Self-assembling nanoscale systems

James A. Geraets

Doctor of Philosophy

University of York

Biology

September 2015

Abstract

Self-assembly is ubiquitous in different areas of science, for example in crystals and viruses, and also plays crucial roles in nanotechnology. Many commonalities link these self-assembling systems, in spite of their complexity and different length and time scales. In this thesis, we take an interdisciplinary perspective to gain new insights into self-assembly, exploring ways of modelling self-assembling systems that are relevant across these different fields.

A challenge in nanotechnology is to develop self-assembling systems capable of generating a desired outcome. An example is graphene nanoribbons, which are a novel type of semiconductor material with great potential in the nanotech industry. In this context, it is unclear which strategies are best for controlling the output of a self-assembly process, either by manipulation of the thermodynamic environment of the assembling system, or other methods of directing self-assembly. We use quantitative modelling of the kinetics of self-assembly as a tool to predict experimental results in self-assembling systems that are too complex for detailed experimental investigation.

Self-assembly of viral protein shells is an example from biology. Viruses have evolved niche methods of assembly that are both robust and highly efficient, as the virus mutation rates are very high, especially in RNA viruses. The viruses discussed in this thesis have an added layer of complexity; it is thought that sequence-specific interactions between viral genomes and the protein building blocks of the viral capsids have a strong impact on the assembly process. We have developed here novel analysis techniques for the modelling of this co-assembly scenario. We use these mechanistic insights to develop new theoretical tools to analyse structural data, providing unprecedented insights into the asymmetric organization of the packaged genome.

Contents

List of Tables

List of Figures

To my parents Jane and Yavuz, brother Joel, sister Deniz, and partner Mercedes, who have all been steadfast in their support of me throughout my time in research.

Acknowledgements

Firstly, I would like to thank my supervisor from the Department of Physics, Dr Yvette Hancock, for all her help, enthusiastic discussions, and mentoring throughout my time in York; I am very grateful for having had the opportunity of working together on some really captivating research. It has been a wonderful three years in the department, researching alongside fellow postgraduate students Jack, Rebecca, Raquel, Andre, and Sam. I will miss our time together—from our group meetings and scientific discussions, to our extremely frequent caffeination breaks, many of which were facilitated by an espresso machine next to my desk.

I will also miss the coffee breaks (and colleagues, naturally) in the York Centre for Complex Systems Analysis (YCCSA), where I was fortunate enough to have another espresso machine adjacent to my desk! In particular, I would like to mention Emilio, Nick, Eric, Richard, Pierre, Giuliana, Tom, Eva, Jenny, Chris, Adam, Simon, Alex, Motiejus, and Frances, all of whom have been a part of the Mathematical Virology group at some time or another. It has been a real pleasure spending time with you all.

The support of staff from YCCSA and the Departments of Biology and Physics has been outstanding. I heartily thank my colleagues in these centres, including Julie, Anne, Darren, Annette, Sarah, Caryn and Lydia. I couldn't have survived in the dog-eat-dog world of a university campus without your constant help with administrative problems, and help with PhD life in general.

I must also thank the innumerable teaching staff (too many to name) that I have had the pleasure of facing the undergraduate students with! Teaching has been a real highlight of my time at York: it's been a genuinely rewarding experience, and I'm very glad that I've spent time contributing to the university community in this way.

Also to be acknowledged here is the crucial help and guidance I have received from my thesis advisor panel members, Dr Stephen Cowling and Prof. Mark Leake. It has been particularly great spending time focussing on my development: I have actually looked forward to my formal progression meetings with pleasure. Their counsel and impartial advice has made the process of completing my PhD much less stressful, and has helped me find gainful employment quickly!

Thanks so much to my collaborators Dr Neil Ranson, Prof. Peter Stockley, Prof. Arwen Pearson, Dr James Ross, Dr Kyle Dent, and all the other guys from the Astbury Centre for Structural Molecular Biology at the University of Leeds for hosting me as a visitor. Thanks also to Prof. John Goodby, Dr Isabel Saez, and Prof. Peter Knight from the Liquid Crystal research group, based in the Department of Chemistry, University of York, for fruitful discussions about graphene and graphene nanoribbon synthesis.

A huge vote of thanks to all those who have had to tolerate my presence in the lab, and who have trained me on various aspects of laboratory work. Although the wet laboratory didn't turn out to be the primary aspect of the doctoral project (and I don't know what that says about my laboratory skills), I am very grateful for the time and attention: I am currently putting the expertise into practice in a virology lab. These patient tutors are Dr Marika Kullberg, Dr Dimitris Lagos, and Dr John Moore from the University of York, as well as Dr Pradeep Luther at Imperial College, and Dr Peiyi Wang and Martin Fuller at

the University of Leeds.

Most especial thanks go to Prof. Debbie Smith, Prof. Paul Kaye, Dr Leo Caves, Prof. Fred Antson, members of the CIDCATS programme executive committee, and others involved in the administration of the CIDCATS doctoral training programme. I have had a wonderful experience on the programme, and have come a long way since the start in 2011! Many thanks to Fred for being my mentor throughout my time in York. I must also credit here all of my fellow CIDCATS students for being such great sports and inspiring colleagues, especially Ahmad, Angela, German, and William from my cohort. The many fruitful discussions we have shared shall always be remembered, as will the friendship and camaraderie.

I must thank and recognize the Wellcome Trust for funding this doctoral research, in the form of research and travel expenses, and a doctoral studentship (number 097326/Z/11/Z). Additionally I gratefully acknowledge supplementary funding for travel from the Institute of Physics; specifically from the Research Student Conference Fund and the C. R. Barber Trust Fund.

Finally; the deciding factor in my decision to come to York for my doctoral research was the prospect of working with the inspirational Prof. Reidun Twarock, who has been a huge support and a great friend. Thank you, Reidun, for all your time and effort in supervising my research.

Declaration

This thesis has not previously been accepted in substance for any degree and is not being concurrently submitted in candidature for any degree other than Doctor of Philosophy of the University of York.

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by explicit references. Some of the material presented within this thesis forms part of the following papers/manuscripts:

I. J. A. Geraets, E. C. Dykeman, P. G. Stockley, N. A. Ranson, and R. Twarock, "Asymmetric genome organization in an RNA virus revealed *via* graph-theoretical analysis of tomographic data," *PLoS Computational Biology*, vol. 11, no. 3, p. e1004146, 2015.

II. J. A. Geraets, J. Baldwin, R. Twarock, and Y. Hancock, "A proposed method for directed self-assembly of graphene nanoribbons," in prep.

I hereby give consent for my thesis, if accepted, to be made available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organizations.

Chapter 1

Introduction

Self-assembly is defined as the spontaneous formation of organized structures by specific interactions between pre-existing components. Spanning scales ranging from atomic to macroscopic, self-assembling systems systematically generate ordered structures in biological, chemical and material formation processes [1–3]. From the interaction of often simple components, emergent complex assembly behaviour can arise dependent on the system characteristics, although interestingly the physical manifestation of many transpiring self-assembly phenomena can be observed in parallel across these different fields [3–5]. Within the scientific community there is currently research into diverse examples of self-assembly [3], including DNA[1] cages [6, 7], viruses [8–12], quasicrystals [13, 14], and graphene [15, 16]. Understanding the behaviour and mechanisms of self-assembling systems therefore lends itself to an interdisciplinary collaboration between multiple disciplines, with results impacting on biology, chemistry, engineering, computer science and mathematics.

Future nanotechnology development is predicated on the development of highly-engineered self-assembling systems: *bottom up* rather than *top down*

[1]For all abbreviations see the glossary on page 222.

methods of manufacture have several advantages in the production of nanoscale devices [1, 5], and it is perhaps inevitable that future nanotechnology will come to rely on atomically-precise novel assembly methods, as the properties of the materials are highly sensitive to disorder [3, 17, 18]. Indeed, self-assembly has an intrinsic advantage over mechanically-directed forms of assembly, as it requires no machinery to move, orient and combine the components. Instead selective binding between uniquely matching surfaces, driven by stochastic interactions, brings the components together. Additionally, the "hands off" nature of self-assembly lends itself to scalable production, and thus there is the potential for quick easy dissemination of successful applications of self-assembly to a wide range of technology.

Of course, the challenge when designing self-assembly systems is that the structure of the components must somehow encode the final combined structure [3, 15]. This necessity naturally increases the complexity of the designed parts, which is a major constraint on their fabrication. Another consideration in supramolecular self-assembly is the weakened partitions between the internal interfaces in the final structure, that have little or no operational function[2]: unless they are subsequently strengthened, the combined product will have to tolerate this weakness.

Overcoming these difficulties and limitations is essential for realizing the promise of self-assembly for the synthesis of new materials and technology. Even the design of the requisite highly-specific interactions that enable self-assembly is a challenging task. It is also unclear which strategies are best for externally directing the self-assembly process, by controlling the thermodynamic environment of the assembling system to regulate its output. We thus look to nature for

[2] A counter-example from virology is that genome release can occur through a widening of these partitions, creating pores, rather than a full decapsidation; yet external conditions would precipitate this structural change [19]. These partitions also facilitate structural transformations in the capsid lattice tiling [20, 21].

inspiration [5]: specifically, as detailed in this thesis, to the world of viruses in order to gain an understanding of the physical processes that underpin successful self-assembling systems.

In viruses, nature has evolved such niche systems—which are highly optimized for efficiency and fidelity of assembly—due to the immense evolutionary and competitive pressure that these infectious agents undergo [8, 11]. As the viral cycles of infection are very short, and mutation rates are generally very high, especially in RNA viruses [22], the search space for optimizing the assembly process far exceeds what is possible for engineers and chemists to achieve, or indeed even to experiment with, in the context of the laboratory environment.

Fortunately, by a synthesis of a deep technical understanding of the physical principles that govern successful self-assembly with computational modelling, it is possible to explore generalized models of self-assembly in the non-virus setting [11, 23]. In this thesis, an example is shown with the use of a kinetic self-assembly (KSA) algorithm, namely the Gillespie algorithm [24, 25], coupled with a new treatment of molecules within the simulation, to model the assembly of nanographene—specifically porous graphene nanoribbons (GNRs)—from carbon-based precursor molecules. With this theoretically-driven approach, we demonstrate that inroads can be made into the development of new paradigms for the production of novel materials.

But self-assembly is not just an enabling methodology for the nascent nanotechnology industry. In fact, the science of self-assembly can also be applied to understanding more about the natural world, understanding biological and physical phenomena, and perhaps ultimately understanding the origin of life itself[3].

[3]Of interest is the "RNA world" hypothesis of abiogenesis, in which self-replicating RNA molecules on the primordial Earth are proposed as the precursors to all known life: furthermore, scenarios have been envisaged in which self-assembling ribonucleoproteinoid virus-like particles could have predated cells [26].

> Καὶ τὸ ὅδον τοῦ μέρους μεῖζον [ἐστιν].
>
> And the whole [is] greater than the part.
>
> (Euclid, *Elements*, Book I, Common Notion 5)

With systems thinking, we can examine the interconnectedness or *systemicity* of complex systems (such as self-assembling systems), so that commonalities in their operation can be inferred [27, 28]. The systems approach advocates consideration of the holistic characteristics of a complex system, rather than the details of their components[4].

At some boundary condition (up to and including the universe) all physical systems are integrated, and science cannot embrace the total complexity of this interconnectedness. Consequently, science operates with models of various kinds—from *in vivo* experimentation to computer simulation—which invariably encapsulate idealizations and abstractions. Observable statements about these systems will have inferential relations to other statements, and systemicity is a quantification of the number of these inferential relations. One of the most important inferential relations is derivability: that physical laws become derivable from a few fundamental laws.

The concept of *emergence* posits that once a given system is of a sufficient complexity, it can exhibit phenomena that cannot be deduced solely from the small-scale behaviour of the individual system components. However, once a complex behaviour has emerged, reasons for that behaviour can often be drawn from examination of the component properties. Additionally, computer simulation of the local rules of a system can be successfully used to recover the emergent phenomena of complex systems. An example emergent behaviour is that of water: it exhibits wetness, which cannot be predicted solely on the basis of the

[4]For an interesting examination of systems thinking, the work of Bogdanov is recommended [29].

properties of hydrogen and oxygen. Another example is temperature: from the consideration of the fundamental laws of physics, temperature as a macroscopic observable measurement can be considered an emergent phenomenon. Similarly, emergent phenomena are present in self-assembling systems, and characterization of these behaviours is crucial when seeking to perturb or mimic the systems, to ensure that the desired result is achieved.

If seeking to classify systems by emergent observables, and determine similarities about their mechanisms, the internal systemicity of each system can be used as a comparator with other systems [30]. Any similarity in the emergent phenomena of systems can be used to infer shared mechanisms and component behaviours in these systems. It can be extrapolated that reconstructions with partial components of an interconnected system can recreate the behaviour of the whole. This idea is portrayed by way of a hologram as an analogy, insomuch as a partial hologram encodes a reconstruction of the entire system. Thus, desirable behaviour at a system-wide level can be achieved by examination of similar systems that exhibit the desired outcome.

This being said, systemicity and emergence are partially irreconcilable: emergent phenomena cannot be fully explained by examination of their components, as we are simply not capable of understanding hierarchical systems at the level of their most basic constituents. Furthermore, these complex nonlinear systems cannot normally be fully simplified, decomposed or generalized without losing their essential characteristics, though redundancy in their subprocesses can be used to simplify their description [31]. This concept can be connected to a quantitative measure of the complexity of a system [32,33]: the amount of information needed in order to fully describe the system. We should view investigation into different self-assembly processes, particularly with regards to biological systems, apropos of this consideration.

For physicists, it is all too easy to ultracrepidate[5] and criticize biologists for their perceived lack of interest in developing a reductive framework for understanding phenomena, and not pursuing the commonalities in biological mechanisms in particular. However, reductionism has limitations; knowing chemistry does not mean that we understand life. A reductionist approach also disregards the wide enriched "ecosystem" of complex systems present in the natural world, from which much can be learnt and challenge our existing understanding of systems and system engineering.

Reductionism aside, there are common features in self-assembling systems that can be studied without loss of generality on an abstract level. Chapter 2 is devoted to a review of such approaches. At the end, we identify two important problems that require further theoretical developments:

(i) novel approaches to tackle the complexity of the reactions between self-assembly building blocks, and

(ii) the integration of other components important for assembly efficiency.

My contributions to (i) will be covered in Chapter 3 and illustrated with applications to nanographene in Chapter 4. (ii) is pertinent to self assembly in virology, which will be the topic of Chapters 5–7, where I am presenting tools I have developed to integrate interactions of the self-assembly protein building blocks with the viral genome. This is described for two viruses in particular, bacteriophage MS2 and satellite tobacco necrosis virus (STNV), for which specific stem-loops, packaging signals (PSs), within the genome mediate the assembly of the capsid into an infectious virion [8–10,34–36]. Investigation of the evolved assembly methodologies of these and other single-stranded RNA (ssRNA) viruses

[5]To go beyond one's scope or province, especially to criticize beyond one's sphere of knowledge. From the Latin *ultra crepidam*, literally "beyond the sandal", alluding to the response of the Greek painter Apelles to a cobbler's criticism.

reveals many similarities in the strategies and mechanisms exploited. PSs have been identified in many other viruses, also in the form of a defined element of secondary structure, such as a stem-loop or collection of stem-loops [37–53]. During assembly, interactions between PSs and capsid protein (CP) building blocks effect conformational changes in the capsid, and thus enable their efficient assembly. They also control the order of assembly: assembly pathways that ensure fidelity and efficiency of completion are thus preferably selected by the virus. Perhaps the most important feature of RNA-centric virus assembly, this is responsible for their ability to package their genome selectively, even in competition with a background of cellular RNA molecules.

In the conclusion (Chapter 8), I will discuss how mimicking nature, in particular viruses, can result in new developments in nanographene synthesis and other areas of nanotechnology. There are clear commonalities between all self-assembling systems: this thesis is about breaking new ground *via* a synergistic consideration of these systems. Although the mechanisms exploited by the viruses far surpass what is currently transferable to designed systems, a better understanding of the specifics of viral assembly is highly desirable due to the significant impact of viruses on human and animal health. The physico-chemical properties of the viral components mediate basic biological function in ways that are not fully understood, and have implications as to their structure and dynamics in the cellular environment. As we have discussed, the knowledge gained by experimental and theoretical consideration of these highly evolved systems can also provide insight into the design of small components for self-assembly of novel materials.

Chapter 2

Kinetics of self-assembly

Self-assembling systems are processes that create incrementally complex hierarchical spatial organizations, the shape of which are entirely derived from the local rules between the self-assembling components. For any chosen set of components, a lexicon of local rules describes the decentralized interactions occurring concurrently at several different time and space scales, which together dynamically produce the self-assembled structure. As discussed in Chapter 1, the emergence of global structure and pattern cannot be deduced from the individual composing elements alone. However, simulation models can be used to gain deeper insights into these complex self-assembling systems. Discussed in this chapter are the principles of the modelling and simulation of self-assembling systems from their local rules, as well as approaches to representing the underlying space and handling of complex spatial structures formed within this space.

2.1 Theory of chemical kinetics

What exactly happens in a chemical reaction when the reactants are turned into products? Given the physical laws underpinning the existence of matter and energy, much of the field of chemistry is dedicated to understanding the

mechanisms by which reactions occur, and how the reactions depend on their environment. Of particular interest in this thesis is the application of kinetics to the optimization of chemical synthesis, manufacturing and engineering.

2.1.1 Arrhenius equation

A large portion of the practical consideration of chemical kinetics uses the *Arrhenius equation*, which provides a conceptual framework in which to interpret kinetic data by relating the rate constant, k, of a reaction to the absolute temperature, T:

$$k = k_0 e^{\frac{-E_a}{k_B T}} \tag{2.1.1}$$

with k_B as the Boltzmann constant, and E_a is the activation energy [54].

The Arrhenius equation is not based on a *first principles* analysis of chemical kinetics, but rather is a description of the relation of kinetics to system variables [55]. It allows experimental data from different reactions to be compared, and can be used predictively (as demonstrated in Chapter 4) by combining experimental guidance with theoretical insight.

Arrhenius made the analysis that underpins the field of kinetics when he postulated that the chemical reaction of molecules was not possible for all reactant molecules, but rather only those that possessed a certain minimum energy, known as activation energy [56], and therefore E_a is critical in the analysis of Equation (2.1.1). In Brownian motion at thermodynamic equilibrium, the fraction of molecules possessing this critical energy can be calculated from the Maxwell-Boltzmann distribution:

$$f(v) = \sqrt{\left(\frac{m}{2\pi kT}\right)^3} 4\pi v^2 e^{-\frac{mv^2}{2kT}} \tag{2.1.2}$$

where v is the velocity and m is the mass of the molecule.

According to equipartition, the total average energy for molecules described by the Maxwell-Boltzmann distribution is distributed equally among three types of energy:

(i) translational energy of the molecule in space,

(ii) rotational and vibrational energy, and

(iii) electronic energy.

In bi- and ter-molecular reactions the accumulation of translational energy has the most impact on increasing the speed of reactions. Conversely, for intramolecular reaction mechanisms an increase and redistribution of vibrational energy across the normal modes of vibration is preferable [56]; an alteration of rotational states often accompanies this redistribution [56]. Molecules can only react when they have acquired the activation energy necessary for that reaction to commence, but they do not react immediately upon becoming activated. The activated molecules possess a finite lifetime during which they can either react or become deactivated [56]. Proceeding to a reaction can be usually described as a collisional process, as can the activation and deactivation steps in energy, in which the chance of a reaction occurring depends on the cross-section of the reactants. In essence, the entire mechanism of molecular interactions and reactions are underpinned by thermal Brownian collisional processes.

2.1.2 Energy minimization

Self-assembly is the outcome of a random collisional motion of molecules in conjunction with the affinity of the binding sites they share for each another. For self-assembly to be exploited to manufacture new materials for the nanotech industry, it is critical to formulate simple and efficient means of organizing molecules and clusters of molecules into precise, pre-determined structures.

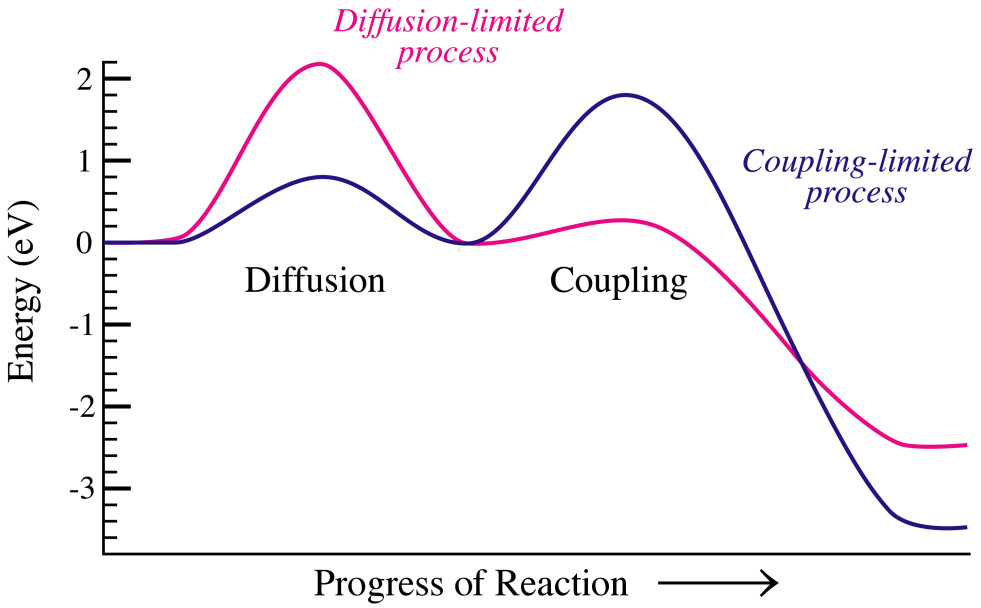


Figure 2.1: Coupling-limited and diffusion-limited reactions: coupling-limited reactions face a greater energy barrier to bond formation than to diffusion, whereas reactions that occur on the same order as diffusion or more easily are likely to be diffusion-limited.

From a thermodynamic perspective, self-assembly is driven by a principle of *energy minimization* [57–59]. In the example systems discussed in this thesis, it is the Gibbs free energy state space that is sampled by the systems (due to constant temperature and pressure), eventually making their way to a local or global energy minimum. The driving force for the sampling of the state space is thermal noise: random Brownian motion of molecules in a coupling-limited regime allows energy minimization to be predominant in the process of forming structures (Figure 2.1). This minimization of Gibbs free energy can be attained by a maximization of the number of molecular interactions [59].

Thermodynamically, the self-assembly process is described by the Gibbs free energy equation [60, 61]:

$$\Delta G = \Delta H - T\Delta S \tag{2.1.3}$$

The Gibbs free energy, ΔG, determines whether the self-assembly occurs

spontaneously or not: if ΔG is negative then the process is spontaneous. ΔH denotes the calculated enthalpy change of the self-assembly process, corresponding to the potential energy and intermolecular forces between the assembling units before and after the reactions. The change in entropy, ΔS, associated with the formation of order within the assembly is normally highly negative for a self-assembly reaction as the self-organization increases order. Thus in order for ΔG, and hence the assembly to occur spontaneously, the enthalpy term must both be negative and in excess of the magnitude of the entropy term. Equation (2.1.3) determines a critical temperature, for which self-assembly will not occur spontaneously; more generally it predicts that as the magnitude of $T\Delta S$ nears that of ΔH, self-assembly is progressively less likely to manifest.

Usually, the magnitude for the enthalpic and entropic terms are finely balanced, so that non-ideal organizations can be rearranged and thus the global free energy minimum reached [62]. In most cases relatively weak intermolecular interactions provide the thermodynamic driving force, with background thermal energy present to allow escaping from non-ideal local minima. However, it is still possible to create self-assembled structures with strong intermolecular interactions, though this will have implications for the number of defects and the ability of the system to self-correct *kinetic traps*: organizations for which continued assembly occurs extremely slowly [63].

2.1.3 Nucleation and growth

At the initiation of a self-assembly reaction, the precursor molecules nucleate into small assemblies as the self-assembled growth commences. These small assemblies are seen to form due to a lowering of the Gibbs free energy in their aggregative state, which results in an increased lifetime [54, 62]. As the reaction continues, and more molecules are recruited into the assemblies, the Gibbs free

energy further decreases until the assembly stabilizes over long timescales.

Depending on the concentration of precursor molecules and assembly particulates in the system, and additionally the speed of assembly (arising from the balance of equilibrium), the normal self-assembly growth will either be monomeric or oligomeric; *i.e.* assembly proceeds solely *via* the addition of single precursors, or else larger assemblies can form through the binding together of smaller assemblies of precursors. In this thesis we discuss examples of both of these scenarios of self-assembly: the dominant mechanism for each depends on the types of interaction present within the self-assembling system, and the environment in which assembly occurs.

2.1.4 Defects

For any self-assembly system, a reasonable concentration of defects can be expected to occur as the structure assembles [64]. These defects arise from non-optimal binding of molecules onto the assembly, by introducing disruptions into the self-assembly pattern. Some chemical systems have also been known to be self-correcting, such as graphene sheets [65], filling any gaps using a non-self-assembly process. However, the corrections seen in these systems do not tend to seamlessly preserve the pristine structure produced *via* self-assembly. Instead, to ensure seamless integration into the assembly, self-assembly reactions must be engineered to correct their own defects (or minimize the occurrence of defects).

Such correction of defects is possible during self-assembly reactions, especially if the forward and backward rates of assembly are finely balanced: we say that equilibration of the assembly is required [64]. If components join together irreversibly upon a collision, they often are unable to form into a regular structure. The ability for components to adjust their relative positions to each other once in an aggregate is the simplest way of ensuring the minimum free

energy configuration is reached. In order that a self-assembling system is able to equilibrate in this way, there must be sufficient thermal energy to allow mass transport of the components [54]. The effect of defects can also be mitigated by templating the self-assembly process; the use of boundaries and surfaces can reduce defects and control structures [64].

Self-correction of defects is of utmost importance in virology, as incompletely formed viruses can be immunogenic and trigger a host response [11, 66]. It is therefore evolutionarily favourable for the virus to be efficient at correcting assembly defects.

2.2 Simple assembly model

2.2.1 Assembly of dimers

The simplest example of assembly considers two species, A and B, of which B is a dimer formed of two A molecules: A_2. The simple reactions that govern the process are:

$$\begin{aligned} \mathrm{A} + \mathrm{A} &\xrightarrow{k_1} \mathrm{B} \\ \mathrm{B} &\xrightarrow{k_2} \mathrm{A} + \mathrm{A} \end{aligned} \tag{2.2.1}$$

and are normally written as:

$$\mathrm{A} + \mathrm{A} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{B}\ . \tag{2.2.2}$$

We can consider the reactions to be *elementary*, in that they have a single transition state, and thus no stable intermediate states.

The *Law of Mass Action* is an empirical finding on which kinetic theory is based, and specifies that the rate of an elementary reaction is proportional to

the concentrations of the reactants, raised to the powers of the stoichiometric coefficients [54]. The law holds independently of other concentrations and reactions occurring. Under the Law of Mass Action, the rate equations relating the populations of A and B, namely n_A and n_B, with the rates of the forward and reverse reactions as r_1 and r_2, are:

$$\begin{aligned} r_1 &= k_1 n_A^2 \\ r_2 &= k_2 n_B \end{aligned} \tag{2.2.3}$$

$$\begin{aligned} \frac{\mathrm{d}n_A}{\mathrm{d}t} &= -2k_1 n_A^2 + 2k_2 n_B \\ \frac{\mathrm{d}n_B}{\mathrm{d}t} &= -k_2 n_B + k_1 n_A^2. \end{aligned} \tag{2.2.4}$$

In the steady state, the derivatives of n_A and n_B with respect to time will be zero:

$$\frac{\mathrm{d}n_A}{\mathrm{d}t} = \frac{\mathrm{d}n_B}{\mathrm{d}t} = 0 \tag{2.2.5}$$

and thus the equilibrium constant, K_c is:

$$K_c = \frac{n_B}{n_A^2} = \frac{k_1}{k_2}. \tag{2.2.6}$$

We assume that nothing is added to or removed from the tank during the synthesis process. Thus the number of molecules, n, is constant throughout the simulation, leading to:

$$n_A + 2n_B = n \tag{2.2.7}$$

$$\frac{\mathrm{d}n_A}{\mathrm{d}t} + 2\frac{\mathrm{d}n_B}{\mathrm{d}t} = 0 \tag{2.2.8}$$

as expected.

These calculations hold for rate equations where the order of the reaction for each reactant is equal to its stoichiometry, *i.e.* elementary reactions.

2.2.2 Configurational entropy

Consider a well-mixed homogeneous system consisting of M uniform cells, through which the molecules are scattered, one per cell: the size of each cell is small enough such that this is the case. The combination of possible locations of A in M, or W_A is [67, 68]:

$$W_A = \binom{M}{n_A} = \frac{M!}{n_A!(M-n_A)!}. \tag{2.2.9}$$

For the molecule B, the remaining number of cells is $M - n_A$. Combinations through this space is:

$$W_{B|A} = \binom{M-n_A}{n_B} = \frac{(M-n_A)!}{n_B!(M-n_A-n_B)!}. \tag{2.2.10}$$

Thus the total combinations are:

$$W_{A,B} = \binom{M}{n_A, n_B} = W_A W_{B|A} \tag{2.2.11}$$

$$= \frac{M!(M-n_A)!}{n_A!n_B!(M-n_A)!(M-n_A-n_B)!} = \frac{M!}{n_A!n_B!(M-n_A-n_B)!}. \tag{2.2.12}$$

The configurational entropy, *via* Boltzmann's entropy formula, is [69, 70]:

$$S = k_B \ln W_{A,B}. \tag{2.2.13}$$

2.2.3 Partition function

The partition function is given as:

$$Z = \sum_{n=n_A+2n_B} W_{A,B} Z_1^{n_A} Z_2^{n_B}. \tag{2.2.14}$$

The energy of the dimer B is set relative to a pair of monomers, as $-\epsilon$. Thus $Z_1 = e^{-\beta 0} = 1$ and $Z_2 = e^{\beta\epsilon}$, where the thermodynamic β is defined as $\frac{1}{k_B T}$.

We assume a single term dominates.

$$R = \ln Z = \ln W_{A,B} + n_A \ln Z_1 + n_B \ln Z_2 \tag{2.2.15}$$

$$= \ln M! - \ln n_A! - \ln n_B! - \ln (M - n_A - n_B)! + n_A \times (0) + n_B \beta\epsilon \tag{2.2.16}$$

Using Stirling's approximation [71, 72]:

$$\ln N! \approx N \ln N - N \tag{2.2.17}$$

we thus obtain:

$$\begin{aligned} R = [M \ln M - M] - [n_A \ln n_A - n_A] - [n_B \ln n_B - n_B] \\ -[(M - n_A - n_B) \ln (M - n_A - n_B) - (M - n_A - n_B)] + n_B \beta\epsilon. \end{aligned} \tag{2.2.18}$$

Therefore,

$$\frac{\partial R}{\partial n_A} = \ln (\frac{M - n_A - n_B}{n_A}) \tag{2.2.19}$$

$$\frac{\partial R}{\partial n_B} = \ln (\frac{M - n_A - n_B}{n_B}) + \beta\epsilon. \tag{2.2.20}$$

We then maximize $R + \lambda(n_A + 2n_B - n)$ as follows:

$$\frac{\partial R + \lambda(n_A + 2n_B - n)}{\partial n_A} = 0,$$

implying

$$\ln\left(\frac{M - n_A - n_B}{n_A}\right) + \lambda = 0$$

and thus

$$\frac{n_A}{M - n_A - n_B} = e^{\lambda}. \tag{2.2.21}$$

Moreover,

$$\frac{\partial R + \lambda(n_A + 2n_B - n)}{\partial n_B} = 0,$$

implying

$$\ln\left(\frac{M - n_A - n_B}{n_B}\right) + \beta\epsilon + 2\lambda = 0$$

and thus

$$\frac{n_B}{M - n_A - n_B} = e^{2\lambda} e^{\beta\epsilon}. \tag{2.2.22}$$

Define the concentrations C_A and C_B as $C_A := n_A/M$ and $C_B := n_B/M$. Then we find, eliminating e^{λ}:

$$\frac{1}{1 - C_A - C_B} \frac{C_A^2}{C_B} = e^{-\beta\epsilon}. \tag{2.2.23}$$

In the case that there is a sufficiently large number of cells, we have $M \gg A, B \Rightarrow C_A, C_B \ll 1$. Thus:

$$\frac{C_A^2}{C_B} = e^{-\beta\epsilon}, \tag{2.2.24}$$

and finally:

$$\frac{k_2}{k_1} = e^{-\beta\epsilon}. \tag{2.2.25}$$

These expressions can be used for modelling dimerization in self-assembly, thus simplifying the calculation of rates in elementary reactions. These expressions can be generalized for modelling the self-assembly of larger systems, for example the viral capsid from its protein building blocks [73]. Equation (2.2.25)

shows that for a given forward rate constant, k_1, we can easily calculate the backwards rate constant. In the example of virology, considering single proteins being recruited into an assembling capsid at a time, these can be modelled as independent of ϵ, and be held constant. Instead, the rate of disassembly can include the consideration of the energetics of interaction. This approach can be used to simplify the modelling of self-assembly in the case of elementary reactions, even in more complicated systems.

2.3 Polymers

Polymerization can be viewed as a self-assembling system in many instances [74–76], though in many cases the resulting polymers include more disorder than would typically be expected in a true self-assembling system. Polymers form by sequential addition of monomers: in this sense the system resembles an extension of the dimer addition process detailed above. However, polymers usually are fairly flexible in macromolecular form, and in complete polymers there is often branching or cross-linking between chains [3,74]. The structure of polymer chains becomes even more complex when the polymers are formed from more than one type of monomer, in a scenario known as *copolymerization*. Types of copolymer known as *block* and *alternating copolymers* are of particular interest to this thesis, as we describe in Chapter 4 the synthesis of a rigid 2-D[1] assembly from alternating precursor molecules.

In polymer science, the number of repeating units (usually simply the monomers themselves that created the polymer) are referred to as the *degree of polymerization*. Here we shall denote it by l as we explore the kinetics of polymerization. With monomers denoted as M, and polymers as P_l (with length l) the following

[1]For all abbreviations see the glossary on page 222.

simple relationship holds:

$$\mathrm{M} + \mathrm{P}_l \xrightarrow{k} \mathrm{P}_{l+1} \ . \tag{2.3.1}$$

There are two possible scenarios: either (i) the monomers are of a different chemical form to the polymers (as in classic polymerization, where monomers attach to the polymer and undergo a chemical transformation), or (ii) akin to self-assembly, monomers are effectively identical to polymers of length 1, *i.e.* $\mathrm{M} = \mathrm{P}_1$. In the latter scenario, it is possible to have coupling of polymers, as well as monomeric addition. The kinetics of the two processes are different, as we explain below.

2.3.1 Case 1: Monomeric assembly scenario

If only monomer addition is allowed (the first case), then some interesting analytical results can be found, as an extension of the dimerization scenario in §2.2. The only species that we assume to be present at t=0 are monomers (M) and initiation points, considered as zero-length polymers (P_0). With $C_l(t)$ and $C_{\mathrm{M}}(t)$ as the concentration of P_l and M at time t respectively, conservation of monomer units implies:

$$\sum_{l=1}^{\infty} l C_l(t) = C_{\mathrm{M}}(t{=}0), \tag{2.3.2}$$

and conservation of the number of polymers implies:

$$\sum_{l=0}^{\infty} C_l(t) = C_0(t{=}0). \tag{2.3.3}$$

We can now apply the Law of Mass Action to describe the assembly process

in terms of the concentrations of the molecules. In general:

$$\frac{\mathrm{d}C_l(t)}{\mathrm{d}t} = kC_{\mathrm{M}}(t)[C_{l-1}(t) - C_l(t)], \tag{2.3.4}$$

and in the case of the initiation of polymerization:

$$\frac{\mathrm{d}C_0(t)}{\mathrm{d}t} = -kC_{\mathrm{M}}(t)C_0(t). \tag{2.3.5}$$

Of course, for non-trivial initial conditions, $C_{\mathrm{M}}(t=0) \neq 0$ and $C_0(t=0) \neq 0$.

We can also write the following sum, from the perspective of the monomers:

$$\frac{\mathrm{d}C_{\mathrm{M}}(t)}{\mathrm{d}t} = -kC_{\mathrm{M}}(t)\sum_{l=0}^{\infty} C_l(t), \tag{2.3.6}$$

which we can solve easily, using Equation (2.3.3) and separation of variables:

$$\int \frac{\mathrm{d}C_{\mathrm{M}}(t)}{C_{\mathrm{M}}(t)} = \int -kC_0(t=0)\,\mathrm{d}t$$

$$\ln C_{\mathrm{M}}(t) = -kC_0(t=0)t + \text{const.}$$

$$\ln C_{\mathrm{M}}(t) - \ln C_{\mathrm{M}}(t=0) = -kC_0(t=0)t$$

$$C_{\mathrm{M}}(t) = C_{\mathrm{M}}(t=0)e^{-kC_0(t=0)t}. \tag{2.3.7}$$

We can then, as outlined by Pelesko [3], eliminate C_{M} from the equations for $P_l(t)$, by defining a new variable $y(t)$ such that $\frac{\mathrm{d}y}{\mathrm{d}t} = kC_{\mathrm{M}}(t)$ and with $y(t=0) = 0$. Equations (2.3.4) and (2.3.5) become:

$$\frac{\mathrm{d}C_l(y)}{\mathrm{d}y} = C_{l-1}(y) - C_l(y) \tag{2.3.8}$$

$$\frac{\mathrm{d}C_0(y)}{\mathrm{d}y} = -C_0(y) \tag{2.3.9}$$

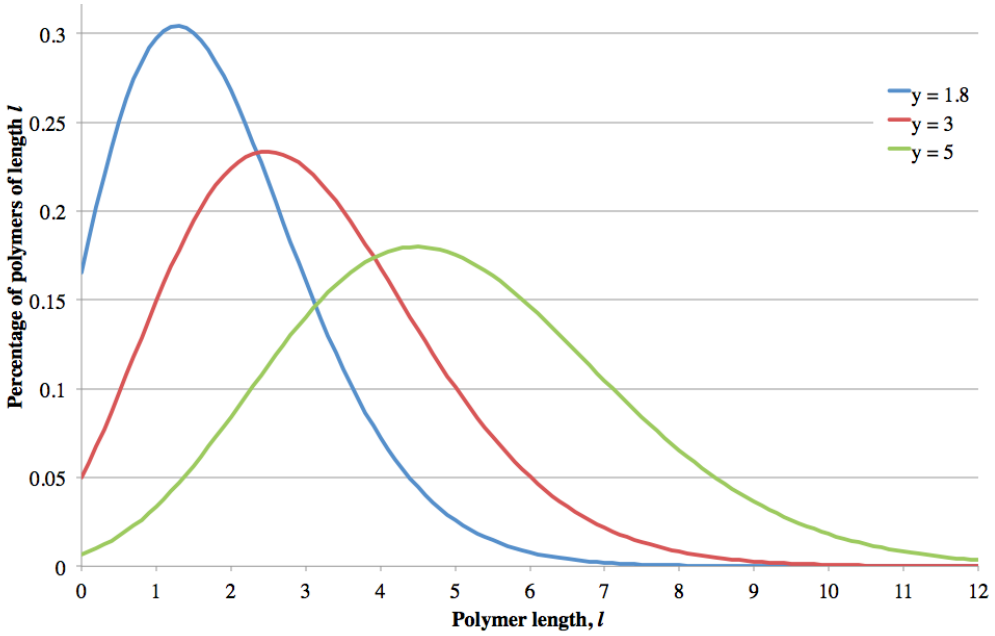


Figure 2.2: Distribution of polymers in a polymerization reaction, in the model described by Equation (2.3.10).

which can be solved to find:

$$\frac{C_l(y)}{C_{0(y=0)}} = \frac{y^l e^{-y}}{l!}. \tag{2.3.10}$$

This analytical solution is a distribution of polymer chain lengths (l) as a factor of y. It is a *Poisson* distribution, with mean and variance y. A plot of this is shown in Figure 2.2, calculated for (non-physical) non-integer values of l by reference to the Gamma function:

$$\Gamma(n) = (n-1)!, \qquad n \in \mathbb{N}_0. \tag{2.3.11}$$

As y is a function of time, this distribution shows the lengths of polymers instantaneously at a given time. Interestingly, when sampled stochastically, the system will tend towards a Poisson distribution of lengths, which implies that

there will not be one product at the end of the reactions. Of course, this is to be expected as there are no reactions controlling the lengths of any of the extending polymers. This behaviour is seen experimentally for some 1-D polymerization reactions [77, 78].

2.3.2 Case 2: Oligomeric assembly scenario

In the second case, assuming that monomers (M) are equivalent to polymers of length 1 (P_1), thus allowing oligomeric addition, with the same initial condition:

$$\sum_{l=1}^{\infty} lC_l(t) = C_1(t=0) = C_{\mathrm{M}}(t=0) \tag{2.3.12}$$

We denote concentration of monomer by $C_{\mathrm{M}}(t)$ (which is equivalent to $C_1(t)$). From here, similar to the dimer assembly above, we can apply the Law of Mass Action to describe the assembly process in terms of the concentrations of the molecules.

However, as each polymer can interact with (and extend) each other polymer, the summations are non-trivial. Instead of attempting an analytic solution, numerical approaches are preferred for this and similar systems. Refer to Chapter 4 for a numerical analysis of self-assembly, in the more interesting case of 2-D growth.

Next we consider approaches to kinetics when modelling more complicated reactions that are not necessarily elementary.

2.4 Catalyst kinetics

Unlike elementary reactions, *complex* reactions such as catalysed reactions are more difficult to analyse in terms of kinetics. As we have seen, under the Law of Mass Action, the rate of elementary reactions is proportional to the concen-

trations of the reactants raised to the powers of their stoichiometric coefficients. This is not the case in complex, composite reactions. The rate equation may potentially not be simply related to the overall stoichiometry of the reaction [54].

Clearly, complex reactions can be modelled to a level of accuracy dependent on the understanding of the chemical reaction mechanism. A true simulation of a reaction mechanism would involve an exact description of the molecular mechanics of the reaction and of any enzymatic or catalytic cycle. However, for many reaction mechanisms some important details remain to be elucidated, even for reactions that are utilized extensively in industrial synthesis [79]. In particular, this is true of the self-assembly coupling reactions utilized in Chapter 4, for which little is known about their mechanism [80]. Yet to an arbitrary level of accuracy, it is possible to characterize reaction mechanisms using a description with fewer degrees of freedom in certain circumstances, with an explicit analytic function of reactants and products describing the rates [81]. Derivation of the corresponding rate equations is usually achieved using a quasi-steady-state or rapid equilibrium (Henri-Michaelis-Menten) approach [61, 81].

Consider the following reaction mechanism for the catalysed conversion of reactant R into product P by catalyst C:

$$\mathrm{R} + \mathrm{C} \underset{k_b}{\overset{k_f}{\rightleftharpoons}} \mathrm{RC} \xrightarrow{k_t} \mathrm{C} + \mathrm{P} \tag{2.4.1}$$

assuming that this single catalysed reaction rate is fastest, and much faster than the non-catalysed reaction:

$$\mathrm{R} \xrightarrow{k_u} \mathrm{P}\ . \tag{2.4.2}$$

As before, we apply the Law of Mass Action to derive a system of non-linear ordinary differential equations defining the rate of change of the concentration

of actors in the system:

$$\frac{\mathrm{d}C_{\mathrm{R}}}{\mathrm{d}t} = -k_f C_{\mathrm{C}} C_{\mathrm{R}} + k_b C_{\mathrm{RC}} \tag{2.4.3}$$

$$\frac{\mathrm{d}C_{\mathrm{C}}}{\mathrm{d}t} = -k_f C_{\mathrm{C}} C_{\mathrm{R}} + k_b C_{\mathrm{RC}} + k_t C_{\mathrm{RC}} \tag{2.4.4}$$

$$\frac{\mathrm{d}C_{\mathrm{RC}}}{\mathrm{d}t} = k_f C_{\mathrm{C}} C_{\mathrm{R}} - k_b C_{\mathrm{RC}} - k_t C_{\mathrm{RC}} \tag{2.4.5}$$

$$\frac{\mathrm{d}C_{\mathrm{P}}}{\mathrm{d}t} = k_t C_{\mathrm{RC}} \tag{2.4.6}$$

There is conservation of catalyst within the system, but the concentration of free catalyst, C_{C}, does change; in fact the number of complexes including catalyst remains constant:

$$C_{\mathrm{C}} + C_{\mathrm{RC}} = (C_{\mathrm{C}} + C_{\mathrm{RC}})|_0 \equiv C_C^0. \tag{2.4.7}$$

In this derivation we assume that the reactant is in a Henri-Michaelis-Menten instantaneous chemical equilibrium with the catalytic complex [82], implying:

$$k_f C_{\mathrm{C}} C_{\mathrm{R}} = k_b C_{\mathrm{RC}}. \tag{2.4.8}$$

Using Equation (2.4.7):

$$k_f (C_C^0 - C_{\mathrm{RC}}) C_{\mathrm{R}} = k_b C_{\mathrm{RC}}, \tag{2.4.9}$$

which can be further simplified to provide:

$$C_{\mathrm{RC}} = \frac{C_C^0 C_{\mathrm{R}}}{K_d + C_{\mathrm{R}}}, \tag{2.4.10}$$

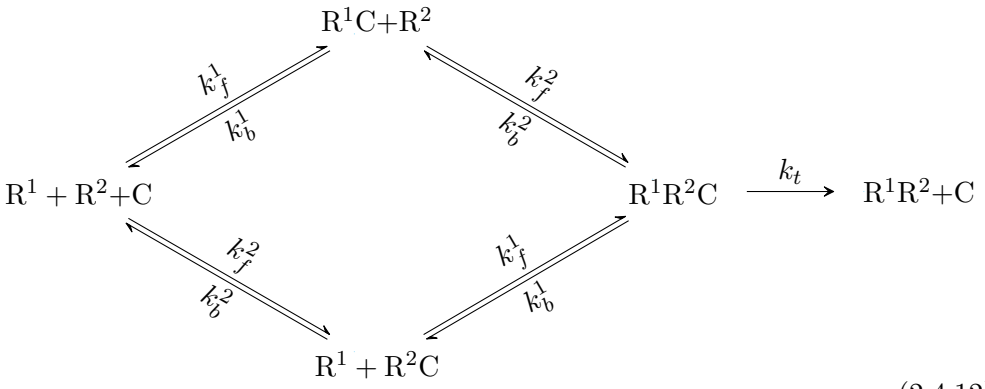
where $K_d = k_b/k_f$ is the dissociation constant for the catalytic complex.

Hence:

$$r = \frac{\mathrm{d}C_\mathrm{P}}{\mathrm{d}t} = \frac{r_\mathrm{max} C_\mathrm{R}}{K_d + C_\mathrm{R}} \tag{2.4.11}$$

where $r_\mathrm{max} = k_t C_C^0$ is the maximum reaction rate. If K_d is large and dominates the denominator, then the reaction can be approximated by a first order reaction with a modified rate constant.

In Chapter 4 there are catalysed symmetric reactions used in the synthesis of nanographene that are kinetically modelled: the Ullmann and Suzuki-Miyaura reactions [83]. These are even more complicated mechanisms than the catalysed reaction described above (see also Figure 2.3). For example, with two functionalized molecules, R^1 and R^2 being coupled by a single catalyst, a simple version of the reaction scheme would be:



(2.4.12)

where $\mathrm{R}^1\mathrm{R}^2$ is the desired product. An accurate derivation of a kinetic rate equation for this reaction pathway would be difficult, but moreover would run the risk of overfitting the negligible data available to describe the mechanisms. A further simplification is therefore preferred.

As we have seen in Equation (2.4.11), under conditions where K_d is large, the rate equation of one catalysed reactant (Scheme 2.4.1) becomes first order. In a similar way, the rate equation of two catalysed reactants (Scheme 2.4.12) can be approximated as second order. Under certain conditions, the second

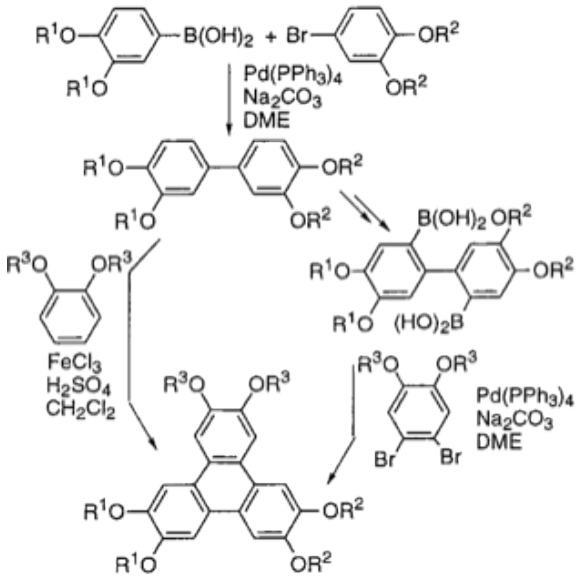


Figure 2.3: Suzuki-Miyaura coupling reaction. As shown here, the coupling reaction can be used to make small aromatic carbon molecules with a wide variation possible in terms of R^1, R^2, and R^3: the palladium-catalysed coupling mechanism is very tolerant toward functional groups. The reaction can also be used to couple together similarly sized molecules into porous nanographene (Chapter 4). Figure taken from [83].

order reaction kinetics are able to be approximated further, and be described by pseudo-first order kinetics. Quantification of the reaction dynamics becomes greatly simplified when approximating to first order.

2.4.1 Pseudo-first order reactions

A second-order reaction is:

$$\mathrm{A} + \mathrm{B} \xrightarrow{k} \mathrm{P} \tag{2.4.13}$$

which has the rate equation:

$$r = kC_AC_B \tag{2.4.14}$$

It can be difficult to experimentally determine the rate of second-order reactions, as the two reactants must be measured simultaneously, and more measure-

ments have to be taken at different concentrations of the two reactants [54]. The reaction can be simplified, in experimental and theoretical terms, by approximation as a pseudo-first order reaction [84]. In order to consider the reaction as pseudo-first order, there must be a disparity in the initial concentrations of the two reactants. For example, of the above reactants, $C_A \gg C_B$. It would then be possible to assume that the concentration of reactant A effectively remains constant throughout the reaction, as any change in concentration due to the reaction proceeding would be negligible to the overall concentration. It is thus possible to define a new rate constant, $k' = kC_A$, that will simplify the rate equation to:

$$r = k'C_B \tag{2.4.15}$$

In this way the new rate constant allows us to simplify the second order reaction and treat it as a first order reaction under the condition of $C_A \gg C_B$.

2.5 Self-assembly catalysed by a surface

Many important industrial reactions are *heterogeneously* catalysed by solid materials [85], with the reaction able to take place more readily on a surface than in a homogeneous gaseous or liquid environment (*e.g.* the Haber process) [86]. At the surface of a solid (composed of atoms or molecules) there is a high propensity for reactions to occur: reactions that in particular may act as part of a catalytic cycle.

To determine the time evolution of concentrations of molecules adsorbed on a surface, varying in both space and time, there are three chemical processes that need to be considered: reaction, diffusion, and adsorption/de-adsorption.

2.5.1 Surface reactions

Here we must introduce *transition state theory* [87], which seeks to characterize reactions pertaining to the Arrhenius equation (§2.1.1, Equation (2.1.1)) from first principles. The theory provides a useful framework for examining experimentally-discovered mechanisms and determining complicated rate equations, but has several shortcomings (details of which are beyond the scope of this cursory examination) [88, 89]. Suffice to say, the theory can provide a nice qualitative perspective in certain conditions. For surface reactions, of note is the Langmuir-Hinshelwood mechanism [87], which is a well-characterized description of a bimolecular reaction in an adsorbed layer on a surface:

$$\mathrm{A_{(s)}} + \mathrm{B_{(s)}} \longrightarrow \mathrm{P} \tag{2.5.1}$$

for which the time course is described by:

$$\frac{\mathrm{d}\theta_A}{\mathrm{d}t} = \frac{\mathrm{d}\theta_B}{\mathrm{d}t} = -\nu e^{\frac{-E_a}{k_B T}} \theta_A \theta_B \tag{2.5.2}$$

where θ is the surface coverage of a particle, and ν is the standard pre-exponential factor predicted by transition state theory [87] (*n.b.* this is derived from first principles using partition functions, and usually approaches the correct magnitude determined experimentally). In this case, the calculation is that $\nu \approx 10^{11}\mathrm{s}^{-1}$–$10^{19}\mathrm{s}^{-1}$ [87].

To calculate in terms of numbers of adsorbed molecules, the equation can be converted using the relation $N = N_0\theta$, where N_0 is the density of surface states in which particles can be positioned: in dimensional terms $[N_0] = \frac{1}{l^2}$. For small species, $N_0 \approx 10^{15}\mathrm{cm}^{-2}$ [87]. However, for larger adsorbates, N_0 would decrease substantially, depending on their size; feasibly for large self-assembly

intermediates, N_0 could approach $10^{12}\,\mathrm{cm}^{-2}$. In this way:

$$\frac{\mathrm{d}N_A}{\mathrm{d}t} = \frac{\mathrm{d}N_B}{\mathrm{d}t} = -k_0 e^{\frac{-E_a}{k_B T}} N_A N_B \tag{2.5.3}$$

with $k_0 \approx 10^{-4}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$–$10^{8}\,\mathrm{cm}^2\,\mathrm{s}^{-1}$, depending on the size of species.

2.5.2 Surface diffusion

Surface diffusion is a very important consideration for accurately determining the mechanisms of reactions on surfaces, due to its influence on the form and structure of products [87, 90–92]. Phenomenologically, diffusion on a surface is described by Fick's laws [87]:

$$J = -D\nabla c \qquad \text{(Fick's first law)} \tag{2.5.4}$$

$$\frac{\partial c}{\partial t} = \nabla(D\nabla c) \qquad \text{(Fick's second law)} \tag{2.5.5}$$

where c is concentration and J is the flux. At low coverages, the diffusion coefficient is simplified as:

$$D = \Gamma\langle s^2\rangle/2d \tag{2.5.6}$$

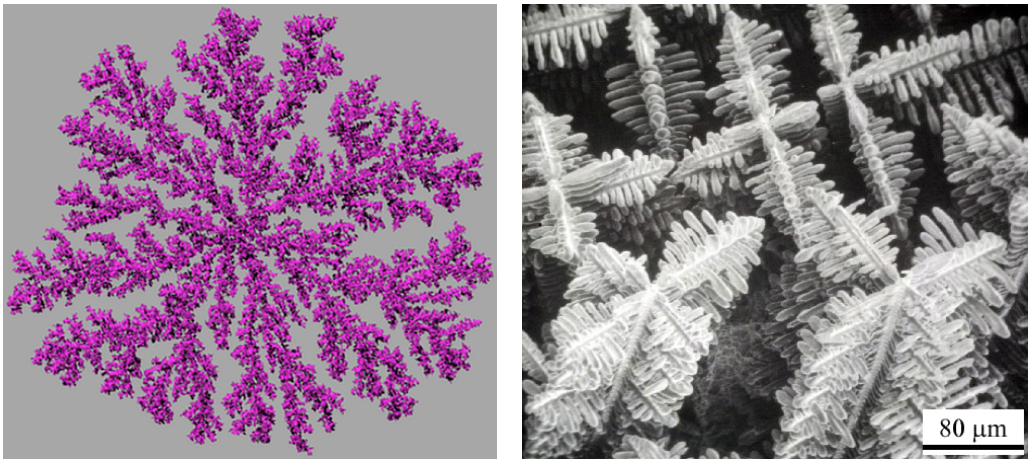
where Γ is the jump rate, s^2 is the average square of the jump length, and d is the dimensionality of the diffusive motion [87].

There are three main factors that influence the rate and mechanism of adsorbed particle diffusion on a surface [55, 91]:

(i) the exchange of energy between adsorbed particles and the substrate;

(ii) the topology of the conformational potential energy surface;

(iii) any substrate lattice relaxation to new positions of equilibrium, as the adsorbed particle moves between cells.



(a) DLA model (b) Dendritic growth of Co-Sm-Cu

Figure 2.4: Dendritic growth. (a) DLA model, utilizing a random walk approach, showing dendritic growth. (b) Scanning electron microscope (SEM) micrograph of the 3D structure of dendrites in a cobalt-samarium-copper alloy, showing inherent order but uncontrolled growth directions. Adapted from licensed work: [96].

The most direct way of testing and describing these factors, and thus faithfully modelling diffusion, is molecular dynamics (MD) simulation of the movement of an adsorbed particle over atoms in the substrate lattice [87]. This is an accurate, but laborious, method that takes all the three factors into account.

There exist both coarse-grained deterministic [90] and stochastic models of diffusion in the context of self-assembly: of the latter diffusion limited aggregation (DLA) models are of particular note. These models were first introduced by Witten and Sander [93,94] as a general model of aggregative processes, based on random walks of particles on a lattice. DLA models give rise to dendritic growth, as shown in Figure 2.4(a). Without constraint, these growth processes leave gaps within the structure, and in spite of their inherently ordered binding interactions, thus exhibit disorder [95]. An example of the pseudo-crystalline structures that can be formed in diffusion limited processes is shown in Figure 2.4(b).

Methods for the creation of pristine crystalline or other non-disordered struc-

tures are predicated upon non-diffusion limited assembly procedures [97]. In the context of the coupling of self-assembling structures on the surface, this means that diffusion across the surface is viewed as free, to a first-order approximation [3]. Moreover, in terms of kinetics, the kinetics of adsorption/deadsorption as well as diffusion are likely to be energetically more favourable than the coupling reactions. Both of these considerations mean that the system can be considered well-mixed, homogeneous and isotropic. In conclusion, diffusion-limited self-assembly processes do not normally yield long-range order: coupling-limited approaches to self-assembly are favoured, and are present throughout the literature [97–99].

2.5.3 Surface adsorption/deadsorption

There are principally two mechanisms for the attachment of adsorbate molecules onto the surface of an adsorbent: *physiadsorption* and *chemiadsorption*. Physiadsorption, often occurring initially, involves weak physical interactions and is driven by long-range van der Waals interactions [87]. On the other hand, chemical adsorption is considerably stronger and results in chemical bonds forming between the substrate and adsorbate [100]. These bonds are on the order of -1.5eV, and as mentioned sometimes form after an initial van der Waals interaction, depending on any activation energy required [86]. Strong activation energies often make such adsorptive process irreversible, and thus we shall focus here on a thermodynamic examination of weaker reversible adsorption/deadsorption processes, which will be largely physical in nature, and which are (as a first-order approximation) more applicable to the experimental conditions featured in Chapter 4.

We commence with the first law of thermodynamics:

$$\Delta U = \delta Q - \delta W. \tag{2.5.7}$$

where U is the internal energy of a closed system, with infinitesimal quantities δQ as the heat added to the system and δW as the work done by the system.

Assuming the pressure at the surface is uniform, we can split the work done into expansion work ($p\,\mathrm{d}V$) and other work ($\delta W'$):

$$\Delta U = \delta Q - p\,\mathrm{d}V - \delta W'. \tag{2.5.8}$$

From the definition of enthalpy $H = U + PV$, we can calculate in terms of enthalpy:

$$\Delta H = \delta Q + V\,\mathrm{d}p - \delta W'. \tag{2.5.9}$$

Hence at constant pressure, and if only expansionary work is done by the system, the last two terms go to zero, and the increase in enthalpy of a system is equal to the added heat:

$$\Delta H = \delta Q. \tag{2.5.10}$$

Consider the following definition of entropy in a homogeneous system, derived from the Clausius equality:

$$\Delta S_{\text{heat}} = -\frac{\delta Q}{T}. \tag{2.5.11}$$

Using Equation (2.5.10), we obtain:

$$\Delta S_{\text{heat}} = -\frac{\Delta H}{T}. \tag{2.5.12}$$

Examining the adsorbing process entropically, the entropy will decrease as

the free adsorbent becomes localized on the surface:

$$\Delta S_{\text{ad}} < 0. \tag{2.5.13}$$

From a universal perspective, $\Delta S_{\text{tot}} \geq 0$. Thus

$$\Delta S_{\text{tot}} = \Delta S_{\text{ad}} + \Delta S_{\text{heat}} \geq 0. \tag{2.5.14}$$

This is equivalent to

$$\Delta S_{\text{ad}} - \frac{\Delta H}{T} \geq 0, \tag{2.5.15}$$

implying

$$T\Delta S_{\text{ad}} \geq \Delta H. \tag{2.5.16}$$

This condition may be satisfied with both $\Delta S_{\text{ad}} < 0$ and $\Delta H < 0$, such that:

$$|\Delta H| > |T\Delta S_{\text{ad}}|, \tag{2.5.17}$$

i.e. adsorption is always exothermic: $\Delta H \leq 0$.

We observe experimentally that energy is released during adsorption as the enthalpy of adsorption is negative, on the order of -0.4eV [86]. Also, examining the process thermodynamically, using Equation (2.1.3), the adsorption will only be able to occur when ΔG is negative, which is possible when $\Delta H < T\Delta S$.

For an adsorption process, the condition is met initially, but as the adsorption proceeds the ΔH value decreases in magnitude, whereas $T\Delta S_{\text{ad}}$ increases in magnitude, and finally $\Delta H = T\Delta S_{\text{ad}}$, so that $\Delta G = 0$. This is the point at which the system reaches the state of adsorptive equilibrium.

Note that the presence of multiple mechanisms of adsorption can result in counterintuitive effects when, for example, increasing the temperature [100]. At

low temperatures, physiadsorption prevails, due to the low activation energy barrier, eventually equilibrating at an adsorptive equilibrium. Small increases in temperature will result in more free adsorbate, and less surface coverage. However, large increases in temperature could allow chemiadsorptive processes to commence: this would reverse the trend, and increase the surface coverage. At still higher temperatures there would be accelerated deadsorbtion and thus a resumption of the overall trend towards lower adsorbtion.

2.6 Simulation of self-assembly

Computational techniques can be used to gain insight into assembly reactions and to focus the scope of investigation, as they require less time and expense than laboratory experimentation. *Stochastic modelling* provides a method of exploring variations in self-assembling systems as diverse as viruses [11] and quantum dots [101], with a minimal cost outlay. In Chapter 4 we describe an accelerated coarse-grained simulation, using a Gillespie KSA algorithm [24], to examine the self-assembly of GNRs from covalently-bonding aromatic molecules on an inert metallic surface. The KSA algorithm simulates the kinetics of self-assembly by generating a network of possible reactions between *intermediate molecules*, and firing reactions stochastically. Our model uses a new technique for determining binding between molecules, introduced in Chapter 3. Reaction pathways can be traced through the reaction network to determine assembly mechanisms, and the time-evolution of the system studied. By interrogating the model with sets of parameters—specifically reaction temperature, precursor concentration, and activation energies for bond formation—certain combinations of parameters will appear amenable for directed self-assembly, whilst others can be ruled out. Moreover, the trends observed in parameter space can inform further experimental design and selection of precursor molecules, for example by

predicting which coupling reactions are most suitable for recreating an observed behaviour.

