Autoassemblage des virus enveloppés : dynamique théorique et méthodes d'analyse des mesures par fluorescence

Timothée Verdier

N° National de Thèse : 226 - 2015

THÈSE

en vue de l'obtention du grade de

Docteur de l'Université de Lyon,
délivré par l'Université Claude Bernard Lyon 1

Discipline : **Physique**

Laboratoire de Physique de l'ENS de Lyon

École Doctorale de Physique et d'Astrophysique de Lyon

présentée et soutenue publiquement le 13 novembre 2015
par **M. Timothée VERDIER**

Self-assembly of enveloped virus : theoretical dynamics and methods for fluorescence measurements analysis

---

Directeur de Thèse : M. Martin Castelnovo

Devant la commission d'examen formée de :

M.	R.	HENRIQUES	*Rapporteur*	-	UCL, London
M.	P.	van der SCHOOT	*Rapporteur*	-	TU, Eindhoven
M.	L.	FORET	*Examinateur*	-	UPMC, Paris
M.	J-F.	JOANNY	*Examinateur*	-	UPMC, Paris
M.	J-P.	RIEU	*Examinateur*	-	UCBL, Lyon
M.	M.	CASTELNOVO	*Directeur de thèse*	-	ENS, Lyon

Remerciements

Je tiens à remercier Martin Castelnovo pour la confiance et le temps qu'il m'a accordé durant les trois ans où il a encadré cette thèse. Martin, ton enthousiasme et ta disponibilité ont été deux puissants moteurs dans ce travail de recherche. Le contenu de ce manuscrit a largement bénéficié de nos discussions stimulantes et des perspectives que tu m'a ouvertes par les travaux et les conférences vers lesquelles tu m'as orienté !

Je veux également remercier les chercheurs avec qui j'ai eu le plaisir de collaborer au cours de ce projet : Lionel Foret –qui m'a appris à apprivoiser les processus d'agrégation des protéines à la membrane– pour ses explications claires et détaillées et son accueil chaleureux sur Paris ; ainsi que Suliana Manley et Julia Gunzenhäuser qui ont fourni les données expérimentales superrésolues et confronté leur savoir à mes idées quant au traitement approprié.

Merci aux membre du jury de thèse : Lionel Foret, Ricardo Henriques, Jean-François Joanny, Jean-Paul Rieu, Paul van der Schoot qui m'ont fait l'honneur de leur relecture attentive du manuscrit et pour l'intérêt qu'ils ont porté à mon travail comme ma soutenance ainsi que pour leur suggestions d'amélioration. Je remercie tout particulièrement Paul van der Schoot et Ricardo Henriques d'avoir accepté d'assumer la tache de rapporteurs.

Merci à toute l'équipe du labo de physique de l'ENS ! Mention toute particulière bien sûr pour mes co-bureau Julien Bernaud docteur es enthousiasme dans les difficultés de la recherche, Jean-Baptiste Blanc pour tes remarques ciselées sur l'art de vivre, Jean-Yonel Chastaing avec qui j'ai appris à contourner les montagnes en riant et sur la pointe des pieds, Enesto Horne qui initie le monde au mate et Franco Tapia mon aîné dans les hauts et les bas de thèse (! Te debo muchos buenos asesoramientos y alegria compadre ! Seguramente güey !). Merci encore à Jean-Baptiste Flament, on aura bouclé des déménagements et un thèse ensemble, et à Caroline Crauste pour sa reclecture et ses conseils motivants sur la rédaction.

La liste est grande de tous ceux qui, amis, professeurs, famille, colocs, collègues, m'ont accompagné, orienté, entrainé, ou soutenu jusqu'ici. À chacun j'adresse un merci !

Merci à Étienne Rosenstielh et aux Benoit d'avoir été mes point de chute sur Lyon le temps de trouver encore d'autres merveilleux colocs et amis. Merci Simon, Célia, et Alphonse Robichu pour tout les bons moments passé dans Lyon et les alentours. Merci à toute la bande du théatre d'improvisation de l'ENS et aux enthousiastes de l'escalade.

À Fanny, pour tout ce que l'on a partagé et partage, et pour avoir été encore plus présente quand il a fallu faire naître ce travail...

Résumé

Cette thèse porte sur la description de l'assemblage des virus dans le cadre de la physique statistique ainsi que sur les méthodes de mesure de cet assemblage utilisant les marqueurs fluorescents. Nous nous y attachons à décrire la dynamique de l'agrégation des protéines aux échelles de la population et du virus unique. Nous proposons deux méthodes pour mesurer les grandeurs physiques associées : taille et forme de la structure finale d'une part, taux d'agrégation au cours de la croissance d'autre part.

Les virus sont des entités biologiques à l'interface entre l'inerte et le vivant. En effet ils ne possèdent pas la capacité de se répliquer, et généralement pas de métabolisme pour produire leur énergie, mais ils partagent les supports de code génétique et la constitution protéique du vivant. Sans capacités reproductrices autonomes, ils sont des parasites obligatoires des cellules dont ils détournent l'activité à leurs fins. Cette activité parasite est la source des effets pathogènes du virus qui touche l'ensemble du vivant depuis les bactéries unicellulaires jusqu'au règne animal en passant par les plantes. Si leur petite taille (généralement sub-micrométrique) les a longtemps tenus hors de portée des moyens d'investigation scientifique, l'importance des enjeux sanitaires et économiques des maladies qu'ils causent a stimulé une importante recherche. Les enjeux de la compréhension du fonctionnement viral sont évidents pour la lutte contre les virus pathogènes, mais dépassent largement ce cadre. Les utilisations à des fins de recherches (ou médicales) de leurs capacités ne manquent pas : pour altérer le fonctionnement cellulaire (thérapie génique), pour cibler les bactéries (alternative possible aux antibiotiques), pour cibler une cellule spécifique de l'organisme (nouvelles techniques de marquage pour l'imagerie du vivant, ciblage des traitements médicaux). L'étude et la maîtrise des propriétés d'auto-organisation des éléments constitutifs des virus trouve aussi des débouchées en micro-électronique où des circuits hybrides incluent des éléments organiques et sont parfois conçus de façon biomimétique, en chimie où l'encapsulation sélective de composés à des dimensions réduites est un enjeu et où, une fois encore, la formidable "compétence" des virus à cette échelle peut-être mise à profit.

Dans ce travail, nous nous sommes intéressés à la description physique de l'auto-assemblage des protéines virales. La physique de l'auto-assemblage in-vitro des virus sphériques, dont la structure est déterminée par l'agencement régulier de leurs constituants protéiques, a été théoriquement et expérimentalement caractérisée auparavant par des modèles d'agrégation. Les modèles existants décrivaient l'assemblage à quantité de composants viraux fixée dans un système fermé à partir des constituants élémentaires du virus. In-vivo, la situation est bien entendu différente. Abstraction faite de la grande complexité du milieu cellulaire, les virus s'échappent de la cellule une fois formés pour aller infecter de nouvelles cellules. De plus, la quantité de constituants est sans cesse modifiée par la fabrication ou la dégradation des protéines virales.

Enfin les méthodes de mesures utilisées in-vitro ne sont généralement plus envisageables in-vivo. Nous avons donc étudié les effets d'un flux de matière dans système ouvert via le calcul de l'état stationnaire, et via la résolution numérique des équations d'évolution des populations d'agrégats qui décrivent la cinétique d'agrégation des protéines virales. Dans ce cadre, nous avons mis en valeur le lien entre la description de l'état général du système en termes de populations et le devenir individuel d'un virus en formation pour le suivi duquel des méthodes expérimentales existent. Nous nous sommes alors attachés à proposer un traitement approprié de telles données expérimentales pour déterminer les valeurs des paramètres physiques du modèle.

L'essor des moyens d'imagerie et d'analyse a joué un rôle clef dans le développement de la compréhension de ces objets de tailles si réduites. Jusqu'à très récemment, ces échelles n'étaient accessibles qu'aux équipements de microscopie électronique ou de rayonnements intenses imposants et onéreux. Dans les années 2000, plusieurs techniques cousines ont émergé et placé de telles échelles à portée de la microscopie optique par fluorescence. Une partie de ces travaux a été récompensée par le prix Nobel de chimie en 2014. Ces nouveaux moyens se révèlent particulièrement adaptés à l'imagerie des cellules et sont une source d'information complémentaire aux techniques existantes en ce qui concerne les virus. Cependant, au niveau le plus fin, le procédé même de ces méthodes modifie la notion habituelle de l'image. Les techniques (f)PALM et (d)STORM, dites de microscopie superrésolue, permettent de générer des cartes de probabilité de la position individuelle de protéines marquées avec un niveau de précision de l'ordre de la dizaine de nanomètres. De la même façon que la notion de contour est mal aisée sur un tableau pointilliste, une mesure précise sur un objet délimité par ces positions n'est pas évidente. Dans cette thèse nous nous sommes attachés à développer une méthode quantitative pour mesurer les dimensions des objets ainsi "imagés" en prenant en compte la structure particulière de l'information produite par les techniques de superrésolution. Cette méthode, construite sur la théorie statistique du maximum de vraisemblance, a été caractérisée sur des images de virus de l'immunodéficience humaine (VIH –l'agent responsable du SIDA) simulées ou réalisées expérimentalement pour cerner les forces et les limites d'une telle estimation.

Abstract

In this thesis work, we study the self-assembly of viral particles and focus on the analysis of measurements based on fluorescence labeling of viral proteins. We propose a theoretical model of the dynamic of viral proteins self-assembly at the cell membrane based on previous models developed to describe the in-vitro assembly of spherical viruses. We study the evolution of the populations in the successive stages of viral budding as well as the evolution of single particle within this framework. We also provide various data analysis to measure the physical values involved in the process: rate of aggregation during the bud growth, size and shape of the eventual structure. Viruses are biological objects unable to replicate without infecting an host cell since they lack part of the molecular machinery mandatory for genetic code replication and proteins production. Originally aimed at controlling the diseases they cause, the study of viruses is now rich of applications in medical and technological field (gene therapy, phage therapy, targeted therapy, bio-templating, cargo specific encapsulation, etc.). The existent models describing the self-assembly of viral proteins have successfully captured many features observed in the in-vitro experiments. We study the expected evolution when an open system is considered with an input flux of proteins and an output flux of released virion, characteristic of the in-vivo situation. We derive the population distribution at steady state and numerically study their dynamic under constant viral protein input flux. We also study the case of a single bud evolution which can be followed by its fluorescence emission. We study the possibility to estimate shape parameters at the single viral particle level such as radius and completion for the human immunodeficiency virus (HIV) from single molecule localization superresolution microscopy. These techniques known as (f)PALM or (d)STORM, record labeled proteins position with a precision of few to tens of nanometers. We propose an approach base on the maximum likelihood statistical method which is tested on both real and simulated images of fully formed particles. Our results suggest that it can offer a precision on the determination of the global structure finner than the positioning precision of the single proteins. This efficiency is however tempered when the parameter of interest does not affect the figures of merit to which the method is sensitive such as the apparent area and the image contours.

Contents

Introduction

The term *virus* comes from the Latin word for poison[106] which tells much about the way they are usually perceived. They are mainly associated in the common sense to great historical pandemics such as “Spanish” influenza (from 21 to 100 millions casualties depending on the estimations (Johnson and Mueller[70]), smallpox, and more recently acquired immune deficiency syndrome (AIDS) and Ebola hemorrhagic fever epidemic, or to less serious diseases as chickenpox, common cold, cold sores...

Viral infections on animals are also present in public mind due to the possible transmission of the disease to humans in several cases, as well as the economic issue when farming is concerned. The recent worldwide concern about avian influenza (highly pathogenic asian avian influenza A (H5N1) virus) illustrates both the social, political and economic impact of virus with global losses estimated in billions US$ (Mcleod et al.[93]). Also, many virus have known animal reservoirs, as pigs and birds for influenza, fruit-bats being suspected in the case of Ebola hemorrhagic fever (Laupland and Valiquette[81]). Historically, the term “vaccination” was coined from the from Latin vaccinus, which means "from cows", by the British physician Edward Jenner to designate his technique of inoculating people with cowpox virus to prevent smallpox (Harper[62]).

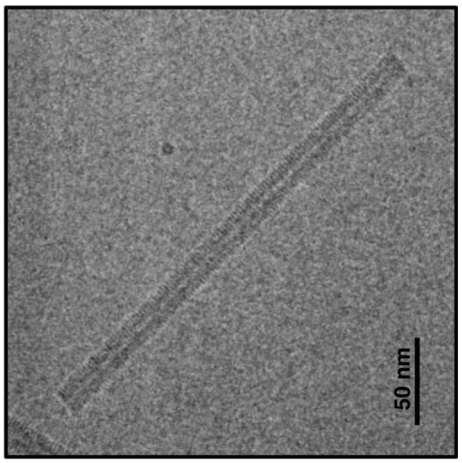
Despite the prominence of animal viruses for public opinion, viral diseases are known to affect a vaster part of the living realm. As a matter of fact, the discovery of virus was made studying a plant disease damaging tobacco crops. As the approach underlines key characteristics of viruses, we will present it in details before we expose the modern definition.

1.1 Viruses

1.1.1 The historical discovery

In the middle of 19th century, bacteria are already known and have been extensively studied, among other by Louis Pasteur (1822–1895). Yet the agent of rabies still escaped the detection in microscope’s observations. The size of bacteria is large enough to visualize them using optical magnification or filter them out from a solution using the filtering protocol designed by Charles Chamberland (1851–1931). In 1884, he successfully removed the typhoid fever bacteria from an infected solution using the filters currently named after him. Later, Dmitry Ivanovsky (1864–1920) working on tobacco plants infected by the tobacco mosaic virus (figure 1.1) showed that extracts from crushed leaves can transmit the disease to healthy plants after filtration on Chamberland filters. He concludes that the pathogenic agent is smaller than known bacteria

(a) A tobacco leaf with symptoms of infection by tobacco mosaic virus (©[111])



(b) The tobacco mosaic virus observed in cryo-electron microscopy (from Sachse et al.[113])

Figure 1.1 – The tobacco mosaic virus was the first discovered

and proposed that the bacteria could somehow produce a toxin causing disease on the healthy plants. The microbiologist Martinus Beijerinck (1851–1931) eventually repeated this experiment showing that further dilutions of the filtered solution remains infectious and that the process can be reproduced using leaves from the newly infected plant. The hypothesis of a toxin produced by bacteria in the former plant to cause the transmission of the disease was thus excluded (Zaitlin[135]). Martinus Beijerinck used the term virus to designate this newly discovered pathogen. He further characterized it as an entity that proliferate inside the living organism but that does not reproduces in the solution itself.

A virus is nowadays defined as a infectious agent unable to replicate without an host cell. The common characteristic of viruses is that they lack part of the molecular machinery mandatory for genetic code replication and proteins production. Viruses therefore depend on their host cell resources and machinery to replicate, and are therefore parasites. As already mentioned, the host cell is not necessarily an animals cell: plant, bacteria, fungi, any cell is potentially a viral target. A virus is generally specific of a (or several) host cell(s) and to the great diversity of cells corresponds a great variety of viruses. The number of known virus is constantly increasing since techniques that enable their visualization and description are quite recent as we will see in the following section.

1.1.2 Virus cycle

The simplest possible image to describe the essential nature of a virus is a box containing instructions to produce kits to assemble the same box altogether with copies of the instructions. If this box is left somewhere in a busy factory, many ready-to-assemble box are likely to be produced... The factory we are speaking about is the host cell which has the cellular machinery to assemble proteins from the nucleic acids sequence (the ribosome, mitochondria, etc.). The viral replication necessitates to access and fool the cellular machinery in order to make it produce viral elements instead of cellular ones. Cell being a closed and often compartmentalized environment, production of new viruses follows a succession of steps known as virus life cycle. Virus life cycle can be extremely complicated, as much as the host cell and its reactions to the

presence of viral material can be complex. Hence the viral strategy will be different whether its host is an eukaryote cell as animal cell where the genetic material is enclosed in a nucleus, or a bacteria without a nucleus. Viruses constitutions are also very variable (see section 1.1.3) leading to different disassembly and assembly strategies. However the virus goal is replication in any case and the following steps are mandatory in all cases figure 1.2:

— The target host cell must be recognized (using cellular membrane receptors).

— The genetic code of the virus has to come to contact with the cellular machinery, which imply its entrance inside the cell through the cellular membrane and its liberation (entry and uncoating).

— The genetic code must be replicated (transcription).

— The genetic code must be translated to produce viral components: the viral proteins (translation).

— Viral components including the replicated genetic code and the viral proteins must assemble to form the viral structure ((self)-assembly).

— The new virus must exit the cell.

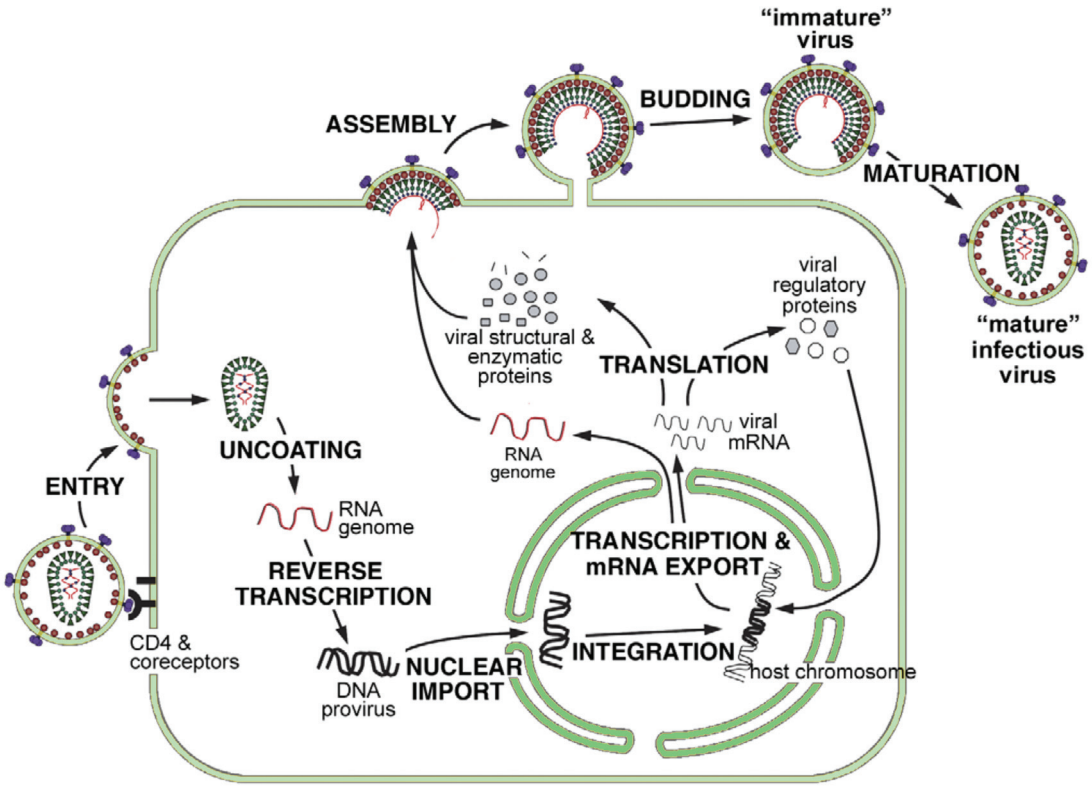


Figure 1.2 – HIV life cycle illustrate the steps of all virus infection and more specific behavior. The host cell outer/nuclear membrane is sketched in light green. As example of specific and complex strategy, the transcription of the HIV genome carried by the virus coded on RNA into DNA for integration into the host chromosome (retroviral strategy) and the production of non packaged regulatory proteins that interfere with the cell regulation system and a maturation step (from Ganser-Pornillos et al.[49])

Each step could of course be achieved through very different strategies, and a consecutive important variability of structure and interactions. However, in spite of this factors of diversity, the virion remains a simple structure at the interface between the protein level and the complex cellular machinery.

Interactions between cell and virus involve a rich physics with a promising range of fruitful questions from polymer physics, mechanics of fluid membrane, solvation, charge interactions, molecular bounds, crystallization, nucleation... Physical models based on simple concepts have already proven to be successful, for instance in explaining how capsid elements assemble (pioneer structural work by Caspar and Klug[23], Crick and Watson[31], Finch and Klug[42], see section section 1.1.3) and paved the way for many applications relying in "encapsidation" (see section 1.2.1). We will therefore detail the rich structure of viruses and how from simple protein component so interesting polyvalent containers auto-assembly is possible.

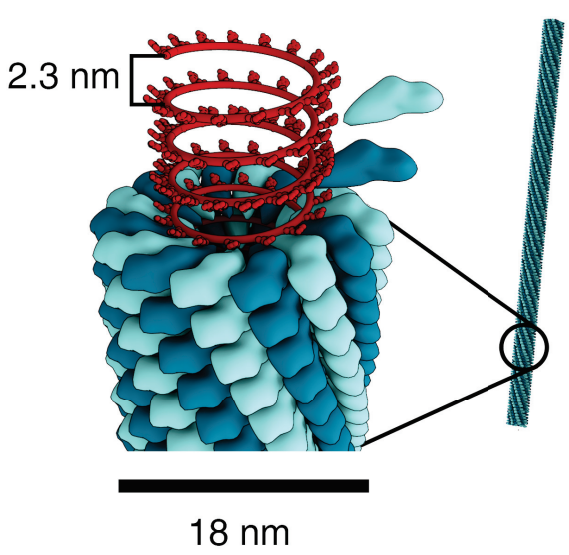
1.1.3 Viruses structures

The complete virus as it can be seen before it infects a cell is called a virion. The general organization consists in a genetic media made of nucleic acids (either DNA or RNA, simple or double stranded, (+/–)-polarized, depending on the virus), that code for the viral genome, protected by a coating of proteins called the capsid. This organization have been hypothesized by Crick and Watson[31] in 1956 –six year after they determined the correct structure of DNA– and completed in 1957 (Crick and Watson[32]) with the suggestion of the cubic and cylindrical symmetries as the most economic choice in term of construction : most of the positions in those constructions are equivalent so less protein variety are required to build it thus limiting the size of the genome to enclose. This geometric structure made of a large number of identical subunits is consistent with the crystallization of viruses first observed on the tobacco mosaic virus in 1935 by Stanley[123]. The proposition was to be confirmed two years later with the icosahedral structure of poliomyelitis solved by Finch and Klug[42] using the X-rays diffraction pattern.

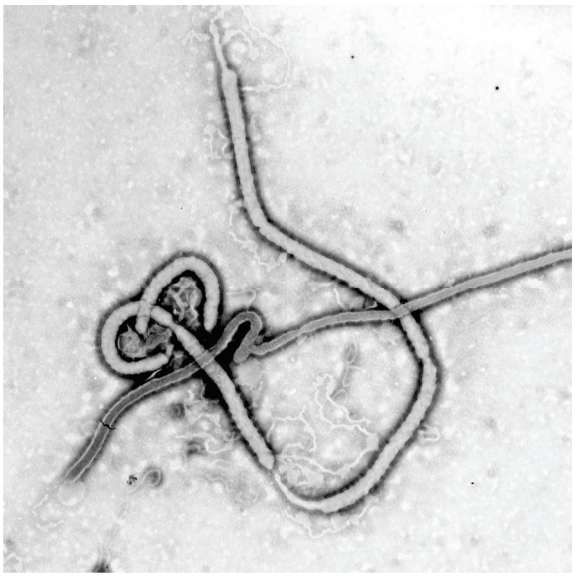
The self-assembly of the capsid is to a large extent akin to a 2D crystallization process in which capsid protein aggregate to form an ordered layer. The hexagonal (triangular) lattice is the planar crystal that allows close packing of identical units maximizing their interactions and is basic canvas adopted by the capsid units. A two dimensional structure is obtained by folding this network in one direction witch generates the tubular and helical structure that we will describe first. Forming a closed structure requires curvature in both dimensions, and pentagonal sites are necessary. They are indeed found in closed shape such as icosahedral, prolate, and irregular capsids.

