a unit. To be explicit, the model can choose to make one monomer move in a single Monte Carlo step, but it chooses to sample multiple monomer movements in one Monte Carlo move. This is done mainly to conserve computational time until convergence to equilibrium conditions. It should be noted that this is, to our best knowledge, the only non-lattice simulator of this kind, as objects can take any position in 3D space that is not occupied already.

To put it simply, the simulator is trying to map the defined Hamiltonian by randomly selecting states from the pool of possible states. The scheme explores this state space and slowly recreates it trying to find the minimum, or at least the local region where the energy is at a minimum.

This scheme is implemented in C++ in which all data structures are self defined. A compressed code base for the simulator is included in our Github and can be used for the avid reader to recreate the simulations in this manuscript. [6]

[6] Due to the random nature of the initial configuration and moves, it is actually improbable that the simulated system will be exactly the same, however, the general properties should be recovered

2.3.2 Tracking Clusters of Polymers Along the Simulation

Since the purpose of the simulation is to observe the behaviour of these systems of polymers through time, there needs to be a metric with which the user can extract the global properties of the system. This metric depends what the user values and is looking for in the simulation. For the purposes of the questions outlined previously, the best way to describe phase-separating systems is by using a distance metric or looking at correlation of monomers in space. The first one is straight forward: the sum of distances in space, considering periodic boundary conditions, between all monomer pairs. Explicitly, it can be described with the following equations:

$$T_1 = (M_{k,i} - M_{j,i})^2$$

$$T_2 = min((M_{k,i} - M_{j,i} \pm L)^2)$$

$$D_{Total} = \sum_{j,k,j\neq k}^{N} \sqrt{\sum_{i}^{3} min(T_1, T_2)}$$

Where j, k run through all monomer pairs until N unique permutations, L is the system size [7] , i runs through the three spacial coordinates and M is the spacial location of a monomer in the system.

For the second method the pair correlation function needs to be defined as the probability density to find a particle at a distance r from a reference point.

[7] Size of the simulation box.

Mathematically, it can be stated as:

$$C(r) = \frac{1}{N} \sum_{n=1}^{N} \sum_{m=1}^{N} \langle \delta(r - (R_n - R_m)) \rangle$$

where R_n are position vectors of the particles. [11] The second metric uses the property of correlations in space and finds the points at which these correlations decay to become insignificant. Namely, the correlation length of the system of spatially organized monomers can display the size of any potential clusters that occur during the simulations. Away from criticality the pair correlation function can be approximated as a power law and exponential decay as follows:

$$C \propto \frac{1}{r^a} e^{-\frac{r}{l}}$$

[11, 12] where r is the distance away from the reference particle, a is a dimensional scaling term, and l is the correlation length depicting the decay rate of the correlation function.

To implement this metric, the simulator chooses sample states along its simulation steps and calculates the pair correlation function from the reference of each single monomer and average to find a single correlation function. This is then fit to the above equation to extract values of l which are then monitored over the course of the simulation.

2.3.3 Testing Simulator for Expected Behaviours

To make sure that the simulator behaves as expected when the system is initially ordered, we simulate a system with 20 polymers of 10 monomers each with 3 different charge patterns. This is shown in figure 2-5, where the all positive charge pattern is

used to model a system which is expected to repel itself at all time points and confer a high distance metric. This is used as a sanity check to make sure the simulator behaves as expected. The other two patterns: blocky and alternating are the test set which allow us to judge the behaviour of the calibrated simulator. The tests are run for 10^6 Monte Carlo steps each. Any supplemental simulation shown in the results includes the steps run, the charge setting along with the interacting polymer species. If the time steps is not mentioned it should be assumed to be 10^6.

2.4 Results and Discussion

2.4.1 Behaviour of Test Samples

The expected behaviour of the all positive charge pattern [8] is to repel each other and be situated at the maximum distance possible from each other. In terms of intra-polymer bonds the maximum distance is set by the simulation to be 1 unit of the simulation box. With this in mind this simulation setup should show spacing between monomers of a polymers to be ~ 1, while those between polymers depend on the system size and number of polymers. Figure 2-5 (A), (B) shows the simulation output for the first scenario, in which all charges are positive [9] . The total time is 10^6 and a total of 100 evenly distributed samples are taken. Using the distance metric we find that the initial configuration starts at a low value and quickly rises to a steady-state value. This is expected since the system will try to find the farthest distance,

[8] All positive was chosen to avoid repeating with a negative set, but the behaviour is the same.

[9] They are all +5

in terms of periodic boundary conditions, at which the Hamiltonian is minimized. Compared to the initial state of the system in Figure 2-5 A, the configuration in 2-5 B shows similar extending chain conformations, due to the repulsive forces of like charges on the chain. This makes sense because the optimal structure for this system of charges should be some sort of ordered crystal lattice which does not move drastically. The result can also be seen when looking at the correlation length fit extracted at each sample from the average correlation function. The correlation length decays rapidly until it reaches a steady state value which is identical to the spacing between the monomer units near the end of the simulation.

Using this test as a baseline, we can compare the behaviour of the distance and correlation length metrics for the remaining charge profiles. Das et al show that blocky charges along a polymer species give that species a higher propensity to phase separate when compared with an alternating charge system. [44] They look at the polymer density and occurrence of percolation cluster to study the phase boundaries of their system and find drastically different phase boundaries given these two metrics to assess phase separation. They show that percolation clusters can exist above the critical temperature gained by the phase profile using the density metric for the blocky system. Surprisingly their data indicates percolation clusters form at lower concentrations for the blocky system which then transitions into polymer dense regions, while the opposite is true for the alternating charges. [44] Although the phase states are different for these two species both have the ability, given appropriate critical temperatures and concentrations, to phase separate. Figure 2-7 (A),(B) show the result of this alternating distribution at the initial state and at a state near the

end. It should be noted that the overall behaviour of the system is noticeably different to the all positive case in which the system does not move drastically and takes a crystal lattice-like shape. There seems to be a large cluster in (B) which makes sense for this small system. Simulating larger systems shows creation of multiple small clusters locally which aggregate over time to form larger clusters. The equilibrium state for these clusters is to have a single large cluster and this is usually achieved when the simulation is run for sufficient time. Running large simulation [10] requires a lot of computational power. The nature of the MCMH protocol does not allow the simulation to be parallelizable and as such it must run on one core. This limitation put the bottleneck for the simulator on the processing power of a single core.

The distance metric (Figure 2-7 C) shows a rapid decrease in the distances between polymers which makes sense because polymers are actually aggregating. The correlation length (Figure 2-7 D) seems to predict the size of the clusters: ~ 0.8 units. Since the correlation function is defined as the average correlation length calculated using all points in the system, the absolute value depends on the number of small clusters in the system. In this case the average is brought down as the smaller clusters are composed of $2-4$ monomers. With the limiting case when there is one stable cluster the values accurately represent the size of the cluster. Options to account for multiple clusters in our analysis is discussed later.

[10] Large in this context is on the order of 200 polymers with over 20 monomers running for over 1 million steps.

Figure 2-6 shows the same system but now with blocky charges along the polymers and we see a similar story to that of the alternating part. Both of these systems are able to come together and create aggregated states. In terms of DAO-5 this is promising because we show that it is possible for the protein to be a part of these aggregates using only a homopolymer solution of itself. This is not to claim that *in vivo* this is the mechanism and no other species are required, we have just shown that it is possible for this protein with its charge patterning to undergo this transition without the need for interacting cations or other polymer species. Essentially, we have recreated the system in the Das paper on a smaller scale and shown that alternating charge system seem to behave identically to the blocky system. Although this needs to be extended to look at phase diagrams of the two species, but the same temperatures and concentrations they seem to behave identically.

2.4.2 Diffusion Kinetics of Polymer Species

So far this simulator has only shown the ability to claim if a system can aggregate by comparing total distance and correlation length values with the all positive cases (with same concentration and temperature), but it is also desirable to know the kinetics of these proteins in the different stages. The diffusion profiles of these system should vary across polymer dense and dilute states, as the monomers in the system have different an exploration space. This can serve as a future metric with which to classify the material properties of any phases in the system. This simulator can also accurately recover expected diffusion dynamics of free floating molecules and also polymer chains in crowded conditions.

Since the absolute values of the diffusion coefficients will depend on the random step size in the MCMH scheme all future simulations are done with the same step size. [11] For each sampled system the trajectories will be used to calculate the mean squared displacement along a varying time step. This is identical to changing the sample step from the previous simulations and calculating the displacements from there. The basics of this metric is that one calculates the square displacements at each timestep in the trajectory relative to a reference point and takes the average. One can write the mean squared displacement as a function of this time step:

$$msd(\tau) = \langle (r(t+\tau) - r(t))^2 \rangle = 2dD\tau^{alpha}$$

Where $r(t)$ is the position at time t, d is the dimension of the system, D is the constant diffusion coefficient and $alpha$ is the scaling factor that determines the nature of the movement. $alpha = 1$ implies random diffusion while $alpha < 1$ is sub-diffusive movement and $alpha > 1$ implies directed movement.

For the MSD, a simple simulation of a single polymer of only one monomer [12] should show a scaling factor of 1, while for polymers in a compacted state should recover $alpha < 1$. Figure 2-8 shows the behaviour of the former case, a set of single point charges (no connections) moving in space. The simulation is over 10^6 steps with

[11] This is a random value between 0 and 1 units, picked uniformly, of the simulation box. Over many moves the dynamics are the same.

[12] A single point charge.

every step being sampled. [13] In a log-log plot of the mean squared displacement as a function of the displacement τ shows a scaling of $alpha \sim 1$ which is expected for a single point charge moving freely in space. It should be noted that at large τ the function starts to become noisy, and reaches a steady state which is equivalent to the maximum displacement possible in a box with periodic conditions. For this system this maximum is half of the diagonal of the $3D$ box.

To determine how this scaling changes when multiple monomers are added to a single polymer, we simulated a single polymer with 5 monomers of no charge and allowed it to freely move in $3D$ using the simulation architecture. Figure 2-9 (A) in blue, shows the average mean squared displacement for a single monomer along the chain. There are two distinct regimes in the plot, an initial linear (in log-log) scaling with a plateau and another linear scaling, culminating with the characteristic steady state at long τ. This is an expected result in accordance to the Rouse model of polymer chains, in which monomer units are connected by hookian springs in a chain. [11–13] This is exactly the system we are simulating if charges are taken out of the monomer units. [14] To simplify matters, the mean squared displacement equation for the Rouse model has two limits: when τ is small and when it is large. Using these limits, one can find two scaling relations for the squared displacements: $\tau^{\frac{1}{2}}$ at low

[13] A full trajectory in respects to every Monte Carlo move.