Previous models of planar self-assembly include Monte Carlo (MC) models, which can generate similar results to KSA models, but are not structured around a reaction network. One uses a *hard sphere* approximation for assembling monomers, with functional groups attached to the surface of the sphere [102]. Another model examines interactions between tripod-shaped molecules [103, 104]. Both of these similar MC simulations are suitable for modelling 2-D (planar) self-assembly in the systems for which they were designed, and they operate by allowing movement of molecules within a simulated real space, and both the formation and breaking of bonds. There are also existing MC models of graphene growth, focussed on the kinetics of the attachment of small carbon species to the edge of graphene monolayers [105, 106]. Intermediate molecules and prominent reaction pathways are not explicitly calculated by MC algorithms unlike KSA: these would have to be monitored separately. Thus KSA modelling offers transparency in determining the reaction pathways, as well as a substantial computational advantage. The KSA algorithm we present also disregards the positions and movement of the molecules in real space, instead working solely with a probabilistic approach. This further streamlines the computation and thus allows more simulations to be conducted within a given time.

Other popular computational approaches to study planar self-assembly utilize MD and density functional theory (DFT) [107]. However, these techniques are limited in scale by the number of atoms simulated: on the order of several thousand or a few hundred, respectively. In contrast, KSA (or alternatively MC) methods can simulate the assembly of hundreds of molecules at once. For any system, MD or DFT provides a more meticulous model of self-assembly interactions than KSA [25]. However, even in coarse-grained MD it is nor-

mal to have time steps on the order of ps [108], and short simulation times of around 100ns [109]. The techniques are unsuitable for systems with infrequent events [110]: even on short timescales, some coarse-grained simulations last several weeks and require vast computational resources [108, 111]. The computational resources required mean it is inefficient to collect statistics using MD over different experimental setups, even as part of a multi-scale model (see [112]), especially as some experiments could last on the order of minutes [108, 110]. However, results from a theoretically-informed KSA model could subsequently be validated by MD, in a process known as reverse-hybrid MC modelling [113], before testing in the laboratory.

2.7 Refinement of self-assembly processes

Principally there are two main avenues along which modelling can improve self-assembly experiments, and thus enable efficient use of self-assembling processes in engineered synthesis pathways:

(i) identification of complex components with desired self-assembly outcomes

(ii) incorporation of components (*e.g.* catalysts, biological cellular material) that regulate the self-assembling process in a desired way.

Generalizations of the complexity of the building blocks and the assembly products are necessary requirements of modelling. Abstraction of the real-world nature of the self-assembling system happens at every stage of the modelling process. For example, the association rate between components is often modelled as uniform. The simulation environment is frequently modelled as homogeneous: *e.g.* the presence of any electric or magnetic fields are ignored. Many of these abstractions are consequences of the inability to measure these initial conditions exactly in the first place. It is understandable, therefore, that an exact authentic

copy with the same initial conditions, and boundary conditions, is unrealistic. The question is therefore to the *parameterization* of the system: when we are fitting a modelled system to experimental data theoretically or computationally, this needs to be considered, and will be discussed further in the next two chapters.

It is challenging to implement regulating components (*e.g.* catalysts, biological cellular material) in a theoretical setting, as these actors vary greatly in their form and amount, yet can substantially affect the outcomes of self-assembly in ways that are difficult to predict. There are two distinct approaches to resolving this difficulty that merit consideration. Should the regulating effects be considered from first principles, in the same way that self-assembly is? Or instead, should the existence of such regulating mechanisms be assumed when fitting models to data? For the latter case, the parameterization of the system is simplified [114, 115], thus it is particularly hard to find mechanisms, which can be obscured by averaging effects in the fitting process. Even ephemeral biological cellular material can have substantive regulatory effects on the self-assembling process. Learning from these regulatory mechanisms is a key output from a systems-based examination of self-assembly.

Indeed, complex biological systems in particular provide inspiration for incorporating self-assembly into productive synthetic chemical systems. This is because the advantages of using self-assembling pathways are particularly highlighted in biology, where there are many examples of different mechanisms and structures that utilize self-assembly principles to great success [59], on many length scales. The common themes in biological self-assembly, to be considered by synthetic chemists, are principally that these processes have evolved to be rapid, economical, and competitive. This is achieved by keeping the precursors simple, which (i) minimizes configurations that are not completable, *i.e.* kineti-

cally trapped, and (ii) also helps ensure that the component construction is not energetically uneconomical for the larger biological system.

Indeed, the self-assembly system can be engineered around the production of the self-assembly components, as the prevailing concentration and timing of their introduction greatly influences the morphology of the products. Component construction is a limiting factor in virology, for example in some ssRNA viruses, whereby the virus assembly has evolved to occur around a single nucleation site on the genome, and be most efficient under a gradual increment of components into the system [11].

Chapter 3

Cross-correlation algorithms for modelling self-assembly

In this chapter we introduce a novel approach to simulating self-assembling processes that can model complex multi-facial reactions between building blocks by using cross-correlation functions. The pared-down algorithm performs large calculations in complex reciprocal space to improve performance, and is best framed as a stochastic rather than deterministic method. The algorithm can be applied to many distinct systems for which a mathematical lattice or tilings can describe interactions between building blocks, to an arbitrary number of dimensions. Results generated by the algorithm show interesting emergent phenomena such as non-linear bifurcating behaviour in some systems, as seen in Chapter 4.

3.1 Definition of cross-correlation

The circular cross-correlation of two N-dimensional matrices, $\mathbf{f}$ and $\mathbf{g}$, is defined as follows:

$$^{\text{circ}}\{\mathbf{f} \star \mathbf{g}\}(i_1, i_2, \ldots, i_N) = \sum_{l_1=0}^{L_1-1} \sum_{l_2=0}^{L_2-1} \cdots \sum_{L_N=0}^{L_N-1} \mathbf{f}(l_1, l_2, \ldots, l_N)\mathbf{g}(l_1 - i_1, l_2 - i_2, \ldots, l_N - i_N) \quad (3.1.1)$$

where L_d is the greater magnitude of the lengths of $\mathbf{f}$ and $\mathbf{g}$ in each dimension $d \in \{1, 2, \ldots, N\}$ from the two matrices (and l_d as indices of that dimension), with the output matrix size being the product $\prod_{d=1}^{N} L_d$ [116]. The circular correlation can be evaluated using discrete Fourier transforms, applying the correlation theorem [116]:

$$^{\text{circ}}\{\mathbf{f} \star \mathbf{g}\}(i_1, i_2, \ldots, i_N) = \mathcal{F}_D^{-1}[\mathbf{F}^*(k_1, k_2, \ldots, k_N)\mathbf{G}(k_1, k_2, \ldots, k_N)] \quad (3.1.2)$$

using Fourier transforms, namely $\mathbf{F}(k_1, k_2, \ldots, k_N) = \mathcal{F}_D[\mathbf{f}(l_1, l_2, \ldots, l_N)]$ and $\mathbf{G}(k_1, k_2, \ldots, k_N) = \mathcal{F}_D[\mathbf{g}(l_1, l_2, \ldots, l_N)]$; $\mathbf{F}^*$ is the complex conjugate of $\mathbf{F}$, k_d are indices in reciprocal space, and $\mathcal{F}_D$ is a discrete Fourier transform [116] that can be performed separably over each axis:

$$\mathcal{F}_D(k_1, k_2, \ldots, k_N) = \frac{1}{\prod_{d=1}^{N} L_d} \sum_{l_1=0}^{L_1-1} \sum_{l_2=0}^{L_2-1} \cdots \sum_{l_N=0}^{L_N-1} \mathcal{F}_D^{-1}(l_1, l_2, \ldots l_N) e^{-\frac{2\pi i k l_1}{T} - \frac{2\pi i k l_2}{T} \cdots - \frac{2\pi i k l_N}{T}}. \quad (3.1.3)$$

The inverse of the discrete Fourier transform is:

$$\mathcal{F}_D^{-1}(l_1, l_2, \ldots, l_N) = \sum_{l_1=0}^{L_1-1} \sum_{l_2=0}^{L_2-1} \cdots \sum_{l_N=0}^{L_N-1} \mathcal{F}_D(k_1, k_2, \ldots, k_N) e^{\frac{2\pi i k l_1}{T} + \frac{2\pi i k l_2}{T} \cdots + \frac{2\pi i k l_N}{T}}. \quad (3.1.4)$$

For a linear correlation,

$$^{\text{linear}}\{\mathbf{f} \star \mathbf{g}\}(i_1, i_2, \ldots, i_N) = \sum_{l_1=-\infty}^{\infty} \sum_{l_2=-\infty}^{\infty} \cdots \sum_{L_N=-\infty}^{\infty} \mathbf{f}(l_1, l_2, \ldots, l_N) \mathbf{g}(l_1 - i_1, l_2 - i_2, \ldots, l_N - i_N). \quad (3.1.5)$$

In contrast to a circular correlation, any indexed elements that lie outside the original range do not contribute to the sum. Therefore, to utilize the correlation theorem, which corresponds to the circular correlations, the input matrices are padded with zeros, preventing the overlap of matrices at the opposite site when calculating elements near the edge. Thus, the size of the linear correlation result matrix, for each dimension d, is $\text{len}(\mathbf{f}|_d) + \text{len}(\mathbf{g}|_d) - 1$. We can therefore calculate the linear correlation with:

$$\{\mathbf{f} \star \mathbf{g}\}(i_1, i_2, \ldots, i_N) = \mathcal{F}_D^{-1}[\mathbf{F}^*_{\text{pad}}(k_1, k_2, \ldots, k_N) \mathbf{G}_{\text{pad}}(k_1, k_2, \ldots, k_N)], \quad (3.1.6)$$

where $\mathbf{F}_{\text{pad}}(k_1, k_2, \ldots, k_N) = \mathcal{F}_D[\mathbf{f}_{\text{pad}}(i_1, i_2, \ldots, i_N)]$ and $\mathbf{G}_{\text{pad}}(k_1, k_2, \ldots, k_N) = \mathcal{F}_D[\mathbf{g}_{\text{pad}}(i_1, i_2, \ldots, i_N)]$.

3.2 A simple, one-dimensional assembly model

The concept can be easily understood by building up a simple 1-D[1] model into additional dimensions. We initially consider a growing polymer, which forms by

[1] For all abbreviations see the glossary on page 222.

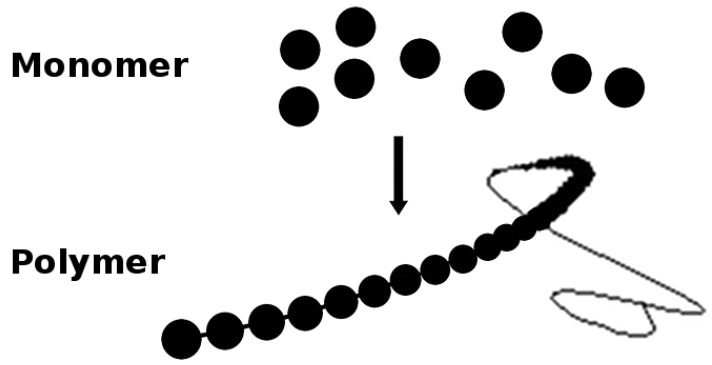


Figure 3.1: Polymer formation by monomeric addition of monomers. The polymer grows at both ends by sequential addition of monomeric units.

aggregation of single monomers onto both ends of the polymer. The scheme for this simple assembly can be seen in Figure 3.1.

The simplest cross-correlation that we could use to represent the coupling of two monomers would be a 1-D cross-correlation:

$$\boxed{1 \quad 1} \star \boxed{1 \quad 1} = \boxed{1 \quad 2 \quad 1} \tag{3.2.1}$$

as arrays containing only one element can only overlap each other in a single position, not allowing the 1-D space to be searched by the cross-correlation function. The function searches the dimensional space represented by the arrays in a manner that is often described as a *sliding window* (Figure 3.2).

Stored in the result array is the number of overlapping points at each offset of the sliding window (*n.b.* this is a result of the input arrays being binary, *i.e.* being composed of only zeros and ones). Although not strictly necessary for cross-correlation, in all our calculations we actually run the calculations with arrays of the same size: padding with zeros if necessary.

The result demonstrates that if the two arrays are fully overlapped, *i.e.* they are occupying the same position, then two binding positions would overlap. Indeed for identical arrays, in what is known as *auto-correlation*, there will be a

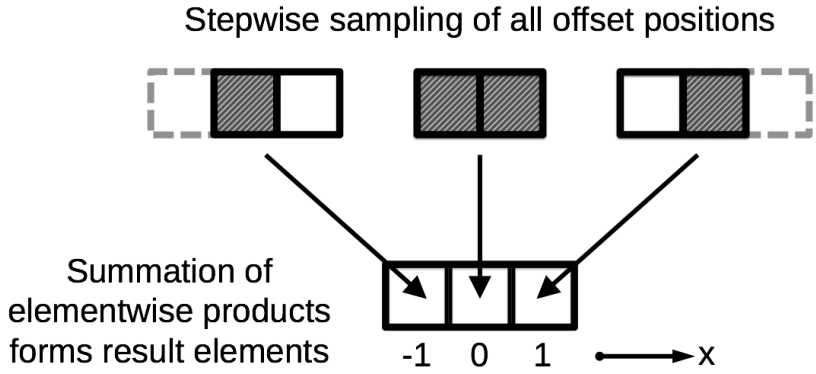


Figure 3.2: Sliding window description of cross-correlation in 1-D. The sliding window calculates the similarity of two series at a particular offset, or lag, of one with relation to the other, in a stepwise manner. This is the cross-correlation algorithm. In an auto-correlation (the cross-correlation of a signal with itself), there is a peak at a lag of zero that corresponds to the signal power.

peak at the offset of 0. However, this is a trivial non-physical result, so must be removed from the result. Thus the final result is:

$$\boxed{1 \quad 0 \quad 1} \tag{3.2.2}$$

which shows that monomers can join together if and only if one is offset from the other, by a distance of a monomer length.

In addition to monomeric assembly, we can consider the addition of partially complete polymers: *i.e.* the assembly of oligomeric units (Figure 3.3). Clearly, the rate of interaction between these pseudo-1-D units is still the same, assuming a well-mixed diffusion-free reaction condition. This can easily be also considered by the algorithm.

3.3 Introducing a pseudo-second dimension

Perhaps the monomers in the polymer can have two types of orientation, with some twisted to allow twisting of the polymer. Assume the twist is confined to

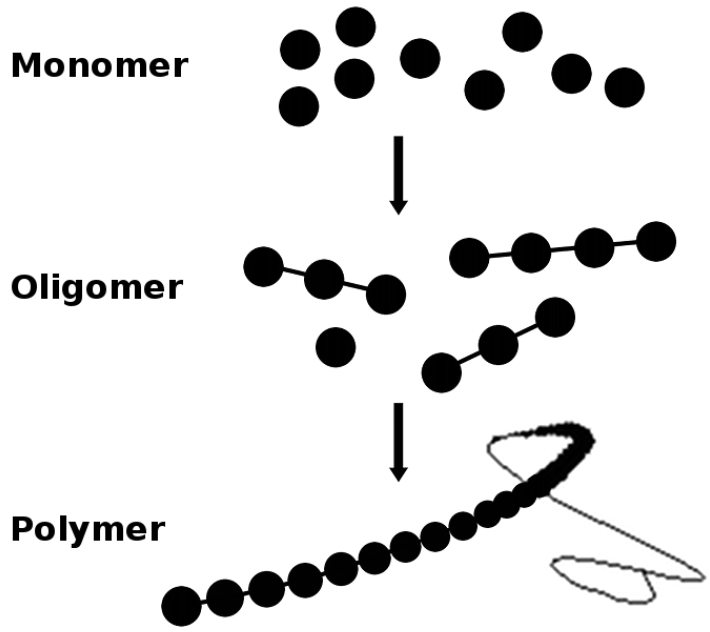


Figure 3.3: Polymer formation by the addition of oligomers. The polymer grows at both ends principally by the joining together of partially complete polymers.

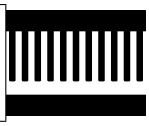
a single unit. We can generalize this 1-D monomer into a second dimension, by considering the second dimension as an additional degree of freedom: in this case, the ability to shift simply into a different rotation. Though the different rotations are at this moment indistinguishable, they may lead to differences in packing if this affects polymer-polymer packing in 2- or 3-D. So it is useful to keep track of the rotation of the polymer:

(3.3.1)

In this example, we will allow only two possible orientations of the polymer. Although there are therefore two different types of monomer, twisted and untwisted, there are more representations in the arrays. They represent the different orientations, or rotations of the 1-D ribbon through the pseudo-second dimension.

The following is a representation of the units in an untwisted ribbon, with a pair of two arrays describing the molecule. There are two strands to the ribbon depicted in this example, dashed and non-dashed, and note that only the dashed half of the ribbon is represented in the arrays: the conjugate non-dashed side of the ribbon is implicitly present but not specified. The left side array provides positions (marked as 1s) for the potential binding points between the dashed strand of monomers, whereas the right side array shows the location of the whole of the dashed strand in the monomer:

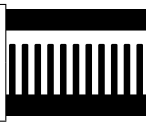
1	0	1	1	1	1
0	0	0	0	0	0



(3.3.2)

As an alternative orientation of the monomer:

0	0	0	0	0	0
1	0	1	1	1	1



(3.3.3)

By contrast, for a twisted ribbon, we have a swapping of the sides of the dashed and non-dashed strands:

1	0	0	1	1	0
0	0	1	0	1	1

(3.3.4)

In an alternative orientation:

0	0	1	0	1	1
1	0	0	1	1	0

(3.3.5)

So if we take an example orientation of a non-twisted (Equation (3.3.3)) and

twisted (Equation (3.3.5)) ribbon section, the cross-correlation produces:

$$\left[\begin{array}{ccc|ccc} 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 \end{array}\right] \star \left[\begin{array}{ccc|ccc} 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \end{array}\right] = \left[\begin{array}{ccccc|ccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 2 & 2 & 1 & 0 \end{array}\right] \tag{3.3.6}$$

of which the central row is what is important, as the polymer has to join end-to-end, and the central row represents no change in the rotation between joins. There is only one point at which these monomers can join together cleanly. This is circled in the central row, shown alone here:

$$\left[\begin{array}{cccc|cccc} 0 & 0 & 1 & 0 & \textcircled{1} & 0 & 1 & 2 & 2 & \textcircled{1} \end{array}\right] \tag{3.3.7}$$

After compensating for the translational difference between these monomers, we have the following resulting polymer:

$$\left[\begin{array}{ccccc|ccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{array}\right] \tag{3.3.8}$$

In order for the algorithm to proceed to extending this ribbon, which is now two monomers long, we remove the binding point from the left hand array where these two monomers joined together (shown circled):

$$\left[\begin{array}{ccccc|ccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 \\ 1 & 0 & \textcircled{0} & 0 & 0 & 1 & 1 & 1 & 1 & 0 \end{array}\right] \tag{3.3.9}$$

This ribbon can be used again iteratively in the algorithm, within a population of monomers and ribbons of different lengths.

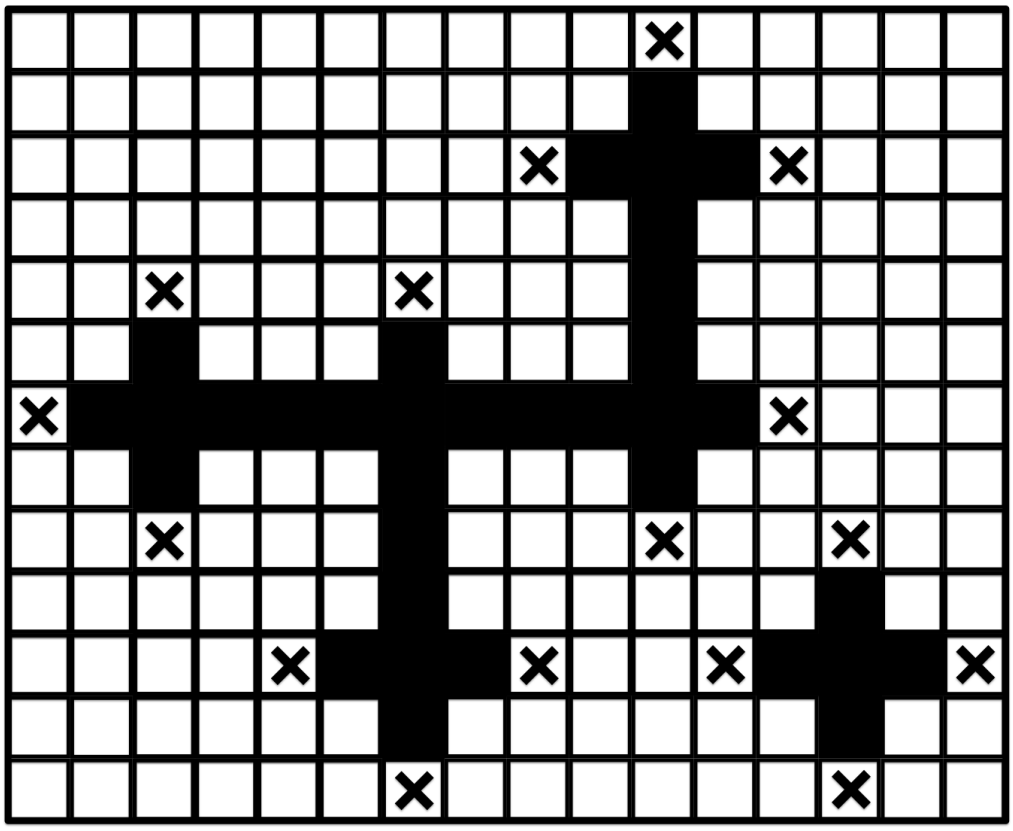


Figure 3.4: Tiling in 2-D: a suitable application for the cross-correlation approach. If molecules (shaded squares), moving around on the lattice, can interact at the binding positions (crosses), then you can have a 2-D tiling scenario.

3.4 Two-dimensional tiling model

The approach can be easily generalized into 2-D, for suitable tiling systems such as that described in Figure 3.4. The sliding window description of cross-correlation in 2-D (Figure 3.5) still holds, but is described more formally in §3.1.

We shall further investigate the 2-D case in the context of an application to modelling the self-assembly of GNR molecules, which involves permuting the square lattice of this example into a honeycomb lattice.

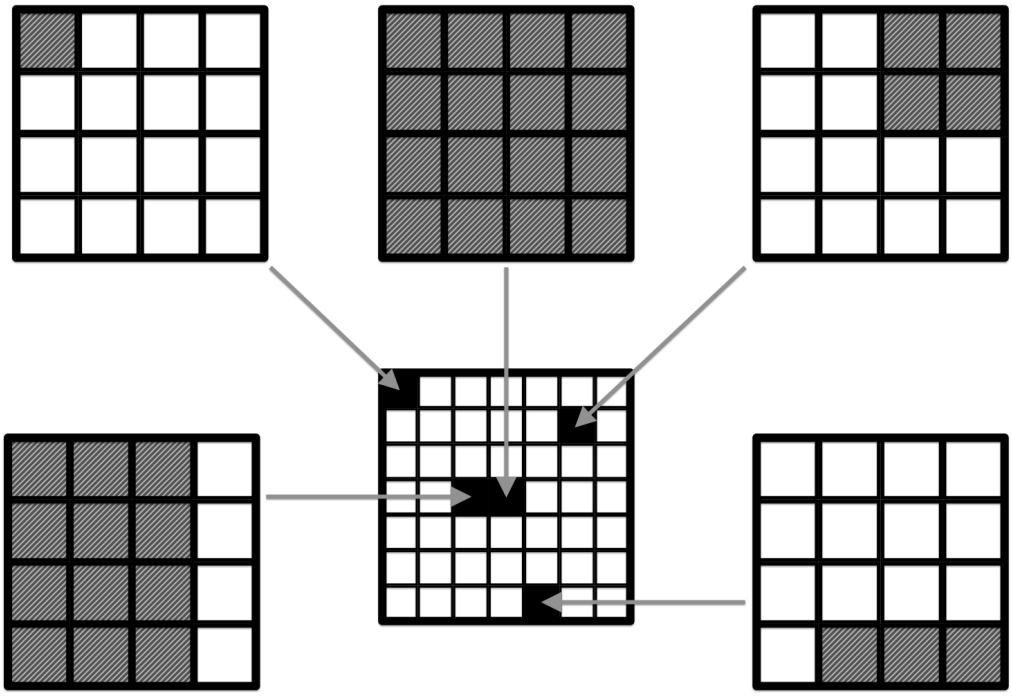


Figure 3.5: Sliding window description of cross-correlation in 2-D. A pictoral description of how the cross-correlation algorithm works, for two 2-D arrays of size (4,4). One array slides over the other, and the corresponding matching elements are multiplied together, and the products are subsequently summed. The sum is stored as an element in the output array. Thus every possible overlapping of these two arrays (of which five examples are given) corresponds to an element in the output array, determined by the geometry of how the two arrays overlap.

3.5 Modelling of nanographene self-assembly

The developed coarse-grained algorithm utilizes a fast probabilistic approach, constraining degrees of freedom of molecular movement and interaction. A honeycomb symmetry coordinate system is used to describe the molecules: it is appropriate to refer to coordinates in the interactions by their location, with respect to nonorthogonal honeycomb lattice vectors, seen in Figure 3.6. Importantly, benzene rings are generalized to a single coordinate representing the entire group of atoms (also shown in Figure 3.6). Thus, individual atoms are not simulated explicitly: molecules can be easily compared by reference to group coordinates rather than individual atomic positions (Figure 3.7(a)). For compu-

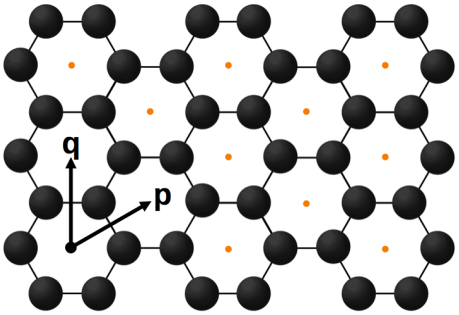
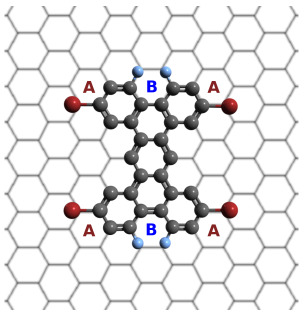
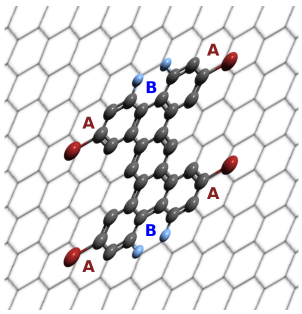


Figure 3.6: Coordinate system: non-orthogonal lattice vectors (p, q) describe a single coordinate (orange) for each ring of six carbon atoms, equivalent in size to a benzene molecule.



(a) Functionalized tetrabenzanthracene molecule



(b) Sheared molecule

0	0	0	0	1
0	0	1	1	0
1	1	1	0	0
0	0	1	0	0
0	0	1	1	1
0	1	1	0	0
1	0	0	0	0

(c) Array representation

Figure 3.7: Representation of molecules in the model. A functionalized tetrabenzanthracene molecule (a) is represented in the model using non-orthogonal basis vectors (b), thus appearing sheared. Every atomic position is accounted for by at least one hexagon, and the relative positions of the hexagons are not changed by the shearing of the lattice. Computationally the positions of the complete benzene rings and the positions of binding functional groups (notated by letters) are stored in binary arrays (c) with a true bit.

tational efficiency and minimization of storage the coordinates are stored in an array (Figure 3.7(c)), though the array appears sheared (Figure 3.7(b)) as the coordinating vectors of the honeycomb lattice are not orthogonal.

Shown in Figures 3.8 and 3.9 are functionalized benzene and tetrabenzanthracene molecules. In these Figures, the *Full* array represents the molecule as a whole; *i.e.*, if aligned on a honeycomb lattice, where the molecule would cover the lattice. The other arrays show the positions in the full array where functional

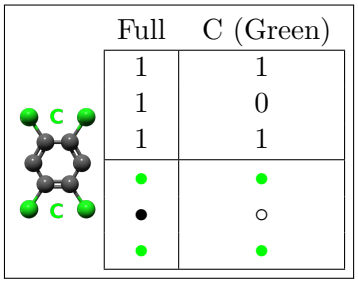


Figure 3.8: Tetrachlorobenzene molecule.

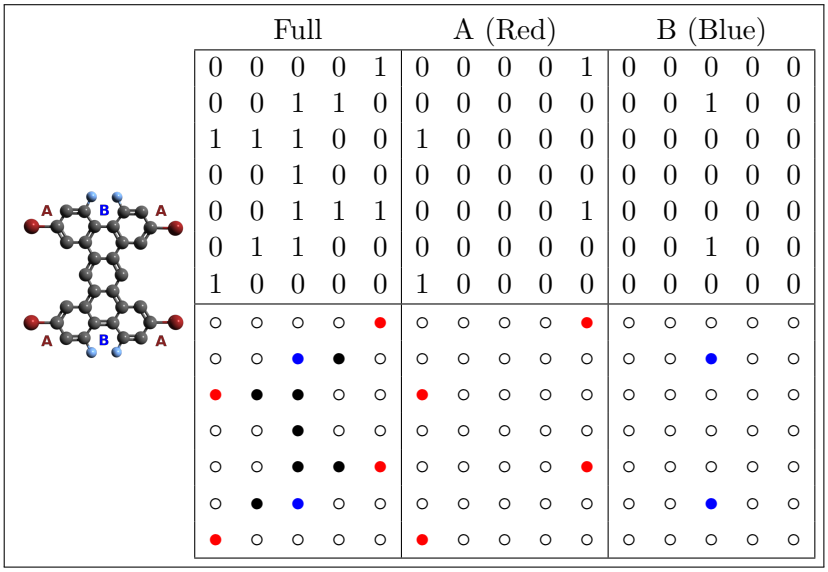


Figure 3.9: Tetrabenzanthracene molecule with functional groups.

binding groups *A, B, C* (*Red*, *Blue* and *Green*, respectively) are present.

Orientations of molecules that do not map directly onto the coordinate system are not considered. Thus, simple lattice transformations can effectively rotate the molecule in the plane of the lattice, and also through the plane of the lattice. Rotations on the lattice are thus in increments of 30°, with also a 2-fold plane inversion transformation.

Only two dimensions are considered in the simulation. Stable three dimensional structures could form stochastically in some systems, but as we are interested in understanding the underlying trend of GNR formation for given

experimental conditions, we ignore these and other off-lattice malformations. Breaking apart of the formed molecular aggregates is not possible in the model. Because the covalent bonds between molecules form by catalysis, disassembly of any carbon-carbon bonds will not occur at standard conditions for the coupling reactions (room temperature, atmospheric pressure) [83]: in fact these same reactions are used to create initiators for self-assembly [117].

We have developed a coarse-grained algorithm that geometrically describes potential interactions between molecules by a standard mathematical operation: cross-correlation. Cross-correlation of arrays representing two molecules (for example those in Figures 3.8 and 3.9) determines how the molecules can fit together on the honeycomb lattice, and calculates in what positions the molecules may form bonds together. Essentially each molecule is translated, checking for possible contacts and thus binding events between molecules, but the correlation method solves every translation simultaneously and can discount any binding events that result in unwanted overlaps between molecules. However, different orientations/rotations of the molecules with respect to each other need to be checked sequentially.

Each molecule is represented by several arrays: one array represents the entire molecule lattice positions (Full array), and others the binding positions on the Full array where the molecules may interact, with different arrays representing different types and positions of binding.

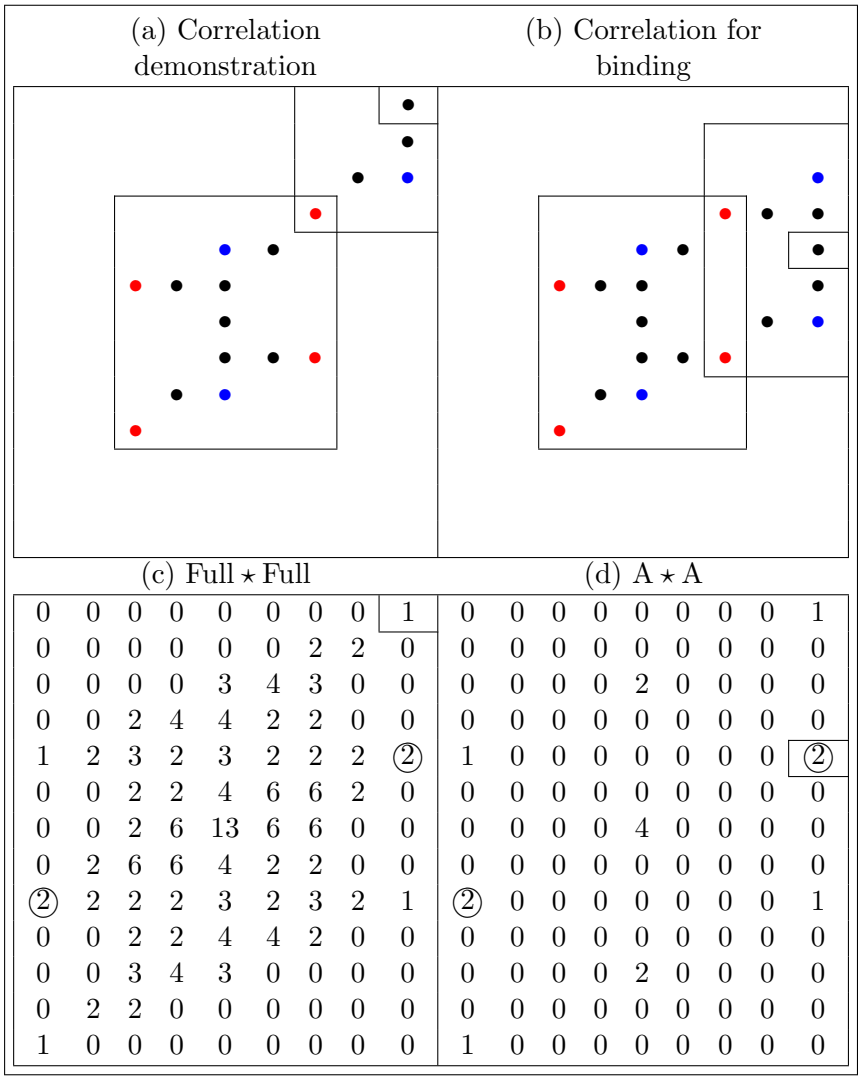
Cross-correlation of the Full arrays of two molecules (Figure 3.10(a)) gives the number of overlapping benzene rings (hexagons) for each translation overlap between the two (Figure 3.10(c)). The position of the value in the array is the physical separation between the molecules on the honeycomb lattice for the overlap calculation. Cross-correlation of binding arrays (Figure 3.10(b)) yields a sparse array similar to what might be expected when convolving Dirac

combs (Figure 3.10(d)), where the array values represent the number of binding events. However, to rule out overlapping and impossible binding configurations, the corresponding positions in the cross-correlation arrays are compared. If the values are the same, then any overlap of the molecules is due to functional groups that are able to form bonds (and no overlapping from non-binding points present): thus the binding is valid.

The number and type of bonds able to form between intermediate molecules is dependent on the possible rotation of molecules around single bonds. Therefore, an additional rule in the case of Figure 3.10(d) is imposed, in that solitary bond formation will not occur, as the molecule would be able to instantaneously form a second bond. We simplify the consideration of such events by ruling out the (unlikely) scenario that a competitor molecule would bind to the nearby site simultaneously. Effectively, adjacent bonds are not allowed to form with third party molecules until all possible bonds have been made. This stabilizes the configuration in the plane of the growth, preventing rotations of bound molecules through the plane. Hence for each binding there is a minimum of two bonds formed: this directly corresponds to the number in the binding array (Figure 3.10(d)), and two is set as the minimum number of bonds allowed to form.

A sample binding event is shown in Figure 3.11, based on the translation/overlap shown in Figure 3.10(b & d) between the model molecules (Figures 3.8 and 3.9). Arrays representing the product are produced by cross-correlating the arrays of the two reactants. For the full molecule array (Product Full), all nonzero elements will be subsequently normalized to one. For the functional groups arrays (Product A,B), all elements greater than one will be removed: these binding positions have already been considered.

Of the two matrices, $\mathbf{f}$ and $\mathbf{g}$, when a binding event is recognized as possible



0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	2	2	0
0	0	0	0	3	4	3	0	0
0	0	2	4	4	2	2	0	0
1	2	3	2	3	2	2	2	②
0	0	2	2	4	6	6	2	0
0	0	2	6	13	6	6	0	0
0	2	6	6	4	2	2	0	0
②	2	2	2	3	2	3	2	1
0	0	2	2	4	4	2	0	0
0	0	3	4	3	0	0	0	0
0	2	2	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0

0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	②
0	0	0	0	0	0	0	0	0
0	0	0	0	4	0	0	0	0
0	0	0	0	0	0	0	0	0
②	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	0	0	0
0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0

Figure 3.10: Autocorrelation of two tetrabenzanthracenes (Figure 3.9). The correlation function features a sliding window of one molecule over another that counts overlap between arrays at each position. In (a) and (b) examples are shown of a particular evaluation of the sliding window, with molecule arrays given as boxes. For (a), the overlap is 1, which is stored in the top right hand corner of the correlation array, whereas for (b), the overlap is 2, which is stored in the marked cell. The full autocorrelation results are shown for (c) the full arrays, and (d) the A (red) array. Any numerical matches between the full array and a combination of the binding arrays, indicates that a binding event is possible. In the simulation, we impose the constraint that two molecules cannot bind to the same site simultaneously, and moreover that any binding between molecules for which two bonds can form to prevent subsequent rotation, must be considered to occur immediately. Hence a minimum of two binding events, and two as number is the condition in the array—this happens twice and is circled.

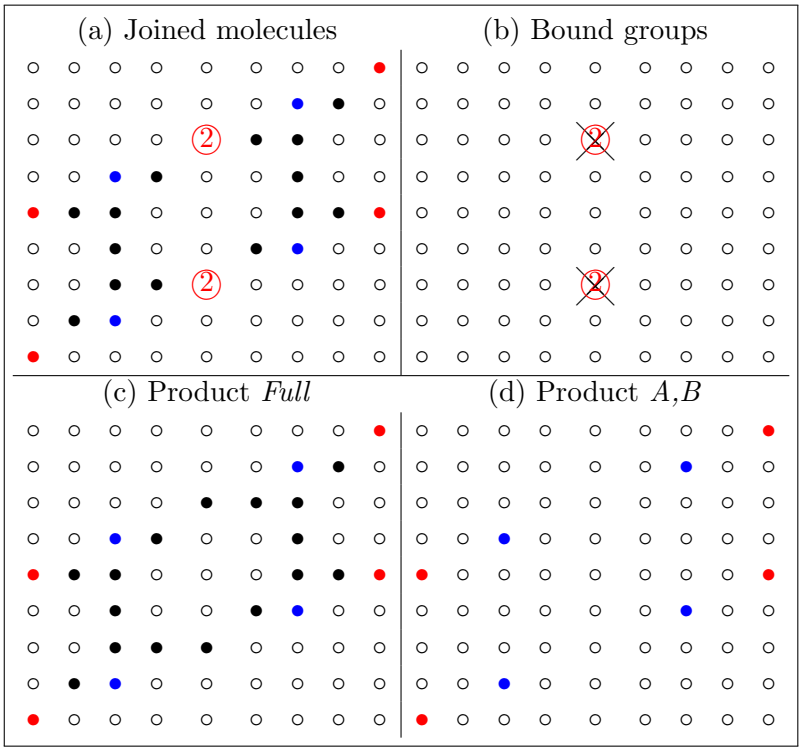


Figure 3.11: Specimen product of a binding event predicted in Figure 3.10. (a) The molecules are gathered into a larger combined array. The translation required to overlay the molecules to effect the binding is calculated from the position of the number within the correlation array. (b) Interacting functional groups occur at values not equal to 0 or 1 in the array. (c) Binding events are counted, then corrected to values of 1 in the Full array, (d) whilst being removed from the binding arrays.

in $\mathbf{f} \star \mathbf{g}(i, j)$ at element (b_i, b_j), due to the binding arrays matching the full array, a translation can be made to offset them, such that they are moved to recreate the overlap of the sliding window for that element.

The vector $(b_i - 1, b_j - 1)$ provides the translation for the matrix indices for matrix $\mathbf{g}$:

$$\mathbf{g}(i, j) \mapsto \mathbf{g}(i + b_i - 1, j + b_j - 1). \tag{3.5.1}$$

The new combined product $\mathbf{z}$ is formed thus:

$$\mathbf{z}(i', j') = \mathbf{f}(i, j) \cup \mathbf{g}(i + b_i - 1, j + b_j - 1). \tag{3.5.2}$$

Regular patterns can be simpler to calculate in reciprocal space, because convolution is elementwise multiplication in reciprocal space. Thus the binding algorithm is particularly quick as it can largely operate in reciprocal space, and reciprocal space interpretations (2-D Fourier transforms) of the molecules are stored preferentially. This allows very quick calculation of cross-correlations. The model is designed as a generalized model of molecular interactions, and is extensible to probe other model systems based on any starting mixture of molecules.

3.6 Packing of viruses

Crystallographic packing of viruses is a suitable scenario for exploiting the characteristics of the cross-correlation function.

As discussed in Chapter 5, the protein capsids of most viruses have icosahedral symmetry: the regularity of the virus symmetry is derived from quasi-equivalent subunits formed from a small number of proteins. Furthermore, this regularity of shape means that viruses often can crystallize, even *in vivo* [118]. However, icosahedral symmetry is incompatible with seamless 3-D crystalliza-

tion due to the *crystallographic restriction theorem* [119], and thus cannot form a perfect crystal lattice. Quasi-equivalence is needed because the number of subunits exceeds the order of the finite symmetry group. It would also occur in a crystallographic setting, *e.g.* for octahedral symmetry.

In 1984, Shechtman's discovery of icosahedral quasicrystals led to the development of a different crystallographic characterization, based on a geometrical embedding of the structure into a higher-dimensional Euclidean space in which icosahedral symmetry is crystallographic [120]. This development led to new mathematical tools being developed in order to study structures with noncrystallographic symmetry, and these have been used to great effect to study virology in a large range of contexts, from virus structural transitions during maturation or infection [121], to capsid structure [122, 123]. Here we will consider the latter application embedded into 6-D Euclidean space, the lowest dimension for which there exists an icosahedrally symmetric lattice and the cut-and-project method is possible[2] [120].

The ability of icosahedral non-enveloped viruses to crystallize by clustering *via* non-covalent interactions depends on electrostatic interactions between them and hence largely on the properties of the capsid surface. The work of Janner [123] determined that the space-group symmetry adopted on crystallization depends not only on the overall morphology of a virus, but on the level of detail of the surface moieties present at select positions on the capsid. Differences in the surface characteristics between strains are well studied in virology, due to the predominance of antigenic epitopes on accessible external moieties, and the important role these components can play in the viral life cycle. The amount of detail available facilitated a detailed examination of the connection between the geometries of viral crystallization and their antigenicity, which went beyond

[2]The icosahedra could also be embedded into a 5-D Euclidean space, but then the representation theory does not allow for the two 3-D spaces required for cut-and-project.

what was expected for similar particles.

The crystallization is considered as sphere packing; positions in the packing where two spheres of the same size touch are known as kissing points [124]. An example of kissing points, for human rhinovirus B (HRV-B) determined by Janner is given in Figure 3.12. This approach holds for the symmetric case when asymmetric electrostatic effects do not apply: either for virions that are symmetrically-charged, or for viruses for which any asymmetric charge distribution is not shared across the population. Note that the second stipulation would not hold in the case of a small number of conserved asymmetric charge distributions. As we shall discuss in Chapter 5, the case of a conserved asymmetric genome organization within a viral capsid is a result of the role the genome plays in the assembly of the capsid. This asymmetric conformation may induce electrostatic effects in the stacking of viruses, resulting from unshielded genome negative charge. If only a few organizations are present in the population of viral particles, this would allow a non-icosahedrally-symmetric crystal packing.

A necessary condition for the asymmetry to propagate through the crystal and not be heterogeneously dispersed is that a static unit must form between a small number of virions, followed by packing of these units (that act akin to a unit cell). This is explored further in Chapter 6, where this argument has been used to identify an asymmetric X-ray crystal structure for STNV.

There are principally two scenarios for which cross-correlation can be used in the context of an asymmetric genome organization. Firstly, in reference to the function's textbook use (a gold-standard measure of similarity between signals), it could be used to determine the similarity of predicted genome organizations. Consensus organizations, orientated with respect to kissing points, could be correlated with each other and compared to the asymmetric experimental data. This could unravel the genome organizations present within the unit cell. Sec-

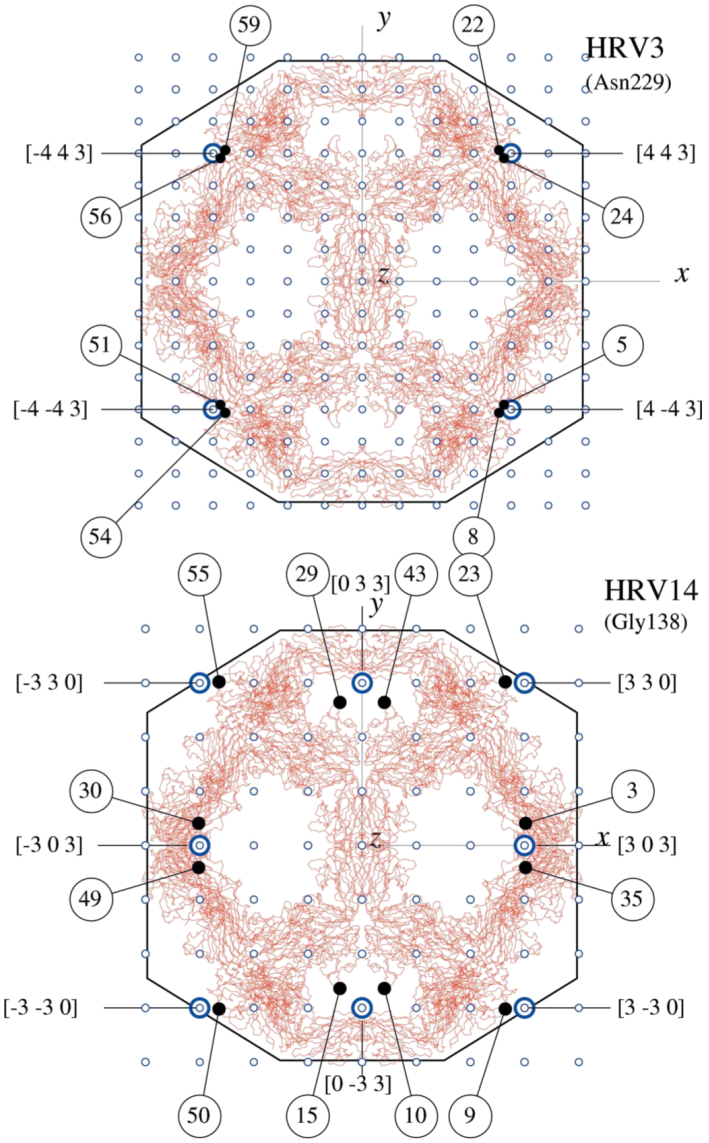


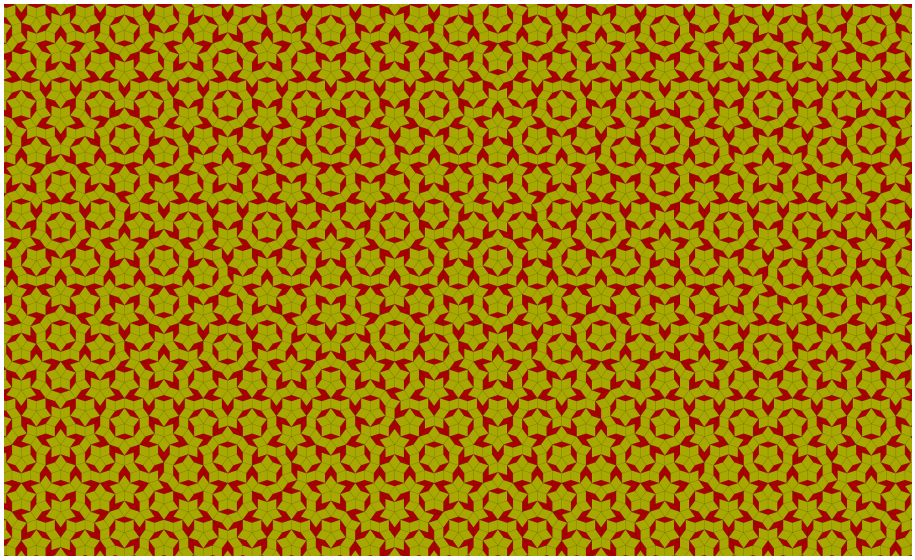
Figure 3.12: Kissing point positions (blue double circles) belonging to the packing lattices of two strains of HRV-B, which differ even though the overall architecture of the serotypes is the same, as demonstrated by the VP2 (see Figure 5.3) morphology and arrangement (red). Also shown are the position of the VP2 residues at minimal distance to the kissing points (black dots), which are Asn229 for HRV-B3 and Gly138 for HRV-B14. Figure is taken from [123].

ondly, a self-assembly stochastic approach could be taken, with crystallization occurring at kissing points, and with the strength of interactions given by a weighted function based on genome organization in proximity to capsid inferred from the knowledge of RNA-CP interactions. A "virtual X-ray diffraction" on the crystal would recreate an asymmetric structure, and this could be compared to the experimental data.

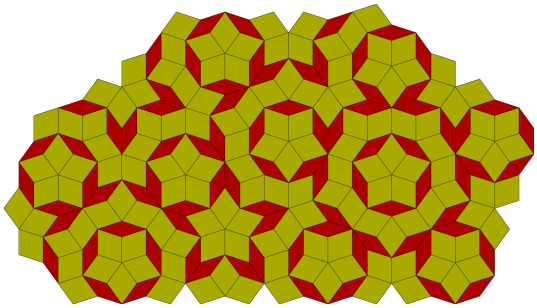
The latter approach is identical to the others presented in this chapter, and would be an interesting investigation, particularly as it would be best to carry out the simulations in 6-D, as an embedding of a 3-D process into 6-D, where the problem is crystallographic. This is because icosahedra tile 6-D space, so the cross-correlation operations could be carried out on sparse matrices, as in the other examples in this chapter. However, the former approach provided a simpler, deterministic rather than stochastic examination of the asymmetric density; we provide results for this in Chapter 6.

3.7 Discussion

Cross-correlation as a mathematical function appears in many guises throughout spatial aspects of scientific research, as it is the gold standard for measuring spatial or structural similarity between data. As an algorithm to study self-assembly, it shows promise when the necessary information can be reduced to operations on a lattice, as the function requires aligned matrices. Interpolation between non-aligned data would be possible, but would reduce the performance of the algorithm significantly. Also, one of the benefits of the algorithm, that of relying on integer Cartesian coordinates, would be lost in this approach and also sampling and rounding errors would become of concern. It is therefore best suited for interactions occurring regularly, positioned with respect to a lattice. As we have seen, the lattice does not need to have an orthogonal basis (§3.5),



(a) Penrose tiling, aperiodic in 2-D



(b) Self-assembling aperiodic cluster of rhombs

Figure 3.13: Penrose tiling, as an example for using algorithm in higher dimension. (a) In 2-D the tiling is aperiodic, but in 5-D the tiling is crystallographic. (b) Kinetic self-assembly of the tiles could still be achieved, but would require 5-D space for a perfect lattice to tile.

or even have to correspond to spatial orientation: the second dimension in §3.3 only represents the twisted orientation of the subunits. More complexity can be included by representing additional degrees of freedom as additional dimensions. This is particularly seen in icosahedral stacking, which is crystallographic in 6-D. There are many other examples of tiling that could be approached in this manner, for example aperiodic Penrose tiling in 2-D could be modelled in 5-D or more (Figure 3.13) [125].

Chapter 4

Synthesis of nanographene *via* self-assembly

The future possibility of high-performance electronics based on graphene has been contemplated since the discovery of the material's extraordinarily high crystal and electronic qualities. A 2-D[1] semi-metallic material, its precisely ordered honeycomb lattice of carbon atoms results in an unusual band structure and unparalleled electron mobility [126, 127]. With widths on the order of tens of nm, GNRs have band gaps due to quantum confinement and edge effects [128, 129]. The band gap of GNRs is tuneable by the ribbon width [130–132], indicating that GNR semiconductors could one day replace silicon in commercial transistors [133].

GNR production has been attempted using many different techniques, but all are limited by the high dependence of the characteristic properties on the GNR width, patterning and edge disorder [129, 134, 135]. In particular, any disorder within the GNR or at its edge will have a great impact on the electronic properties of the material [136]. Lithographic etching from bulk graphene sheets

[1] For all abbreviations see the glossary on page 222.

is limited in both width and smoothness by the resolution of the lithography equipment [129]. Unzipping of carbon nanotubes faces similar problems with edge disorder [137]. Existing bottom-up synthesis procedures such as epitaxial growth (on a substrate) [135] or carbon vapour deposition [138, 139] do not control crystallinity or shape. None of these techniques can currently manufacture GNRs to with the edge quality required for commercial electronic applications.

Patterning of the GNR will also adjust the band gap: these *nanoporous* GNRs are useful for many applications [140]. Methods of manufacturing patterned GNRs that preserve regularity are a desirable alternative to those for pristine GNRs, as they still allow engineering of the band gap [141–143]: patterning (such as the example in Figure 4.1) can lead to a greater band gap that is conserved between GNRs of different widths [144, 145]. This can be seen in Figure 4.2, which has been calculated with a *tight-binding* (TB) model[2], parameterized as in Hancock *et al.* [145]. At present, no techniques have been developed to controllably manufacture these patterned GNRs with atomic precision, though there are recent experimental breakthroughs that partly fulfil the necessary requirements. One such approach pioneered by Klaus Müllen entails building an aromatic seeding structure, to which benzene-based compounds can attach (covalently), forming GNRs with high purity [17, 146, 147]. By configuration of the initial backbones, the structure of GNRs can be controlled. GNR formation is by a two-step surface-assisted process of dehalogenation followed by cyclodehydrogenation, which is well characterized in theoretical papers [17, 148]. However, this technique is limited in the GNRs that are possible to synthesize by the producibility and shape of the initial "backbone" of the two-step reaction.

We suggest that of the possible alternate approaches to manufacturing pat-

[2]The TB model code used for this calculation, incorporating a Hubbard-U term, was developed by Jack Baldwin and Yvette Hancock at the University of York, and is parameterized as in Hancock *et al.* 2010 [145].

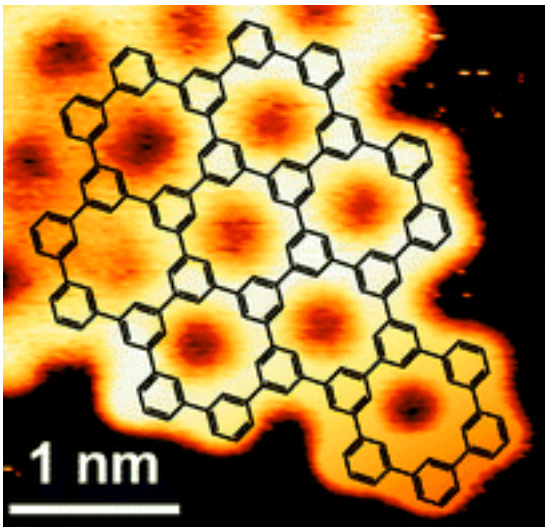


Figure 4.1: Patterned nanoscale graphene. There exists the opportunity to engineer patterns into nanoscale graphene and GNRs, such as this example from Bieri and colleagues [144].

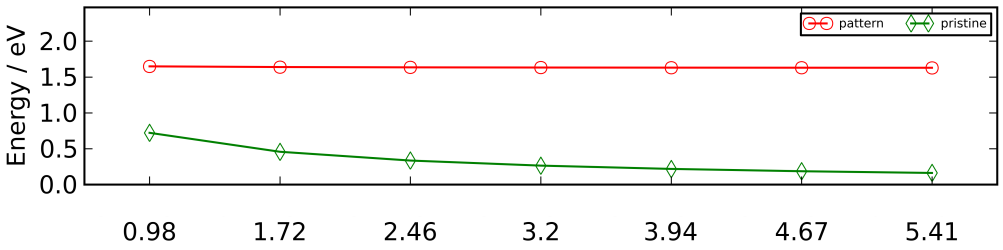


Figure 4.2: Armchair-orientated GNR band gaps (eV) change with the width of the ribbon (nm). Patterned ribbons from Figure 4.3 (red) maintain a band gap irrespective of width, whereas for pristine GNRs (green), wider ribbons have smaller band gaps than narrower ribbons. Calculated using a TB model developed by Baldwin and Hancock [145].

terned GNRs, a self-assembly system [10, 97, 103, 104, 149–151] offers the largest scope for engineering a GNR with desirable characteristics. Suitably designed liquid- or gas-phase aromatic precursor molecules could self-assemble by aryl-aryl coupling into a GNR, on a metal substrate [97, 152]. Covalent binding between the precursors would be *via* so-called *click-chemistries* (modular, stable, high-yielding reactions) between functional groups, such as the Suzuki-Miyaura and Ullmann methods of biaryl coupling [83, 152]. The GNR products would depend on the experimental setup: *e.g.* temperature, concentration of initiator molecules, catalysis, and time of reaction.

Control over the size and patterning of synthesized GNRs is desirable; for each possible GNR self-assembly synthesis method, several different types of GNR could be produced, depending on the experimental setup. Designing the synthesis protocols is a very broad problem, and the search space is suitable for computational approaches. For the purpose of demonstrating the KSA method applied to the directed self-assembly of GNRs, we have chosen to run the algorithm on a single model system. The initiators of self-assembly in this system are based on the molecules of tetrabenzanthracene and benzene (Figure 4.3), and are designed so that they will tessellate into regular patterns, and thus assemble into patterned ribbons. The covalent bonds formed between molecules in the model system are assumed to be stable in the conditions of the coupling reactions to allow the coupling to occur, and thus do not break apart once formed. The use of tetrabenzanthracene is motivated by the possibility of its synthesis with functional groups [117, 153], and previous studies on its aggregates [154, 155]. Another particular incentive for the model system is the predicted band gap of the patterned GNR it creates: it is higher than a pristine GNR of the same width, and also is less dependent on the width of the GNR (Figure 4.2).

Here, we demonstrate a KSA computation in the context of GNR formation

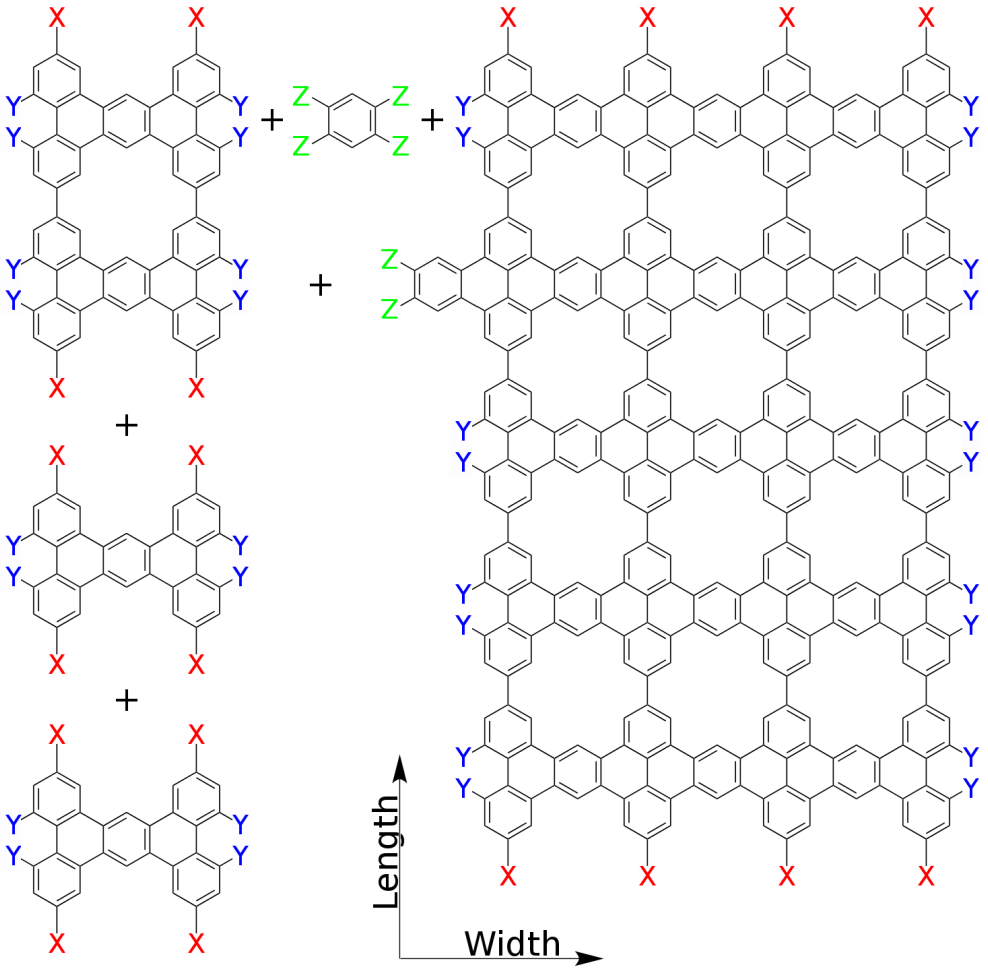


Figure 4.3: Nanporous GNR produced by tetrabenzathracene and benzene coupling. The design calls for A–A homocoupling (symmetric) and B–C heterocoupling (asymmetric) click-chemistry reactions, though these are not explicitly modelled. Preferred ribbon growth direction is indicated as the length direction, corresponding to A–A coupling. Identical small holes are spaced regularly in the GNR. This will allow an armchair edge to the GNR, and a band gap of 1.6eV to transport of electrons this direction (see Figure 4.2).

on a model system of molecules, using a Gillespie algorithm to form a network of possible coupling reactions, and exploiting a novel way of representing their geometries to speed up the simulations. The model does not allow aberrations such as folding to form within the GNR. There exist computational models of GNR formation, that utilize a MD approach [156,157]. These studies consider folding and twisting of GNRs as their primary objective rather than self-assembly: additionally buckling of existing GNRs is considered by other studies [158,159]. Our computation is coarse-grained and much better suited to predictive approach in combination with experimental data. Assuming free non-directional diffusion on a homogeneous surface and a coupling-limited reaction (see 4.1) [97], we investigate the model system with the new implementation of the Gillespie algorithm, showing how synthesis of GNRs can be controlled by laboratory variables.