Helical viruses We have already introduced the tobacco mosaic virus which is a helical virus family representative. The capsid proteins are stacked around the genetic material in helical structure giving a rod shape as the tobacco mosaic virus or a filamentous shape such as Ebola virus (figure 1.3b). Other well known animals viruses having a capsid with helical symmetry are rabies virus and measles virus, but they are also surrounded by a lipid envelope and described as enveloped viruses.

Spherical (icosahedral) viruses They are probably the best known and described viral capsids. The proposition of Crick and Watson[32] was that one type of identical capsid subunits



(a) Protein structure of the tobacco mosaic virus shown in figure 1.1b is a rod with right handed helical pitch of 23 Å with 16.3 protein per helix turn, inner diameter of 4nm and outer diameter of 18nm (Mateu[91]) (©[122])



(b) En election micrograph of an Ebola viral particle showing its filamentous structure. Filament diameter is 60-80 nm

Figure 1.3 – Ebola virus and tobacco mosaic virus are two helical virus

only (not necessarily a single protein, an assembly of protein is also possible) assemble so that each sees the same environment (see figure 1.4). This provides a good container requiring a minimal amount of genetic information. This theory was latter developed and generalized by Caspar and Klug[23] under the name of quasi-equivalence theory to allow more flexibility in the constitution than the sixty identical units inconsistent with the mass experimentally observed. Equivalent Interaction between neighbors are termed "quasi-equivalent" as one admits that the building block can withstand a slight modification of their conformation to adapt to a different environment depending on their position in the capsid so that hexamers as well as pentamers can be formed from the capsid proteins. When hexameric units are inserted between the pentameric ones in a concerted manner, so that those are maintained as summit of a equilateral triangle, the result is still an icosahedron with larger facets and involving more building blocs. In those constructions, the number of subunits involved is still a multiple of sixty: $n = 60 \times T$. T is called the triangulation number and must verify $T = h^2 + k^2 + hk$, where $(h, k) \in \mathbb{N}^*$ define the steps to go from one pentameric summit to the next in the lattice. T also counts the number of different binding conformations that are necessary to form the structure (figure 1.4). The T number provides a classification of capsids inside the icosahedral family. Of course, if one relax the genetic parsimony constraint to produce the least variety of different proteins, the large icosahedral capsids can be build out of T different building bloc for each type of binding site rather than one adaptive one. Several viruses are known to have adopted this strategy such as picornavirus. Furthermore, large variations are observed among viruses. In prolate capsids where the initial icosahedron is extended in one direction forming a cylinder constructed on pure hexameric sites (Mateu[91]). This leads us to the next family.

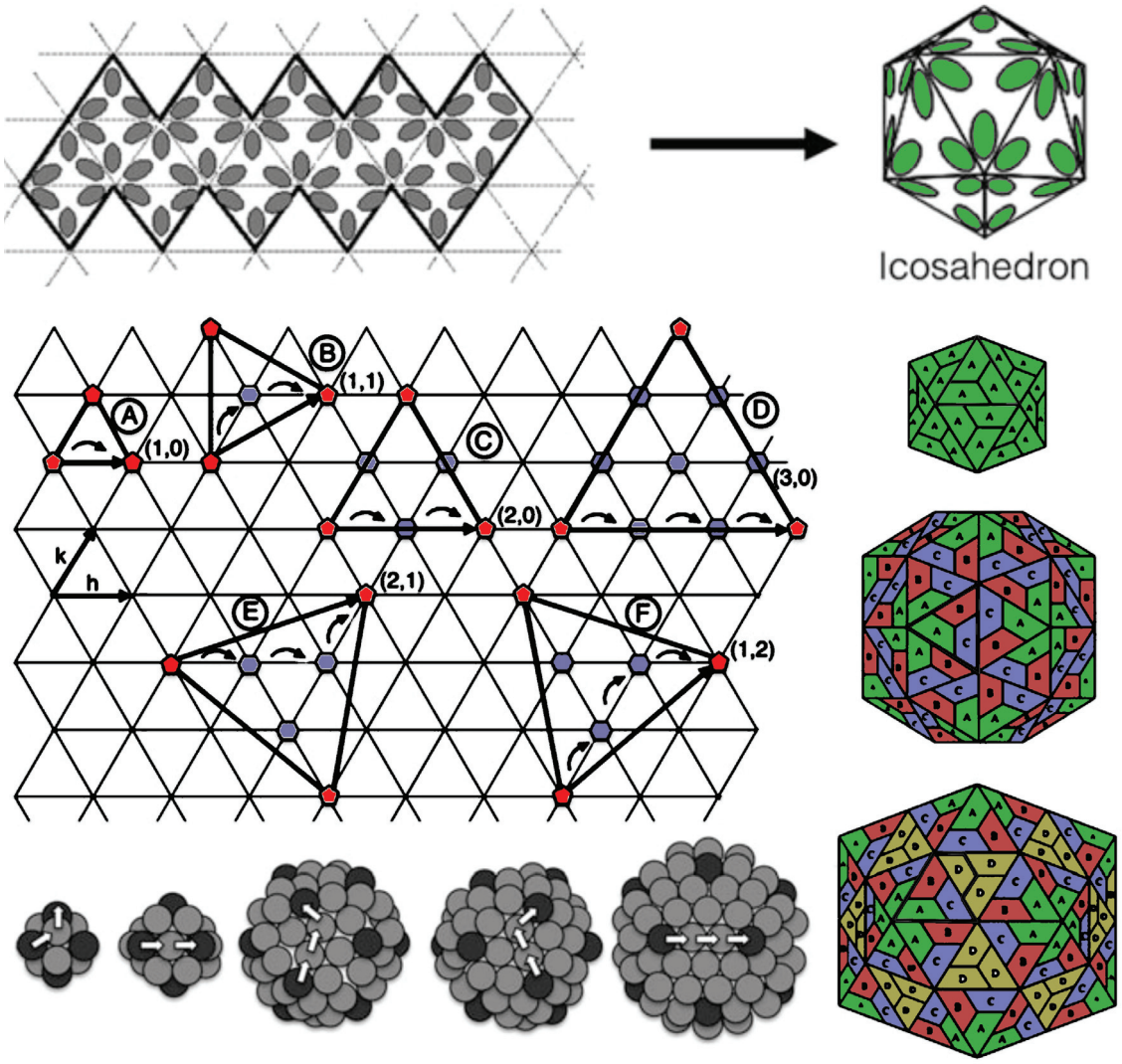
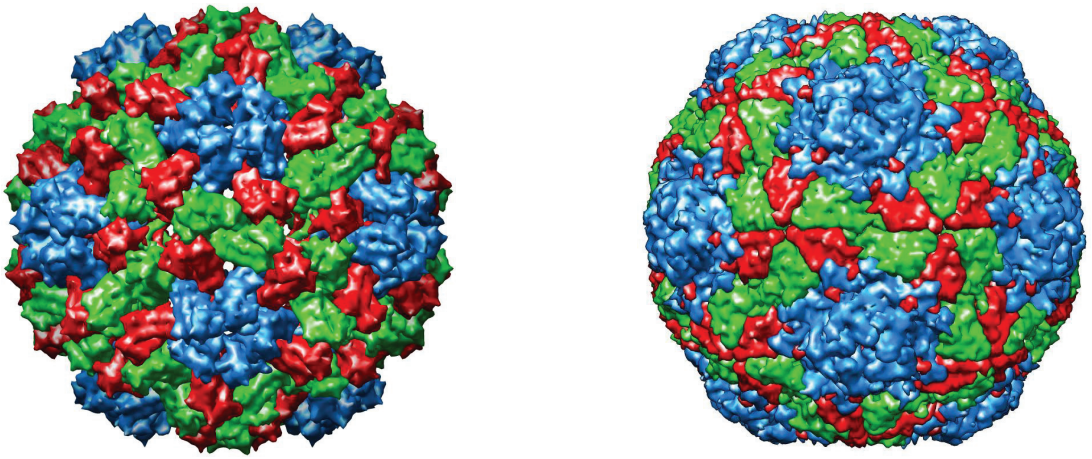


Figure 1.4 – (top) A portion of 2D triangular lattice of sixty identical capsid units folds in an icosahedron for adapted binding angle. Each capsid subunits, or capsomer, has the same environment in the resulting icosahedron, so it can self-assemble from identical subunits forming a container with a good surface/volume ratio. (middle left) The quasi equivalence theory assumes that capsomers can adapt to slightly different geometric binding configuration (not any more equivalent but "quasi") so that larger triangular facets (B,C,E or F, and D) can be made including hexameric coordination sites to enlarge the resulting icosahedral capsid (bottom). The new constructs, can be labeled by their T numbers.(middle left) in $T = 1$ all capsomer occupy an equivalent position (original idea), in $T = 3$ their are three different binding configuration, and four in a $T = 4$ capsid (from Mateu[91])



(a) The Cowpea chlorotic mottle virus icosahedral capsid (diameter 300 Å) is made of 180 identical proteins that assemble into dimers (capsomers) which in turn form the capsid (quasi-equivalence rule).

(b) The picornavirus capside (diameter 308 Å) is made of 60 replica of 3 different proteins (capsomers) designed to bind in non equivalent environments.

Figure 1.5 – Icosahedral capsids following $T = 3$ symmetry either following quasi equivalence (same subunits) or using different proteins (from Viperdb[22]).

Irregular viruses Members of this family show either complex capsid, such as the assembly of different part helical and (pseudo)icosahedral such as several bacteriophages that infect bacteria, or as the mature Human Immunodeficiency Virus (HIV-1) capsid which does not adopt a prescribed shape but rather varies between a closed cylinder or a conical shape. Recently discovered Pandora virus has a shape similar to a lemon.

Enveloped viruses Capsid is not the only structural feature of virion. Enveloped virus exhibit in addition, as the name suggest, a lipid bilayer taken away from the cellular host in the late stage of the virus life cycle. It can be either a portion of the cellular membrane or of the cellular compartment where assembly took place, or lipids recruited during assembly stage. Viral proteins are generally embedded in the lipid bilayer that are typically needed by the virus to recognize and enter its target cell. In addition a supplementary layer of proteins called the matrix mediates the interaction between the membrane and the inner capsid. The virus matrix, also formed of a viral protein lattice partly embedded in the lipid membrane, is generally less organized than the capsid.

Human immunodeficiency virus immature structure In this thesis, we focus specifically on the case of enveloped virus that assemble under the cell membrane such as Human Immunodeficiency Virus (HIV-1). This section presents a closer look at the structure of human immunodeficiency virus virion shortly after it has left its host cell.

As other viruses, human immunodeficiency virus undergoes a maturation phase that brings strong modification to its structure. It is only once it has passed the maturation phase that the virus particle is able to infect a new cell. Before this stage, the human immunodeficiency virus capsid is still not assembled and the virion is called immature. The human immunodeficiency virus is an enveloped virus. Just after its release from the cell, the external lipid

membrane is spherical, enclosing the viral proteins and two identical copies of the viral (-)RNA. A strong variability is observed: the outer diameter of the immature viral particles varies at the level of the product from a single experiment: 133 ± 17 nm (Wilk et al.[133]), 126 ± 10 nm or 130 ± 17 nm (Carlson et al.[20] using tomogram or cryoEM)). The standard deviation indicated is the dispersion of the particle size distribution (much larger than the measure precision in electron microscopy). Inside the immature particle, Wilk et al.[133] reported a incomplete spherical shell of the structural protein gag[1] enveloped by the lipid membrane (figure 1.6). The gag protein of the human immunodeficiency virus structure is a very long polyprotein of

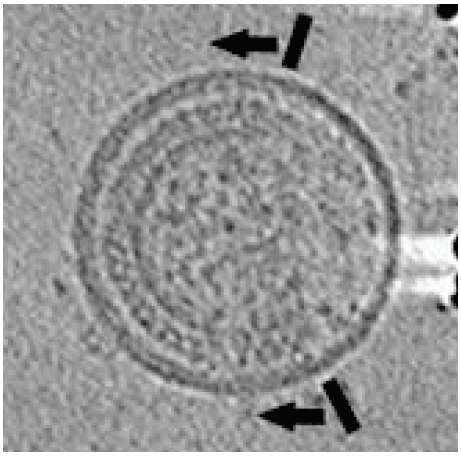


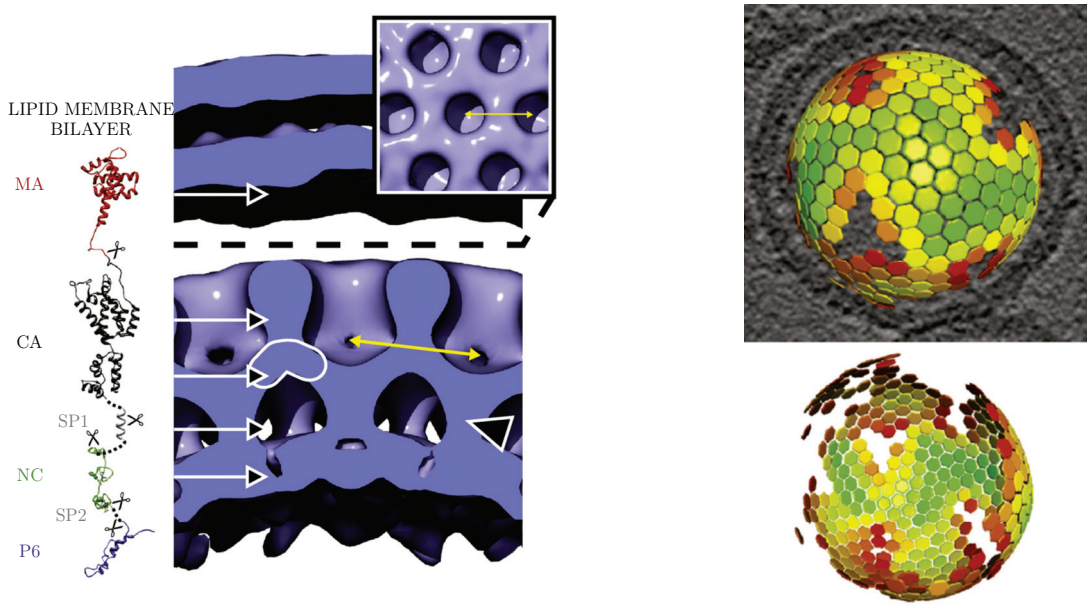
Figure 1.6 – Cut in the reconstructed tomogram of an immature virion (from Wright et al.[134]). The radial organization and incompleteness (black arrows) of the gag layer appears clearly

$\simeq 25$ nm in its conformation inside the immature capsid (Wright et al.[134]) with well identified domains (figure 1.7a):

- MA (matrix domain), binds directly to the inner leaflet of the plasma membrane by insertion of a hydrophobic group and binding with several phospholipids. It is however not necessary to observe particle formation (Ganser-Pornillos et al.[49]).
- CA (capsid domain) tightly associates in an organized lattice under the membrane and is critical to observe particle formation. Cleaved and liberated in the maturation process, this domain is responsible for the capsid formation inside the particle.
- SP_1 is a spacer domain also contributing to gag-gag interaction and mandatory for lattice formation.
- NC (nucleo-capsid) domain also participates to the gag-gag interaction to assemble the gag lattice and also binds the viral RNA for packaging in the forming particle. It latter folds the viral RNA in a compact structure during maturation steps.
- SP_2 is second spacer domain.
- P_6 does not affect the *in-vitro* assembly but is proved necessary for the proper release of the virus at the very end of the viral assembly (Göttlinger et al.[55]).

Inside the immature particle Wilk et al.[133] showed that the gag proteins are anchored by the MA domain into the lipid layer and radially oriented with the P_6 domain pointing towards

1. gag stands for "group specific antigen" which designates the gene sequence that code for the structural protein, as well as this protein itself, in the retrovirus family. The precise form of these proteins as well as their properties are not identical from one virus to another (see Briggs et al.[18])



(a) The different domains of the gag polyprotein (left) along with the electronic density inside a immature virion (adapted from Briggs and Kräusslich[15]). The MA domain is partially embedded in the protein bilayer, making if thicker. Underneath CA domain organize in an hexameric lattice (inset) with a 8 nm spacing (yellow arrow)

(b) fit of an hexameric lattice in the tomogramm density shows that gag organize into a continuous coating inside the immature virion with several defects that allow curvature of the lattice and an important gap (from Briggs and Kräusslich[15]) as showed by Carlson et al.[20]

Figure 1.7 – Organization of the protein coat underneath the lipid membrane bilayer

the center of the particle. The number of gag forming the incomplete inner protein layer is of the order of 10^3, and more accurate estimations vary much, depending on assumptions and experimental technique used: the gag proteins content of a virion was reported to be 1400 gag proteins (Zhu et al.[138] –using proteins ratio assumptions), 4900 (Briggs et al.[17] –assuming full coverage of a 145 nm particle), 2400 ± 700 (Carlson et al.[20] –assuming 60-70% coverage, using tomography and scanning transmission electron microscopy). The hexagonal organization of the gag lattice was suggested by Nermut et al.[104] and identified by Briggs et al.[18], Wright et al.[134] with a measured spacing of 8 nm between two hexamers centers using cryo-electron microscopy tomography. Furthermore, the gag lattice coats only $^2/_3$ of the inner surface of the lipid membrane inside the immature particle (Carlson et al.[20]). This organization suggests a revised model of the viral self assembly.

1.1.4 Human immunodeficiency virus cycle late stages: budding and release

The building of the viral particle inside the cell with its exit and maturation constitutes the late stage of the human immunodeficiency virus cycle. As many enveloped virus, its builds directly on the membrane. During the process, the lipid membrane is locally curved around the protein coat in a structure called a bud. This step of virus cycle prior to the release outside

the cell is therefore called the budding of the virus. Based on mutation studies, cryo-electron microscopy (Carlson et al.[20]) and fluorescence observations (Jouvenet et al.[72]), a sequence of the human immunodeficiency virus formation steps from the beginning of viral proteins production inside the cell can be hypothesized (figure 1.8a).

Yet this scheme is not sufficient to explain the variability and incompleteness of the immature protein coat. The issue was thus challenged by proposition including a more detailed description with many exit channels as showed on figure 1.8b. Those scenarios give a synthesis of the information extracted from fluorescence and structural characterization experiments. The various actors that take part in the viral life cycle are more and more exhaustively determined and the succession of their intervention within the virus cycle described. Going deeper in the details of the viral assembly now requires a quantitative approach to describe the consequence of a physical parameter modification, such as a constituent concentration or the temperature, on the measurement that is the form of the size distribution or the time evolution of the concentrations.

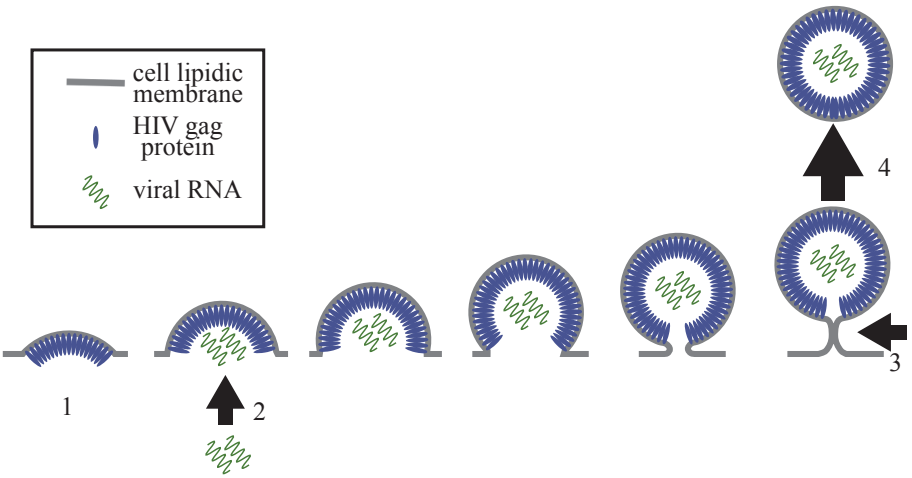
The development of a quantitative approach of the observations and models based on physical rules enriches considerably the understanding of the virus. The next section is aimed at broader overview of the recent advances on this field.

1.2 Physical Virology

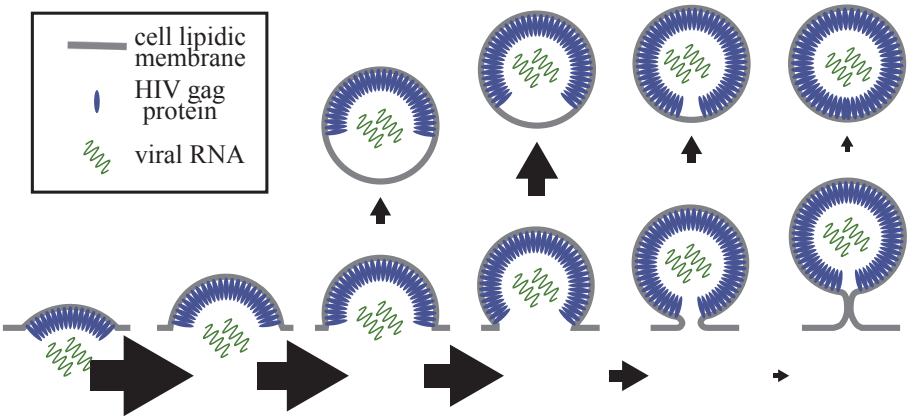
The development of new means of measurement based on the 20^{th} century discoveries (among other X-rays, electron wave-particle duality...), discussed in section 1.3, contributed for a large part in a precise description of the viruses. It granted access to scales unreachable until then. Also physical based models took also an important share in the progress with a powerful description of polymer physics that rules conformations of the genetic media, of the mechanical properties at low scales with lipid membrane elasticity theory for instance (Helfrich[63]), and of large population statistical behavior such as the micellization phenomenon. The use of the physical concepts of forces, deformations, energy, entropy and equilibrium gives a universal framework to understand the observations on the organization and dynamic of the virus in more and more complex situations. The cell, that is the reference biological entity, is still a very complex system regarding physical description, however the physics rules still apply and the description of simpler sub-systems at the scale of the virus is not any more a thought experiment.

1.2.1 Many exploration directions and applications

There are many interesting points related to the structures of viruses. Promising applications largely use it. For instance vaccination is more efficient using non pathogenic viruses "decorated" on their surface with antigens of the target pathogen agent, than with antigen alone. Vaccination relies on a successful recognition of a pathogen antigen that generates immune memory. Since immune system has co-evolved with viruses in order to recognize and react to viral presence, its reaction is much stronger than to the antigen alone (Andrade et al.[4]). A good knowledge of viral structure is a key factor to find a protein residue exposed at its surface as anchor. This also apply for functionlizing viral surface with markers or molecules in order to image (Steinmetz et al.[124]) or deliver a treatment to their specific target cells (Ashley et al.[6]); as well as for antiviral drug design. But outgrowth of the virus structure studies fall



(a) Late stages of the human immunodeficiency virus cycle: budding an release. From left to right are shown different stage of the bud growth: the gag structural proteins bind to the membrane, diffuse and aggregate. This initiates a budding site (1). As more free proteins aggregate the bud grows. The viral RNA and other non structural viral proteins are selectively packaged in the course of bud growth (2) and the protein aggregation deforms the cell membrane. Active steps is needed to trigger the scission of the virion from the membrane (3) eventually releasing the immature particle outside the cell (4).