[14] To reduce tangents in this manuscript, a mathematical description of Rouse theory is not outlined, but can be found in introductory polymer physics textbooks. [11–13]

and τ^1 at high. [15] [11–13] However, if one uses the centre of mass of this polymer chain then the expected τ^1 scaling is recovered. To verify the $\tau^{\frac{1}{2}}$ scaling at low τ the initial 10% of the blue curve is fit with the mean squared relation and a scaling factor of $alpha = 0.54 \pm 0.47$ is recovered. Figure 2-9 (A) shows the average of the individual monomer displacement functions along with the function calculated using the centre of mass of the chain shown in green. Of course, when one adds charges to the system the multiple scaling regimes still remain because they are still ball and spring simulations. To outline the nature of the diffusion coefficient difference along the same polymer chain figure 2-9 (B) shows diffusion coefficients along the index of the polymer. As expected the system shows more freedom for monomers at the ends of the polymer than those trapped in the interior. These results are consistent with the expectation of these limit cases, as such these dynamics show that this simulator is accurately modelling theoretical polymer behaviour and is accurate to use for larger systems where the expected behaviour is not intuitive. Further work on using this to classify regions within a simulation are being explored.

2.5 Conclusion and Outlook

It has been shown that using a simple ball and chain system to model protein sequences in solution is an adequate enough approach to recover multiple known properties of polymer dynamics. By using the novel approach of a pair correlation function as a metric for the phase separation of polymer species in this type of

[15] There are actually multiple scaling relations depending on the relaxation times, but this is not in the scope of the discussion.

simulation, we have attempted to identify cluster sizes as a function of simulation time. The simulation also is able to accurately describe protein dynamics with respect to theoretical formulations of the Rouse model for protein chains. Namely, it has been shown that intrinsically disordered proteins with a rapidly alternating charge distribution are able to undergo the same type of clustering as proteins with a blocky charge distribution.

The objective of this project was to create a robust yet simple simulator of proteins which could mimic and eventually predict properties of these proteins *in vivo*. That is an ambitious goal but as it has been shown, the proposed simulator is adaptable to any protein system the user wants to investigate, and can quantifiably predict the behaviour of the coarse grained system. It should be clear that this does not guarantee that the modelled polymer structures behave accordingly *in vivo*. Nevertheless, the simulator can be used to generate hypothesis that aid in the design of *in vivo* and *in vitro* experiments. By using this abstract coarse graining the system loses most of the chemical and steric information of actual protein folding and interactions. One can extend this framework to model explicit atoms in the protein, akin to GROMACS simulations, but the timescale and amount of proteins has to be limited due to the computational bottleneck. [46]

The group is currently working to create phase diagrams for the systems in question to determine for these systems of polymers, the behaviour of the clustering

as interactions energies and temperature vary. [16] This manuscript has only shown some of the capabilities of this simulator, but it can be extended to model different biological scenarios. For example, filament structures can be modelled by changing the free movement of monomers to satisfy rigid rods, while also exchanging Coulomb's interactions with another interaction term to mimic filament monomer interactions.

The program is available to download on GitHub and is regularly updated. To access the files and make feature suggestions or to report bugs the user should visit the GitHub repository. [17]

[16] The interaction energy can be modulated by changing the coefficient of the coulombs interaction term while the temperature plays a part in the acceptance or rejection of a specific move, in accordance to the Boltzmann distribution outlined previously.

[17] $https://github.com/joemans3/C\_Polymer$ included is a python module for the analysis of the trajectory output.

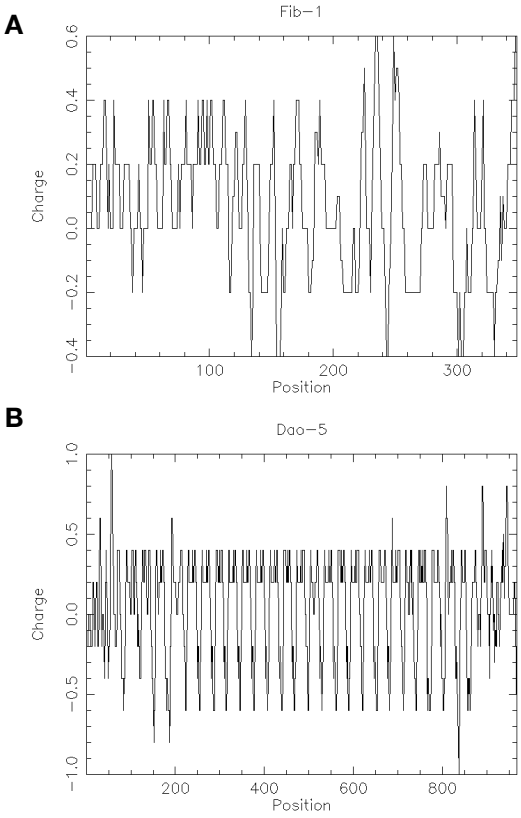

Figure 2–1: Plots depicting the charge profiles of Fib-1 and Dao-5, A and B respectively. Charges are calculated using a running window of 5 residues and the average of this window is associated to the middle residue of the set. Figure were created using the EMBOSS code base available at http://www.bioinformatics.nl/cgi-bin/emboss.

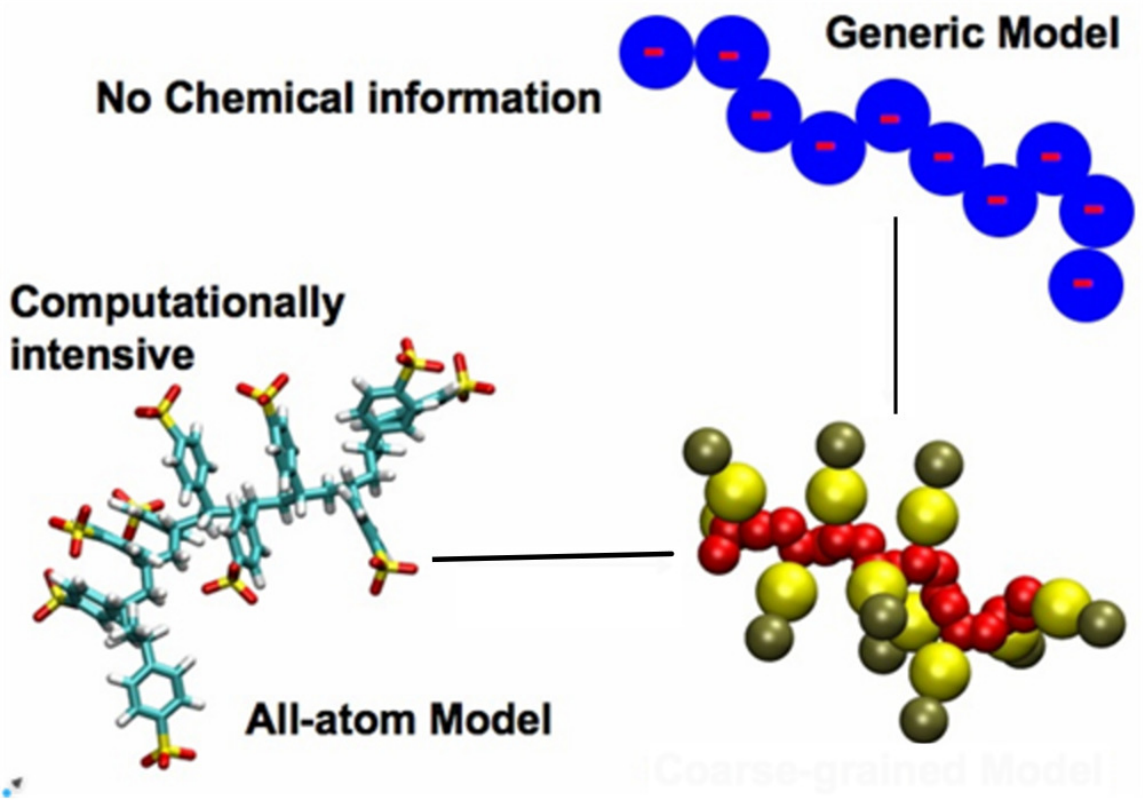

Figure 2–2: Example of the various levels of detail to model a simulator from. The bottom left shows explicit bonds and side chains of amino acids, while the top right is a ball and chain model which has no chemical information. The latter is used as a base for the proposed simulator. This figure is adapted from *Multiscale, Hybrid and Coarse-Grained Methods*. [47]

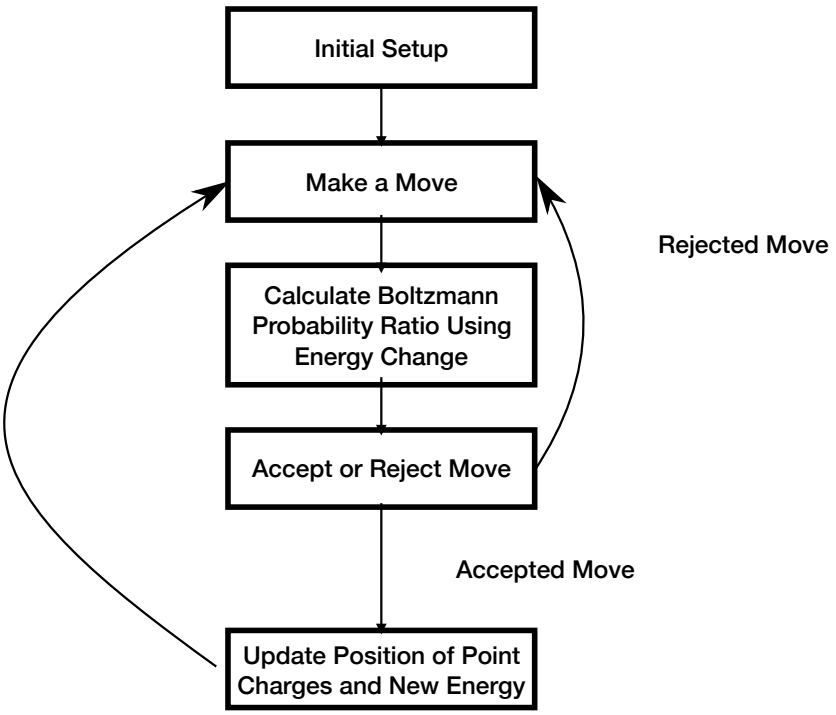

Figure 2–3: Flow diagram for a generic MCMH (Monte Carlo Metropolis Hastings) scheme.