4.1 Model system

Functionalized polycyclic aromatic hydrocarbons based on tetrabenzanthracene and benzene are used for the initiators of the model self-assembly system. Coupling occurs between the functional groups present on all initiators. A combination of several coupling mechanisms is needed to allow control over the direction of GNR growth: there needs to be two or more different reactions to control growth in two dimensions. The Ullmann reaction has long been the method of choice for chemists to generate a carbon-carbon bond between two aromatic nuclei [83], as steric hindrance prevents the standard technique of oxidative substitution in the presence of Fe(III)Cl_3 [160]. Aryl halides in the presence of activated copper will form a biaryl and a copper halide [83]. The reaction will only produce symmetrical biaryls, and so other coupling methods would be needed were asymmetric initiators to be used. For example the Suzuki-Miyaura reaction [83] is an asymmetric coupling between halide and boronic acid functional

groups, catalysed by a palladium(0) complex. The exact method of couplings in the system remains unspecified in the computational model, as the energetics of coupling remains one of the parameters in the kinetic simulation. Rather, suitable coupling reactions will be chosen that can fulfil the requirements for the synthesis process suggested by the kinetic modelling. Ideally a small number of pristine GNRs will be produced once self-assembly has equilibrated.

The chemical scheme for the Ullmann reaction is:

$$\underset{\text{Aryl–Halide}}{\mathrm{Ar^1{-}X}} + \underset{\text{Aryl–Halide}}{\mathrm{Ar^2{-}X}} \xrightarrow{\mathrm{Cu}} \underset{\text{Biaryl}}{\mathrm{Ar^1{-}Ar^2}} \tag{4.1.1}$$

and the scheme for the Suzuki-Miyaura reaction is:

$$\underset{\substack{\text{Aryl–Boronic}\\ \text{acid}}}{\mathrm{Ar^1{-}B(OH)_2}} + \underset{\text{Aryl–Halide}}{\mathrm{Ar^2{-}X}} \xrightarrow{\mathrm{Pd(0)}} \underset{\text{Biaryl}}{\mathrm{Ar^1{-}Ar^2}} \tag{4.1.2}$$

where X represents any halide, and Ar groups refer to the rest of the aryl molecules.

The initiators are designed such that they are small enough to be soluble in an organic solvent: a well-mixed diffusing process for the initiators is envisaged. This is because the self-assembly interactions ought to be *coupling-limited*, not diffusion-limited [97, 99]. Coupling-limited reactions, where the rate-limiting step of GNR growth is the bond breaking/forming, allow fine-tuning of the resulting products [97]. Conversely, diffusion-limited reactions are universally unordered [93, 94, 99], as there is no regulation of growth, which happens spontaneously. An unordered molecule cannot subsequently reform to be more kinetically stable, due to the strong intramolecular covalent bonds. The initiators and growing GNRs ordered on the inert substrate can diffuse across the surface,

with the rate of diffusion dependent on the surface and molecule [161,162]. Free diffusion is expected for coupling-limited reactions [97]: the energy of activation for diffusion (0.2eV–0.4eV [97]) is in the regime that the assembly would be coupling-limited. The coupling-limited nature of the interactions justifies a well-mixed free-diffusion approximation in the simulation.

When modelling, we do not consider the interactions of initiators with the surface, as the energetics of absorption, diffusion, and desorption are orders of magnitude less than the coupling reactions. We also disregard any directional diffusion, which can occur with certain functionalization and surfaces [163,164].

For planar self-assembly, the polycyclic aromatic hydrocarbon initiators used must have a strong surface interaction: a weak interaction would mean that intermolecular interactions would dominate, and the planar orientation would not be preserved [97, 165]. For example, at graphite surfaces initiators may self-assemble "edge-on" due to weak molecule-surface interactions. In contrast, on Au(111) the stronger interaction between the surface and the molecules allows flat planar growth [165]. We therefore assume an innately strong interaction with the inert surface, or else the presence of functional group anchors: functionalization with thiols or other groups may be required to anchor the GNR to a substrate as it assembles out of solution [95, 166–168]. It is noted that any functional groups or adatoms could change the electronic structure of the resulting GNRs, but are not specified explicitly in the model.

Various studies have demonstrated that polycyclic aromatic hydrocarbons can diffuse across inert metallic surfaces, with the rate of diffusion dependent on the surface and molecule [97, 140, 162, 163, 169]. For coupling-limited self-assembly, diffusion should be energetically more favourable than coupling. The Ag(111) surface has the largest energy barrier for the recombination of halides, but a small diffusion barrier [97,162], making it particularly suitable for covalent

self-assembly.

Both Ag(111) and Au(111) have similar barriers in terms of sliding diffusion, with the sliding diffusion of phenyl calculated at around 0.25eV [162]. Pentacene has a very small diffusion energy barrier at 2.6×10^{-2}eV on Au(111) [161]. These diffusion values are in the regime such that the assembly would be coupling-limited.

Halide functional groups are separated from initiators upon binding to metallic surfaces. The dissociation energy for iodine is lower than that of bromine [162]: precursors functionalized with bromine and iodine can be utilized on Au(111), Ag(111) and Cu(111). Preferential dissociation of these functional groups could be utilized to enable directional growth. Though the halogens are able to desorb atomically from Ag(111) and Au(111), halogen by-products could interfere with the formation of ordered structures.

Unidirectional diffusion, where the movement of the adsorbed molecule on the surface is limited to a certain path, has been demonstrated with dithiolanthracene on Cu(111) [163]. Thiols are well-known methods of attaching molecules to metallic surfaces, where a gold-sulphur bond is formed with a strength of 1.9eV. However, there is no evidence for utilizing thiol attachments for planar surface assembly when halide and borate groups are present [140,169], the functional groups used for Suzuki-Miyaura and Ullmann coupling reactions. This does not preclude the possibility of thiol-terminal ribbons grown perpendicular to the plane of the surface using the same coupling reactions.

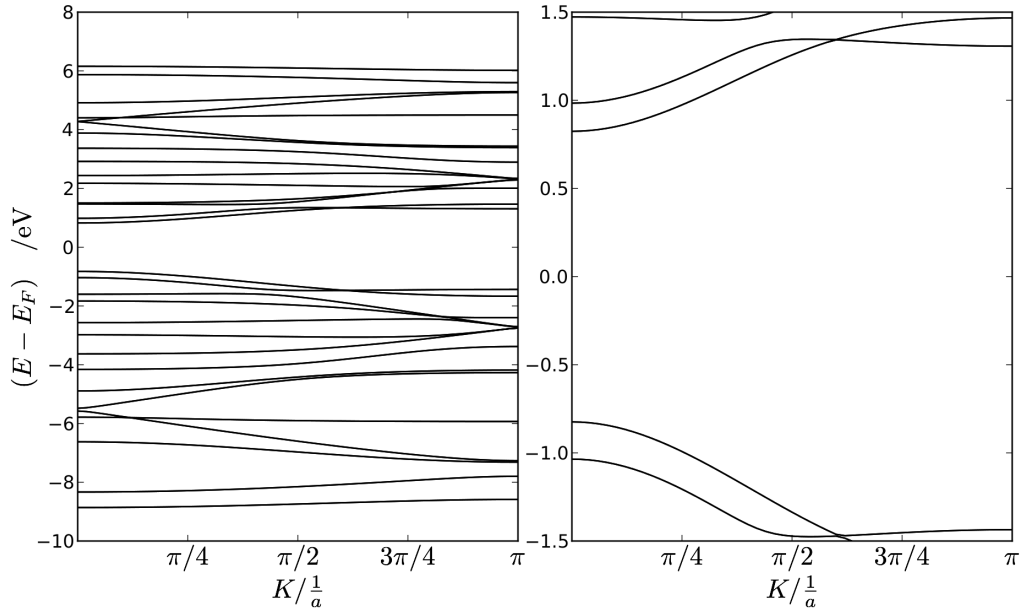
In summary, the model surface was assumed to be energetically homogeneous, with a uniform energy of interaction existing between a molecule and an absorption site across the surface [103, 104]. Notwithstanding the surface interaction, free diffusion is still assumed on the surface; as a consequence of the surface homogeneity, the total energy of interaction with the surface does not

change. For this reason, the molecule-surface interaction is disregarded in the model.

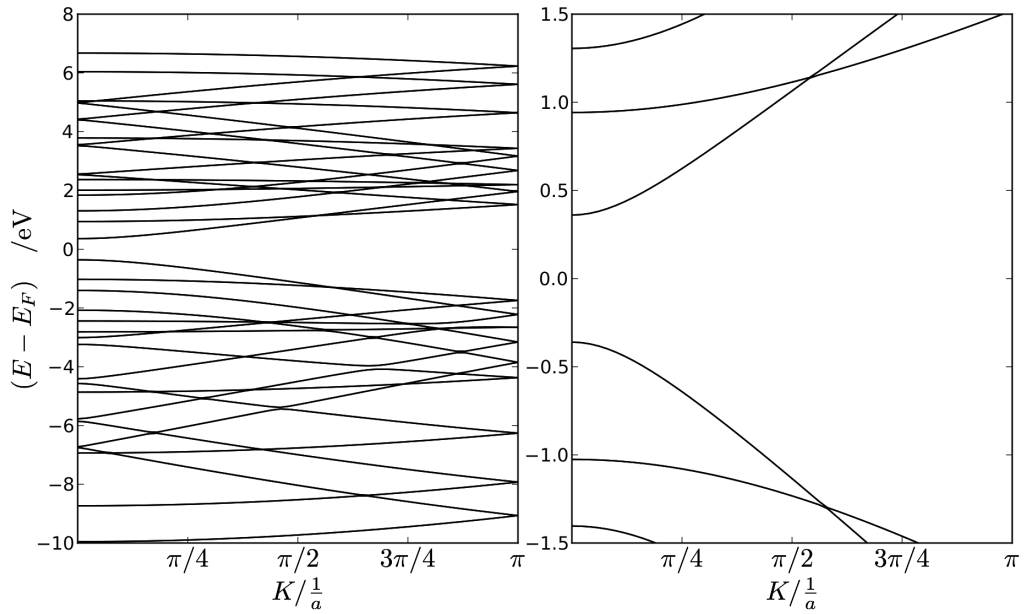
4.2 Designer GNRs

As discussed, the properties of GNRs are highly sensitive to the width, patterning, and edge structure of the ribbon, making these systems amenable to engineering. The patterned GNR introduced in this chapter was designed to meet a specific electronic requirement: its band gap (Figure 4.2). It was possible to design the GNR by predictive modelling of patterned GNR band gaps using a generalized minimal TB quantum mechanical model for nanographene, that has DFT accuracy [145]. The TB model is more computationally tractable than DFT and can be used to efficiently calculate the electronic band structure of large experimentally-relevant structures such as those introduced in this chapter. From a design perspective, accurate and efficient calculation of GNR electronic properties is advantageous for assisting the experimental engineering of GNRs with desirable characteristics.

At single width the ribbon is calculated to have a predicted band gap that is approximately 1.6eV: it is higher than a pristine GNR of the same width (Figure 4.4). The band gap is also is less dependent on the width of the GNR, as can be seen in the example widened band structures in Figure 4.5. The general trend for this ribbon is reported (as discussed previously) in Figure 4.2, where it is shown that the band gap remains in the patterned ribbons, irrespective of their width. For other designs of ribbons, this would not necessarily hold: it is a key feature of the ribbons we introduce in this chapter.

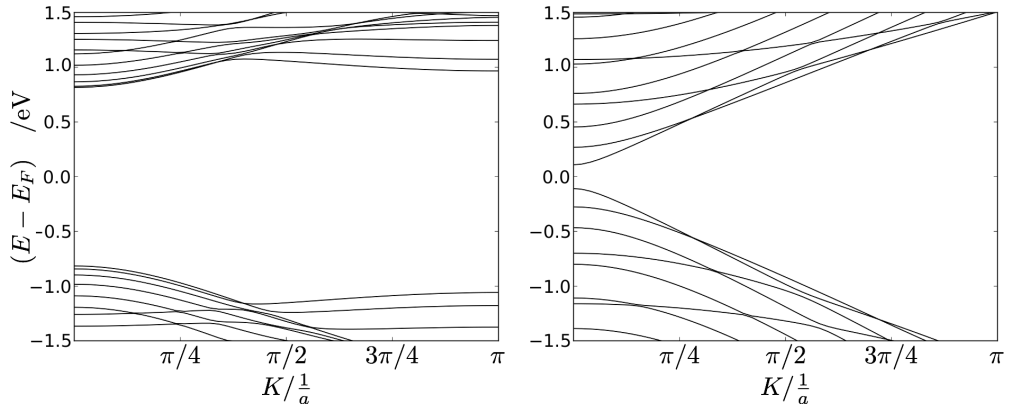


(a) Patterned ribbon: electronic band structure



(b) Pristine ribbon: electronic band structure

Figure 4.4: Electronic band structure of patterned and pristine single width ribbons, *i.e.* 0.98nm wide. Band structures calculated with a generalized TB model for GNRs, parameterized as in [145]. Shown is the band structure, and also the zoomed-in band structure to the area around the Fermi energy. The patterned ribbon has a band gap of approximately 1.6eV, whereas the pristine ribbon only has a band gap of 0.75eV.



(a) Patterned (b) Pristine

Figure 4.5: Example electronic band structures around the Fermi energy of patterned and pristine ribbons that have grown to a width of 3.94nm. Calculated with a generalized TB model for GNRs and parameterized as in [145]. At this width, the patterned ribbon maintains a band gap of approximately 1.6eV, whereas the pristine ribbon band gap has reduced to 0.25eV.

4.3 Methods

As discussed in Chapter 3, we have developed a coarse-grained algorithm using the cross-correlation function that geometrically describes potential interactions between molecules. The cross-correlation function is used between arrays representing the molecules in the simulation, and determines how the molecules can assemble on the honeycomb lattice (Figure 4.6). In particular, the cross-correlation approach finds the coordinate positions on the lattice where molecules may form bonds together. The advantage of using a cross-correlation approach is that the function can be carried out in reciprocal space: under the convolution theorem, real space convolution is elementwise multiplication in reciprocal space. This is exploited by the algorithm, and the quick calculation of cross-correlations allows fast simulation of binding events.

Even with increased speed arising from generalizing interactions into reciprocal space, a deterministic approach to solving differential equations describing

$$\begin{array}{c} \\ \text{TBA} \\ \text{TCB} \end{array} \begin{array}{c} \begin{array}{cc} \text{TBA} & \text{TCB} \end{array} \\ \begin{bmatrix} 8 \times (AA)^2 & 8 \times BC \\ \text{——} & \text{None} \end{bmatrix} \end{array}$$

Figure 4.6: Geometries of interaction: types of bonds able to form between the TBA (tetrabenzanthracene) and TCB (tetrachlorobenzene) molecules, found by the cross-correlation algorithm. TBA–TBA binding is possible in 8 different rotations and translations, each time forming $2 \times AA$ bonds. TBA–TCB (BC) bond formation is possible in 8 different rotations.

the system interactions is not possible: there is an infinite number of possible products with unlimited size, not controlled by any end step. A Gillespie algorithm [24], a stochastic modelling approach can be implemented at $O(n)$ where n is the number of trial molecules in the simulation population. In the algorithm, only interactions possible within the population at that moment are considered, forming a reaction network: reactions are then chosen stochastically, using a random number generator to pick from a weighted list of reactions; the population changes over time. The approach decreases the search space of the computation, because fewer interactions are considered as the number of molecules decreases over time. For our implementation of the algorithm, the relative likelihood of binding due to the mobility and size of the components was not considered; we assumed that the reactions were not diffusion-limited. These factors are not usually included in Gillespie simulations, which assumes a well-mixed homogeneous volume, but the model can be extended to include diffusion [170, 171].

For kinetic simulation of the system, the rate constants for bond formation between the self-assembling molecules are needed. The rate constants are found from consideration of the Arrhenius equation (Equation (2.1.1)), which links the activation energy of the reaction with the temperature and rate constant. Though the coupling chemistries considered in the simulation often require catalysis, the activation energy for each reaction in the Gillespie network is calculated without catalysts being considered explicitly. As this is a minimal model, we do

not include catalysts explicitly; perturbing the activation energy is a good first-order approximation for the effects of the catalysts. Thus the sum of binding activation energies alone is used to calculate the rate constant for each reaction.

Each pair of molecules will have several orientations of reaction; these are found using the cross-correlation algorithm (Chapter 3). An example for the model system before any reactions have taken place is given in Figure 4.6: the types of bonds able to form are eight different rotations and translations that allow tetrabenzanthracene-tetrabenzanthracene coupling, and eight different orientations/translations facilitating tetrabenzanthracene-benzene coupling. Once there are other molecules/aggregates in the simulation, the interactions between all of these species will also be found using the cross-correlation algorithm.

From here, we will have two snapshots to explain the rest of the algorithm: one of the system before any reactions (Figure 4.7), and one at an example snapshot in time of a general system of four molecules (Figure 4.8). As discussed, all the possible interactions that form bonds are calculated as to the rate constants using the Arrhenius equation. The rate constants are calculated with a pre-exponential factor of $10^9 \mathrm{s}^{-1}$ [164]. These rate constants of all of the different orientations between two molecules are added together to be saved in an array (Figures 4.7(a) and 4.8(a)). Note that in these figures, the array showing the rate constants of interaction is of size $N \times N$, even though the system contains fewer molecules at that moment; N is chosen to be larger than the number of different intermediates possible during the simulation, and is kept constant. The rest of the $N \times N$ array is blank, and during the simulation rows and columns are utilized in the case of new molecules, as needed.

The time-dependent concentration of reactant molecules is stored as a single-dimensioned array of length N (Figures 4.7(b) and 4.8(b)). Assuming a well-mixed and diffusion-limited system, the rate of pairs of system intermediates

	TBA	TCB	$\cdots$	N
TBA	3.661×10^{-02}	8.583×10^{-03}	$\cdots$	0.0
TCB		0.0	$\cdots$	0.0
$\vdots$			$\ddots$	$\vdots$
N				0.0

(a) Rate constant array.

TBA	TCB	$\cdots$	N
1000	1000	$\cdots$	0

(b) Unit number.

	TBA	TCB	$\cdots$	N
TBA	499500	1000000	$\cdots$	0
TCB		499500	$\cdots$	0
$\vdots$			$\ddots$	$\vdots$
N				0

(c) Encounter array.

	TBA	TCB	$\cdots$	N
TBA	1.829×10^{04}	8.583×10^{03}	$\cdots$	0.0
TCB		0.0	$\cdots$	0.0
$\vdots$			$\ddots$	$\vdots$
N				0.0

(d) Binding rate array.

Figure 4.7: Example of probability calculation, for the initial configuration of the experiment, before any reactions have occurred. For the molecules TBA (tetrabenzanthracene), and TCB (tetrachlorobenzene). Throughout, the upper half-triangle of arrays is used to prevent duplicate counting when summing over the whole array. (a) Rate constant array: elements in the array show calculated rate constants of binding, calculated using the Arrhenius equation for the test activation energies $Ea_{AA} = 0.45$ and $Ea_{BC} = 0.95$. (b) Number of units: a snapshot of the number of molecular units for the sample calculations. The concentration is implied from the constant volume of the simulation. (c) Encounter array: derived from the (b) concentration array using Eqns. 4.3.1 and 4.3.2. (d) Binding rate array: elementwise calculation of rate of reaction using (a) rate constants and (c) likelihood of encountering different molecules.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & \cdots & N \end{array} \\ \begin{bmatrix} 0.2 & 2.3 & 0.0 & 0.6 & \cdots & 0.0 \\ & 3.5 & 2.2 & 1.1 & \cdots & 0.0 \\ & & 0.0 & 2.1 & \cdots & 0.0 \\ & & & 0.3 & \cdots & 0.0 \\ & & & & \ddots & \vdots \\ & & & & & 0.0 \end{bmatrix} \end{array}$$

(a) Rate constant array.

$$\begin{array}{c} 1 \quad 2 \quad 3 \quad 4 \quad \cdots \quad N \\ \begin{bmatrix} 202 & 11 & 26 & 2 & \cdots & 0 \end{bmatrix} \end{array}$$

(b) Unit number.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & \cdots & N \end{array} \\ \begin{bmatrix} 40602 & 2222 & 5252 & 404 & \cdots & 0 \\ & 110 & 286 & 22 & \cdots & 0 \\ & & 650 & 52 & \cdots & 0 \\ & & & 2 & \cdots & 0 \\ & & & & \ddots & \vdots \\ & & & & & 0 \end{bmatrix} \end{array}$$

(c) Encounter array.

$$\begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & \cdots & N \end{array} \\ \begin{bmatrix} 8120.4 & 5110.6 & 0.0 & 242.4 & \cdots & 0.0 \\ & 385.0 & 629.2 & 24.2 & \cdots & 0.0 \\ & & 0.0 & 109.2 & \cdots & 0.0 \\ & & & 0.6 & \cdots & 0.0 \\ & & & & \ddots & \vdots \\ & & & & & 0.0 \end{bmatrix} \end{array}$$

(d) Binding rate array.

Figure 4.8: Example of probability calculation at a midway point during a simulation run. Throughout, the upper half-triangle of arrays is used to prevent duplicate counting when summing over the whole array. (a) Rate constant array: elements in the array show calculated rate constants of binding, where each corresponding row/column index represents one of N molecules in the mixture. Example of four molecules summed rate constants of interaction with each other. The information is stored as a symmetric size $N \times N$ array, where there are less than N molecules in the population, and where each corresponding row/column index represents one of the N molecule types in the mixture. (b) The time-dependent number of molecules for the example four types. The concentration is implied from the constant volume of the simulation. (d) Binding rate array: elementwise calculation of rate of reaction using (a) rate constants and (b) concentrations.

forming bonds (*e.g.* arbitrary molecules P&Q) is proportional to their concentrations, though reactions of homogeneous species (P&P) have a correction term:

$$r_{PQ} = k_{PQ}[P][Q] \tag{4.3.1}$$

$$r_{PP} = k_{PP}\frac{[P][P-1]}{2} \tag{4.3.2}$$

Calculation of the rates using Equations (4.3.1) and (4.3.2) is demonstrated in Figures 4.7(d) and 4.8(d), where a 2-D array stores the rates between pairs of reactants. Only the upper quadrant of the array is needed, as it holds degenerate information: each quadrant represents a single count of each reaction rate. The sum of the quadrant represents the current total reaction rate of the system, from which the time until the next reaction can be simulated. Within this timestep until the next reaction, the rates are normalized to give the probabilities for each reaction in the range $(0, 1)$.

These probabilities are used in the Gillespie algorithm to stochastically select the next occurring reaction. This reaction is then simulated in the system: the reactants form bonds at a stochastically sampled position (see Chapter 3), and the simulation timer is advanced by the timestep. This completes the Gillespie simulation step.

The simulation finishes when either there are no possible reactions, or any possible further reactions would be likely to take more than the characteristic timestep (120s) to occur; this would afford the experimentalist time to terminate the synthesis process. The simulation can be viewed as a network of reaction pathways between intermediate states. At regular intervals during the simulation, the state of the system is saved. This stores data about the time evolution of the system and reaction network, which can be subsequently analysed. We only analyse GNRs present at the end of each simulation run: the final products

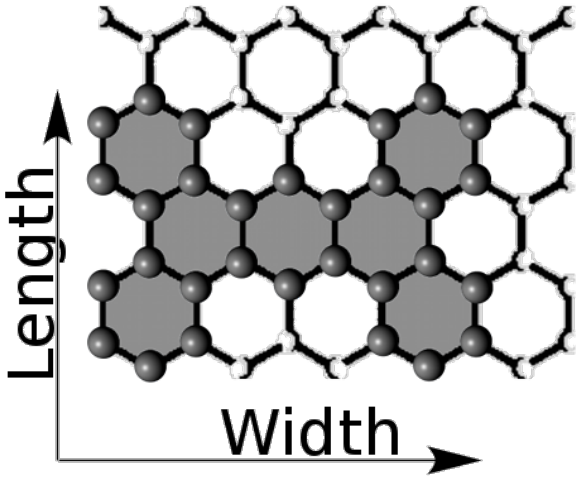


Figure 4.9: Measuring Width and Length of the growing GNR. Width, W, and length, L, are defined by the number of whole hexagons covered in each direction marked: in this example, $W = 4$ and $L = 3$. Area is defined as total number of whole hexagons: $A = 11$, and can be calculated from L and W by Equation (4.3.4). The number of occupied benzene rings (shaded), $N = 7$, allows the occupancy O to be calculated by Equation (4.3.5).

of the synthesis experiment.

Both the product and growing GNRs' properties are described mathematically, in order to assess their characteristics for favourable traits. Ribbon length, L, and width, W, are monitored, as is the number of benzene rings, N, covered by the molecule (Figure 4.9). The GNRs produced should have a high length/width ratio, R (Equation (4.3.3)). From length and width we define area, A (Equation (4.3.4)), as the full space around the ribbon including any gaps (Figure 4.9), where ceil is a function that rounds up a real number to the next higher integer. We define occupancy, O, as the completeness of the ribbon formation: a criterion of successful assembly, measured as proportion of area covered (Equation (4.3.5)). We also use a general cost function, C, to track the overall quality of GNRs produced by each simulation (Equation (4.3.6)).

The functions describing the ribbons are:

$$R = \frac{L}{W} \qquad \text{(Length/width ratio)} \qquad (4.3.3)$$

$$A = \text{ceil}(L \times (W - 0.5)) \qquad \text{(Area)} \qquad (4.3.4)$$

$$O = \frac{N}{A} \qquad \text{(Occupancy)} \qquad (4.3.5)$$

When $L > 3, W > 4$, the cost function is:

$$C = O\frac{L}{W} = OR \qquad \text{(Cost function)} \qquad (4.3.6)$$

else: $C = 0$.

As the simulation is stochastic, several simulations need to be run at each set of experimental conditions in order to understand the underlying behaviour of the system at those conditions. For this reason, at least 20 simulations were run at each sampling point. For the simulations detailed in this thesis, temperature was set at 400K [172]. Between 400 and 2000 initiator molecules were used in each simulation. The volume of the reaction chamber was set such that the initial concentration of initiators was 1000l^{-1}.

The model is designed as a generalized representation of molecular interactions, and is extensible to probe other systems: any starting mixture of molecules can be chosen. For example, defective initiators can be included in the initial stages of synthesis. Additionally, the interaction of the molecules and the surface could be explicitly included.

4.4 Results

In the simulation, the activation energies were investigated while other parameters, such as temperature and concentration, were held constant. Activation

Observable	Min	Max
Length	3	1599
Width	4	2399
Ratio	0.00125	399.75
Occupancy	0.15	1.00

Table 4.1: Minimum and maximum values of observables over wide sampling.

energies were set in a physical range between 10^{-2}eV–10^{1}eV: a range in which the energies are physically plausible [173–176]. As a coarse-grained abstraction of the laboratory, experimental features essential for controlling coupling reactions (such as catalysis) were ignored for the efficient exploration of parameter space (see §4.3). Ignored parameters effectively are incorporated into the activation energy needed for each reaction to occur [176, 177]. We simulated coupling of initiator molecules to completely form GNRs, when varying the activation energy required for each reaction, ending up with an interesting phase space (Figure 4.10). The maximum and minimum values of observables in the system (*i.e.* Length, Width, Occupancy of ribbons) are shown in Table 4.1, and the distribution of the observables over the sampled parameter space are shown in Figure 4.11. Ea_{AA} and Ea_{BC} refers to energies of activation in eV for A–A and B–C bond formation, from the interaction in Figure 4.3. Phase diagrams show the length (Figure 4.11(a)), width (Figure 4.11(b)), and completion (Figure 4.11(d)) of GNRs formed. GNR length and width is measured according to the directions indicated in Figure 4.9: the completion of the ribbon is defined as the number of benzene rings occupied within the box traced out by length and width (Equation (4.3.5)). A phase diagram of the combined fitness score, calculated according to Equation (4.3.6), is also provided (Figure 4.11(d)). Simulations were conducted 20 times for sample points over four orders of magnitude for Ea_{AA} and Ea_{BC}. The most complete GNR formation with desirable proportions was noted in the bordered region (upper right of phase diagrams).

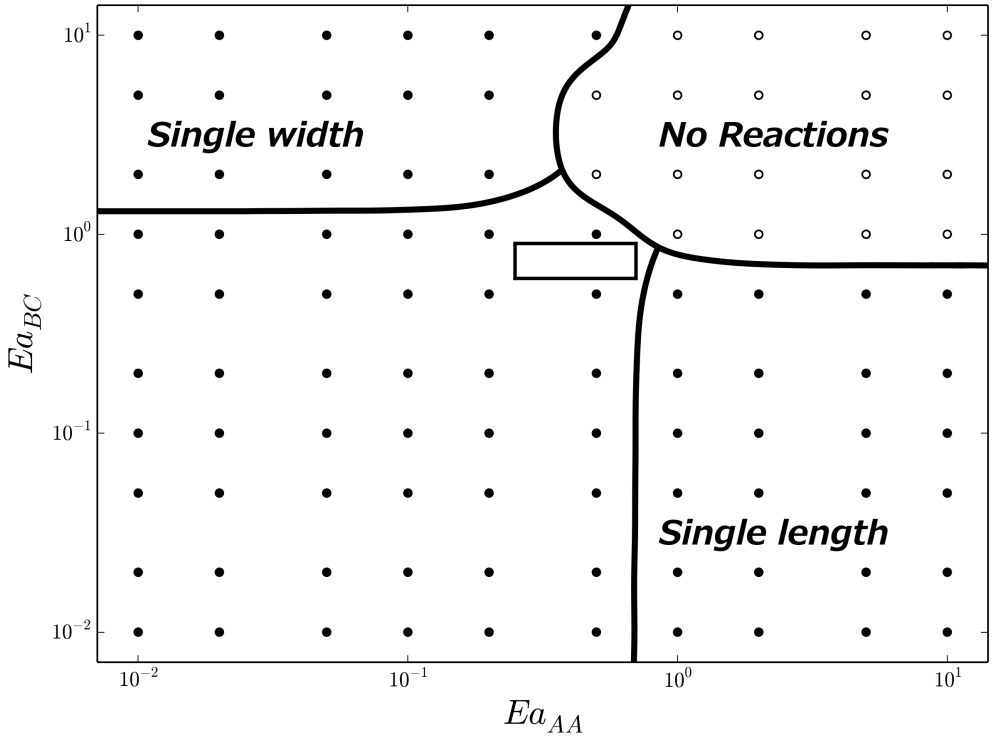
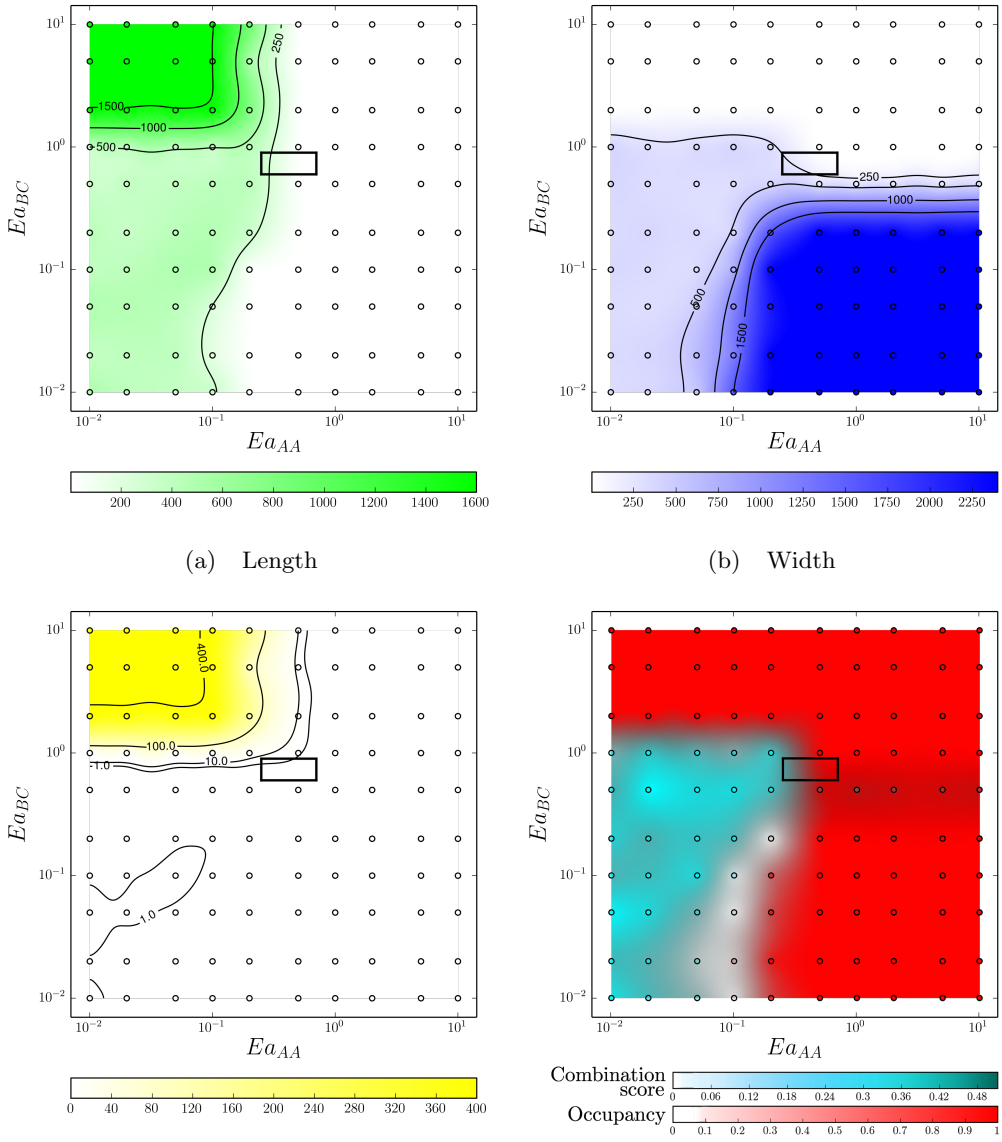


Figure 4.10: Phase diagram for tetrabenzanthracene and benzene self-assembly. Temperature: 400 K. Ea_{AA} and Ea_{BC} refers to energies of activation in eV for A–A and B–C bond formation, from the interaction in Figure 4.3.

If the activation energies for A–A and B–C coupling reactions (Ea_{AA} and Ea_{BC} respectively) are both high, such as 1.0eV, then no reactions occur within the characteristic timescale (120s). Without bond making, no GNR forms, and the initiators are left unreacted. Conversely, if both the activation energies are very low, then any possible reaction will occur very quickly, exhausting the molecules, and resulting in uncontrolled growth. At these energies, a single disordered GNR is produced for each simulation, as the molecules bind at many points to the growing ribbon. The number of holes within the ribbon increases at the end of the simulation, leading to reduced occupancy score: see Figure 4.11(d).

If Ea_{AA} is much larger than Ea_{BC}, then the B–C reaction will be preferred,



(a) Length

(b) Width

(c) Ratio (Equation (4.3.3))

(d) Occupancy (Equation (4.3.5)) & cost function (Equation (4.3.6))

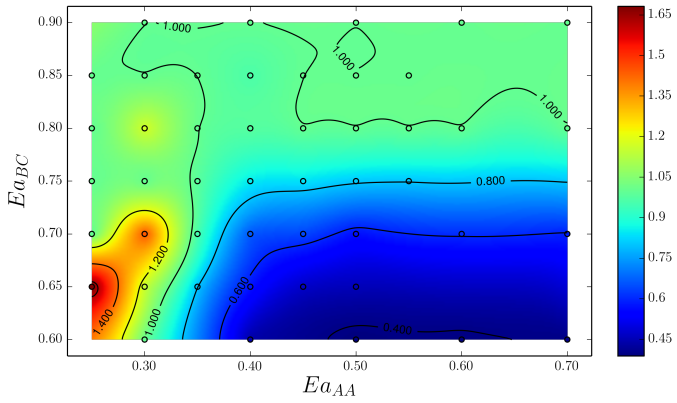
Figure 4.11: Observables for tetrabenzanthracene and benzene self-assembly. Temperature: 400 K. Ea_{AA} and Ea_{BC} refers to energies of activation in eV for A–A and B–C bond formation, from the interaction in Figure 4.3. Simulations were conducted 20 times for sample points over four orders of magnitude for Ea_{AA} and Ea_{BC}. The length (a), width (b), and their ratio (c) of the resulting ribbons are shown in individual phase diagrams, as also is the occupancy and cost function (d). Long GNRs with small widths and high occupancy were noted in the region outlined.

and GNRs will be wider than longer, as evidenced in Figure 4.11(b). As expected, low Ea_{AA} and high Ea_{BC} results in GNRs with armchair edges. In order to get "wide ribbons" and avoid single-width GNRs, the activation energies for Ea_{AA} and Ea_{BC} should be set around 0.3eV–0.6eV and 0.3eV–0.9eV respectively (box on Figure 4.11). This subset of parameter space is investigated with finely spaced sampling points (Figure 4.12).

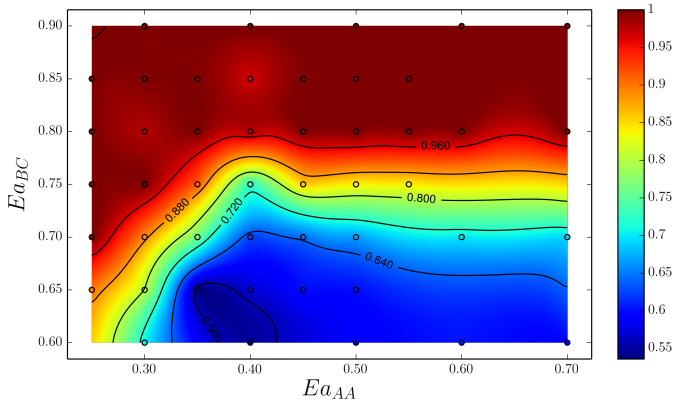
Figure 4.12(a) shows the length/width ratio of GNR products: not all of the sampling point activation energies are sufficiently biasing the A–A direction of growth. However GNRs from this domain do maintain few defects (Figure 4.12(b)), and also few copy numbers of large GNRs (Figure 4.12(c)). Particularly desirable GNR length/width ratios and completion are seen when Ea_{AA} is 0.25eV–0.30eV and Ea_{BC} is 0.60eV–0.70eV. The convergence of these simulations after 40 sampling runs is good: the standard error on the mean for each of these measurements is shown in Figure 4.13. As previously discussed, at low values for both activation energy parameters, there is reduced occupancy score, as the GNR bonds are produced more quickly and thus there are fewer constraints on unfavourable bonds forming. This is evidenced by a higher standard error for this regime (Figure 4.13(b)). At these values, the activation energy for two or more bonds forming will be multiples of that for diffusion (0.2eV–0.4eV [97]): rates of coupling will thus be several orders of magnitude less than rates of diffusion (Equation (2.1.1)), confirming that we are still in a coupling-limited domain.

An even finer search could be undertaken to reveal precise phase separations within the activation energy phase space.

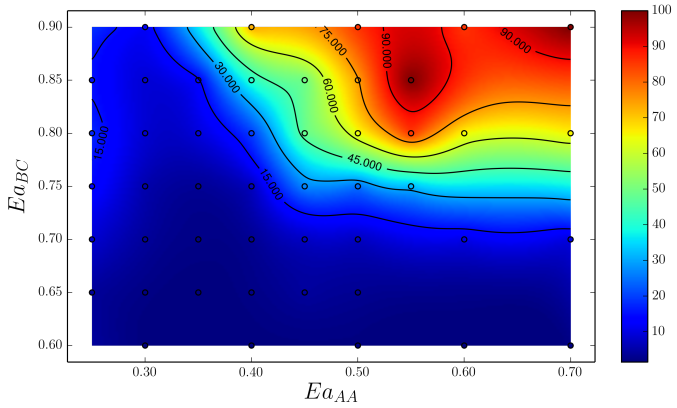
To engineer the formation of GNRs predicted by these specific kinetics in the lab, the energetics simulated here could be emulated by alteration of catalyst concentration or functional groups for the coupling reactions. Alternate coupling



(a) Mean Ratio: Length / Width

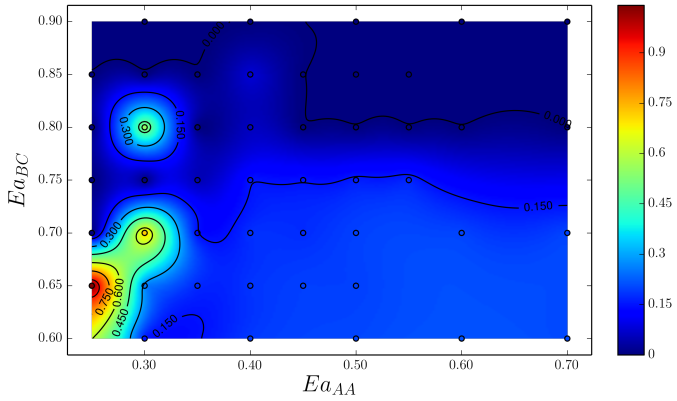


(b) Mean Occupancy

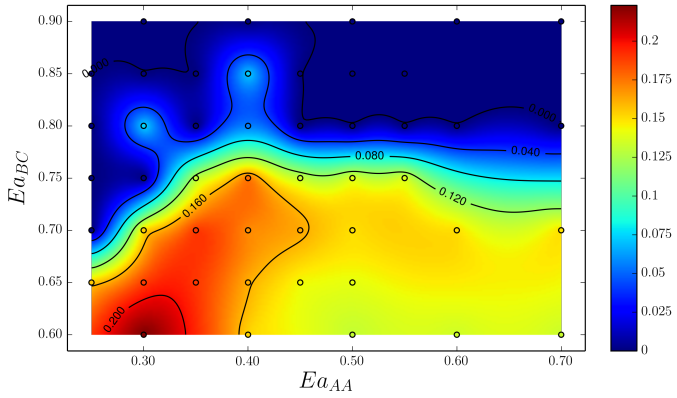


(c) Mean Number of products

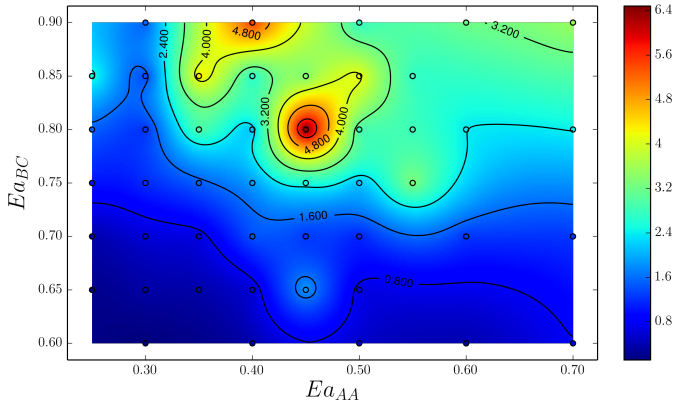
Figure 4.12: Detailed phase diagrams for tetrabenzanthracene and benzene self-assembly. Temperature: 400 K. Ea_{AA} and Ea_{BC} refers to energies of activation in eV for A–A and B–C bond formation. Simulations were conducted 40 times for sample points over a fine sampling of Ea_{AA} and Ea_{BC}. The ratio of length to width (a) and occupancy (b) of the resulting ribbons are shown. The number of product ribbons (c) are also given.



(a) Ratio standard error



(b) Occupancy standard error



(c) Number standard error

Figure 4.13: Phase diagrams for tetrabenzanthracene and benzene self-assembly showing the standard error on the mean for the observables in Figure 4.12.

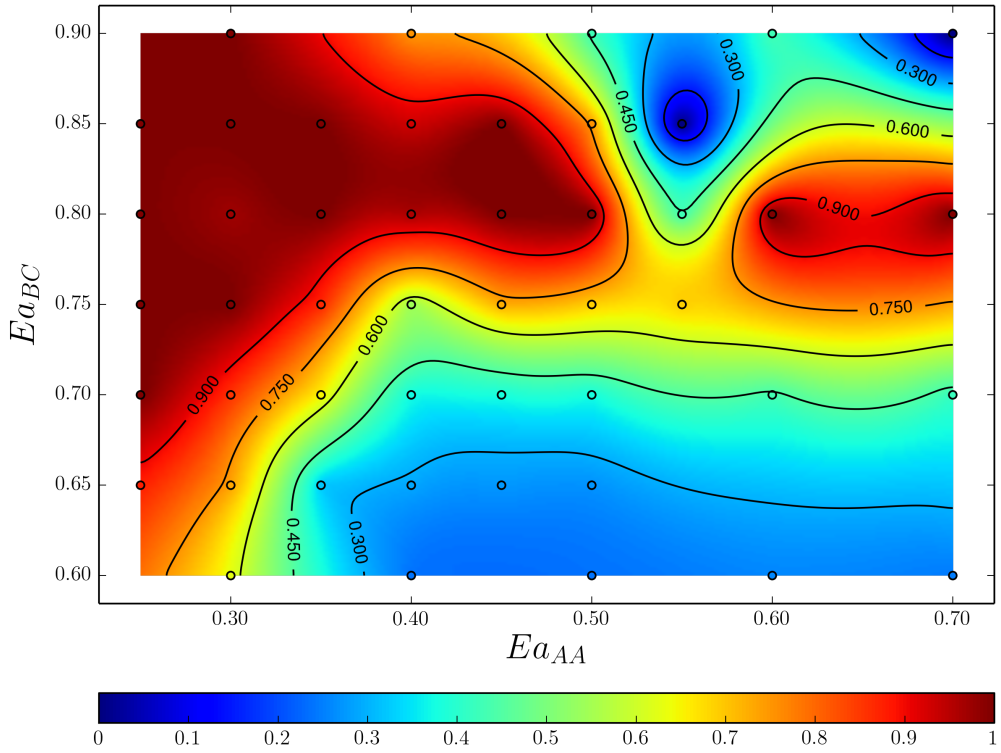


Figure 4.14: Detailed phase diagrams for the cost function describing tetrabenzanthracene and benzene self-assembly. Temperature: 400 K. Ea_{AA} and Ea_{BC} refers to energies of activation in eV for A–A and B–C bond formation. Simulations were conducted 40 times for sample points over a fine sampling of Ea_{AA} and Ea_{BC}.

reactions with different activation energies could also be introduced. A system of physical reactions could thus be designed in the lab to exploit the favourable outcomes predicted by the simulation. However, these length/width ratios of the GNRs are still lower than would be ideal for use in nanoscale devices (*cf.* Figure 4.12(a)). Adjusting the activation energies will affect the completion as well as length/width ratio, but there other perturbations to the system can be used to bias the directional growth. Here we modify the simulation to address one of these perturbations.

GNR formation in experiment can also depend on how the reaction is started. For example, the initial mixing of the initiators is a possible technical limita-

tion. The simulation can be modified to introduce different reactants at different times, to test the response of the system to these perturbations. As an exemplar, we describe a gradual mixing at the start of the reaction, introducing functionalized benzene at different rates into the simulation until it is equal to the starting amount of tetrabenzanthracene (Figure 4.15). The width of the produced GNRs vary in each simulated benzene addition rate (Figure 4.16). We use parameters for the activation energies near the region providing good GNR outputs: namely $Ea_{AA} = 0.3\text{eV}$ and $Ea_{BC} = 0.75\text{eV}$. It is seen that if the benzene is all added at the start of the reaction, only single width GNRs are formed: polymerization in the length direction is preferred. This is also true for very quick rates of benzene addition (Figure 4.16, red). When the addition rate approaches 10^4s^{-1}, wider ribbons are seen (Figure 4.16, blue). At low addition rates around 10^{-1}s^{-1}, there is not enough benzene in the system to start a reaction before the simulation terminates—as could the experiment be ended early—producing single width GNRs (Figure 4.16, green).

At the higher rates, benzene binding to tetrabenzanthracene is not unfavourable, though single-width polymerization is still preferred. However, after a few benzenes are bound to tetrabenzanthracene scaffolds, any tetrabenzanthracene polymers are likely to not bind as benzenes will clash. This is why, at lower benzene addition rates, the lower effective concentration of benzene enables the wider GNRs to form. Notably, these widened GNRs are fairly complete: without too many gaps, occupancy is maintained above 90% (Figure 4.16, red). Closing every gaps in all GNRs produced would require a further step after the model system has equilibrated: specially designed “completer” molecules could be introduced to the system. For the model system, edges of the ribbon could be terminated by a wash in excess tetrachlorobenzene (Figure 3.8) to compete all B–C bonds, followed by the introduction of specially designed tetrabenzan-

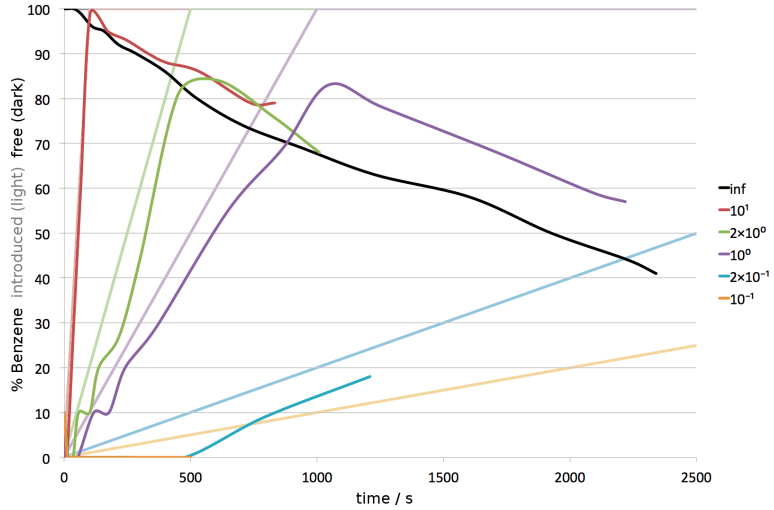


Figure 4.15: Precursor ramp: free benzene. The functionalized benzene is introduced into the reaction at different rates, or all at once initially (∞), here plotted as different colours, at a single sampled point (Ea_{AA} is 0.3eV, Ea_{BC} is 0.75eV). The amount of free benzene (dark lines) is strictly equal or less than the amount of benzene added to the simulation (light lines). Reaction termination time (indicated by the end of the dark lines) is dependent on the number of free reactions, not just on free benzene. Slower rates of addition do not necessarily indicate slower completion of reaction, but rather would indicate less benzene being included in final product.

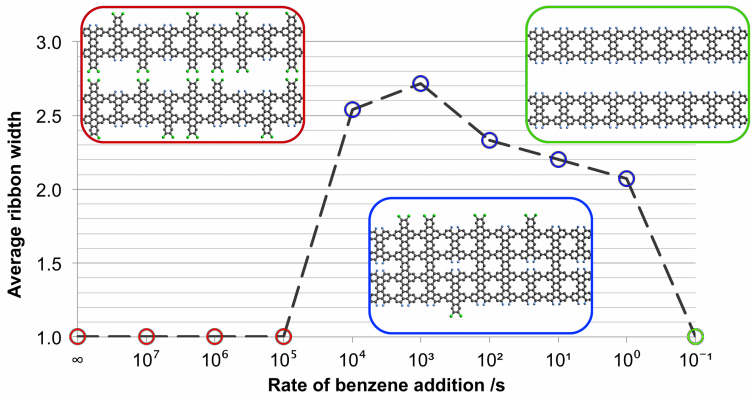


Figure 4.16: Precursor ramp: ribbon width. Ribbon width is dependent on the rate of the addition of functionalized benzene into the reaction. The quicker rates (red) yield single-width ribbons, as the ribbons are unlikely to bind together, due to the sporadic attachments of benzene to the ribbon. For slow rates (green), the reaction is essentially terminated before any benzenes are bound to the ribbon. At intermediate rates (blue), wider GNRs are able to be formed. Data from a single sampled point (Ea_{AA} is 0.3eV, Ea_{BC} is 0.75eV).

thracene molecules with only one side of B functionalization. Further simulation will be required to determine this procedure.

4.5 Discussion

We have developed a new course-grained algorithm for simulating GNR production by self-assembly. The kinetics-based model of the synthesis process allows specification and refinement of experimental protocols for preferentially producing GNRs with desirable characteristics. In this study, we demonstrate the simulation using two assembly initiators that form a patterned GNR with a useful electronic property: a band gap of 1.6eV that is stable across many widths of the ribbon. Experimental variables in the synthesis process correspond to parameters in the simulation, and we interrogate combinations of these parameters to determine how each effects the synthesized products. A phase diagram of these parameters is produced to guide experimental setup, such that GNRs with specific desirable properties can be designed. However, further investigation of this parameter space is needed to increase the consistency of producing the desired GNRs in experiments. Importantly, experimental questions such as initiator mixing, and temperature changes can be explored by the simulation. The model can be adapted to replicate and gain insight into the synthesis process, as required.

For the presented model system, we have predicted that if no other alteration of the system occurs, GNRs with desirable characteristics could be best produced with activation energies in the range $Ea_{AA} = 0.25\text{eV}–0.30\text{eV}$ and $Ea_{BC} = 0.60\text{eV}–0.70\text{eV}$ (Figures 4.12 and 4.14). The convergence of the simulations within this range of activation energies also hints towards the possibility of robustly and efficiently producing GNRs experimentally (Figure 4.13).

We predict that the nucleation of self-assembly in this and similar systems

could be best performed with a gradual introduction of species into the reaction (Figure 4.16). Further simulation is required to determine whether this is essential for achieving optimal self-assembly for different assembly parameters. Other self-assembling systems are known to exploit the timing of introducing initiators into the system to control outcome. An example from nature is that of viruses: it has been shown that the main structural viral coat protein fails to fully assemble into complete viruses unless it is is gradually introduced into the assembly environment [11]. Interestingly, it is highly evolutionarily incumbent on viruses to completely form, else they have a greater chance of being detected and eradicated by host immune response. This suggests that viruses have evolved to exploit the same outcomes as the physical behaviour experienced by this model simulation. The role of the mixing rate in controlling product fidelity also promises to be of value for future work on the synthesis of GNRs.

Strikingly, our model predicts that synthesizing correctly-assembled complete products may require a time-dependent experimental protocol. The required time to halt assembly (included in the model as the characteristic timescale of 120s) is proportional to the activation energies (data not shown), and could be altered to compensate for unadjustable activation energies, in order to realize a predicted model scenario experimentally.

Further extensions of the model could include the inert surface explicitly: perturbations to the surface could be incorporated, and also the diffusion by molecules across the surface in different directions. The diffusion could be included by modelling the orientation of the molecules persistently throughout the simulation, which could be used to mediate the rate of interaction and thus the sampling of the reaction network (*n.b.* in contrast to a MC simulation [102–104] the physical location of the molecules would still not have to be considered). Catalytic effects could be explicitly modelled by including catalyst molecules

within the reaction network: interactions between these catalysts and the initiator molecules would be best modelled as reversible reactions. Both of these changes to the modelling would yield additional insight, but would require substantial additional steps of computation.

Chapter 5

Self-assembly in virology

Biological systems have evolved sophisticated self-assembly processes to facilitate the accurate and efficient synthesis of structurally diverse assemblies from relatively simple subunits. Among the many examples of self-assembly in biology, the assembly of virus capsids is perhaps the most studied system due to its high symmetry, which makes it more tractable for theoretical analysis than a heterogeneous system [178]. The study of viral assembly is often a means for understanding complex self-assembly systems in general, in seeking to understand the highly economical assembly processes arising from a minimal number of different subunits. Our understanding of complex self-assembling systems, including capsid assembly, is limited by the difficulty of probing the reaction dynamics at nanometer scales, particularly *in vivo*, so theory and computation have been indispensable for gaining new insights into self-assembly [11,178,179].

In contrast to our work detailed previously on synthesis in organic chemistry (Chapter 4), capsid self-assembly is driven by molecular recognition between proteins by means of weak non-covalent interactions, which facilitates self-correcting dynamic assembly pathways towards a thermodynamic minimum: *i.e.* a fully-formed capsid (Figure 5.1) [73, 180]. Many viruses have evolved robust self-

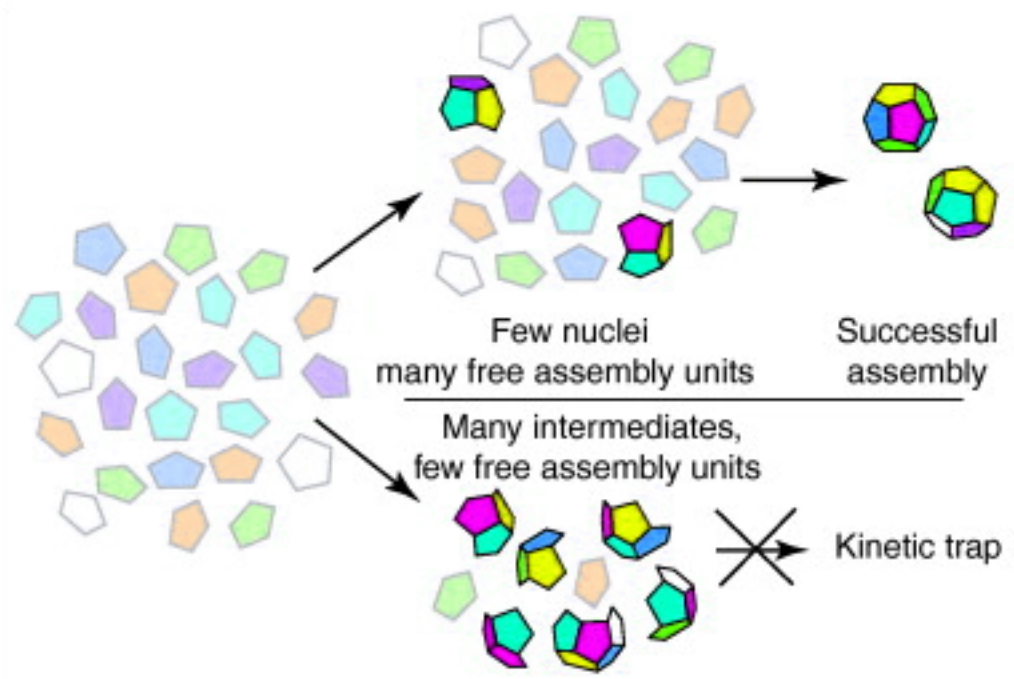


Figure 5.1: How do viruses assemble? Figure adapted from [180].

assembly pathways that are tuned for efficiency and fidelity of assembly: in many cases the viral capsid can even assemble *in vitro* outside of the cellular environment [181]. Of particular interest are the co-assembling processes seen primarily in ssRNA[1] viruses, where capsid assembly is mediated by interactions with the concurrently-packaged viral genome, increasing the efficiency and fidelity of formation [8, 9]. In the co-assembled viruses, the formation of an infectious virion requires the viral RNA to be selectively packaged against a background of cellular RNAs, due to the evolutionary pressure against misencapsidation [11, 182].

[1]For all abbreviations see the glossary on page 222.

5.1 Virus structure

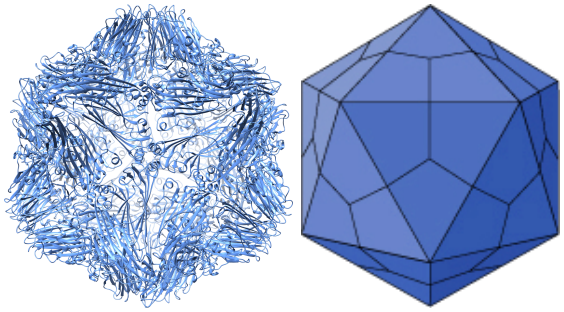
Viruses are remarkable examples of symmetry and self-assembly at the nanoscale. In order to minimize the size of the genome fragment needed to code for the viral capsid while maximizing its volume/surface area ratio (known as the principle of genetic economy [183]), several copies of the same or similar proteins assemble into a symmetric structure, in which all subunits occupy the same environment (*i.e.* icosahedral or helical capsids, seen in the vast majority of viruses). First proposed in 1956 by Watson and Crick [183], this idea was later further developed by Caspar and Klug for icosahedrally-symmetric viral capsids with more than 60 subunits, where they can be described in terms of icosahedral surface lattices [184]. Their 'quasi-equivalence theory' stated that for capsids with more than 60 subunits, each subunit occupies a similar but not identical environment in the capsid [184]. Many different icosahedral capsid structures can be formed using a quasi-equivalent approach: some example capsids formed by quasi-equivalent proteins are shown in Figure 5.2.

Under quasi-equivalence, the T-number describes the number of structural subunits per asymmetric unit that forms the structure. Furthermore, Caspar and Klug deduced that the icosahedral viral capsid must consist of $60T$ protein subunits, with T being restricted to:

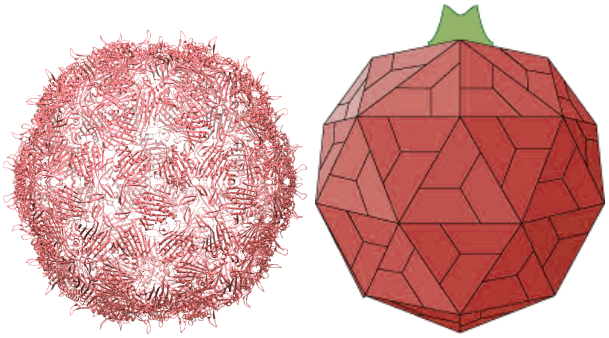
$$T = h^2 + hk + k^2, \qquad h \in \mathbb{Z}^+, \qquad k \in \mathbb{Z}^*, \tag{5.1.1}$$

where $\mathbb{Z}^+$ is the set of positive integers and $\mathbb{Z}^*$ is the set of nonnegative integers.

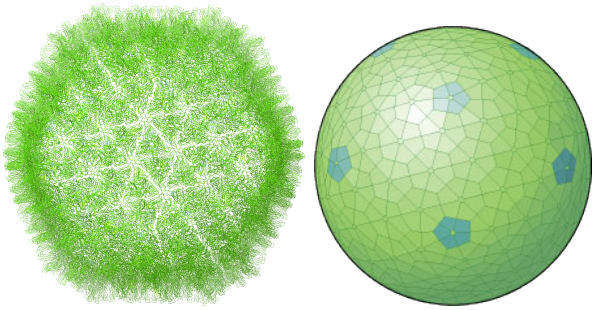
Interestingly, the T-number cannot always uniquely characterize a capsid, since for some $T \geq 49$ more than one pair of (h, k) can share the same T. For instance, $(7, 0)$ and $(5, 3)$ both give $T = 49$. Certain T-numbers can be skew on the lattice (where (h, k) do not have a common divisor, and $T \geq 7$), see for



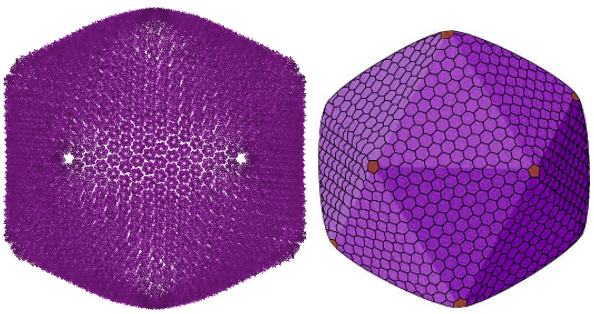
(a) STNV ($T = 1$)



(b) MS2 ($T = 3$)



(c) IBDV ($T = 13l$)



(d) PBCV-1 ($T = 169d$)

Figure 5.2: Capsids formed of quasi-equivalent proteins. (a) STNV, a 194.0Å diameter $T = 1$ capsid, from PDB:4V4M [49]. (b) MS2, a 288.0Å diameter $T = 3$ capsid, from PDB:2MS2 [185]. (c) Infectious bursal disease virus (IBDV), a 748.0Å diameter $T = 13l$ capsid (laevo form), from PDB:1WCE [186]. (d) Paramecium bursaria chlorella virus 1 (PBCV-1), a 1858.0Å diameter $T = 169d$ capsid (dextro form), from PDB:1M4X [187].

example Figures 5.2(c) and 5.2(d). These viruses are described as possessing either right (*dextro*) or left (*laevo*) handedness. The Caspar-Klug theory does not permit formations with certain T-numbers, for example $T = 2$ and $T = 6$ [184]. Most icosahedral viruses follow the Caspar-Klug rules, but there are exceptions and complications. Indeed, viruses which do not fit the Caspar-Klug model in Equation (5.1.1) are known to exist; these can be modelled with *viral tiling theory* [188, 189].