(b) An amended description of the human immunodeficiency virus release in the cell to explain the variability of released particles completeness. Each bud stands as representative of the population present on the membrane and arrows area are proportional to the net flux between population (adapted from Carlson et al.[20])

Figure 1.8 – The budding of enveloped virus such as the human immunodeficiency virus

outside of the medical field. The viruses well defined structures at the manometer scale can be used to template nanoelectronic devices fabrication (Bruckman et al.[19], Moon et al.[97]).

Yet, the "static" view of the structure evolved to a dynamical description. Fluctuations of the capsid conformation, called breathing, have been shown (Lewis et al.[85]) as well as strong, irreversible conformational changes called maturation. This fosters the study of capsid as mechanical objects that deform, can withstand pressure or break under strong stress, and how this properties are affected during virus cycle (Castellanos et al.[24], Mateu[90]).

The genome conformation inside the capsid is not frozen either. The pressure it applies on capsid walls have been predicted with model from the confined polymer theory, osmotic pressure and charge interaction (Angelescu and Linse[5], Muthukumar[103], Siber et al.[117]). Interaction with the native viral package, RNA or DNA inside the capsid has been extensively studied. In several cases, packaging inside the fully formed capsid is an active process requiring a motor and consuming adenosine triphosphate, whereas in other cases, electrostatic interaction between the negatively charged polymer and the capsid is likely to promote assembly. Packaging and expulsion forces have been experimentally measured at the scale of the single virus on Φ29 phage using optical tweezers (Smith et al.[120]) emphasizing the strength (57 pN) and yield of the molecular motor, as well as the experimental precision available to manipulate single polymers and viruses.

A remarkable point in the viral dynamic is that the viral organization emerges, in most of the cases, from the individual properties of the constituents. The structure assembly and self organization has been early demonstrated *in-vitro* without the need of adenosine triphosphate the biologic energetic unit that fuels active process. This was demonstrated for instance with tobacco mosaic virus showed in figure 1.1 (Fraenkel-Conrat and C.Williams[48], and cowpea chlorotic mottle virus pictured in figure 1.5a (Bancroft et al.[9]). These experiments showed an important dependency on the structural proteins concentration, the pH and the ionic strength of the solution. Other viral proteins in contrast do not assemble in the absence of their genome.

The physical rules that preside over virus formation set the basis of self-assembly and its constraints. The capsid structure must be stable enough to protect the genome and yet weak enough to allow for a versatile assembly/disassembly and deliver the genome during infection. The assembly must be in addition exceptionally efficient to proceed in the cellular complex environment whereas competitive events are likely to perturb it sequence. We will now look at the physical principles that rule this self-assembly.

1.2.2 Viral assembly

At the global level self assembly of any viral structure is controlled by the specific binding of its components (figure 1.9). The components are furthermore restricted to a limited small number of varieties by the biological constraint (size of the genome, efficiency, etc.). As a rule, viruses are made from the repetitive aggregation of the same subunits. The successful idea of Caspar and Klug[23] that the constitutive units should be closed packed and maximize contact is an optimization concept. The isolated constituent are actually free to build whatever structure. Anyhow a structure with too few and weak coordination between the constituent is unstable and will disappear shortly after it appears. Over a long observation time or equivalently for a large sized system, the more stable assemblies are the prominent ones. This is at least true at null temperature. In contrast, at non zero temperature, thermal fluctuations tend to favor less ordered structures that can result from a larger number of possible combinations or

assembly paths. This competitive effect are described by the enthalpy and entropy contribution. This is the principle of statistical physics and thermodynamics. Quantitative derivation of this principle is made using the thermodynamics potentials. The optimum of the potential indicates equilibrium between the driving force of the system corresponding to the measure at the level of the global system.

Application of the thermodynamic principle under the specified constraints of given volume and temperature leads to the minimization of the free energy potential. How to derive this potential from the microscopic physics of the system component is theoretically established by statistical physics, but the complexity of the task is tremendous because of both the physical needs for a proper description of all the interactions and degrees of freedom and the calculation itself. Useful models rely on approximations and effective descriptions that model the properties of the components in the system at the desired level of description.

1.2.2.1 *In-vitro* equilibrium assembly of empty icosahedral capsids

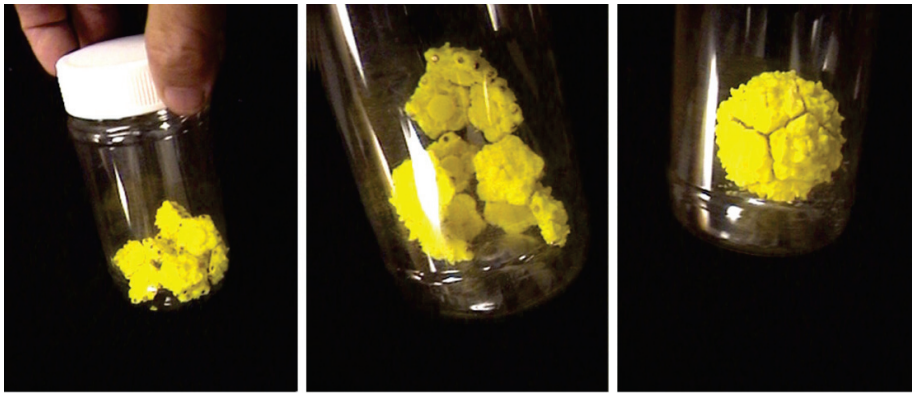


Figure 1.9 – Self assembly of an empty capsid driven only by the capsomers interaction and agitation. Small magnets on the edges of the pieces mimic the interaction between the capsomers and shaking the box mimic thermal agitation (from Olson et al.[105])

The quantitative distribution of self-assembly products of empty icosahedral capsid *in-vitro* is dominated by the isolated capsomers and the complete capsid. The capsomer being the most stable assembly of capsid proteins that can be found in solution (for instance the dimers for figure 1.5a), and they are elemental block of the capsid assembly. It agrees well with the thermal equilibrium state of a free energy potential dominated by the energy involved in close distance binding of capsomers otherwise not interacting and thus described as ideal gas for the entropic contribution. Such description is equivalent to a succession of chemical reactions where capsomers are added or removed from the forming capsid (Zlotnick[139]). The binding energy hence described can be tuned changing the pH of the solution with an acid or its ionic strength by addition of salts. Consistently, the binding energy main contributions are identified, at the microscopic level, as the electronic and hydrophobic interactions (Kegel and Schoot Pv[75]). Both come from the properties of the residues of the proteins forming capsomers that are exposed at their surfaces: basic or hydrophobic. Basic residues often present at the inner face of the capsid create excess of positive charges leading to a repulsion that can be diminished by the screening of added salt or a lower pH whereas hydrophobic interaction becomes potentially stronger at higher temperature. These observations are supported by the *in-vitro* assembly of

hepatitis B virus (Zlotnick et al.[142]). However, such an equilibrium description fails to explain the observed hysteresis of the capsid stability. Experiences starting from pure capsomers or from pure complete capsids do not yield the same final concentrations following equilibrium rule (Luque Santolaria[86]). This last result suggests a kinetics effect.

1.2.2.2 *In-vitro* kinetics of empty capsids

The equilibrium state derived from thermodynamics corresponds to an dynamic equilibrium where flux between reacting species are still present at the microscopic level (capsids keep forming and disappearing) but balance at macroscopic level. This establishes a compatibility relationship between the kinetics rates of all the reactions generating or consuming a specie and the equilibrium concentrations called global balance. A more restrictive relationship, where each reaction equilibrates with its reverse one independently (Van Kampen[131]), holds when reactions have time reversibility property. The description from Zlotnick[139] thus also provides a basis to describe the kinetics of the system with additional assumptions on the rate constants. A model, constructed accordingly in two steps: formation of a stable nucleus and elongation towards completion, was found to agree with the observed evolution (Kegel and Schoot Pv[75], Zlotnick et al.[140], Zlotnick and Stray[141], Zlotnick et al.[142]). However refinement of the description of the nucleation in two steps was needed to quantitatively fit the concentration evolution of brome mosaic virus (Chen et al.[29]). The framework of classical nucleation theory was proposed by Zandi et al.[136] as a unified approach to the capsid formation. All those works were aimed at the description of the well defined *in-vitro* experimental condition where the number of capsomer is fixed in a closed system. Our work is build on the same ground that these models. We will detail them in chapter 2. We consider the extension of this approach to model the aggregation process of the viral protein in the cell.

1.2.2.3 Towards *in-vivo* models

The cell is an open system for viral proteins: new monomers are constantly introduced in the system, build by the cellular machinery, and formed virions escape the cell to infect new cells. We considered the extension of the *in-vitro* approach to model the aggregation process of the viral protein under the membrane in such an open system. We studied more precisely the onset of viral production, when the system has not reached its steady state. Indeed, in contrast with the closed system laboratory experiment, an infected cell producing virions consumes its resources and can eventually die. Therefore the long time limits might not be the relevant quantities to study. In chapter 2 we will look at how this flux in the system affect the kinetics of the capsid formation.

A complete description of the capsid formation would include the effect of other constituents. The more obvious one is the genome which is compulsory to obtain self-assembly in several cases *in-vitro* and not necessary in others (like cowpea chlorotic mottle virus or B hepatitis virus) however mandatory *in-vivo*. The complexity of the problem increases when the dynamic of the capsomers interacting with the polymer is to be taken into account. In silico-experiments as molecular dynamic and coarse grained simulation have thus been used to study the co-assembly. Elrad and Hagan[39] used a coarse grain simulation approach to provide a detailed phase diagram of the assembly success depending on the polymer length and binding energy. They conclude to an enhanced efficiency of capsid growth when a polymer binds the capsomers. If the binding between units is low enough and the adsorption on the polymer is strong enough,

the phenomenon resemble more condensation, where all the proteins aggregate first and then reorganize in the most stable conformation as around a template (Hagan[60], McPherson[94]). However for a large range of concentration and binding energies, effect of the polymer binding capsid subunit remains consistent with the classical nucleation. A latter work from Mahalik and Muthukumar[87] reports also that the effect of the polymer binding capsid subunit is consistent with the classical nucleation theoretically at higher effective concentration of capsomers. Indeed capsomers bound by the polymer are maintained closed in a smaller volume.

For enveloped viruses, the assembly takes place at the membrane which must be deformed in order to form the bud. The elastic properties of the fluid membranes such as the cell lipid bilayer are described by the general Helfrich theory (Helfrich[63]). The free energy associated to the deformation by a rigid coating during the budding process as been computed (Deserno[36]) and analytically derived in different limits (Foret[46]).

To confront the description of the model to the situation in cells, we envisaged fluorescence microscopy as we were in contact with the team of Suliana Manley at the Ecole Polytechnique Federale de Lausanne, who developed superresolution fluorescence microscopy techniques to monitor human immunodeficiency virus HIV-1 budding. For obvious reasons the transfection is made so that culture cells produce non infectious viral particles, with the same structure as the authentic wild type virion, designated as "virus like particles". Different techniques based on fluorescence are promising for the *in-situ* observation of the virus cycle. They are introduced in the next section.

1.3 Imaging viruses life

Understanding of the viruses cycle and interactions with the host cell relies on numerous experimental techniques providing the different measures the biophysicist needs: atomic structure, deformations, forces, velocities, diffusive constant, concentrations, etc. We briefly introduce here a non exhaustive list of classical techniques related to the imaging of the structure of the viruses. Our aim is to indicate in which context the analysis methods that we develop in this manuscript find their place. The description of the fluorescence microscopy and its specific use in virology is then developed in more details before we present our work in this "in this light".

1.3.1 Comparative overview

The figure 1.10, shows the different useful length scales in the study of viral world: organism, cell, virus, proteins and long polymers, or atoms. Depending on the scale of interest, techniques with the adapted resolution must be chosen. When alternatives exist a choice can be made according to strengths and drawbacks of each method depending on the goal pursued. Good biophysical characterization generally depends on their combination.

Nuclear magnetic resonance Structure of molecules can be resolved by probing how the environment of atom nuclei affects the resonance frequencies associated to their magnetic momentum inversion and precession while relaxing (Larmor frequency). A magnetic momentum is associated with any non zero spin nuclei such as H^1 and C^{13}. During the measure, a high static magnetic field imposes an energy difference between aligned and anti-aligned states. As sample nuclear spins are aligned in the static field, a carefully chosen sequence of pulses is

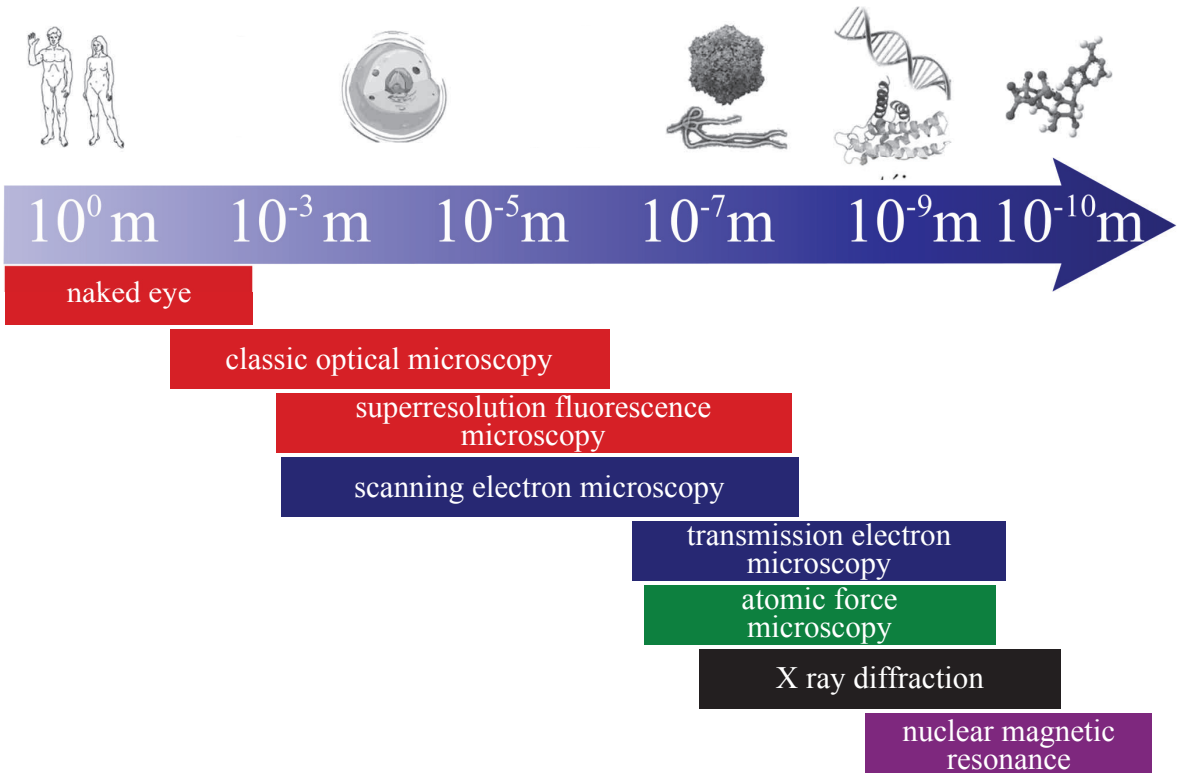
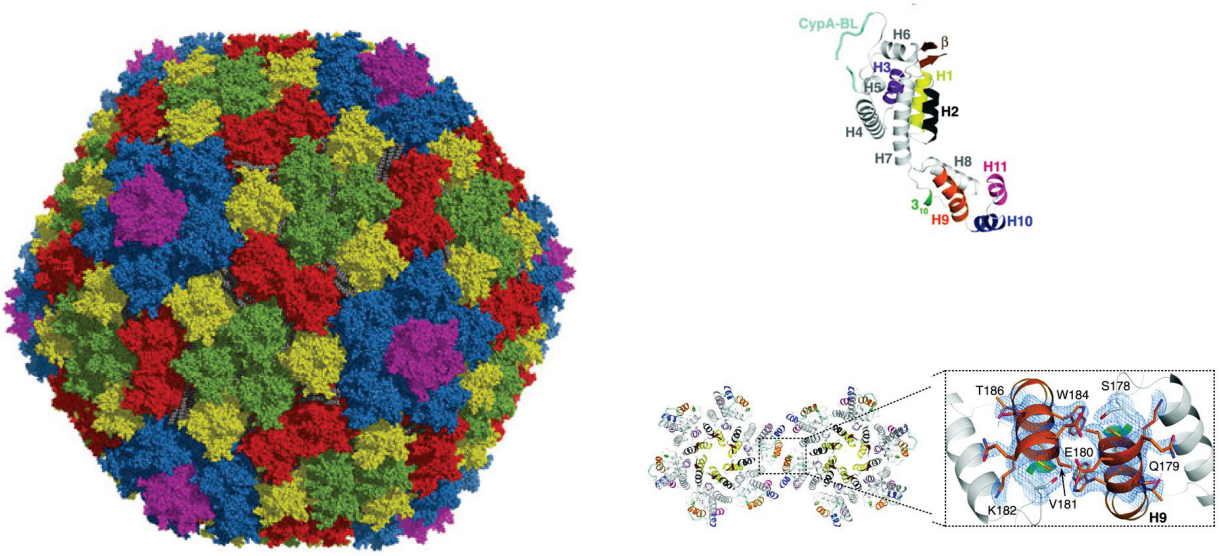


Figure 1.10 – The various methods to get an insight of the structural features of an object. Optics (red) imply labeling of small objects with dyes to enhance contrast and is limited by the diffraction bound. This limit was recently overcome using superresolution techniques (see section 1.3.3). High energy electrons (blue) have a smaller wave length at the expense of data post-treatments, and demand an arbitration between low contrast, labeling and risk for the sample integrity. At lower scale, only indirect imaging methods apply, either by probing local interactions (atomic force microscopy) or by solving the structure identified by its response signal: diffraction pattern in X-rays or nuclear spin relaxation spectra in nuclear magnetic resonance.

sent at a frequency resonant with the transitions between aligned to anti-aligned spin of the atom of interest thus modifying the initial orientations. A scan in frequency would provide a spectrum of resonances, but a pulse excites a window of frequencies around the frequency of the carrier signal allowing to probe all the spectrum space at once. Magnetic fluctuations due to sample relaxation can then be recorded. The spectra generated by the procedure is then used to deduce the molecular structure that satisfy the atomic environment recorded. The technique uses purified molecules generally in solution. More details can be found in Cavanagh et al.[27]. As the response time of the technique is very short, it is possible to use it to study dynamical properties and conformational changes. A detail monograph of nuclear magnetic resonance and virus structure is available in Mateu[91]. Nuclear magnetic resonance has been applied for the determination of structures of different viral proteins or protein sub-domains such as the human immunodeficiency virus gag MA and NC (Turner and Summers[129]) in conjugation with X-ray diffraction on crystallized subdomains.



(a) Reddy et al.[109] reported the larger structure determined by Xray diffraction: the full human adenovirus, diameter 95 nm at 3.5 Å resolution (illustration from Mateu[91])

(b) Structure of MA gag isolated subdomain (top) reported by Gres et al.[56] from crystal Xray scattering: the resolved electronic density (blue mesh) at 2.4 Å is precise enough to position the protein residues (bottom)

Figure 1.11 – The Xray scattering technique

Diffraction methods (X-rays, neutrons...) A crystallized structure diffracts the radiations in distinct directions selected by constructive or destructive interferences. The diffraction pattern obtained is used to solve the structure at atomic resolution. The inverse problem is not simple since only the light intensity is recorded and the phase information has to be reconstructed, based on experimental techniques and chemical models. The technique demands regular enough structures that can be purified and crystallized in a fairly large amount. Generally portions of the full structures are used : part of a large protein, proteins figure 1.11b... But regular viruses also have been successfully crystallized and reconstructed with a resolution allowing the determination of protein tertiary conformations figure 1.11a. This is still the most precise method at a time when crystal can be made, even if improvements in electron microscope captors tend to reduce the gap between the two techniques. A known drawback is that conformation can change between the native structure *in situ* and the crystallized form. In solution, the intensity of the signal is much lower and spectra are broaden and the precision is impaired (Small Angle X-ray Scattering).

Electron microscopy Two different set-ups are used with very different resolution powers and possibility. Both of them need to image in vacuum so that liquid phases are excluded, stability is essential at this resolution and object must be fixed. Surface of conductive bulk samples can be imaged using the re-emitted electrons excited by the electron incident beam. This is known as scanning electron microscopy and demands a metallic coating of most biological samples that are insulators. It unveils surfaces shapes with a strong contrast and a large field of view. However it is by far less precise than transmission electron microscopy and not much used to study viruses. Transmission electron microscopy is in contrast limited to the study of very thin samples (less than 1 µm) but reaches precisions of few angstroms. This second method

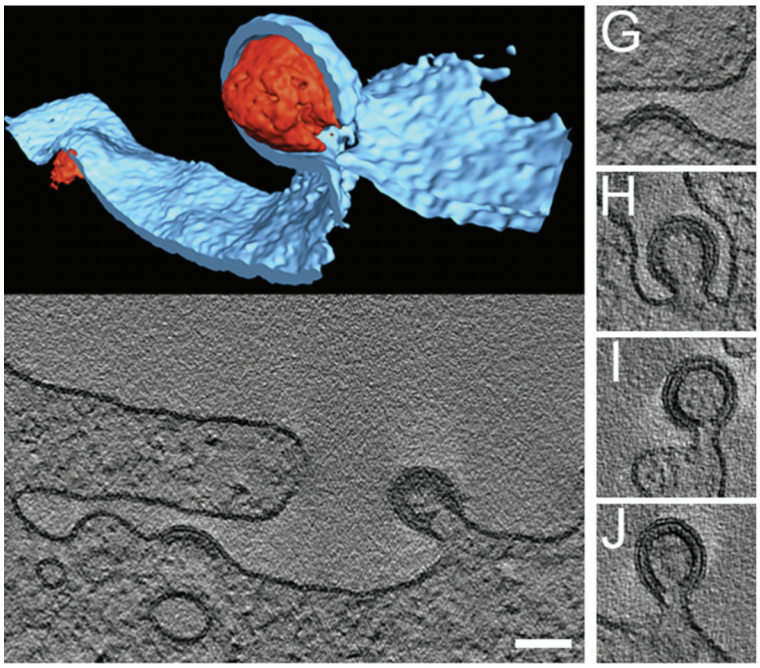


Figure 1.12 – Tomogram showing two budding sites of the human immunodeficiency virus gag protein (red) at a cell membrane (blue) (top left). A slice through the above tomogram –scale bare 100 nm (bottom). Column on the right (G,H,I,J) shows tomogram slices through other budding site (from Carlson et al.[20])

however suffers from the very weak contrast between organic molecules and, since electron beam damages the sample, a trade-off is necessary between contrast, focus and magnification. Production of contrasted images can be achieved via staining labeling the object with heavy atoms absorbing electrons with an inherent limit in the resolution of 20 Å an the inner part of the structure is not accessible. Very thin sample can be obtained by ultramicrotomy (50-100 nm). Fine study of the structure under the micrometer range is largely improved when the sample is frozen in a liquid film ($\simeq 0.2\,\mu m$) at $T < 100\,K$ with sufficient speed to vitrify the water into amorphous ice therefore maintaining the structures in their native conformations and avoiding electrons scattering. Two important issues arise: the signal to noise ratio is extremely poor, and the image is a projection of the object density in the focal plan. Image orientation and averaging is used in the Fourier space to produce high contrast images when the structures studied are mono-disperse or (quasi identical) enough and show clear symmetries (icosahedral or helical). This method leads to high resolution density maps reaching near atomic resolutions, but necessitates tens of thousands of individual images and demands both experimental time and computational power. Extension to non regular viruses has been carried out in cryo-electron microscopy tomography where the same virus is imaged under many different angles and reconstructed by back projection (see figures 1.7 and 1.12). However, the resolution is limited by the number of acquisitions as imaging degrade the sample. An extended review of the technique and its application to virus is made by Castón[26].