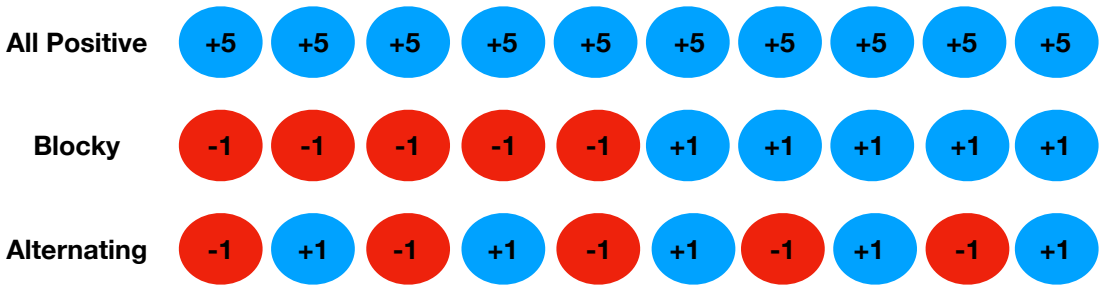

Figure 2–4: The charge pattern of the three types of test run to ensure validity of the simulator. The first is all positive charges, the second is a blocky charge pattern and the final one is an alternating pattern.

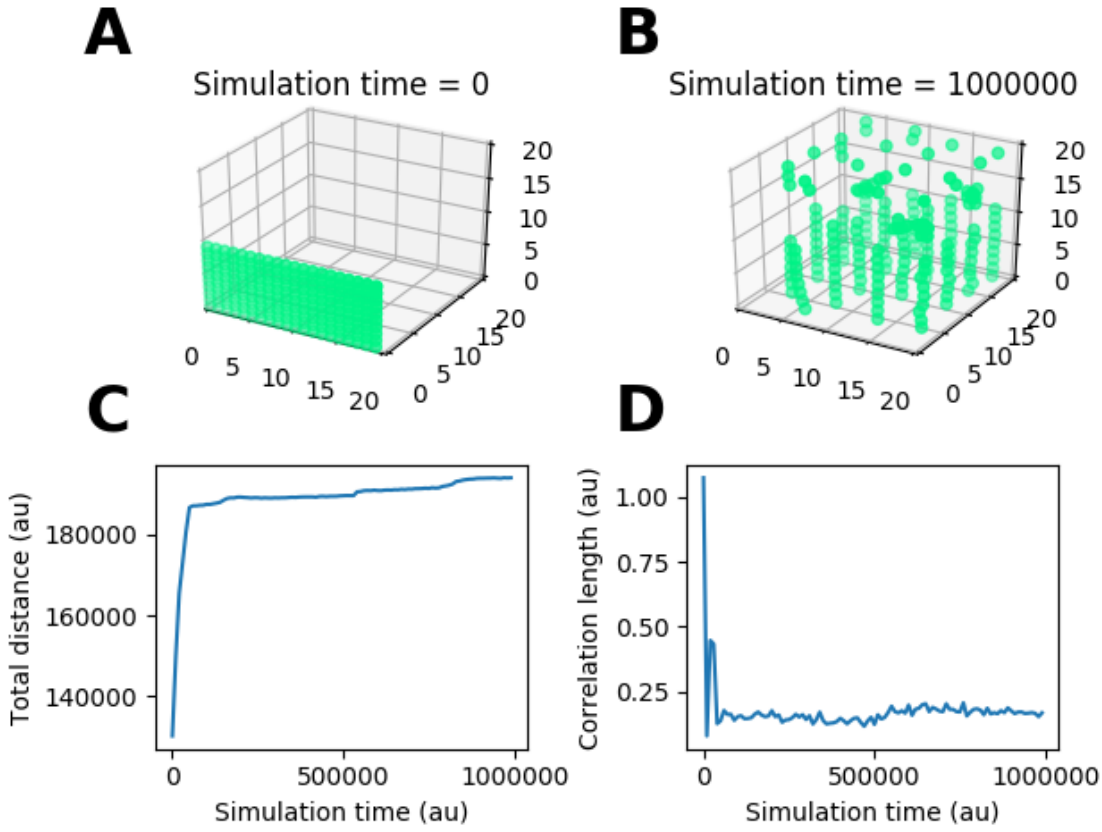

Figure 2–5: Simulation results for the all positive charge (green spheres) distribution at initial (A) configuration and final (B). The distance metric (C) of the simulation shows an increase over time of the distances between monomers. The correlation length fit (D) extracted from the average correlation function over time shows a decrease to a steady state value which indicates small scale spatial correlations.

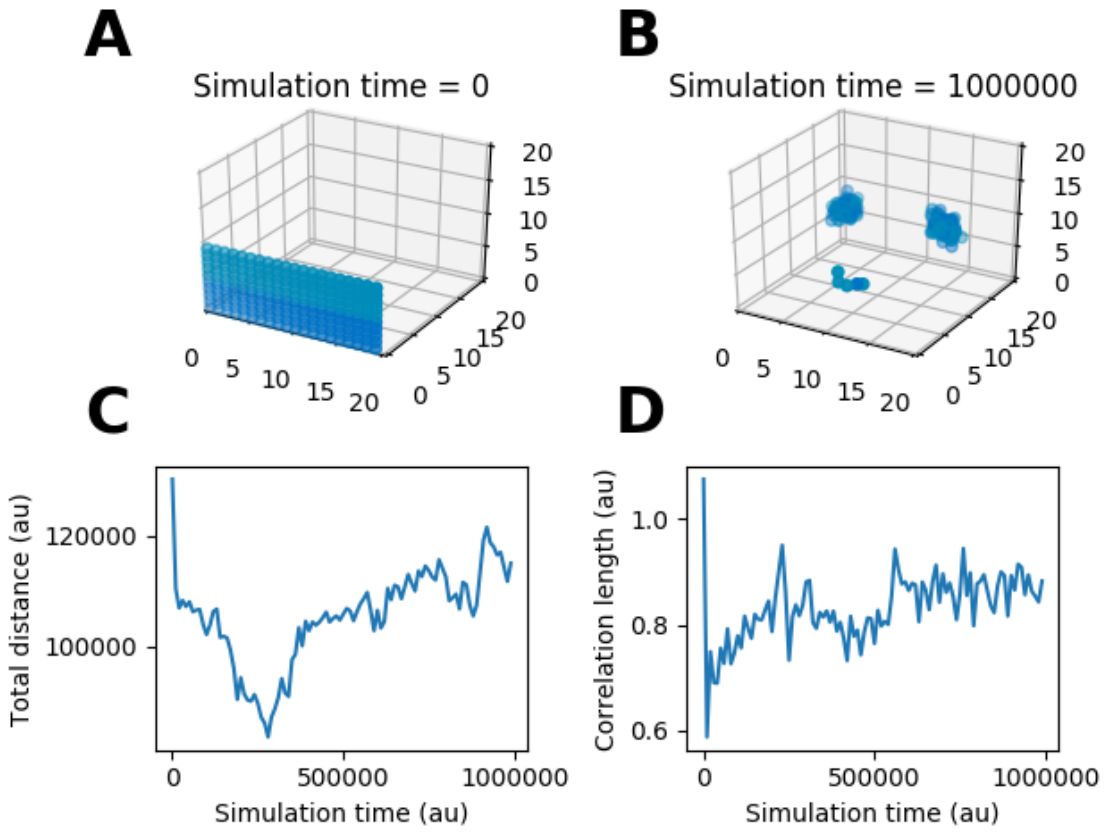

Figure 2–6: Simulation results for the blocky (dark and light blue spheres) distribution at initial (A) configuration and near the final (B). The distance metric (C) of the simulation shows a decrease over time of the distances between monomers. The correlation length fit (D) extracted from the average correlation function over time shows an increase to a steady state value implied the averaged cluster size of the system.

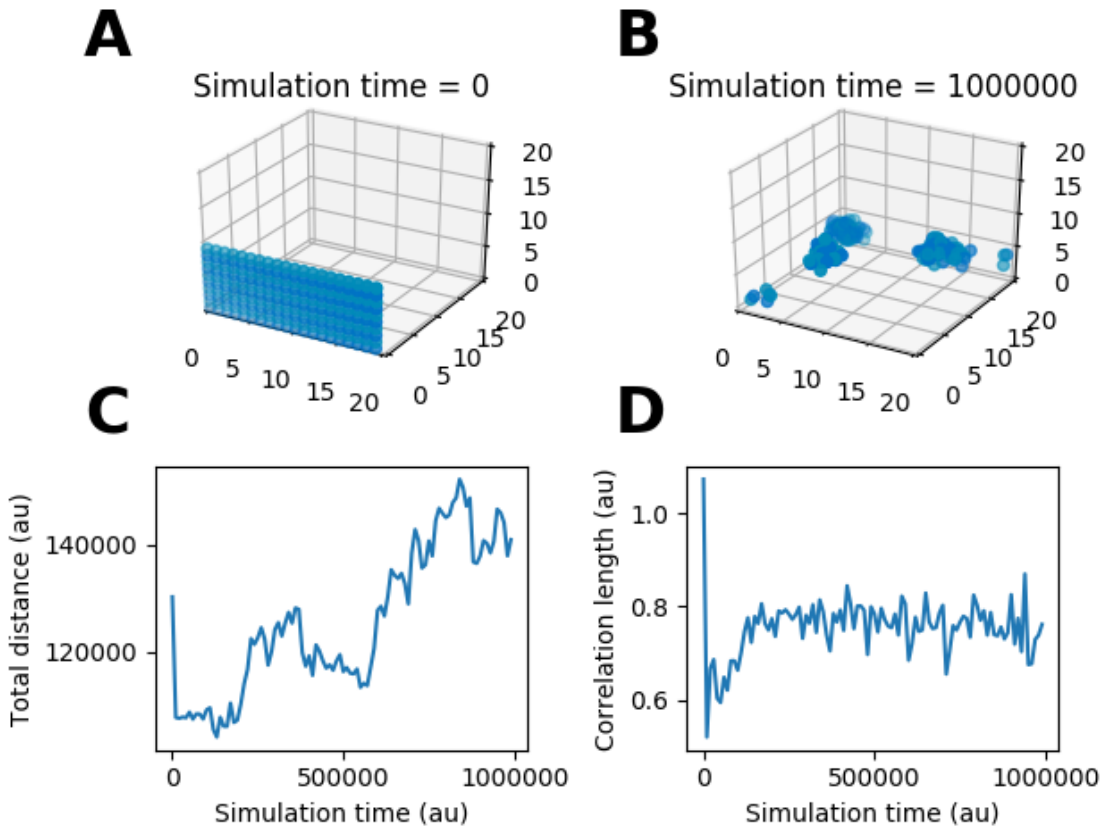

Figure 2–7: Simulation results for the alternating charge distribution at initial (A) configuration and near the final (B). The distance metric (C) of the simulation shows an initial decrease and subsequent incremental increase of the distances between monomers. The correlation length fit (D) extracted from the average correlation function over time shows an increase to a steady state value implied the averaged cluster size of the system.