The fundamental subunit in the Caspar-Klug construction is not necessarily a single protein. For example, the icosahedral capsid of the *Picornaviridae* have $T = 1$ with pseudo $T = 3$ symmetry, meaning that each asymmetric capsid unit is composed of three similar non-identical proteins [190] (see Figure 5.3). Another example is the bluetongue virus (BTV) core, an icosahedral shell composed of 120 proteins, corresponding to a forbidden triangulation number ($T = 2$), but if protein dimers are considered the fundamental building blocks of the capsid, then the 60 asymmetric dimers are effectively organized as a $T = 1$ capsid [191] (Figures 5.4(a) and 5.4(b)). *Saccharomyces cerevisiae* virus L-A (L-A) also has a similar $T = 2$ structure [192, 193].

Other viruses, including *Papillomaviridae* such as human papillomavirus (HPV), and *Polyomaviridae* such as murine polyomavirus (MPyV) and simian virus 40 (SV40), have a capsid structure similar to $T = 7$ viruses, but using a different number and organization of subunits than predicted by Caspar-Klug (Figures 5.4(c) and 5.4(d)) [188, 198]. In a beautiful paper, Twarock unveiled non-triangular tilings that allowed these forbidden geometries using a generalized principle of quasi-equivalence [188]. Twarock found that, similar to a Penrose tiling of the plane in which aperiodic tiling with darts and kites occurs, the capsid can be viewed as a spherical section of an aperiodic tiling, allowing five-fold symmetry axes to form (see Figure 5.4(e)).

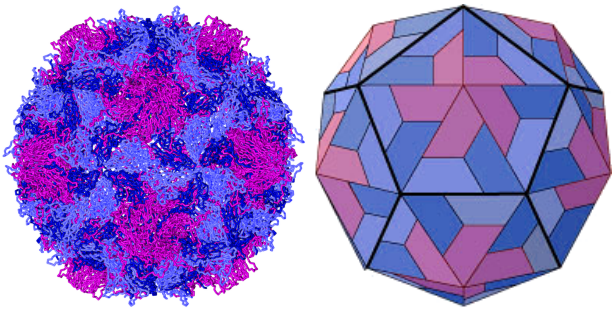


Figure 5.3: Poliovirus: a member of the *Picornaviridae*, which have pseudo $T = 3$ capsids [190]. Diameter of 324.0Å, from PDB:1HXS [194]. The capsid is composed of 12 pentamers and 20 hexamers for a total of 180 proteins, but are not $T = 3$ symmetry as described by Caspar-Klug theory [184] because the basic unit of the exterior is composed of three different proteins (coloured differently here). The three proteins, VP1, VP2, and VP3, share no sequence homology, but the topology of all three is the same, in the form of an eight-stranded antiparallel β-barrel jelly roll [195]. As the three subunits are morphologically very similar, the structure is denoted as pseudo $T = 3$. Note there is another structural protein, VP4, that has an extended conformation and lies on the inner surface: its presence is essential for the stability of the virion [195].

The structures of the CPs and the interactions between them are responsible for the overall capsid structure. They are highly engineered to bind to each other with high specificity, and we say that there are local rules that govern their interactions, due to symmetry, binding energy, *etc.*, that result in self-assembly of the capsid. The principle of quasi-equivalence means that identical protein types can play different roles at different points in the capsid. For each virus, CPs have evolved to fulfil quasi-equivalent roles in their capsid, while meeting the inter-protein interactions and forming the shell with the evolved T-number. Often, the CPs need to undergo changes in conformation in order to meet the roles dictated by quasi-equivalence (*e.g.* in the phage MS2 [199,200]). BTV, with its forbidden $T = 2$ structure, is interesting in terms of quasi-equivalence in that it has evolved to have a monomer that can adopt two distinct conformations (Figure 5.4(b)), allowing circumvention of the normal rules of Caspar-Klug theory (Figure 5.5).

We have seen that although the individual protein monomers are asymmetrical, through their organization into capsomeric subunits and finally capsids, the

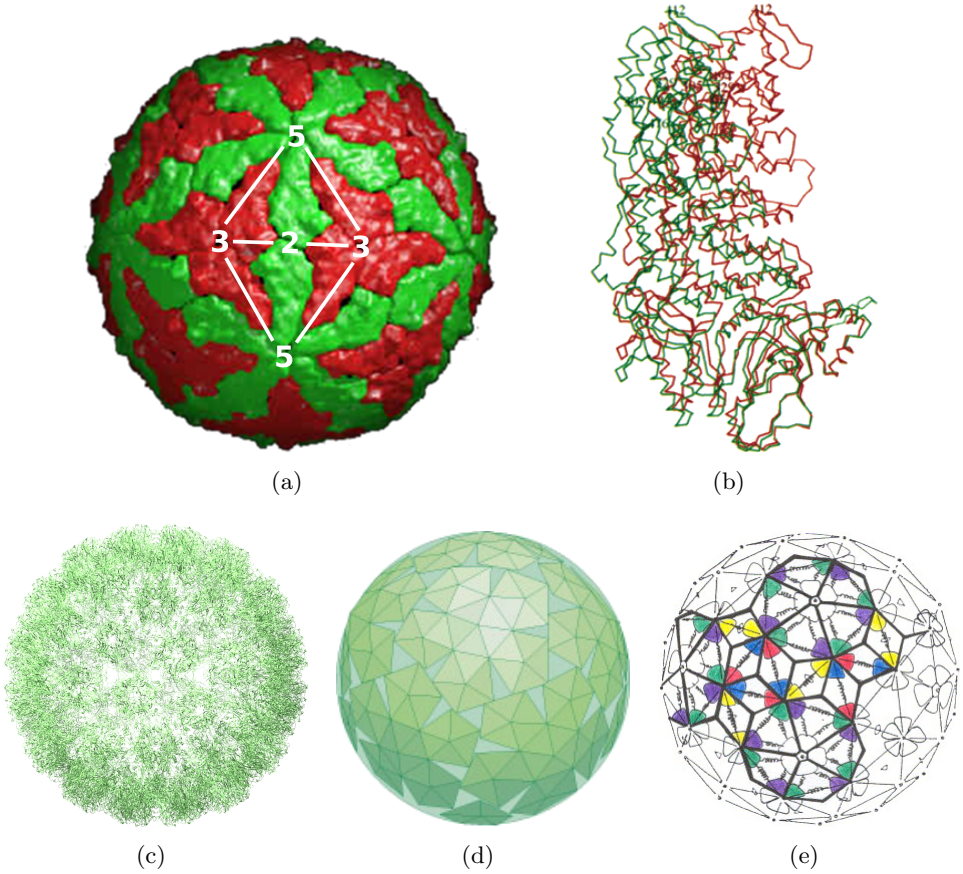
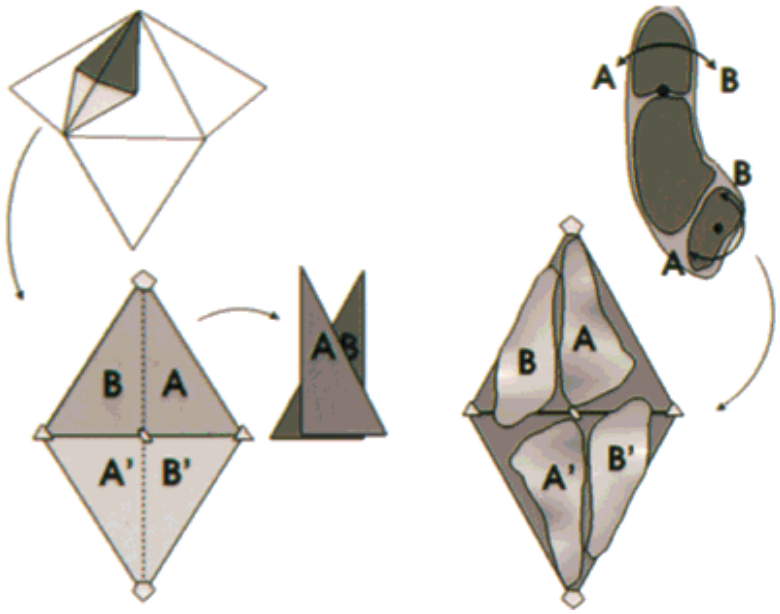


Figure 5.4: Exceptions to Caspar-Klug theory. (a) BTV core VP3 adopts an icosahedral capsid made of 120 CP, with pseudo $T = 2$; if a dimer is considered as the fundamental CP subunit, the structure is effectively $T = 1$ [191]. (b) BTV monomer distortion: the C_α chain of BTV VP3 protein, showing the difference in conformation between molecules A (green) and B (red), with the structural elements that form the quasi-two-fold dimer interface superimposed [191]. (c) Human papillomavirus (HPV), a $T = 7d$ capsid with a diameter of 602.0Å (PDB:3J6R [196]). (d) Diagrammatic representation of HPV capsid, similar to *Polyomaviridae* such as MPyV and SV40 [188,197]: 72 pentamers comprising 360 capsid proteins, arranged as $T = 7d$, leaving gaps. (e) Tessellation with darts and kites for *Papillomaviridae* and *Polyomaviridae* superimposed on the $T = 7d$ lattice [188].



(a) (b)

Figure 5.5: The problem with $T = 2$ structures under quasi-equivalence. (a) The two $T = 2$ dimers cannot be translationally overlaid: they have mirror symmetry. Such geometries cannot be represented *via* triangulation as the local environments are different, and moreover mirroring (reflections through the surface) is not possible due to the curvature of the capsid. (b) These restrictions may be overcome by an internal conformational shift of the monomers, which allows them to interlock and thus close up the capsid, as shown for BTV [191].

virus develops a high degree of symmetry. However, here we should make a distinction between the symmetry and the shape of a capsid: icosahedral symmetry does not necessarily imply a virus being shaped as an icosahedron. This can be observed in Figure 5.2, where we can see that the viruses with lowest T-numbers have icosahedral symmetry, but are not icosahedral in shape, but rather exhibit a spherical morphology. With increasing T-number, the capsids become more icosahedral, as can be observed in Figure 5.2(d). Such buckling transitions are expected due to the elastic properties of thin protein shells [201]. Higher T-number viruses are normally of a larger size, because they require more proteins than the lower T-numbers, with the assumption that the molecular mass of a

protein subunit remains approximately the same for all viruses [202]. Thus the link between sphericity and radius can be explained, with reference to continuum elasticity theory, by the Föppl-von Kármán number, γ:

$$\gamma = \langle R \rangle^2 \frac{Y}{\kappa} \tag{5.1.2}$$

with $\langle R \rangle$ as the average capsid radius, Y the 2-D Young modulus, and κ the bending rigidity. Note that the thickness of the capsid is particularly important for the elastic properties, *i.e.* the Young modulus and bending rigidity. Most viruses either have $\gamma \leq 150$ (implying a form approaching a sphere) or $200 \leq \gamma \leq 1500$ (noticeably buckled) [201]. This analysis implies that the larger viruses (with greater icosahedron shape tendencies) are more likely to possess structural deviations such as cones and ridges. Conversely, the smaller viruses are well approximated as perfect spheres, even though the capsids are technically icosahedral in symmetry [201].

Symmetry plays a pivotal role in understanding virus structure. Symmetry averaging techniques are used to refine viral structures obtained *via* X-ray diffraction, and they have enabled the reconstruction of capsid structures from cryo transmission electron microscopy (cryo-EM) data [203]. In the best cases, the resolution of cryo-EM structures rivals that of X-ray diffraction studies, yielding detailed insights into the form and function of symmetrical viral components.

By comparison, asymmetric components normally contribute too weakly to the images obtained by cryo-EM to allow the refinement of an asymmetric model [204]. Note that in a crystal packing of viral particles, the asymmetric features of individual viruses usually do not dictate crystal packing contacts, and are therefore averaged out by the lattice. The important functional roles of such viral components in the viral life cycle are therefore difficult to characterize.

An example is the single-copy maturation protein (MP) in bacteriophage MS2 that is hypothesized to replace a protein dimer in the capsid [205], and attaches to the bacterial receptor during the infection to facilitate genome extraction.

Moreover, the asymmetric organization of the viral genome inside a capsid is at present difficult to reconstruct. Recent work based on bioinformatics [206] and kinetic modelling [10] demonstrates that some viral genomes, notably those of ssRNA viruses, have conserved asymmetric structures in proximity to their capsids, consistent with the important roles of such genomes in virus assembly [8]. Imaging techniques based on symmetry averaging reveal density consistent with ordered genome segments in many viruses, including the *Leviviridae* [35] (Figure 5.12). This suggests that there are conserved structural elements in the organization of the genomes of these and other viruses, but icosahedral averaging washes out details. Cryo electron tomography (Cryo-ET) provides structural information on these asymmetric features, albeit at much lower resolution (35-50Å) [207]. Sub-tomographic (asymmetric, non-icosahedral) averaging can be used to increase resolution, especially if prominent features such as phage tails are present [208, 209], but generally it is not sufficient to determine the asymmetric structure of a virus at atomic resolution [12, 205].

It is therefore important to develop new analysis techniques that are able to reveal genome organization based on the low resolution information contained in 3-D structural data. In Chapter 6 an analysis of the asymmetric X-ray crystallography structure of STNV is discussed, where interestingly the genome electrostatics within the capsid affect the overall particle morphology, and influence the orientation of virions within the crystalline virus stacking. This results in a non-random orientation to be preferred; we see a dual unit cell, presumably because a pairing of two virions in a specific orientation can best neutralize the overall charge.

In Chapter 7 we introduce a new method that uses RNA location information from icosahedrally-averaged maps, as well as knowledge of the contact sites between genomic RNA and CP to analyse the low resolution, tomographic density maps *via* a constraint optimization technique to reveal the putative asymmetric genome organization in proximity to the capsid. We use MS2 as a model system to demonstrate the new technique since an asymmetric tomogram (that uses the bacterial receptor for alignment) as well an icosahedrally-averaged map are available. Although the biochemical properties and structural characteristics of MS2 (described in detail in Chapter 7) are used to constrain and define the approach, the method can be applied to any virus for which similar information is available [12].

5.2 Genome packaging

Formation of infectious ssRNA virus particles requires that the genome is selectively encapsidated out of a milieu of competing non-cognate cellular RNAs, which most viruses manage with remarkably high specificity (*e.g.*, 99% [11,182]). It is all the more remarkable that the genomes of many RNA viruses are composed of several distinct strands, rather than as a single, long genome, all of which are required for continued infection. This gives rise to two strategies for the virus: these distinct strands must either (i) each be packaged into separate virus particles (with the consequence of a higher multiplicity of infection being required at all stages of the life cycle) as in the *Bromoviridae* [41], or (ii) all be packaged into every virion (requiring a gateway selection mechanism to delay maturation of the particle until all strands have been packaged), as for example in the *Nodaviridae* [210].

There are several explanations for how ssRNA viruses may selectively package their cognate genomic RNAs in the presence of competitor cellular RNAs [11,

182]. One possible method is that the viral components are separated from other cellular material, *e.g.* in viral factories. However, there is evidence of misencapsidation occurring frequently in viruses that assemble in this manner [211, 212], suggesting that there are other more successful means of ensuring selective packaging. An alternative explanation is that the selective packaging of RNAs occurs as the result of sequence-specific interactions between assembling CPs and their cognate viral RNAs, arising as a result of evolutionarily optimized recognition strategies [8, 9, 11]: this is the scenario that will be discussed here.

Indeed, there is mounting evidence that the genome plays a critical role in the promotion and direction of viral assembly for many ssRNA viruses, with more specificity than a purely electrostatic model would suggest [8–10, 34]. During assembly, the RNA interacts with the CPs and a precise fold is selected from the ensemble of possible ones; the central region of the RNA is liquid crystalline, but in proximity to the capsid the RNA plays a role in both the nucleation of CP assembly and determining the subsequent sequential order of CP assembly (see Figure 5.6). From a genome packaging perspective, the proteins effectively work as chaperones, facilitating the folding of the RNA molecule into a structure consistent with the capsid geometry, and collapsing the RNA into a tighter conformation (Figure 5.7) [213]. The possible geometries of the RNA are constrained by the existence of interactions between specific secondary structures in the genome and protein capsid—*packaging signals* (PSs)—that can mediate protein conformation between quasi-equivalent conformers, thus minimizing the energy required to form the icosahedral capsid [199]. The binding affinities and positions of these PSs determine the assembly pathways [10]. Sequence specific interactions are expected for most PSs, but in the example of pariacoto virus (PaV) (Figure 5.8) sequence specificity is not required [214]; but cognate genomes would presumably facilitate utilization of more efficient assembly path-

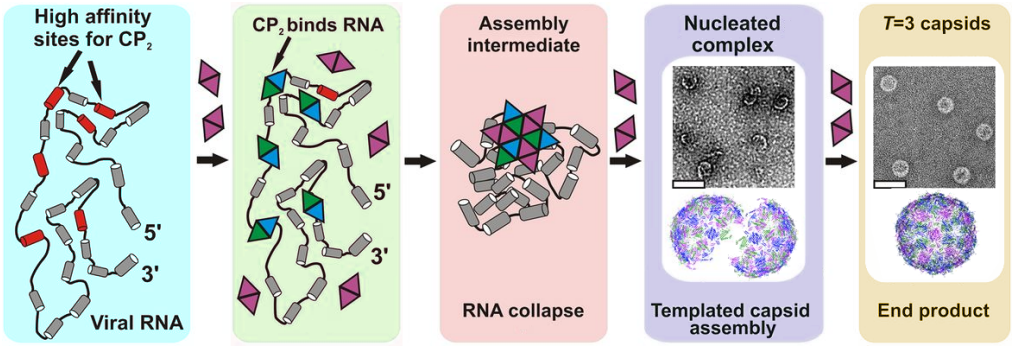
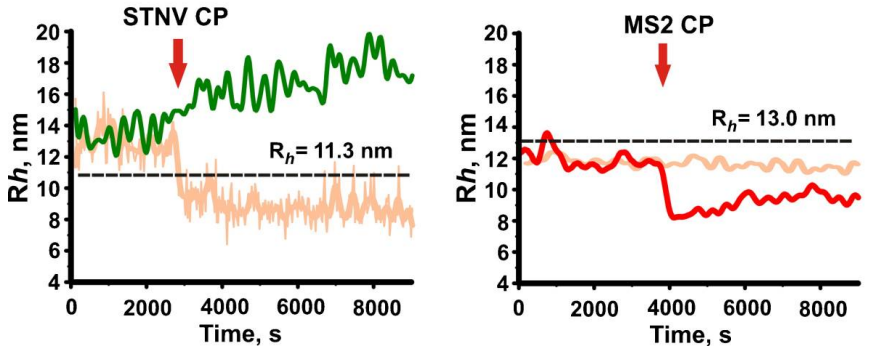


Figure 5.6: Cartoon of a PS-mediated assembly process, adapted from [213].

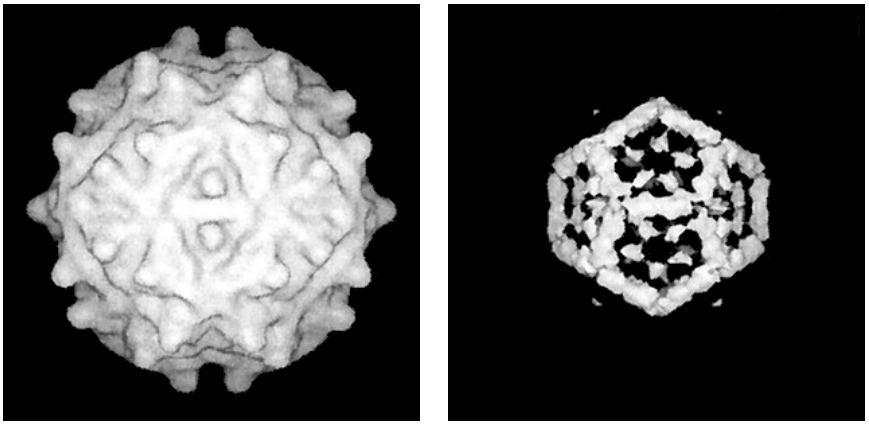


(a) STNV collapse (b) MS2 collapse

Figure 5.7: Collapse of STNV and MS2 genomes on addition of cognate CP: addition of CP to the genome forces a collapse in the hydrodynamic radius (R_h, *i.e.* size) of the genome, as it adopts a new conformation for packaging. The R_h for genomic RNAs (at 1nM) (a) STNV and (b) MS2, derived from smFCS[2], is seen to reduce rapidly on the addition of cognate CP subunits (red arrow), which allow co-assembly of the capsids. Before CP addition, multiple genome conformers are present in equilibrium; many conformers are larger than their respective capsids (capsid external radii given by dashed black line); adapted from [9].

ways.

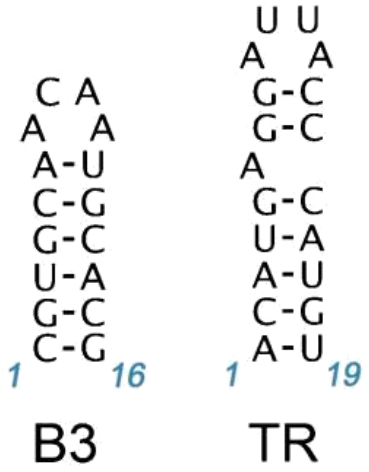
In this thesis, we will focus on the well-characterized assembly methods for two ssRNA viruses: MS2 and STNV. For these viruses, the PSs within the genome are short RNA stem-loops (see Figure 5.9) [8–10, 34–36, 215]. Indeed, STNV and MS2 are thought to encompass up to 30 and 60 degenerate PSs, respectively, of varying CP affinity within each respective genome [34, 206]. Investigation of the evolved assembly methodologies of these and other ssRNA



(a) PaV particle structure (b) PaV genome organization

Figure 5.8: Structure of WT PaV, (a) particle determined to 23Å resolution by cryo-EM, and (b) the double-stranded RNA cage resolved *via* a difference map approach. Adapted from [214].

viruses reveals many similarities in the strategies and mechanisms exploited. In fact, PSs have been identified for other RNA viruses, also in the form of a defined element of secondary structure, such as a stem-loop or collection of stem-loops. Non-enveloped viruses for which the PS assembly paradigm has either been established, or evidence exists suggesting that it is likely to occur (specifically the existence of a high affinity PS) include those from a number of different families, including *Picornaviridae* [37], *Secoviridae* [38], *Bromoviridae* [39–41], *Leviviridae* [35, 36], *Nodaviridae* [42, 210], *Virgaviridae* [43, 44], *Tombusviridae* [45], *Tymoviridae* [46], and satellite viruses [47–49]. PS assembly mechanisms are also implicated by evidence of a PS in large enveloped ssRNA viruses, including members of the *Coronaviridae* such as mouse hepatitis virus (MHV) [50], Rift Valley fever virus (RVFV) [51], and the etiologic agent of SARS: SARS coronavirus (SARS-CoV) [52]. PSs are also believed to play an important role in the assembly of Gag in human immunodeficiency virus type 1 (HIV-1), a member of the ssRNA-RT retroviruses [53], and hepatitis B virus (HBV), a dsDNA-RT virus with a single-stranded pregenomic RNA (pgRNA) cargo at the moment



(a) STNV PSs (b) MS2 PSs

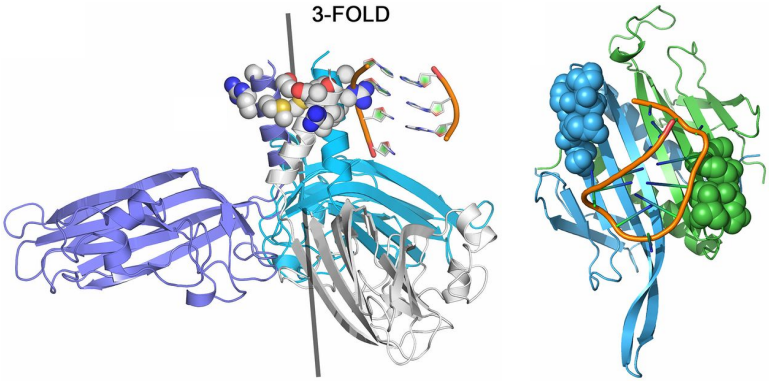
Figure 5.9: Structure of high affinity STNV and MS2 PSs: the sequence identity and secondary structures of the first known PSs of STNV and MS2, B3 [34] and TR [215] respectively. Other PSs with similar sequence and structural motifs are seen at positions within each respective genome [34, 206].

of packaging [216]. However, the most comprehensive evidence for PS-mediated assembly in viruses comes from the model system MS2.

For each virus, the structural elements comprising the PSs are finely tuned in terms of affinity and specificity for CP: each PS within the genome has a variation on a common stem-loop motif, each with a different affinity for the binding site [8]. Identifying these shared PS motifs in the genome requires consideration of the primary and secondary structure of the RNA [34, 206, 217], yet efficiently and accurately predicting the secondary structure of a nucleotide sequence is difficult, and remains an active area of research in computational biology [217]. With respect to identifying PS motifs, of interest is the folding kinetics and stability of the predicted local secondary structures likely to be formed by a genome, which relate to the local environment (*i.e.* the surrounding protein, nucleic acids, ions, *etc.*). In solution, the thermodynamic stability of folded RNA secondary (and, by extension, tertiary) structure is particularly

dependent on interactions between the negatively charged RNA polyelectrolyte phosphate backbone and surrounding positive counterions, which shield the electric charge and preserve overall electrical neutrality [218]. In particular, divalent (+2) cations such as Mg^{2+} are more important for the stabilization of the densely charged RNA structure than univalent (+1) ions, even at sub-millimolar concentrations [219, 220]. Counterions can also create a favourable environment for the formation of PSs, by producing a net attractive free energy for folding of secondary structure motifs [218, 219]. Within the context of the densely-packed assembling virus, charged nucleotides and charged amino acid residues originating from the capsid both also contribute to the overall electrostatic environment, and thus to the stability and folding kinetics of the encapsidated genome secondary and tertiary structures. Furthermore, the protein surfaces in proximity to the PS binding sites are often positively charged, indicating a role in neutralizing the RNA negative charge [221, 222]. Importantly, secondary structure predictions of the identified PSs of several viruses show that only a minority of the predicted PS stem-loops are folded in the lowest-energy conformation, implying that the folding of PSs is kinetically driven [34, 206].

It is also possible to find information on PSs by examination of the evolutionary conservation of viral genomes over time: PS positions, especially the high affinity ones, have been seen to be conserved. For some of these viruses, PS locations within the genome are known to encode high-affinity CP binding sites that pair with these PSs on assembly, and thus contribute to selective packaging [211] (Figure 5.10). This facilitates evolution within the PS domains, as there are more possibilities for mutations that preserve binding affinity within this region. These viruses have evolved to be resilient (in evolutionary terms) to challenges to their assembly mechanisms, and the PS-mediated co-assembly process provides a means of doing this. To examine this further, it is conve-



(a) STNV PS3 encode CP binding sites

(b) MS2 PS SL-2 encode CP binding sites

Figure 5.10: PSs in STNV and MS2 encode CP binding sites, from [211]. (a) PS3 in the STNV genome, identified in Patel *et al.* 2015 [211]. (b) PS SL-2 in the MS2 genome identified in Dykeman *et al.* 2013 [206].

nient to consider the viruses infecting a host at a specific time as a quasispecies: the stationary state of a set of interrelated genotypes that evolve *via* selection and mutation within the niche that is the particular host [23, 223–225]. Upon a perturbation to the system, such as a challenge by drug or immune response, the environment of viral replication, and in our case assembly, will change. A different assembly pathway could be required, relying on a different order (and *in extremis*, type) of PS-capsid interactions. The adaptability of the system is contingent upon the degenerate nature of PSs within the viral genome, which allow the order and strength of these PS-capsid interactions to be varied under different conditions; over the quasispecies as a whole, infectious progeny can still be formed under increasing or decreasing pressure from the host.

It is therefore difficult to disrupt viral assembly due to the flexibility of the PS assembly mechanisms. Interceding in all PS-CP interactions equally would partially attenuate assembly, but the degenerate nature of the PSs means that others with different affinities could bias assembly towards stable pathways [11, 23]. Therefore an understanding of the different PSs, their distribution in the

viral genomes, and their role in controlling the order of assembly is required. One approach could be to direct assembly into unstable pathways, where aggregating CP could become stalled in kinetic traps.

5.3 Structure informing viral mechanisms

Solving the structure of full virus particles, as well as that of viral components, is a well-established area of scientific research that is notable for its interdisciplinary collaboration of biologists, chemists, physicists, and mathematicians. Over the years, novel biochemical and theoretical methods have been frequently used to gain extra insight into virus structure, as well as the mechanisms of capsid formation and the pathways of self-assembly [8, 11, 206, 226–229]. However, the principal techniques used for determining virus structure remain X-ray crystallography and electron microscopy, the latter of which is recently breaking new ground with a rejuvenation of cryo-EM and cryo-ET studies [230]. In these techniques, the 3-D virus structure can be reconstructed from electron micrographs of virus particles, which are projection images, using computer image analysis protocols [203].

For viruses that use PS-mediated assembly mechanisms, 3-D structural information of the asymmetric distribution of PSs within the capsid can be used to determine the order of assembly (*e.g.* MS2 in Figure 5.11) [12, 205]. The general organization of genome within viruses has been probed experimentally, including that of MS2 [35,36,205]. Studies have shown that the genetic material is distributed inhomogeneously inside the capsid, in a multi-shell arrangement. In MS2, based on the space occupied by the nucleotides of the translation repressor (TR) seen in virus-like particle (VLP) crystal structures as a guide, approximately 90% of the genome fills an estimated 25-30% of the interior of the capsid. Because of the wide range of estimates of RNA density that have been

published, not all the genome could be accounted for [36]. More generally, the RNA appears very dense close to the inner surface and less dense in the centre, which in many reconstructions appears completely devoid of ssRNA. Indeed, the central region of RNA is not thought to adopt a particular structure; its organization—pseudo-liquid crystalline—is a consequence of the self-repulsion of the negatively charged RNA [231, 232].

On symmetric icosahedral averaging of cryo-EM reconstructions, the ssRNA genome in proximity to capsid is often seen as a well-ordered polyhedral cage, as seen for the *Leviviridae* in Figure 5.12; this region contains the PS RNA-CP interactions. The positions of the PSs in the genome determined for MS2 suggest that the connections between PSs are single-stranded [206, 233]. The polyhedral cage arrangement reflects the PS binding sites that are made by the genome to CP, which appear in what is known as the *outer shell* of two RNA shells in these reconstructions, intimately associated with the interior of the CP shell [36]. The central liquid crystalline region of RNA in the core of the virus is thought to be largely double-stranded, as this would allow the densest packing possible between the charged RNA molecules, and connections between the shells are consistent with duplex RNA segments. The start and end parts of the double-stranded central RNA portions, where the segment connects to the outer shell, are thought to be located at the same five-fold vertex: indeed at the same PS position [36].

It can therefore be deduced that every PS is connected by the single RNA molecule to two other PSs in the outer shell, as any double-stranded regions extending into the central core do not affect the connectivity of the outer shell [12, 206]. Thus the outer shell alone can be used as a topological tool to rationalize connectivity between PSs. As a whole, the RNA in the outer shell can be viewed as forming a connected path between all of the PS positions, representing the

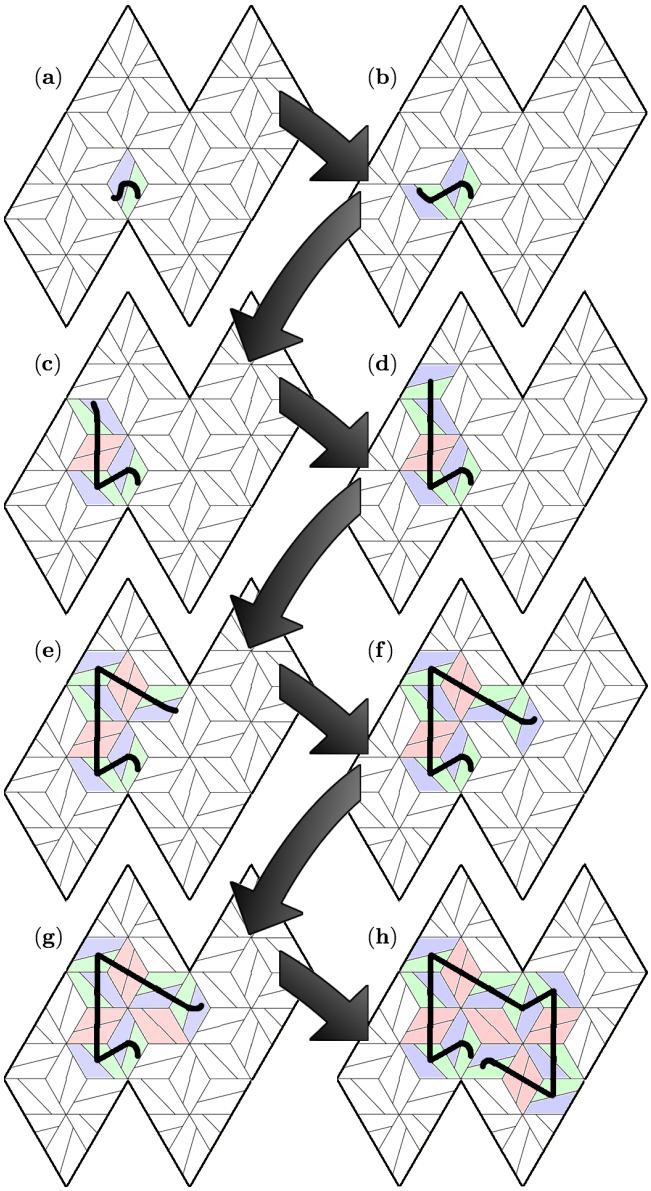


Figure 5.11: MS2 assembly pathway implied by RNA conformation: the order of CP addition maps to the asymmetric structure of the RNA molecule (in black) inside the MS2 virion, in proximity to the capsid. (a) Nucleation of the ssRNA-CP complex at a CP dimer, in which a PS stem-loop acts as an allosteric effector of conformational switching of the dimer to an asymmetric conformation (blue/green). (b) A second dimer is recruited, and binds to the adjacent PS position on the genome. (c) Two more dimers are recruited to the complex, but one does not bind a PS stem-loop and remains as a symmetric homodimer (pink). (d–g) The complex continues to recruit both homo- and heterodimers to satisfy the capsid geometry. (h) The path can reach a dead end, whereby correction of the intermediate conformation is necessary to achieve completion.

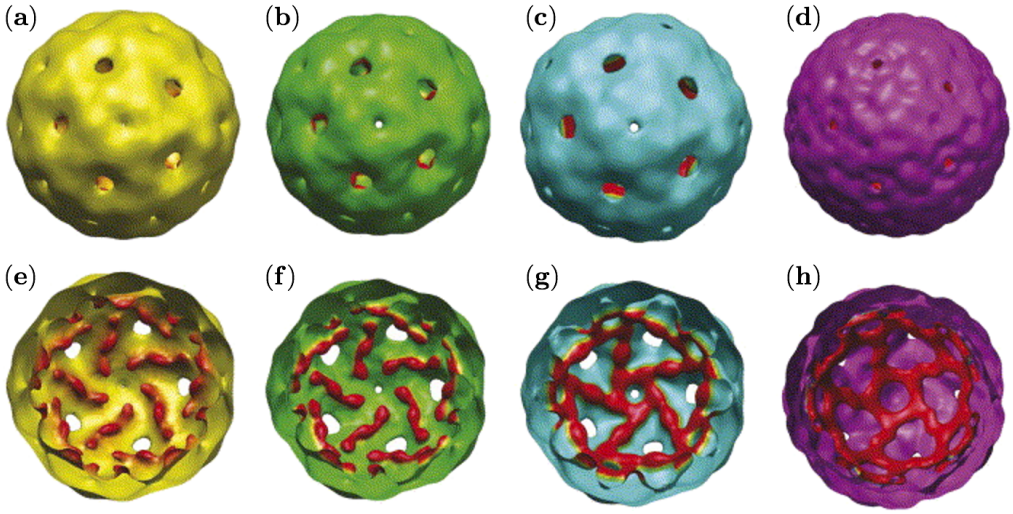


Figure 5.12: Genome organization within the *Leviviridae* (red, RNA). (a–d) Exterior and (e–h) interior of bacteriophages (a) MS2, (b) Qβ, (c) PP7 and (d) AP205, orientated to the five-fold axis [35].

order in which CP-PS contacts are formed during assembly (Figure 5.11). In this scenario, the outer RNA shell is necessarily asymmetric, as perfect icosahedral symmetry is not possible in the case of a connected linear genome [234].

Connectivity of the outer shell is justified by the reasoning that if neighbouring PS did not map to neighbouring positions along a path, PSs at different five-fold vertices would have to be connected directly *via* RNA passing through the capsid interior, which is not consistent with the liquid crystalline nature of the core [36]. This alternative scenario would require a complicated network of long-distance interactions, and would be inconsistent with the analysis that the core RNA is formed of stem-loops that are kept contained in the core by electrostatic interactions, provided that there is sufficient negatively-charged RNA present [36, 206]. The ssRNA remaining fully connected in the outer shell also keeps the RNA folding uncomplicated, in that predominantly the RNA only needs to fold simple stem-loop structures, ranging in size from the small PS stem-loops to the large internalized liquid crystalline stem-loops. Furthermore,

the genome needs to be packaged in such a way that no knots are formed: knot formation could prevent genome ejection and subsequent translation.

Observed in asymmetric reconstructions (and to some extent in icosahedral reconstructions) is an asymmetric distribution of RNA within the symmetric CP shell, with an ordered asymmetric outer shell of ssRNA intimately associated with the interior of the protein capsid, for which the organization of the ssRNA and hence organization of PSs has implications for the order in which CP is recruited to the growing capsid shell [12].

Importantly, this asymmetric distribution of viral genomes within a virion may also be an essential factor in the extrusion/uncoating of these genomes as the first step in subsequent infection [19, 205, 235–239]. Thus, the ssRNA genome plays many concurrent functional roles during both assembly and disassembly/uncoating, underpinning the efficiency of vital stages of the life cycle.

5.4 Hamiltonian paths

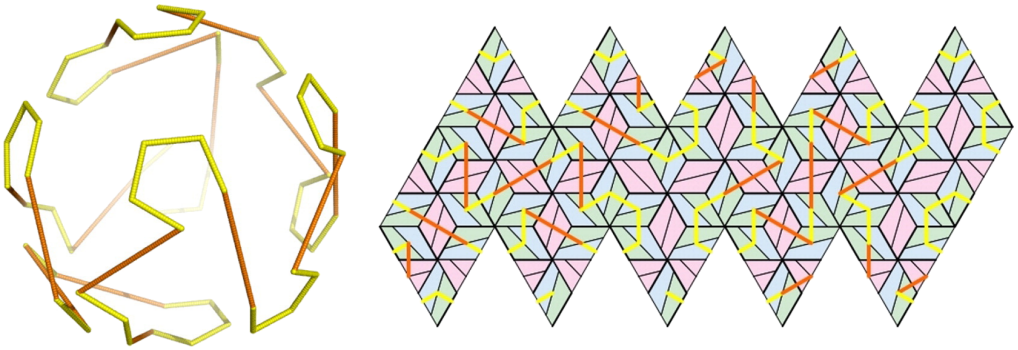
A Hamiltonian path is a concept from the mathematical field of graph theory. It is defined as a path over a connected graph, either directed or undirected, that visits each and every node exactly once. A Hamiltonian cycle is a Hamiltonian path that is cyclic: *i.e.* a path that is both continuous and unbroken. Hamiltonian paths can be applied to the field of virology as a description of part of the genome inside some viruses, as it is an appropriate concept to explain the combinatorial ways in which different PSs can be connected in a way that is consistent with the observed cryo-EM density corresponding to RNA [12].

As we have discussed, in many viruses there is a ssRNA outer shell in contact with CP that is seen as a polyhedral cage under icosahedral averaging [36], and that the inner shell can effectively be ignored when considering the connectivity of the PSs. The organization of the RNA in the outer shell can therefore be

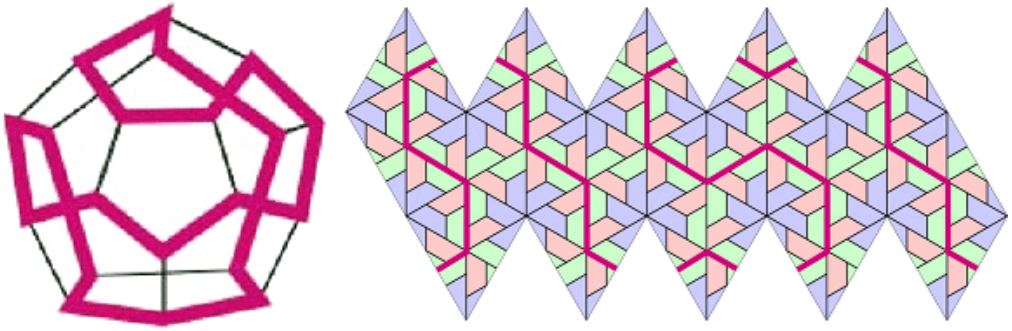
equated to a Hamiltonian path on that polyhedron, *i.e.* a path that meets all the PS positions located at vertices of the cage. Importantly, every possible organization of the RNA can be found by computing all possible Hamiltonian paths on the polyhedral cage.

Hamiltonian path approaches have been used in a number of studies of PS-CP interaction, for example: (i) kinetic modelling of MS2 capsid self-assembly [10], (ii) a bioinformatics analysis of PSs in MS2 and the related phage GA [206], and (iii) structural studies (Figure 5.13). Specifically, RNAs seen within virus structures have been modelled by Hamiltonian paths in studies of MS2 [12, 206] and the double-stranded RNA (dsRNA) PaV (Figure 5.8) [234]. A new example is given in Chapter 6, where we enumerate Hamiltonian paths derived from local rules that govern the self-assembly of STNV. For each virus, the specifics of the viral assembly mechanisms must be built into the formulation of a constraint set: namely, the local rules on how to move between PSs. However, similar constraints are likely to apply across viral families for which the same assembly mechanism has evolved. For example, the Hamiltonian paths derived for MS2 (Chapter 7) would likely be shared across the rest of the *Leviviridae*, because the topologies of the averaged polyhedral RNA densities are the same (Figure 5.12). It must be noted here that the PaV example shown in Figure 5.13(b) due to the viral dsRNA is very different to the PS examples discussed above: for the main assembly there is a Hamiltonian cycle based on a duplexed triconnected graph running under two-fold axes, with vertices under each three-fold axis and additional side branching RNA portions. These side branches take the form of bulging stem-loops and pass under unoccupied two-fold axes, connecting to the N-termini of CP where the PS is thought to bind [234].

For other viruses, in which occupation of the majority of the PS binding sites is likely due to their function in assembly, and for which the PSs are positioned



(a) Hamiltonian paths on MS2



(b) Hamiltonian paths on PaV

Figure 5.13: Hamiltonian paths describe the outer shell of RNA, in MS2 [12, 206] and PaV [234].

at the vertices of the RNA cage corresponding to the icosahedrally-averaged map of the genome in proximity to capsid, the constraint set is also given by Hamiltonian paths. Naturally, the set of Hamiltonian paths will depend on the numbers of PS binding sites and the connectivity between them. If there is evidence that the $5'$ and $3'$ ends are in proximity in the packaged genome, then the set of constraints can be reduced to only circular Hamiltonian paths; this represents a large reduction in the complexity of the search space.

In Chapters 6 and 7 it is demonstrated that the method of using constraint sets inspired by insights from PSs can allow refinement of structural data. Specifically, asymmetric structural information is compared to Hamiltonian paths connecting the PS contact sites. The asymmetric organizations of the packaged

genomes are able to be revealed in some detail using a Hamiltonian path-based analysis, by comparing constraints encoding all possible ways of connecting PSs with asymmetric structural data [12], in the cases of STNV and MS2. However, as we argue above, the method of interrogating structural data *via* constraint sets inspired by PS-mediated assembly mechanisms is applicable to many classes of viruses.

Note that this method also applies if some of the potential binding sites remain unoccupied in random positions across the ensemble of particles used to generate the tomographic data, as such random mistakes would not be reinforced during averaging over different particles: hence it is sufficient that the majority of PS binding sites are occupied. For each virus, multiple possible constraint sets arise from different assumptions and considerations on the specifics of the PS-mediated mechanism of assembly. In the case of insufficient information being available to decide *a priori* between these constraint sets, asymmetric structural data could be interrogated against the different possible options, to indicate which proposed PS-mediated assembly mechanism is likely to occur, thus validating hypotheses on assembly mechanisms indirectly in an interdisciplinary approach.

Chapter 6

Analysis of STNV self-assembly

An important characteristic of ssRNA[1] viral genomes is the negative charge that they carry, which is due to the phosphate groups in the RNA backbone [240]. Upon viral assembly, compacting this highly negatively charged molecule into the protein capsid requires positively charged residues on the interior of the viral capsid, to coordinate the folding and encapsidation of the genome.

In the densely packed environment of ssRNA viruses, contributions to the overall electrostatic field arise from all charged constituents of the capsid, including charged nucleotides and charged amino acid residues, and these electrostatic interactions are, by their nature, of a long range [241]. As discussed, the capsid charge is mainly positive, as is the exterior, whilst the charge inside is negative; however, some of the negative charge is not shielded and can be found outside (as can positive charge inside) [242].

In the RNA-capsid co-assembly paradigm, the ssRNA genome directly binds to capsid proteins, facilitating the assembly of the capsid [8]. The stable asym-

[1]For all abbreviations see the glossary on page 222.

metric organizations of the mature viral genome predicted by this process have been well discussed in this thesis. This asymmetric organization of RNA will lead to an asymmetric distribution of charge within the capsid. The consequences of this asymmetric distribution of charge with respect to the crystallization of mature virions will be discussed in this chapter, with respect to a small ssRNA virus, STNV.

Similar to most ssRNA viruses, STNV is known to be overcharged, implying an absolute polyion charge exceeding that of the capsomers [243]. However, its CP has an isoelectric point (pI) of 10, meaning it is a polycation at physiological pH [49]. As we shall discuss here, unshielded electrostatic charge from conserved genome tertiary structure (arising from RNA-CP co-assembly) biases the orientation of STNV virus-like particles (VLPs) during crystal formation, and has allowed an asymmetric crystal structure to be determined.

6.1 STNV

With a diameter of just 19 nm, STNV is one of the smallest known viruses. As its name suggests, STNV is a satellite virus of tobacco necrosis virus (TNV), *i.e.* it relies on a coinfection or superinfection of TNV for its replication [244, 245]. *In vivo*, STNV and TNV are transmitted by the soil fungus *Olpidium brassica*, from where they infect the root cells of tobacco plants. The STNV genome, which is ssRNA and 1239 nt in length, encodes a single 22 kDa coat protein monomer, of which 60 copies arrange into a $T = 1$ icosahedral capsid. The STNV genome encompasses three main functions in the viral life cycle: it operates as a substrate for replication and translation, and assists in the viral co-assembly processes outlined in Chapter 5.

STNV and its helper virus TNV have little sequence similarity at both amino acid and nucleotide levels, and their coat proteins show no antigenic cross-

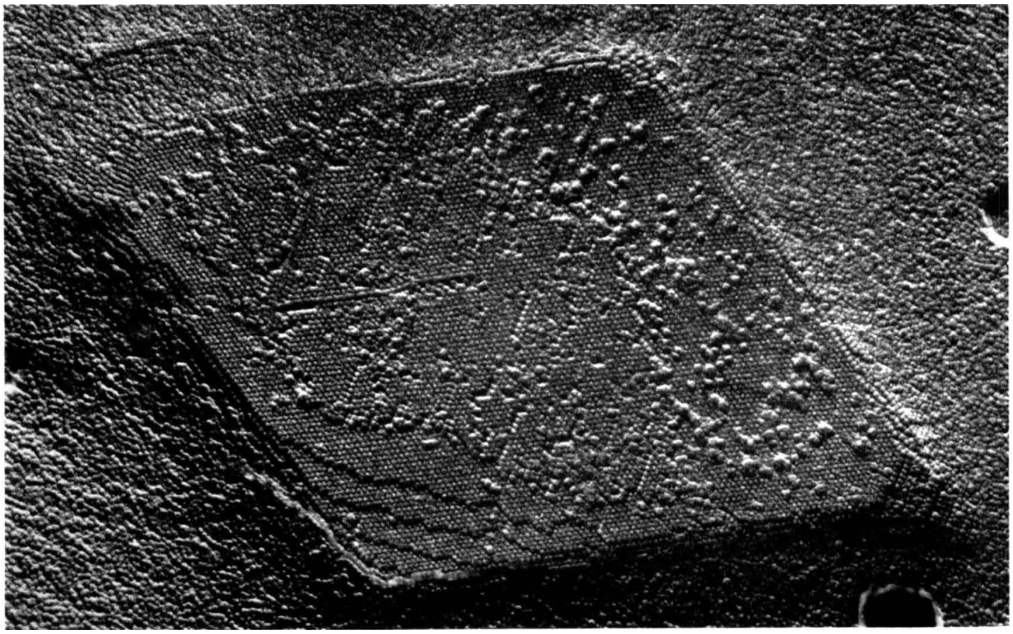


Figure 6.1: Micrograph of crystalline STNV (platinum shadowed carbon replica at 54,000× magnification), taken from [244].

reactivity [118]. However, due to STNV not possessing its own replicase, TNV RNA-dependent RNA polymerase (RdRp) binds to the STNV genome at the 3′ end, across a stretch of 100–150 nt with some sequence similarity.

The structure of STNV has been determined by various 3-D imaging techniques. In particular, due to the inclination of STNV to crystallize (Figure 6.1), observed *in vivo* [118], X-ray crystallography has been used to great effect, resolving the WT structure to 2.45Å [246]. In spite of this, the encapsidated genome could not be visualized by crystallographic methods: in the high-resolution structure, no interpretable density could be associated with RNA, suggesting that there is no repeating organization within the encapsidated genome relative to the icosahedral symmetry of the capsid.

This is a conundrum—as discussed previously, if RNA-CP contacts are important for assembly, why aren't they resolved in the high-resolution structures of this and other viruses, and why is there so little RNA density within these

reconstructions? A possible partial explanation is the absence of low-resolution terms in the Fourier synthesis. Moreover, whilst the RNA molecule packaged within a single virion does not strictly have icosahedral symmetry, on averaging the symmetry will be imposed. Note that even if the conformation of the genome was unique, it would be impossible to confirm this under these circumstances as different symmetry axes would be aligned in different particles, thus averaging out this information, *i.e.* icosahedrally averaging the data to high resolution would eliminate much of the density for the pseudo-icosahedrally organized viral RNA.

However, low-resolution neutron diffraction did reveal genome structural information, using a ${}^{1}_{1}H_2O/{}^{2}_{1}H_2O$ contrast matching approach [47]. At 16Å, the resolution of this density was too poor for any RNA structure to be resolved in any great detail. Nonetheless, RNA density was determined in the vicinity of five-fold symmetry axes, between clusters of CP N-terminal arms. The genome density is localized in a region radially positioned at the same level as the protein capsid, *i.e.* inside indentations and crevices, implying a significant amount of interaction between the two components, and is thus suggestive of a PS-mediated assembly process. Although the genome structure and its interaction with CP is yet to be determined in detail, sequence-specific PSs have been found in the RNA genome [34]. The corresponding RNA-CP interactions occur in symmetry related positions and thus features of the density are expected to be retained in the icosahedrally averaged map, as discussed in Chapter 5.

6.2 STNV VLP model

STNV VLPs formed by integration into an *E. coli* recombinant system were examined recently for the structures of their RNA-CP contacts [49]. The STNV CP assembles *in vivo* into capsids that closely resemble the WT virus, pack-

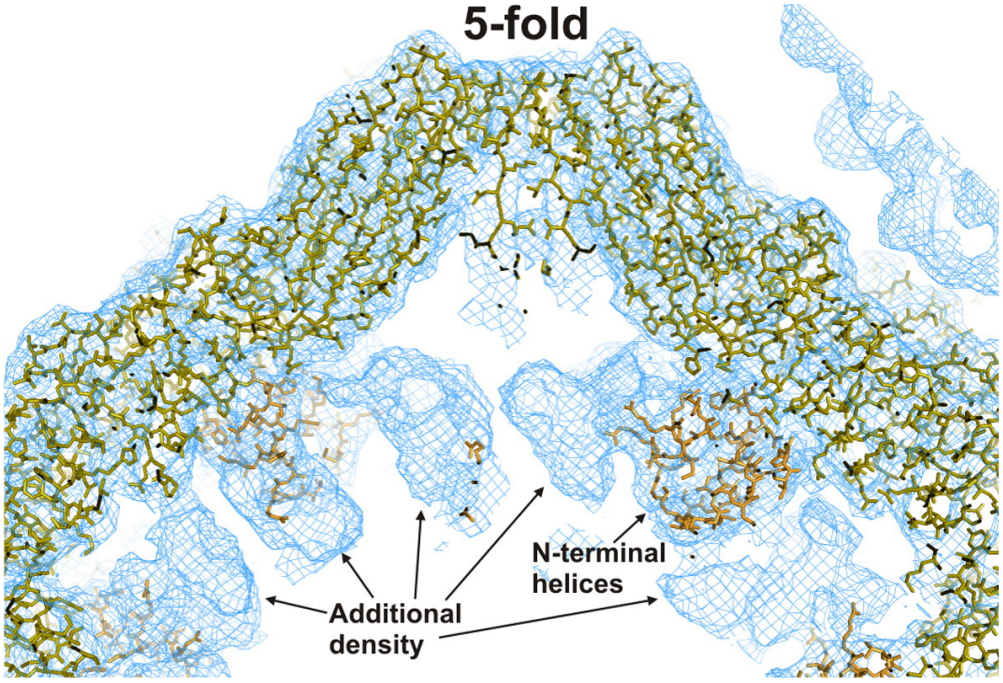
aging the recombinant messenger RNA (mRNA) transcript. The 2.45Å resolution STNV WT structure (PDB:2BUK [246]) was used as a basis for the rigid body refinement and icosahedral averaging, leading to a final resolution of 1.45Å (PDB:4V4M [49]). The densities of the models match well over the CP (Figure 6.2(a)), but a large area of electron density attributed to RNA PS positions is unaccounted for in the high resolution WT structure, perhaps because PSs could be bound in slightly different orientations due to the (albeit limited) degree of structural variation across the PS ensemble. The PS-CP contacts are positioned between the N-terminal triple-helical arms of the CP, and overlap with regions of RNA density that appear duplexed, *i.e.* are double-helical in structure (PDB:3S4G, Figure 6.2(b)). Note that about 72% of the total RNA in STNV appears duplexed. We will further examine this structure of the encapsidated B3 aptamer [34, 49, 247] to test the nature of STNV PS-CP interactions: moreover, it is a suitable model to predict the likely genome organization within the WT virus.

Viral genome packaging is understood as taking place on nascent +ssRNA strands as they are released from the RdRp (which for WT STNV is the TNV RdRp). In the VLP case, the RNA-CP complex is formed around the an RNA fragment rather than the full genome, and replicated by an *E. coli* T7 RNA polymerase; both of these components are different to the WT case [211].

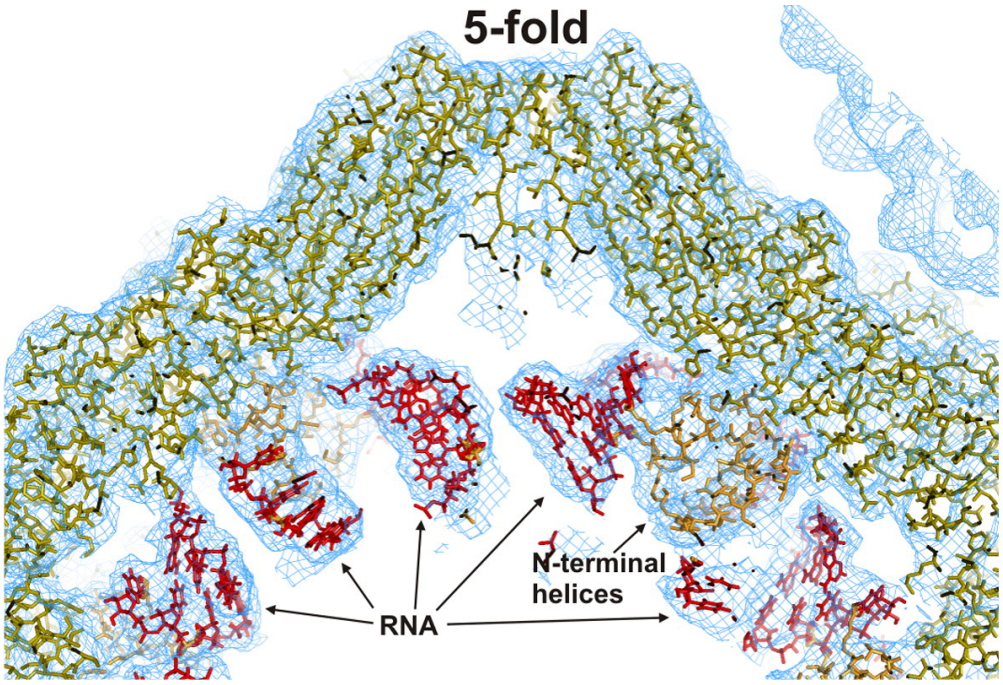
6.3 PS of STNV

Multiple degenerate PSs have been identified in the STNV genome. Using the VLP model, up to 30 small stem-loop PSs were identified with a characteristic -AxxA- motif predicted *via* bioinformatics and RNA SELEX[2], where x is any nucleotide [34, 211, 247]. The putative STNV PSs within the genome were

[2]Systematic evolution of ligands by exponential enrichment.



(a) Low-resolution averaged electron density map



(b) dsRNA fragments built into additional density

Figure 6.2: STNV RNA structure. (a) A low-resolution electron density map, with STNV CP fitted into corresponding density for the shell and the N-terminal helical protrusions. (b) There is strong additional non-CP density between the helix clusters, into which RNA fragments are fitted: there is no strong RNA density directly under the 5-fold vertices or in the centre of the particle (PDB:3S4G [49]).

identified *via* sliding of a small window, followed by folding of the genomic sequence overlapping with the window, and a subsequent search for the -AxxA- motif within the folded secondary structures. All such stem-loops with the motif, with a negative free energy of formation, were identified [34]. Within the PS-mediated assembly paradigm, there is an essential requirement for the CP affinities of the PSs to be hierarchical, with the 30 PSs exhibiting a range of distinct binding affinities for CP.

As discussed above, each RNA PS stem-loop is positioned in the packaged genome adjacent to two CP subunits, and contacts the N-terminal triple-helical arms of the CPs. Each stem-loop position is close to a neighbouring stem-loop position across a particle two-fold axis. The protein surfaces surrounding the PSs have excess positive charge [221], suggesting a role in charge neutralization.

6.4 Local rules

6.4.1 Connectivity

Hypotheses on the order of protein association in STNV assembly can be derived from homology modelling[3] of the possible RNA connections between PS binding sites in the interior of the STNV capsid (using PDB:3S4G [49]). We have formulated our hypotheses on connectivity between RNA-CP binding sites in local rules capturing the assembly process, called in the following the *permissible moves*, or moves in short. These are displayed in Figure 6.3, and correspond to all connections between RNA-CP binding sites that we positioned with a distance below 51Å, see Table 6.1. It is unlikely that moves corresponding to significantly larger distances would occur. The consequences for protein assembly (in terms of the assembly pathways, and the organization of the packaged

[3]Homology modelling was undertaken in S2S-Assemble2 [248].

№	Approx. distance	
1	28.5Å	≈9 nt
2	50.8Å	≈15 nt
3*	—	—
4	31.5Å	≈9 nt
5	50.8Å	≈15 nt
6	8.6Å	≈5 nt
7	8.6Å	≈5 nt
8	49.8Å	≈15 nt
9	49.8Å	≈15 nt

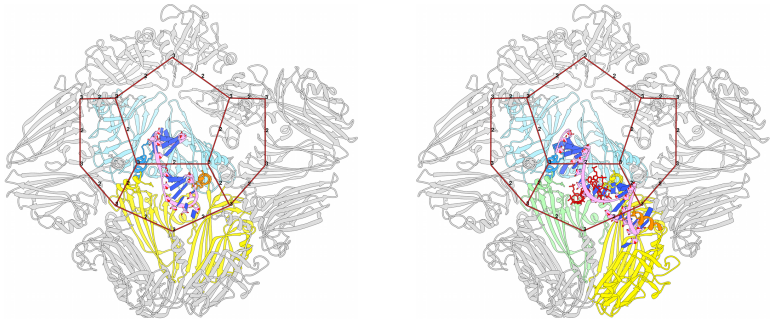
Table 6.1: Move distances: distances between RNA-CP binding sites corresponding to the permissible moves. Move 3 is omitted, as steric clashes with CP helical arms prevented the move (Figure 6.4).

genome in proximity to capsid) for different hypotheses can be determined *via* modelling and benchmarked against experimental results, thus providing a platform to indirectly decide which hypotheses are correct.

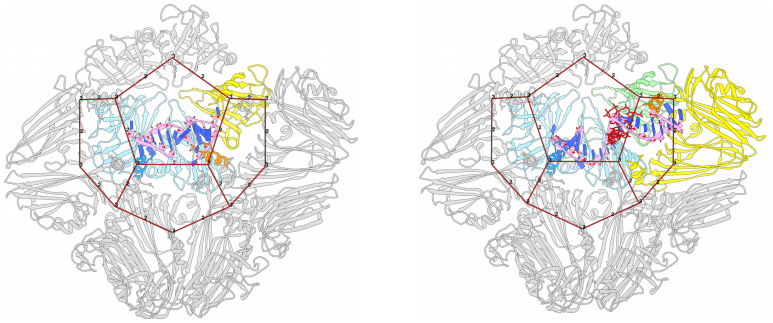
In particular, an understanding of these connections is essential as it has implications for the assembly pathways. These connections serve to recruit CP onto the nucleating capsid, in strict order, as seen in Figure 6.4. As discussed, each PS can attach to two CP monomers in a binding pocket. Thus each move of the RNA to realize a connection can recruit up to two CP monomers into their correct places on the growing capsid.