Fluorescence microscopy Fluorescent microscopy is an attractive technique to easily and directly record the spatial organization of a selected target at small length scales with a high contrast.

A molecule that absorbs an incoming radiation transits from its ground state to an excited state. There are many different ways for an excited molecule to return to the ground state and dissipate the absorbed energy. Fluorescence consists in the emission of light during the transition to a lower energetic state. As there can be several non radiative transitions before this emission, the exciting transition often corresponds to a higher energy (lower wavelength) than the fluorescence emission, allowing for a non visible ultra-violet excitation and a visible green emission for instance.

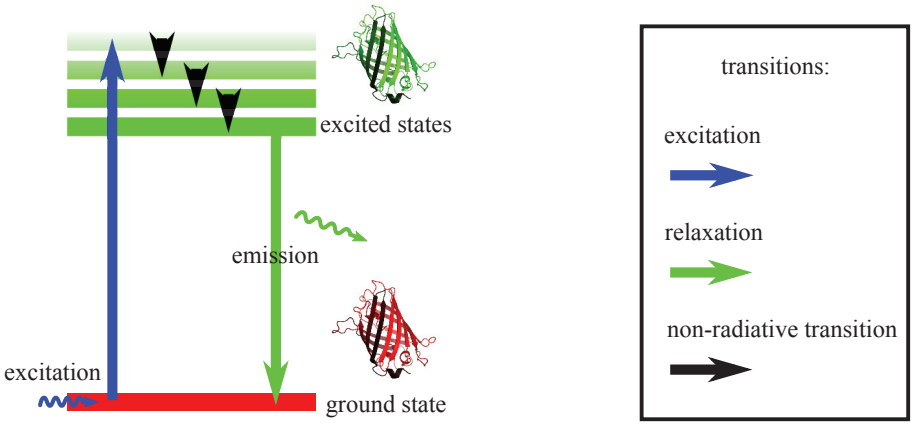


Figure 1.13 – Fluorescence principle: excitation light brings the molecule to excited state which almost instantaneously returns to ground state with a fluorescent light emission

As improvements have been made in the optical devices, and in the ability to label the object of interest *in situ* by anti-body labeling or modification of the expression gene to obtain a label with fluorescent properties fused in the expressed protein structure, the technique has become more valuable and widely used in the cell biology field. But the spatial resolution that can be achieved with an optical device is fundamentally bounded by the diffraction phenomenon. The image of a point-like emitter produced is not a point but a spot, called the point spread function (psf) of the imaging device. As a result, the image is blurred with a typical length equal to the width of the point spread function. This width is of the order of magnitude of the average emission wave length λ as given by Abbe's law which links it to the lateral resolution $R_{x,y}$ in the image plane and the numerical aperture of the microscope η (Abbe[1]):

$$R_{x,y} \simeq \frac{\lambda}{2\eta} \tag{1.1}$$

This limit therefore establishes both the typical distance between two simultaneous emitters that it is possible to distinguish and the minimal excitation spot that can be created in order to selectively excite the fluorescent labels in a region of interest with the mean excitation wave length. The numerical aperture is of order one ($\eta \simeq 1.5$ for most modern highly corrected microscopes (Sengupta and Lippincott-Schwartz[114])) and visible light has wave length in the hundreds of nanometers range. The resolution limit fixed by the optical microscopy is therefore of the order of few hundreds of nanometers. This boundary to accessible spatial features lays above typical sizes of organelles trans-membrane receptors, microtubules of the order of tens of nanometers and of great interest in understanding the inner cell organization, and even further apart from the typical size of their constitutive proteins of few nanometers.

Actually, in fluorescence microscopy, diffraction affects both the excitation and the consecutive fluorescent light emitted. It thus also sets the limit of the excitation pattern that can be produced.Reduction of the excitation pattern is one solution to enhance image performance by depleting the background emission. This is the spirit of total internal reflection fluorescence microscopy (TIRF). The volume of excitation is however still limited by the diffraction phenomenon. In the same logic, superresolution techniques such as stimulated emission depletion (STED) microscopy overcome this limitation by building an effective area of emitting labels smaller than the excitation spot. This is achieved by combining two diffraction limited excitation and depletion patterns at the expense of scanning the sample.Other techniques rely on the post treatment of the emission pattern such as the point localization superresolution. In the following, we will concentrate on the details of total internal reflection fluorescence and the point localization superresolution. Those techniques indeed find natural applications in the *in situ* assembly of viruses study and we will propose in this work different methods to extract meaningful information from the data they produce.

1.3.2 Total Internal Reflection Fluorescence

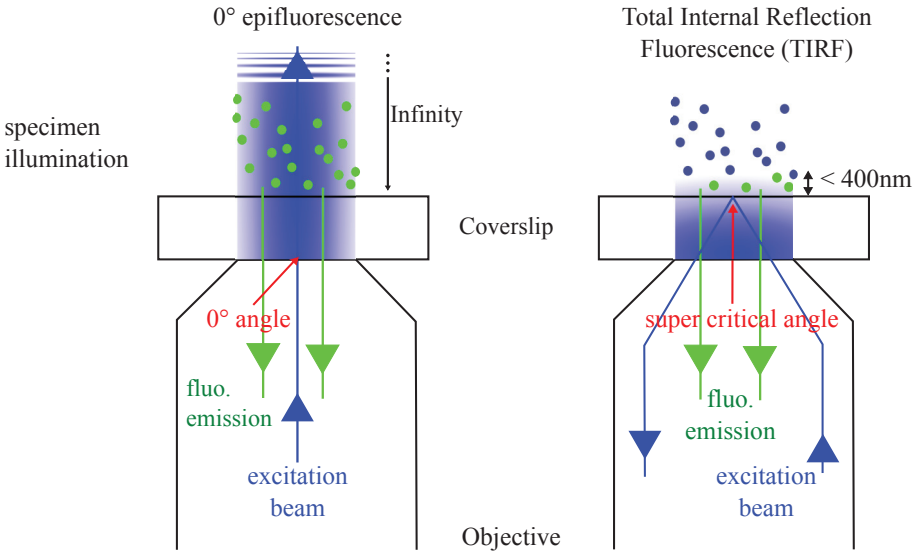


Figure 1.14 – Total Internal Reflection Fluorescence principle: compared to classical display (left) the excitation beam is fully refracted at the dioptre between cover slip and sample so that only an evanescent wave with exponential decrease is seen in the sample. Only the labels within excitation wavelength are efficiently excited (right)

The near field methods take benefit of the refraction principle to reduce the spatial extension of the excitation field to the order of the wavelength in the vertical dimension. In this set-up, the excitation beam is sent in the cover slip where the object of interest lays, with an incidence angle greater than the total refraction angle of the interface cover slip-sample. Hence the excitation beam is entirely reflected and only an evanescent wave penetrates the sample. According to refraction law, the number of photons in the sample is exponentially decreasing with penetration depth at a rate fixed by the excitation wavelength. The intensity that can efficiently excite the label is thus found only in the vicinity of the surface. This configuration is particularly useful to image with a very low background noise phenomenon happening at

the membrane of the cell such as protein attachment and diffusion, budding, etc. Event if the detail of the virion image can not resolved, the total emissions from different clusters of proteins can be distinguished if the clusters are spatially separated by at least several times the emission wavelengths. A single cluster intensity can this way be monitored in time to image the dynamic of aggregation or interaction when different molecules are labeled in a distinctive manner. This way insight of the dynamic of the human immunodeficiency virus budding has been gained showing the arrival and cargo of the viral RNA (Jouvenet et al.[72]). As a drawback, the excitation field is not homogeneous and the effect of stoichiometry on the fluorescence intensity cannot be distinguished from the effects of the cluster position fluctuations in depth. This issue can be addressed using more sophisticated imaging schemes with different illumination incidence angles (Ku et al.[80]). Although those experiments do not establish the structure of the protein clusters, they provide information about the assembly dynamic. While early experiments were mainly aimed at establishing the sequence of assembly steps (gag nucleation, RNA packaging,etc.) –figure 1.15; the latest experiments are dedicated to determine the stoichiometry of the budding sites and monitor their time evolution, allowing *in situ* determination of individual cluster dynamic. This new possibility calls for models of the aggregation process in cell at the individual level of each protein cluster and establishment of the relationship with the population behavior to foster quantitative interpretation of the results.

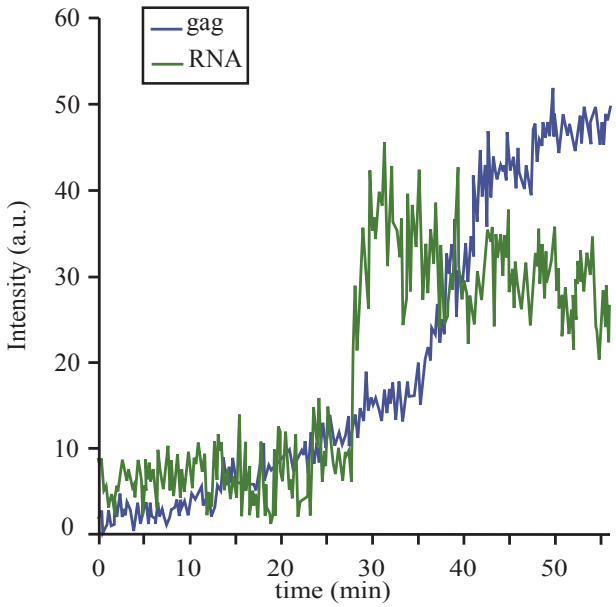


Figure 1.15 – Intensity of the fluorescence signals from labeled RNA (green) and gag protein (blue) from a single budding site at cell membrane using the total internal reflection configuration. (from Jouvenet et al.[72])

1.3.3 Superresolution point localization imaging

The physical diffraction limit of light microscopy can however be bypassed using more complex imaging scheme. In this work we focus on the superresolution family of single molecules localization techniques that take benefit of specially designed fluorescent dies with two or more

"states" between which that can be photo-switched to control their emission behavior figure 1.16.

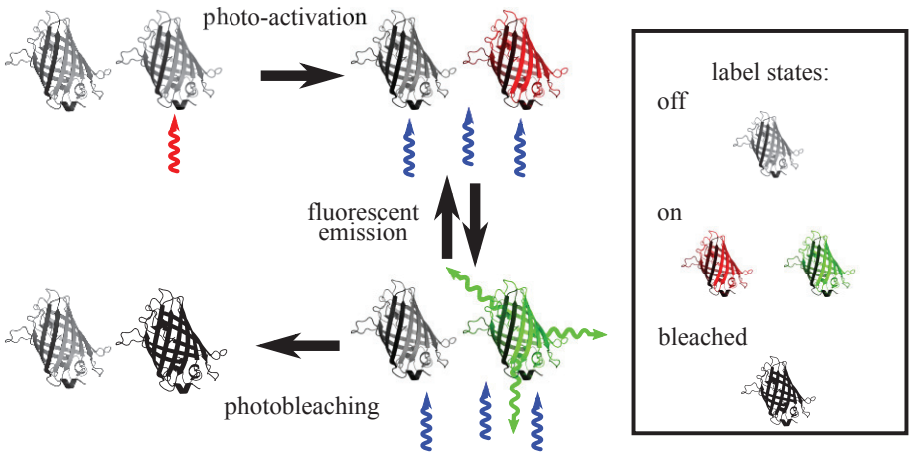


Figure 1.16 – Photo-switching principle: a low intensity photo-activation laser (red) switches a sparse subset of the labels from dark ("off") to bright ("on") state. Switched on label is fluorescent at the excitation wave length (blue) and emits (green) by cycling between ground to excited states (figure 1.13). Only the "on" label is imaged until it bleaches enabling a time separation of the emission signals from the two neighboring labels. Indeed the second label remains "off" and will be switch "on" at latter time

Using the scheme shown on figure 1.16, microscopist can photo-activate a subset of fluorescent dyes within the sample by choosing the appropriate intensity of photo-activation laser. Only a few dies interact with the photons and are stochastically turned "on". If the density of labels in the bright state is low enough, the probability is strong that they are spatially distant from more than the typical diffraction length given by Abbe's law (equation (1.1)) and this sparsity guarantees that the diffraction spots are distinct on the camera sensor (CCD) as shown on figure 1.17. The positioning of each emitter can then be achieved with a finer precision than the diffraction spot dimension since emitter position image (corrected from diffraction) is located at the center of the diffraction-limited spot it generates. The positioning precision depends on how well the diffraction spot of the emitter is defined. This is function of the number of photons received on the camera sensor, the size of the pixels on the sensor on which photons are counted, and eventually, the signal to background noise ratio. The estimation theory gives the following variance $\sigma_i|^2_{x,y}$ of the center position for the diffraction spot i (Thompson et al.[127] –assuming a Gaussian points spread function fitted by the least square method in 2D):

$$\sigma_i|^2_{x,y} \simeq \frac{s^2 + a^2/12}{N_i} + \frac{8\pi s^4 b_i^2}{a^2 N_i^2} \tag{1.2}$$

Where s is the standard deviation of the point spread function, a is the pixel size in the image (taking into account the system magnification), N_i is the total number of photons measured from emitter i, and b_i is the number of background photons collected in the fitting window used for molecule i. By choosing a pixel size comparable to the standard deviation of the point spread function, we see that the localization scales as $\sigma_i|^2_{x,y} \propto \sqrt{\frac{s^2}{N_i}}$ which underlies the importance of a good photon yield and let anticipate a consequent amelioration of the spatial resolution by an order of magnitude for a hundreds of collected photons. In the early 2000s, analogous methods

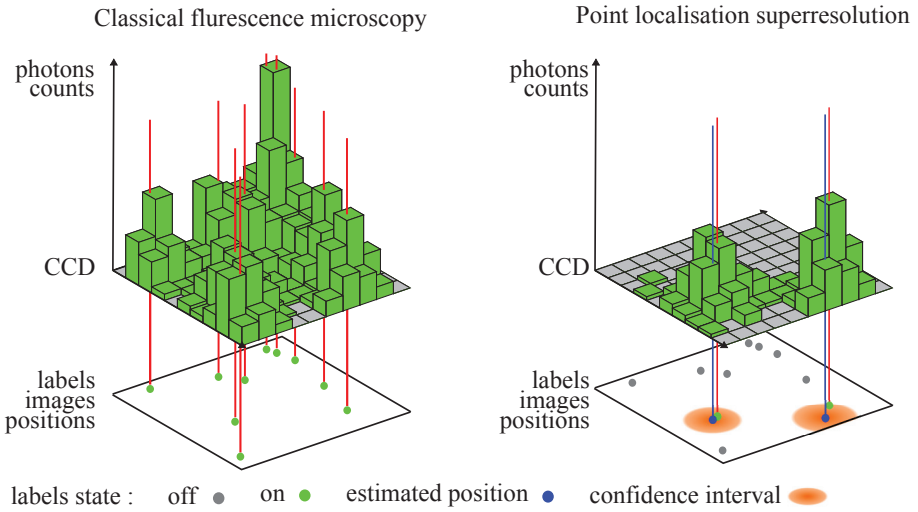


Figure 1.17 – Principle of point localization superresolution microscopy: the classic fluorescence image is the sum of the simultaneous emission of all the labels emissions. Individual diffraction spots cannot be distinguished (left). In point localization microscopy only a sparse subset of the label is switch on at a time and non overlapping contributions can be assigned to individual emitters. Each emitter can in turn be localized at the center of the point spread function by fitting its point spread function with a centroid finding algorithm. The localization precision depends on how well the point spread function is defined which is directly linked to the photon yield of the emitter and the background noise.

based on such single molecule localization have been independently demonstrated. Variously known as (fluorescence) photo-activated localization microscopy ((f)PALM, Betzig et al.[12], Hess et al.[64]) or stochastic optical reconstruction microscopy (STORM, Rust et al.[112]). The main difference between the techniques lies in the type of photoswiching dye used: synthetic dies for STORM and fluorescent proteins for (f)PALM, and the sparse photo-activation design (continuous low intensity laser in fPALM compared to short time laser pulses for the others). As a result the spatial resolution of the final image composed of all of their molecular locations is substantially improved (from 18 nm to 40 nm). Synthetic have generally a larger photon yields of the order of 10^3 before bleaching when fluorescent proteins are in the order of 10^2. The positioning precision is thus better but at the expense of a higher spontaneous emission rate that degrades the contrast ratio and of larger sizes and to anti-body labeling or click chemistry that might affect the tagged structure more than a $\simeq$ 5 nm fused fluorescent protein (Bates et al.[11]).

In 2D superresolution point microscopy, a time or a spacial separation of the emissions combined with the knowledge of the optical point spread function in the image plane are the two necessary elements to calculate the emitters positions in (x, y) plane. In a classic set-up, the point spread function changes, it width increasing, as the imaged point is moved out of focus. The knowledge of the point spread function variations with the source distance from focus plan is a natural measure to find the object axial position. However as this evolution is symmetric with respect to the displacement $\pm\delta z$ it is not sufficient. In order to achieve 3D imaging supplementary information must be provided. Different solutions have been demonstrated to achieve this purpose (see figure 1.18). Huang et al.[68] introduced a cylindrical lens in the optical path. The difference of curvature between x and y directions generate an elliptical point spread

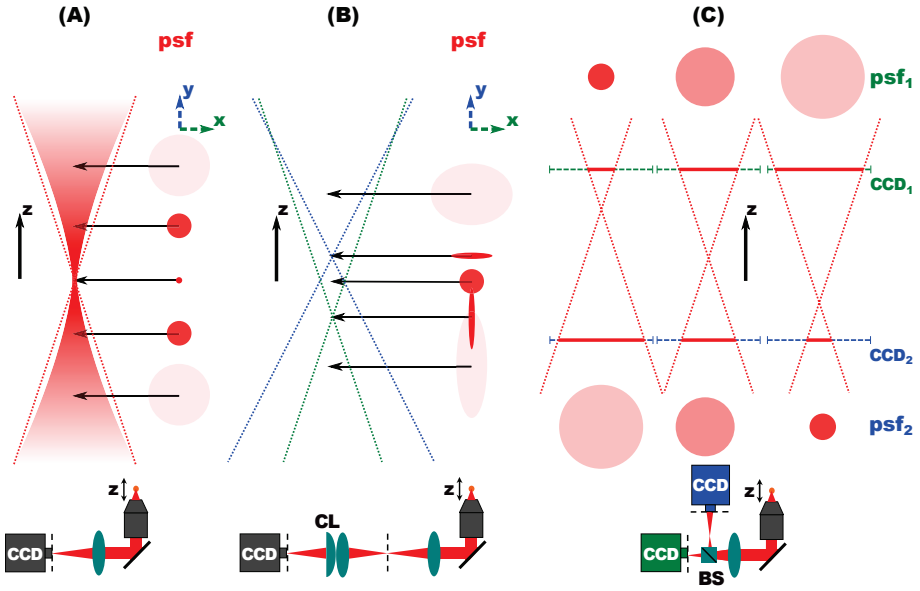


Figure 1.18 – Principle of two classically used 3D superresolution imaging solution: in a classical fluorescence microscope setup (A) the shape evolution of the point spread function (psf) is symmetric for a $\pm z$ displacement of the source from the focus plan (red surfaces) or equivalently a CCD displacement. (B) Introducing astigmatism with a cylindric lens (CL) combined with an spherical lens before the CCD, focus is achieved on the CCD in x (green dashed shape) or y (blue dashed shape) for two different source positions in z. In between, those two positions, the width of the psf in each direction gives the value of the z position of the emitter. (C) In the two planes setups, the psf is simultaneously recorded in two different image planes using a beam splitter (BS), this allows to interpolate the 3D-psf and deduce the source position.

function on the CCD with major extension in x or y depending on the displacement sign. The price is a degraded precision in (x, y) directions ($\sigma_i|_{x,y} \simeq$20-30 nm and $\sigma_i|_z \simeq$50-60 nm. Juette et al.[74] used a beam splitter to simultaneously image the point spread function in two different plans to eventually fit the 3D-point spread function and extract the position ($\sigma_i|_{x,y} \simeq$30 nm and $\sigma_i|_z \simeq$70 nm). Other techniques exists involving more delicate set-ups such as interferometry that has demonstrated the best resolution at the time but is more ($\sigma_i|_{x,y} \simeq$20-30 nm and $\sigma_i|_z \simeq$20-10 nm demonstrated by Shtengel et al.[116]). Further important improvements have been showed such as multi color channels (Bates et al.[10], Bossi et al.[14], Dedecker et al.[35]), 3D and two color channel (Jones et al.[71]), or particle tracking Hess et al.[65], Manley et al.[89], Shroff et al.[115].

Superresolution on virus It is a striking fact that among the different systems chosen to test the new methods, one was the human immunodeficiency virus gag protein aggregation at the membrane reproduced on figure 1.19! The use of superresolution microscopy is for a part inherited of the classic fluorescence microscopy expertise, leading to natural co-localization studies of different molecules of interest labeled with different dyes. The independent migration of the human immunodeficiency virus envelope glycoproteins to gag budding sites and its incorporation in the viral particle as been shown by Muranyi et al.[102] using two color superresolution imaging. The dynamic of the influenza HA envelope protein with membrane specific domains as also been analyzed this way by Hess et al.[65], demonstrating in addition the interest

of superresolution for dynamical studies in identifying the movements of the proteins inside the membranes clusters. Beyond the gain in the typical objects sizes that the technique can probe, it seems clear that the appealing interest of the technique is to unveil structural informations of the object and their evolution. In this spirit, Lelek et al.[83] used the superresolution microscopy to study the delivery of the human immunodeficiency virus genome inside the nucleus of the cell in the early steps of infection. The integrase protein, in charge of integrating the viral genome in its host's, was imaged in the cell cytoplasm. Resolution however is not sufficient to clearly identify the capsid shape to deduce whether the protein had been liberated despite apparent variations of the geometry. Structure was inferred by clustering the measures in two different groups according to shape proximity providing evidence that the capsid remains intact across the cytoplasm and delivers its content close to the nucleus. This illustrates the current state of the art in term of structure identification and the scientific interest of inference methods. Quantitative analysis of the proteins clustered on the membrane using statistical test function as been proposed by Kiskowski et al.[78], Sengupta and Lippincott-Schwartz[114] to extract typical organization scales at the level of the population. The second method has been recently adopted by Malkusch et al.[88], Muranyi et al.[102] in the study of the human immunodeficiency virus gag assembly at the membrane to deduce the average cluster size. Furthermore Malkusch et al.[88] pointed out that label density fluctuation of several assembly sites were well resolved enough to possibly reconstruct the gag radius of curvature of selected clusters. The development of a method to do so would undoubtedly be of large interest in further studies. Gunzenhäuser et al.[59] working on the same system proposed the use of photo-activated light microscopy technique to estimate the number of protein imaged in a gag cluster (which is not a simple task as dyes are known to blink and can be counted more than once). They studied the shape and size dispersion of the budding sites distribution on the membrane incidentally pointing out the need of such a method to measure individual cluster sizes regarding a structural model.