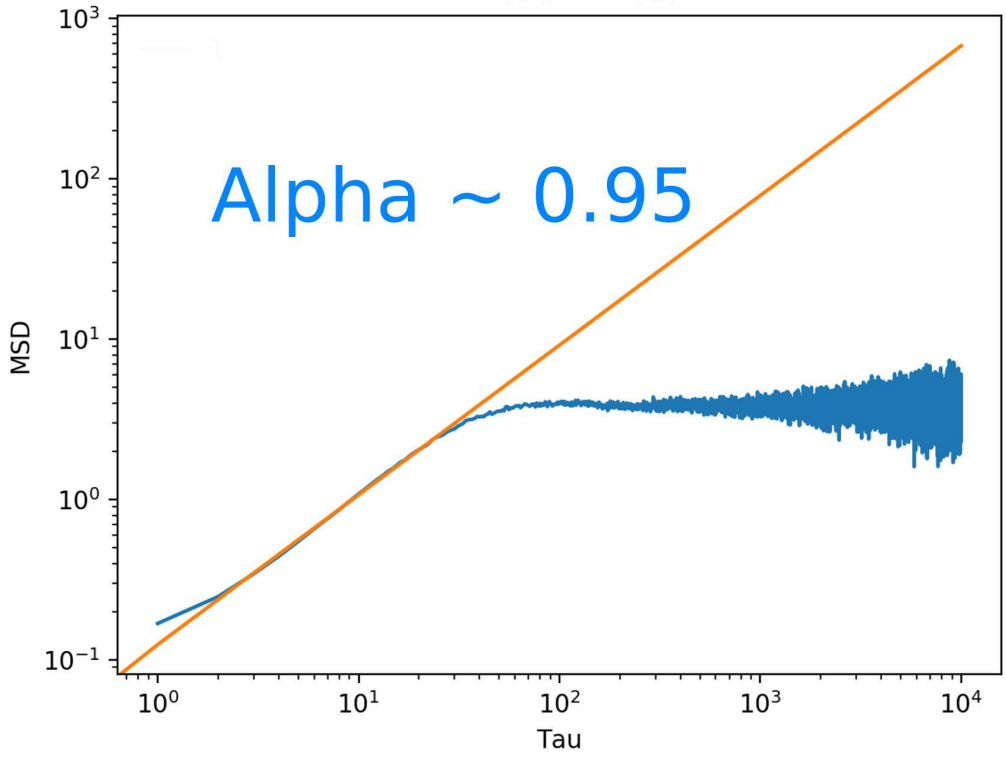

Figure 2–8: Shown is the mean square displacement (blue), for a simulation of 5 uncharged monomer (not connected) moving in 3D using the MCMH scheme, as a function of τ (timestep). The fitted line (orange) is the linear fit to the scaling relation outlined in the discussion. The fit is only applied to the first 10% of the data to avoid including noise at high τ. The fitted value for alpha is 0.948 ± 0.13. Error is calculated using the covariance matrix of the linear squares fit.

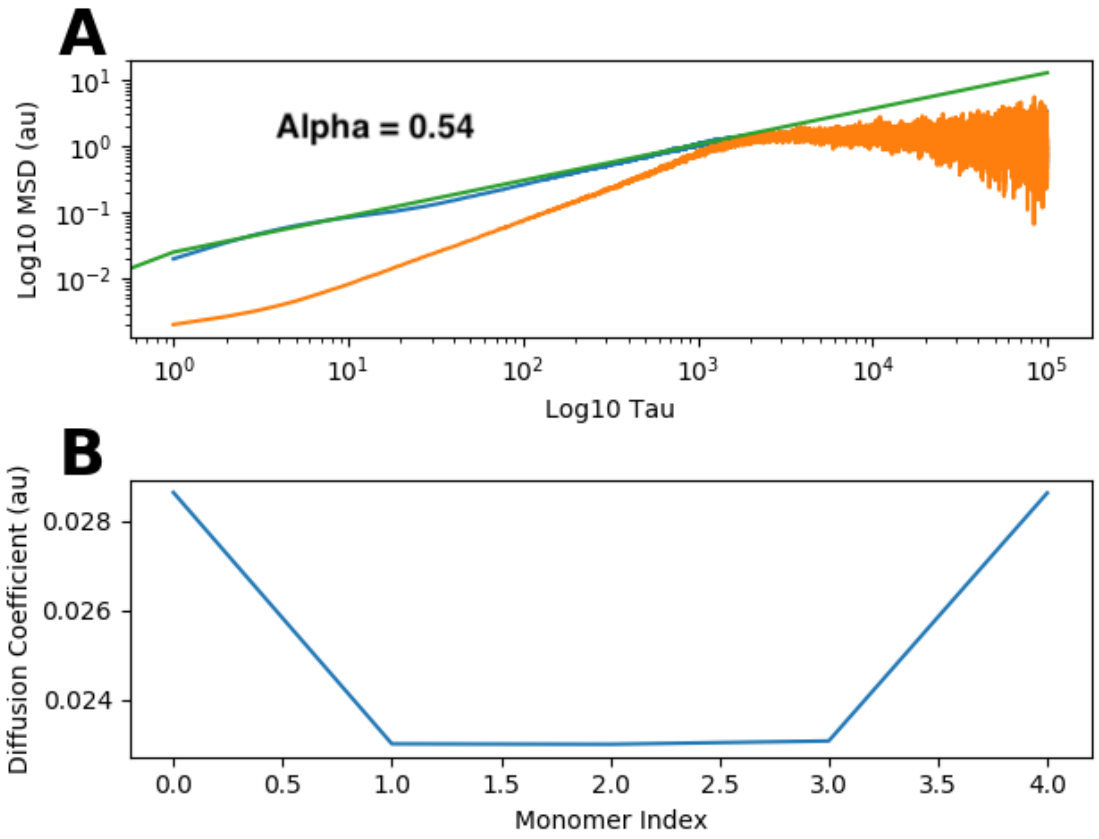

Figure 2–9: (A)Shown is the average mean square displacement for a monomer along a chain (blue), for a simulation of 1 polymer with 5 monomers, moving in 3D using the MCMH scheme, as a function of τ (timestep). The fitted line (green) is the linear fit to the scaling relation outlined in the discussion. The fit is only applied to the first 10% of the data to avoid including noise at high τ. The fitted value for alpha is 0.54 ± 0.47. Error is calculated using the covariance matrix of the linear squares fit. The orange curve is the mean squared displacement calculated using the centre of mass of the polymer rather than a monomer point. (B) Shows the values of the diffusion coefficients along the polymer chain as a function of monomer index.

CHAPTER 3
Determining the Nature of RNA Polymerase Clusters in *E. coli*

3.1 Introduction and Background

Since the first association of liquid-like condensates in cell biology as a product of phase separation, groups have been keen to find examples of the universality of this phenomenon. [1] Although it is not always possible to associate the occurrence of these phases to the behaviour of the cell, it is evident that order is occurring in a classically disordered system. Much of the phase separation literature focuses on examples in eukaryotic systems, in which the cell already has a mechanism with which to segregate itself into distinct components. As such, the ability to also create distinct subregions, with such abundance, which are not membrane bound but exist through phase separation, is interesting. This mechanism may be prominent in the less studied prokaryotic system in which membrane bound organelles are not found and proteins are diffusing in a mixed state as it gives the cells a method to organize its constituents. Recent works have shown that the prokaryotic system also shows examples of this mechanism at work. [48]

Many groups have shown a clustering of DNA-directed RNA polymerase (RNAP) in *E. coli* in optimal growth conditions, namely, rich media and 37°C. [49–53] In bacteria RNAP controls all of the transcription in the cell, and as such the occurrence of these clusters has been linked to the transcriptional activity in the cell, which is modulated by the growth conditions. [51–53] Namely, the synthesis of rRNA in the

cell during high growth is proposed to be the stabilizing factor for these clusters of RNAP, and treatments with rifampicin, a global transcription inhibitor, have shown to disperse the clusters. [51–53] However, Weng et al. have recently shown that these clusters are also able to form when rRNA production is abolished. [53] They found that although in fast growing cells the RNAP clusters are spots for transcription, the occurrence and organization of the clusters is independent of rRNA synthesis and more dependent on the structure of the chromosome. [53]

The formation of these clusters is important for the organizational abilities of this cell. Is this simply a $1:1$ binding of the RNAP to DNA in a dense location or are these proteins existing in a phase separated state around these binding sites in an effort to increase initiation rate? Both these scenarios imply different kinetics within these clusters, which can be tested using fluorescence recovery after photobleaching (FRAP) and observe the exchange between the clusters and the environment. DNA bound RNAP is restricted in movement until it leaves the DNA and as such the time for recovery of the fluorescence will be much longer when compared to RNAP which are freely moving. Due to the sub-micron size of these clusters it is not always possible to bleach a small area to look for recovery, as such FRAP methods are not applicable and another method needs to be used. We approach this by tracking individual molecules of RNAP in different regions of the cell to determine the changes, if any, in the diffusion pattern. Groups have used super resolution techniques to observe single molecules of RNAP in and outside these clusters before, but classified only transcribing and freely diffusion molecules. [51, 52] They were using this to compare movements with the chromosome and specially map the distributions of these clusters

along the body of the cell. [51,52] We are more interested in using the single molecule imaging to quantify the movement of proteins in different compartments of the cell and determine whether molecular motion inside the clusters is consistent with DNA binding and/or phase separation.

3.2 Methods and Materials

3.2.1 Cell Culture Preparation

E. coli strains expressing RpoC::mMaple (DNA-directed RNA polymerase subunit beta') created by Dr. Anne-Marie Ladouceur are used as a fluorescent marker in order to track the RNAP complex as a whole. Colonies of RpoC::mMaple tagged cells were kept in agar plates under $4°C$ conditions. Cultures were started with a single colony grown at 37° C in a 1ml EZ-rich solution with 20% glucose as a carbon source, for a period of 16 hours in a shaking incubator.