6.4.2 Nucleation of assembly

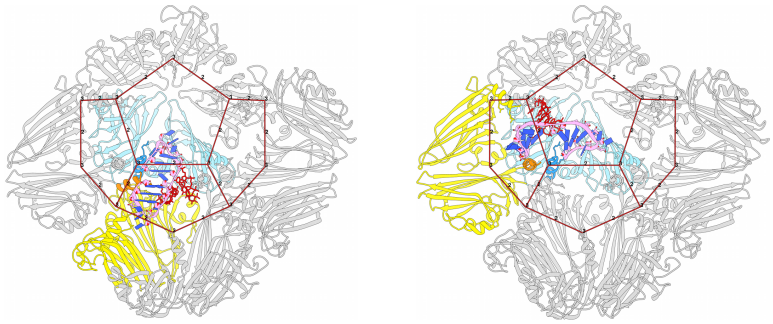
In the STNV trimer, there is a stabilizing effect of Calcium ions in the N-terminal arms [49], which is thought to create a pre-stable form before initiation of RNA/CP co-assembly. Indeed, in solution we see mainly monomers, but additionally with trimers free to initiate assembly, albeit at less than 1% of the total (Prof. Peter Stockley, personal communication). Our hypothesis is that packaging will commence with the nucleation of assembly at a free trimer, with



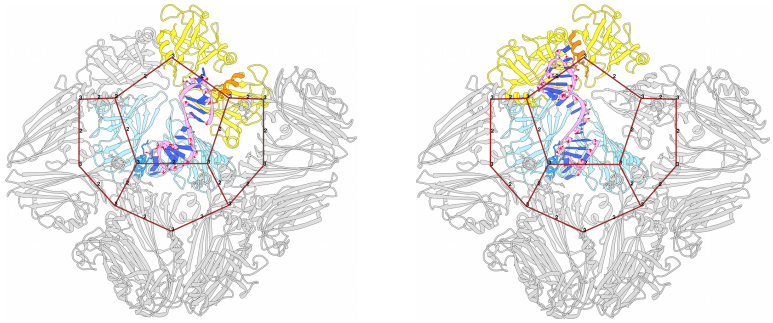
(a) Move 1 (b) Move 2



(c) Move 4 (d) Move 5



(e) Move 6 (f) Move 7



(g) Move 8 (h) Move 9

Figure 6.3: Illustration of permissible moves, encoding the local rules capturing the assembly process. The moves were identified using homology modelling of RNA connecting PS binding sites, in an X-ray structure of STNV VLPs containing PS B3 30-mer fragments (PDB:4V4M [49]). Move descriptions are in Table 6.1.

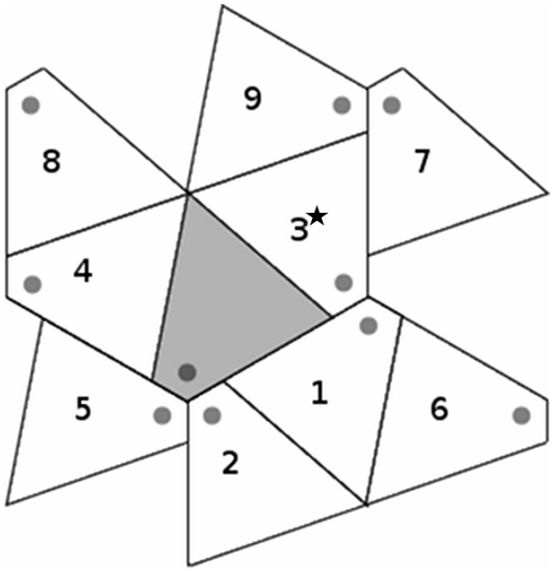


Figure 6.4: Recruiting of CP onto the nucleating capsid. Moves from an RNA-CP binding site on the grey protein to binding sites on proteins 1-9 were determined *via* homology modelling (Figure 6.3). No move to protein 3 was found, due to steric clashes between the RNA molecule and a N-terminal arm of an adjacent protein.

a PS interacting with one of the trimer PS positions. Note that we do not know the nucleation site on the genome, nor whether it is unique. We compare nucleation at any point with nucleation at the extreme $5'$ end of the genome. However, we assume that a trimer will nucleate.

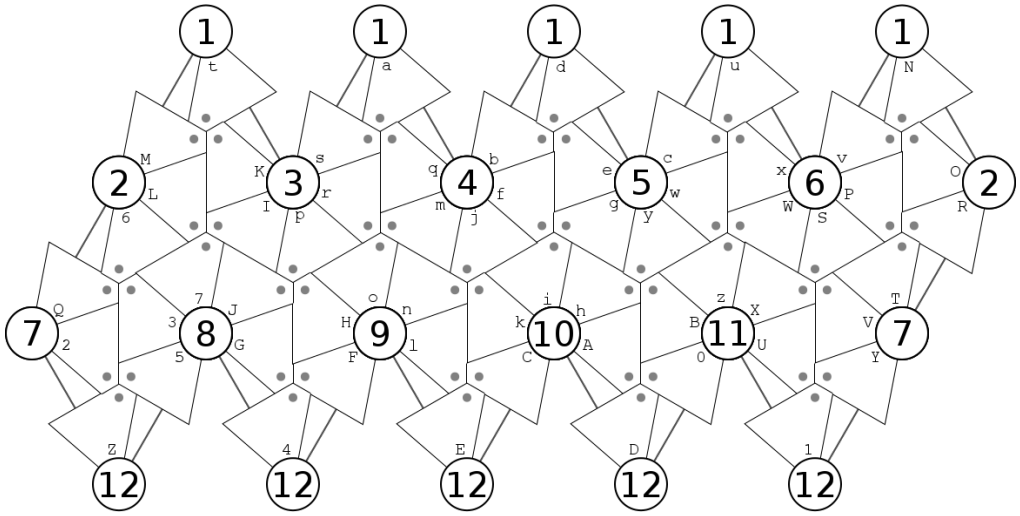
6.5 Connectivity paths

6.5.1 Generating paths

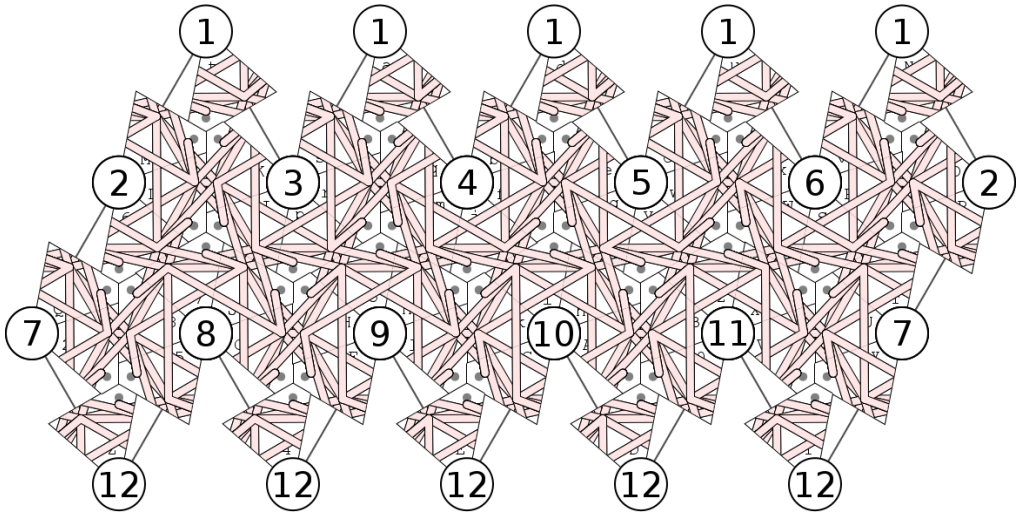
Paths representing connectivity between PSs are calculated with reference to the structural organization of the capsid. For this, a labelling system is introduced, in which proteins are labeled as (*cf.* Figure 6.5):

`abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSUVTWXYZ01234567`.

Paths are calculated starting at CP monomer `a`, without loss of generality. Then, for each incomplete path of length n, starting at $n = 1$, we determine all



(a) Scaffold



(b) Connectivity map

Figure 6.5: Geometry map: an icosahedral projection representation of the capsid and RNA permissible moves, from the inside of the capsid. (a) The arrangement of CP in the capsid, marked with their labels and N-terminal arm positions. (b) A diagrammatic overlay of all eight permissible moves between all RNA-CP binding points.

possible ways of completing the path by incrementing every possible single move from the final node. The allowed moves correspond to all, or a subset of, the eight moves derived from the homology modelling discussed above.

If trimers can form, we allow them to form: we preferentially allow trimers at the 3-fold axes to form if able to during the search for path solutions. If there are possible moves that result in trimer formation, either by extension towards the $5'$, or towards the $3'$, then these are selected in preference to other moves. This helps constrain the computational search space, which is very large given the number of possible moves from every point.

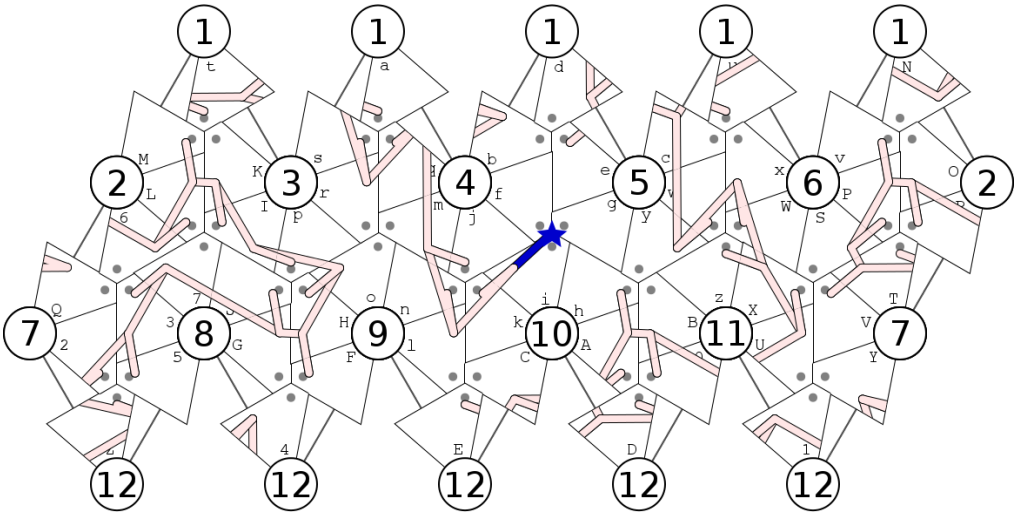
Nucleation, as discussed before, has implications for the initiation of the paths. Example paths that nucleate either at the $5'$ end of the genome, or elsewhere in the genome, are given in Figure 6.6.

6.5.2 Generalizing paths

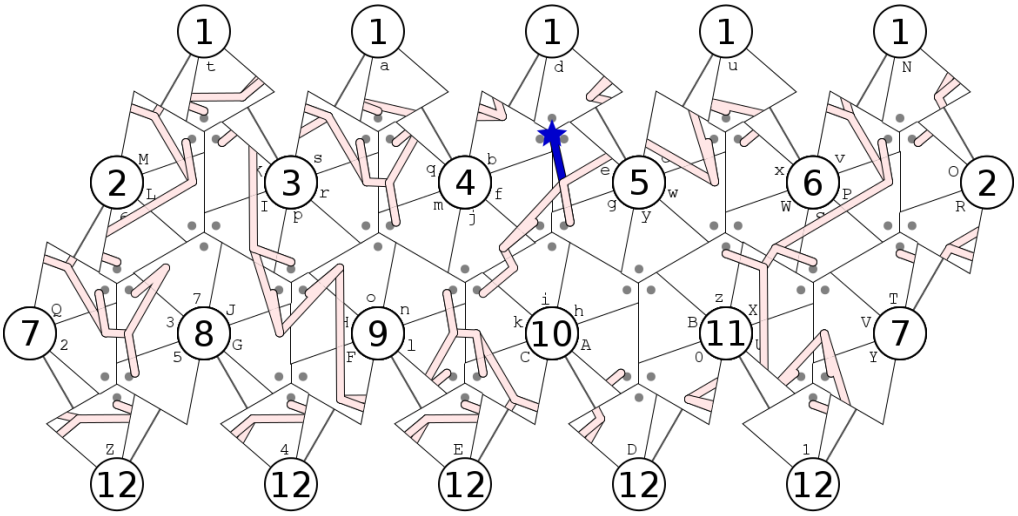
Once paths have been generated on the protein neighbour map, they can be generalized into moves, *i.e.* described as the sequential order of mapping operations between proteins (appearing in Figure 6.4), rather than the proteins themselves (as described in Figure 6.5). This allows an easier way to recreate the paths starting from any given protein. Also, calculation of symmetric and mirror paths is much easier with reference to the generalized moves, as these degenerate paths can then be calculated with simple string manipulations. The moves are referred to by the symmetry rotations given in Table 6.2.

6.5.3 Analysing paths

There are 156 paths from nucleation at any position in the genome, whereas only 27 paths if nucleation is allowed only at one end (Figure 6.7). The paths that are complete, *i.e.* lead to all proteins being recruited, are between 33 and



(a) Nucleation at PS at 5′ end



(b) Nucleation at PS three positions from the 5′ end

Figure 6.6: Nucleation impacts connectivity. Nucleation at (a) the PS at the 5′ end or (b) elsewhere at other PS positions leads to quantifiable differences in the connectivity of the genome, implying that the overall genome organization in proximity to capsid is qualitatively different for these scenarios.

№	Sub-moves	Description
0		Nucleation
1	DS	2-fold
2	DS, A5	3-fold clockwise
4	C5	5-fold clockwise
5	C5, DS	3-fold anticlockwise
6	DS, C5	Next-nearest 2-fold
7	A5, DS	Next-nearest 5-fold anticlockwise
8	C5, C5	2× 5-fold clockwise
9	A5, A5	2× 5-fold anticlockwise

Table 6.2: Move representation in terms of symmetry operations: DS is a jump across the nearest two-fold axis, A5 and C5 are anti/clockwise rotations about the relevant five-fold axis, respectively.

37 moves long (including nucleation), implying between 33 and 37 bound PSs. If nucleation is restricted to the $5'$ PS, the complete paths include a smaller range between 34 and 37 PSs. This is illustrated in Figure 6.8. An assembly mechanism favouring CP recruitment at the $5'$ end, which is highly likely here as genomes are packaged upon synthesis, would therefore bias genome organization to this smaller range.

It is clear that the latter is just a subset of the wider sampling, but this was not a foregone conclusion. The preferential allowing of trimers to form could have feasibly led to different paths appearing as solutions, simply as a result of restricting the search space. Interestingly, all complete paths initiated within the set of the first four PSs of the genome (Figure 6.8). This means that there is a geometric reason why assembly initiates at the $5'$ end, which enforces the packaging of RNA exiting the polymerase, and thus enhances assembly production.

If we quantify the similarity of the paths we can learn more about the mechanisms of capsid protein addition that are exploited during assembly, because the geometry of the paths has a direct correlation with the geometry of the capsid

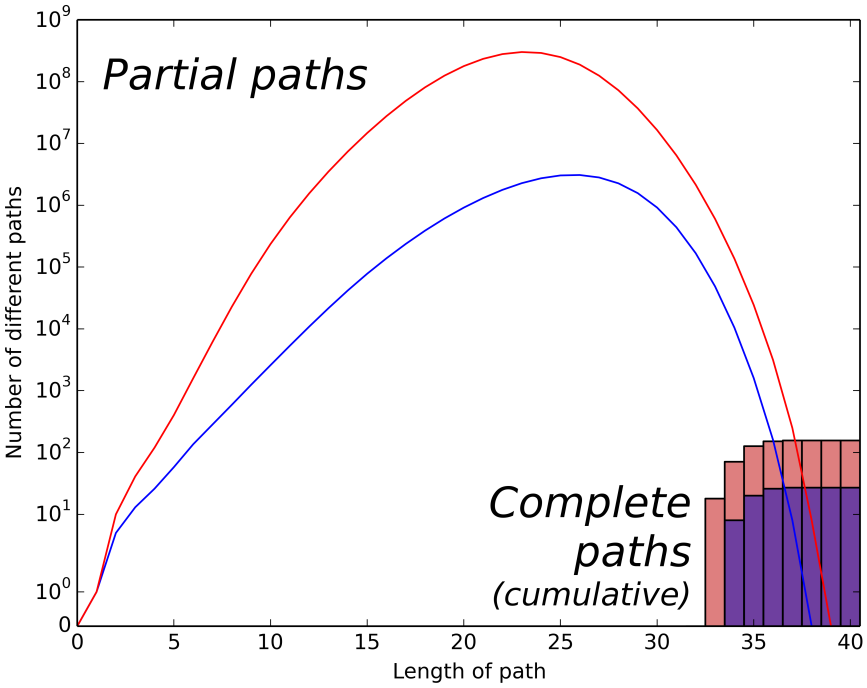


Figure 6.7: Number of paths by path length, assuming either nucleation at the 5′ end (blue), or nucleation at any PS position in the genome (red). Partial paths calculated for both scenarios, and are shown as the curves. It can be seen that initially the number of partial paths grows exponentially, as the moves are not constrained by pre-existing moves. As the algorithm enters later stages, the move opportunities become very constrained. Finally, some paths will be able to complete, *i.e.* when all CP has been recruited onto the capsid; these complete paths are shown as the bar chart.

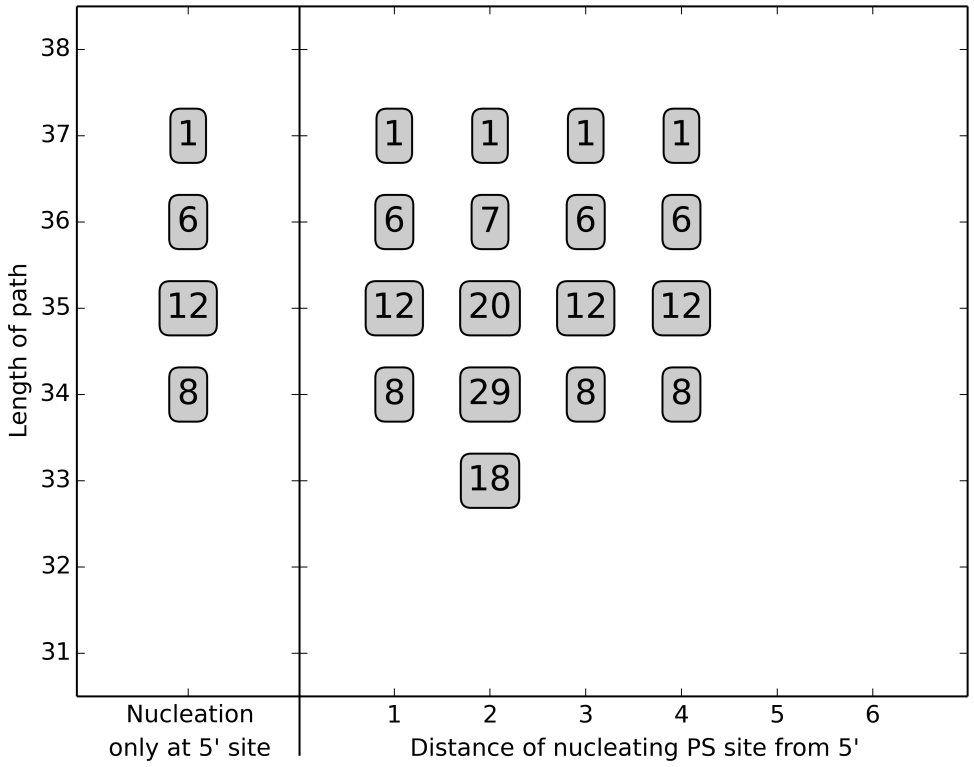


Figure 6.8: Path length and nucleation site of complete paths: if nucleation is allowed only at the PS site at the extreme 5′ end of the genome, then only 27 paths are possible, varying in length between 34 and 37 PSs. Alternatively, if we consider nucleation at any PS site in the genome, then it is calculated that only paths nucleating at the first four PSs from the 5′ end will lead to complete paths. These paths will have between 33 and 37 PSs bound.

№	Frequency
0	156
1	1005
2	12
4	1760
5	75
6	378
7	984
8	959
9	75

Table 6.3: Frequency of moves across solutions.

intermediates. In particular, across all the solution paths, we can quantify the number of times that a move is utilized (see Table 6.3).

Move number 4 appears the most in the 156 paths, at 1760 times. It is particularly common for the move to appear in clusters of one or three moves: this is analysed in Table 6.4. Move 4 appears alone 1235 times, and a group of three moves (444) appears 175 times, but no other pairings of the move appears, which suggests that this move (4) and its triple repetition (444) are defining elements of the assembly mechanism that permits completion of the capsid. Interestingly, moves 9 and 5 always appear in the same paths together, implying that they are required together. In the pairing, move 9 always appears first in the sequence, which occurs just 75 times.

6.5.4 Choosing a subset

Here we seek to choose a subset of the resulting paths for further analysis, by looking for similarities in the paths for possible shared mechanisms.

We cluster the paths using a pairwise comparison of Sørensen-Dice coefficient

	x4444x	x444x	x44x	x4x	Cumulative
4444	0				0
444	0	175			175
44	0	350	0		350
4	0	525	0	1235	1760

Table 6.4: Non-cumulative analysis of repeated occurrence of Move 4, which was noticed to occur repeatedly together. In total there were 1760 instances of 4 moves in the result paths, as part of larger subpaths x4x, x44x, x444x, or x4444x, where x represents a move other than 4. The combination 44444 cannot occur as this would result in an incomplete path: the fifth consecutive move around a five-fold axis would return to the starting position. There are no occurrences of x4444x or x44x within the result paths: these move combinations have been calculated to not result in complete encapsidation. There are however 175 incidences of x444x and 1235 occurrences of x4x.

scores, s, calculated using a gram (substring) size of 2:

$$s = \frac{2n_t}{n_x + n_y} \tag{6.5.1}$$

where n_t is the number of grams that are present in both the strings, and n_x and n_y are the number of grams in each of a pair of two strings (x and y). In our case, the strings either correspond to the spatial order by which proteins are recruited in the $5'$ to $3'$ direction along the RNA, or the order of generalized moves performed by the RNA. The Sørensen-Dice coefficient is particularly appropriate for scoring the string similarity as we are concerned with measuring the conservation between strings of small fragments, but the order is of less importance, because the principal interest is of conserved features shared by different assembly scenarios (*i.e.* small sequences of generalized moves) or conservation in connectivity between PS positions (*i.e.* the same pairs of protein positions occurring, indicating a possibility of conserved density in any imaging-based comparison of the RNA conformation of these paths).

Dendrograms were produced using the dendrogram function from the python scipy.cluster.hierarchy module, utilizing the Euclidean distance metric from the

linkage function within that module. The dendrograms produced are shown in Figure 6.9 and Figure 6.10: Figure 6.9 groups paths by similar groups of moves generalized to start and finish at arbitrary proteins, whereas Figure 6.10 groups paths by the proteins connected in a capsid, minimized to the lowest possible score of all orientations for each pair of paths. This is necessary because, unlike the generalized moves, these actual moves are indicated with reference to specific positions on the capsid, so the orientation of the paths makes a difference to the proteins connected by RNA, and thus the clustering algorithm. This is particularly useful for the remainder of this chapter, when seeking to analyse asymmetric structural data.

The same large cluster of Groups B, C, G, and H appears in each dendrogram, containing 48 paths.

These related groups share features of their assembly scenarios, revealing conserved features characterizing the group, and we expect that not a single scenario but a group of assembly pathways sharing important features occur in the virus. This group, with dendrogram branches in green, is an example of this, where all of these paths have initiation at the second PS position from the $5'$ end.

6.5.5 Variation analysis

The clustering of paths into groups creates the opportunity for a detailed comparison of the similar paths. A simple alignment of grouped paths can unlock extra information about the paths (for an example, see Table 6.5). From the alignment it was possible to identify regions where the paths differ within the aligned paths, which were composed of moves 1, 4, and 8. Each variable region of the path represents a possible parallel mechanism for the addition of a few proteins to the capsid. However, we are not so interested in the detail of each

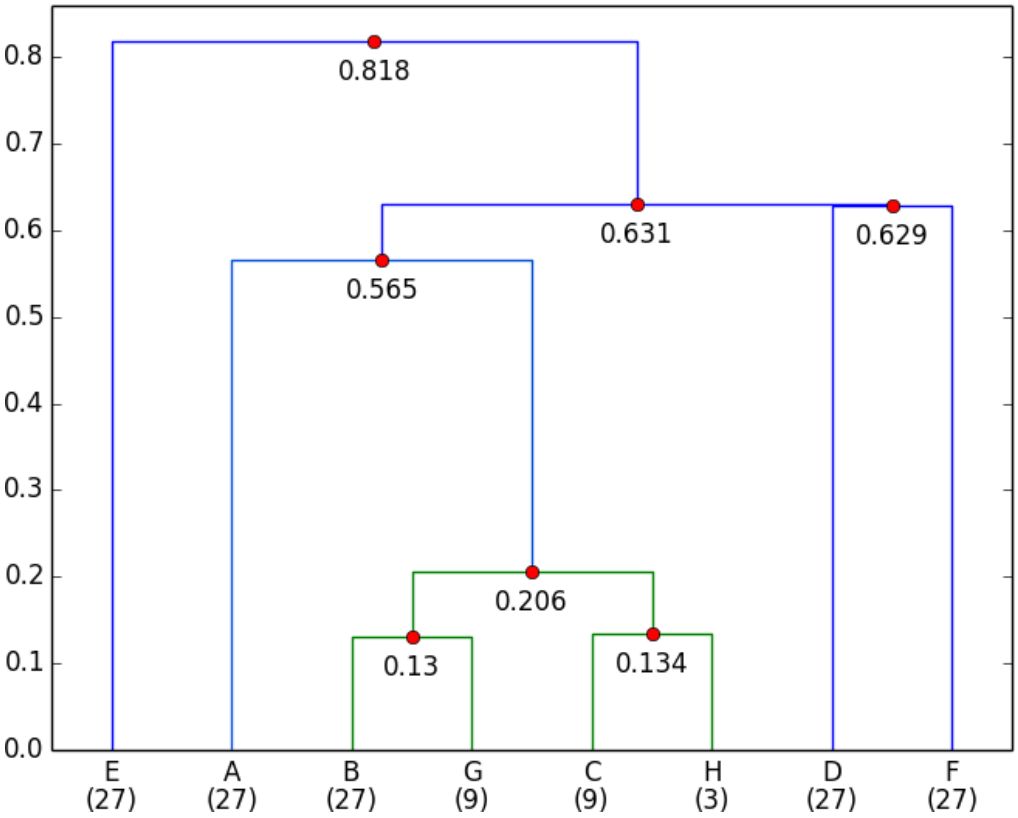


Figure 6.9: Clustering of paths, by move order, *i.e.* by the sequential ordering of the generalized moves 1-9. Paths with similar successions of moves are identified, and broken down into 8 groups (marked A-H along the x-axis); each group is labelled by the number of paths contained within the group. The clustering algorithm utilizes the Sørensen-Dice coefficient with a gram size of 2, performing a pairwise comparison of all paths and using a Euclidean distance metric; this forms the basis of clustering, and is displayed on the y-axis. As the grouped paths have similar successions of moves, they could be viewed as sharing common mechanisms or common sub-conformation structures.

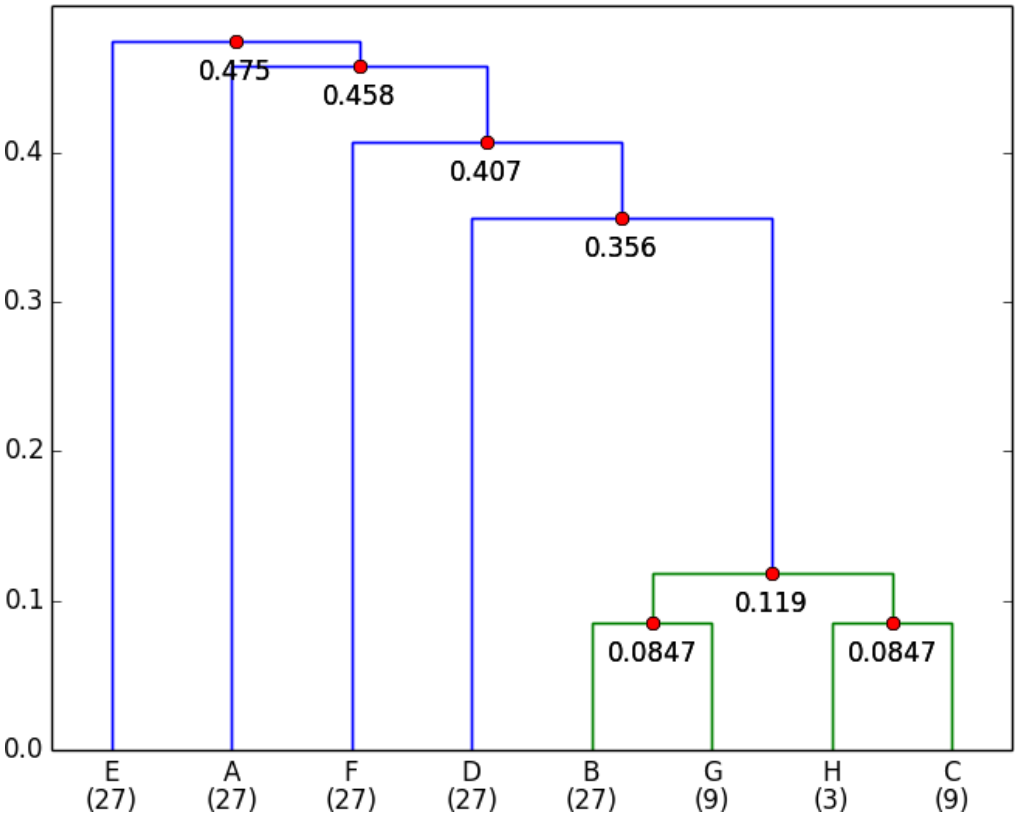


Figure 6.10: Clustering of paths, by order of CP addition. Paths with similar real-space RNA connections are identified, and broken down into the same 8 groups as in Figure 6.9 (marked A-H along the x-axis); each group is labelled by the number of paths contained within the group. As before, the clustering algorithm utilizes the Sørensen-Dice coefficient with a gram size of 2, performing a pairwise comparison of all paths and using a Euclidean distance metric (displayed on the y-axis). As the grouped paths have many of the same proteins being linked by RNA, the grouped paths are likely to appear very similar across most of the RNA outer shell in low- or mid-resolution asymmetric reconstructions.

path but on the overall characteristic geometry of a likely group of paths that could describe the formation of PS interactions within the assembling capsid. Therefore all moves of type 1, 4, and 8 were removed from the sequences, in order to generate a consensus sequence of moves (Table 6.6).

Because we have now discarded some of the moves to form consensus paths, we cannot use the generalized moves to index our position on the capsid. Therefore we have to use the realized RNA binding order, by converting the consensus sequence of moves into a consensus sequence of RNA-CP binding events (Table 6.7, see Appendix A for full set). These consensus-based organizations do not describe the whole path organization or order of protein addition, but rather their common, conserved features. However, the full paths can be subsequently determined: examples from each group are given in Table 6.8 and Table 6.9.

Note that in a comparison of two paths, the best structural alignment of RNA segments in the virion is not implied by the first PS occurring at protein `a`; neither does the nucleation point being in the same point necessarily lead to the greatest conservation of segments with respect to the orientation of the virion. All possible orientations need to be compared between groups. Testing every possible orientation of pairs of sequences, the greatest number of shared edges possible is shown in Table 6.12.

This analysis does not discriminate between the sense of RNA predicted to connect proteins (*i.e.* $5'$ to $3'$ or $3'$ to $5'$): in the model this is equivalent to the connections `AB` and `BA` being viewed as identical RNA segments between proteins `A` and `B`. However, no other connections are viewed as equivalent in the subsequent analysis.

We therefore have to orientate each consensus sequence so that its realization within the capsid leads to the greatest conservation of RNA tertiary structure within the virion, as measured by the maximal number of shared edges in Ta-

077877	44418	77	8414148	77	44414	641481414
077877	44418	77	8414148	77	4814	641481414
077877	44418	77	8414148	77	8414	641481414
077877	44418	77	4814148	77	44414	641481414
077877	44418	77	4814148	77	4814	641481414
077877	44418	77	4814148	77	8414	641481414
077877	44418	77	44414148	77	44414	641481414
077877	44418	77	44414148	77	4814	641481414
077877	44418	77	44414148	77	8414	641481414
077877	4818	77	8414148	77	44414	641481414
077877	4818	77	8414148	77	4814	641481414
077877	4818	77	8414148	77	8414	641481414
077877	4818	77	4814148	77	44414	641481414
077877	4818	77	4814148	77	4814	641481414
077877	4818	77	4814148	77	8414	641481414
077877	4818	77	44414148	77	44414	641481414
077877	4818	77	44414148	77	4814	641481414
077877	4818	77	44414148	77	8414	641481414
077877	8418	77	8414148	77	44414	641481414
077877	8418	77	8414148	77	4814	641481414
077877	8418	77	8414148	77	8414	641481414
077877	8418	77	4814148	77	44414	641481414
077877	8418	77	4814148	77	4814	641481414
077877	8418	77	4814148	77	8414	641481414
077877	8418	77	44414148	77	44414	641481414
077877	8418	77	44414148	77	4814	641481414
077877	8418	77	44414148	77	8414	641481414
077877	[44418] [4818] [8418]	77	[44414148] [4814148] [8414148]	77	[44414] [4814] [8414]	641481414

Table 6.5: Illustration of the variation within Group E: These 27 sequences from Group E have shared features and variable regions, composed of moves 1, 4, and 8. Interestingly, there are specifically three different options for each variable region; each can be thought of as representing a possible parallel mechanism for the addition of a few proteins to the capsid, until a common scheme resumes. There are three variable regions within the move sequence, giving an upper bound of $3^3 = 27$ possible paths: all of the possible degenerate options have thus been sampled. This implies that the region is not constrained, allowing for any possible order of protein recruitment in this window before specific constraints kick in again. The resulting 27 assembly scenarios form one of the groups in the dendrogram clustering analysis (Figure 6.9 and Figure 6.10), and form Group E (Table 6.6). Note that Group E does not form part of the subset of similar groups which undergo further analysis, but serves here as an illustrative example.

Group	Remove 148	Consensus moves
A	X066X69X77X7X6X6X657X7	444066X14691X177178146464657X17
B	X0966X77X7X7X5X7X	40966X1877X18874148187X5481874
C	X0966X7X72X7X65X7X	40966X1417X17214148187865481874
D	X077X77X77X6X77X6X	4077877X1877X614877X14641481414
E	077X77X77X77X6X	077877X1877X1414877X14641481414
F	X70X77X77X77X6X	170877X1877X1414877X14641481414
G	X096X7X77X7X7X5X6X	409617877X18874148187X548181464
H	X096X7X7X72X7X65X6X	409617417X172141481878654818146 4

Table 6.6: Moves 1, 4, and 8 removed from sequences (replaced here by X) groups the 156 sequences into the same eight groups identified by clustering (Figure 6.9 and Figure 6.10).

Group	Consensus moves	Consensus RNA binding order
A	444066X14691X177178146464657X17	aduNLs..IJGoFE..ZY3RQSxwyXU0BClm..feh
B	40966X1877X18874148187X5481874	adNLs..IJ5HE..ZYT2ROPSxwgz..0ACijqnl
C	40966X1417X17214148187865481874	adNLs..IJ5HE..ZY2QROPSxwgzU0ACijqnl
D	4077877X1877X614877X14641481414	advwgzA..ijqnp..sNMLR72..TSWV01ZEFHG5
E	077877X1877X1414877X14641481414	acfmil..opKJ6..MNuvPWTV..zygACD14532Y
F	170877X1877X1414877X14641481414	abtuMP..xwgzA..ijmnlHE5..JIK6RQ2UV01Z
G	409617877X18874148187X548181464	adNLKJ5HE..ZYT2ROPSxwgz..0ACijqrpnl
H	409617417X172141481878654818146 4	adNLKJ5HE..ZY2QROPSxwgzU0ACijqrpnl

Table 6.7: Consensus moves and RNA binding order to protein. From the groups found using dendrogram clustering (Figure 6.9 and Figure 6.10) and confirmed by detailed analysis of the predicted sequence of spatially generalized RNA-protein interaction mechanisms, consensus moves that reference the organization of proteins present within the capsid have been formed. These consensus sequences are divided into *Consensus moves*, which track the interaction of RNA PS with CP, and *Consensus RNA binding order*, which show the order that proteins are recruited and attach to the capsid. The entire aligned sequences for these groups are given in Appendix A.

Group	Nucleation	Example RNA binding order
A	M	R6LMuPSWxwyXBACijeabrIJGonlFD0VT234
B	u	dutrbfjmnFkD1457GIKL6OPWTQ3YU0zyehA
C	0	MO6JKsrponmihCDE4532YTSxX0Bygcdtbf
D	b	fbrtuMPWxwgzAkimnlHE536JIK7RQ2VU01Z
E	h	hjewdvSWX0VZ4EDCklFoGJsrqbtNMLRQ237
F	z	UV0zAgcwxPuM6RQ237GIolFEDkijmabtsr
G	d	adNLKJ5HEDZYT2ROPSxwgzXV0ACijqrpnl
H	q	bqjhgzXWTQYZ1DCklFopKJ536RMNdcwvP

Table 6.8: Example order of protein binding by group: this is a spatial rather than temporal order, representing the locations on the genome to which the protein binds from 5′ (left) to 3′ (right), denoted by the position of nucleation which determines the overall spatial organization.

Group	Nucleation	Example protein addition
A	M	R6LMtKONuvPSWxcwyzX1BhACkifjgedaqbsrpI7JGHonmlFEDU0YVTQ253Z4
B	u	duxcNatsrqbfjmonlFCkED1Z4H527JGpIKML6ROvPSWVTQ3YXU0BzwygeihA
C	0	MOvNRL67JIKsrpHonjmkihACFDlEZ4G53Q2YVTPSWxzXU01Bwygecudatqbf
D	b	fbdeqsratNuOMvPSWxcwygBzhACkijmonlFH4EG536JpIKL7RTQ2YVXU0D1Z
E	h	hyBifjgecwudxvPSWzXU0YV1Z4EDACknlFHoIGpJKsrmqbatNOML6RTQ2537
F	z	U01ZVXSYBzhAygecwWxvPNuOML6RTQ2537JGpIHonlF4EDCkifjmqadbtKsr
G	d	adebutNMLIK7JG5FH4ED1Z3YUTQ26ROvPSWxcwygBzXV0hACkifjmqsrponl
H	q	bqsamfjihygBzXSWVTQ2Y0ZU1EDACknlFHorpIK7JG543L6ROMtNudecwxvP

Table 6.9: Example order of protein assembly by group: the order of addition of protein to the assembling capsid, corresponding to the example RNA-protein binding order provided for each group in Table 6.8.

ble 6.12. The orientations providing the greatest conservation between organizations in the subset groups (Groups B, C, G, H) is calculated to be unchanging: *i.e.* the maximum shared edges occur when the organizations start at the same nucleation protein, and are therefore not rotated with respect to each other (Table 6.13). Note that trying to combine these four favoured groups further in the analysis does not lead to favourable results, as although the number of shared edges is high, the paths still are too different to cluster together. At this point, it is possible to combine with experiment to see which group is the most likely to occur.

We have also compared the optimum orientation of the four other groups to the orientation best shared by the subset groups (Table 6.14). There is much more disorder amongst all of these groups together; for some pairs of groups, there is more than one orientation that results in the same number of shared edges. Based on Table 6.12, groups B, C, G, and H have the largest level of coherence, with the highest level of conservation of the corresponding paths. We will persist with an analysis of the subset groups alone, as they share many of the same moves. Frequency of the moves utilized by paths in this cluster are shown in Table 6.10, along with a comparison to the move frequency of the paths as a whole, normalized to the number of paths.

6.6 Crystallography

While the inherently asymmetric genome within a single virus particle cannot strictly be icosahedrally symmetric, the crystal lattice will impose icosahedral averaging upon the density in the X-ray Fourier maps. As discussed above, the prevailing view is that even if there were a unique conformation of the RNA present in all virions, the nucleic acid density would exhibit static disorder in the crystal structure analysis due to the stochastic nature of crystallization. Thus

№	Frequency	Comparison
0	48	100%
1	276	89.25%
2	12	325.00%
4	464	85.68%
5	48	208.00%
6	108	92.86%
7	228	75.30%
8	392	132.85%
9	48	208.00%

Table 6.10: Frequency of moves across the subset of solutions. Comparison with respect to the full result move frequency, normalized at 100% to the change in move 0, *i.e.* nucleation and thus the number of solutions.

	x4444x	x444x	x44x	x4x	Cumulative
4444	0				0
444	0	40			40
44	0	80	0		80
4	0	120	0	344	464
Comparison		74.3%		90.5%	

Table 6.11: Non-cumulative analysis of repeated occurrence of Move 4, which was noticed to occur repeatedly in consecutive moves. Incidences of x4x, x44x, x444x, and x4444x are noted (and combination x44444x cannot occur). The subset is found to have 40 incidences of x444x and 344 occurrences of x4x, where x represents a move other than 4. The comparison row shows that the subset has 74.3% of the x444x subpaths and 90.5% of the solitary x4x moves expected. These values have been derived by normalizing to the expected number from the wider population: the subset contains 30.8% of the paths, but only 22.9% of the x444x subpaths and 27.9% of the solitary x4x moves compared to their incidences in the full list of path solutions.

Maximized shared edges								
A	B	C	D	E	F	G	H	
27	8	10	9	9	9	12	10	A
	26	23	9	9	9	23	20	B
		28	10	10	10	20	25	C
			27	22	19	9	9	D
				27	22	11	10	E
					27	10	10	F
						28	25	G
							30	H

Table 6.12: Alignment comparison: for each pair of consensus paths from each group, the calculated number of shared connections of RNA between proteins in the structure, at different orientations of the predicted RNA organizations within the virion. Interestingly the Groups B, C, G, and H have particularly high comparison of shared edges.

Shared edges					Best orientation				
B	C	G	H		B	C	G	H	
26	23	23	20	B	a	a	a	a	B
	28	20	25	C		a	a	a	C
		28	25	G			a	a	G
			30	H				a	H

Table 6.13: Subset of 48 3-D alignment comparison. All initiate assembly at 2nd PS position from the 5′ end.

Best orientation								
A	B	C	D	E	F	G	H	
a	dU	d	5	opvY	Z	N	N	A
	a	a	ms	CM	y	a	a	B
		a	s	bCM	R	a	a	C
			a	t	f	M	M	D
				a	b	6	6	E
					a	P7	P	F
						a	a	G
							a	H

Table 6.14: Best alignment of all groups to each other. The values in each cell represent the offset starts between the groups, that upon superposition of the two structures will result in the greatest number of shared edges. The greatest number of shared edges is the value displayed in Table 6.12. For the subset groups (analysed in Table 6.13), and clearly also the principal diagonal which represents a self-comparison, the maximum shared edges occurs for no offset, at position `a`.

this information is at first glance insufficient to deduce whether there are multiple conformations of the genome present within the crystallized virion population.

However, recent evidence has suggested that the crystallization of STNV is effected by the conformation of RNA within the capsid, as described in Figure 6.11 (Prof. Arwen Pearson, personal communication). Specifically, it is thought that the overwhelmingly negative charge of the RNA is not fully shielded by the positively-charged protein capsid. Thus electrostatics could have an impact on the overall charge distribution, by biasing the orientation of virions in the crystal lattice, and thus leading to preferred stacking in the crystal.

In an asymmetric X-ray structure[4] of STNV VLPs containing the B3 aptamer [34, 49], there is not a full density for the RNA, suggesting that the unit cell for the stacking is small (Prof. Arwen Pearson, personal communication). If there was an unclear stacking unit, this would be lost in the Fourier maps. Indeed, from experiment a unit cell size of two is expected; *i.e.* the virions stack in pairs in a defined orientation, thought to be 120° (Prof. Arwen Pearson, personal

[4]This is material, as yet unpublished, from a collaborator. Data not shown.

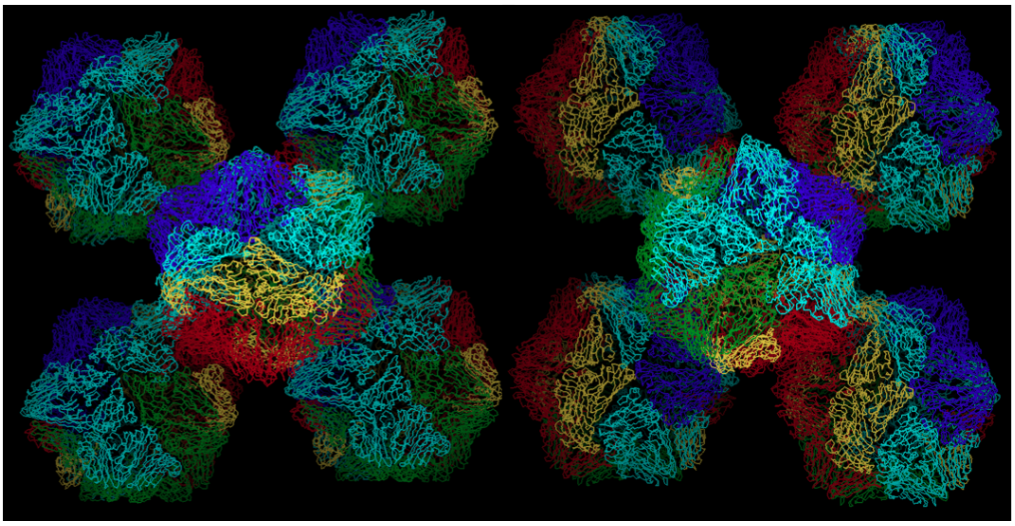


Figure 6.11: Visualization of asymmetric stacking. If the STNV capsid's positive charge does not effectively shield the negatively charged RNA, it is possible that external charge could occur (exemplified here by different coloured capsid components). If this followed specific patterns, *e.g.* if RNA organization in proximity to capsid shows specific patterns, as discussed in the previous section, then it is possible that the orientation of the crystalline virions would not be random, and some sort of preferred stacking could result.

communication). This opens an opportunity for analysing the asymmetric density attributed to RNA inside the capsid, based on the knowledge of the stacking of particles. In particular, as expected the RNA density at PS positions does conform to the icosahedral symmetry conferred by the capsid, and thus a shared motif is expected at these positions in the asymmetric organization. We develop here ways of using knowledge of the particle stacking to refine our analysis in the previous sections.

In order to index lattice positions, addition to the 3-D orthonormal basis $\mathbf{e} = (e_1, e_2, e_3)$, a symmetry-adapted icosahedral basis $\mathbf{a} = (a_1, a_2, a_3, a_4, a_5, a_6)$ can be defined; the six vectors point to six vertices of an icosahedron that are

not aligned to each other [122,249]. The icosahedral basis vectors are as follows:

$$\begin{aligned}
\mathbf{e_1} &= (1,0,\tau) & \mathbf{a_1} &= (1,0,0,0,0,0)\\
\mathbf{e_2} &= (\tau,1,0) & \mathbf{a_2} &= (0,1,0,0,0,0)\\
\mathbf{e_3} &= (0,\tau,1) & \mathbf{a_3} &= (0,0,1,0,0,0)\\
\mathbf{e_4} &= (-1,0,\tau) & \mathbf{a_4} &= (0,0,0,1,0,0)\\
\mathbf{e_5} &= (0,-\tau,1) & \mathbf{a_5} &= (0,0,0,0,1,0)\\
\mathbf{e_6} &= (\tau,-1,0) & \mathbf{a_6} &= (0,0,0,0,0,1)
\end{aligned} \tag{6.6.1}$$

where τ is the golden ratio:

$$\tau = \frac{1+\sqrt{5}}{2} \tag{6.6.2}$$

Any vector, r, in 3-D can be represented in terms of the icosahedral basis given in Equation (6.6.1);

$$\mathbf{r} = (x,y,z) \mapsto (n_1,n_2,n_3,n_4,n_5,n_6) \tag{6.6.3}$$

Visualization in 3-D of course is best performed in the orthonormal basis in 3-D, **e**. However, the crystallographic nature of the icosahedral group becomes clear in the icosahedral basis, **a**, as icosahedral symmetry is crystallographic in 6-D.

Rotations that preserve the group symmetry are given by the following op-

eration matrices, in both the orthonormal and icosahedral bases:

$$R_3^e = \frac{1}{2}\begin{bmatrix} \tau & 1-\tau & 1 \\ \tau-1 & -1 & -\tau \\ 1 & \tau & 1-\tau \end{bmatrix} \tag{6.6.4}$$

$$R_5^e = \frac{1}{2}\begin{bmatrix} 1 & -\tau & \tau-1 \\ \tau & \tau-1 & -1 \\ \tau-1 & 1 & \tau \end{bmatrix} \tag{6.6.5}$$

$$R_3^a = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix} \tag{6.6.6}$$

$$R_5^a = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix} \tag{6.6.7}$$

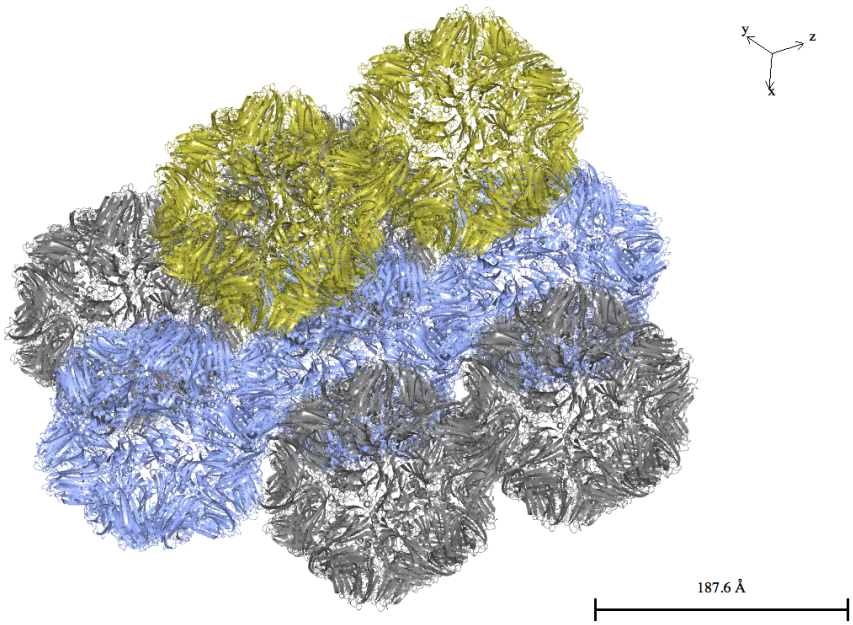


Figure 6.12: Crystal stacking of STNV virions.

The projection operators that are used for the cut-and-project method are:

$$P_{\text{parallel}} = \frac{1}{\sqrt{2(2+\tau)}} \begin{bmatrix} \tau & 0 & -1 & 0 & \tau & 1 \\ 1 & \tau & 0 & -\tau & -1 & 0 \\ 0 & 1 & \tau & 1 & 0 & \tau \end{bmatrix} \tag{6.6.8}$$

$$P_{\text{perpendicular}} = \frac{1}{\sqrt{2(2+\tau)}} \begin{bmatrix} \tau & -1 & 0 & 1 & -\tau & 0 \\ 0 & \tau & -1 & \tau & 0 & -1 \\ 1 & 0 & \tau & 0 & 1 & -\tau \end{bmatrix} \tag{6.6.9}$$

The STNV crystal (Figure 6.12) has the lattice parameters given in Table 6.15, and the space group of the crystal is that of crystallographic space group number 5 (which can be described by the notations in Table 6.16).

Parameter	Value
a	315.02Å
b	300.55Å
c	183.51Å
α	90°
β	94.37°
γ	90°

Table 6.15: Lattice parameters of STNV asymmetric crystal, which is monoclinic and base centred on the C face.

Notation	Value
Crystallographic group	5
Hermann-Maugin	$C2 : b1 = C121$
Schoenflies	C_2^3
Hall	$C2y$

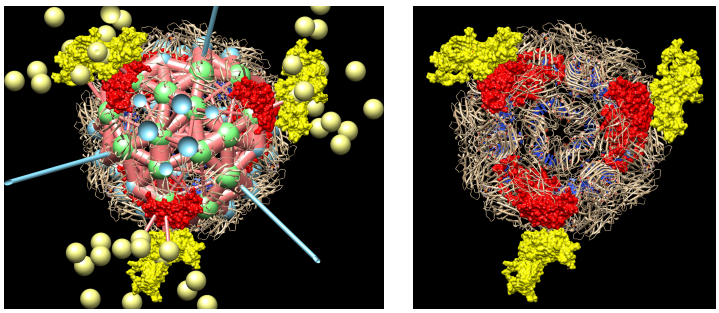
Table 6.16: Space group notation of the STNV crystal lattice.

6.7 Kissing points

Kissing points, *i.e.* the points of contact between different particles within the stack [123,250], were identified using UCSF Chimera Crystal Contacts tool [251] (see Figure 6.13). The tool was written to display steric clashes between different asymmetric crystal structures, in order to validate crystallography data, usually saved in the file header during crystallographic model refinement. It serves as a useful tool here to show the closest points (kissing points) in the crystal of STNV.

The location of the identified kissing points on an icosahedral STNV capsid map are shown in Figure 6.14. There are eight interactions with surrounding capsids; six of these interactions are largely between single proteins in each capsid (denoted here as Green-Blue kissing points), but two, at opposing 3-fold axes, are interactions between groups of three proteins (Yellow-Red kissing points).

We consider the RNA positions calculated above that are in proximity to the



(a) (b)

Figure 6.13: Identification of STNV kissing points. (a) The Crystal Contacts tool of UCSF Chimera [251] shows the location on crystalline capsid proteins that intersect at different distances of interpolation between the capsids. The blue spheres are in close proximity to the three central green spheres. These are identified as Green-Blue kissing points that occur within the STNV unit cell, which consists of two virions. (b) Additionally displayed in red and yellow are the proteins that interact with other adjacent unit cells: we denote these interactions Red-Yellow kissing points.

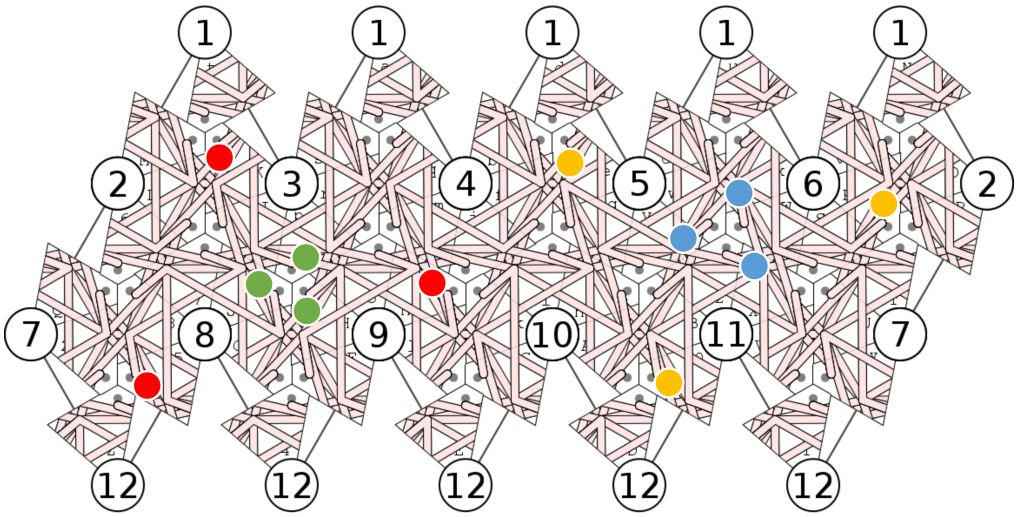
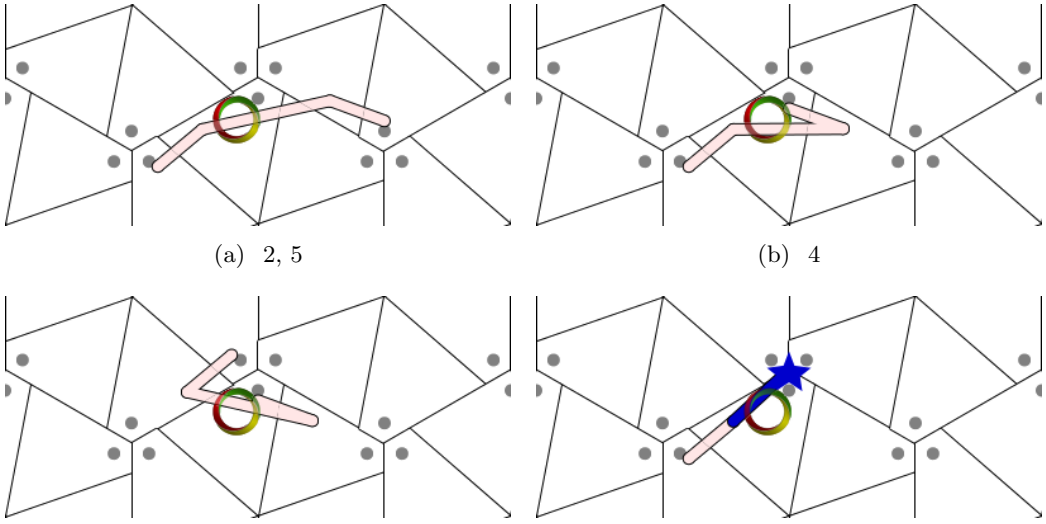


Figure 6.14: Location of STNV kissing points. A total of eight interactions were determined (using UCSF Chimera Crystal Contacts tool, for which the process is illustrated in Figure 6.13): six of these interactions (orange/yellow) are largely between single proteins in each capsid, but two (blue/green), at opposing 3-fold axes, are interactions between groups of three proteins.

kissing points identified. The RNA connections between proteins in proximity to the kissing points are interrogated to check for overlapping with the kissing point region. This is to check whether the negative charge of the RNA molecule could influence the packing of the virions within the unit cell. The RNA connections (in terms of moves) between proteins adjacent to the Yellow, Red and Green (type I) kissing points that have been identified as influencing the kissing point electrostatics are shown in Figure 6.15. These kissing points are all at the same geometric symmetry position on a capsid protein, so share the RNA moves that could change the region's overall electric charge; these are moves 2, 4, 5, 6, 7, and 0 (nucleation) in specific orientations and between specific proteins.

The RNA connections that influence the charge density at Blue (type II) kissing points are shown in Figure 6.15. The Blue kissing point lies where a stem-loop PS would be positioned, interacting with an N-terminal arm on an adjacent protein. Therefore, any PS positioned at the kissing point would effect charge (Figure 6.16(a)). Alternatively, moves 2 and 5 between neighbouring proteins have also been identified as overlapping with the Blue kissing point position (Figure 6.16(b)).

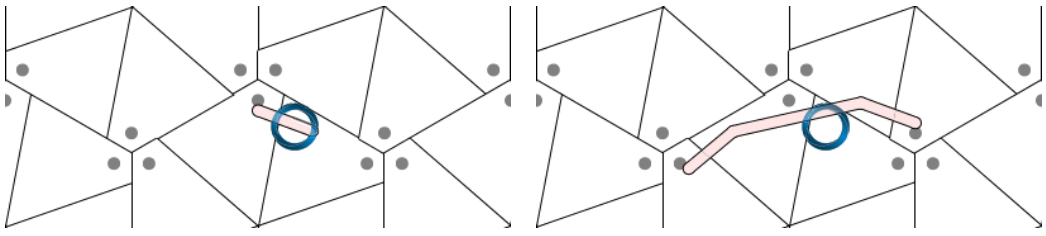
The moves that influence the electric charge at kissing points Red, Yellow, Blue and Green are compared to the consensus move sequences for the subset groups identified in §6.5.4. However, the problem with the consensus sequences are that there are gaps at the variable regions. We cannot have gaps, as that would infer that there was no RNA, and hence positive charge from bare protein. Instead, we average the move sequences for all paths within a group, to yield an RNA occupation (*i.e.* negative charge) score, c, between 0.0 and 1.0. Upon calculation, due to the three different options for each variable region of moves 1, 4, and 8 (displayed in Table 6.5), values for occupation were found to be



(a) 2, 5 (b) 4

(c) 6, 7 (d) 0

Figure 6.15: Moves occurring adjacent to kissing point type I (Yellow, Red, Green): (a) moves 2 and 5, (b) move 4, (c) 6 and 7, (d) move 0 (nucleation).



(a) 0, 1, 2, 4, 5, 6, 7, 8, 9 (b) 2, 5

Figure 6.15: Moves occurring adjacent to kissing point type II (Blue): (a) any moves start or stop at the stem-loop position indicated as binding to the N-terminal arm of capsid protein (small circle), or (b) moves 2 or 5 in proximity to this location.

multiples of one third, being drawn only from the following amounts:

$$c \in \{0, \frac{1}{3}, \frac{2}{3}, 1\} \tag{6.7.1}$$

Calculation of electrostatic interactions between Yellow-Red pairs of kissing points that meet in the lattice was performed by comparison of the occupation score, using the following equation:

$$E_{YR} = |c_Y - c_R|, \tag{6.7.2}$$

and the score S_{YR} for a unit cell:

$$S_{YR} = \sum_{\text{all } YR} E_{YR}. \tag{6.7.3}$$

Note that the sign of E_{YR} is not of interest, solely that there is a difference between the two kissing points. For a single unit cell, the maximum score, indicating good stacking, of S_{YR} is 3.0, as there are three kissing points. For a dual unit cell, the maximum score of S_{YR} is 6.0; for an even larger unit cell the maximum score would be proportionally higher. Also note that non-integer scores close to the maximum value mean that the ideal conditions are reached in some of the paths in that group: *i.e.* from a S_{YR} score of $5\frac{1}{3}$ in a dual unit cell it can be deduced that a third of the paths in the group meet ideal conditions. Conversely, either another third of paths experience electrostatic repulsions at two kissing points, or two-thirds of the paths experience an electrostatic repulsion at a single kissing point.

Calculation of S_{GB} scores, indicating the interaction between groups of paths at the Green-Blue kissing points (determined using the UCSF Chimera routine), were calculated in a similar manner. However, the crystal packing of STNV

means that these kissing points are very close together. As these electrostatics operate in concerto between the same particles, we consider the electrostatics of these positions in a combined manner: *i.e.* we calculate the overall charge for Blue and compare to the overall charge for Green, using the equation:

$$S_{GB} = E_{GB} = |(c_{G1} + c_{G2} + c_{G3}) - (c_{B1} + c_{B2} + c_{B3})|. \tag{6.7.4}$$

The maximum score possible for S_{GB} is always 3.0 for unit cells that do not include a Green-Blue interaction within the cell, such as those that we are studying here.

Analysis of the individual paths making up each group was not performed, due to the unlikelihood that a single path would be dominant in the ensemble of many paths (*i.e.* of the same group) sharing similar sequences of moves. This is principally due to the number of different proteins available to bind to for a PS stem-loop on a nascent genome being packaged, which creates a large computational search space of possibilities; the path is a single route through a connected graph of eight possible moves. This is in contrast to the example of MS2 (Chapter 7) where there are only three moves evidenced in the symmetric structure [10,12]: a complete path is a single route through a triconnected graph, and therefore a maximum of two moves are possible at any time (as one of the moves would have been used previously). Thus, resulting paths in MS2 share fewer similarities than in STNV, as there are no degenerate parallel options available (such as those in Table 6.5).

6.8 Embedding of asymmetrically charged units into a crystal lattice

6.8.1 Single unit cell

First we consider a *single unit cell*: *i.e.* the same orientation of the virus throughout the crystal. In this case, the same Red and Yellow kissing points will meet at every gap between virions in the crystal lattice (see Table 6.17 and Figure 6.16 for a description of these kissing point interactions). The Yellow 0 and Red 0 kissing points always form a pair of adjacent parts of the capsid; similarly Yellow 1 / Red 1 and Yellow 2 / Red 2 remain as pairs. Additionally we consider that the Blue and Green kissing points form a stable, unchanging composite structure due to the unchanging orientation of virions in the crystal.