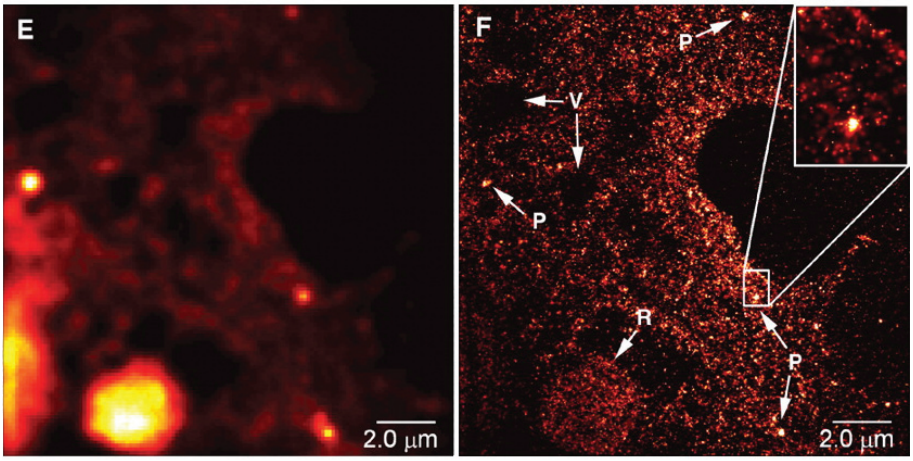


Figure 1.19 – The very first PALM image of a viral protein at the membrane. "Summed-molecule TIRF (E) and PALM (F) images, respectively, of a COS-7 cell expressing the retroviral protein gag tagged with dEos. The PALM image highlights voids (arrows labeled V), a higher density region (arrow R), and probable condensation at several points (arrows labeled P) into VLPs of $\sim$ 100 to 150 nm size (inset)." (reproduced from Betzig et al.[12])

Viruses, but also many cellular machines, such as the multi-protein structures involved in membrane fission or fusion, transport across membranes, cell division, and more, lie below the

resolving power of fluorescence microscopy. Superresolution fluorescence imaging promises to finally directly reveal their organization *in situ*. Indeed, recent microscopy studies of the midbody (Elia et al.[38], Guizetti et al.[58]), centriole (Keller et al.[76], Mennella et al.[95], Sillibourne et al.[118]), and nuclear pore (Szymborska et al.[125]) have advanced the models for how such machines are assembled. At the same time, these studies reveal some of the current limitations in interpreting superresolution images.

A good estimation of the structure on which lye the labels is the goal of imaging. To construct a good estimation of the labels density from the photon recorded on the camera captor, several different strategies ha been proposed. The most common consists in estimating the most likely position of each single emitter as originally proposed. This leads to a image made of the collection of all the estimated positions. The information of the positioning uncertainty attached to each estimated position is sometimes included substituting each point for a Gaussian with standard deviation given by equation (1.2) thus generating continuous intensity images more alike the classical fluorescence ones. Various operations can be performed on the photon measurement to obtain such results: they are called statistical estimators and the operation is generally pointed out as deconvolution. They are implemented in image production algorithm (see Fitzgerald et al.[44], Mukamel et al.[100]) or provide theoretical measure of the spatial information included in the image (Mukamel and Schnitzer[99]).

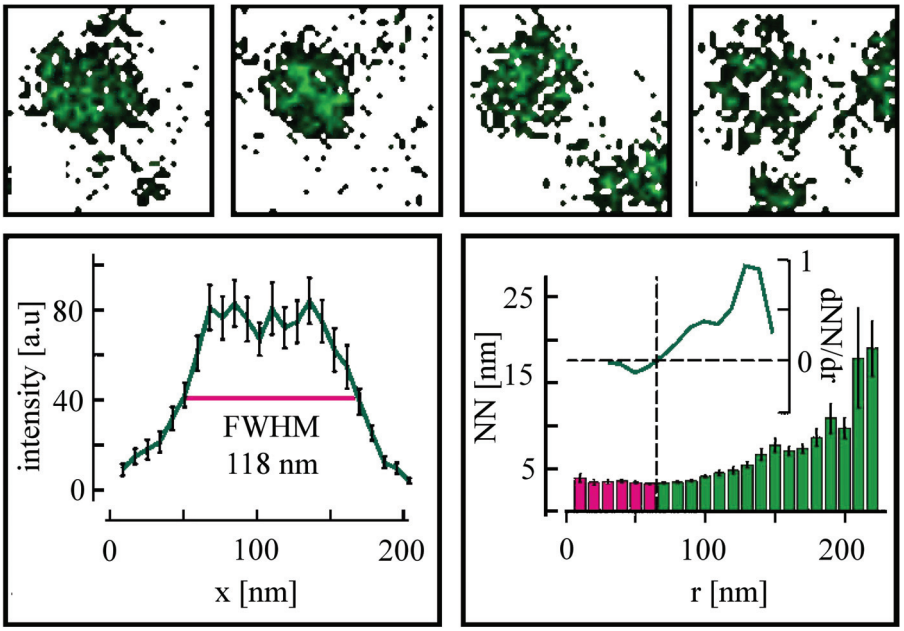


Figure 1.20 – dSTORM images (300 nm × 300 nm) of gag clusters observed at the membrane are potentially viruses budding sites (top row). 25 clusters were centered and averaged radially from the center to produce the intensity distribution (down left) and give a typical cluster width, thus neglecting clusters variance. Pair statistics such as distance to nearest neighbor is used to identify close packing area (pink bars) where nearest neighbor distance is constant indicating the typical mean cluster width of $\simeq$ 70 nm (from Malkusch et al.[88])

Indeed because of the stochastic nature of the method, images are inherently noisy and the amount of information they carry is limited by the labeling density. Labeling density is even lower than the dye density inside the imaged sample because all dyes cannot be read due to misfolding, failed photo-conversion, or premature photo-bleaching for instance. The

effect on the information in the image is generally described as a consequence of Nyquist sampling theorem, which states that any spatial variations smaller than the distance between two neighboring dyes is not accessible. The question of which features of the labeled structure are conserved in the position image and what should be accounted as distortion is crucial for those methods. The theoretical information content of such a point localization image was estimated by Mukamel and Schnitzer[99] using the Fisher theory of unbiased estimation that we will discuss further in the course of our work in section 3.6.3. It gives the minimal spatial frequency that is meaningful given both the sample labeling density and the imaging point spread function width and photon yield that set the localization precision bound.

Furthermore, when the localization precision of single molecules is not much smaller than the size of the object, the structural information content of the image is distorted or obscured by positioning errors. This positioning error is intrinsic to optical imaging and depends mainly on the fluorophore quantum yield. Thus, individual images lack sufficient quality to test or build structural models. Instead, significant particle averaging is often necessary to use the data to prove a model (see the works from Keller et al.[76], Malkusch et al.[88], Szymborska et al.[125], Van Engelenburg et al.[130]).

When molecular structures or particles are identical, statistical averaging over a large set of images is a valid way to address these limitations. However, in the most general case, there can be genuine structural variability as it has been stressed in the case of the human immunodeficiency virus (see Carlson et al.[20], Gunzenhäuser et al.[59]), which is completely discarded by averaging.

We propose to address this issue by introducing an estimation method applied at the single particle level. This assumes that we have a geometrical parametrization of the structure –an *a priori* structural model–, and that we want to determine its parameters from the superresolution point localization images. Tomo-electron microscopy is of course a method of choice to obtain a insight of *in situ* object shapes but does not allow to label specific protein. The complementarity of the two techniques have already been used in correlation microscopy study. Methods designed to use a prior knowledge on the shape of the imaged object at the single particle level and its proper characterization is not available in literature to our knowledge. Yet the "inverse problem" of finding the best model parameters given the measure is not new. We believe that refining the analysis of superresolution image in regard of model comparison would open large perspectives in the field.

It is however not obvious how model information can be transferred to the analysis of superresolution images. For instance, in the case of the human immunodeficiency virus like particles, tomo-electron microscopy has shown that the gag protein coat in each particle forms a spherical and continuous lattice with a large gap of uncovered lipid membrane. We know however that the size of this gap as well as the radius vary from one particle to another. In this work we tested the possibility to estimates radius and completion of single particles imaged by stochastic localization superresolution using the maximum likelihood approach. The method calculates a score for all possible values of parameters. The score corresponds to the probability of obtaining the observed data from the structure parametrized by this set. Thus, we can identify the highest scoring set of parameters, which corresponds to the most probable structure underlying the measured data. It is precisely the approach followed when superresolution image is reconstructed by replacing each photon peak by the putative emitter at the estimated center of the spot: its most likely position.

We will illustrate the interest of this approach by validating it on PALM images of budded,

immature virus-like particles (VLPs) formed from the fluorescently labeled polyprotein HIV-1 gag. Cryo-electron microscopy has shown that VLPs are formed from an incomplete spherical protein shell beneath the lipid viral envelope and are highly polydisperse (Briggs et al.[17], Carlson et al.[20]). gag proteins from individual VLPs are therefore expected to lie on spherical shells of variable radii and closure angles, features that would be obscured by particle averaging. We compute the maximum likelihood 3D geometry, thereby estimating the particle radius and protein coverage for individual VLPs. We apply this strategy to both simulated and real PALM data. Comparison between the parameters used as inputs to simulate the data and the output parameters of the estimation procedure gives an estimate of the precision reached at a statistical level. This way, we are able to estimate the radii and the closure angle that best explains the measured data from a given particle, and also provides a mean to extract the uncertainty on this estimation.

1.4 Outline of the thesis

The manuscript is organized as follows: in the first chapter we detail the state of the art concerning the models of viral capsid self-assembly and we introduce the model for an open system submitted to a constant flux of proteins. We then describe the model behavior in two different scenarios for the proteins aggregation, we give the steady state size distribution and detail the transient regime. The last chapter section establish the link between the description of the evolution in term of concentration and the time evolution of the size of a single aggregates. A direct simulation of the size of the aggregate in time is given and a statistical treatment of the associated fluorescent data is proposed. In a second chapter we describe our treatment of the superresolution images. We first detail the procedure to properly orient and center the images. In a second time, we detail the maximum likelihood principle to estimate the parameters of a model describing the object imaged and study the method efficiency on simulated data. Eventually we test the algorithm on real photoactivated light microscopy images.

Chapter 2

Self assembly dynamic

In this chapter we study the dynamic of the viral proteins aggregation in an open system when an constant input flux of protein is maintained at the entry. We first detail the existing studies of *in-vitro* closed system, then we extend the model to open system and study its behavior. Eventually we make a connection with the growth of a single viral bud which can be tracked experimentally. The budding dynamic of a specific virus is unlikely to strictly follows a universal model. The virus formation is affected by many different factors:

— A virus target a specific host cell. Cells organization and activities vary much. Cells are indeed very different between two living organism (for instance plant cell and a bacteria) as well as between two differentiated cells inside a complex organism (for instance a neuron and a red blood cell), and even between two cell of the same tissue at different phases of the cell life cycle... The environment in which budding occurs is therefore strongly different for one virus to another.

— A virus interact with the host cell, either inside the nucleus by pontificating the host genetic program and its regulation, or in the cytosol altering the cell protein production and activity which modifies the parameters of environment of the budding.

— viral protein have a rich design that can undergo important modifications through conformational changes or scission of a part. Virus can benefits from the structural plasticity of constitutive proteins at various stages of the process which modifies its physical properties (affinity of the neighbors)

Despite those specificities, virus budding strategies share commons features like the repeated addition of elementary subunits the capsomers (a single protein or a proteins assembly), in order to obtain a closed container enclosing the genome.

Enveloped viruses assemble at the membrane of the cell or of one of its compartment. This scenario differs from most of the assembly of icosahedral viruses, for which all assembly intermediates (from simple subunits to complete capsids) share the same volume. In the case of enveloped viruses, there is a subunit flux towards the membrane, and completed capsids leave the membrane. While in the case of non-enveloped virus, the self-assembly can be modeled by a closed system, with constant subunit mass, the self-assembly in the case of enveloped virus is likely to be modeled as an "open" system, with one input (subunits) and one output (completed capsids).

In order to understand the features of this self-assembly process, simplified approaches have already shown good results in providing a general canvas to understand the dynamic in a closed

system.

2.1 Closed system

We summarize in this section the main results of this work, using the same notation for the sake of simplicity.

Pioneer work was made by Zlotnick[139]. He looked at the case of experimental in-vitro assembly of cowpea chlorotic mottle virus (CCMV) demonstrated by Bancroft[8]. In the modeled situation, an initial number of capsid capsomers are introduced in the system in the presence of viral material and the final state is observed to be "either intact [capsids] or completely disassembled into stable poly- or monomeric subunits". In such a closed system, the self-assembly of virus polyhedral capsid is described as a set of chemical reaction. Each reaction consisting on a single equivalent building block (called a capsomer) accretion or loss on the growing structure:

$$\begin{aligned} 1 + 1 &\overset{K_1}{\rightleftharpoons} 2 \\ &\dots \\ N_i + 1 &\overset{K_i}{\rightleftharpoons} N_{i+1} \\ &\dots \end{aligned} \tag{2.1}$$

Two key ideas, are proposed to simplify the model until it can be solved:

1. to restrict the assembly to the final capsid canvas: each building block being added with the proper orientation and position it should eventually have in the complete capsid.
2. to consider only the formation of the most stable species at each step among all the accretion possibilities.

In the case of dodecahedron capsid formed of twelve identical pentagonal subunits with five identical edges he studied, the assembly pathway is a linear graph (a unique minimal free energy configuration is found at each step). The equilibrium constant for each reaction is deduced from the free energy difference between products and reactants:

$$\underset{i,i+1}{\Delta} G^0 = n_{i,i+1} \Delta G^0 \tag{2.2}$$

where $n_{i,i+1}$ count the number of newly formed bonds in the capsomer addition and ΔG^0 is the free energy involved in one bond formation. The resulting equilibrium constant at temperature T is:

$$\begin{aligned} K_i(T) &\equiv \frac{[i+1]_{\text{eq}}}{[i]_{\text{eq}}[1]_{\text{eq}}} \\ &= S_1 \frac{C_{i,i+1}}{C_{i+1,i}} e^{-\frac{\Delta G^0}{ij}/k_B T} \end{aligned} \tag{2.3}$$

Where S_1 and $C_{i,i+1}/C_{i+1,i}$ are entropic contributions that take into account rotational degrees of freedom:

— S_1 account for the number of equivalent ways for the building block to bind (five in this case for each equivalent edge)

— c_{ij}/c_{ji} is the ratio of the number of binding sites on structure i where a binding event leads to structure j and building units in structure j that can be removed to form i, as $j = i + 1$.

Key characteristics of the model are that even for very small binding energies, the exponential dominates the other contribution. As binding is considered favorable, $\Delta G^0 < 0$ and the effective potential a growing capsid shows no barrier to full-completion(see figure 2.1).

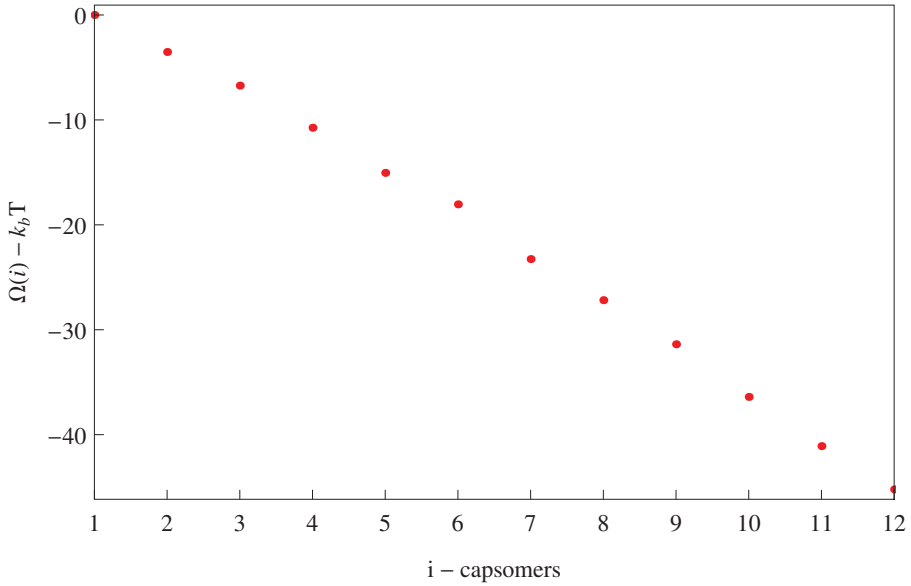


Figure 2.1 – Free energy landscape from Zlotnick model $\Omega(i) = \sum_{j \leq i} e^{-\log K_j}$ for $\Delta G^0 = -k_B T$

The equilibrium concentrations can be recursively deduced from the concentration of free capsomers $[1]_{\text{eq}}$ and the law of mass action:

$$[n]_{\text{eq}} = \left(\prod_{i=2}^{n} K_n \right) [1]_{\text{eq}}^n \tag{2.4}$$

The equilibrium size distribution is therefore a compromise between the enthalpy dominated term $\prod_{i=2}^{n} K_n$ growing at least exponentially for $\Delta G^0 \ll 0$ and the geometric evolution of the second part $[1]_{\text{eq}}^n$. The intermediates are always disadvantaged, and as a result, the equilibrium is dominated by the extreme species, namely capsomers and full capsids. The greater ΔG^0, the stronger such behavior. A specific concentration emerges which is defined by $[1]_{\text{eq}} = [12]_{\text{eq}} = C_{app}$, that is the turning point from which product fraction is not any more dominated by capsomers, but by complete capsid (see figure 2.2).

The capsomers concentration at equilibrium can be deduced from the conservation of material in the closed system:

$$\sum_{n=1}^{N} n.[n] = [1]_{\text{total}} \tag{2.5}$$

The work also implements a study of the kinetics of the aggregation following the pathway suggested by the scheme showed on equation (2.1):

$$1 \underset{k_2^{off}}{\overset{k_1^{on}}{\rightleftharpoons}} 2 \underset{k_3^{off}}{\overset{k_2^{on}}{\rightleftharpoons}} \ldots \underset{k_{n-1}^{off}}{\overset{k_{n-2}^{on}}{\rightleftharpoons}} n-1 \underset{k_n^{off}}{\overset{k_{n-1}^{on}}{\rightleftharpoons}} n \tag{2.6}$$

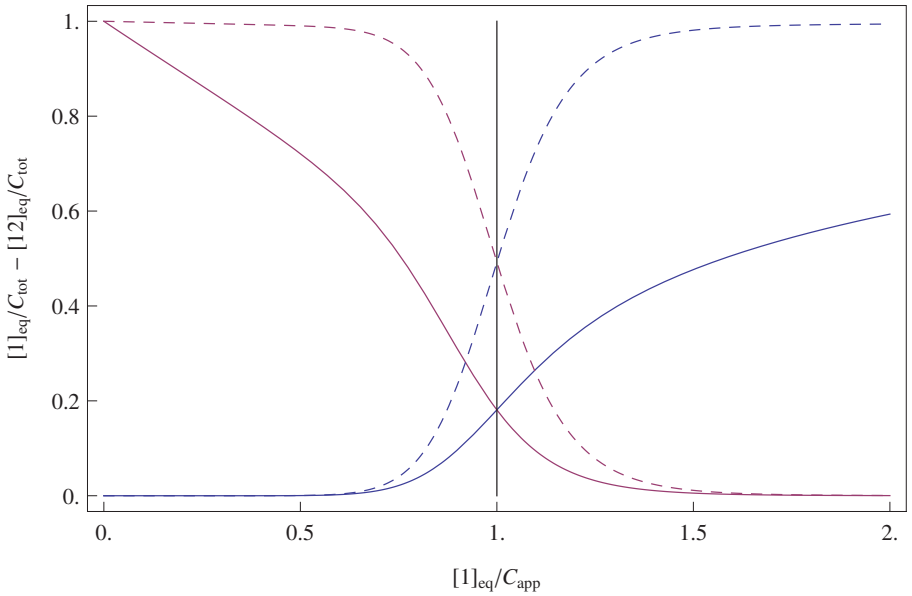


Figure 2.2 – The equilibrium capsids (blue) and capsomers (purple) concentration fraction $[i]_{\text{eq}}/C_{\text{tot}}$ in Zlotnick model for two different $\Delta G^0 \in \{-1, -3\}k_BT$ (resp. solid and dashed lines) as a function of the dimensionless parameters $[1]_{\text{eq}}/C_{\text{app}}$ and with $C_{\text{tot}} = \sum_{j=1}^{12}[j]_{\text{eq}}$ the total concentration in solution. The inversion concentration is marked by the vertical line. For $\Delta G^0 = -3k_BT$, only capsomer and capsid can be present in solution at equilibrium and everything looks like there is a single equilibrium reaction between monomer and closed virus $[1]_{\text{eq}} + \left(\prod_{i=2}^{12} K_n\right)[1]_{\text{eq}}^{12} = C_{\text{tot}}$ which gives ${}^{[1]_{\text{eq}}}/_{C_{\text{tot}}} = {}^{[1]_{\text{eq}}}/_{([1]_{\text{eq}} + (\prod_{i=2}^{12} K_n)\,[1]_{\text{eq}}^{12})}$ and ${}^{[12]_{\text{eq}}}/_{C_{\text{tot}}} = (\prod_{i=2}^{12} K_n)\,[1]_{\text{eq}}^{12}/_{([1]_{\text{eq}} + (\prod_{i=2}^{12} K_n)\,[1]_{\text{eq}}^{12})}$. On the contrary, low binding energy ($\Delta G^0 = -k_BT$) leads to mixture of intermediates in the neighborhood of C_{app}, so that the concentrations of monomers and viruses do not sum to the total concentration in the system.

With corresponding rates-equation:

$$\begin{aligned}
\frac{\partial [1]}{\partial t} &= -k^{on} \sum_{j \neq 1} [j][1] + \sum_{j>2} k_n^{off}[j] + 2k_2^{off}[2] \\
&\dots \\
\frac{\partial [n]}{\partial t} &= k_{n+1}^{off}[n+1] - k_n^{off}[n] + k^{on}\left([n-1][1] - [n][1]\right) \\
&\dots \\
\frac{\partial [N]}{\partial t} &= -k_N^{off}[N] + k^{on}[N-1][1]
\end{aligned} \tag{2.7}$$

with a constant accretion rate $k_{on} = 10^8 \text{mol}^{-1}.\text{s}^{-1}$ modeling a diffusion limited association of two capsomers and the off rate k_{n+1}^{off} is set accordingly to the n$^{\text{th}}$ reaction constant K_n. Indeed, by definition of the thermodynamic equilibrium of a reaction, on and off flux balance each other and therefore:

$$\begin{aligned}
k_{n+1}^{off} \left(\prod_{n=\text{products}} [p] \right) &= k_{on} \left(\prod_{r=\text{reactans}} [r] \right) \\
k_{n+1}^{off} &= \frac{k_{on}}{K_n}
\end{aligned} \tag{2.8}$$

which is the "detailed balance" property. Interestingly, evolution of the different concentration is non monotonic with the initial concentration of free monomers introduced in the system at $t = 0$. At low initial capsomers concentration, all intermediates concentrations are kept low as assembly proceeds and a lag-time is observed between the introduction and the formation of the first capsids. The distribution asymptotically converges to the equilibrium distribution. As the initial concentration is increased, on one hand the lag-time reduces but, on the other hand, reaching the equilibrium takes more time. This kinetics trap is due to the proportionality of the intermediate creation rate with monomers concentrations. Lots of intermediate species are initially made at high concentration and consumes most of the monomers to assemble until very few are left. For the assembly to carry on, free units must first be released in the system by disassembling intermediates, process which faces energy barrier as shown by the potential in figure 2.1.