After the growth period, 1% of the cells are incubated again with 10ml EZ-rich solution with 20% glucose for a period of 90 minutes. To quantify 1% of the cells, the concentration of the culture is calculated using a spectrophotometer at OD_{600}. The appropriate volume of the sample to use is then calculated by conservation of mass. This is done to ensure the cells are growing optimally in a non crowded and rich media. The aim is to have these cells in mid-log phase such that the growth rate is the highest. This is the point at which the brightest foci have been observed.

3.2.2 Live Cell Imaging

Slides are created using Thermo Fisher gene frames (1.0 x 1.0 cm) and treated cover slips to counteract dust. Cover slips are soaked in versa-clean overnight, and undergo a methanol and acetone wash cycle before being sonicated for 30 minutes.

The cover slips are then put into a plasma oven for 15 minutes. Finally the slips are flamed before being used on the preparation slide. The gene frames are filled with a 1% agarose EZ-rich solution and kept at $37°C$ to simulate the optimal temperature and media for growth.

Imaging was done using an inverted Olympus IX83 using a 100x oil objective lens. The camera used to capture the images was a Hamamatsu Orca-Flash 4.0 sC-MOS and excitation was done using an iChrome Multi-Laser Engine from Toptica Photonics and a 405/488/561/640nm filter set. Initially, the target area was photobleached by continuous excitation using 561nm. 5000 frame movies spaced with 50ms intervals were taken on the bleached field of view while continuously activating with 405nm.

We used data obtained by Nicolas Soubry from the Reyes lab using a lacO array expressing LacI::mMaple to control for tight continuous binding. Data was taken in the same conditions and analysis occurred identically.

3.2.3 TrackMate to Create Molecular Trajectory Data

Two types of image are collected from the microscope: bright field fluorescence and at $561nm$ excitation. The bright field images are used to segment cells to make the trajectory analysis easier and reduce external noise produced outside the cell from dust particles. TrackMate uses a laplacian of the gaussian (LoG) method for spot detection. [54] The subsequent filtering of the spots based on intensity thresholds is done to eliminate unlikely spots. The radius of the spots is estimated by the fitting of a gaussian to the centre and extracting the standard deviations. Spots are then linked between frames using a simple nearest neighbours search. To account for

blinking molecules a gap parameter is used to allow linking to occur even if a spot does not occur in a subsequent frame. The output from this analysis gives the centre of each spot in a track along with the average intensity and estimated radius.

3.2.4 Method for Analyzing Trajectory Data

To define a local concentration of tracks in space, we take long frame samples from one movie and create a max projection onto one frame. This is repeated N number of times; this N can vary but all the analysis presented is using $N = 1000$. This implies that every 1000 frames are projected onto one frame. Since the total frame count of the movie is 5000, the new projected space consists of 5 frames. Independently of the tracking done on the 5000 frames, these 5 frames are run through the LoG detection for spots in TrackMate. This allows us to define a droplet as a local concentration of RNAP tracks in time. To account for drops being classified on single trajectories, namely if a drop contains only one track, a threshold of 3 tracks localized to a drop is set. If the proposed drop has less than 3 tracks it is not used. It should also be noted that only tracks which have at least 10 time points are taken into account, the rest are discarded as noise.

Using this definition we create 2 spatial regions in the cell. One which is the inside of the drop, and the other, outside. If the independently computed tracks localize only to the inside of the drop, which are defined by TrackMate using a candidate radius, for the duration of the track, they are classified as In tracks. If the track spends $50 - 99\%$ of its track duration inside these drops they are classified as In and Out tracks, aptly named for moving across the boundary of the droplet. The rest are classified as Out tracks, which are never inside a drop. From a theoretical

standpoint, this classification maps the drops to be these clusters of RNAP while the tracks on the outside are non-specifically associated with the nucleoid. Using this we can extract the diffusion dynamics of all three fractions.

There are two methods to extract the diffusion coefficients of these trajectories, both of which use the mean squared displacement metric. Mathematically this is expressed as:

$$msd(\tau) = \langle (r(t+\tau) - r(t))^2 \rangle = 2dD\tau^a$$

where $r(t)$ is the position at time t, d is the dimension of the system, D is the constant diffusion coefficient and a is the scaling factor that determines the nature of the movement. $a = 1$ implies random diffusion while $a < 1$ is confined movement and $a > 1$ implies directed movement.

The first method uses the average D over the duration of the track, in which case, a is assumed to be 1. The other method calculates D from the time-averaged MSD for each track. This time step is τ and the mean squared displacement of these new sub-tracks is recalculated at each of the τ. Using these two variables we are able to extract both the diffusion coefficient and the scaling factor for all tracks. The different classified tracks are then subjected to these two methods to extract both the diffusion coefficients and scaling factors. This is done to show consistent results even with simplifying assumptions.

It is possible that this arbitrary method of classifying droplets is a circular argument. To control for this, we use the track data and apply both the average and full mean squared displacement method to gain a probability distribution of diffusion coefficients for all tracks without any classification. Then by fitting multiple

gaussian functions to this probability density, we are able to find significant peaks in the distribution which are consistent with the classifications made earlier. This is further expanded on in the discussion. This gaussian matrix fitting is done using the Scikit-learn implementation using a machine learning approach. [55]

3.3 Results and Discussion

The method of classification for these drops uses an arbitrary number of frames in the max projection. If this number is large, then movements of the drops in time will overestimate the boundaries, while setting it too low will reduce to the tracking regime. So naturally the effect of the number used needs to be controlled for. One metric is to look at each viable track in the system and calculate the fraction of time it spends inside a drop. Tracks which are classified as In will all have values of 1 and Out tracks will be 0 by definition. In and Out tracks will take on a value between 0 and 1. This is shown in figure 3-1 where each coloured line represents one of the segmented frames [1] , and the proportion of all viable tracks which exist in some fraction within a drop. One should note that the profile of this criterion is identical for all such segmented frames, which implies that the behaviour of the RNAP is consistent throughout the duration of the movie.

Figure 3-2 (A) depicts the diffusion coefficients of these tracks calculated using the average mean square displacements of a track after being classified into the three categories. Although there is considerable overlap in the distributions it should be clear that they have distinct peaks. The scale is semi-log in the diffusion coefficients

[1] Namely, the first 1000, second, third, fourth, and fifth projected frames.

which are shown in pixel units of the camera capture. The mean values for the 3 classified distributions are: In $= 0.166 \pm 0.160 \frac{\mu m^2}{s}$, Out $= 0.812 \pm 0.772 \frac{\mu m^2}{s}$, and In and Out $= 0.440 \pm 0.360 \frac{\mu m^2}{s}$. For completeness these values are provided in Table 1.

It can be seen that in the RpoC data the In fraction seems to the slowest, while the diffusion coefficients increase for In and Out and Out. This suggests that the local densities of RNAP are either creating a confined region which has a higher density than the outside or that there exists a fraction of the In RNAP which are actually tightly bound to the DNA (actively transcribing) and bring down the effective diffusion coefficients. The latter is unlikely because of the unimodality of the In distribution suggests one type of movement. It is possible that the fraction of these bound and transcribing RNAP are underrepresented in the distribution. Namely, they may exist but relative to the copy number of unbound they are lost in the distribution. To control for this we use the control *lacO*-LacI system to gain the expected diffusion coefficients if a protein is always tightly bound to the DNA. This is shown as an overlay in figure 3-2 (A) in teal. We can see that there is not significant overlap in the distributions and the majority of the In fraction RNAP are moving much faster than the control, but a small portion of the left tail overlaps with the control. Further analysis is required to determine the nature of the In distribution which is discussed later.

A consequence of this classification procedure is that any track that is In has to be a track in a local density area. If for instance TrackMate is unable to identify a local density of tracks into a drop the tracks which should be classified as In will now be classified as Out. By this logic, the distribution of In is unique while the

Out is actually a mixture of all three classifications. This is seen outright in figure 3-2 (A) where the distribution of the Out has a characteristic left tail bleeding into the In distribution. Of the total viable tracks over the 20 trials, the In fraction has 5.7%, In and Out at 22.1% and Out has 72.2% of the total tracks. This supports the over-representation of Out as seen in figure 3-2 (A).

It should be noted that there is also considerable overlap in the distributions of the In and Out and Out fractions. Figure 3-2 (B) shows a box plot representation of the same system in (A), where the distributions of these two fractions can be compared easily. They have the same range and share a similar mean despite being defined as existing in two different regimes. This is mainly due to the use of the arbitrary 50% threshold to classify between In and Out and Out [2] . If it is lowered the distributions of the In and Out fraction resembles that of the Out fractions, which makes sense since the threshold controls the amount of time a track spends inside a drop compared to the outside. If this is low, the majority of the time the track is outside the defined drops and will resemble the average dynamics of the Out fraction. If the threshold is raised, then the tracks spend most of their time in the drop and their dynamics will resemble that of the In fraction.

In an effort to reinforce the concepts of Figure 3-1, that the step size in the frame projection shows identical dynamics, we decompose the results of Figure 3-2 (B) into its 5 component frames. To be specific, for each 1000 frame projection, the respective distribution of the diffusion coefficients are displayed for all three classifications. It

[2] See methods.

is evident that the mean and range of these distributions is consistent throughout each of the 1000 steps.

As stated before, the method to classify drops can be considered circular, namely, we define a drop and localize a track to it and claim that it is in the drop. The discrepancy in the diffusion profiles can then be explained by the low fraction of tracks classified as In and the case reduces to under sampling. To control for this, we can take the total distribution without classification, sum of the In, Out and In and Out, and try to use a gaussian matrix fit to estimate the mean of a 3 gaussian mixture on the data. Figure 3-3 (A) shows the result of this analysis, where the blue curve is the distribution of RpoC track data and the vertical red line are the estimated means of 3 gaussians used to describe the data. The fitted means of the gaussians are, in no order: $0.428 \pm 0.412 \frac{\mu m^2}{s}$, $0.118 \pm 0.135 \frac{\mu m^2}{s}$, and $0.913 \pm 0.636 \frac{\mu m^2}{s}$. SI units are shown in Table 1. Comparing these gaussian means with the means of the three classified distributions we find the values are quite similar. So, using two different methods, one which calculates diffusion coefficients of pre-classified tracks and the other which doesn't, the dynamics result to be similar. Namely, the 3 classified fractions and their means are recovered from applying an independent analysis on the track data as a whole.