The electrostatics resulting from the organization of RNA within the virions inside the single cell of the crystalline lattice are shown in Table 6.18. Only RNA organizations in Group G of the subset sampled form strong Red-Yellow kissing point interactions at all three positions, with strong interactions ($S_{YR} = 3.0$) occurring when the paths are orientated starting at proteins `h`, `N` and `5`. The score S_{YR} represents a maximum difference in charge at each of the three kissing points occurring in the single unit cell (*i.e.* 3×1.0). However, the score $S_{GB} = 0.333$ indicates a very weak (possibly repulsive) interaction between the Green and Blue kissing points. These electrostatics between Green and Blue are disappointingly poor. However, the lattice is further spaced in this direction, at about a minimum of 36Å, so it is possible for a single unit cell to dominate a crystal structure if this greater distance makes the stacking position viable. In most of the orientations that maximize the Yellow-Red electrostatic interactions, the electrostatics between Green and Blue kissing points are dominated by protein and non-existent RNA rather than pairing strongly negatively charged

Yellow	Red
0	0
1	1
2	2

Table 6.17: Yellow-Red kissing points for single cell stacking, predicted by icosahedral tiling of STNV virions.

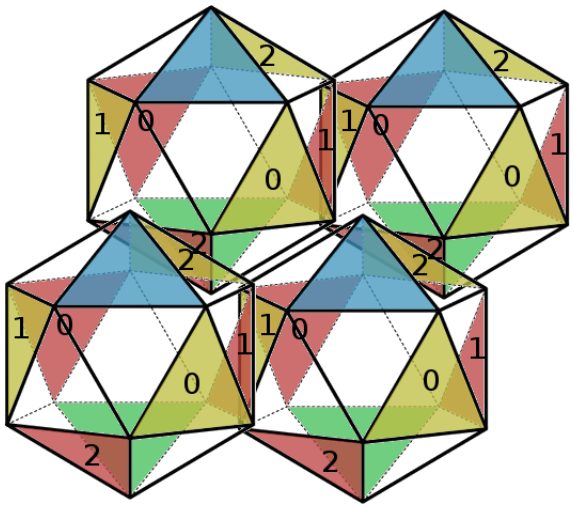


Figure 6.16: Single cell stacking: an example interaction between kissing points in the STNV single unit cell.

Group	Orientation	S_{GB}	S_{YR}
B	-		
C	-		
G	h	0.333	3.0
	u	0.333	3.0
	V	0.333	3.0
H	-		

Table 6.18: Resulting orientations for single cell stacking that maximize electrostatic interaction between virions ($S_{YR} = 3.0$). Only Group G RNA organizations allow electrostatically optimized stacking of virions in a single-virion unit cell paradigm.

RNA; *i.e.*, the paths do not visit positions that would account for the strong electrostatic interactions expected at this positions.

In this single unit cell regime, Group G is the only cluster that can account for these strong Yellow-Red interactions (Table 6.18). Although the S_{YR} is only maximized in Group G, other groups also approach this level of interaction, but there are inevitably electrostatic clashes, no matter what the orientation of the organized RNA is within the single unit cell virion. A super group encompassing a combined superposition of all subset clusters (Groups B, C, G, H) that ignores flexible segments also does not produce favourable electrostatic interactions, suggesting that if a single unit cell is prevalent within the crystal, then the organization of the RNAs within the crystal will be G-like rather than a mixture of different organizations from different groups (as this would lead to increased electrostatic clashes).

6.8.2 Dual unit cell

If we consider a *dual unit cell* instead of a single unit cell, then we envisage two virions making up a combined repeated unit within the crystal (Figure 6.17). If the Green-Blue interactions are to be preserved (and these kissing points only occur at one place, respectively, on the capsid) then there can only be rotations about the 3-fold Green-Blue axis between the pair of virions: an offset of 120°.

We need to test both rotation offsets of 120° and 240° as the paths are asymmetric, and the Green-Blue rotation axis does not have mirror symmetry due to the Green and Blue kissing points being different (and being composed of different moves: see Figures 6.15 and 6.15).

There are also two different ways in which the unit cells can be arranged, in terms of the stacking between the unit cells, which manifest itself in terms of vertical stacking (Green-Blue interactions) between the dual virion unit cells.

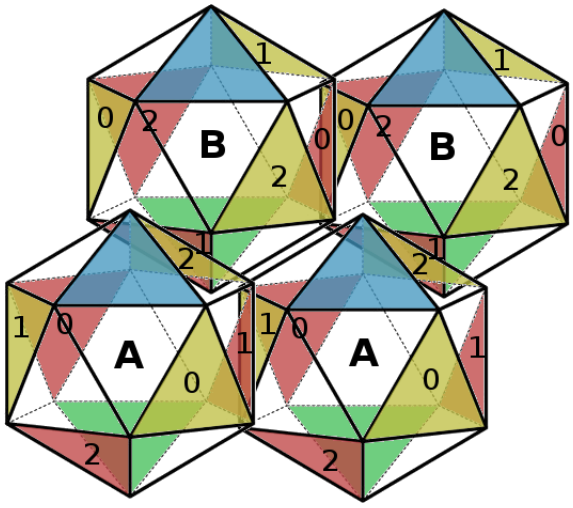
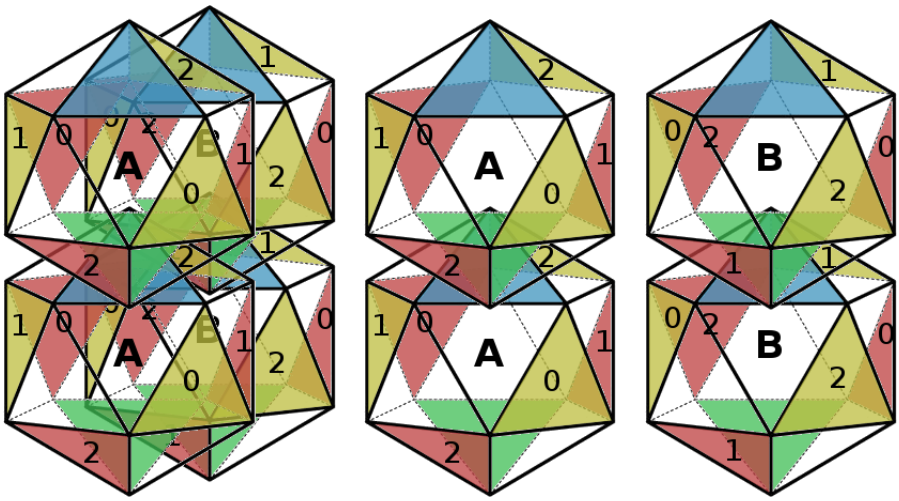


Figure 6.17: Dual stacking. Shown here is a rotation offset of 120° between the two virions in the unit cell, A and B. We proceed to analyse the dual unit cell here on the assumption that the same organization of the RNA is conserved in the particles, excepting for the offset rotation between the virions. Note that there are two ways in which the four virions shown here can proceed to fill the space: in terms of vertical stacking (Green-Blue interactions) A can sit atop A, and B atop B, producing columns of A and B, or alternatively the vertical arrangement can alternate between A and B.

Either virion A can meet the adjacent virion A with a Green-Blue interaction, and virion B meet virion B, producing columns of A and B (Figure 6.18), or alternatively the vertical arrangement can alternate between virions A and B (Figure 6.19). We denote the first case as AA/BB stacking, and the second case as AB stacking. The kissing point interactions associated with the AA/BB and the AB stackings differ, and are listed in Tables 6.19 and 6.21, respectively.

In the first case, where the AA/BB stacking occurs, the best outcomes are for RNA organizations within the predicted organizations of Group C (Figure 6.20). For three different orientations (paths starting at proteins `c`, `B` and `S`) there are fairly high S_{GB} and S_{YR} scores of interaction, at both rotations of 120° and 240° equally. Note that, as discussed previously, the scores of $S_{GB} = 2.333$ means that perfect stacking was achieved for a third of the component paths of Group C. This is similar to the case for $S_{YR} = 5.333$.



(a) (b) (c)

Figure 6.18: AA/BB stacking: (a) stacking for AA/BB in the lattice, (b)-(c) stacking of AA/BB individually outside the lattice to show Green-Blue contacts.

120°		240°	
Yellow	Red	Yellow	Red
A0	B2	A0	B1
A1	A1	A1	A1
A2	B1	A2	B0
B0	B0	B0	A2
B1	A2	B1	A0
B2	A0	B2	B2

Table 6.19: Yellow-Red kissing points for AA/BB stacking, at rotation offsets of 120° and 240°.

Group	Orientation	Rotation	S_{GB}	S_{YR}
B	-			
C	c	120°, 240°	2.333	5.333
	B	120°, 240°	2.333	5.333
	S	120°, 240°	2.333	5.333
G	-			
H	-			

Table 6.20: Resulting orientations for AA/BB stacking that maximize electrostatic interaction between virions ($S_{YR} \geq 5\frac{1}{3}$). Only the Group C consensus paths meet the required electrostatic interaction requirements, for paths starting at proteins c, B, and S. These protein positions are three-fold degenerate on the capsid map, and hence the resultant paths are related *via* three-fold rotations with each other.

For the second case, namely AB stacking, the S_{YR} score is optimized in Group G alone, at orientations of h at 120° and V at 240°. However, the S_{GB} score for these organizations in this lattice are very low. These AB optimized orientations both also occurred in the single cell lattice case (unlike the result u from the AB calculation). This is powerful because, effectively, the organization can either adopt a single or AB-like dual unit cell (with rotation). Essentially this result demonstrates that either of these two organizations could proliferate throughout the crystal in three different rotational geometries (*i.e.* 0°, 120° and 240°), with perfect electrostatic interactions between Yellow and Red kissing points. Additionally, if any orientation is allowed, abandoning the requirement for repeated Green-Blue interactions (with just Red-Yellow interactions remaining), then h, u and V most frequently form favourable interactions with other orientations (Table 6.23). Therefore, they would be perhaps the most proficient at dealing with symmetry mismatches occurring at polycrystalline grain boundary regions. However, as discussed above, these organizations could only dominate a perfect monocrystalline structure if the greater distance between the Green-Blue crystal contacts makes the stacking position viable.

Interestingly, the same orientations from Group C in AA/BB stacking occur

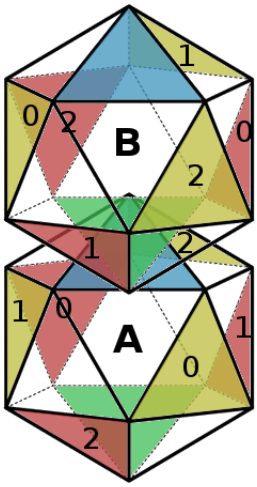


Figure 6.19: AB stacking.

120°		240°	
Yellow	Red	Yellow	Red
A0	A0	A0	A0
A1	A1	A1	A1
A2	B1	A2	B0
B0	B0	B0	A2
B1	A2	B1	B1
B2	B2	B2	B2

Table 6.21: Yellow-Red kissing points for AB stacking, at rotation offsets of 120° and 240°.

additionally in the AB stacking (Table 6.22), with the exact same scores for S_{YR} and S_{GB}. It is therefore likely that if Green-Blue electrostatic interactions cannot be discounted (as expected *a priori*), then organizations from Group C would dominate the crystal. This is tested further by considering any possible unit cells that preserve the Green-Blue axis of interaction, and calculating the minimum S_{YR} score possible for every orientation of the group (Table 6.24). It is found that these realizations (starting at proteins `c`, `B` and `S`) allow the best stacking irrespective of the unit cell if the Green-Blue interaction cannot be ignored.

Group	Orientation	Rotation	S_{GB}	S_{YR}
B	x	120°	0.0	5.333
	y	240°	0.0	5.333
C	c	120°, 240°	2.333	5.333
	B	120°, 240°	2.333	5.333
	S	120°, 240°	2.333	5.333
	x	120°	0.0	5.333
	y	240°	0.0	5.333
G	h	120°	0.333	6.0
	V	240°	0.333	6.0
	c	120°	2.333	5.333
	B	240°	2.333	5.333
	j	120°	1.0	5.333
	Z	240°	1.0	5.333
H	-			

Table 6.22: Resulting orientations for AB stacking that maximize electrostatic interaction between virions. Displayed are orientations for which $S_{YR} \geq 5\frac{1}{3}$.

Group	Orientation	Favourable orientations at YR
B	-	
C	-	
G	h	0 2 4 6 7 H I L M O R V W a g h m u
	u	1 4 D E F G L Q V Z e h k m p u v z
	V	4 C J L P U V b f h j m o q r t u w
H	-	

Table 6.23: Pairs of orientations maximizing Yellow-Red interactions in a single unit cell. We check 60-degenerate pairs of orientations of each Group consensus paths, noting any pairs of orientations that could preserve the maximum Yellow-Red interactions ($S_{YR} = 6$). These were only found for Group G. Essentially, degenerate orientations of these paths could be found in the single unit cell, throughout the crystal lattice, orientated with respect to each other as noted.

Group	Orientation	Min S_{YR}	S_{GB}
B	-		
C	c	5.333	2.333
	B	5.333	2.333
	S	5.333	2.333
G	-		
H	-		

Table 6.24: Orientations that maximize Yellow-Red interactions ($S_{YR} \geq 5\frac{1}{3}$) for all possible unit cells that preserve the Green-Blue interactions (in practice 3-fold degenerate). Green-Blue interactions will be the same regardless of the unit cell: thus the values calculated are for any possible unit cell that preserves the Green-Blue orientation. The only possible results occur from Group C.

6.8.3 Other unit cells

For increasing unit cell size, it is increasingly unlikely that the ensemble would form, as the interactions would require a fine balancing of electrostatics, and the possible disorder in the system is very high. Discriminating between these larger cells also becomes more difficult as they become more numerous. Ultimately the best approach would be a stochastic model of crystal formation using these asymmetric orientations of the group consensus sequences, instead of a continued deterministic investigation of optimum orientations. A kinetic self-assembly model, as introduced in Chapter 3, is being developed to simulate the formation of crystalline STNV. Putative unit cells can be validated against the simulation data; the simulated crystal can be probed for both symmetry and asymmetry. The kinetics of formation are essential to include, as they can discriminate between unit cells that may be energetically advantageous to form, and those that are not, reducing the search space and providing insight into the formation of the crystal. Importantly, the resultant simulated crystal and its packing can be compared directly against the experimentally derived asymmetric crystal structure.

6.9 Discussion

We have performed a theoretical analysis using deterministic modelling of the possible RNA organization inside STNV capsids, and the probable crystalline arrangements of virions containing conserved organizations thereof. The goal of this continued work is to compare the predictions from this modelling to an asymmetric X-ray crystallography structure of STNV VLPs containing RNA fragments [49, 247], which must include some asymmetry for the asymmetric structure to have formed. The asymmetric crystal structure is of a low resolution, and in addition to static disorder, there are polycrystalline regions throughout that also result in some averaging of the observed RNA orientations. For the unit cells we have probed, we have checked the implications on crystal symmetry as a result of electrostatics arising from the genome organization: without some aspect of regularity, any details will probably be washed out, so this can be used to identify conserved structural features in the genome organization of different viral particles. Expanding the search to larger unit cells (to probe resulting crystal asymmetry) has not been performed yet. This is a stretch goal, but demands a shift in approach from deterministic modelling of unit cells, to a stochastic kinetic assembly model of the crystal. This is crucial, because the unit cells cannot simply be abstract descriptions of the crystal: the overall crystal structure must form kinetically, and the unit cell reproducibility must be considered.

Unlike MS2 (Chapter 7), there is less evidence to suggest a rigid small number of moves (only 3 options in MS2), so we have worked with groups of paths rather than individual paths. The groups have common mechanisms, *i.e.* use common moves and sets of moves which are likely to occur because they involve small distances and help recruit a full capsid. Two possible groups of paths provided results which deserve further study: Group G and Group C. Of the

two, Group G had the most favourable dominant Yellow-Red interactions, and it is likely that these will force the crystallizing virions to form a dual unit cell (in the AB stacking regime) as indicated by the results in Table 6.22. However, similar electrostatics between Red and Yellow kissing points can occur at many orientations of the virions for Group G (as shown in Table 6.23), and this is exacerbated by poor stabilizing Green-Blue interactions; the data suggest that although the crystal may be dominated by pseudo-dual unit cells, there will be polycrystalline non-symmetry throughout.

In the case of Group C, the organizations were not perfect, and we were unable to determine an electrostatically optimal stacking of the Yellow-Red kissing points. However, if the S_{YR} and S_{GB} metrics are considered together in conjunction, Group C has the best predicted stacking. Group C organizations starting at proteins `c`, `B` and `S` allow the least electrostatic clashes irrespective of the unit cell, if the Green-Blue interaction cannot be ignored. It is entirely possible that although some electrostatic clashes persist, they may be able to be mitigated by a rearrangement of the RNA organization inside the capsid upon crystallization, possibly even by following alternate assembly mechanisms, characterized by different sequences of moves (as outlined in Table 6.5). MD simulation could be used to check whether partial RNA rearrangement within the capsid at crystallization requires consideration as a compensating mechanism for electrostatic repulsion.

MD could also be used to check how much the charge differences for certain moves actually affect the formation of kissing points. In general, MD would be an excellent tool to explore other possible stacking paradigms for STNV; in conjunction with the 6-D cross-correlation algorithm we could predict other stacking paradigms that would work if the RNA was organized differently.

The stacking of particles featuring the RNA organization described here

could be benchmarked against other VLPs encapsidating synthetic RNAs or partial genomes, which is possible as the assembly construct is a true *in vitro* model [49]. The crystal symmetry group and the asymmetric unit cell could be probed to uncover differences in stacking, which could be compared to the packaged RNA sequence identity.

Further work could involve collection of an asymmetric tomogram, that could be analysed in a similar approach as Chapter 7, providing higher-resolution data to resolve the organizations of the STNV genome in more detail; this will be the topic of our future work. We also seek to compare the spacing of the putative PSs identified in the genome by Bunka and colleagues [34], to the approximate distances predicted in Table 6.1, in a similar manner to the analysis published recently for two related bacteriophages [206]. Lastly, we hope to discover more about the assembly of STNV, and the stability of capsid assembly pathways, by analysing the assembly kinetics and efficiency implied by different types of genome organizations in proximity to capsid, based on the classification of move sequences presented here.

Chapter 7

Analysis of MS2 self-assembly

In capsid-RNA[1] co-assembly, the kinetic pathways of capsid assembly and the genome organization impact on each other, due to the imposition of the order of RNA-CP interactions during the capsid formation. Thus, we can learn more about assembly by studying the static genome organization in the mature virion. Here we present a new method developed for the analysis of cryo-EM data, and demonstrate it for the example case, bacteriophage MS2.

7.1 MS2

Bacteriophage MS2, which infects *E. coli* (Figure 7.1), is a member of the *Leviviridae* with $T = 3$ icosahedral symmetry, with a capsid composed of 90 protein dimers (Figure 7.2(a)), and a genome of 3569 nt of ssRNA. In 1976, the RNA of this phage became the first ever full genome to be sequenced [252]. Since then, the phage has been well characterized by a number of structural [185, 205, 253], biochemical [213, 254, 255], and theoretical studies [10, 199, 200, 206, 231, 256], and is thus often used as a model viral system. In this chapter, similar to the previous, we will focus on the sequence specific nature of RNA-CP interactions

[1]For all abbreviations see the glossary on page 222.

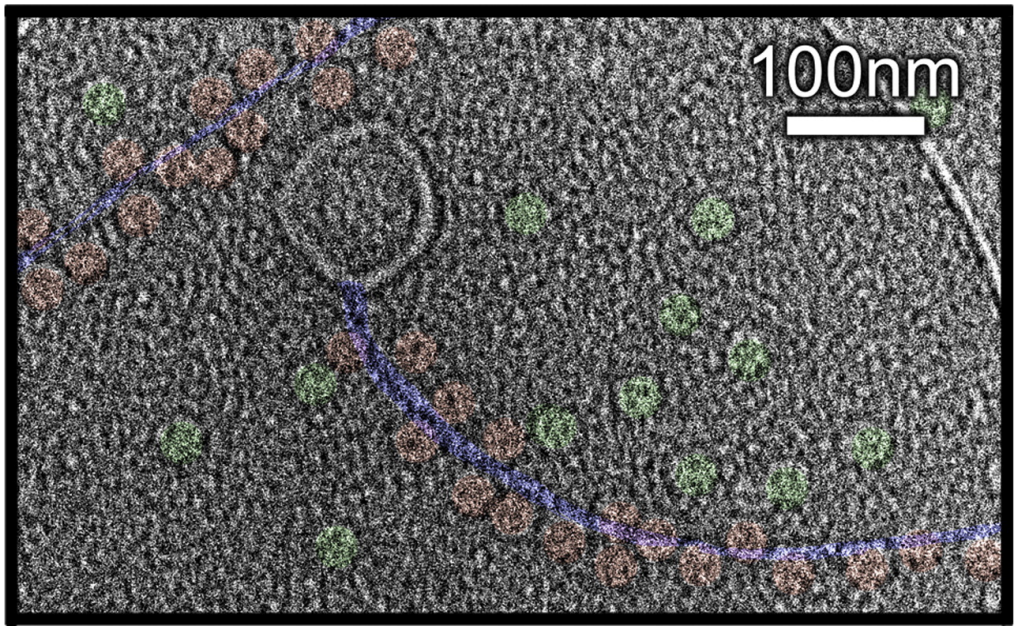


Figure 7.1: MS2 particles infecting *E. coli* F-pilus *via* binding using its single-copy MP. Coloured in red are the virions attached to the *E. coli* F-pilus (in blue), whilst the unattached viral particles are coloured green. Micrograph is adapted from [205].

in MS2, which have become recognized as fulfilling important functional mechanisms for both controlling assembly and disassembly/uncoating. In particular, a low nanomolar affinity complex has been identified between a dimer of CP and TR [213, 215], a 19 nt stem-loop that is a translational operator of the MS2 replicase cistron, and also serves to trigger the assembly of the $T = 3$ CP shell.

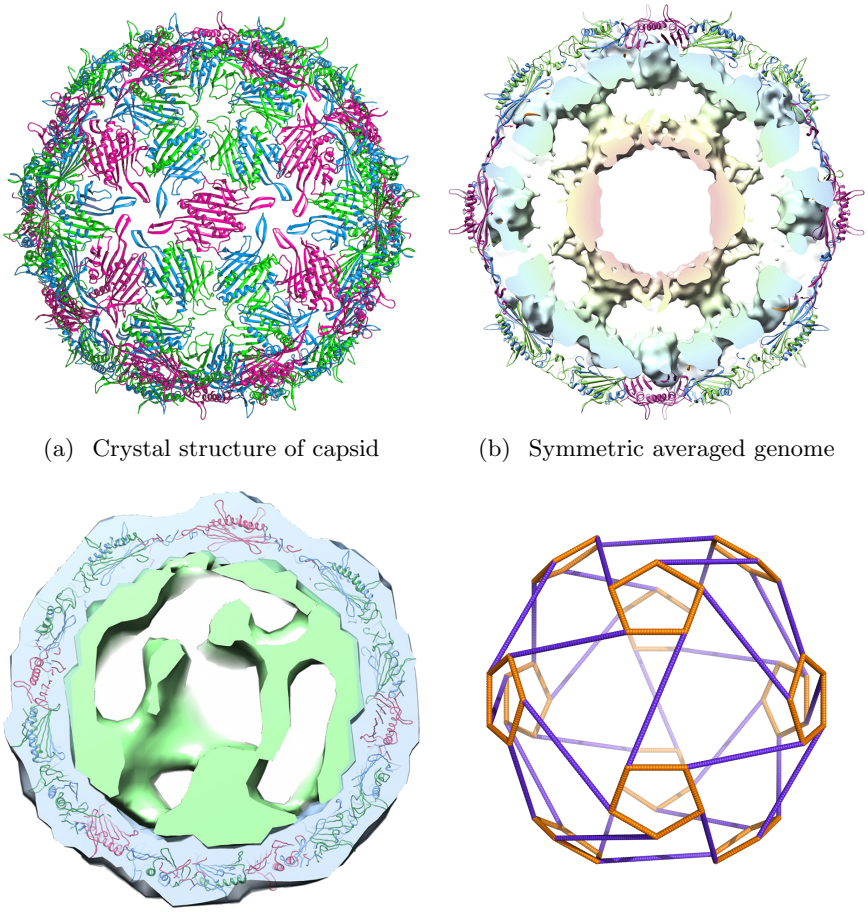
In spite of the importance of PS RNA-CP contacts for viral assembly, the electron density profile of the genome often is attenuated in high-resolution structures of MS2, and also in other RNA viruses [185, 253, 257]. Conversely, in cryo-EM structures of MS2 solved to a moderate resolution, the RNA genomic material is seen be organized pseudo-icosahedrally similar to the other *Leviviridae* (Figure 7.2(b), *cf.* Figure 5.12), tracing the shape of a polyhedron (Figure 7.2(d)) inside the capsid [35, 36, 258]. This leads to the conclusion that icosahedral averaging to high resolutions can often eliminate substantial portions

of the genome, which is organized pseudo-icosahedrally despite being intrinsically asymmetric (Figure 7.2(c)). Therefore, determining the structure of the genome to a high resolution is a challenge, and there is an opportunity to utilize additional insights and techniques, which is the goal of this chapter.

For MS2, we can make a distinction between the different shells of RNA seen within the capsid in the medium-resolution icosahedrally-averaged reconstructions. As discussed in Chapter 5, the central portion of the genome within the capsid is thought to be liquid crystalline, and its properties governed largely by the substantial repulsion expected between the overwhelmingly negatively-charged RNA. This is borne out in the structure, as the centre is mostly unresolved: therefore the overall structure is not even pseudo-icosahedral, but truly asymmetric.

We can use Hamiltonian paths to describe the outer shell of RNA in contact with the capsid [36], which appears under icosahedral averaging as a cage with 60 nodes: each node is situated where a PS is bound to one of the 60 asymmetric protein dimers within the capsid (Figure 7.2(a)) [10,199]. The PS RNA-CP node positions are connected in the virion by the single ssRNA molecule packaged inside MS2, but under icosahedral averaging, the polyhedron appears as an averaging of these pseudo-icosahedral connections. The icosahedral cage is seen to have two different length edges: short, around the five-fold axes, and long, across the two-fold axis. The graph describing the icosahedral cage is given in Figure 7.2(d).

There are thus structural implications in the mature virion for the organization of the PS connections (of ssRNA) within proximity to capsid. These may have implications for the uncoating/extrusion of the virion [34,259], yet directly result from the simultaneous co-assembly of CP and RNA into an intact particle. The RNA-CP binding mediates the order in which CP is recruited onto



(a) Crystal structure of capsid (b) Symmetric averaged genome

(c) Asymmetric genome (d) Graph describing icosahedral cage

Figure 7.2: Bacteriophage MS2: capsid and genome structure. (a) Crystal structure of MS2 capsid (ignoring single-copy MP). The viral capsid is formed from 60 asymmetric and 29 symmetric copies of the CP dimers, with one MP that takes the place of a symmetric dimer (PDB:2MS2 [185]). (b) The genomic RNA is organized inside the particles in two shells, with the outer shell adopting the shape of a polyhedral cage in icosahedrally-averaged reconstructions [12, 36]. (c) Asymmetrically averaged tomogram of bacteriophage MS2 bound to its receptor, the bacterial F-pilus [205]. The portion of the electron density corresponding to the CP shell (and bacterial pilus) is shown in blue; green depicts the density for genomic RNA (and presumably some elements of the MP), which forms the basis of the analysis described in this chapter. The RNA density forms a shell that is intimately associated with the inside surface of the capsid. (d) Graph of the polyhedral cage describing the icosahedrally averaged genome, showing long (purple) and short (orange) PS-PS RNA connections. The cage also partially describes the outer shell of the asymmetric genome structure.

the growing capsid: the order of assembly is crucial to maintaining an efficient pathway of assembly [11].

The first conformational change occurs at the highest-affinity PS in the MS2 genome: the TR site [215]. Initiation of assembly begins when C/C homodimer CP binds to the ssRNA at the TR stem-loop, enabling a conformational change into an A/B heterodimer that facilitates formation of five-fold axes [199, 200] by preventing the electrostatic repulsion of the F-G loops of the CP dimers that would otherwise clash around the five-fold axis (Figure 7.3). This switching occurs allosterically, and its magnitude is predicted to be independent of the primary structure of the interacting stem-loop [199]. As the CP aggregates into partially-formed capsid, the CP ensemble is decorated with RNA connecting the PS positions: effectively a partially formed RNA polyhedral cage, with each vertex corresponding to a PS *in situ* (see Figure 7.4). As the RNA binds to CP and the assembly proceeds the remaining parts of the ssRNA are rearranged: the secondary structures of some stem-loops will already have formed, allowing for immediate PS interaction, whereas some other stem-loops might become accessible as the RNA conformation changes. This reorganization of the RNA allows further PS-CP binding events to proceed, in an order that is controlled by the RNA conformation (Figure 5.11). Thus, interactions between CP and RNA mediate the order of capsid assembly.

Consideration of MS2 biology can reduce the number of possible Hamiltonian paths that could describe the outer shell of RNA. For MS2, it has long been known that the $5'$ and $3'$ ends of the genome (at 388–398 nt and 3510–3520 nt respectively) both bind to the single-copy MP (Figure 7.5) [260], which is an important site in infection (see Figure 7.1). This bidentate binding circularizes the genome, with TR located in the centre of the long-looped segment of the genome. As discussed, CP assembly then initiates at TR, and in the final capsid

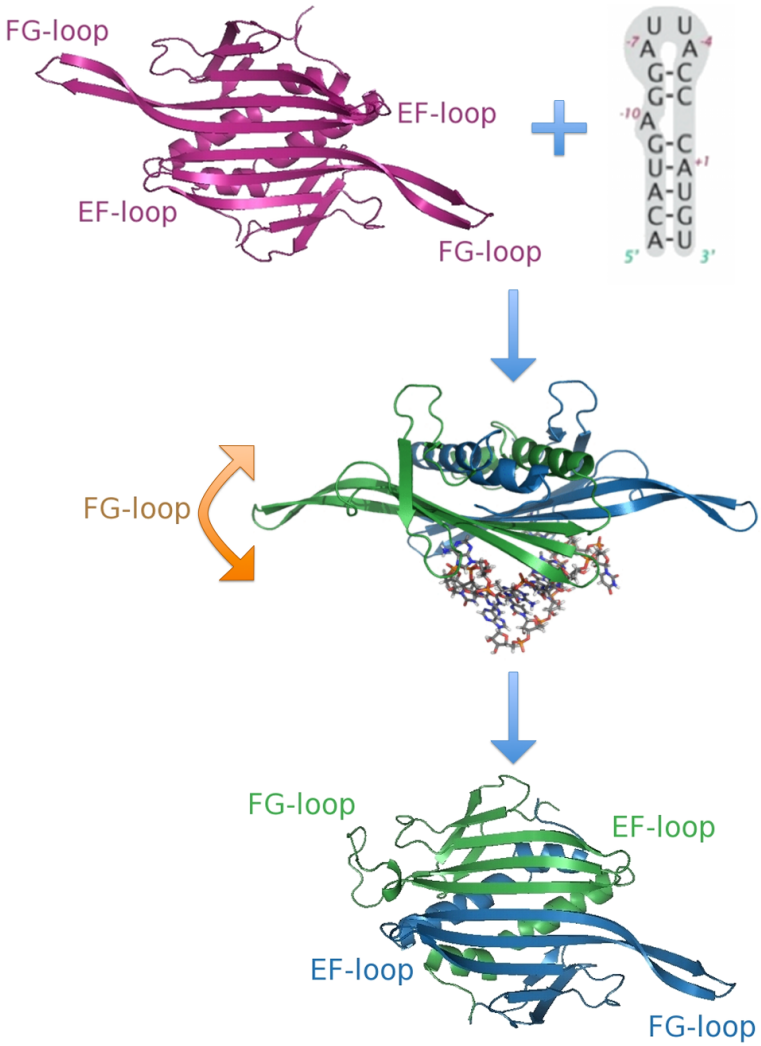


Figure 7.3: MS2 quasi-conformers: allosteric switch. Efficient assembly requires a quasi-conformational change from the dominant symmetric C/C (magenta/magenta) dimer form into asymmetric A/B (blue/green) dimers. RNA-CP interactions on the C/C dimers is a trigger of this, specifically PS binding to CP (*n.b.* this can also happen stochastically, albeit very slowly [200]). After PS binding, major rearrangements of the C/C dimer take place within the FG-loops: one FG-loop becomes more dynamic, whilst the other FG-loop becomes more stable. This enables folding of one of the dynamic FG-loops into the B conformer of the A/B heterodimer. More than 12Å separates the binding site of PS and the FG-loops: thus the switching effect is described as allosteric [199, 200].

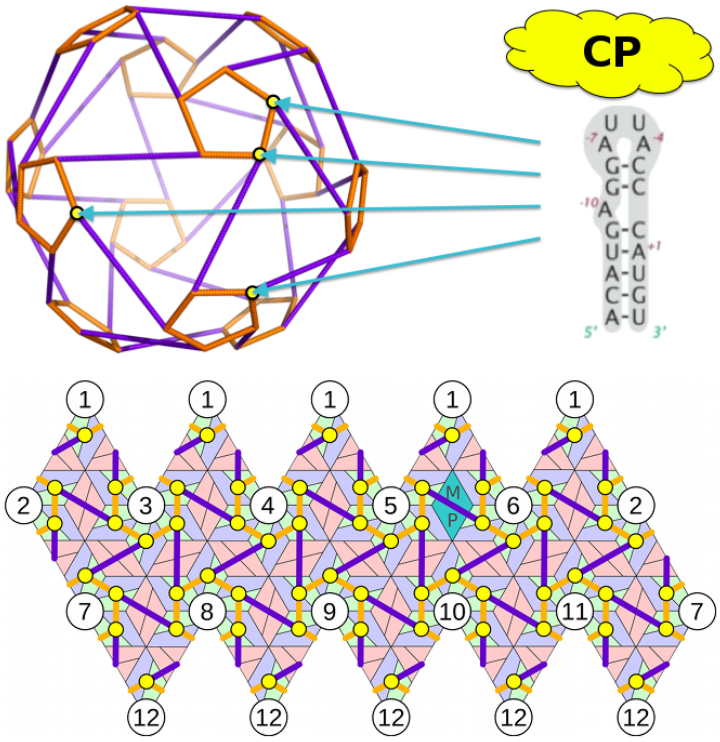
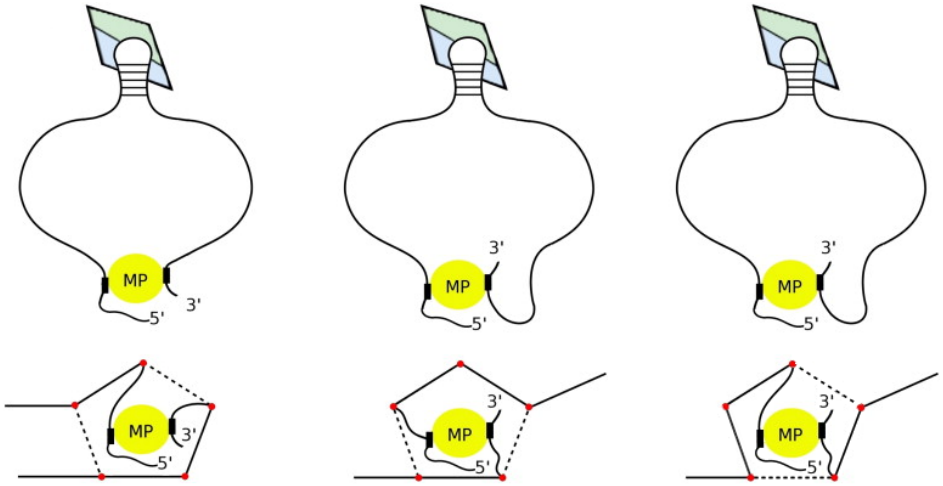


Figure 7.4: Packaging signal binding positions at the vertices of the RNA polyhedral cage, which is shown here both in 3-D, and in a planar representation superimposed on a model of the capsid protein structure. Up to 60 PS sites are present under each asymmetric dimer of the MS2 capsid: 5 around each five-fold axis [12]. However, only 53 PS have been found in a secondary-structure analysis of the genome; extras have been inferred [10]. RNA connects the PS positions in the outer shell of RNA visible under cryo-EM reconstruction [12, 36, 205].



(a) Cycle (b) Pseudo-cycle (c) Pseudo-cycle

Figure 7.5: Hamiltonian cycles and pseudo-cycles. RNA Hamiltonian paths of MS2 start and finish at the same five-fold axis, as (a) Hamiltonian cycles, or alternatively as (b) or (c) Hamiltonian pseudo-cycles.

MP is thought to take the place of a symmetric C/C homodimer [12, 205]. The implications of these results are that the Hamiltonian paths start and end in the same area of the polyhedron, thought to be the same five-fold [206]. Paths that start and end at the same five-fold vertex are necessarily either Hamiltonian cycles (Figure 7.5(a)), whereby the paths start and end on adjacent dimers, or else we define them to be pseudo-cyclic (Figures 7.5(b), 7.5(c)), where the paths start and end on opposite dimers across the five-fold axis. In either case, after binding to the final dimers, both ends remain attached to the MP.

7.2 Geometric constraints on genome organization

Paths are calculated with reference to the structural organization of the capsid. For this, a labelling system is introduced, in which different protein dimers in the capsid are labelled as (*cf.* Figure 7.6):

`abcdefghijklmnopqrstuvwxyzABCDEFGHIJKLMNOPQRSUVTWXYZ01234567`.

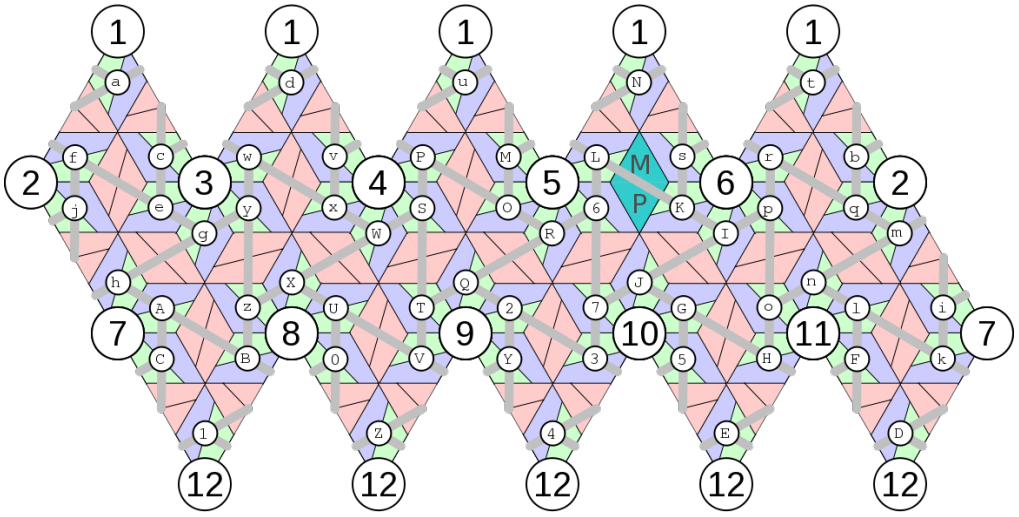


Figure 7.6: Planar representation of capsid geometry. Protein heterodimers in MS2 are labelled, with a single character from the range a—z + A—Z + 0—7. 5-fold symmetric vertices are labelled 1—12.

Paths are calculated starting at dimer a. There are two ways that the paths can start and end around a single five-fold axis, shown in Figure 7.5, and these do not need to be explicitly calculated: this will save computation time. Instead, moves around the five-fold vertex 1 are ignored. Thus the first move possible is only to dimer b. The start of the path, corresponding to an incomplete path of length 1 is given by: ab.

Starting at $n = 1$, for each incomplete path of length n, the path is attempted to be increased by every possible single move from the last added node. Since paths are not allowed to backtrack, and all nodes are 3-coordinated, *i.e.* have three edges, there are at most two possible moves. Any moves that are possible, *i.e.* do not return to a previously visited node, and do not fall around 5-fold vertex 1, are included in a list of paths of length $n+1$. This algorithm continues until all nodes (apart from those surrounding vertex 1) have been visited. At this stage, all possible Hamiltonian paths have been generated, excepting the start/end of the paths.

Taking into account the start/end of the paths, there are in total 132 path options (see Appendix B), which are equally valid to be realized in $5'$ to $3'$ order as in $3'$ to $5'$ order.

Once paths have been generated on the protein neighbour map, they can be generalized into moves, *i.e.* described as successive geometric operations mapping protein dimers onto neighbouring dimers, rather than the protein dimers themselves. This allows an easier way to recreate the paths starting from any given protein dimer. Also, calculation of symmetric and mirror paths is much easier when the protein positions are disregarded, as such degenerate paths can be calculated with simple string manipulations.

Given the calculated generalized paths, we have to superimpose these paths onto the RNA density, starting and finishing at defined points with reference to the viral capsid. Note that the starting points are referred to by the same protein labelling map as before, see Figure 7.6. These starting points come from consideration of the binding of the RNA $5'$ and $3'$ end regions to the MP, which localizes the ends of the RNA in the vicinity of MP. Since the resolution of the averaged tomogram, obtained *via* alignment and averaging of individual tomograms, was not sufficient to unambiguously identify the location of the MP, and the binding sites of the RNA were difficult to identify, we bookmarked all paths which started and finished within the eight five-fold axes closest to MP. This was a very conservative overestimate, which ensured that no possible path was missed in our analysis. As the paths can start at any of the five PS positions (nodes) at those vertices, there are thus $132 \times 8 \times 5 = 5,280$ possible path solutions.

These form a library of constraints on genomic RNA organization in proximity to capsid, that we are using in the following to interrogate cryo-ET data (provided in Appendix B). Note that for viruses with different polyhedral RNA

organizations the same method can be applied by computation of the Hamiltonian paths on the corresponding polyhedral density. Moreover, since Hamiltonian path computations only depend on the topology of the polyhedron, *i.e.* the network of connections between vertices irrespective of the lengths and orientations of the edges, the same library of Hamiltonian paths can be used for wider classes of viruses, such as those studied by van den Worm *et al.* [35] or bacteriophage GA [206].

7.3 Asymmetric sub-tomographically averaged structure

The analysis is based on a sub-tomographically-averaged, asymmetric structure of MS2 [205], obtained by imaging mature MS2 bound to its natural receptor, the F-pilus of *E. coli*. A total of 22 tomograms were taken with 2374 bound viral particles. The 1500 best correlating virion subtomograms (63% of the total) were normalized, low-pass Fourier filtered to 30Å, and then averaged to produce a structure at 39Å resolution. The data was presented as a density map of 643 pixels, sampled to 9.12Å per pixel (EMD:2365 [205]).

7.4 Difference map between tomogram and X-ray protein structure

A difference map between the asymmetric EM reconstruction [205] and the X-ray structure of the protein capsid (PDB:2MS2 [185]) was determined as follows: The protein structure was filtered to 39Å resolution to match the EM data. Then the pixel size and orientation of the two maps were made equivalent by trilinear interpolation of the reduced-resolution X-ray structure with Chimera [251]. Radial

plots compare the distribution of density in the protein map and the tomogram, with the pilus/MP complex masked away for the calculation (see Figure 7.7). The radial distributions are expected to be similar in the radial ranges corresponding to the CP shell, but different elsewhere at radial levels corresponding to viral RNA (which is organized as a two-shell architecture, see [36]) and the 44 kDa single-copy MP. Note that the radial distributions are not identical in the area overlapping with CP—this is due to the low resolution of the map as the CP shell density cannot easily be accounted for in the asymmetric map. Therefore, a contour mask of the tomogram with the protein map is used to sample the low-resolution map and used to eliminate the protein density *via* the UCSF Chimera mask routine [251], rather than a direct subtraction of the normalized maps. A mask of 0.5σ best isolates the RNA whilst excluding protein. Finally, two icosahedral masks are applied: the inner core of RNA is masked away under radius 80Å, and an outside mask of radius 120Å removes noise resulting from masking artefacts and the pilus/MP complex. The resultant pruned density (Figure 7.7) contains information about (i) the outer RNA shell in contact with CP, (ii) MP, and (iii) potential traces of CP lying within the shell that were not captured by the masking process.

7.5 Difference map between icosahedrally-averaged EM density and the X-ray protein structure

A difference map is created between the icosahedrally-averaged map [36] and the X-ray structure of the protein capsid (PDB:2MS2 [185]). We base our analysis on density map EMD-1431 of mature MS2, calculated using single particle analysis of 9,335 separate images, equating to c. 560,000 sample points with icosahedral averaging [36]. We use a procedure analogous to the one described above for

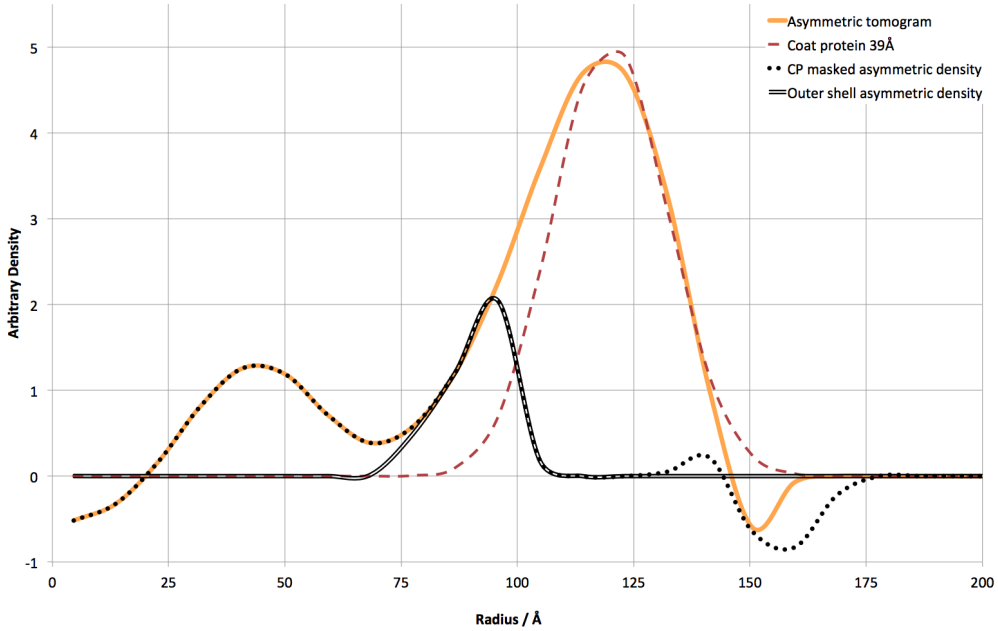


Figure 7.7: Radial plot of the difference map between tomogram [205] and Fourier-filtered X-ray protein structure [185].

the tomogram to isolate the RNA. The protein structure is filtered to 9.5Å resolution, with a grid spacing of 1.26Å, to match the symmetric map, and normalization of the resultant protein map to the CP area of the symmetric map is performed. The resampled filtered protein is then subtracted from the symmetric map, yielding a symmetric cage of RNA with a polyhedral shape as in Figure 7.2(d). The outer shell of RNA is isolated by icosahedral masking with vertex radii of 80Å and 120Å. The resulting map for the outer shell of RNA in the icosahedrally-averaged map (Figure 7.8) was aligned with that for the asymmetric RNA organization in the tomogram by reference to the X-ray protein structure used to create each difference map using UCSF Chimera [251]. After normalization the aligned maps had similar average, standard deviation and maximum density values.

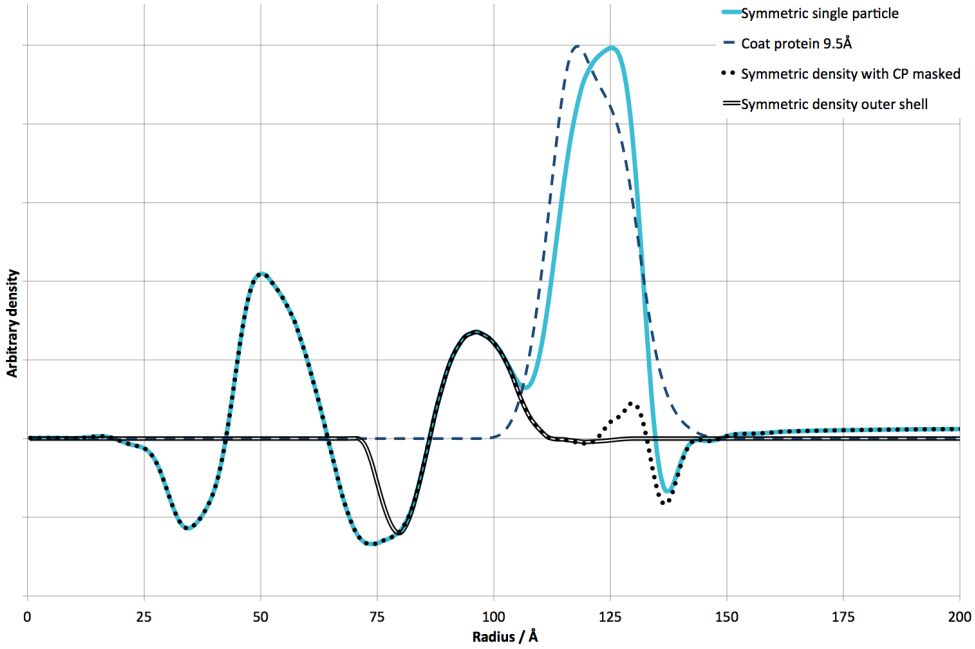


Figure 7.8: Radial plot of the difference map between icosahedrally-averaged map [36] and Fourier-filtered X-ray protein structure [185].

7.6 Mapping data onto the geometric model

The UCSF Chimera Segment Map tool [261] was used to perform a watershed segmentation (Figure 7.9) on the symmetric RNA cage density, which partitioned the polyhedral density into segments attributed to its edges. Each long edge of the cage in Figure 7.2(d) was represented by three segments as shown in Figure 7.10. The same watershed segmentation is applied to the asymmetric RNA outer shell map. Hence pixels from the asymmetric RNA map can be associated with defined segments on the polyhedral shell, and each connection thus has a density profile associated with it.

Of the segmented data (Figure 7.10), only tomographic density overlapping the middle segment (pink) is retained for analysis, as density overlapping with the outer segments (red) may potentially also sample density associated with short edges and RNA-CP connections (PS stem-loops), and can therefore not

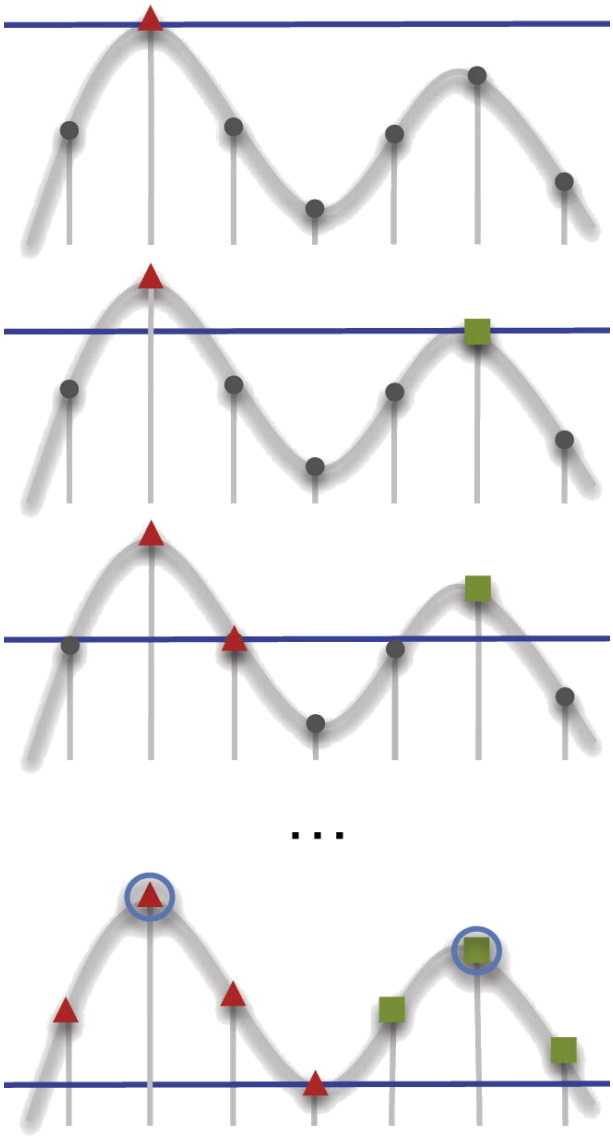


Figure 7.9: Cartoon explaining watershed segmentation in 1-D, adapted from [261]. The diagram illustrates the first three and the final step for a watershed segmentation algorithm performed on a 1-D density map. The smooth curve shared represents the underlying true density function with the y-axis representing the density value of the map, which in the reconstructed data is discretized *via* sampling onto an evenly spaced grid. Each sampled point is considered in order of decreasing density: if it is positioned adjacent to a point that already has been placed in an existing region, then it is also added to that same region; else it starts a new region. In this example, two regions result: that of the triangular and square points. It can be seen that each region corresponds to a local maximum in the density (circled). Importantly, the process is repeated on successively smoothed maps that have undergone Fourier filtering, in order that the regions obtained by the unfiltered iteration can be grouped together [261].

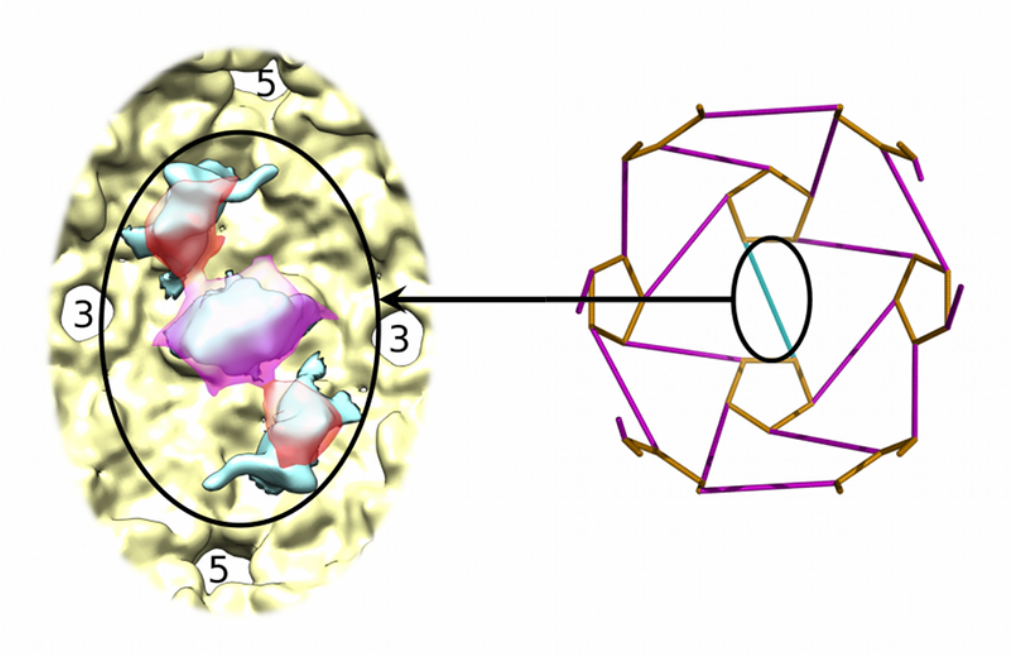


Figure 7.10: Illustrations of the segmentation procedure. Each long edge of the polyhedral density (corresponding to the icosahedrally-averaged map) is partitioned into three segments *via* the UCSF Chimera SegmentMap tool [251, 261]. The middle segmented density (pink) will contain RNA density originating from the long edge alone, whereas the outer density segments (red) could additionally include density arising from neighbouring short edges and RNA-CP connections (*i.e.* PSs). The segments shown in this figure are from a representative single connection (coloured cyan), not an average of all the connections, and are shown viewed from inside the virion along a particle two-fold axis. CP in the background is shown in beige.

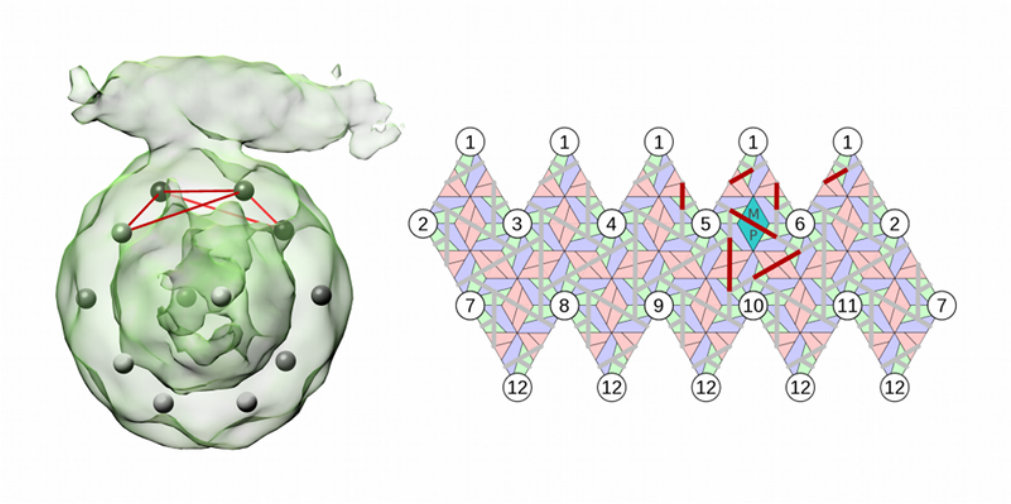


Figure 7.11: Positions of the long edges ignored in the analysis due to proximity to MP. Five long edge connections, shown as solid lines between five-fold vertices, are omitted as their proximity to MP makes association of a corresponding RNA density distribution ambiguous.

be unambiguously attributed. The segments shown in this figure are from a representative single connection (coloured cyan), not an average of all the connections; they are shown viewed from inside the virion along a particle 2-fold axis. CP in the background is shown in beige.

We decided to make a very conservative decision on how much data to include, and thus only use the density encoded by the middle segment to represent a long edge. This is because the short segments close to the polyhedral nodes, as well as the short edges themselves, may contain density corresponding to the RNA-CP contact—the PS complex—located at a polyhedral node bordering the edge, which could distort the analysis. Moreover, connections between PS positions adjacent to the MP/pilus, see Figure 7.11, are discarded as they may contain unmasked density arising from the MP, genomic RNA, or a combination of both.

7.7 The density profiles

The library of putative path organizations was used as a set of constraints in the analysis of the asymmetric electron density for the outer RNA shell, which we isolated from the tomogram as described in §7.4. Note that any path in the library provided information on which edges were likely to be occupied, given that occupation of some of the edges—or the lack thereof—could be confirmed based on the tomogram. The first step was therefore to determine a subset of the 90 edges of the averaged map (with reference to the polyhedron in Figure 7.2(d)) that were likely occupied or unoccupied given the density distribution of the tomogram. We excluded all short edges as they were too short to distinguish unambiguously whether density represented the RNA-CP contact (*i.e.* PS) positioned at the vertex, or a connection between two PSs along a short edge. As discussed above, we also disregarded the five long edges around the MP (Figure 7.11), as it was not possible to ascertain whether all density in these regions was attributable to RNA.

The edges were named with reference to the nodes they connect. In particular, we labelled the nodes with reference to the capsid proteins, using again the labelling system introduced earlier, *i.e.* PS positions being labelled as proteins as in Figure 7.6, which shows the relation between the heterodimers and nodes. The capsid has been orientated such that the MP maps (arbitrarily and without loss of generality) to the homodimer on the two-fold axis between vertices 5 and 6, as is marked on the path outputs.

To determine which long edges are occupied, we analysed the density distributions as follows. We attributed tomographic density to each of the 25 long edges of the polyhedral cage representing the icosahedrally-averaged density considered in this analysis and fitted it to a normal distribution, using the norm.fit function from the scipy.stats python library. The normal fitting function auto-

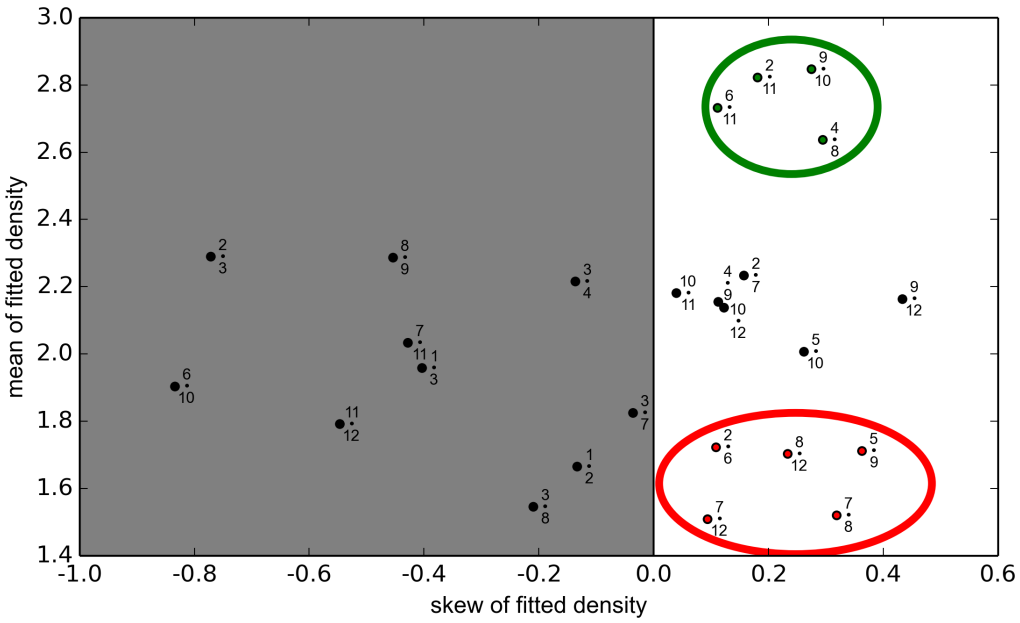


Figure 7.12: Classification of polyhedral edges as occupied and unoccupied: a comparison of the density profiles of the sampled long edge connections. The mean of a fitted normal distribution (y-axis) is scattered with a skewness parameter (x-axis). Connections with negative skew are disregarded as no statement about occupancy can be deduced in this case. From the remainder, two groups of four and five connections are identified as occupied (in the green circle) and unoccupied (red circle), respectively. These are used as constraints in the analysis.

matically calculates the best positioning of a unimodal normal distribution for the dataset. For a sparse dataset the mean of a fitted normal distribution is less affected by outliers than a simple arithmetic mean of the raw data. Note that connections occupied in the RNA density are expected to have a substantially higher mean density than unoccupied connections. Thus a ranking of the level of density associated with these edges was achieved using the mean of the fitted normal distribution. Using the fitted mean, four connections stood apart from the others, with mean densities of 2.6–2.9, see Figure 7.12, suggesting that these four edge connections were likely occupied by RNA in the virion. These were denoted as *occupied* connections, and were used as constraints in the analysis of the asymmetric structure.