The characteristics of the concentration distribution obtained in this model are consistent with key observations from in-vitro experiments. Yet if the most stable intermediates are the obvious choice to capture the long time stationary features of the system, it is less clear that they must have the prominent role in the kinetic mechanism.

Concerns about the choice of a given pathway have been addressed either by a mean field statistical treatment by Endres et al.[41], as well as simulations by Zhang and Schwartz[137].

Endres et al.[41] enumerated all the possible configurations of the previously described model of dodecahedral capsid and consecutively the pathways between each. The number of configurations growing exponentially with the total number of unit in the final capsid. The most stable intermediate can indeed be different from the one which is the more connected and kinetically favored at comparable reaction rates. The authors of the article concludes that early stage of the assembly are dominated by most "probable" species whereas the most stable arise as times goes on. In spite of this variation an accurate description of the evolution can be

produced considering only a relatively small subset of all the possible configurations at low concentrations.

Zhang and Schwartz[137] studied, for their part, the effect of limiting the assembly to capsomers addition on the overall kinetics. They simulated assembly of the pentameric capsomers as proposed by Zlotnick[139], implementing a second scenario granting to species formed of more than a single capsomer (called aggregates) the possibility to aggregates if producing a valid intermediate. The assembly was followed using local rules and N-fold algorithm –or Gillespie algorithm (Gillespie[51]). In a local rule assembly scheme unlike in classic molecular dynamic simulations, the spatial features of the interactions between the building blocks are ignored and each accretion event is described as a chemical reaction where two identified reactant build a prescribed product –therefore shape is again prescribed by the capsid canvas. The Gillespie algorithm was originally designed to sample stochastic time trajectories of a system ruled by chemical reaction and disregarding spatial fluctuations. At each step, it iteratively draws the next reaction to proceed as its waiting time from an exponential distributions with characteristic time equal to the inverse of the reaction rate to actualize the system state. The study compared various scenarios of assembly playing on association and dissociation rates assumed equal for all the reactions. Despite those constant rates, a nucleation step is implicit in the implemented assembly as "loops" formed by three capsomers are considered infinitely stable and can't be broken any more. However since the nucleating rate is of order of the elongating rate, it does not prevent the system from falling in kinetic traps described before. As rule of thumb, the assembly proceeds faster to the equilibrium when there is more available paths to reach the final product. Obviously, kinetics traps are absent when intermediates can aggregate. Thus allowing intermediate to bind together leads to better capsids formation yields in reachable times.

The sensibility of the model to kinetic trap at high concentration seemed somehow an important drawback in the viral replication strategy. So it was hypothesized in the following works from Endres and Zlotnick[40], Zlotnick et al.[142], that a first nucleation step would enhance robustness towards kinetics trap. Indeed a first energy barrier preventing the initial strong capsomer concentration to be consumed in forming early intermediate would avoid the stalling of the reaction. In a closed system with nucleation steps the consumption of monomers is controlled by the order of the first nucleation reaction (the number of capsomer needed to form the nucleus), and this rate limiting steps control the capsid formation rates since elongation follows easily. The measure of the capsid formation rate was further demonstrated to be a convenient tool to identify the stoichiometry of the nucleus formed by pentamers of dimers in cowpea chlorotic mottle virus (Zlotnick et al.[140]). The kinetic approach models also the stability of capsids observed by Singh and Zlotnick[119], that is greater than an equilibrium model predict. Indeed dilution of the capsid solution does not lead to the size distribution expected, suggesting a kinetic control of the reactions. The study of the dynamic in the assembly line model starting from a solution of pure capsomers was completed by Morozov et al.[98], showing that the size distribution looks like an assembly front that will propagate in time until either it reaches the complete size and mimic an equilibrium distribution, or stops before full capsid are formed and evolve extremely slowly towards equilibrium by diffusion.

A thermodynamic derivation of the assembly on a prescribed spherical canvas, akin to classic nucleation theory, has been proposed by Zandi et al.[136], in the case where intermediate species concentrations can be assumed negligible compared to fully formed capsid and that capsids are formed of a large number of subunits. The assumptions support the presence of a nucleation

barrier in the system and allow to view capsid growth as the continuous filling of a sphere.

The idea that underlies classical 2D-nucleation theory is to treat all the capsomers aggregates as a mixture of perfect gaz in an system of fixed size S at temperature T in contact with a reservoir of single capsomers with chemical potential μ_1. We will now briefly gives the statistical physics derivation of the thermodynamic potential and equilibrium concentrations.

We first consider the full system including free capsomers reservoir. The single capsomers are considered as aggregates of size $n = 1$. The perfect gaz hypothesis implies that low range internal interactions are present only inside aggregates but they do not interact with long range interactions. Within this framework, the canonical partition function accounting for all the internal energy level of aggregate of size n fixed in space is denoted z_n. The partition function for states with a given number of capsomers in the system is the sum over the partition function of each possible size distribution $\{N_n\}$ that satisfies the conservation of the total number of capsomers n_t writes (N_n being the number of aggregate of size n in the system):

$$Z = \sum_{\{N_n\}} Z(\{N_n\}) = \sum_{\{N_n\}} \prod_{n=1}^{n_{\max}} \frac{(\frac{Sz_n}{\lambda_l^2})^{N_n}}{N_n!} \quad | \quad \sum_n n.N_n = n_t \tag{2.9}$$

Where a classical integration over the translational degrees of freedom of the aggregates gives a factor $\frac{S}{\lambda_l^2}$ times z_l, with λ_l the de Broglie thermal wave length of the aggregate, and indistinguishably of the different aggregates of equal size is given by classic combinatoric formula ($q^{N_l}/N_l!$) accordingly to the supplementary assumption that the number of thermally reachable energy states is huge compared to the number of aggregates in the system –so that probability to find many particles in the same state is negligible. As each gaz of aggregate with a given size is independent from the other, count of the number of states of the mixture is given by the product over sizes. A classic treatment assumes that in thermodynamic limit the sum is strongly dominated by the term with equilibrium size distribution:

$$Z \simeq \prod_{n=1}^{n_{\max}} \frac{(\frac{Sz_n}{\lambda_l^2})^{N_n^{\text{eq}}}}{N_n^{\text{eq}}!} \tag{2.10}$$

Which eventually gives the free energy

$$\begin{aligned} \mathcal{F}(T, S, \{N_n\}) &= -k_B T \ln Z \\ &= \sum_{n=1}^{n_{\max}} k_B T N_n (\ln(s_0.N_n/S) - 1) + N_n f(n) \qquad (2.11) \\ \text{where :} \quad f(n) &= -k_B T \ln z_n + kT \ln(s_0/\lambda_l^2) \qquad (2.12) \end{aligned}$$

We denoted by $f(n)$ the free energy contribution of all internal degrees within an aggregate of size n and introduced s_0 the typical surface area of a capsomer in an aggregate (the momentum contribution with Broglie wave length contribution is included in f).

Now, if capsomers are not considered part of the system, but instead that a reservoir of single capsomers with chemical potential μ_1 is in contact with the system as proposed in the

article, the thermodynamic potential to minimize is the following:

$$\mathcal{F}(T,S,\{N_n\}) - \mu_1 \sum_n n.N_n = -k_B T \ln Z - \mu_1 \sum_n n.N_n$$
$$= \sum_{n=1}^{n_{\max}} k_B T N_n (\ln(s_0.N_n/S) - 1) + N_n \Delta f(n) \tag{2.13}$$
$$\text{with:} \quad \Delta f(n) = f(n) - \mu_1.n \tag{2.14}$$

Minimization of the free energy gives the following size distribution:

$$C_n^{\text{eq}} = \frac{N_n^{\text{eq}} s_0}{S} = e^{-\beta \Delta f(n)} \tag{2.15}$$

where $\beta = k_B T$ is the classical unit of thermal energy with the absolute temperature T and the Boltzmann constant k_B. Defining $C_1^{\text{eq}} = e^{-\beta \Delta f(1)}$, the equation transforms into:

$$\mu_1 = f(1) + k_B T \ln C_1^{eq} \tag{2.16}$$
$$C_n^{\text{eq}} = (C_1^{\text{eq}})^n e^{-\beta(f(n) - n.f(1))} \tag{2.17}$$

Which is consistent with the equilibrium mode proposed by Zlotnick[139].

To describe the effect of assembly through nucleation barrier, the phenomenological liquid drop model is used. The proposed free energy difference between monomeric unit and the growing capsid is written as:

$$\Delta f(n) = n \Delta \mu + \Gamma L(n) \tag{2.18}$$

The first term is the mean energy gained through the subunit association process in the bulk, while the second term is the cost associated to missing contacts at the free rim of the partial capsid. It is a line tension term proportional to the length of the rim.

In the continuous limit, the following relation links the surface to the perimeter of a sphere cap defined in geometrical coordinates by $\{r = R, \phi \in [0, 2\pi], \theta \in [0, \theta_m]\}$:

$$L = 2\pi R \sin\theta_m \qquad S = 2\pi R^2 (1 - \cos\theta_m) \qquad \sin\theta = \sqrt{1 - \cos^2\theta_m}$$
$$\Rightarrow L = 4\pi R \sqrt{\frac{S}{4\pi R^2}(1 - \frac{S}{4\pi R^2})} \tag{2.19}$$

Back to the discrete case, under the classic assumption that the density is constant in the aggregate and defining N the total number of capsomers to cover the full sphere, r_p the radius of the average disk occupied by a capsomers in the full aggregate and n_L the number of capsomers at the periphery of the bud:

$$S(n) = n \frac{4\pi R^2}{N}, \quad \pi r_p^2 = \frac{4\pi R^2}{N}, \quad L = 2 r_p n_L \tag{2.20}$$

Eventually:

$$L(n) = \frac{4R}{\sqrt{N}}.\pi \sqrt{n\left(1 - \frac{n}{N}\right)} \tag{2.21}$$

Where the first factor gives the portion of the cap rim occupied by a capsomer on average $(2r_p)$ and second factor is equal to n_L, the approximate number of capsomers at the periphery with unsatisfied coordination. Equation (2.22) can be rewritten as:

$$\Delta f(n) = \Delta \mu.(n-1) + \Gamma \frac{4R}{\sqrt{N}}.\pi \sqrt{(n-1)\left(1 - \frac{n}{N}\right)} \tag{2.22}$$

Where the shift in the indices ensures consistency with $\Delta f(1) = 0$ without changing the general behavior in n (indeed our continuous description makes little sense at low n anyway), and $\gamma = {}^{4\Gamma R}/\sqrt{N}$ is the average line tension per periphery capsomer : the gain of free energy compared to the bulk due to the absence of several of its neighboring bounds.

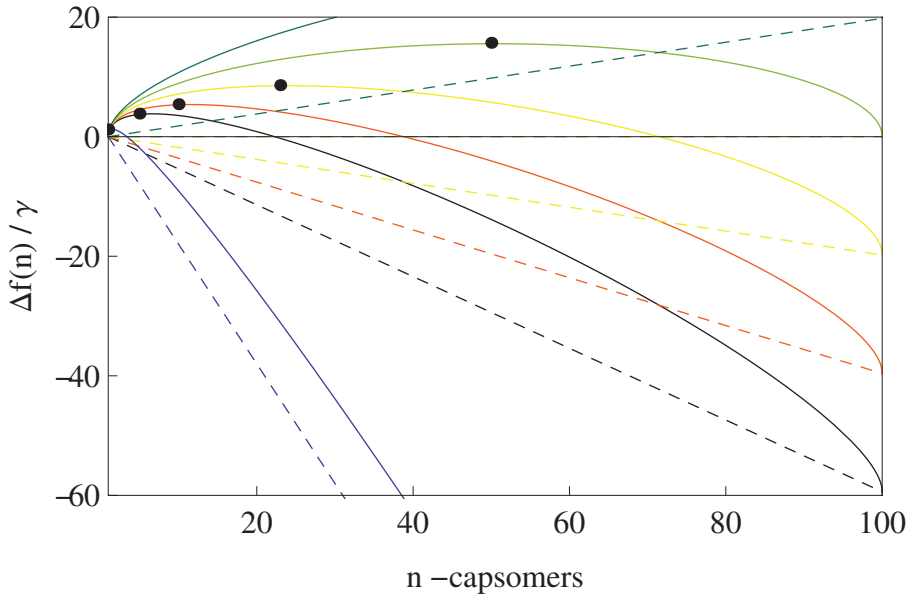


Figure 2.3 – ${}^{\Delta f(n)}/\gamma = (n-1)\rho + \pi\sqrt{(n-1)(1-\frac{n}{N})}$, variation of free energy of aggregates formation in units of the maximum line tension per periphery capsomer $\gamma = {}^{R\Gamma}/\sqrt{N}$ of a truncated sphere (continuous limit description of the capsid assembly) for various ratio $\rho = {}^{\Delta\mu}/\gamma$ of the bulk to periphery energetic contribution. With $N = 100$ and $\rho \in \{0.2, 0, -0.2, -0.4, -0.6, -2\}$ (resp. cyan, green, yellow, orange, black, and blue curves). Black dots show the position and height of the energy barrier, dashed lines give the linear contribution without line tension.

This is the continuous limit for the free energy of a regular polyhedron build from many subunits. The characteristics of energy barrier shown on figure 2.3, are easy to calculate solving $\frac{\partial \Delta f(n)}{\partial n}|_{n_b} = 0$ for a chemical potential difference favoring assembly $\Delta\mu < 0$:

$$n_b = 1 + \frac{N}{2}\left(1 - \frac{1}{\sqrt{1 + \frac{\pi^2\gamma^2}{N\Delta\mu^2}}}\right) \tag{2.23}$$

$$\Delta f(n_b) = \frac{N\Delta\mu}{2}\left(1 - \sqrt{1 + \frac{\pi^2\gamma^2}{N\Delta\mu^2}}\right) \tag{2.24}$$

As the difference between the mean energy per capsomer in the full capsid and the free capsomer in solution $\Delta\mu$ increases at fixed line tension Γ, the position and the height of the energy barrier are lowered.

2.2 Open System

2.2.1 Assembly model

In collaboration with Lionel Foret, we worked on a simplified model of self assembly in an open system with both entrance and exit flux. The key features of the assembly dynamic that we want to capture are the repetitive aggregation of identical units and constraints from a prescribed geometry. These two criteria are found in Human Immunodeficiency Virus assembled virions. HIV budding proceed by accumulation of gag proteins at the inner part of the cell lipid membrane. Once gag-proteins have attached the membrane, the monomers diffuses until they meet and bind together. So formed patches can be considered immobile compared to the gag-proteins for their diffusion coefficient are different by at least on order of magnitude (Manley et al.[89]). Area of membrane where a patch of proteins forms is reshaped during the process into a spherical structure called bud until a scission event occurs and the bud is release outside of the cell.

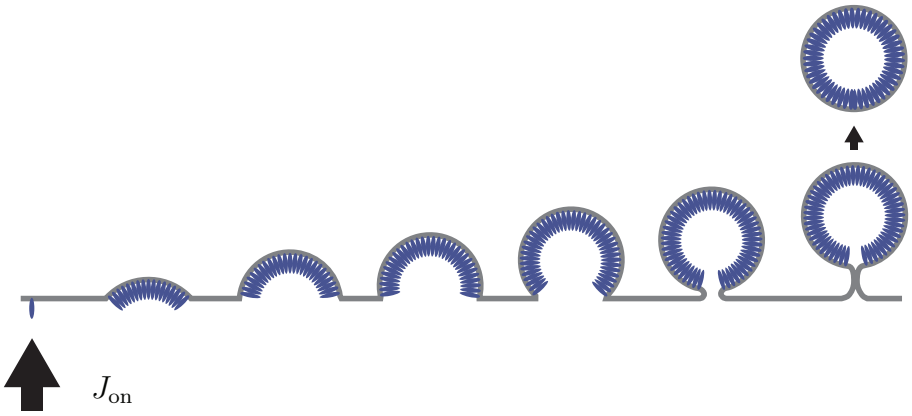


Figure 2.4 – HIV assembly and budding as it appears from cryo-electron microscopy

To model this process we consider identical subunits that can attach to a surface and diffuse on. Such subunits aggregate with one another as they meet. Proteins are produced through the translation of the viral DNA, which produces RNA. RNA strands are replicated and the proteins production from the RNA sequences starts. Further maturation steps are often required for the protein to reach its active form. We will not go into the details of the production of the gag-proteins and instead consider that the whole production results in a mean flux of proteins binding to the membrane $J_{\text{on}}(t)$. In order to be able to infect new cells, the formed virus must have packed among other things the viral genetic code. Different strategies of encapsidation, for instance initiation of the capsid nucleation on RNA strands, can modulate the general aggregation scheme that we focus on here.

A protein bound to the membrane is our building block and we note $c_1(t)$ the concentration of monomers bound on the membrane per unit area. As a gag protein reaches a formed patch to aggregate, it feels an interaction that depends on the configuration of the latter. This interaction is a complex mix of different force sources as the shape of the membrane in the proximity of the aggregate, and protein-protein interaction relatives to the positions of the proteins in the aggregate. Assuming that the bud (protein patch and membrane) equilibrates with one another between each binding event and disregarding the fluctuations, the buds consisting of a given number of proteins are all in the same state of minimum energy and equivalent from the point of view of a incoming protein. Under this simplification the interaction only depends on the

size of the aggregate. Considering the low diffusion coefficient of the buds, fusion of forming buds should be very rare events that we will neglect, as well as scission of buds in two child buds is an unlikely event involving to break many proteins bounds and requires a very strong fluctuation to provide the energy. We therefore adopt the assembly line mechanism proposed by Zlotnick[139] to build our description:

$$1+1 \underset{k_2^-}{\overset{k_1^+}{\rightleftarrows}} 2\ ;\ 2+1 \underset{k_3^-}{\overset{k_2^+}{\rightleftarrows}} 3\ \ldots\ \mathrm{n_{mer}}+1 \underset{k_{n+1}^-}{\overset{k_n^+}{\rightleftarrows}} \mathrm{n{+}1_{mer}}\ \ldots\ \mathrm{N\text{-}1_{mer}}+1 \underset{k_N^-}{\overset{k_{N-1}^+}{\rightleftarrows}} \mathrm{N_{mer}} \tag{2.25}$$

As a bud equilibrates on a small time scale compared to the aggregation time scale, its evolution follows a canvas prescribed by its successive equilibrium states. Each equilibrium configurations giving the position of the protein in the patch and the shape of the patch that minimizes the energy of the lipid membrane and the underlying protein patch. Prompted by the shape of gag buds observed in cryo-EM (see Carlson et al.[20,21]) we approximate the shape of the growing protein patch to a portion of sphere growing with a fixed radius of curvature R as more proteins are recruited. The aggregation of proteins is reported to form a hexameric lattice with few defects. We model the proteins under the membrane as packed with a constant density per unit area of membrane d_0. A fully closed bud of gag protein is thus a sphere of radius R containing $N_{\max}$ gag-proteins with $R \approx 70 \pm 10$nm and $N_{\max} \approx 3.10^3 \pm 1.10^3$ gag-proteins (Carlson et al.[21]).

Those elements are close to the model presented by Zandi et al.[136] to study the assembly of a spherical capsid. We will therefore adopt a similar description of the bud energetic. The deformation of the membrane is assumed to add a energetic contribution to the bud internal energetic disregarding any type of membrane mediated interactions between the buds. The energy gain for adding a new protein to the patch is function of the energy of the formed bounds and the energy involved in reshaping the bud to its new shape. Unlike Zandi et al.[136], we include the freely diffusing monomers attached to the membrane in our system as bud of size $n = 1$ (canonical ensemble). We model the different contributions to the free energy of the bud as follows

— free energy involved by bounds between proteins of the patch follows the droplet model as proposed by Zandi et al.[136]. Note that however the linear part differs in physical meaning in the canonical formalism: here it is purely the internal free energy of the bud and not its formation free energy (for which the chemical potential of each of the isolated constituents is subtracted). Two terms are involved:

  — ϵ_b: the difference between the mean free energy for internal degrees of freedom of a protein inside the patch and the free energy of an unbound protein diffusing on the membrane. In other words, ϵ_b is the mean bounding energy and $\epsilon_b < 0$.

  — γ: the correction term for the average number of unformed bounds for each proteins at the periphery of the aggregates with perimeter $L(n)$

— a supplement of energy is involved to bend the membrane around the patch. The membrane is a bilayer lipid membrane on which we can apply a continuous elastic model as shown for instance in Foret and Sens[47] work. (A detailed calculation can be found in Foret[46]). In the simplified model where the protein coat imposes it preferred shape and membrane deformation is restricted to the bud, two terms are generally taken into account:

— σ: the tension of the membrane, it is the work against the membrane tension to extract a unit area of membrane from the surrounding membrane in order to augment the amount of membrane in the forming bud. And ΔS is the excess of membrane needed to build the bud. Under our hypothesis of spherical cap, $\Delta S = S(n) - \pi(L(n)/2\pi)^2 = s_0 \frac{n^2}{N}$

— κ_m: is the bending modulus of the lipid bilayer, and the term explicit how much energy is involved in shaping the membrane to the bud curvature.

$$\Delta e_{\text{meca}}(n) = \sigma \Delta S + \frac{8\pi\kappa_m}{R^2} S = \sigma \frac{s_0}{N} n^2 + \frac{8\pi\kappa_m s_0}{R^2} n \tag{2.26}$$

The bud internal free energy writes, as reference free energy of unbound protein is set to 0:

$$f(n) = (n-1).\epsilon_b + \gamma.\pi \sqrt{(n-1)\left(1 - \frac{n}{N}\right)} + \Delta e_{\text{meca}}(n) \tag{2.27}$$

In order to derive the rate equation for the evolution of the bud size distribution, we need the rate of attachment/detachment of proteins to a bud of given size (shape). In classical kinetics rate theory, rates of attachment events depends on the local concentration of free proteins around the bud that are likely to attach and the local interaction with the bud when the protein is close enough to react. Within the framework of Eyring absolute rate theory (see Hill[66]), the ratio of forward and backward reaction are given by the equilibrium constant of the reaction:

$$\begin{aligned} \frac{k_n^+}{k_{n+1}^-} &= K(T) \\ &= e^{-\beta.\Delta_r F^0} \\ &= e^{-\beta(f(n+1)-f(n)-f(1))} \end{aligned} \tag{2.28}$$

where $\Delta_r F^0$ is the free energy difference between a aggregate formed of $n-1$ proteins plus an isolated monomer and a aggregate formed of n proteins bound together. The free energy difference derived from the binding energy is expressed in equation (2.27). Oriented flux in the system can be labeled with the following convention:

$$\overset{J_{\text{on}}}{\rightarrow} c_1 \overset{J_2}{\rightleftarrows} c_2 \overset{J_3}{\rightarrow} c_3 \ \dots \ c_n \overset{J_{n+1}}{\rightarrow} c_{n+1} \ \dots \ c_{N-1} \overset{J_v}{\rightarrow} \text{virion} \tag{2.29}$$

The kinetics equations write then:

$$\partial_t c_n = J_n - J_{n+1} \quad \text{with} \quad J_n = k_{n-1}^+ c_1 c_{n-1} - k_n^- c_n \quad \text{for} \quad 2 < n < N \tag{2.30}$$

$$\partial_t c_1 = J_{\text{on}}(t) - J(2) - \sum_{n=2}^{N} J(n) \tag{2.31}$$

$$J_v = k_{N-1}^+ c_{N-1} c_1 \equiv J_N \quad \text{with} \quad k_N^- = 0 \tag{2.32}$$

An interesting limit for the interpretation of the dynamic is the theoretical situation in which, the monomer concentration is maintained constant. Under this regime, the nature of the equations ruling the aggregation becomes obvious.