As we have established the validity and robustness of our analysis methods, we can return to describing the nature of these RNAP clusters. In figure 3-2 (A), the In fraction distribution overlaps slightly with the control suggesting that there might be two fractions inside the In distribution, differing by their diffusion coefficients. These may correspond to a fraction of RNAP bound and transcribing DNA and

another fraction moving in this dense drop. One way to approach this is to apply the gaussian matrix fit on the isolated In fraction distribution and see if we can fit 2 gaussians that return means which make sense. Figure 3-3 (B) shows this gaussian matrix fit of order 2 applied to only the In distribution and the recovered means are: $0.184 \pm 0.133 \frac{\mu m^2}{s}$, and $0.056 \pm 0.049 \frac{\mu m^2}{s}$. The larger value is actually similar to the overall mean diffusion of the In distribution calculated before, whereas the lower mean occurs in the region of overlap between the In fraction and the control. This fact actually becomes easier to see when only the In distribution is plotted, as in figure 3-3 (B). There seems to be a second smaller peak in the left tail of the distribution. [3] Of course, it is possible that this is an ad-hoc explication but the fact that two gaussians are able to fit this distribution and the fact that the In fraction contains overlap with the control is evidence of the two distinct fraction idea.

3.4 Conclusion and Outlook

By utilizing single molecule tracking methods we were able to quantify the diffusion kinetics of RNAP across *E. coli*. We created a robust procedure to classify local densities of RNAP and map individual RNAP tracks to them. Using this method three distinct fractions of RNAP are discovered: a DNA bound fraction inside these drops that is possibly transcribing, a free floating fraction, and one which exists in these drops but is not bound. The fact that the intermediate In fraction diffusion

[3] This mean diffusion coefficient of the bound RNAP are consistent with those found by Bakshi et al. [52]

coefficients are slower than the Out but faster than the LacO control suggests that those RNAP exist in a distinct environment.

The extent to which RNAP is responsible for initiating this clustering is still unclear. It is possible that the RNAP is not the primary protein undergoing phase separation and rather another protein, a member of the transcription initiation/elongation complex, is the one creating these clusters and the RNAP is just localizing to these regions because of binding interactions. This is similar to the interacting modules model for phase separation [4] in which a scaffold protein is undergoing phase separation and due to interacting domains with another protein, it is dragged into the system as well. To test this we are looking at candidate transcription factors, namely NusA which seems to show evidence of these clusters in rich media alongside RNAP. With sufficient diffusion data on NusA, we would like to compare the distributions of these classified fractions between the two proteins. We also want to look at how the distribution of NusA and RNAP occurs in these clusters. Initial fixed cell data suggests that NusA localizes across the whole area of the cluster while RNAP localizes in the interior. This is not conclusive to which protein is the scaffold and we need to observe the dynamics of both in real time to see if NusA localizes first and brings RNAP along or vice versa.

[4] Outlined in Chapter 2.

Table 3–1: Table of converted diffusion coefficients for data calculated for Figures 3-2(A), 3-3(A), and 3-3(B). Units are given in $\frac{\mu m^2}{s}$, and association to different RNAP fractions is made. For calculations which do not map to a certain fraction, the entry is NA.

	In Mean ($\frac{\mu m^2}{s}$)	**In Bound Mean** ($\frac{\mu m^2}{s}$)	**Out Mean** ($\frac{\mu m^2}{s}$)	**In:Out Mean** ($\frac{\mu m^2}{s}$)
Figure 3-2 (A)	0.166 ± 0.160	NA	0.812 ± 0.772	0.440 ± 0.360
Figure 3-3 (A)	0.118 ± 0.135	NA	0.913 ± 0.636	0.428 ± 0.412
Figure 3-3 (B)	0.184 ± 0.133	0.056 ± 0.049	NA	NA

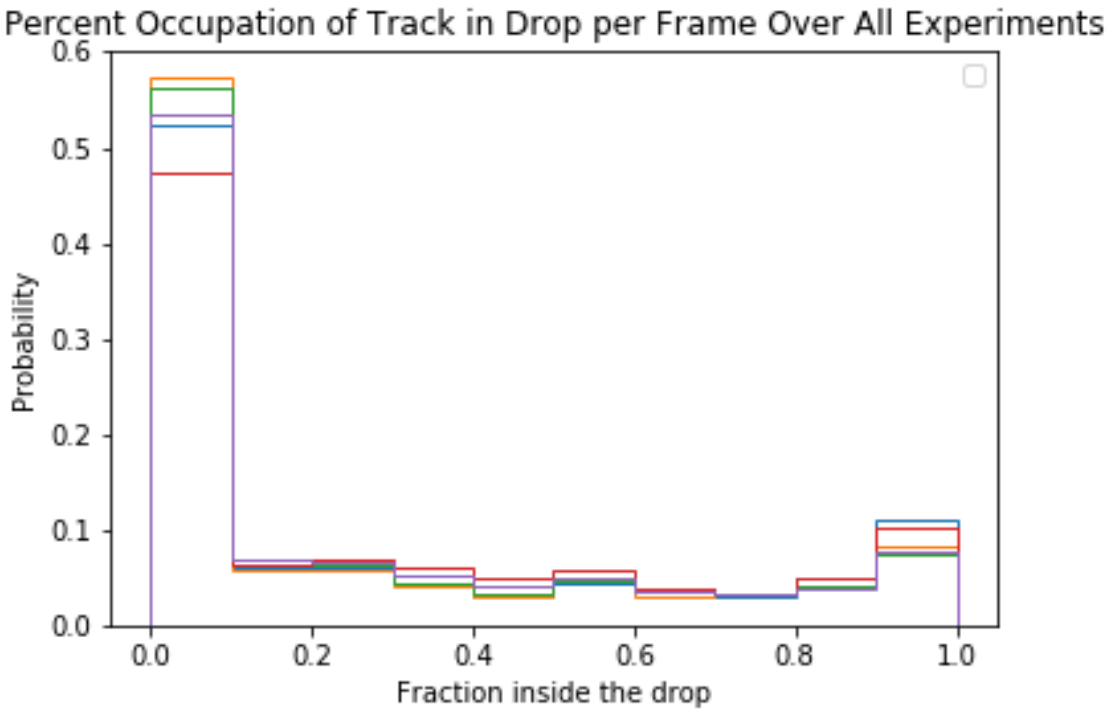

Figure 3–1: Shown in different colours are the segmented frames and the vertical axis represents the fraction of viable tracks in the system that exists at varying proportions inside a defined drop. The total 5000 frame movie is decomposed into sections of 1000 frames each, where each of these sections is max projected into one frame. These 5 frames are used to classify drops and tracks as In, Out and In and Out. The data is gathered from 20 of these 5000 frame movies.

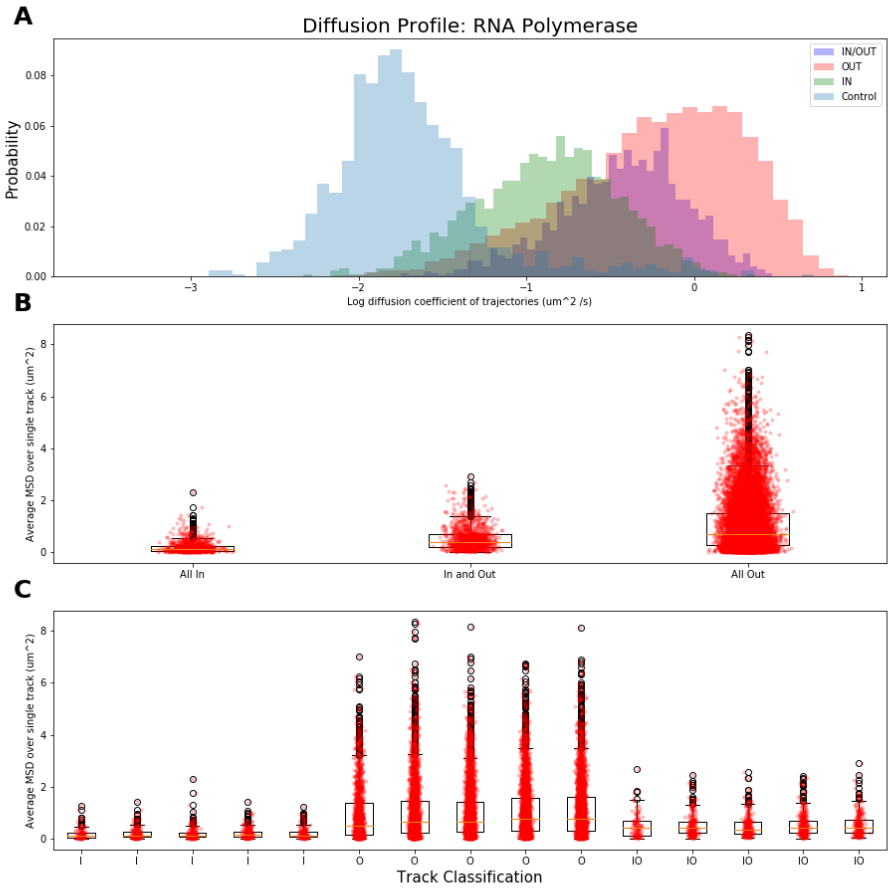

Figure 3–2: (A) Shows the time averaged diffusion coefficients calculated for 20 movies of rpoC::mMaple after classification of tracks as In, Out, and In and Out as a probability distribution. The diffusion data, without classification, of the control experiments is shown in teal as an overlay. (B) Shows the same data of the Rpoc in (A) projected onto a single dimension. Overlaid is the box plot representation of the raw data, where one track's diffusion coefficient is one red point and black disks are outliers. These are defined as points being larger than Q3 +1.5IQR where Q3 is the 3rd quartile and IQR is the interquartile range or smaller than Q1 −1.5IQR, with Q1 defined as the first quartile. (C) Shows the same data in (B) but now along the axis of the segmentation done of the 5000 frame movies into 1000 sets. Each I represents the dynamics of the In tracks belonging to the ith 1000 set. Out for O and In and Out as IO.