To determine which connections could be classed *unoccupied*, we used the skew parameter of the sampled distributions to examine smearing of density (computed *via* scipy.stats.skew). Skewness characterizes the balance of a distribution to either side of the peak density. As expected, the group of connections classed occupied had a skew between 0.1–0.3. Negatively skewed connections were disregarded from the analysis, because a negative skew meant that there were only a very limited number of high-density points, which made up the cumulative density. Because of their low copy numbers, small fluctuations in sampling made a big difference to the overall density, and we therefore did not want to make a judgement of occupancy based upon these data. We therefore did not place any constraints on edges with negatively skewed distributions. The remaining data were separated into distinct groups. The five data points shown in the red circle in Figure 7.12, with mean values between 1.5–1.8, were adjudged unoccupied, *i.e.* characterized by an absence of density corresponding to RNA.

Although we were unable in this analysis to make use of any data regarding short edges, due to the low resolution of the asymmetric tomogram, information regarding long edges turned out to be sufficient for conclusions to be drawn. There were nine constraints on RNA organization that were used to compare the asymmetric structure with the library of all possible Hamiltonian path organizations: four long edges were deemed occupied, and five unoccupied.

7.8 Determining RNA organization in proximity to capsid

Only five members of the library of all possible Hamiltonian paths were consistent with these nine constraints. In Figure 7.13 we display the occupation of long edges with reference to the two five-fold vertices they connect, following the

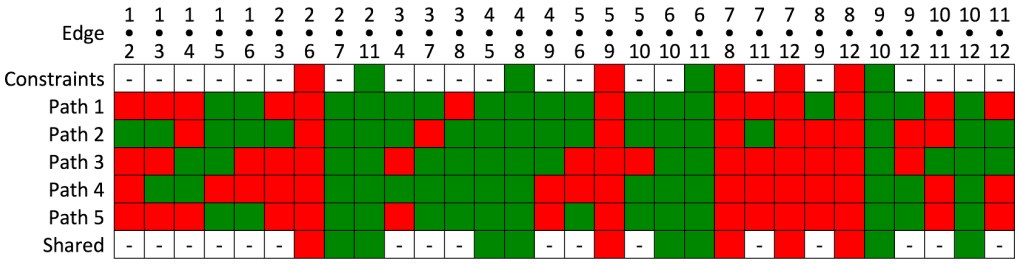


Figure 7.13: Constraints on the RNA organization consistent with the tomogram. Each possible RNA organization is characterized by which long edges (Figure 7.2(d), purple edges) are occupied in the polyhedral shell of the icosahedrally-averaged density. Long connections are labelled by the numbers of the five-fold vertices (Figure 7.6) they connect (x•y connecting five-fold vertices x and y). Constraints imposed in the analysis are indicated in the first row, with green indicating an occupied edge, and red an unoccupied edge. The five paths meeting these constraints are characterized according to occupied and unoccupied edges. The last row shows edges shared by all five paths.

numbering scheme of vertices given in Figure 7.6. Intriguingly the paths match for 13 of the 30 long edges, suggesting that the structure common to all paths is likely to be a prevalent feature in different viral particles.

Each path was a roadmap of connectivity between RNA-CP contacts. In order to decide if any of these putative RNA organizations was more likely to occur than another, we used the following criterion: We associated with each option a density distribution by ascribing density to occupied edges in proportion to their lengths and computed the density obtained by averaging around the five-fold axis adjacent to MP. We used this as a characteristic to benchmark against the five-fold averaged density determined experimentally [262] (Figure 7.14(h), adapted from [10]). Path 4 (Figure 7.15(a)) closely matched (Figure 7.14(f)) this distribution, whereas the other paths did not.

Path 3 is an alternative embedding of the same geometric path as Path 4, but with a different orientation relative to MP, which starts and finishes at vertex 9. The occupation of the connections for Path 3 is different to Path 4, even though the overall geometry of the path is the same. Interestingly, these are the only two path solutions that do not have a connection that could be impeded by clashing

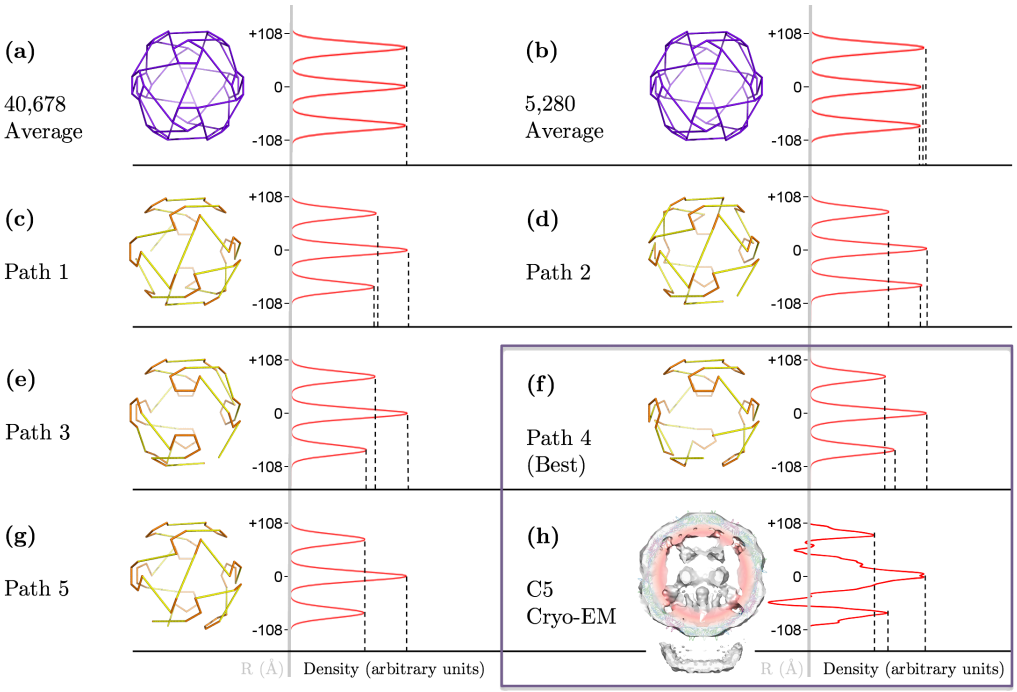


Figure 7.14: Symmetry averaging identifies Path 4 as the correct solution. C5-averaged densities in 1-D projection for tomographic data and the path solutions listed in Figure 7.13 are compared. The vertical axis shows the radial distance from the centre of the capsid in angstrom, and the horizontal axis corresponds to the C5-averaged density at that radial distance in arbitrary units; density profiles for tomogram and path solutions are normalized by equalizing the maximum densities. Density profiles are shown for: (a) the average of all possible 40,678 Hamiltonian paths; (b) the average of all paths consistent with RNA interaction with the MP; (c–g) the density profiles for the five paths in Figure 7.13 individually; (h) the C5 cryo-EM reconstruction from the tomogram, adapted from [10]. Path 4 (*cf.* Figure 7.15(a)), identical to Path 3 (Figure 7.15(b)) from a geometric point of view but positioned differently within the density with respect to MP, provides the closest fit with the cryo-EM data.

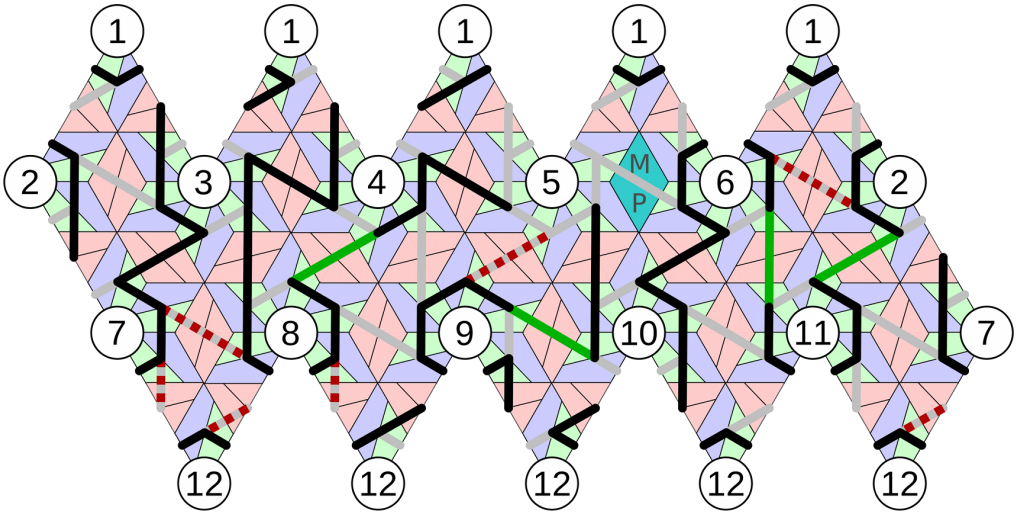
with the MP; *i.e.* they do not have long-edge connections between vertices 5 and 6 (see Figure 7.13).

This strongly suggested that Path 3 and Path 4 were indeed the correct models for the organization of the RNA in MS2, and that Path 4 would be most likely to be observed in any virion (as it matches the averaged density best): *i.e.* it is dominant in the ensemble of organizations including those that did not match to the density data in this analysis, and misfolded RNA conformations.

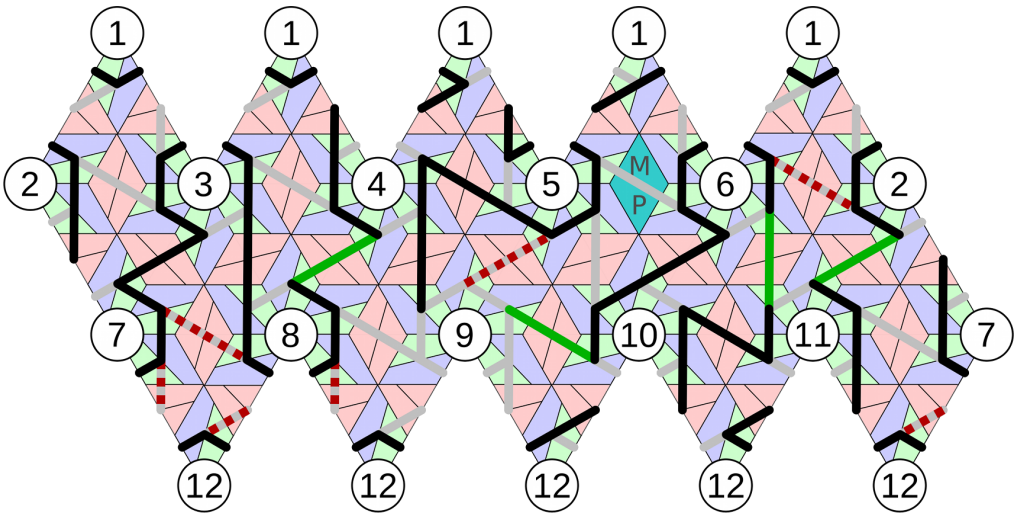
Remarkably, Paths 3 and 4 are also consistent with results of two independent studies: the assembly pathways determined *via* kinetic modelling of capsid self-assembly [10], and the PS positions identified *via* a bioinformatics analysis of RNA SELEX data [206]. Our analysis here represents a completely independent reconfirmation that the organization of the viral genome in proximity to capsid is highly constrained and likely identical in the vast majority of viral particles, taking on the organization depicted in Figure 7.15(a).

7.9 Discussion

The analysis method introduced here has for the first time identified the conformational path taken by a viral genome in proximity to its capsid from the low-resolution density map of an asymmetric, averaged tomogram. Previously, a model of the asymmetric genome organization in satellite tobacco mosaic virus (STMV) has been built [263]. That work relied on the icosahedrally-averaged crystal structure which revealed ≈70% of the viral genome to be in contact with the protein shell *via* a series of dsRNA segments ≈9 bp long [48, 264, 265]. The X-ray structure provided the first definition of RNA PSs [266]. In addition to the X-ray density the modelling used predictions of the most likely secondary structure elements within the genome to identify the sequences forming the double-stranded segments [267]. Ours is the first direct analysis of an asymmet-



(a) Solution 1: Path 4



(b) Solution 2: Path 3

Figure 7.15: Hamiltonian path solution identified by the method. (a) The best match with the C5 averaged data (Path 4) starts and finishes at vertex 5 adjacent to the MP (cyan). Following the colouring convention in Fig. 3, red dashed lines show unoccupied and green lines occupied constraints; other occupied connections implied by our analysis are shown in black. The position of TR, the strongest PS, is denoted in yellow [15]; heterodimers are coloured in blue/green and homodimers in pink. (b) An alternative embedding of the same (geometric) path with a different orientation relative to MP. The path (Path 3) starts and finishes at vertex 9; hence the occupation of the connections differs from Path 4 in (a), even though the overall geometry of the path is the same.

ric map containing RNA density. The method introduced here can be used to analyse any asymmetric dataset of a viral genome organization, provided that a distinct shell of density is seen in proximity to capsid in the averaged cryo-EM density, the contact sites between genomic RNA and capsid protein are known, and information regarding their positions and function can be used to formulate a constraint set on the connectivity between the PSs. Insights into PSs are becoming available for a number of ssRNA viruses *via* the use of CLIP-seq[2] techniques [268]. In addition, there is a growing body of work directed at obtaining asymmetric structures for this class of viruses in order to understand how their genomes are released during infection. Our approach is therefore likely to provide important insights into genome organization in wider groups of RNA viruses. In particular, many RNA viruses show order in the organizations of their genomes in icosahedrally-averaged cryo-EM and X-ray structures [207], for example bean pod mottle virus (BPMV) [269], STMV [48] and PaV [270]. In such cases, constraint sets in terms of paths with appropriate combinatorial properties can be used to map the putative asymmetric organization of their genomes into the corresponding symmetrically averaged densities and hence provide information on connectivity between the RNA-CP contact sites.

This example illustrates that structural information on genome organization obtained *via* the method introduced here has important implications for our understanding of the functional roles of viral genomes in virus assembly. More broadly, the method applies to any virus for which RNA-protein contacts are important for virus assembly, *i.e.* all viruses that follow a PS-mediated assembly process [8]. PSs are known to exist in a number of viral families including those infecting humans, *e.g.* alphaviruses [271], and plants [45], so this method is applicable to wider groups of RNA viruses. PS-RNA interactions are able

[2]Cross-linking and immunoprecipitation with high-throughput sequencing.

to bias assembly towards a subset of the possible assembly pathways due to differential PS-CP affinities for different PS [11]. Specific PS binding moreover enhances assembly efficiency by triggering a collapse in the hydrodynamic radius of the genome below the inner radius of the virus protein shell [213], enabling the assembly of the protein shell around the compacted genome.

Knowledge of the precise locations of the PSs and connectivity between them, which is provided by the analysis presented here, is therefore an important component in understanding the mechanisms by which viruses achieve the observed assembly fidelity and efficiency *in vivo*. This, in turn, is a prerequisite for the development of novel antiviral strategies that target virus assembly. As demonstrated in [11], drugs interrupting PS-CP interactions can slow down the assembly process and decrease viral yield *via* misencapsidation of cellular RNAs. Moreover, a better understanding of conserved features in the genome organization within a viral family provides novel insights into the selective pressures on viral evolution. The method described here enables the identification of such features, and therefore also has profound implications for our understanding of viral evolution.

Chapter 8

Conclusion

Ἐκ μέρους γὰρ γινώσκομεν

Unser Wissen ist Stückwerk

For we know only in part

(I Corinthians 13:9, translated by Martin Luther)

8.1 Biomimicry

Biomimicry is a revolutionary means of scientific advance that analyses and adapts nature's best ideas for technology [5]. In terms of self-assembly at the nanoscale, biological systems truly demonstrate the possibilities for chemical systems to consider when engineering synthesis routes, exploiting the advantages of self-assembly. In particular, there are five principal considerations for chemical engineers engaged in biomimicry, when seeking to replicate the stunning examples of structure and function found in self-assembled biological systems:

(i) Efficient and accurate assembly of products can be achieved without central coordination from simple precursors, if these precursors are designed to be able to self-organize.

(ii) Due to convergence, self-assembly processes are particularly economical.

(iii) A self-checking and self-correcting kinetic pathway for synthesis can be achieved by modification of the energetics of bonds formation.

(iv) The lexicon of interactions in the system can be simplified by the use of a few types of identical subunits, which reduces the amount of information needed to encode the structures of the components.

(v) Control mechanisms can be introduced that assist in the biasing of the self-assembling system towards kinetically favourable pathways, that operate without central coordination, and thus do not undermine the distributed nature of self-assembly.

To build complex structures, nature adopts a hierarchic approach, progressively increasing length scales whilst maintaining the order [272]. In terms of controlling self-assembly, the principle consideration is the nucleation of assembly. This is in clear contrast to a scenario where order appears from a mixture of assembly precursors spontaneously. The capsid assembly kinetics of many ss-RNA[1] viruses suggest a strong nucleation event followed by a rapid completion phase [11]. In this case, the strength of nucleation has evolved to be enforced by PSs in the genome and the controlling of the timing of CP synthesis and its availability in the vicinity of assembly. We have seen in Chapter 4 that transferring the latter approach to the organic chemistry laboratory could pay dividends in the synthesis of GNRs.

In the case of the genome, in the PS viral co-assembly paradigm it can be thought to act as a kind of scaffold that serves to recruit and stabilize the assembly components [180]. Approaches that utilize nucleic acids as a scaffold (though less finely tuned, and without PSs) have been realized in 1-D, 2-D and

[1]For all abbreviations see the glossary on page 222.

3-D for soft-matter systems [273–275]; this is also a possible avenue to controlling the synthesis of nanographene. Interestingly, in terms of control using PSs, a recent paper on the self-assembly of 3-D DNA bricks found that heterogeneity in bond energies when forming 3-D structural organizations improves the nucleation kinetics, and can be required for completion of some assemblies [276]. This is similar to the heterogeneous distribution of PS in a viral capsid: instead of a scaffold, it may be possible to use building blocks with different affinities for each other.

Indeed, designed complex self-assembled systems are trending toward the design of multiparticle systems, that consist of many distinct building blocks. Enriching the complexity of the system will allow fine-tuning of the factors that influence the kinetics of self-assembly; single assembly pathways could be selected by modifying individual strengths of interaction within the system [276].

As we have discussed in this thesis, there are shared mechanisms between all self-assembling systems, many of which still require a detailed analysis. Certainly, there is plenty of inspiration from nature that we could help develop new methods of controlling material synthesis methods, and many new developments in nanotechnology already seem to mimic existing biological processes [272,277]. It is clear that self-assembling systems used in the manufacturing of novel materials have a bright future.

8.2 Linking function and structure

A better understanding of the asymmetric organization of viral genomes is vital if we are to properly understand the functional roles of genomes in RNA viruses. Recent research has revealed that far from being a passenger in the assembly of the viral particle, genomes critically enhance the efficiency of virus assembly *via* multiple dispersed, sequence-specific contacts with capsid protein [8]. These

PSs act collectively in a cooperative manner [11, 66], and their relative placement in the tertiary structure of the genome is important for their function. In particular, it is the relative affinities of the PSs for CP at defined positions in the packaged genome that impact on the geometries of the assembly intermediates, *i.e.* on the structures of the partially assembled protein shells on pathway to capsid. For the viruses discussed here, it had previously been shown that this interplay of PS affinities and capsid geometry results in a highly ordered genome organization in proximity to capsid. It has moreover been established that the same overall organization of the packaged genome can occur in evolutionarily related viruses, *e.g.* MS2 and GA [35, 206], suggesting that there is a selective advantage for a specific genome organization in this family of viruses. This advantage can be explained in terms of assembly pathways: since PSs are instrumental in recruiting CP to the growing nucleus during PS-mediated self-assembly, the positions of the PS-CP contacts impact on the geometry of the assembly intermediates and hence on the assembly pathways. For the conserved genome organization identified in MS2 and GA earlier [206], this corresponds to an assembly pathway through the most stable intermediates, *i.e.* those forming a maximal number of CP-CP bonds [10].

8.3 Antivirals targeting assembly

It is no simple task to interrupt the PS-mediated assembly of viruses: there is no magic bullet that could interrupt all the PSs equally to attenuate assembly, due to degeneracy in the PSs and their binding affinities for CP. Determining the correct way in which to interrupt or disrupt the assembly system with antivirals is a research goal of the Twarock group at York, in collaboration with the Stockley group at Leeds and Butcher group at the University of Helsinki. An understanding of the different PS, their distribution in the viral genomes,

and their role of controlling the order of assembly is required [11]. 3-D structural information of the asymmetric distribution of PS within the capsid can be used to determine the order of assembly; this approach forms a major part of this thesis. Currently, these findings are being tested in viruses with direct clinical significance: the PS-assembly mechanism has been demonstrated in human viruses. However, more needs to be elucidated on the robustness of the PS assembly mechanisms to determine the optimal strategy for attenuating viral assembly.

Appendix A

STNV paths

Table of possible RNA-CP[1] binding order for STNV assembly, aligned and grouped as discussed in §6.5.5. For each group, the consensus sequence of RNA-CP binding events was calculated by conversion of the consensus sequence of moves.

Group	Consensus moves	Consensus proteins
A	444066X14691X177178146464657X17	
		aduNLs..IJGoFE..ZY3RQSxwyXU0BClm..feh
		aduNLs pIJGoFE 1ZY3RQSxwyXU0BClm bfeh
		aduNLs pIJGoFE 1ZY3RQSxwyXU0BClmqbfeh
		aduNLs pIJGoFE 1ZY3RQSxwyXU0BClmq feh
		aduNLs pIJGoFED1ZY3RQSxwyXU0BClm bfeh
		aduNLs pIJGoFED1ZY3RQSxwyXU0BClmqbfeh
		aduNLs pIJGoFED1ZY3RQSxwyXU0BClmq feh
		aduNLs pIJGoFED ZY3RQSxwyXU0BClm bfeh

[1]For all abbreviations see the glossary on page 222.

		aduNLs pIJGoFED ZY3RQSxwyXU0BClmqbfeh
		aduNLs pIJGoFED ZY3RQSxwyXU0BClmq feh
		aduNLsr IJGoFE 1ZY3RQSxwyXU0BClm bfeh
		aduNLsr IJGoFE 1ZY3RQSxwyXU0BClmqbfeh
		aduNLsr IJGoFE 1ZY3RQSxwyXU0BClmq feh
		aduNLsr IJGoFED1ZY3RQSxwyXU0BClm bfeh
		aduNLsr IJGoFED1ZY3RQSxwyXU0BClmqbfeh
		aduNLsr IJGoFED1ZY3RQSxwyXU0BClmq feh
		aduNLsr IJGoFED ZY3RQSxwyXU0BClm bfeh
		aduNLsr IJGoFED ZY3RQSxwyXU0BClmqbfeh
		aduNLsr IJGoFED ZY3RQSxwyXU0BClmq feh
		aduNLsrpIJGoFE 1ZY3RQSxwyXU0BClm bfeh
		aduNLsrpIJGoFE 1ZY3RQSxwyXU0BClmqbfeh
		aduNLsrpIJGoFE 1ZY3RQSxwyXU0BClmq feh
		aduNLsrpIJGoFED1ZY3RQSxwyXU0BClm bfeh
		aduNLsrpIJGoFED1ZY3RQSxwyXU0BClmqbfeh
		aduNLsrpIJGoFED1ZY3RQSxwyXU0BClmq feh
		aduNLsrpIJGoFED ZY3RQSxwyXU0BClm bfeh
		aduNLsrpIJGoFED ZY3RQSxwyXU0BClmqbfeh
		aduNLsrpIJGoFED ZY3RQSxwyXU0BClmq feh
B	40966X1877X18874148187X5481874	
		adNLs..IJ5HE..ZYT2ROPSxwgz..0ACijqnl
		adNLs pIJ5HE 1ZYT2ROPSxwgz V0ACijqnl
		adNLs pIJ5HE 1ZYT2ROPSxwgzX 0ACijqnl
		adNLs pIJ5HE 1ZYT2ROPSxwgzXV0ACijqnl
		adNLs pIJ5HED1ZYT2ROPSxwgz V0ACijqnl

		adNLs pIJ5HED1ZYT2ROPSxwgzX 0ACijqnl
		adNLs pIJ5HED1ZYT2ROPSxwgzXV0ACijqnl
		adNLs pIJ5HED ZYT2ROPSxwgz V0ACijqnl
		adNLs pIJ5HED ZYT2ROPSxwgzX 0ACijqnl
		adNLs pIJ5HED ZYT2ROPSxwgzXV0ACijqnl
		adNLsr IJ5HE 1ZYT2ROPSxwgz V0ACijqnl
		adNLsr IJ5HE 1ZYT2ROPSxwgzX 0ACijqnl
		adNLsr IJ5HE 1ZYT2ROPSxwgzXV0ACijqnl
		adNLsr IJ5HED1ZYT2ROPSxwgz V0ACijqnl
		adNLsr IJ5HED1ZYT2ROPSxwgzX 0ACijqnl
		adNLsr IJ5HED1ZYT2ROPSxwgzXV0ACijqnl
		adNLsr IJ5HED ZYT2ROPSxwgz V0ACijqnl
		adNLsr IJ5HED ZYT2ROPSxwgzX 0ACijqnl
		adNLsr IJ5HED ZYT2ROPSxwgzXV0ACijqnl
		adNLsrpIJ5HE 1ZYT2ROPSxwgz V0ACijqnl
		adNLsrpIJ5HE 1ZYT2ROPSxwgzX 0ACijqnl
		adNLsrpIJ5HE 1ZYT2ROPSxwgzXV0ACijqnl
		adNLsrpIJ5HED1ZYT2ROPSxwgz V0ACijqnl
		adNLsrpIJ5HED1ZYT2ROPSxwgzX 0ACijqnl
		adNLsrpIJ5HED1ZYT2ROPSxwgzXV0ACijqnl
		adNLsrpIJ5HED ZYT2ROPSxwgz V0ACijqnl
		adNLsrpIJ5HED ZYT2ROPSxwgzX 0ACijqnl
		adNLsrpIJ5HED ZYT2ROPSxwgzXV0ACijqnl
C	40966X1417X17214148187865481874	
		adNLs..IJ5HE..ZY2QROPSxwgzU0ACijqnl
		adNLs pIJGHE 1ZY2QROPSxwgzU0ACijqnl

adNLs pIJGHED1ZY2QROPSxwgzU0ACijqnl

adNLs pIJGHED ZY2QROPSxwgzU0ACijqnl

adNLsr IJGHE 1ZY2QROPSxwgzU0ACijqnl

adNLsr IJGHED1ZY2QROPSxwgzU0ACijqnl

adNLsr IJGHED ZY2QROPSxwgzU0ACijqnl

adNLsrpIJGHE 1ZY2QROPSxwgzU0ACijqnl

adNLsrpIJGHED1ZY2QROPSxwgzU0ACijqnl

adNLsrpIJGHED ZY2QROPSxwgzU0ACijqnl

D 4077877X1877X614877X14641481414

advwgzA..ijqnp..sNMLR72..TSWV01ZEFHG5

---

advwgzAC ijqnpIKsNMLR72 UTSWV01ZEFHG5

advwgzAC ijqnpIKsNMLR72Y TSWV01ZEFHG5

advwgzAC ijqnpIKsNMLR72YUTSWV01ZEFHG5

advwgzAC ijqnpI sNMLR72 UTSWV01ZEFHG5

advwgzAC ijqnpI sNMLR72Y TSWV01ZEFHG5

advwgzAC ijqnpI sNMLR72YUTSWV01ZEFHG5

advwgzAC ijqnp KsNMLR72 UTSWV01ZEFHG5

advwgzAC ijqnp KsNMLR72Y TSWV01ZEFHG5

advwgzAC ijqnp KsNMLR72YUTSWV01ZEFHG5

advwgzACkijqnpIKsNMLR72 UTSWV01ZEFHG5

advwgzACkijqnpIKsNMLR72Y TSWV01ZEFHG5

advwgzACkijqnpIKsNMLR72YUTSWV01ZEFHG5

advwgzACkijqnpI sNMLR72 UTSWV01ZEFHG5

advwgzACkijqnpI sNMLR72Y TSWV01ZEFHG5

advwgzACkijqnpI sNMLR72YUTSWV01ZEFHG5

advwgzACkijqnp KsNMLR72 UTSWV01ZEFHG5

advwgzACkijqnp KsNMLR72Y TSWV01ZEFHG5

advwgzACkijqnp KsNMLR72YUTSWV01ZEFHG5

advwgzA kijqnpIKsNMLR72 UTSWV01ZEFHG5

advwgzA kijqnpIKsNMLR72Y TSWV01ZEFHG5

advwgzA kijqnpIKsNMLR72YUTSWV01ZEFHG5

advwgzA kijqnpI sNMLR72 UTSWV01ZEFHG5

advwgzA kijqnpI sNMLR72Y TSWV01ZEFHG5

advwgzA kijqnpI sNMLR72YUTSWV01ZEFHG5

advwgzA kijqnp KsNMLR72 UTSWV01ZEFHG5

advwgzA kijqnp KsNMLR72Y TSWV01ZEFHG5

advwgzA kijqnp KsNMLR72YUTSWV01ZEFHG5

E 077877X1877X1414877X14641481414

acfmil..opKJ6..MNuvPWTV..zygACD14532Y

---

acfmilFHopKJ6 OMNuvPWTV0BzygACD14532Y

acfmilFHopKJ6 OMNuvPWTV0 zygACD14532Y

acfmilFHopKJ6 OMNuvPWTV BzygACD14532Y

acfmilFHopKJ6R MNuvPWTV0BzygACD14532Y

acfmilFHopKJ6R MNuvPWTV0 zygACD14532Y

acfmilFHopKJ6R MNuvPWTV BzygACD14532Y

acfmilFHopKJ6ROMNuvPWTV0BzygACD14532Y

acfmilFHopKJ6ROMNuvPWTV0 zygACD14532Y

acfmilFHopKJ6ROMNuvPWTV BzygACD14532Y

acfmilF opKJ6 OMNuvPWTV0BzygACD14532Y

acfmilF opKJ6 OMNuvPWTV0 zygACD14532Y

acfmilF opKJ6 OMNuvPWTV BzygACD14532Y

acfmilF opKJ6R MNuvPWTV0BzygACD14532Y

acfmilF opKJ6R MNuvPWTV0 zygACD14532Y

acfmilF opKJ6R MNuvPWTV BzygACD14532Y

acfmilF opKJ6ROMNuvPWTV0BzygACD14532Y

acfmilF opKJ6ROMNuvPWTV0 zygACD14532Y

acfmilF opKJ6ROMNuvPWTV BzygACD14532Y

acfmil HopKJ6 OMNuvPWTV0BzygACD14532Y

acfmil HopKJ6 OMNuvPWTV0 zygACD14532Y

acfmil HopKJ6 OMNuvPWTV BzygACD14532Y

acfmil HopKJ6R MNuvPWTV0BzygACD14532Y

acfmil HopKJ6R MNuvPWTV0 zygACD14532Y

acfmil HopKJ6R MNuvPWTV BzygACD14532Y

acfmil HopKJ6ROMNuvPWTV0BzygACD14532Y

acfmil HopKJ6ROMNuvPWTV0 zygACD14532Y

acfmil HopKJ6ROMNuvPWTV BzygACD14532Y

F 170877X1877X1414877X14641481414

abtuMP..xwgzA..ijmnlHE5..JIK6RQ2UV01Z

---

abtuMPSWxwgzAC ijmnlHE537JIK6RQ2UV01Z

abtuMPSWxwgzAC ijmnlHE53 JIK6RQ2UV01Z

abtuMPSWxwgzAC ijmnlHE5 7JIK6RQ2UV01Z

abtuMPSWxwgzACkijmnlHE537JIK6RQ2UV01Z

abtuMPSWxwgzACkijmnlHE53 JIK6RQ2UV01Z

abtuMPSWxwgzACkijmnlHE5 7JIK6RQ2UV01Z

abtuMPSWxwgzA kijmnlHE537JIK6RQ2UV01Z

abtuMPSWxwgzA kijmnlHE53 JIK6RQ2UV01Z

abtuMPSWxwgzA kijmnlHE5 7JIK6RQ2UV01Z

abtuMPS xwgzAC ijmnlHE537JIK6RQ2UV01Z

abtuMPS xwgzAC ijmnlHE53 JIK6RQ2UV01Z

abtuMPS xwgzAC ijmnlHE5 7JIK6RQ2UV01Z

abtuMPS xwgzACkijmnlHE537JIK6RQ2UV01Z

abtuMPS xwgzACkijmnlHE53 JIK6RQ2UV01Z

abtuMPS xwgzACkijmnlHE5 7JIK6RQ2UV01Z

abtuMPS xwgzA kijmnlHE537JIK6RQ2UV01Z

abtuMPS xwgzA kijmnlHE53 JIK6RQ2UV01Z

abtuMPS xwgzA kijmnlHE5 7JIK6RQ2UV01Z

abtuMP WxwgzAC ijmnlHE537JIK6RQ2UV01Z

abtuMP WxwgzAC ijmnlHE53 JIK6RQ2UV01Z

abtuMP WxwgzAC ijmnlHE5 7JIK6RQ2UV01Z

abtuMP WxwgzACkijmnlHE537JIK6RQ2UV01Z

abtuMP WxwgzACkijmnlHE53 JIK6RQ2UV01Z

abtuMP WxwgzACkijmnlHE5 7JIK6RQ2UV01Z

abtuMP WxwgzA kijmnlHE537JIK6RQ2UV01Z

abtuMP WxwgzA kijmnlHE53 JIK6RQ2UV01Z

abtuMP WxwgzA kijmnlHE5 7JIK6RQ2UV01Z

G 409617877X18874148187X548181464

adNLKJ5HE..ZYT2ROPSxwgz..0ACijqrpnl

---

adNLKJ5HE 1ZYT2ROPSxwgz V0ACijqrpnl

adNLKJ5HE 1ZYT2ROPSxwgzX 0ACijqrpnl

adNLKJ5HE 1ZYT2ROPSxwgzXV0ACijqrpnl

adNLKJ5HED1ZYT2ROPSxwgz V0ACijqrpnl

adNLKJ5HED1ZYT2ROPSxwgzX 0ACijqrpnl

adNLKJ5HED1ZYT2ROPSxwgzXV0ACijqrpnl

adNLKJ5HED ZYT2ROPSxwgzV 0ACijqrpnl

		adNLKJ5HED ZYT2ROPSxwgzX 0ACijqrpnl
		adNLKJ5HED ZYT2ROPSxwgzXV0ACijqrpnl
H	409617417X17214148187865481814 64	
		adNLKJ5HE..ZY2QROPSxwgzU0ACijqrpnl
		adNLKJGHE 1ZY2QROPSxwgzU0ACijqrpnl
		adNLKJGHED1ZY2QROPSxwgzU0ACijqrpnl
		adNLKJGHED ZY2QROPSxwgzU0ACijqrpnl

Appendix B

MS2 paths

This appendix provides Hamiltonian paths, given in terms of moves, for the MS2[1] CP-RNA assembly consist of cycles and pseudo-cycles [10], see Figure 7.5. The paths, of which there are 66 move order in total (132 protein dimer order due to 2× handedness) [10], are given in the separate tables below.

In move order, move 1 indicates a connection across a two-fold axis, and moves 2 and 3 represent moves around a five-fold axis: the mirror symmetry, due to the degeneracy of moves 2 and 3, results in two protein order paths realized for each move path, but each with opposite handedness. The paths are 56 moves (57 proteins) as three dimers of the five around the first/last 5-fold vertex (containing dimers `a`, `t`, `N`, `u`, and `d`) are connected by small edges, which are ignored in our analysis. Thus each of the paths listed here can be considered as ending and starting with either the cycle or pseudo-cycle organizations given in Figure 7.5. For the protein paths, of which there are two provided for each move order as described, the labelling of asymmetric dimers is shown in Figure 7.6.

For the structural data analysis, we consider each of the 132 protein path solutions transposed to start at each PS location on an asymmetric dimer. Dur-

[1]For all abbreviations see the glossary on page 222.

ing the analysis, 8 of 12 five-fold vertices were selected as start/end locations, as these are near the MP. There are, of course, 5 PS sites at each of these 8 vertices. This provides the total set of path solutions: $132 \times 8 \times 5 = 5,280$. Each path is equally valid to be realised in $5'$ to $3'$ order as in $3'$ to $5'$ order.

B.1 Cycles

There are 37 cycle move paths, which represent 74 paths in terms of protein dimer mappings, due to handedness:

Moves	Proteins
12212213312121312222122122221221331212131222212212122221	
	abfjihAB0UVTSWXzygecwxvPOML6RQ2YZ1DCklFE4537JGHonmqrpIKst
	abqmnoHG532QR67JIprsKLMOPvxWSTVYZ4EFlkCD10UXzBAhijfegywcd
121222213122122122131212133122221331221222221221212222121	
	abfecwyghijmqrpIJG54ZYVU01DEFHonlkCABzXWxvPSTQ2376ROMLKst
	abqrsKIponmjfegyzB01ZY2354EDCAhiklFHGJ76LMORQTVUXWSPvxwcd
133331222213333122221333312212222122131312121331333313 31	
	abqmjfecwyghikCABzXU01DE4ZYVTSWxvPOMLKIJ76RQ235GHFlnoprst
	abfjmqrsKIponlFHGJ7354ED1ZY2QR6LMOPvxwyzXWSTVU0BACkihgecd
122121222212222122122133121213122221221222221221331212131	
	abfjihgecwyzXU0BAClkFHGJ76RQ2354ED1ZYVTSWxvPOMLKIponmqrst
	abqmnoprsKIJ735GHFlkCABzXWSTVU01DE4ZY2QR6LMOPvxwyghijfecd
122122133133133331213121312222122122133133133333121312131	
	abfjihAB0UVY235GJ76RQTSWXzygecwxvPOMLKIpoHFE4Z1DCklnmqrst
	abqmnoHG532YVU0BzXWSTQR67JIprsKLMOPvxwyghACD1Z4EFlkijfecd
121221221221331221331331331313333121331213121312131213122131	
	abfecwxvPOMLKIponlkCAB0UVY2354Z1DEFHGJ76RQTSWXzyghijmqrst

abqrsKLMOPvxwyghiklFHG532YVU01Z4EDCABzXWSTQR67JIponmjfecd

12222122121222213122121221331222213312121312222122122121

abfjmqrpIJGHonlFE45376RQ2YZ1DCkihAB0UVTSWXzygecwxvPOMLKst

abqmjfegyzBAhikCD10UXWSTVYZ4EFlnoHG532QR67JIprsKLMOPvxwcd

12222133331222213333122122221221313121213313333133122221

abfjmqrsKIponlFHGJ7354EDCkihABzXWSTVU01ZY2QR6LMOPvxwygecd

abqmjfecwyghikCABzXU01DEFlnoHGJ76RQ2354ZYVTSWxvPOMLKIprst

12122221212222131221221221312121331222213312212222122121

abfecwyghABzXU01DCkijmqrpIJ76RQ235GHonlFE4ZYVTSWxvPOMLKst

abqrsKIpoHGJ7354EFlnmjfegyzXWSTVU0BAhikCD1ZY2QR6LMOPvxwcd

13312213333122131221221221313121312131212133133133121331

abqmnlFE4Z1DCkijfecwxvPOMLKIJ76RQTSWXzyghAB0UVY235GHoprst

abfjikCD1Z4EFlnmqrsKLMOPvxwyzXWSTQR67JIpoHG532YVU0BAhgecd

13312121312222122121222212122221221312222122122122122133 1

abqmnlkijfecwyghACD10BzXUVTSWxvPOMLKIJG5376RQ2YZ4EFHoprst

abfjiklnmqrsKIpoHFE45GJ732QR6LMOPvxwyzB0UXWSTVYZ1DCAhgecd

13312212122131222212122221221221212222122222122131212133 1

abqmnlFEDCkijfecwyghABzXU01Z45376RQ2YVTSWxvPOMLKIJGHoprst

abfjikCDEFlnmqrsKIpoHGJ7354Z10UXWSTVY2QR6LMOPvxwyzBAhgecd

12213131212133133331331222213333122221333312222133331221

abfjiklnmqrpoHFE4Z1DCAhgecwyzB0UXWxvPSTVY2QROML6735GJIKst

abqmnlkijfeghACD1Z4EFHoprsKIJG5376LMORQ2YVTSPvxWXU0Bzywcd

12213122221212212212213312222133121213122221221212222121

abfjiklFHonmqrpIJG54EDCAhgecwyzB01ZYVUXWxvPSTQ2376ROMLKst

abqmnlkCAhijfegyzB01DEFHoprsKIJG54ZY2376LMORQTVUXWSPvxwcd

12222122122133121213122221221222212213312121312222122121

abfjmqrpIJG54Z10BACDEFHonlkihgecwyzXUVY2376RQTSWxvPOMLKst

abqmjfegyzB01Z45GHFEDCAhiklnoprsKIJ732YVUXWSTQR6LMOPvxwcd

1221333312222133133331213122221221333312222133133331213121

abfjikCAhgecwyzB01DE4ZYVUXWxvPSTQ235GJ76ROMLKIpoHFlnmqrst

abqmnlFHoprsKIJG54ED1ZY2376LMORQTVU0BzXWSPvxwyghACkijfecd

1312212212213121213312222133122122221221212222212122222121

abqrpIJG54EDCABzyghiklFHonmjfecwxvPSWXU01ZYVTQ2376ROMLKst

abfegyzB01DEFHGJIponlkCAhijmqrsKLMOR67354ZY2QTVUXWSPvxwcd

1222213313333121312222122133331222213313333121312222122 1221

abfjmqrsKLMOR67JIponlFHG532QTVYZ4ED10UXWSPvxwyzBACkihgecd

abqmjfecwxvPSWXzyghikCAB0UVTQ2YZ1DE45376ROMLKIJGHFlnoprst

1221331333312131222212213313333121312222122133133331213121

abfjikCDE4Z10BAhgecwyzXUVY235GJ76RQTSWxvPOMLKIpoHFlnmqrst

abqmnlFED1Z45GHoprsKIJ732YVU0BzXWSTQR6LMOPvxwyghACkijfecd

1221221212222122221221212222131221212213312222133121213 1

abfjihABzygecwxvPSWXU01ZYVTQ2354EDCklFHGJ76ROMLKIponmqrst

abqmnoHGJIprsKLMOR67354ZY2QTVU01DEFlkCABzXWSPvxwyghijfecd

1221222212213312121312222122121222122222122122133121213 1

abfjihACklFHGJ76RQ2354ED1ZYVTSWXU0BzygecwxvPOMLKIponmqrst

abqmnoHFlkCABzXWSTVU01DE4ZY2QR6735GJIprsKLMOPvxwyghijfecd

1221212222131221212213312222133121213122221221221312122221

abfjihgecwyzBACklFED10UXWxvPSTVYZ4532QROML67JGHonmqrpIKst

abqmnoprsKIJGHFlkCDE45376LMORQ2YZ10UVTSPvxWXzBAhijfegywcd

1331221222212212122221212222121222213122122122131212133 1

abqmnlFED1Z45376RQ2YVTSWXU0BzyghACkijfecwxvPOMLKIJGHoprst

abfjikCDE4Z10UXWSTVY2QR6735GJIpoHFlnmqrsKLMOPvxwyzBAhgecd

1222213313313312131212133131331221212213131213122122122 1

abfjmqrsKLMOPvxwyzBACD10UXWSTVYZ4EFHG532QR67JIponlkihgecd

abqmjfecwxvPOMLKIJGHFE45376RQ2YZ1DCAB0UVTSWXzyghiklnoprst

12222122131212133122221331221212213122221212222122122121

abfjmqrpIJ76RQ235GHonlFE4ZYVTSWXU01DCkihABzygecwxvPOMLKst

abqmjfegyzXWSTVU0BAhikCD1ZY2QR67354EFlnoHGJIprsKLMOPvxwcd

12212212122221222212213121213312222133122121221312222121

abfjihABzygecwxvPSWXU01DCklFE4ZYVTQ235GHonmqrpIJ76ROMLKst

abqmnoHGJIprsKLMOR67354EFlkCD1ZY2QTVU0BAhijfegyzXWSPvxwcd

12212122221212222121222213122122122131212133122221331221

abfjihgecwyzXWxvPSTQROML6732YVU0BACklnmqrpoHFED1Z45GJIKst

abqmnoprsKIJ76LMORQTSPvxWXUVY235GHFlkijfeghACDE4Z10Bzywcd

12212122221212222122131222212122122122133122221331212131

abfjihgecwyzXWxvPSTQ2354ED1ZYVU0BAClFHGJ76ROMLKIponmqrst

abqmnoprsKIJ76LMORQTVU01DE4ZY235GHFlkCABzXWSPvxwyghijfecd

12213312121312222122121222212222122122133121213122221221

abfjikCD10BAhgecwyzXUVTSWxvPOML6RQ2YZ4EFlnmqrpoHG537JIKst

abqmnlFE45GHoprsKIJ732QR6LMOPvxWSTVYZ1DCkijfeghAB0UXzywcd

12122122122133122221331212131222212212122221212222122131

abfecwxvPOMLKIponlFHGJ76RQ2354ED1ZYVTSWXU0BzyghACkijmqrst

abqrsKLMOPvxwyghikCABzXWSTVU01DE4ZY2QR6735GJIpoHFlnmjfecd

12122221221312222121221221221331222213312121312222122121

abfecwyghACDEFHonlkijmqrpIJG54Z10BzXUVY2376RQTSWxvPOMLKst

abqrsKIpoHFEDCAhiklnmjfegyzB01Z45GJ732YVUXWSTQR6LMOPvxwcd

12222133331221222212213131212133133331331222213333122221

abfjmqrsKIponlkihACD1ZY2354EFHGJ76LMORQTVU0BzXWSPvxwygecd

abqmjfecwyghiklnoHFE4ZYVU01DCABzXWxvPSTQ235GJ76ROMLKIprst

13122121221331222213312121312222122122121222212222122121

abqrpIJGHonmjfecwyghiklFEDCABzXU01Z45376RQ2YVTSWxvPOMLKst

abfegyzBAhijmqrsKIponlkCDEFHGJ7354Z10UXWSTVY2QR6LMOPvxwcd

1331212131222212212212122221222212212122213122121221331

abqmnlkijfecwyghACD1ZYVU0BzXWxvPSTQ2376ROMLKIJG54EFHoprst

abfjiklnmqrsKIpoHFE4ZY235GJ76LMORQTVUXWSPvxwyzB01DCAhgecd

1212222122122121222212222122131212133122221331221212213 1

abfecwyghACD1ZYVU0BzXWxvPSTQ2354EFHGJ76ROMLKIponlkijmqrst

abqrsKIpoHFE4ZY235GJ76LMORQTVU01DCABzXWSPvxwyghiklnmjfecd

1221312121331222213312212122131222212122221221221212222 1

abfjiklFEDCAhgecwyzB01Z4532YVUXWxvPSTQROML67JGHonmqrpIKst

abqmnlkCDEFHoprsKIJG54Z10UVY2376LMORQTSPvxWXzBAhijfegywcd

1331331331212131312212122133131331212131312212212213333 1

abqmnoHGJ76RQ2354ZYVTSWXU01DEFlkCABzyghijfecwxvPOMLKIprst

abfjihABzXWSTVU01ZY2QR67354EDCklFHGJIponmqrsKLMOPvxwygecd

B.2 Pseudo-cycles

Pseudo-cycles number 29 move paths, corresponding to 58 protein dimer paths:

Moves	Proteins
1212212213131213121312213333122133133133133333121331331	
	abfecwxvPORQTSWXzyghijmqrsKIponlkCAB0UVY2354Z1DEFHGJ76LMN
	abqrsKLMOPSTQR67JIponmjfecwyghiklFHG532YVU01Z4EDCABzXWxvu
1221221331313312131213333133131333331331213122222122133331	
	abfjihAB0UVYZ1DCklnmqrsKIpoHFE45GJ732QTSWXzygecwxvPOR6LMN
	abqmnoHG532YZ4EFlkijfecwyghACD10BzXUVTQR67JIprsKLMOPSWxvu
1331222213333122122122131222212122221212212222213121213331	
	abqmnlFHoprsKIJG54EDCkijfecwyghABzXU01ZYVTSWxvPORQ2376LMN

abfjikCAhgecwyzB01DEFlnmqrsKIpoHGJ7354ZY2QR6LMOPSTVUXWxvu

12212213313333121312222122133133331313313333312131213313 1

abfjihAB0UVY2QTSWXzygecwxvPOR6735GJIpoHFE4Z1DCklnmqrsKLMN

abqmnoHG532YVTQR67JIprsKLMOPSWXU0BzyghACD1Z4EFlkijfecwxvu

12222133121331212131312212212222121222213122121221331221

abfjmqrsKL67354EFHGJIponlkihgecwyzXWxvPSTVU0BACD1ZY2QROMN

abqmjfecwxWXU01DCABzyghiklnoprsKIJ76LMORQ235GHFE4ZYVTSPvu

12122122133133133331213131221331331333312131213122133131

abfecwxvPOR67354Z1DEFHGJIponlkCAB0UVY2QTSWXzyghijmqrsKLMN

abqrsKLMOPSWXU01Z4EDCABzyghiklFHG532YVTQR67JIponmjfecwxvu

12222133121213122122122221312121331222213312212122131221

abfjmqrsKL67JIponlkihgecwyzBACD10UXWxvPSTVYZ4EFHG532QROMN

abqmjfecwxWXzyghiklnoprsKIJGHFE45376LMORQ2YZ1DCAB0UVTSPvu

12122122133121313222212213312121312222122122122222122133131

abfecwxvPOR67JIponlFHG532QTSWXUVYZ4ED10BzyghACkijmqrsKLMN

abqrsKLMOPSWXzyghikCAB0UVTQR6732YZ1DE45GJIpoHFlnmjfecwxvu

12213313333121312222122133331222213313333131333312133131

abfjikCDE4Z10BAhgecwyzXUVY2QTSWxvPOR6735GJIpoHFlnmqrsKLMN

abqmnlFED1Z45GHoprsKIJ732YVTQR6LMOPSWXU0BzyghACkijfecwxvu

12213313133121333313133331331213122221221333312222133331

abfjikCDEFlnmqrsKIpoHGJ7354Z10BAhgecwyzXUVY2QTSWxvPOR6LMN

abqmnlFEDCkijfecwyghABzXU01Z45GHoprsKIJ732YVTQR6LMOPSWxvu

12122221221312121331222213312213122121221331221312222121

abfecwyghACDEFHG54Z10BzXUVY237JIponlkijmqrsKL6RQTSWxvPOMN

abqrsKIpoHFEDCAB01Z45GJ732YVUXzyghiklnmjfecwxWSTQR6LMOPvu

12213131213312133331313313333133122221333312222133331221

abfjiklnmqrsKL6735GJIpoHFE4Z1DCAhgecwyzB0UXWxvPSTVY2QROMN

abqmnlkijfecwxWXU0BzyghACD1Z4EFHoprsKIJG5376LMORQ2YVTSPvu

1331221333312213122122131312131212133133133121333313133 1

abqmnlFE4Z1DCkijfecwxvPORQTSWXzyghAB0UVY235GHoprsKIJ76LMN

abfjikCD1Z4EFlnmqrsKLMOPSTQR67JIpoHG532YVU0BAhgecwyzXWxvu

1222213333122122122221222212122122122221312131212133133 1

abfjmqrsKIponlkihgecwyzXU0BACD1ZYVTSWxvPORQ2354EFHGJ76LMN

abqmjfecwyghiklnoprsKIJ735GHFE4ZY2QR6LMOPSTVU01DCABzXWxvu

1212222121222212122122221312131221221221333312222133133 1

abfecwyghABzXU01ZYVTSWxvPORQ2354EDCkijmqrsKIponlFHGJ76LMN

abqrsKIpoHGJ7354ZY2QR6LMOPSTVU01DEFlnmjfecwyghikCABzXWxvu

1331313333133133121213131221212213313133121312212213333 1

abqmnoprsKIJ732QTSWXUVYZ10BzyghACDE45GHFlkijfecwxvPOR6LMN

abfjihgecwyzXUVTQR6732YZ45GJIpoHFED10BACklnmqrsKLMOPSWxvu

1222213312213121213312213122122221312121331221222212212 1

abfjmqrsKL6RQTSWXUVY237JIponlkihACDEFHG54Z10BzygecwxvPOMN

abqmjfecwxWSTQR6732YVUXzyghiklnoHFEDCAB01Z45GJIprsKLMOPvu

1222213333122122221221221212222122222122131213121213313 1

abfjmqrsKIponlkihACD1ZYVTSWXU0BzygecwxvPORQ2354EFHGJ76LMN

abqmjfecwyghiklnoHFE4ZY2QR6735GJIprsKLMOPSTVU01DCABzXWxvu

1221222213121331213121213313131331221221222212222121221221 1

abfjihACklnmqrsKL6732YZ45GJIpoHFED10BzygecwxvPSWXUVTQROMN

abqmnoHFlkijfecwxWXUVYZ10BzyghACDE45GJIprsKLMOR6732QTSPvu

1331313333133331331213122221221333312222133331222213333 1

abqmnoprsKIJ735GHFlkijfecwyghACDE4Z10BzXUVY2QTSWxvPOR6LMN

abfjihgecwyzXU0BACklnmqrsKIpoHFED1Z45GJ732YVTQR6LMOPSWxvu

1221212222121222213122122122131312121331213312222133122 1

abfjihgecwyzXWxvPSTVU0BACklFHGJIponmqrsKL67354ED1ZY2QROMN

abqmnoprsKIJ76LMORQ235GHFlkCABzyghijfecwxWXU01DE4ZYVTSPvu

1331221313121333313313312131212133131331221312212213331

abqmnlFE45GHoprsKIJ732QTSWXzyghAB0UVYZ1DCkijfecwxvPOR6LMN

abfjikCD10BAhgecwyzXUVTQR67JIpoHG532YZ4EFlnmqrsKLMOPSWxvu

1313333133133122121221313121213313133122122131213333133 1

abqrsKIpoHFE4ZYVTSWXU01DCABzyghiklnmjfecwxvPORQ235GJ76LMN

abfecwyghACD1ZY2QR67354EFHGJIponlkijmqrsKLMOPSTVU0BzXWxvu

1212212213121312222122121222212122221221333312222133133 1

abfecwxvPORQ2354ED1ZYVTSWXU0BzyghACkijmqrsKIponlFHGJ76LMN

abqrsKLMOPSTVU01DE4ZY2QR6735GJIpoHFlnmjfecwyghikCABzXWxvu

1221212222122131222212122122122133122221331213121213313 1

abfjihgecwyzXUVYZ4ED10BAClFHG532QTSWxvPOR67JIponmqrsKLMN

abqmnoprsKIJ732YZ1DE45GHFlkCAB0UVTQR6LMOPSWXzyghijfecwxvu

1331222213333122122221221212222121222213122122131212133 1

abqmnlFHoprsKIJG54ED1ZYVTSWXU0BzyghACkijfecwxvPORQ2376LMN

abfjikCAhgecwyzB01DE4ZY2QR6735GJIpoHFlnmqrsKLMOPSTVUXWxvu

1221222212212213312121312222122122221221331213121213313 1

abfjihACklFHG532QTSWXUVYZ4ED10BzygecwxvPOR67JIponmqrsKLMN

abqmnoHFlkCAB0UVTQR6732YZ1DE45GJIprsKLMOPSWXzyghijfecwxvu

1313333133331331222213333122221333312212222131213333133 1

abqrsKIpoHFlnmjfecwyghikCABzXU01DE4ZYVTSWxvPORQ235GJ76LMN

abfecwyghACkijmqrsKIponlFHGJ7354ED1ZY2QR6LMOPSTVU0BzXWxvu

1212222122121222213122121221331222213312131221212213313 1

abfecwyghACD10BzXUVYZ4EFHG532QTSWxvPOR67JIponlkijmqrsKLMN

abqrsKIpoHFE45GJ732YZ1DCAB0UVTQR6LMOPSWXzyghiklnmjfecwxvu

Glossary

1-D One-dimensional.

2-D Two-dimensional.

3-D Three-dimensional.

5-D Five-dimensional.

6-D Six-dimensional.

Å Ångström. 10^{-10} m.

AP205 Bacteriophage AP205 (of *Acinetobacter* spp.).

β Thermodynamic beta. The reciprocal of the thermodynamic fundamental temperature of a system.

B3 B3 aptamer of STNV. A 16 nt stem-loop PS, B3 matches the STNV genome across the central 10 nt. It displays the STNV PS motif AxxA, and is shown to bind to STNV CP and allow encapsidation.

bp Base pairs.

BPMV Bean pod mottle virus.

BTV Bluetongue virus.

CLIP-seq Cross-linking and immunoprecipitation with high-throughput sequencing. Identifies binding sites of RNA-binding proteins.

CP Capsid protein.

cryo-EM Cryo transmission electron microscopy. A form of transmission electron microscopy where the sample is studied at cryogenic temperatures, usually in liquid nitrogen.

cryo-ET Cryo electron tomography. An implementation of cryo-EM in which samples are tilted during imaging, producing a series of 2-D micrographs that can be computationally combined into a 3-D reconstruction.

DFT Density functional theory. Computational quantum mechanical modelling of the electronic structure of many-body systems.

DLA Diffusion limited aggregation. Particles in Brownian motion cluster to form disordered aggregates.

DNA Deoxyribonucleic acid.

dsRNA Double-stranded RNA.

E. coli *Escherichia coli.*

$\boldsymbol{E_a}$ Activation energy.

ΔG Gibbs free energy.

GA Bacteriophage GA (of *E. coli*).

GNR Graphene nanoribbon.

ΔH Change in enthalpy.

HBV Hepatitis B virus.

HIV-1 Human immunodeficiency virus type 1.

HPV Human papillomavirus.

HRV-B Human rhinovirus B.

IBDV Infectious bursal disease virus.

$\boldsymbol{k}$ Rate constant.

$\boldsymbol{k_B}$ Boltzmann constant. $1.381 \times 10^{23}\,\mathrm{J\,K^{-1}}$.

$\boldsymbol{K_c}$ Equilibrium constant.

$\boldsymbol{K_d}$ Disassociation constant.

KSA Kinetic self-assembly. An algorithm that simulates self-assembly by generating a network of possible reactions between molecules and firing reactions stochastically, with a probability based on their kinetics.

$\boldsymbol{l}$ Degree of polymerization.

L-A *Saccharomyces cerevisiae* virus L-A.

MC Monte Carlo. Algorithms relying on repeated random sampling.

MD Molecular dynamics. Computer simulation of atomic trajectories by numerical integration of Newton's equation of motion with respect to interatomic potential.

MHV Mouse hepatitis virus.

MP Maturation protein. Single copy protein in bacteriophage MS2, also called A-protein.

MPyV Murine polyomavirus.

mRNA Messenger RNA.

MS2 Bacteriophage MS2 (of *E. coli*).

nt Nucleotides.

PaV Pariacoto virus.

PBCV-1 *Paramecium bursaria* chlorella virus 1.

pgRNA Pregenomic RNA. Single-stranded mRNA, packaged in viruses such as HBV, and serves as a template for reverse transcription and genomic DNA formation within the viral nucleocapsid.

pI Isoelectric point. The pH for which a molecule carries no net electric charge.

PP7 Bacteriophage PP7 (of *Pseudomonas aeruginosa*).

PS Packaging signals. Sequence-specific genomic secondary structures that interact with capsid protein to mediate assembly.

Qβ Bacteriophage Qβ (of *E. coli*).

$\boldsymbol{r}$ Rate of reaction.

RdRp RNA-dependent RNA polymerase.

$\mathbf{R}_h$ Hydrodynamic radius. Derived from smFCS curves, it is the apparent size of the entire dynamic solvated molecule, in this case, the genome in complex with assembling CP.

RNA Ribonucleic acid.

RVFV Rift Valley fever virus.

σ Standard deviation.

ΔS Change in entropy.

SARS Severe acute respiratory syndrome.

SARS-CoV SARS coronavirus. The etiologic agent of SARS.

SELEX Systematic evolution of ligands by exponential enrichment.

SEM Scanning electron microscope.

smFCS Single-molecule fluorescence correlation spectroscopy. Sensitive method of measuring the R_h of a single molecular complex, in this case a viral genome surrounded by CP, to observe the collapse of the genome conformation on the addition of CP.

ssRNA Single-stranded RNA.

STMV Satellite tobacco mosaic virus.

STNV Satellite tobacco necrosis virus.

SV40 Simian virus 40.

T Absolute temperature.

t Time.

***T*-number** Triangulation number. Describes the number of structural subunits per asymmetric unit in a capsid.

TB Tight-binding. A minimal quantum mechanical model of electronic band structure.

TNV Tobacco necrosis virus.

TR Translation repressor. A 19 nt RNA stem-loop PS in the MS2 genome, encompassing the start codon of the viral replicase. Shown to bind to MS2 CP and effect encapsidation.

VLP Virus-like particle. Non-infectious particles of viral structural proteins from a recombinant heterologous system. By definition they do not contain full viral genome, but can sometimes package other nucleic acid, such as the recombinant mRNA [49].

WT Wild-type. The phenotype of the relevant species as typically observed in nature.

References

[1] D. Philp and J. F. Stoddart, "Self-assembly in natural and unnatural systems," *Angewandte Chemie International Edition in English*, vol. 35, no. 11, pp. 1154–1196, 1996.

[2] S. Zhang, "Fabrication of novel biomaterials through molecular self-assembly," *Nature Biotechnology*, vol. 21, no. 10, pp. 1171–1178, 2003.

[3] J. A. Pelesko, *Self assembly: The science of things that put themselves together*. CRC Press, 2007.

[4] M. Gross, *Travels to the nanoworld*. Basic Books, 2008.

[5] J. M. Benyus, *Biomimicry*. William Morrow New York, 1997.

[6] E. Winfree, F. Liu, L. A. Wenzler, and N. C. Seeman, "Design and self-assembly of two-dimensional DNA crystals," *Nature*, vol. 394, no. 6693, pp. 539–544, 1998.

[7] N. C. Seeman, "Molecular craftwork with DNA," in *Culture of Chemistry* (B. Hargittai and I. Hargittai, eds.), pp. 141–152, Springer, 2015.

[8] P. G. Stockley, N. A. Ranson, and R. Twarock, "A new paradigm for the roles of the genome in ssRNA viruses," *Future Virology*, vol. 8, no. 6, pp. 531–543, 2013.

[9] P. G. Stockley, R. Twarock, S. E. Bakker, A. M. Barker, A. Borodavka, E. Dykeman, R. J. Ford, A. R. Pearson, S. E. V. Phillips, N. A. Ranson, and R. Tuma, "Packaging signals in single-stranded RNA viruses: Nature's alternative to a purely electrostatic assembly mechanism," *Journal of Biological Physics*, vol. 39, no. 2, pp. 277–287, 2013.

[10] E. C. Dykeman, N. E. Grayson, K. Toropova, N. A. Ranson, P. G. Stockley, and R. Twarock, "Simple rules for efficient assembly predict the layout of a packaged viral RNA," *Journal of Molecular Biology*, vol. 408, no. 3, pp. 399–407, 2011.