2.2.2 Dynamics under a constant monomeric concentration

As c_1 is assumed to be constant and independent of time, the set of equation reduce to the following: The kinetics equations write:

$$\partial_t c_1 = 0 \tag{2.33}$$

$$\partial_t c_n = J_n - J_{n+1} \quad \text{with} \quad J_n = K_{n-1}^+ c_{n-1} - K_n^- c_n \quad \text{for} \quad 2 < n < N \tag{2.34}$$

$$J_v = K_{N-1}^+ c_{N-1} \quad \text{with} \quad k_N^- = 0 \tag{2.35}$$

Where we have defined the apparent reaction rates: $K_n^+ \equiv k_n^+ c_1$ and $K_n^- \equiv k_n^-$. In this scheme, a aggregate of size n during $\mathrm{d}t$ grows to size $n+1$ with probability $K_n^+ \,\mathrm{d}t$ or shrinks to size $n-1$ with probability $K_n^+ \,\mathrm{d}t$. If the aggregates reaches the maximum size N, it leaves the system whereas the constant concentration of monomer dispense a constant flux of nucleated dimers at the entrance of the assembly line $K_1^+ c_1$. Overall, at the individual level, the growth process is in this limit a Markovian random walk since the probability for each step is fully determined by the size of the aggregate ($K_n^+ \,\mathrm{d}t$) independent of the system history.

At the level of a population, the equation (2.34), simply conveys conservation of the "material" (here the number of aggregates and not the number of monomers) in the system: variation of the concentration is given by the balance between onward and outward flux. At this level the one step random walk has to well known physical contributions: diffusion locally characterized by the diffusion coefficient $D(x)$ and drift associated to local velocity $V(x)$, which can be identified by establishing a Fokker-Plank equation as the continuous limit. The Fokker-Plank equation:

$$\partial_t c = -\partial_x V.c + \frac{1}{2}\partial_{x^2} D.c \tag{2.36}$$

can be identified from the discrete equation:

$$\begin{aligned} \partial_t c_n &= K_{n-1}^+ c_{n-1} - K_n^- c_n - (K_n^+ c_n - K_{n+1}^- c_{n+1}) \\ &\equiv \frac{1}{2}\left(\frac{D_{n+1}c_{n+1} + D_{n-1}c_{n-1} - 2D_n c_n}{\Delta x^2}\right) - \frac{1}{2}\left(\frac{V_{n+1}^+ c_{n+1} - V_n^+ c_n}{\Delta x}\right) \\ &\quad - \frac{1}{2}\left(\frac{V_n^- c_n - V_{n-1}^- c_{n-1}}{\Delta x}\right) \end{aligned} \tag{2.37}$$

Term by term identification and ± 1 shift of the indices leads to the system:

$$\begin{cases} \frac{1}{2}(V_n^+ \Delta x^{-1} - D_n \Delta x^{-2}) &= K_n^- \\ D_n \Delta x^{-2} + \frac{1}{2}(V_n^- \Delta x^{-1} - V_n^+ \Delta x^{-1}) &= -(K_n^+ + K_n^-) \\ \frac{1}{2}(V_n^- \Delta x^{-1} - D_n \Delta x^{-2}) &= K_n^+ \end{cases} \tag{2.38}$$

First and last equations of the system of equations (2.38) lead to the intermediate one, this is not a system of independent equations. Under the condition of a locally well defined velocity $V_n^- = V_n^+ \equiv V_n$, the unique solution is:

$$\begin{cases} V_n &= \Delta x(K_n^+ - K_n^-) \\ D_n &= \Delta x^2(K_n^+ + K_n^-) \end{cases} \tag{2.39}$$

With the continuous limit sets as $N \to \infty$ by $x = \frac{n}{N}$ and $\Delta x = \frac{1}{N}$. In our framework, the velocity (resp. diffusion) corresponds to the average displacement (resp. square displacement) per unit of time (see Van Kampen[131] for detailed discussion).

So the concentration profile of aggregate will evolve with a drift proportional to the difference between onward and backward rates, meanwhile it will be spread by diffusion to the sum of onward and backward rates. Onward rate is proportional to monomer concentration $c_1(t)$ whereas backward rates are constant. Therefore, the model results in a classic drifted diffusion controlled by the evolution of the single gag proteins in time. Complexity arises from the coupling between the evolutions of the monomers population and growth of the aggregates as nucleation and growth consumes the monomers and thus alters its own evolution.

We are now ready to relax the constraint of fixed monomeric concentration to look at the evolution of the system submitted to a constant input flux. Before starting the investigation on the population dynamic, we will make a short detour to the steady state solution analysis. This second limit is analytically tractable in full generality and helps to determine the relevant parameters which might also be involved in the control of the dynamic.

2.2.3 Steady state under a constant flux

The steady state corresponds to the situation where the concentrations of any specie in the system do not evolve any more in time:

$$\frac{\partial c_n}{\partial t} = 0 \quad \text{for} \quad 1 \leq p < N \tag{2.40}$$

which gives once replaced in the previous system of equations:

$$J_n = J_{n+1} = J_s \quad \text{for} \quad 2 < n < N-1 \tag{2.41}$$

$$J_{\text{on}} = J_s + \sum_{n=2}^{N} J_s = N J_s$$

$$\Rightarrow J_s = \frac{J_{\text{on}}}{N} \tag{2.42}$$

Equation (2.30) with result equation (2.42) produces a recursive formula for the concentrations of each aggregate sizes:

$$c_n = \frac{k_{n-1}^+}{k_n^-} c_1 c_{n-1} - \frac{J_{\text{on}}}{k_n^- N} \tag{2.43}$$

which can be solved starting from the free monomeric concentration, to express the concentration of any aggregate size in term of free monomeric concentration:

$$\begin{aligned} c_n &= c_1^n \left(\prod_{q=1}^{n-1} \frac{k_q^+}{k_{q+1}^-} \right) - \frac{J_{on}}{N} \sum_{q=2}^{n} (k_n^-)^{-1} c_1^{n-q} \prod_{r=q}^{n-1} \frac{k_r^+}{k_r^-} \\ &= c_1^n \left(\prod_{q=1}^{n-1} \frac{k_q^+}{k_{q+1}^-} \right) - \frac{J_{on}}{N} \sum_{q=1}^{n-1} c_1^{n-q-1} (k_q^+)^{-1} \frac{k_q^+}{k_n^-} \prod_{r=q+1}^{n-1} \frac{k_r^+}{k_r^-} \\ &= c_1^n \left(\prod_{q=1}^{n-1} \frac{k_q^+}{k_{q+1}^-} \right) - \frac{J_{on}}{N} \sum_{q=1}^{n-1} c_1^{n-q-1} (k_q^+)^{-1} \prod_{r=q}^{n-1} \frac{k_r^+}{k_{r+1}^-} \end{aligned} \tag{2.44}$$

with the convention on empty product $\prod_n^{n-1} = 1$. The input flux can be found in term of c_1 also by inserting the previous solution in the boundary condition at $n = N$ (equation (2.32)):

$$\begin{aligned}
\frac{J_{\text{on}}}{N} &= k_{N-1}^+ c_1 c_{N-1} \\
&= k_{N-1}^+ c_1^N \left(\prod_{n=1}^{N-2} \frac{k_n^+}{k_{n+1}^-} \right) - \frac{J_{\text{on}}}{N} k_{N-2}^+ \sum_{q=1}^{N-2} \left(c_1^{N-q-1} (k_q^+)^{-1} \prod_{r=q}^{N-2} \frac{k_r^+}{k_{r+1}^-} \right) \\
\Rightarrow \frac{J_{\text{on}}}{N} &= \frac{k_{N-1}^+ c_1^N \left(\prod_{n=1}^{N-2} \frac{k_n^+}{k_{n+1}^-} \right)}{1 + k_{N-1}^+ \sum_{q=1}^{N-2} \left(c_1^{N-q-1} (k_q^+)^{-1} \prod_{r=q}^{N-2} \frac{k_r^+}{k_{r+1}^-} \right)} \\
&= \frac{1}{\sum_{q=1}^{N-1} \left(c_1^{-q-1} (k_q^+)^{-1} \prod_{r=1}^{q-1} \frac{k_r^+}{k_{r+1}^-} \right)}
\end{aligned} \tag{2.45}$$

We can replace J_{on} in equation (2.44) by the value derived in equation (2.45), to obtain the analytic expression of the steady state concentration in the system submitted to a constant flux of monomers:

$$\begin{aligned}
c_n &= c_1^n \left(\prod_{q=1}^{n-1} \frac{k_q^+}{k_{q+1}^-} \right) - \frac{\sum_{q=1}^{n-1} c_1^{n-q-1} (k_q^+)^{-1} \prod_{r=q}^{n-1} \frac{k_r^+}{k_{r+1}^-}}{\sum_{q=1}^{N-1} \left(c_1^{-q-1} (k_q^+)^{-1} \prod_{r=1}^{q-1} \frac{k_r^+}{k_{r+1}^-} \right)} \\
&= \frac{\sum_{q=n}^{N-1} c_1^{n-q} (k_q^+)^{-1} \prod_{r=n}^{q-1} \frac{k_r^+}{k_{r+1}^-}}{\sum_{q=1}^{N-1} c_1^{-q} (k_q^+)^{-1} \prod_{r=1}^{q-1} \frac{k_r^+}{k_{r+1}^-}}
\end{aligned} \tag{2.46}$$

The equation (2.45) provides the stationary concentration of monomers at the membrane under a constant flux J_{on}: $c_1(J_{\text{on}})$, which in turn can be used in equation (2.46) to find the steady state concentration of any other specie. So that the steady state is fully solved.

Kinetics results can be linked to the classic thermodynamics of self assembly. Let us reintroduce the classic expression of Helmholtz free energy density of an ideal dilute solution in the linear approximation (see equation (2.11) or equivalently Zandi et al.[136] derivation):

$$\frac{\mathcal{F}}{S} \simeq k_B T \sum_i c_i (\log c_i - 1 + \beta f(i)) \tag{2.47}$$

where c_i represent the dimensionless concentration of specie i in the aggregate solution on the membrane $\frac{N_i s_0}{S}$. The chemical potential of the free monomers is defined as:

$$\mu_1 = \left. \frac{\partial \frac{\mathcal{F}}{S}}{\partial c_1} \right|_{V,T,C_{i\neq 1}} = k_B T \log c_1 + f(1) \tag{2.48}$$

and we use it to define the energy cost to create a aggregate of proteins compared to the reference state of freely diffusing proteins. This aggregate formation free energy writes:

$$\Delta f(n) = f(n) - n.\mu_1 \quad \text{with} \quad \mu_1 = k_B T \log c_1 + f(1) \tag{2.49}$$

and the previously exposed steady states equations can be rewritten in terms of the free energy cost using the detailed balance condition equation (2.28):

$$\begin{aligned} \frac{k_n^+}{k_{n+1}^-} &= e^{-\beta(f(n+1)-f(n)-f(1))} \\ \Rightarrow \quad \prod_{r=n_1}^{n_2} \frac{k_r^+}{k_{r+1}^-} &= e^{-\beta\sum_{r=n_1}^{n_2} f(r+1)-f(r)-f(1)} \\ &= e^{-\beta(f(n_2+1)-f(n_1)-(n_2-n_1+1)f(1))} \end{aligned}$$

which gives in equations (2.45) and (2.46):

$$\frac{J_{\text{on}}}{N} = \frac{c_1^\infty}{\sum_{q=1}^{N-1}\left(e^{\beta\Delta f(q)}/k_q^+\right)} \tag{2.50}$$

$$c_n^\infty = e^{-\Delta f(n)} \frac{\sum_{q=n}^{N-1} e^{\beta\Delta f(q)}/k_q^+}{\sum_{q=1}^{N-1} e^{\beta\Delta f(q)}/k_q^+} \tag{2.51}$$

Where superscript ∞ emphasis that these relations hold at steady state. Now that we are able to calculate the final composition of the system submitted to a constant flux, we will apply our results to understand the dynamical evolution starting with the simpler case of constant attachment rates.

2.2.4 Linear energy and fixed on-rate

In a first approximation, the forward rate is assumed to be constant for all sizes of proteins aggregates:

$$k_n^+ = k^+ \quad \forall n \tag{2.52}$$

Furthermore we keep only the linear terms in the expression of the energy of the bud (equation (2.27)), $f(n) = (n-1)\epsilon_b$ leading to a very simple detailed balance expression equation (2.28) and to a constant backward rate:

$$\frac{k_{n+1}^-}{k_n^+} = e^{\beta\epsilon_b} = c_* \tag{2.53}$$

In this constant case (constant chemical rates, constant flux), a scaling law can be found out. Let's consider for instance a modification of the value of the input flux $J_{\text{on}} \longrightarrow K_{\text{on}}$. We are going to look under which condition the new solution can be found from a rescaling of a former solution of the dynamic equation by applying a constant dilatation of time and a re-normalization of all the concentration in the system:

$$\begin{cases} t & \longrightarrow & \eta t \\ c_n & \longrightarrow & d_n : t \mapsto \alpha c_n(\eta t) \end{cases} \tag{2.54}$$

The new set of equation is:

$$\partial_t\, d_n = K_n - K_{n+1} \quad \text{with} \quad K_n = k_d^+(d_1 d_{n-1} - d_* d_n) \quad \text{for} \quad 1 < n < N \tag{2.55}$$

$$\partial_t\, d_1 = K_{\text{on}} - K(2) - \sum_{n=2}^{N} K(n) \tag{2.56}$$

and transforms into:

$$\partial_t c_n = \frac{1}{\alpha\eta}(K_n - K_{n+1}) \quad \text{with} \quad K_n = \alpha^2 \frac{k_d^+}{k_c^+} k_c^+ (c_1 c_{n-1} - \frac{d_*}{\alpha c_*} c_* c_n) \quad \text{for} \quad 1 < n < N \tag{2.57}$$

$$\partial_t c_1 = \frac{1}{\alpha\eta}\left(K_{\text{on}} - K(2) - \sum_{n=2}^{N} K(n)\right) \tag{2.58}$$

We want the new equations to be equivalent to the first set of equation from which c_n is a solution equations (2.30) and (2.32), identification of α and η leads to the condition for the two solutions to collapse to the same master curve as shown on figure 2.5:

$$\begin{cases} \dfrac{K_{\text{on}}}{d_*^2 k_d^+} = \dfrac{J_{\text{on}}}{c_*^2 k_c^+} = \Omega \\ \dfrac{d_n}{d_*} = \dfrac{c_n}{c_*} \\ \tau . k_d^+ d_* = t . k_c^+ c_* \end{cases} \tag{2.59}$$

As a consequence all problems with the same Ω are equivalent up to a rescaling of the solution and we can fix k^+ and c_* without any loss of generality. Please note that this demonstration holds only for a fixed maximum size of aggregates (defining the number of equation in the set). Should N change the solutions will not be comparable apart asymptotically. The ratios will obviously appear naturally in the following calculations.

2.2.4.1 Steady state

The steady state equations can be further simplified as they end up to be geometric series:

$$\frac{J_{\text{on}}}{N} = \frac{k^+ c_1^\infty (c_1^\infty - c_*)}{1 - (c_1^\infty / c_*)^{1-N}} \tag{2.60}$$

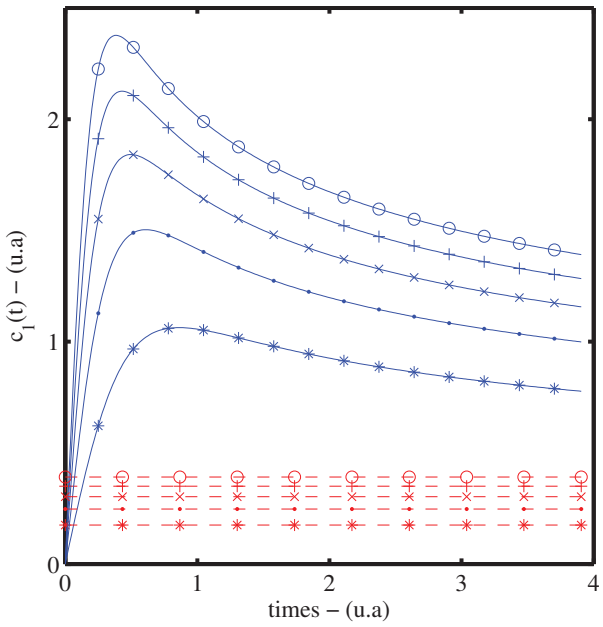
$$c_n^\infty = c_1^\infty \frac{1 - (\frac{c_1^\infty}{c_*})^{n-N}}{1 - (\frac{c_1^\infty}{c_*})^{1-N}} \tag{2.61}$$

As shows figure 2.6 picturing the concentrations profiles at steady state for different input flux, the system switches between two regimes with very different behaviors. Either only the small sizes are formed and a very small flux pass through the system or all the different sizes are present in equivalent proportions with a larger flux throughout the system. The relationship between those two different states can be understood by studying the mean growth velocity as steady state is reached (see equation (2.39)). At steady state, the change in size for protein patch reduces to a biased random walk with average growth velocity and diffusion coefficient:

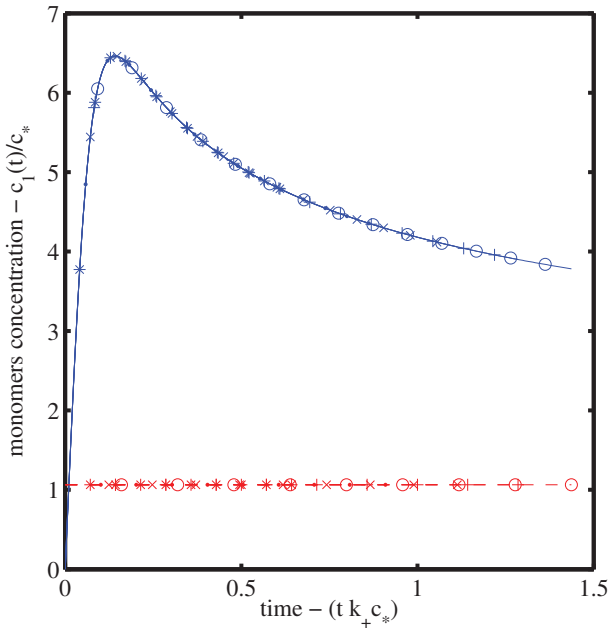
$$\bar{v} = k^+ c_1^\infty - k^- \tag{2.62}$$

$$\bar{D} = k^+ c_1^\infty + k^- \tag{2.63}$$

When $c_1^\infty < c_*$, this average growth speed is negative and the system is non-propagating: aggregates formed by a fluctuation tend to shrink and the chances to reach the fully closed



(a) Five different solutions for the evolution of monomers concentrations (plain blue lines) with $\Omega = 10$ and $k^+ = 1$ and $J_{\text{on}} \in \{20(+), 40(*), 60(.), 80(\times), 100(\circ)\}$. Their respective steady state limits c_1^∞ calculated from expression equation (2.45) are shown (red dashed)



(b) Same solutions with rescaled time and concentrations: all the curves are equivalent and collapse on the same master curve

Figure 2.5 – Scaling of the kinetic equation with constant rates for $\Omega = \frac{J_{\text{on}}}{k^+ c_*^2} = 10$.

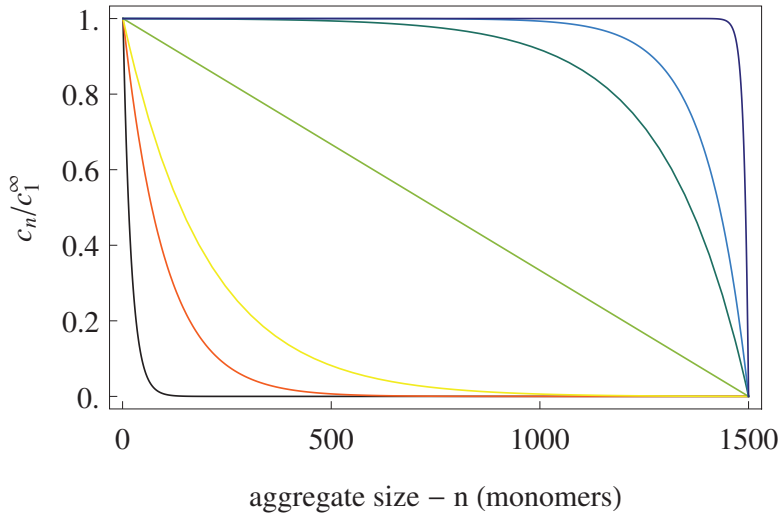


Figure 2.6 – Steady state concentration of the various sizes of aggregates in the system normalized by the monomer concentration, for different input flux of monomers in the cell (from lower left to upper right, $\Omega \in \{0.99, 0.995, 1, 1.005, 1.01, 1.1\}$ for a number of proteins in the fully formed virions $N = 1500$).

state to escape are low. Oppositely, when $c_1^\infty > c_*$, the average growth speed turns positive and the system is propagating. The condition can be rewritten as $c_1^\infty > c_*$. In terms of energetic, the condition writes: $\frac{c_1}{c_*} > 1 \Leftrightarrow e^{\beta(\mu_1+\epsilon_b)} > 1 \Leftrightarrow \mu_1 > -\epsilon_b$, proteins will regroup in aggregates as far as binding reduces their free energy.

With this description it appears that if the system submitted to a input flow of monomers is in its non-propagating state, the monomers will accumulate until propagation starts, in contradiction with the establishment of a steady state. There is indeed a mechanism left to assure the transportation material throughout the assembly line: the diffusion which balances the drift, resulting in a net flux. Only low flux are expected to be sustainable in such a regime, and this is what one can see from the relation $J_{\text{on}}(c_1^\infty)$ equation (2.60) that is plotted on figure 2.7. We note however that the existence of a steady state with positive input flux and negative mean velocity is restricted to finite N. In fact, figure 2.7, illustrates that eligible J_{on} tends to 0 for c_1^∞ with the power of N (dominant term of denominator in equation (2.60)). Hence a continuous description such as proposed in equation (2.38), somehow equivalent to taking $N \to \infty$, will collapse all those states to $J_{\text{on}} = 0$. It is however possible to link discrete and continuous situation, by looking for steady states with fixed concentration c_1^∞ at $x = 0$ and $c_N = 0$ at $x = N$. Two successive integrations lead to:

$$c(x) = c_1^\infty \frac{1 - e^{\bar{v}(x-L)/\bar{D}}}{1 - e^{\bar{v}L/\bar{D}}} \tag{2.64}$$

Which looks very much like our discrete expression and from which it can be inferred that steady states concentrations are constant but on a small layer close to the boundary of typical size $\bar{D}/\bar{v}$[110]. But finding c_1^∞ for a given Ω is assembly length dependent and cannot be done in the continuous approximation.