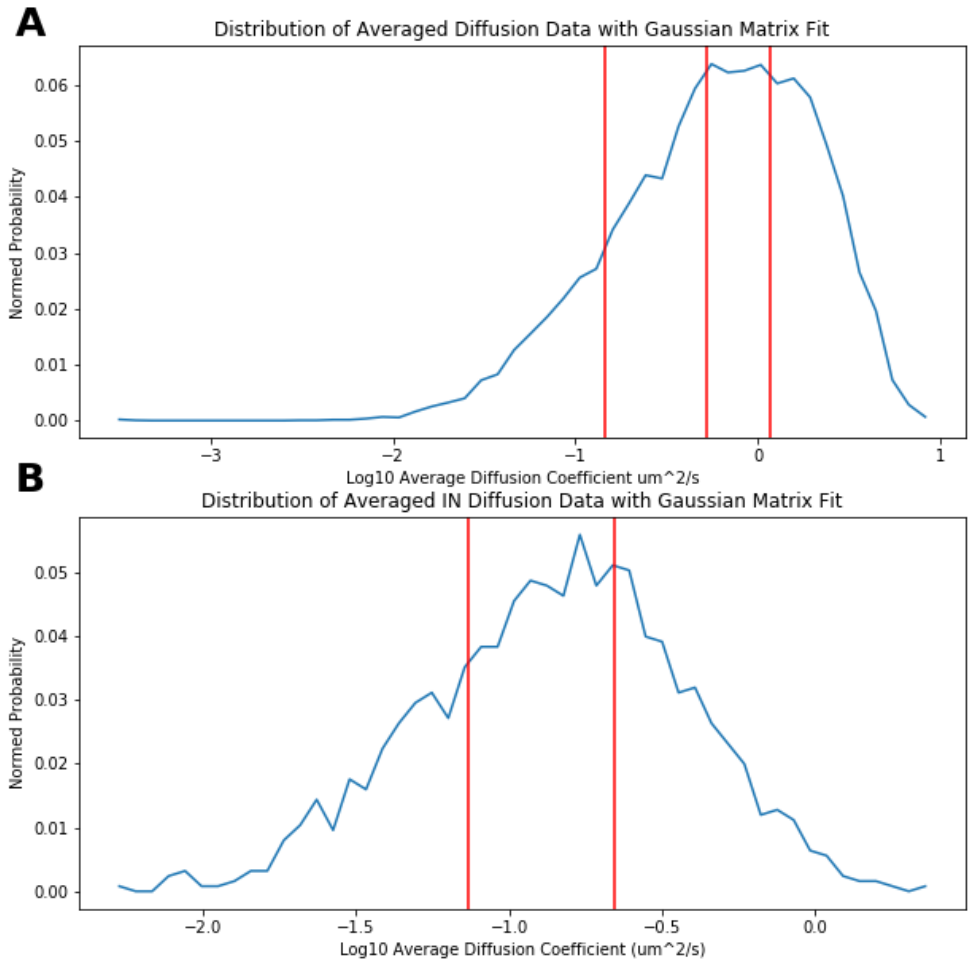

Figure 3–3: (A) The full distribution of the diffusion coefficient of all viable tracks, this is equivalent to adding the distributions of the In, Out, and In and Out fractions from figure 3-2. The red vertical lines show the means of the 3 gaussians fit onto this distribution using the gaussian matrix method. The means are: 0.118 ± 0.135, 0.428 ± 0.412, 0.913 ± 0.636 in $\frac{\mu m^2}{s}$. (B)The In distribution of the diffusion coefficient of all viable tracks. The red vertical lines show the means of the 2 gaussians fit onto this distribution using the gaussian matrix method. The means are: 0.056 ± 0.049 and 0.184 ± 0.133 in $\frac{\mu m^2}{s}$.

CHAPTER 4
Discussion and Conclusion

To gain insight into the weak promiscuous polymer interactions which underlie some of these phase separated systems we created a coarse grained molecular simulation. The relationship between charge patterns in polymer species and their propensity to aggregate/phase separate as explored in point mutation studies with FUS and similar proteins, seems to indicate that certain charge patterns provide conformational heterogeneity to these polymers which has an effect on the protein's ability to phase separate. Our simulations attempt to emulate this heterogeneity by our ability to modulate the types of polymers available to the system and also their charge patterns. The probabilistic backbone provided by Markov Chain Monte Carlo schemes is an accurate emulation of the noise which exists in cells. By observing both the small scale movements of polymer chains and the global dynamics of the system as a whole we gain a complete picture of this artificial system. By defining how we believe the particles interact, through the Hamiltonian, we can explore the whole set of energetic configurations and pick the most likely properties given this Hamiltonian. Since we wanted to study electrostatic interactions between monomer species, the Hamiltonian is only defined with the respective potentials. In other scenarios the user should be able to extend the Hamiltonian easily in an attempt to model their own systems.

While figure 2-5 acts as a control check, the two charge patterns in figures 2-6, 2-7 show that although blocks of charged residues perform the best, in terms of clustering, the alternating charge distributions are not all that far behind. We can judge this not only by the absolute values of the distance and correlation metrics but also the stability of these both over time. Even though the alternating distribution allows for clustering the fluctuations in the system are higher than that of the blocky system, suggesting that pure entropic and energetic forces favour the blocky system when it comes to clustering via these promiscuous electrostatic interactions [1] . [9] With respect to the alternating charge profile of DAO-5, we can predict that the charge pattern alone should be sufficient for the protein to phase separate *in vitro*. Of course, the simulations need to be extended to larger polymer species to accurately assess the role of polymer length in creating polymer meshes within the system. In a sense one can think of each monomer as an interaction hub, and the more there are the more interaction permutations are possible in the system. Also since this protein has not been purified and studied *in vitro* it is hard to access the accuracy of our predictions so far. Contrary to the experiments with Ddx4, we find that both charged systems are capable of clustering, suggesting that the phase separating behaviour may be further modulated by the way in which these charges are achieved (amino acids) or the overall length. [42] By increasing the length of these systems to the scale of $\sim$100s of monomers per polymer one should be able to accurately assess

[1] This is an expected result based on the mutation studies done *in vitro* on proteins such as FUS [9, 30]

these properties. Computational efficiency impedes this due to the vast amounts of interactions the simulator needs to iterate through. Further work needs to be done regarding the efficiency of the simulator's algorithms and also see if parallel computing can be used in conjunction with the Monte Carlo scheme.

Although not shown the simulator also supports changes made to the distribution of charges and addition of any new polymers into the system at any time during the simulation. This allows the user to effectively model the role of post translational-like modifications on the behaviour of these polymer systems. Once a cluster forms, the user can disrupt all the charges in the system to be all the same and try to recreate figure 2-5, as an exercise. Although an extreme example, this shows the extendability of our model and applicability to actual biological experiments. In a sense, the simulation examples presented are akin to the *in vitro* studies with purified phase separating proteins, the only difference being that system variables are much easier to change. Ideally, one would want to simulate all species in a cell and recreate life *in silico* but being able to reduce these complex systems to the core components gives a fundamental understanding of the systems, and is vastly easier and computationally feasible. If the length of monomers can be extended in the future, it would be informative to simulate existing point mutations studies of FUS, Ddx4, and hnRNPA1 within the confines of this simulator to assess the accuracy and also explore new interaction modes for the species. How do the these IDPs conform spatially once they are inside these droplets? Within these weak interactions between monomers, are some interactions more stable/likely to exist for a longer time than others? These are just a few questions which are easy to explore in these types of

simulations but tend to raise experimental complexity. In this way we have created an extendable model to explore the behaviour of possibly charged polymer chains *in silico*.

These weak electrostatic interactions in IDPs are shown to impart a liquid-like material state to phase separated systems leading us to hypothesize that this could explain the dynamic clustering and dissolution of RNAP in *E. coli*. [7, 9, 56] Using the property of liquid-like systems to be more dynamic compared to gels and solids we used single particle diffusion of RNAP to elucidate the nature of these clusters. There are two models to explain the dynamics we have seen with regards to RNAP: the proteins are bound directly to the DNA, and some of the proteins are bound while the others exist in a liquid-like phase around the region. [2] By comparing the diffusion profile of the RNAP to that of a DNA bound system we show in figure 3-2 A that even the slowest fraction of RNAP only slightly overlaps in diffusion coefficients with the numbers expected for tight DNA binding. We can then conclude that in well grown *E. coli* there exists at least three distinct fractions of RNAP: one tightly bound to DNA, one which diffuses in the surrounding media, and one which is exploring the space of the nucleoid.

Using our segmentation techniques and definitions of a local cluster in time, we show that single particle methods can be a robust method with which to study potential phase separated systems *in vivo*. Even though we are unable to make a distinct

[2] This might be the cell's attempt to increase binding rates for the transcriptional machinery.

conclusion regarding the material properties of these clusters [3] , and if they are indeed a result of phase separation acting in a prokaryotic system, we have shown that our single particle methods and analysis is able to accurately measure the dynamics of these systems. This provides future studies with an extendable experimental procedure which in combination with the *in silico* efforts allows experimenters to bounce between model construction and experimental confirmation effortlessly.

As stated before, the literature is rich in providing examples of these phase separated systems but very little concrete evidence exists for the functional need for them. In the nephrin example, the authors see nucleation of actin filaments through the Arp2/3 complex associated with the phase separating system, but the connection ends there. [24] Similar stories exist for stress granules where the associated proteins are all shown to phase separate but the functional role remains unclear. [30] It has been hypothesized that the enrichment of certain proteins in these systems allows the cell to increase reaction rates, effectively labeling these systems as reaction vesicles the cell can use dynamically. [7, 15] Similarly in *E. coli* the RNAP clusters can be seen as a dynamic system which facilitates fast reinitiation rates for RNAP to the DNA after termination, while these clusters dissociate in times when the cell does not need to maintain a high production of proteins (slow growth). We have not shown any evidence to support this but future studies on disrupting transcription through RNAP-DNA binding may show an effect on the diffusion profile of RNAP in these two growth conditions. Since RNAP and the transcriptional machinery contain very large

[3] As we haven't compared to known dynamics of these states

molecules it is unfeasible to fully simulate them using the simulator in Chapter 2, but coarse graining them into smaller essential regions may provide a good exercise to see if the RNAP polymer can self-associate with itself and if not, what modifications are needed for it?

This manuscript attempts to provide two methods, one *in silico* and one *in vivo*, which hope to identify key features of intrinsically disordered phase separating proteins and their material properties. We hope the protocols on single particle tracking of these molecules becomes a robust way to identify distinct phases in future studies on the topic, while the simulator allows future modellers to design synthetic polymers that may help in ALS and other potential phase separation associated diseases.

Alexander Oparin, the Soviet biochemist, had discussed the occurrence of organic coacervates in the primordial soup as an organizational tool in *The Origin of Life*, but this idea of coacervates in the living cell, forming and dissociating as liquids now leads us to revisit this idea. [57] We are only now starting to understand the role of phase separated compartments in modern life, but an argument can be made that early life consisted of protocells that behaved exactly as reaction vesicles created through phase separation. Preliminary work is being conducted on these protocells, but current efforts are far too simple and should be extended with future studies. [58]

References

[1] Brangwynne C. P., Mitchison T. J., Hyman A. A. 2011 Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proceedings of the National Academy of Sciences* **108**:4334–4339.

[2] Banani S. F., Lee H. O., Hyman A. A., Rosen M. K. 2017 Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* **18**:285–298.