[11] E. C. Dykeman, P. G. Stockley, and R. Twarock, "Solving a Levinthal's paradox for virus assembly identifies a unique antiviral strategy," *Proceedings of the National Academy of Sciences*, vol. 111, no. 14, pp. 5361–5366, 2014.

[12] J. A. Geraets, E. C. Dykeman, P. G. Stockley, N. A. Ranson, and R. Twarock, "Asymmetric genome organization in an RNA virus revealed *via* graph-theoretical analysis of tomographic data," *PLoS Computational Biology*, vol. 11, no. 3, p. e1004146, 2015.

[13] P. F. Damasceno, M. Engel, and S. C. Glotzer, "Predictive self-assembly of polyhedra into complex structures," *Science*, vol. 337, no. 6093, pp. 453–457, 2012.

[14] C. R. Iacovella, A. S. Keys, and S. C. Glotzer, "Self-assembly of soft-matter quasicrystals and their approximants," *Proceedings of the National Academy of Sciences*, vol. 108, no. 52, pp. 20935–20940, 2011.

[15] J. A. Geraets, J. Baldwin, R. Twarock, and Y. Hancock, "A proposed method for directed self-assembly of graphene nanoribbons," in prep.

[16] Y. Xu, G. Shi, and X. Duan, "Self-assembled three-dimensional graphene macrostructures: Synthesis and applications in supercapacitors," *Accounts of Chemical Research*, 2015.

[17] J. Cai, P. Ruffieux, R. Jaafar, M. Bieri, T. Braun, S. Blankenburg, M. Muoth, A. P. Seitsonen, M. Saleh, X. Feng, K. Müllen, and R. Fasel, "Atomically precise bottom-up fabrication of graphene nanoribbons," *Nature*, vol. 466, no. 7305, pp. 470–473, 2010.

[18] M. Fuechsle, J. A. Miwa, S. Mahapatra, H. Ryu, S. Lee, O. Warschkow, L. C. L. Hollenberg, G. Klimeck, and M. Y. Simmons, "A single-atom transistor," *Nature Nanotechnology*, vol. 7, no. 4, pp. 242–246, 2012.

[19] S. E. Bakker, R. J. Ford, A. M. Barker, J. Robottom, K. Saunders, A. R. Pearson, N. A. Ranson, and P. G. Stockley, "Isolation of an asymmetric RNA uncoating intermediate for a single-stranded RNA plant virus," *Journal of Molecular Biology*, vol. 417, no. 1, pp. 65–78, 2012.

[20] P. Cermelli, G. Indelicato, and R. Twarock, "Nonicosahedral pathways for capsid expansion," *Physical Review E*, vol. 88, no. 3, p. 032710, 2013.

[21] G. Cardone, R. L. Duda, N. Cheng, L. You, J. F. Conway, R. W. Hendrix, and A. C. Steven, "Metastable intermediates as stepping stones on the maturation pathways of viral capsids," *MBio*, vol. 5, no. 6, pp. e02067–14, 2014.

[22] J. W. Drake and J. J. Holland, "Mutation rates among RNA viruses," *Proceedings of the National Academy of Sciences*, vol. 96, no. 24, pp. 13910–13913, 1999.

[23] R. J. Bingham, E. C. Dykeman, and R. Twarock, "*In silico* models of viral evolution reveal mechanisms underpinning stability and cooperativity in a viral quasi species," in prep.

[24] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The Journal of Physical Chemistry*, vol. 81, no. 25, pp. 2340–2361, 1977.

[25] D. T. Gillespie, "Stochastic simulation of chemical kinetics," *Annual Review of Physical Chemistry*, vol. 58, pp. 35–55, 2007.

[26] M. D. Nussinov, V. A. Otroshchenko, and S. Santoli, "The emergence of the non-cellular phase of life on the fine-grained clayish particles of the early Earth's regolith," *BioSystems*, vol. 42, no. 2, pp. 111–118, 1997.

[27] G. M. Weinberg, *An introduction to general systems thinking*. Wiley New York, 1975.

[28] I. V. Blauberg, V. N. Sadovskii, and B. G. Iudin, "Philosophical principles of systemicity and the systems approach," *Soviet Studies in Philosophy*, vol. 17, no. 4, pp. 44–68, 1979.

[29] G. Gorelik, "Bogdanov's tektology: Its nature, development and influence," *Studies in East European Thought*, vol. 26, no. 1, pp. 39–57, 1983.

[30] S. E. Wallis, "Abstraction and insight: Building better conceptual systems to support more effective social change," *Foundations of Science*, vol. 19, no. 4, pp. 353–362, 2014.

[31] H. A. Simon, "The architecture of complexity," *Proceedings of the American Philosophical Society*, vol. 106, no. 6, pp. 467–482, 1962.

[32] B. Allen, B. C. Stacey, and Y. Bar-Yam, "An information-theoretic formalism for multiscale structure in complex systems," *arXiv preprint arXiv:1409.4708*, 2014.

[33] H. A. Simon, "Can there be a science of complex systems?," in *Unifying themes in complex systems, vol. 1: Proceedings from the first international conference on complex systems* (Y. Bar-Yam, ed.), pp. 3–14, New England Complex Systems Institute, Westview Press, 2000.

[34] D. H. J. Bunka, S. W. Lane, C. L. Lane, E. C. Dykeman, R. J. Ford, A. M. Barker, R. Twarock, S. E. V. Phillips, and P. G. Stockley, "Degenerate RNA packaging signals in the genome of satellite tobacco necrosis virus: Implications for the assembly of a $T = 1$ capsid," *Journal of Molecular Biology*, vol. 413, no. 1, pp. 51–65, 2011.

[35] S. H. E. van den Worm, R. I. Koning, H. J. Warmenhoven, H. K. Koerten, and J. van Duin, "Cryo electron microscopy reconstructions of the *Leviviridae* unveil the densest icosahedral RNA packing possible," *Journal of Molecular Biology*, vol. 363, no. 4, pp. 858–865, 2006.

[36] K. Toropova, G. Basnak, R. Twarock, P. G. Stockley, and N. A. Ranson, "The three-dimensional structure of genomic RNA in bacteriophage MS2: Implications for assembly," *Journal of Molecular Biology*, vol. 375, no. 3, pp. 824–836, 2008.

[37] J. Seitsonen, P. Susi, O. Heikkilä, R. S. Sinkovits, P. Laurinmäki, T. Hyypiä, and S. J. Butcher, "Interaction of $\alpha_{\mathrm{v}}\beta_3$ and $\alpha_{\mathrm{v}}\beta_6$ integrins with human parechovirus 1," *Journal of Virology*, vol. 84, no. 17, pp. 8509–8519, 2010.

[38] Z. G. Chen, C. Stauffacher, Y. Li, T. Schmidt, W. Bomu, G. Kamer, M. Shanks, G. Lomonossoff, and J. E. Johnson, "Protein-RNA interactions in an icosahedral virus at 3.0Å resolution," *Science*, vol. 245, no. 4914, pp. 154–159, 1989.

[39] J. M. Fox, G. Wang, J. A. Speir, N. H. Olson, J. E. Johnson, T. S. Baker, and M. J. Young, "Comparison of the native CCMV virion with *in vitro* assembled CCMV virions by cryoelectron microscopy and image reconstruction," *Virology*, vol. 244, no. 1, pp. 212–218, 1998.

[40] J. A. Speir, S. Munshi, G. Wang, T. S. Baker, and J. E. Johnson, "Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy," *Structure*, vol. 3, no. 1, pp. 63–78, 1995.

[41] Y. G. Choi and A. L. N. Rao, "Packaging of brome mosaic virus RNA3 is mediated through a bipartite signal," *Journal of Virology*, vol. 77, no. 18, pp. 9750–9757, 2003.

[42] A. J. Fisher and J. E. Johnson, "Ordered duplex RNA controls capsid architecture in an icosahedral animal virus," *Nature*, vol. 361, no. 6408, pp. 176–179, 1993.

[43] D. Zimmern and P. J. G. Butler, "The isolation of tobacco mosaic virus RNA fragments containing the origin for viral assembly," *Cell*, vol. 11, no. 3, pp. 455–462, 1977.

[44] P. Butler and A. Klug, "Assembly of the particle of tobacco mosaic virus from RNA and disks of protein," *Nature*, vol. 229, no. 2, pp. 47–50, 1971.

[45] F. Qu and T. J. Morris, "Encapsidation of turnip crinkle virus is defined by a specific packaging signal and RNA size," *Journal of Virology*, vol. 71, no. 2, pp. 1428–1435, 1997.

[46] B. Böttcher and R. A. Crowther, "Difference imaging reveals ordered regions of RNA in turnip yellow mosaic virus," *Structure*, vol. 4, no. 4, pp. 387–394, 1996.

[47] G. A. Bentley, A. Lewit-Bentley, L. Liljas, U. Skoglund, M. Roth, and T. Unge, "Structure of RNA in satellite tobacco necrosis virus: A low resolution neutron diffraction study using $^{1}H_2O$ $^{2}H_2O$ solvent contrast variation," *Journal of Molecular Biology*, vol. 194, no. 1, pp. 129–141, 1987.

[48] S. B. Larson, J. Day, A. Greenwood, and A. McPherson, "Refined structure of satellite tobacco mosaic virus at 1.8Å resolution," *Journal of Molecular Biology*, vol. 277, no. 1, pp. 37–59, 1998.

[49] S. W. Lane, C. A. Dennis, C. L. Lane, C. H. Trinh, P. J. Rizkallah, P. G. Stockley, and S. E. V. Phillips, "Construction and crystal structure of recombinant STNV capsids," *Journal of Molecular Biology*, vol. 413, no. 1, pp. 41–50, 2011.

[50] J. A. Fosmire, K. Hwang, and S. Makino, "Identification and characterization of a coronavirus packaging signal," *Journal of Virology*, vol. 66, no. 6, pp. 3522–3530, 1992.

[51] S. Murakami, K. Terasaki, K. Narayanan, and S. Makino, "Roles of the coding and noncoding regions of Rift Valley fever virus RNA genome segments in viral RNA packaging," *Journal of Virology*, vol. 86, no. 7, pp. 4034–4039, 2012.

[52] P.-K. Hsieh, S. C. Chang, C.-C. Huang, T.-T. Lee, C.-W. Hsiao, Y.-H. Kou, I.-Y. Chen, C.-K. Chang, T.-H. Huang, and M.-F. Chang, "Assembly of severe acute respiratory syndrome coronavirus RNA packaging signal into virus-like particles is nucleocapsid dependent," *Journal of Virology*, vol. 79, no. 22, pp. 13848–13855, 2005.

[53] A. Lever, H. Gottlinger, W. Haseltine, and J. Sodroski, "Identification of a sequence required for efficient packaging of human immunodeficiency virus type 1 RNA into virions," *Journal of Virology*, vol. 63, no. 9, pp. 4085–4087, 1989.

[54] K. A. Connors, *Chemical kinetics: The study of reaction rates in solution.* John Wiley & Sons, 1990.

[55] E. Hasselbrink and B. I. Lundqvist, *Handbook of surface science, vol. 3: Dynamics.* Elsevier, Amsterdam, 2008.

[56] M. R. Wright, *Fundamental chemical kinetics: An explanatory introduction to the concepts.* Elsevier, 1999.

[57] J. Halley and D. A. Winkler, "Consistent concepts of self-organization and self-assembly," *Complexity*, vol. 14, no. 2, pp. 10–17, 2008.

[58] J. Gerhart and M. Kirschner, *Cells, embryos, and evolution: Toward a cellular and developmental understanding of phenotypic variation and evolutionary adaptability.* Blackwell, 1997.

[59] R. F. Bruinsma, W. M. Gelbart, D. Reguera, J. Rudnick, and R. Zandi, "Viral self-assembly as a thermodynamic process," *Physical Review Letters*, vol. 90, no. 24, p. 248101, 2003.

[60] R. Gutzler, L. Cardenas, and F. Rosei, "Kinetics and thermodynamics in surface-confined molecular self-assembly," *Chemical Science*, vol. 2, no. 12, pp. 2290–2300, 2011.

[61] G. G. Hammes, *Thermodynamics and kinetics for the biological sciences*. Wiley-Interscience New York, 2000.

[62] G. M. Whitesides, J. P. Mathias, and C. T. Seto, "Molecular self-assembly and nanochemistry: A chemical strategy for the synthesis of nanostructures," tech. rep., DTIC Document, 1991.

[63] M. F. Hagan, O. M. Elrad, and R. L. Jack, "Mechanisms of kinetic trapping in self-assembly and phase transformation," *The Journal of Chemical Physics*, vol. 135, no. 10, p. 104115, 2011.

[64] G. M. Whitesides and M. Boncheva, "Beyond molecules: Self-assembly of mesoscopic and macroscopic components," *Proceedings of the National Academy of Sciences*, vol. 99, no. 8, pp. 4769–4774, 2002.

[65] R. Zan, Q. M. Ramasse, U. Bangert, and K. S. Novoselov, "Graphene reknits its holes," *Nano Letters*, vol. 12, no. 8, pp. 3936–3940, 2012.

[66] E. C. Dykeman, P. G. Stockley, and R. Twarock, "Building a viral capsid in the presence of genomic RNA," *Physical Review E*, vol. 87, no. 2, p. 022717, 2013.

[67] R. Schiller, "A derivation of entropy of mixing," *Radiation Physics and Chemistry (1977)*, vol. 23, no. 1, pp. 25–27, 1984.

[68] P. J. Flory, "Thermodynamics of heterogeneous polymers and their solutions," *The Journal of Chemical Physics*, vol. 12, no. 11, pp. 425–438, 1944.

[69] T. E. Ouldridge, A. A. Louis, and J. P. K. Doye, "Extracting bulk properties of self-assembling systems from small simulations," *Journal of Physics: Condensed Matter*, vol. 22, no. 10, p. 104102, 2010.

[70] M. O. Vlad, J. Ross, and D. L. Huber, "Linear free energy relations and reversible stretched exponential kinetics in systems with static or dynamical disorder," *The Journal of Physical Chemistry B*, vol. 103, no. 9, pp. 1563–1580, 1999.

[71] P. J. Flory, "Thermodynamics of high polymer solutions," *The Journal of Chemical Physics*, vol. 10, no. 1, pp. 51–61, 1942.

[72] L. A. Rodriguez-Guadarrama, S. K. Talsania, K. K. Mohanty, and R. Rajagopalan, "Thermodynamics of aggregation of amphiphiles in solution from lattice Monte Carlo simulations," *Langmuir*, vol. 15, no. 2, pp. 437–446, 1999.

[73] A. Zlotnick, "To build a virus capsid: An equilibrium model of the self assembly of polyhedral protein complexes," *Journal of Molecular Biology*, vol. 241, no. 1, pp. 59–67, 1994.

[74] T. Kato, N. Mizoshita, and K. Kishimoto, "Functional liquid-crystalline assemblies: Self-organized soft materials," *Angewandte Chemie International Edition*, vol. 45, no. 1, pp. 38–68, 2006.

[75] T. Aida, E. W. Meijer, and S. I. Stupp, "Functional supramolecular polymers," *Science*, vol. 335, no. 6070, pp. 813–817, 2012.

[76] S. I. Stupp and L. C. Palmer, "Supramolecular chemistry and self-assembly in organic materials design," *Chemistry of Materials*, vol. 26, no. 1, pp. 507–518, 2013.

[77] P. J. Flory, "Molecular size distribution in ethylene oxide polymers," *Journal of the American Chemical Society*, vol. 62, no. 6, pp. 1561–1565, 1940.

[78] G. J. P. Britovsek, S. A. Cohen, V. C. Gibson, P. J. Maddox, and M. van Meurs, "Iron-catalyzed polyethylene chain growth on zinc: Linear α-olefins with a Poisson distribution," *Angewandte Chemie International Edition*, vol. 41, no. 3, pp. 489–491, 2002.

[79] J. A. Keith and P. M. Henry, "The mechanism of the Wacker reaction: A tale of two hydroxypalladations," *Angewandte Chemie International Edition*, vol. 48, no. 48, pp. 9038–9049, 2009.

[80] A. A. C. Braga, G. Ujaque, and F. Maseras, "A DFT study of the full catalytic cycle of the Suzuki-Miyaura cross-coupling on a model system," *Organometallics*, vol. 25, no. 15, pp. 3647–3658, 2006.

[81] O. Demin and I. Goryanin, *Kinetic modelling in systems biology.* CRC Press, 2008.

[82] L. Michaelis and M. L. Menten, "Die kinetik der invertinwirkung," *Biochemische Zeitschrift*, vol. 49, no. 333-369, p. 352, 1913.

[83] J. Hassan, M. Sévignon, C. Gozzi, E. Schulz, and M. Lemaire, "Aryl-aryl bond formation one century after the discovery of the ullmann reaction," *Chemical Reviews*, vol. 102, pp. 1359–470, May 2002.

[84] P. L. Houston, *Chemical Kinetics and Reaction Dynamics.* McGraw-Hill, 2001.

[85] G. Rothenberg, *Catalysis: Concepts and green applications.* Wiley VCH, 2008.

[86] M. J. Pilling and P. W. Seakins, *Reaction kinetics*. Oxford University Press, 1996.

[87] V. P. Zhdanov, *Elementary physicochemical processes on solid surfaces*. Springer Science & Business Media, 1991.

[88] E. Vanden-Eijnden and F. A. Tal, "Transition state theory: Variational formulation, dynamical corrections, and error estimates," *The Journal of Chemical Physics*, vol. 123, no. 18, p. 184103, 2005.

[89] J. Pineda and S. Schwartz, "Protein dynamics and catalysis: The problems of transition state theory and the subtlety of dynamic control," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 361, no. 1472, pp. 1433–1438, 2006.

[90] T. R. Waite, "Theoretical treatment of the kinetics of diffusion-limited reactions," *Physical Review*, vol. 107, no. 2, p. 463, 1957.

[91] V. P. Zhdanov, "Dynamics of surface diffusion," *Surface Science*, vol. 214, no. 1, pp. 289–303, 1989.

[92] S. A. Rice, *Diffusion-limited reactions*. Elsevier, 1985.

[93] T. A. Witten Jr and L. M. Sander, "Diffusion-limited aggregation, a kinetic critical phenomenon," *Physical Review Letters*, vol. 47, no. 19, p. 1400, 1981.

[94] T. A. Witten and L. M. Sander, "Diffusion-limited aggregation," *Physical Review B*, vol. 27, no. 9, p. 5686, 1983.

[95] C. D. Bain, E. B. Troughton, Y. T. Tao, J. Evall, G. M. Whitesides, and R. G. Nuzzo, "Formation of monolayer films by the spontaneous assem-

bly of organic thiols from solution onto gold," *Journal of the American Chemical Society*, vol. 111, no. 1, pp. 321–335, 1989.

[96] DoITPoMS, "Micrograph no 617: deeply etched co-dendrites," 2002. This image is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 2.0 license.

[97] J. Björk and F. Hanke, "Towards design rules for covalent nanostructures on metal surfaces," *Chemistry: A European Journal*, vol. 20, no. 4, pp. 928–934, 2014.

[98] L. Smykalla, P. Shukrynau, M. Korb, H. Lang, and M. Hietschold, "Surface-confined 2D polymerization of a brominated copper-tetraphenylporphyrin on Au(111)," *Nanoscale*, vol. 7, no. 9, pp. 4234–4241, 2015.

[99] M. Bieri, M.-T. Nguyen, O. Gröning, J. Cai, M. Treier, K. Aït-Mansour, P. Ruffieux, C. A. Pignedoli, D. Passerone, M. Kastler, K. Müllen, and R. Fasel, "Two-dimensional polymer formation on surfaces: Insight into the roles of precursor mobility and reactivity," *Journal of the American Chemical Society*, vol. 132, no. 46, pp. 16669–16676, 2010.

[100] K. J. Laidler, *Chemical Kinetics*. McGraw-Hill, New York, 1950.

[101] K. E. Khor and S. D. Sarma, "Quantum dot self-assembly in growth of strained-layer thin films: A kinetic Monte Carlo study," *Physical Review B*, vol. 62, no. 24, p. 16657, 2000.

[102] K. Baek, G. Yun, Y. Kim, D. Kim, R. Hota, I. Hwang, D. Xu, Y. H. Ko, G. H. Gu, J. H. Suh, G. C. Park, B. J. Sung, and K. Kim, "Free-standing, single-monomer-thick two-dimensional polymers through cova-

lent self-assembly in solution," *Journal of the American Chemical Society*, vol. 135, no. 17, pp. 6523–6528, 2013.

[103] P. Szabelski, S. De Feyter, M. Drach, and S. Lei, "Computer simulation of chiral nanoporous networks on solid surfaces," *Langmuir*, vol. 26, no. 12, pp. 9506–9515, 2010.

[104] P. Szabelski, W. Rżysko, T. Pańczyk, E. Ghijsens, K. Tahara, Y. Tobe, and S. De Feyter, "Self-assembly of molecular tripods in two dimensions: Structure and thermodynamics from computer simulations," *RSC Advances*, vol. 3, no. 47, pp. 25159–25165, 2013.

[105] R. Whitesides and M. Frenklach, "Detailed kinetic Monte Carlo simulations of graphene-edge growth," *The Journal of Physical Chemistry A*, vol. 114, no. 2, pp. 689–703, 2009.

[106] P. Wu, H. Jiang, W. Zhang, Z. Li, Z. Hou, and J. Yang, "Lattice mismatch induced nonlinear growth of graphene," *Journal of the American Chemical Society*, vol. 134, no. 13, pp. 6045–6051, 2012.

[107] M. Linares, A. Minoia, P. Brocorens, D. Beljonne, and R. Lazzaroni, "Expression of chirality in molecular layers at surfaces: Insights from modelling," *Chemical Society Reviews*, vol. 38, no. 3, pp. 806–816, 2009.

[108] J. A. Elliott, "Novel approaches to multiscale modelling in materials science," *International Materials Reviews*, vol. 56, no. 4, pp. 207–225, 2011.

[109] R. A. Lawson, A. J. Peters, P. J. Ludovice, and C. L. Henderson, "Coarse grained molecular dynamics model of block copolymer directed self-assembly," in *SPIE Advanced Lithography*, pp. 86801Y–1–86801Y–11, International Society for Optics and Photonics, 2013.

[110] B. P. Uberuaga and A. F. Voter, "Determining reaction mechanisms," in *Handbook of Materials Modeling* (S. Yip, ed.), pp. 1627–1634, Springer, 2005.

[111] M. S. P. Sansom, K. A. Scott, and P. J. Bond, "Coarse-grained simulation: A high-throughput computational approach to membrane proteins," *Biochemical Society Transactions*, vol. 36, no. 1, pp. 27–32, 2008.

[112] R. M. Nieminen, "From atomistic simulation towards multiscale modelling of materials," *Journal of Physics: Condensed Matter*, vol. 14, no. 11, p. 2859, 2002.

[113] R. L. McGreevy, "Reverse Monte Carlo modelling," *Journal of Physics: Condensed Matter*, vol. 13, no. 46, p. R877, 2001.

[114] C.-C. Lee, A. Nayak, A. Sethuraman, G. Belfort, and G. J. McRae, "A three-stage kinetic model of amyloid fibrillation," *Biophysical Journal*, vol. 92, no. 10, pp. 3448–3458, 2007.

[115] M. S. Kumar and R. Schwartz, "A parameter estimation technique for stochastic self-assembly systems and its application to human papillomavirus self-assembly," *Physical Biology*, vol. 7, no. 4, p. 045005, 2010.

[116] E. C. Ifeachor and B. W. Jervis, *Digital signal processing: A practical approach*. Pearson Education, 2002.

[117] J. W. Goodby, M. Hird, K. J. Toyne, and T. Watson, "A novel, efficient and general synthetic route to unsymmetrical triphenylene mesogens using palladium-catalysed cross-coupling reactions," *Journal of the Chemical Society, Chemical Communications*, no. 14, pp. 1701–1702, 1994.

[118] M. W. Rees, M. N. Short, and B. Kassanis, "The amino acid composition, antigenicity, and other characteristics of the satellite viruses of tobacco necrosis virus," *Virology*, vol. 40, no. 3, pp. 448–461, 1970.

[119] M. Senechal, *Quasicrystals and geometry*. CUP Archive, 1996.

[120] D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn, "Metallic phase with long-range orientational order and no translational symmetry," *Physical Review Letters*, vol. 53, no. 20, p. 1951, 1984.

[121] G. Indelicato, P. Cermelli, D. G. Salthouse, S. Racca, G. Zanzotto, and R. Twarock, "A crystallographic approach to structural transitions in icosahedral viruses," *Journal of Mathematical Biology*, vol. 64, no. 5, pp. 745–773, 2012.

[122] A. Janner, "Form, symmetry and packing of biomacromolecules: I. Concepts and tutorial examples," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 66, no. 3, pp. 301–311, 2010.

[123] A. Janner, "Form, symmetry and packing of biomacromolecules: II. Serotypes of human rhinovirus," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 66, no. 3, pp. 312–326, 2010.

[124] J. H. Conway and N. J. A. Sloane, *Sphere packings, lattices and groups*. Springer Science & Business Media, 1988.

[125] P. W. K. Rothemund, "Using lateral capillary forces to compute by self-assembly," *Proceedings of the National Academy of Sciences*, vol. 97, no. 3, pp. 984–989, 2000.

[126] X. Du, I. Skachko, A. Barker, and E. Y. Andrei, "Approaching ballistic transport in suspended graphene," *Nature Nanotechnology*, vol. 3, no. 8, pp. 491–495, 2008.

[127] A. K. Geim and K. S. Novoselov, "The rise of graphene," *Nature Materials*, vol. 6, no. 3, pp. 183–191, 2007.

[128] X. Li, X. Wang, L. Zhang, S. Lee, and H. Dai, "Chemically derived, ultrasmooth graphene nanoribbon semiconductors," *Science*, vol. 319, no. 5867, pp. 1229–1232, 2008.

[129] M. Y. Han, B. Özyilmaz, Y. Zhang, and P. Kim, "Energy band-gap engineering of graphene nanoribbons," *Physical Review Letters*, vol. 98, no. 20, p. 206805, 2007.

[130] A. H. C. Neto, F. Guinea, N. M. R. Peres, K. S. Novoselov, and A. K. Geim, "The electronic properties of graphene," *Reviews of Modern Physics*, vol. 81, no. 1, p. 109, 2009.

[131] T. Enoki, Y. Kobayashi, and K.-I. Fukui, "Electronic structures of graphene edges and nanographene," *International Reviews in Physical Chemistry*, vol. 26, no. 4, pp. 609–645, 2007.

[132] V. Barone, O. Hod, and G. E. Scuseria, "Electronic structure and stability of semiconducting graphene nanoribbons," *Nano Letters*, vol. 6, no. 12, pp. 2748–2754, 2006.

[133] F. Schwierz, "Graphene transistors," *Nature Nanotechnology*, vol. 5, no. 7, pp. 487–496, 2010.

[134] D. Querlioz, Y. Apertet, A. Valentin, K. Huet, A. Bournel, S. Galdin-Retailleau, and P. Dollfus, "Suppression of the orientation effects on bandgap in graphene nanoribbons in the presence of edge disorder," *Applied Physics Letters*, vol. 92, no. 4, p. 042108, 2008.

[135] C. Berger, Z. Song, X. Li, X. Wu, N. Brown, C. Naud, D. Mayou, T. Li, J. Hass, A. N. Marchenkov, E. H. Conrad, P. N. First, and W. A. de Heer,

"Electronic confinement and coherence in patterned epitaxial graphene," *Science*, vol. 312, no. 5777, pp. 1191–1196, 2006.

[136] K. Nakada, M. Fujita, G. Dresselhaus, and M. S. Dresselhaus, "Edge state in graphene ribbons: Nanometer size effect and edge shape dependence," *Physical Review B*, vol. 54, no. 24, p. 17954, 1996.

[137] L. Jiao, L. Zhang, X. Wang, G. Diankov, and H. Dai, "Narrow graphene nanoribbons from carbon nanotubes," *Nature*, vol. 458, no. 7240, pp. 877–880, 2009.

[138] A. Reina, S. Thiele, X. Jia, S. Bhaviripudi, M. S. Dresselhaus, J. A. Schaefer, and J. Kong, "Growth of large-area single-and bi-layer graphene by controlled carbon precipitation on polycrystalline Ni surfaces," *Nano Research*, vol. 2, no. 6, pp. 509–516, 2009.

[139] C. Mattevi, H. Kim, and M. Chhowalla, "A review of chemical vapour deposition of graphene on copper," *Journal of Materials Chemistry*, vol. 21, no. 10, pp. 3324–3334, 2011.

[140] T. Kudernac, S. Lei, J. A. A. W. Elemans, and S. De Feyter, "Two-dimensional supramolecular self-assembly: Nanoporous networks on surfaces," *Chemical Society Reviews*, vol. 38, no. 2, pp. 402–421, 2009.

[141] J. Bai, X. Zhong, S. Jiang, Y. Huang, and X. Duan, "Graphene nanomesh," *Nature Nanotechnology*, vol. 5, no. 3, pp. 190–194, 2010.

[142] M. Kim, N. S. Safron, E. Han, M. S. Arnold, and P. Gopalan, "Fabrication and characterization of large-area, semiconducting nanoperforated graphene materials," *Nano Letters*, vol. 10, no. 4, pp. 1125–1131, 2010.

[143] F. Cervantes-Sodi, G. Csanyi, S. Piscanec, and A. C. Ferrari, "Edge-functionalized and substitutionally doped graphene nanoribbons: Elec-

tronic and spin properties," *Physical Review B*, vol. 77, no. 16, p. 165427, 2008.

[144] M. Bieri, M. Treier, J. Cai, K. Aït-Mansour, P. Ruffieux, O. Gröning, P. Gröning, M. Kastler, R. Rieger, X. Feng, K. Müllen, and R. Fasel, "Porous graphenes: Two-dimensional polymer synthesis with atomic precision," *Chemical Communications*, no. 45, pp. 6919–6921, 2009.

[145] Y. Hancock, A. Uppstu, K. Saloriutta, A. Harju, and M. J. Puska, "Generalized tight-binding transport model for graphene nanoribbon-based systems," *Physical Review B*, vol. 81, no. 24, p. 245402, 2010.

[146] J. Wu, M. D. Watson, and K. Müllen, "The versatile synthesis and self-assembly of star-type hexabenzocoronenes," *Angewandte Chemie*, vol. 115, pp. 5487–5491, Nov. 2003.

[147] L. Zhi and K. Müllen, "A bottom-up approach from molecular nanographenes to unconventional carbon materials," *Journal of Materials Chemistry*, vol. 18, pp. 1472–1484, 2008.

[148] S. Blankenburg, J. Cai, P. Ruffieux, R. Jaafar, D. Passerone, X. Feng, K. Müllen, R. Fasel, and C. A. Pignedoli, "Intraribbon heterojunction formation in ultranarrow graphene nanoribbons," *ACS Nano*, vol. 6, no. 3, pp. 2020–2025, 2012.

[149] R. P. Andres, J. D. Bielefeld, J. I. Henderson, D. B. Janes, V. R. Kolagunta, C. P. Kubiak, W. J. Mahoney, and R. G. Osifchin, "Self-assembly of a two-dimensional superlattice of molecularly linked metal clusters," *Science*, vol. 273, no. 5282, pp. 1690–1693, 1996.

[150] X. Yan and L. Li, "Solution-chemistry approach to graphene nanostructures," *Journal of Materials Chemistry*, vol. 21, no. 10, pp. 3295–3300, 2011.

[151] P. Han, K. Akagi, F. Federici Canova, H. Mutoh, S. Shiraki, K. Iwaya, P. S. Weiss, N. Asao, and T. Hitosugi, "Bottom-up graphene-nanoribbon fabrication reveals chiral edges and enantioselectivity," *ACS Nano*, vol. 8, no. 9, pp. 9181–9187, 2014.

[152] L. Chen, Y. Hernandez, X. Feng, and K. Müllen, "From nanographene and graphene nanoribbons to graphene sheets: Chemical synthesis," *Angewandte Chemie International Edition*, vol. 51, no. 31, pp. 7640–7654, 2012.

[153] M. C. Artal, K. J. Toyne, J. W. Goodby, J. Barberá, and D. J. Photinos, "Synthesis and mesogenic properties of novel board-like liquid crystals," *Journal of Materials Chemistry*, vol. 11, no. 11, pp. 2801–2807, 2001.

[154] M. Randić, D. J. Klein, H.-Y. Zhu, N. Trinajstić, and T. Živković, "Comparative study of large molecules. highly accurate calculation of a limit for infinite systems from data on finite systems," *Theoretica Chimica Acta*, vol. 90, no. 1, pp. 1–26, 1995.

[155] T. Živković, M. Randić, D. J. Klein, H.-Y. Zhu, and N. Trinajstić, "Analytical approach to very large benzenoid polymers," *Journal of Computational Chemistry*, vol. 16, no. 4, pp. 517–526, 1995.

[156] A. L. J. Pang, V. Sorkin, Y.-W. Zhang, and D. J. Srolovitz, "Self-assembly of free-standing graphene nano-ribbons," *Physics Letters A*, vol. 376, no. 8, pp. 973–977, 2012.

[157] D. Konatham and A. Striolo, "Molecular design of stable graphene nanosheets dispersions," *Nano Letters*, vol. 8, no. 12, pp. 4630–4641, 2008.

[158] Q. Lu and R. Huang, "Excess energy and deformation along free edges of graphene nanoribbons," *Physical Review B*, vol. 81, no. 15, p. 155410, 2010.

[159] L. Ortolani, E. Cadelano, G. P. Veronese, C. Degli Esposti Boschi, E. Snoeck, L. Colombo, and V. Morandi, "Folded graphene membranes: Mapping curvature at the nanoscale," *Nano Letters*, vol. 12, no. 10, pp. 5207–5212, 2012.

[160] A. A. O. Sarhan and C. Bolm, "Iron(III) chloride in oxidative C-C coupling reactions," *Chemical Society Reviews*, vol. 38, no. 9, pp. 2730–44, 2009.

[161] W. D. Wheeler, B. A. Parkinson, and Y. Dahnovsky, "The adsorption energy and diffusion of a pentacene molecule on a gold surface," *The Journal of Chemical Physics*, vol. 135, no. 2, p. 024702, 2011.

[162] J. Björk, F. Hanke, and S. Stafström, "Mechanisms of halogen-based covalent self-assembly on metal surfaces," *Journal of the American Chemical Society*, vol. 135, no. 15, pp. 5768–5775, 2013.

[163] K.-Y. Kwon, K. L. Wong, G. Pawin, L. Bartels, S. Stolbov, and T. S. Rahman, "Unidirectional adsorbate motion on a high-symmetry surface: "Walking" molecules can stay the course," *Physical Review Letters*, vol. 95, no. 16, p. 166101, 2005.

[164] G. Pawin, K. L. Wong, K.-Y. Kwon, R. J. Frisbee, T. S. Rahman, and L. Bartels, "Surface diffusive motion in a periodic and asymmetric potential," *Journal of the American Chemical Society*, vol. 130, no. 46, pp. 15244–15245, 2008.

[165] J.-R. Gong, L.-J. Wan, Q.-H. Yuan, C.-L. Bai, H. Jude, and P. J. Stang, "Mesoscopic self-organization of a self-assembled supramolecular rectangle

on highly oriented pyrolytic graphite and Au(111) surfaces," *Proceedings of the National Academy of Sciences*, vol. 102, no. 4, pp. 971–974, 2005.

[166] R. G. Nuzzo, L. H. Dubois, and D. L. Allara, "Fundamental studies of microscopic wetting on organic surfaces. 1. formation and structural characterization of a self-consistent series of polyfunctional organic monolayers," *Journal of the American Chemical Society*, vol. 112, no. 2, pp. 558–569, 1990.

[167] S.-H. Hsu, D. N. Reinhoudt, J. Huskens, and A. H. Velders, "Lateral interactions at functional monolayers," *Journal of Materials Chemistry*, vol. 21, no. 8, pp. 2428–2444, 2010.

[168] H. Wang, X. Wang, X. Li, and H. Dai, "Chemical self-assembly of graphene sheets," *Nano Research*, vol. 2, no. 4, pp. 336–342, 2009.

[169] R. K. Smith, P. A. Lewis, and P. S. Weiss, "Patterning self-assembled monolayers," *Progress in Surface Science*, vol. 75, no. 1, pp. 1–68, 2004.

[170] R. Erban, S. J. Chapman, and P. K. Maini, "A practical guide to stochastic simulations of reaction-diffusion processes," tech. rep., University of Oxford, Oxford, United Kingdom, 2008.

[171] D. Bernstein, "Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm," *Physical Review E*, vol. 71, no. 4, p. 041103, 2005.

[172] E. Sperotto, G. P. M. van Klink, G. van Koten, and J. G. de Vries, "The mechanism of the modified Ullmann reaction," *Dalton Transactions*, vol. 39, no. 43, pp. 10338–10351, 2010.

[173] S. Mukhopadhyay, G. Rothenberg, D. Gitis, H. Wiener, and Y. Sasson, "Kinetics and mechanism of heterogeneous palladium-catalyzed coupling

reactions of chloroaryls in water," *Journal of the Chemical Society, Perkin Transactions 2*, no. 11, pp. 2481–2484, 1999.

[174] Z. Li, Y. Fu, Q.-X. Guo, and L. Liu, "Theoretical study on monoligated Pd-catalyzed cross-coupling reactions of aryl chlorides and bromides," *Organometallics*, vol. 27, no. 16, pp. 4043–4049, 2008.

[175] Z. Li, Y.-Y. Jiang, and Y. Fu, "Theoretical study on the mechanism of Ni-catalyzed alkyl-alkyl Suzuki cross-coupling," *Chemistry: A European Journal*, vol. 18, no. 14, pp. 4345–4357, 2012.

[176] S. Kozuch and S. Shaik, "A combined kinetic-quantum mechanical model for assessment of catalytic cycles: Application to cross-coupling and Heck reactions," *Journal of the American Chemical Society*, vol. 128, no. 10, pp. 3355–3365, 2006.

[177] J.-M. Chern and F. G. Helfferich, "Effective kinetic modeling of multistep homogeneous reactions," *AIChE Journal*, vol. 36, no. 8, pp. 1200–1208, 1990.

[178] B. Sweeney, T. Zhang, and R. Schwartz, "Exploring the parameter space of complex self-assembly through virus capsid models," *Biophysical Journal*, vol. 94, no. 3, pp. 772–783, 2008.

[179] R. Schwartz, P. W. Shor, P. E. Prevelige, and B. Berger, "Local rules simulation of the kinetics of virus capsid self-assembly," *Biophysical Journal*, vol. 75, no. 6, pp. 2626–2636, 1998.

[180] A. Zlotnick and S. Mukhopadhyay, "Virus assembly, allostery and antivirals," *Trends in Microbiology*, vol. 19, no. 1, pp. 14–23, 2011.

[181] H. Fraenkel-Conrat and R. C. Williams, "Reconstitution of active tobacco mosaic virus from its inactive protein and nucleic acid components," *Pro-*

ceedings of the National Academy of Sciences, vol. 41, no. 10, pp. 690–698, 1955.

[182] A. Routh, T. Domitrovic, and J. E. Johnson, "Host RNAs, including transposons, are encapsidated by a eukaryotic single-stranded RNA virus," *Proceedings of the National Academy of Sciences*, vol. 109, no. 6, pp. 1907–1912, 2012.

[183] F. H. C. Crick and J. D. Watson, "Structure of small viruses," *Nature*, vol. 177, no. 4506, pp. 473–475, 1956.

[184] D. L. D. Caspar and A. Klug, "Physical principles in the construction of regular viruses," in *Cold Spring Harbor symposia on quantitative biology*, vol. 27, pp. 1–24, Cold Spring Harbor Laboratory Press, 1962.

[185] R. Golmohammadi, K. Valegård, K. Fridborg, and L. Liljas, "The refined structure of bacteriophage MS2 at 2.8Å resolution," *Journal of Molecular Biology*, vol. 234, no. 3, pp. 620–639, 1993.

[186] F. Coulibaly, C. Chevalier, I. Gutsche, J. Pous, J. Navaza, S. Bressanelli, B. Delmas, and F. A. Rey, "The birnavirus crystal structure reveals structural relationships among icosahedral viruses," *Cell*, vol. 120, no. 6, pp. 761–772, 2005.

[187] N. Nandhagopal, A. A. Simpson, J. R. Gurnon, X. Yan, T. S. Baker, M. V. Graves, J. L. Van Etten, and M. G. Rossmann, "The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus," *Proceedings of the National Academy of Sciences*, vol. 99, no. 23, pp. 14758–14763, 2002.

[188] R. Twarock, "A tiling approach to virus capsid assembly explaining a structural puzzle in virology," *Journal of Theoretical Biology*, vol. 226, no. 4, pp. 477–482, 2004.

[189] R. Twarock, "The architecture of viral capsids based on tiling theory," *Journal of Theoretical Medicine*, vol. 6, no. 2, pp. 87–90, 2005.

[190] J. M. Hogle, M. Chow, and D. J. Filman, "Three-dimensional structure of poliovirus at 2.9Å resolution," *Science*, vol. 229, no. 4720, pp. 1358–1365, 1985.

[191] J. M. Grimes, J. N. Burroughs, P. Gouet, J. M. Diprose, R. Malby, S. Ziéntara, P. P. Mertens, and D. I. Stuart, "The atomic structure of the bluetongue virus core," *Nature*, vol. 395, no. 6701, pp. 470–478, 1998.

[192] J. R. Castón, B. L. Trus, F. P. Booy, R. B. Wickner, J. S. Wall, and A. C. Steven, "Structure of L-A virus: A specialized compartment for the transcription and replication of double-stranded RNA," *Journal of Cell Biology*, vol. 138, no. 5, pp. 975–985, 1997.

[193] H. Naitow, J. Tang, M. Canady, R. B. Wickner, and J. E. Johnson, "L-A virus at 3.4Å resolution reveals particle architecture and mRNA decapping mechanism," *Nature Structural & Molecular Biology*, vol. 9, no. 10, pp. 725–728, 2002.

[194] S. T. Miller, J. M. Hogle, and D. J. Filman, "*Ab initio* phasing of high-symmetry macromolecular complexes: Successful phasing of authentic poliovirus data to 3.0Å resolution," *Journal of Molecular Biology*, vol. 307, no. 2, pp. 499–512, 2001.

[195] R. R. Rueckert, "Picornaviridae: The viruses and their replication," in *Fundamental virology* (B. N. Fields, D. M. Knipe, and P. M. Howley, eds.), ch. 16, pp. 477–522, Lippincott-Raven, 3 ed., 1996.

[196] G. Cardone, A. L. Moyer, N. Cheng, C. D. Thompson, I. Dvoretzky, D. R. Lowy, J. T. Schiller, A. C. Steven, C. B. Buck, and B. L. Trus, "Maturation of the human papillomavirus 16 capsid," *MBio*, vol. 5, no. 4, pp. e01104–14, 2014.

[197] T. Stehle and S. C. Harrison, "Crystal structures of murine polyomavirus in complex with straight-chain and branched-chain sialyloligosaccharide receptor fragments," *Structure*, vol. 4, no. 2, pp. 183–194, 1996.

[198] R. C. Liddington, Y. Yan, J. Moulai, R. Sahli, T. L. Benjamin, and S. C. Harrison, "Structure of simian virus 40 at 3.8Å resolution," *Nature*, vol. 354, no. 6351, pp. 278–284, 1991.

[199] E. C. Dykeman, P. G. Stockley, and R. Twarock, "Dynamic allostery controls coat protein conformer switching during MS2 phage assembly," *Journal of Molecular Biology*, vol. 395, no. 5, pp. 916–923, 2010.

[200] E. C. Dykeman and R. Twarock, "All-atom normal-mode analysis reveals an RNA-induced allostery in a bacteriophage coat protein," *Physical Review E*, vol. 81, no. 3, p. 031908, 2010.

[201] J. Lidmar, L. Mirny, and D. R. Nelson, "Virus shapes and buckling transitions in spherical shells," *Physical Review E*, vol. 68, no. 5, p. 051910, 2003.

[202] B. V. V. Prasad and M. F. Schmid, "Principles of virus structural organization," in *Viral Molecular Machines* (M. G. Rossmann, ed.), ch. 3, pp. 17–47, Springer, 2012.

[203] R. A. Crowther, D. J. DeRosier, and A. Klug, "The reconstruction of a three-dimensional structure from projections and its application to electron microscopy," in *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 317, pp. 319–340, The Royal Society, 1970.

[204] E. V. Orlova and H. R. Saibil, "Structure determination of macromolecular assemblies by single-particle analysis of cryo-electron micrographs," *Current Opinion in Structural Biology*, vol. 14, no. 5, pp. 584–590, 2004.

[205] K. C. Dent, R. Thompson, A. M. Barker, J. A. Hiscox, J. N. Barr, P. G. Stockley, and N. A. Ranson, "The asymmetric structure of an icosahedral virus bound to its receptor suggests a mechanism for genome release," *Structure*, vol. 21, no. 7, pp. 1225–1234, 2013.

[206] E. C. Dykeman, P. G. Stockley, and R. Twarock, "Packaging signals in two single-stranded RNA viruses imply a conserved assembly mechanism and geometry of the packaged genome," *Journal of Molecular Biology*, vol. 425, no. 17, pp. 3235–3249, 2013.

[207] N. A. Ranson and P. G. Stockley, "Cryo-electron microscopy of viruses," in *Emerging topics in physical virology* (P. G. Stockley and R. Twarock, eds.), ch. 1, pp. 1–33, Imperial College Press, 2010.

[208] M. C. Morais, Y. Tao, N. H. Olson, S. Grimes, P. J. Jardine, D. L. Anderson, T. S. Baker, and M. G. Rossmann, "Cryoelectron-microscopy image reconstruction of symmetry mismatches in bacteriophage φ29," *Journal of Structural Biology*, vol. 135, no. 1, pp. 38–46, 2001.

[209] J. E. Johnson and W. Chiu, "DNA packaging and delivery machines in tailed bacteriophages," *Current Opinion in Structural Biology*, vol. 17, no. 2, pp. 237–243, 2007.

[210] P. A. Venter, N. K. Krishna, and A. Schneemann, "Capsid protein synthesis from replicating RNA directs specific packaging of the genome of a multipartite, positive-strand RNA virus," *Journal of Virology*, vol. 79, no. 10, pp. 6239–6248, 2005.

[211] N. Patel, E. C. Dykeman, R. H. A. Coutts, G. P. Lomonossoff, D. J. Rowlands, S. E. V. Phillips, N. Ranson, R. Twarock, R. Tuma, and P. G. Stockley, "Revealing the density of encoded functions in a viral RNA," *Proceedings of the National Academy of Sciences*, vol. 112, no. 7, pp. 2227–2232, 2015.

[212] A. Routh, T. Domitrovic, and J. E. Johnson, "Packaging host RNAs in small RNA viruses: An inevitable consequence of an error-prone polymerase?," *Cell Cycle*, vol. 11, no. 20, pp. 3713–3714, 2012.

[213] A. Borodavka, R. Tuma, and P. G. Stockley, "Evidence that viral RNAs have evolved for efficient, two-stage packaging," *Proceedings of the National Academy of Sciences*, vol. 109, no. 39, pp. 15769–15774, 2012.

[214] K. N. Johnson, L. Tang, J. E. Johnson, and L. A. Ball, "Heterologous RNA encapsidated in pariacoto virus-like particles forms a dodecahedral cage similar to genomic RNA in wild-type virions," *Journal of Virology*, vol. 78, no. 20, pp. 11371–11378, 2004.

[215] K. Valegård, J. B. Murray, P. G. Stockley, N. J. Stonehouse, and L. Liljas, "Crystal structure of an RNA bacteriophage coat protein operator complex," *Nature*, vol. 371, no. 6498, pp. 623–626, 1994.

[216] J. C.-Y. Wang, D. G. Nickens, T. B. Lentz, D. D. Loeb, and A. Zlotnick, "Encapsidated hepatitis B virus reverse transcriptase is poised on an ordered RNA lattice," *Proceedings of the National Academy of Sciences*, vol. 111, no. 31, pp. 11329–11334, 2014.

[217] E. C. Dykeman, "An implementation of the Gillespie algorithm for RNA kinetics with logarithmic time update," *Nucleic Acids Research*, vol. 43, no. 12, pp. 5708–5715, 2015.

[218] P. S. Henke and C. H. Mak, "Free energy of RNA-counterion interactions in a tight-binding model computed by a discrete space mapping," *The Journal of Chemical Physics*, vol. 141, no. 6, p. 064116, 2014.

[219] D. E. Draper, "A guide to ions and RNA structure," *RNA*, vol. 10, no. 3, pp. 335–343, 2004.

[220] J. L. Boots, M. D. Canny, E. Azimi, and A. Pardi, "Metal ion specificities for folding and cleavage activity in the *Schistosoma* hammerhead ribozyme," *RNA*, vol. 14, no. 10, pp. 2212–2222, 2008.

[221] R. J. Ford, *The roles of RNA in the assembly and disassembly of single-stranded RNA icosahedral viruses*. University of Leeds, 2012.

[222] P. Ni, Z. Wang, X. Ma, N. C. Das, P. Sokol, W. Chiu, B. Dragnea, M. Hagan, and C. C. Kao, "An examination of the electrostatic interactions between the N-terminal tail of the brome mosaic virus coat protein and encapsidated RNAs," *Journal of Molecular Biology*, vol. 419, no. 5, pp. 284–300, 2012.

[223] M. Eigen, J. McCaskill, and P. Schuster, "Molecular quasi-species," *The Journal of Physical Chemistry*, vol. 92, no. 24, pp. 6881–6891, 1988.

[224] A. S. Lauring and R. Andino, "Quasispecies theory and the behavior of RNA viruses," *PLoS Pathogens*, vol. 6, no. 7, p. e1001005, 2010.

[225] S. Ojosnegros, C. Perales, A. Mas, and E. Domingo, "Quasispecies as a matter of fact: Viruses and beyond," *Virus Research*, vol. 162, no. 1, pp. 203–215, 2011.

[226] J. Carey, V. Cameron, P. L. De Haseth, and O. C. Uhlenbeck, "Sequence-specific interaction of R17 coat protein with its ribonucleic acid binding site," *Biochemistry*, vol. 22, no. 11, pp. 2601–2610, 1983.

[227] J. Carey and O. C. Uhlenbeck, "Kinetic and thermodynamic characterization of the R17 coat protein-ribonucleic acid interaction," *Biochemistry*, vol. 22, no. 11, pp. 2610–2615, 1983.

[228] P. G. Stockley, O. Rolfsson, G. S. Thompson, G. Basnak, S. Francese, N. J. Stonehouse, S. W. Homans, and A. E. Ashcroft, "A simple, RNA-mediated allosteric switch controls the pathway to formation of a $T = 3$ viral capsid," *Journal of Molecular Biology*, vol. 369, no. 2, pp. 541–552, 2007.

[229] S. Talbot, S. Goodman, S. Bates, C. Fishwick, and P. Stockley, "Use of synthetic oligoribonucleotides to probe RNA-protein interactions in the MS2 translational operator complex," *Nucleic Acids Research*, vol. 18, no. 12, pp. 3521–3528, 1990.

[230] M. Adrian, J. Dubochet, J. Lepault, and A. W. McDowall, "Cryo-electron microscopy of viruses," *Nature*, vol. 308, no. 5954, pp. 32–36, 1984.

[231] K. M. ElSawy, L. S. D. Caves, and R. Twarock, "On the origin of order in the genome organization of ssRNA viruses," *Biophysical Journal*, vol. 101, no. 4, pp. 774–780, 2011.

[232] A. Luque and D. Reguera, "Theoretical studies on assembly, physical stability and dynamics of viruses," in *Structure and Physics of Viruses* (M. G. Mateu, ed.), ch. 19, pp. 553–595, Springer, 2013.

[233] R. F. Bruinsma, "Physics of RNA and viral assembly," *The European Physical Journal E*, vol. 19, no. 3, pp. 303–310, 2006.

[234] J. Rudnick and R. Bruinsma, "Icosahedral packing of RNA viral genomes," *Physical Review Letters*, vol. 94, no. 3, p. 038101, 2005.

[235] S. Hafenstein, L. M. Palermo, V. A. Kostyuchenko, C. Xiao, M. C. Morais, C. D. S. Nelson, V. D. Bowman, A. J. Battisti, P. R. Chipman, C. R. Parrish, and M. G. Rossmann, "Asymmetric binding of transferrin receptor to parvovirus capsids," *Proceedings of the National Academy of Sciences*, vol. 104, no. 16, pp. 6585–6589, 2007.

[236] H. C. Levy, M. Bostina, D. J. Filman, and J. M. Hogle, "Catching a virus in the act of RNA release: A novel poliovirus uncoating intermediate characterized by cryo-electron microscopy," *Journal of Virology*, vol. 84, no. 9, pp. 4426–4441, 2010.

[237] M. Bostina, H. Levy, D. J. Filman, and J. M. Hogle, "Poliovirus RNA is released from the capsid near a twofold symmetry axis," *Journal of Virology*, vol. 85, no. 2, pp. 776–783, 2011.

[238] J. Ren, X. Wang, Z. Hu, Q. Gao, Y. Sun, X. Li, C. Porta, T. S. Walter, R. J. Gilbert, Y. Zhao, D. Axford, M. Williams, K. McAuley, D. J. Rowlands, W. Yin, J. Wang, D. I. Stuart, Z. Rao, and E. E. Fry, "Picornavirus uncoating intermediate captured in atomic detail," *Nature Communications*, vol. 4, p. 1929, 2013.

[239] T. J. Tuthill, E. Groppelli, J. M. Hogle, and D. J. Rowlands, "Picornaviruses," in *Cell Entry by Non-Enveloped Viruses* (J. E. Johnson, ed.), pp. 43–89, Springer, 2010.

[240] A. Šiber, A. L. Božič, and R. Podgornik, "Energies and pressures in viruses: Contribution of nonspecific electrostatic interactions," *Physical Chemistry Chemical Physics*, vol. 14, no. 11, pp. 3746–3765, 2012.

[241] V. A. Belyi and M. Muthukumar, "Electrostatic origin of the genome packing in viruses," *Proceedings of the National Academy of Sciences*, vol. 103, no. 46, pp. 17174–17178, 2006.

[242] D. Zhang, R. Konecny, N. A. Baker, and J. A. McCammon, "Electrostatic interaction between RNA and protein capsid in cowpea chlorotic mottle virus simulated by a coarse-grain RNA model and a Monte Carlo approach," *Biopolymers*, vol. 75, no. 4, pp. 325–337, 2004.

[243] J. D. Perlmutter, C. Qiao, and M. F. Hagan, "Viral genome structures are optimal for capsid assembly," *Elife*, vol. 2, p. e00632, 2013.

[244] B. Kassanis, "Properties and behaviour of a virus depending for its multiplication on another," *Journal of General Microbiology*, vol. 27, no. 3, pp. 477–488, 1962.

[245] P. Babos and B. Kassanis, "Serological relationships and some properties of tobacco necrosis virus strains," *Journal of General Microbiology*, vol. 32, no. 1, pp. 135–144, 1963.

[246] T. A. Jones and L. Liljas, "Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 å resolution," *Journal of Molecular Biology*, vol. 177, no. 4, pp. 735–767, 1984.

[247] R. J. Ford, A. M. Barker, S. E. Bakker, R. H. Coutts, N. A. Ranson, S. E. V. Phillips, A. R. Pearson, and P. G. Stockley, "Sequence-specific, RNA-protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein," *Journal of Molecular Biology*, vol. 425, no. 6, pp. 1050–1064, 2013.

[248] F. Jossinet and E. Westhof, "S2S-Assemble2: A semi-automatic bioinformatics framework to study and model RNA 3D architectures," *Handbook of RNA Biochemistry: Second, Completely Revised and Enlarged Edition*, pp. 667–686, 2014.

[249] A. Janner, "From an affine extended icosahedral group towards a toolkit for viral architecture," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 69, no. 2, pp. 151–163, 2013.

[250] A. Janner, "Form, symmetry and packing of biomacromolecules: III. Antigenic, receptor and contact binding sites in picornaviruses," *Acta Crystallographica Section A: Foundations of Crystallography*, vol. 67, no. 2, pp. 174–189, 2011.

[251] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin, "UCSF Chimera—a visualization system for exploratory research and analysis," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1605–1612, 2004.

[252] W. Fiers, R. Contreras, F. Duerinck, G. Haegeman, D. Iserentant, J. Merregaert, W. Min Jou, F. Molemans, A. Raeymaekers, A. Van den Berghe, G. Volckaert, and M. Ysebaert, "Complete nucleotide sequence of bacteriophage MS2 RNA: Primary and secondary structure of the replicase gene," *Nature*, vol. 260, no. 5551, pp. 500–507, 1976.

[253] K. Valegård, L. Liljas, K. Fridborg, and T. Unge, "The three-dimensional structure of the bacterial virus MS2," *Nature*, vol. 345, no. 6270, pp. 36–41, 1990.

[254] R. A. Kastelein, E. Remaut, W. Fiers, and J. Van Duin, "Lysis gene expression of RNA phage MS2 depends on a frameshift during translation of the overlapping coat protein gene," *Nature*, vol. 295, no. 5844, pp. 35–41, 1982.

[255] H. Lago, A. M. Parrott, T. Moss, N. J. Stonehouse, and P. G. Stockley, "Probing the kinetics of formation of the bacteriophage MS2 translational operator complex: Identification of a protein conformer unable to bind RNA," *Journal of Molecular Biology*, vol. 305, no. 5, pp. 1131–1144, 2001.

[256] K. M. ElSawy, L. S. D. Caves, and R. Twarock, "The impact of viral RNA on the association rates of capsid protein assembly: Bacteriophage MS2 as a case study," *Journal of Molecular Biology*, vol. 400, no. 4, pp. 935–947, 2010.

[257] A. Schneemann, "The structural and functional role of RNA in icosahedral virus assembly," *Annual Review of Microbiology*, vol. 60, pp. 51–67, 2006.

[258] R. Koning, S. van den Worm, J. R. Plaisier, J. van Duin, J. P. Abrahams, and H. Koerten, "Visualization by cryo-electron microscopy of genomic RNA that binds to the protein capsid inside bacteriophage MS2," *Journal of Molecular Biology*, vol. 332, no. 2, pp. 415–422, 2003.

[259] G. Basnak, V. L. Morton, Ó. Rolfsson, N. J. Stonehouse, A. E. Ashcroft, and P. G. Stockley, "Viral genomic single-stranded RNA directs the pathway toward a $T = 3$ capsid," *Journal of Molecular Biology*, vol. 395, no. 5, pp. 924–936, 2010.

[260] T. Shiba and Y. Suzuki, "Localization of A protein in the RNA-A protein complex of RNA phage MS2," *Biochimica et Biophysica Acta: Nucleic Acids and Protein Synthesis*, vol. 654, no. 2, pp. 249–255, 1981.

[261] G. D. Pintilie, J. Zhang, T. D. Goddard, W. Chiu, and D. C. Gossard, "Quantitative analysis of cryo-em density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions," *Journal of Structural Biology*, vol. 170, no. 3, pp. 427–438, 2010.

[262] K. Toropova, P. G. Stockley, and N. A. Ranson, "Visualising a viral RNA genome poised for release from its receptor complex," *Journal of Molecular Biology*, vol. 408, no. 3, pp. 408–419, 2011.

[263] Y. Zeng, S. B. Larson, C. E. Heitsch, A. McPherson, and S. C. Harvey, "A model for the structure of satellite tobacco mosaic virus," *Journal of Structural Biology*, vol. 180, no. 1, pp. 110–116, 2012.

[264] S. B. Larson, S. Koszelak, J. Day, A. Greenwood, J. A. Dodds, and A. McPherson, "Double-helical RNA in satellite tobacco mosaic virus," *Nature*, vol. 361, no. 6408, pp. 179–182, 1993.

[265] S. B. Larson, S. Koszelak, J. Day, A. Greenwood, J. A. Dodds, and A. McPherson, "Three-dimensional structure of satellite tobacco mosaic virus at 2.9Å resolution," *Journal of Molecular Biology*, vol. 231, no. 2, pp. 375–391, 1993.

[266] S. B. Larson and A. McPherson, "Satellite tobacco mosaic virus RNA: Structure and implications for assembly," *Current Opinion in Structural Biology*, vol. 11, no. 1, pp. 59–65, 2001.

[267] S. J. Schroeder, J. W. Stone, S. Bleckley, T. Gibbons, and D. M. Mathews, "Ensemble of secondary structures for encapsidated satellite tobacco

mosaic virus RNA consistent with chemical probing and crystallography constraints," *Biophysical Journal*, vol. 101, no. 1, pp. 167–175, 2011.

[268] P. Ni, R. C. Vaughan, B. Tragesser, H. Hoover, and C. C. Kao, "The plant host can affect the encapsidation of brome mosaic virus (BMV) RNA: BMV virions are surprisingly heterogeneous," *Journal of Molecular Biology*, vol. 426, no. 5, pp. 1061–1076, 2014.

[269] T. Lin, J. Cavarelli, and J. E. Johnson, "Evidence for assembly-dependent folding of protein and RNA in an icosahedral virus," *Virology*, vol. 314, no. 1, pp. 26–33, 2003.

[270] L. Tang, K. N. Johnson, L. A. Ball, T. Lin, M. Yeager, and J. E. Johnson, "The structure of pariacoto virus reveals a dodecahedral cage of duplex RNA," *Nature Structural & Molecular Biology*, vol. 8, no. 1, pp. 77–83, 2001.

[271] A. E. Firth, S. Atasheva, E. I. Frolova, and I. Frolov, "Conservation of a packaging signal and the viral genome RNA packaging mechanism in alphavirus evolution," *Journal of Virology*, vol. 85, no. 16, pp. 8022–8036, 2011.

[272] M. Boncheva and G. M. Whitesides, "Biomimetic approaches to the design of functional, self-assembling systems," in *Dekkar encyclopedia of nanoscience and nanotechnology* (J. A. Schwarz, C. I. Contescu, and K. Putyera, eds.), vol. 1, pp. 287–294, Marcel Dekker, 2004.

[273] Y. Y. Pinto, J. D. Le, N. C. Seeman, K. Musier-Forsyth, T. A. Taton, and R. A. Kiehl, "Sequence-encoded self-assembly of multiple-nanocomponent arrays by 2D DNA scaffolding," *Nano Letters*, vol. 5, no. 12, pp. 2399–2402, 2005.

[274] H. T. Maune, S.-P. Han, R. D. Barish, M. Bockrath, W. A. Goddard III, P. W. K. Rothemund, and E. Winfree, "Self-assembly of carbon nanotubes into two-dimensional geometries using DNA origami templates," *Nature Nanotechnology*, vol. 5, no. 1, pp. 61–66, 2010.

[275] F. A. Aldaye, A. L. Palmer, and H. F. Sleiman, "Assembling materials with DNA as the guide," *Science*, vol. 321, no. 5897, pp. 1795–1799, 2008.

[276] W. M. Jacobs, A. Reinhardt, and D. Frenkel, "Rational design of self-assembly pathways for complex multicomponent structures," *Proceedings of the National Academy of Sciences*, vol. 112, no. 20, pp. 6313–6318, 2015.

[277] A. C. Mendes, E. T. Baran, R. L. Reis, and H. S. Azevedo, "Self-assembly in nature: Using the principles of nature to create complex nanobiomaterials," *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology*, vol. 5, no. 6, pp. 582–612, 2013.