[3] Berry J., Weber S. C., Vaidya N., Haataja M., Brangwynne C. P. 2015 RNA transcription modulates phase transition-driven nuclear body assembly. *Proceedings of the National Academy of Sciences* **112**:E5237–E5245.

[4] Banerjee P. R., Milin A. N., Moosa M. M., Onuchic P. L., Deniz A. A. 2017 Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. *Angewandte Chemie - International Edition* **56**:11354–11359.

[5] Parry B. R., Surovtsev I. V., Cabeen M. T., O'Hern C. S., Dufresne E. R., Jacobs-Wagner C. 2014 The bacterial cytoplasm has glass-like properties and is fluidized by metabolic activity. *Cell* **156**:183–194.

[6] Munder M. C., et al. 2016 A pH-driven transition of the cytoplasm from a fluid- to a solid-like state promotes entry into dormancy. *eLife* **5**:1–30.

[7] Hyman A. A., Weber C. A., Jülicher F. 2014 Liquid-Liquid Phase Separation in Biology. *Annual Review of Cell and Developmental Biology* **30**:39–58.

[8] Weber S. C. 2017 Sequence-encoded material properties dictate the structure and function of nuclear bodies. *Current Opinion in Cell Biology* **46**:62–71.

[9] Gomes E., Shorter J. 2019 The molecular language of membraneless organelles. *The Journal of biological chemistry* **294**:7115–7127.

[10] Hohenberg P., Halperin B. 1977 Theory of dynamic critical phenomena. *Reviews of Modern Physics* **49**:435–479.

[11] Doi M. 1996 *Introduction to Polymer Physics* (Clarendon Press).

[12] Doi M., Edwards S. 1986 *The Theory of Polymer Dynamics* (Oxford Science Publications).

[13] Doi M. 2013 *Soft Matter Physics* (OUP Oxford).

[14] Brangwynne C. P., Eckmann C. R., Courson D. S., Rybarska A., Hoege C., Gharakhani J., Jülicher F., Hyman A. A. 2009 Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **324**:1729–1732.

[15] Boeynaems S., Alberti S., Fawzi N. L., Mittag T., Polymenidou M., Rousseau F., Schymkowitz J., Shorter J., Wolozin B., Van Den Bosch L., Tompa P., Fuxreiter M. 2018 Protein Phase Separation: A New Phase in Cell Biology. *Trends in Cell Biology* **28**:420–435.

[16] Jain S., Parker R. 2015 *Ten years of progress in gw/p body research* (Springer).

[17] Mitrea D. M., Kriwacki R. W. 2016 Phase separation in biology; Functional organization of a higher order Short linear motifs - The unexplored frontier of the eukaryotic proteome. *Cell Communication and Signaling* **14**:1–20.

[18] Anderson P., Kedersha N., Ivanov P. 2015 Stress granules, P-bodies and cancer. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* **1849**:861–870.

[19] Kroschwald S., Maharana S., Mateju D., Malinovska L., Nüske E., Poser I., Richter D., Alberti S. 2015 Promiscuous interactions and protein disaggregases determine the material state of stress-inducible RNP granules. *eLife* **4**:1–32.

[20] Riback J. A., Katanski C. D., Kear-Scott J. L., Pilipenko E. V., Rojek A. E., Sosnick T. R., Drummond D. A. 2017 Stress-Triggered Phase Separation Is an Adaptive, Evolutionarily Tuned Response. *Cell* **168**:1028–1040.e19.

[21] Yoo H., Triandafillou C., Drummond D. A. 2019 Cellular sensing by phase separation: Using the process, not just the products. *The Journal of biological chemistry* **294**:7151–7159.

[22] Brangwynne C., Tompa P., Pappu R. 2015 Polymer physics of intracellular phase transitions. *Nature Physics* **11**:899–904.

[23] Feric M., Vaidya N., Harmon T. S., Mitrea D. M., Zhu L., Richardson T. M., Kriwacki R. W., Pappu R. V., Brangwynne C. P. 2016 Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell* **165**:1686–1697.

[24] Li P., et al. 2012 Phase transitions in the assembly of multivalent signalling proteins. *Nature* **483**:336–340.

[25] Posey A. E., Holehouse A. S., Pappu R. V. 2018 *Phase Separation of Intrinsically Disordered Proteins* (Elsevier Inc.) Vol. 611, 1 edition, pp 1–30.

[26] Harmon T. S., Holehouse A. S., Rosen M. K., Pappu R. V. 2017 Intrinsically disordered linkers determine the interplay between phase separation and gelation in multivalent proteins. *eLife* **6**:1–31.

[27] Schmidt H. B., Görlich D. 2016 Transport Selectivity of Nuclear Pores, Phase Separation, and Membraneless Organelles. *Trends in Biochemical Sciences* **41**:46–61.

[28] Weber S. C., Brangwynne C. P. 2012 Getting RNA and protein in phase. *Cell* **149**:1188–1191.

[29] Boke E., Mitchison T. J. 2017 The balbiani body and the concept of physiological amyloids. *Cell Cycle* **16**:153–154.

[30] Patel A., et al. 2015 A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **162**:1066–1077.

[31] Murthy A. C., Dignon G. L., Kan Y., Zerze G. H., Parekh S. H., Mittal J., Fawzi N. L. 2019 Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nature Structural & Molecular Biology* **26**.

[32] Burke K. A., Janke A. M., Rhine C. L., Fawzi N. L. 2015 Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Molecular Cell* **60**:231–241.

[33] Meshwork F. T.-d. 2006 FG-Rich Repeats of Nuclear Pore Proteins with Hydrogel-Like Properties. **314**:815–817.

[34] Pak C. W., Kosno M., Holehouse A. S., Padrick S. B., Mittal A., Ali R., Yunus A. A., Liu D. R., Pappu R. V., Rosen M. K. 2016 Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein. *Molecular Cell* **63**:72–85.

[35] Moesa H. A., Wakabayashi S., Nakai K., Patil A. 2012 Chemical composition is maintained in poorly conserved intrinsically disordered regions and suggests a means for their classification. *Molecular BioSystems* **8**:3262–3273.

[36] Zarin T., Tsai C. N., Nguyen Ba A. N., Moses A. M. 2017 Selection maintains signaling function of a highly diverged intrinsically disordered region. *Proceedings of the National Academy of Sciences* **114**:E1450–E1459.

[37] Tsuiji H., Iguchi Y., Furuya A., Kataoka A., Hatsuta H., Atsuta N., Tanaka F., Hashizume Y., Akatsu H., Murayama S., Sobue G., Yamanaka K. 2013 Spliceosome integrity is defective in the motor neuron diseases ALS and SMA. *EMBO Molecular Medicine* **5**:221–234.

[38] Wurtz J. D., Lee C. F. 2018 Chemical-Reaction-Controlled Phase Separated Drops: Formation, Size Selection, and Coarsening. *Physical Review Letters* **120**:78102.

[39] Jacobs W. M., Frenkel D. 2017 Phase Transitions in Biological Systems with Many Components. *Biophysical Journal* **112**:683–691.

[40] Schwarz U. 2013 Theoretical Statistical Physics. *HEIDELBERG UNIVERSITY*.

[41] Gibbs J. W. 1878 On the equilibrium of heterogeneous substances. *American Journal of Science* **s3-16**:441–458.

[42] Nott T., et al. 2015 Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell* **57**:936–947.

[43] Silva J., Gallo C., Costa D., Rangel L. 2014 Prion-like aggregation of mutant p53 in cancer. *Trends in Biochemical Sciences* **39**:260–267.

[44] Das S., Eisen A., Lin Y. H., Chan H. S. 2018 A Lattice Model of Charge-Pattern-Dependent Polyampholyte Phase Separation. *Journal of Physical Chemistry B* **122**:5418–5431.

[45] Hastings W. 1970 Monte carlo sampling methods using Markov chains and their applications. *Biometrika* **57**:97–109.

[46] Lindahl E., Hess B., Spoel D. 2001 GROMACS 3.0: A package for molecular simulation and trajectory analysis. *J. Mol. Mod.* pp 306–317.

[47] Fyta M. 2016 *Computational approaches in physics* Vol. 4, pp 1–137.

[48] Al-Husini N., Tomares D. T., Bitar O., Childers W. S., Schrader J. M. 2018 α-Proteobacterial RNA Degradosomes Assemble Liquid-Liquid Phase-Separated RNP Bodies. *Molecular Cell* **71**:1027–1039.e14.

[49] Endesfelder U., Finan K., Holden S., Cook P., Kapanidis A., Heilemann M. 2013 Multiscale Spatial Organization of RNA Polymerase in Escherichia coli. *Biophys J.* pp 172–181.

[50] Jin D. J., Mata Martin C., Sun Z., Cagliero C., Zhou Y. N. 2017 Nucleolus-like compartmentalization of the transcription machinery in fast-growing bacterial cells. *Critical Reviews in Biochemistry and Molecular Biology* **52**:96–106.

[51] Stracy M., Lesterlin C., Garza de Leon F., Uphoff S., Zawadzki P., Kapanidis A. N. 2015 Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proceedings of the National Academy of Sciences* **112**:E4390–E4399.

[52] Bakshi S., Siryaporn A., Goulian M., Weisshaar J. 2012 Superresolution imaging of ribosomes and RNA polymerase in live Escherichia coli cells. *Molecular Microbiology* **81**:21–38.

[53] Weng X., Bohrer C., Bettridge K., Lagda A., Cagliero C., Jin D., Xiao J. 2018 RNA polymerase organizes into distinct spatial clusters independent of ribosomal RNA transcription in E. coli .

[54] Tinevez J. Y., Perry N., Schindelin J., Hoopes G. M., Reynolds G. D., Laplantine E., Bednarek S. Y., Shorte S. L., Eliceiri K. W. 2017 TrackMate: An open and extensible platform for single-particle tracking. *Methods* **115**:80–90.

[55] Pedregosa F., et al. 2011 Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**:2825–2830.

[56] Yi-Hsuan L., Jacob P. B., Julie D. F.-K., Hue S. C. 2017 Charge pattern matching as a 'fuzzy' mode of molecular recognition for the functional phase separations of intrinsically disordered proteins. *New J. Phys* **19**:115003.

[57] Oparin A. I. 1959 *Proceedings of the First International Symposium on the Origin of Life on the Earth, held at Moscow 19-24 August 19578* (Pergamon Press).

[58] Zwicker D., Seyboldt R., Weber C. A., Hyman A. A., Jülicher F. 2017 Growth and division of active droplets provides a model for protocells. *Nature Physics* **13**:408–413.