In large systems, $\Omega(c_1^\infty)$ tends to a step function resulting in a abrupt change in the vicinity

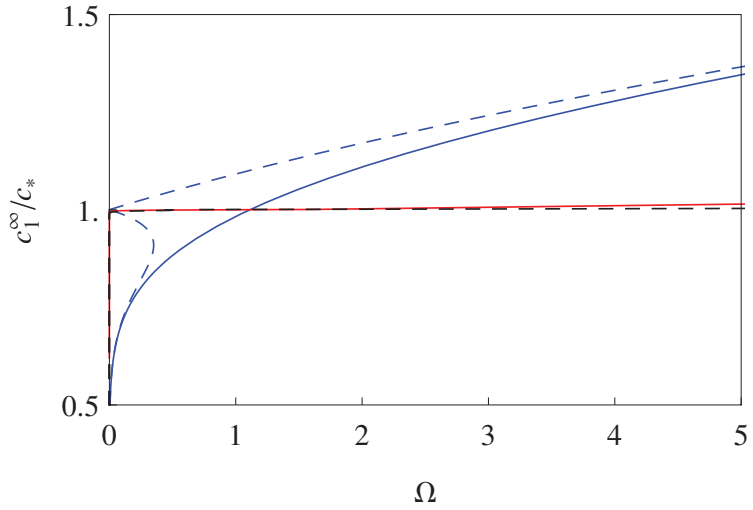


Figure 2.7 – Relation between the steady monomers concentration c_1^∞ and the input flux of aggregates Ω at the entrance of the assembly line. The effect of the system size on the transition is clear around $c_1^\infty/c_* = 1$ (solid lines: $N = 10$–blue and $N = 1500$–red). Approximations (equation (2.65)) for $c_1^\infty < c_*$ and $c_1^\infty > c_*$ are shown (dashed lines). They fail in the neighborhood of $c_1^\infty = c_*$

of $c_1^\infty \simeq c_*$:

$$\frac{\Omega}{N} \simeq \begin{cases} (1-\frac{c_1^\infty}{c_*})\left(\frac{c_1^\infty}{c_*}\right)^N \simeq 0 & \text{for} \quad c_1^\infty < c_* \\ \frac{1}{N-1} & \text{for} \quad c_1^\infty = c_* \\ \frac{c_1^\infty}{c_*}(\frac{c_1^\infty}{c_*}-1) & \text{for} \quad c_1^\infty > c_* \end{cases} \tag{2.65}$$

Around $\Omega = 1$, the two limits of our approximations are different which tells us that our approximation is wrong at the transition between the two regimes discussed for $c_1^\infty \simeq c_*$. Far from the critical point for $\Omega = 1$, we can solve the approximate forms to find $c_1^\infty(\Omega)$:

$$\frac{c_1^\infty}{c_*} \simeq \begin{cases} \left(\frac{\Omega}{N}\right)^{\frac{1}{N}} & \text{for} \quad \Omega \ll 1 \\ \frac{1}{2}\left(1+\sqrt{1+4\frac{\Omega}{N}}\right) & \text{for} \quad \Omega \gg 1 \end{cases} \tag{2.66}$$

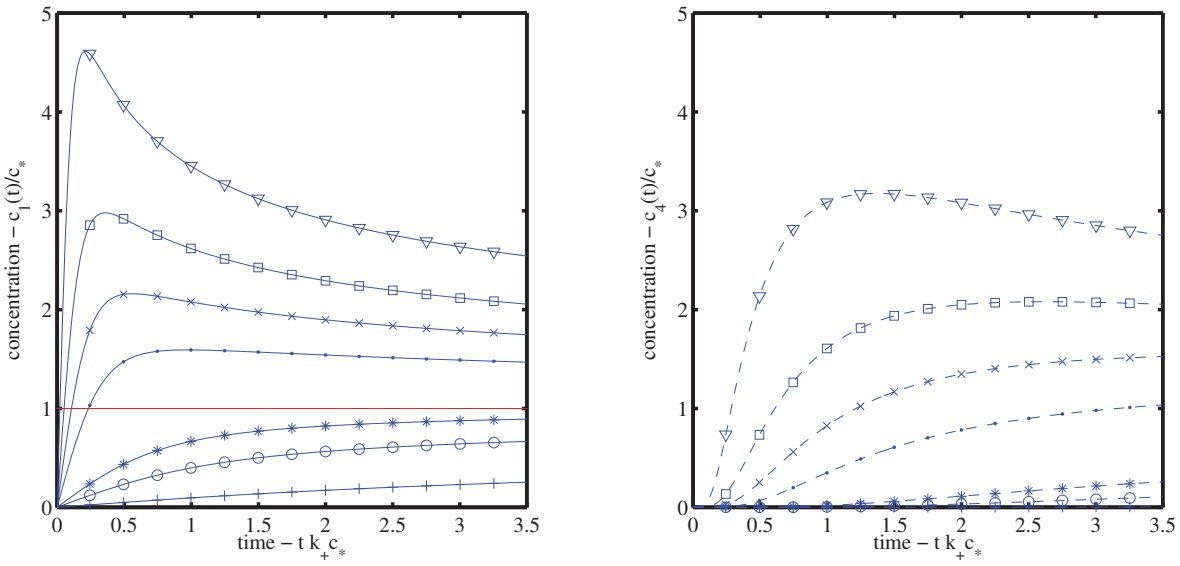
2.2.4.2 Dynamics

We now turn ourself to the time evolution. Steady state study let appear an important quantity controlling the long time evolution of the system: the ratio $\Omega = \frac{J_{\text{on}}}{k^+ c_*^2}$, linked to the critical concentration $c_* = e^{-\beta\epsilon_b}$.

Unlike in the steady state, we have not been able to solve analytically the equation for the dynamic. To study the evolution of the monomers and aggregates population, we used a discrete time step Δt and numerically solved the system of equations using a simple explicit Euler method scheme.

Similarly to what was demonstrated concerning the steady state, dynamic itself behaves in two different ways depending on the value of Ω as shown on figure 2.8:

— $\Omega \leq 1$: the evolution is monotonic. The aggregates concentration exhibit sigmoidal evolutions. Following a lag time, longer as the size of the aggregates of interest is bigger,



(a) Monomers concentration $c_1(t)$ and critical concentration (c_*, red line) normalized by c_*

(b) Tetramers concentration $c_4(t)$ normalized by c_*

Figure 2.8 – The two different behaviors shortly after the input flux J_{on} as been turned on at $t = 0$ depending on $\Omega = \frac{J_{\text{on}}}{k^+ c_*^2} \in \{0.1(+), 0.5(\circ), 1(*), 5(.), 10(\times), 20(\square), 50(\triangledown)\}$, is shown for the early rising populations –monomers $c_1(t)$ (plain) and tetramers $c_4(t)$ (dashed)). The lag time between the arrival of monomer in the system and the building of tetramers depends visibly on $\Omega(J_{\text{on}})$

the concentrations gradually rise in the whole system towards the final steady state value. The stronger the input flux and the smaller the time to reach the steady state.

— $\Omega \geq 1$: the evolution is non monotonic. The concentration of each aggregates rises to eventually outnumber the steady state concentrations followed by heavily damped oscillations (only apparent for very high $\Omega \gg 10^2$) around the steady state value.

As discussed in section 2.2.2, we expect the solutions of our model to result from the transport of the monomer concentration fluctuations throughout the different sizes of aggregates, with deformations due to the diffusion. Consequently the study of $c_1(t)$ is rich of information concerning the time evolution of the whole system. We can also expect a change in the evolution as the first nucleated aggregates reach the final size N and leaves the system. We will begin by focusing on the early evolution of monomer concentration when the propagation of aggregates has not been sufficient for the final boundary condition to affect the solution.

2.2.4.3 Short times: overshot

When $\Omega \gg 1$, the concentration of monomers peaks to a value much larger than the final steady state value. This behavior is not obvious to characterize as it results from the competition between the constant flux of monomers entering the system and their consumption both in dimers nucleation (at a rate proportional to $c_1^2(t)$) and in the growth or shrinkage of the formed aggregates (proportional to the monomer concentration $c_1(t)$ times the sum of all the other aggregates concentrations in the system) as stated by equation (2.31).

As this is a feature of the early evolution, observed for characteristic times $\tau_{\max}$ much shorter than the typical time to reach the end of the system ($\tau_n \sim \frac{N}{k^+ c_*}$) this evolution should be insensitive to the size of the line assembly in our range of interest ($N \gg 10^2$). Simulations show that maximum value scales like $\Omega^{\frac{1}{2}}$ (figure 2.9) as does the steady state value c_1^∞ when the input flux is large compared to the size of the system (Eq equation (2.66) with $\Omega \gg N$). It stresses again that, in the steady state regime, the size of the assembly line indeed plays an important role whereas the early state of the system looks insensitive to N. Furthermore, The time to reach this maximum value also seems to follow $t_{\max} \sim \Omega^{-\frac{1}{2}}$. Since aggregates concentrations play an equivalent role in the monomer consumption compared to the quadratic term of dimers formation, we introduce the concentration of aggregates of all sizes in the system:

$$\Xi(t) = \sum_{n=2}^{N-1} c_n(t) \tag{2.67}$$

To understand the effect of the different factors involved in the early evolution, we look for a rearrangement of the kinetics equations (2.30) and (2.31), into a simpler closed form on c_1 and Ξ:

$$\begin{aligned} \partial_t \, \Xi &= \sum_{n=2}^{N-1} \partial_t \, c_n \\ &= \sum_{n=2}^{N-1} (J_n - J_{n+1}) \\ &= J(2) - J(N-1) \\ &= k^+ c_1^2 - k^- c_2 - k^+ c_1 c_{N-1} \end{aligned} \tag{2.68}$$

This last equation tells that aggregates are added to the system only through nucleation of dimers and removed both when dimers break or fully formed particle exit.

$$\begin{aligned}\partial_t c_1 &= J_{\text{on}} - k^+ c_1^2 + k^- c_2 - \sum_{n=2}^{N} (k^+ c_1 c_{n-1} - k^- c_n) \\ &= J_{\text{on}} - (k^+ c_1 - k^-)\Xi - 2k^+ c_1^2 + k^- c_2\end{aligned} \tag{2.69}$$

This second equation expresses that monomer are introduced in the system by the input flux J_{on}, and they are consumed either by the growing aggregates at the average growing rate $(k^+ c_1 - k^-)$ or to form dimers (two monomers are required to form a dimer or released as it breaks as the last terms account for)... The analysis is simplified with further approximations. We neglect $k^- c_2$ compared to the full aggregate consumption $-k^- \Xi$ in equation (2.69) and consistently neglect the same term in equation (2.68) (thus formed aggregates cannot disappear any more). And as we currently study the early times, we will also assumes that there is no aggregates of great sizes in the system $c_{N-1} \simeq 0$. Under those two assumptions, we obtain a non-linear closed system of two equations for $c_1(t)$ and $\Xi(t)$:

$$\left\{ \begin{array}{ll} \partial_t c_1 & = J_{\text{on}} - (k^+ c_1 - k^-)\Xi - 2k^+ c_1^2 \\ \partial_t \Xi & = k^+ c_1^2 \end{array} \right. \tag{2.70}$$

Unfortunately, we have not been able to solve analytically this set of equation, event if the second one appears obvious and we can have an integro-differential equation on c_1:

$$\left\{ \begin{array}{ll} \partial_t c_1 & = J_{\text{on}} - (k^+ c_1 - k^-) \int_{\tau=0}^{t} k^+ c_1(\tau)^2 \, \mathrm{d}\tau - 2k^+ c_1^2 \\ \Xi(t) & = \int_{\tau=0}^{t} k^+ c_1(\tau)^2 \, \mathrm{d}\tau \end{array} \right. \tag{2.71}$$

However as the system is originally empty, $c_1(0) = 0$ and we have the approximations:

$$\partial_t c_1|_{t\to 0} = J_{\text{on}} \Rightarrow c_1(t) \sim J_{\text{on}}.t \tag{2.72}$$

$$\Rightarrow -2k^+ c_1^2 \sim t^2 \tag{2.73}$$

$$\Rightarrow \int_{\tau=0}^{t} k^+ c_1^2(\tau) \, \mathrm{d}\tau \sim t^3$$

$$\Rightarrow (k^+ c_1(t) - k^-) \int_{\tau=0}^{t} k^+ c_1^2(\tau) \, \mathrm{d}\tau \sim t^3 \tag{2.74}$$

If we drop the highest order term (thus neglecting the aggregate growth compared to their nucleation), we eventually have an equation on monomer concentration which solution can be found by variable separation:

$$\left\{ \begin{array}{ll} \partial_t c_1 & = J_{\text{on}} - 2k^+ c_1^2 \\ c_1(0) & = 0 \end{array} \right. \quad \Rightarrow \quad c_1(t) = \sqrt{\frac{J_{\text{on}}}{2k^+}} \tanh\left(\sqrt{2k^+ J_{\text{on}}}.t\right) \tag{2.75}$$

This solution captures most of the main features we have seen to this point:

1. scaling law is maintained throughout the approximation steps $c_1(t) = c_* \sqrt{\frac{\Omega}{2}} \tanh\left(\sqrt{2\Omega} c_* k^+ t\right)$
2. $\partial_t c_1|_{t=0} = J_{\text{on}}$
3. maximum value is of order $\sqrt{\Omega}$ for fixed c_* and k^+

4. typical time to reach maximum is $\Omega^{-\frac{1}{2}}$ for fixed c_* and k^+

In this case, however, the monomer concentration does not peak as when the aggregates consumption contribution is present. This tells us that the overshot behavior is an effect of this later contribution rather than an effect of the nucleation non linear term ($k^+c_1^2$). Overshot behavior could be observed for similar growth dynamic with different initiation steps, such as initiation on seeds which involve a linear nucleation step with respect to the monomer concentration ($k_{\text{seed}}c_1$) or nucleation steps involving more monomers at a time.

2.2.4.4 Intermediate times: propagation

In the previous section, we have shown that the initiation of growing aggregates as the monomer population was increased by a constant flux was sufficient to overtake the incoming proteins flux and deplete the monomer population. We have identified the limit rate $J_{\text{critical}} = k^+c_*^2$ (equivalently $\Omega = 1$) beyond which the system enters this regime or below which the monomer population gently grows towards its steady state concentration. But in order to perform our analysis of the early stage of dynamic, we looked at the whole population of aggregates without paying more attention at its size distribution and we focused on $c_1(t)$ evolution. This choice was prompted by the results of section 2.2.2, where we pointed out that size evolution of aggregates over time has to be a drift-diffusion process controlled with $c_1(t)$ playing both role of emission source and driving force. We now turn our attention to the implication of the previously described $c_1(t)$ on the bud evolution.

In $\Omega < 1$ regime, we already understood that the system remains non propagating all the time until steady state is reach. When submitted to such low flux, the system gradually fills as diffusion equilibrates populations with consecutive sizes but with a strong attenuation as size increase.

In $\Omega \gg 1$ regime, There is a change in the system state as $c_1(t)$ becomes higher than c_* and formed aggregates begin to grow with a positive mean rate. As all accumulated small aggregates grow with the same mean velocity a front starts to propagate. Meanwhile the consumption of monomers increases faster that its renewal rate. The consequent diminution of the monomer concentration leads to a diminution of the dimers nucleation and eventually of all the aggregates populations at the back of the propagating front. A wave with a large tail propagates towards large sizes at speed $v(t) = k^+c_1(t) - k^-$ as shown on figure 2.11b.

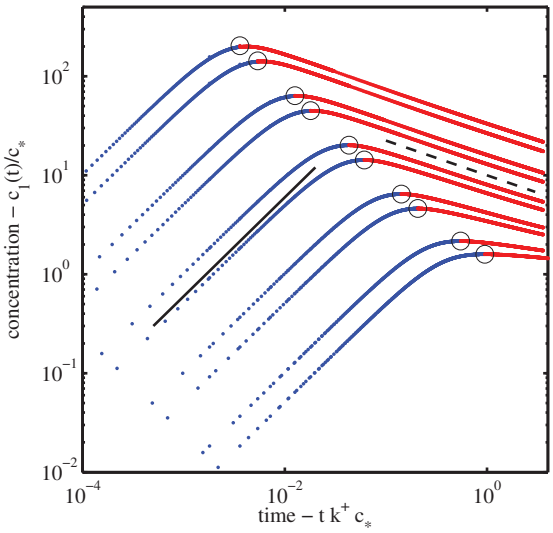
On the qualitative biological ground, the mechanism predicts that the distribution of the growing buds will be peaked close to the maximum observable size until viruses are released outside the cell.

In fact, without diffusion, the solution of the propagation equation with $c_1(t)$ boundary condition would be the monomer concentration translated in space at speed $v(t)$:

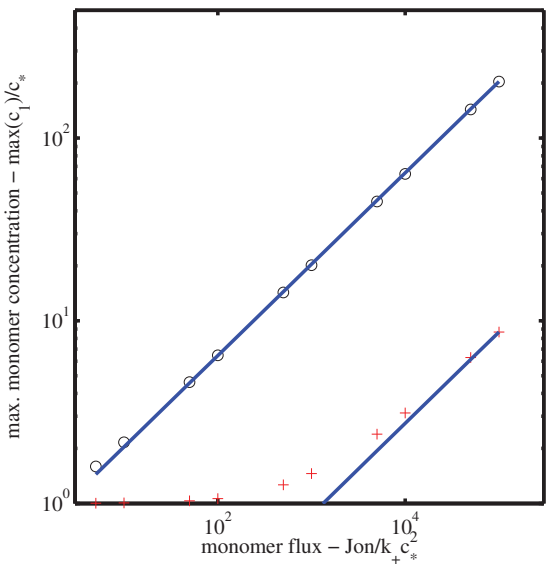
$$c(x,t) = c_1\left(\int_{\tau=0}^{t} v(\tau)\,\mathrm{d}\tau - x\right) \tag{2.76}$$

2.2.4.5 Long times: propagation in large size systems

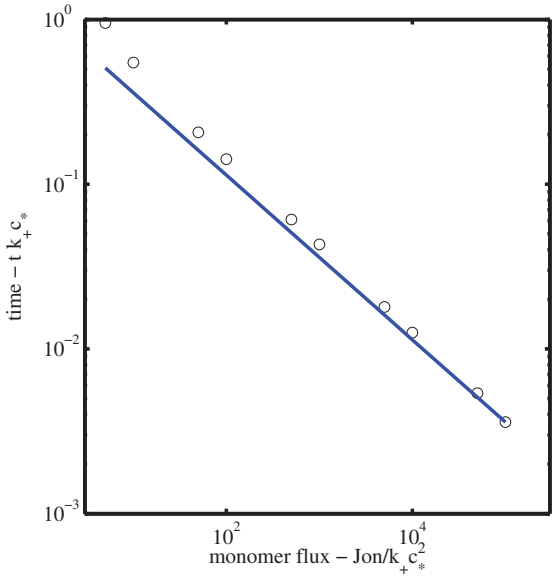
As monomer concentration keeps decreasing, rapidly consumed by the aggregates, the growth speed falls too. Simulations show that $c_1(t)$ seems to eventually reach a “quasi steady



(a) Log-Log plot of the first peak of the monomer concentration as J_{on} is turned on for $\Omega = \{5, 10, 50, 10^2, 5.10^2, 10^3, 5.10^3, 10^4, 5.10^4, 10^5\}$ (from bottom to top) and $N = 1500$. The earliest evolution is linear on time $c_1(t \to 0) \sim J_{\text{on}}.t$ (black plain line), until nucleation of dimers begins. Then the concentration reaches a maximum ($\circ$) and decreases. Very high Ω seems to evolve with the same power law after maximum ($\sim t^{-0.35}$ – dashed black line), which would not hold for values of order 1. However a closer look emphasis that this part of the curve is slightly concave rather than linear.

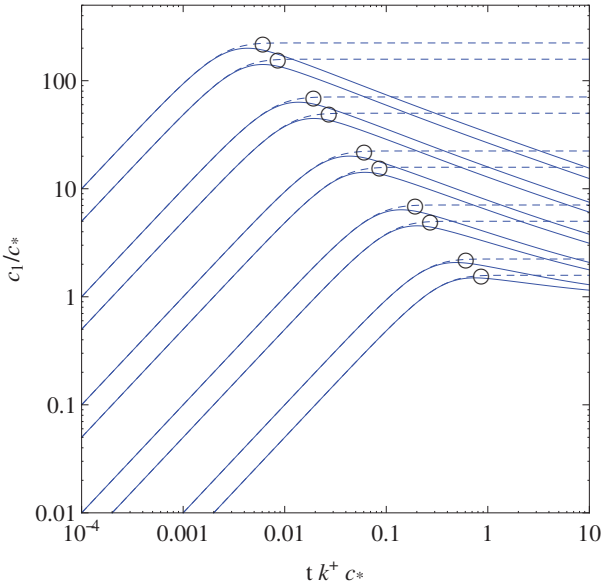


(b) Log-Log plot of the monomer concentration maximum value ($c_1^{\max}$, $\circ$) compared to the steady state final concentration (c_1^{∞}, $+$) as functions of the input flux value (Ω identical as figure 2.9a). The scaling in $\Omega^{\frac{1}{2}}$ expected for c_1^{∞} from equation (2.66) for $\Omega \gg N$, seems to hold also for $c_1^{\max}$ without restriction on N (solid blue lines)

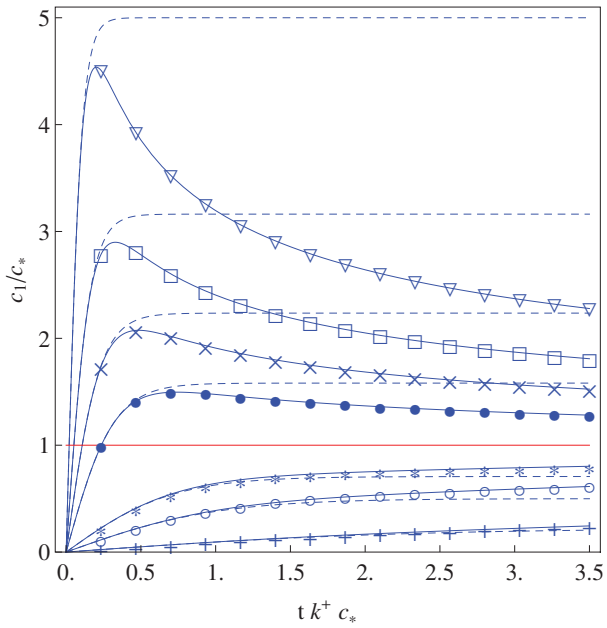


(c) Log-Log plot of the time $t_{\max}$ to reach the peak value ($c_1(t_{\max}) = c_1^{\max}$, $\circ$) as function of the input flux value (Ω identical as figure 2.9a). $t_{\max}$ seems to scale as $\Omega^{-\frac{1}{2}}$ (for $\Omega \gg N$) which manifests itself on figure 2.9a by the maxima aligned with $c_1(t_{\max}) \sim t_{\max}^{-1}$

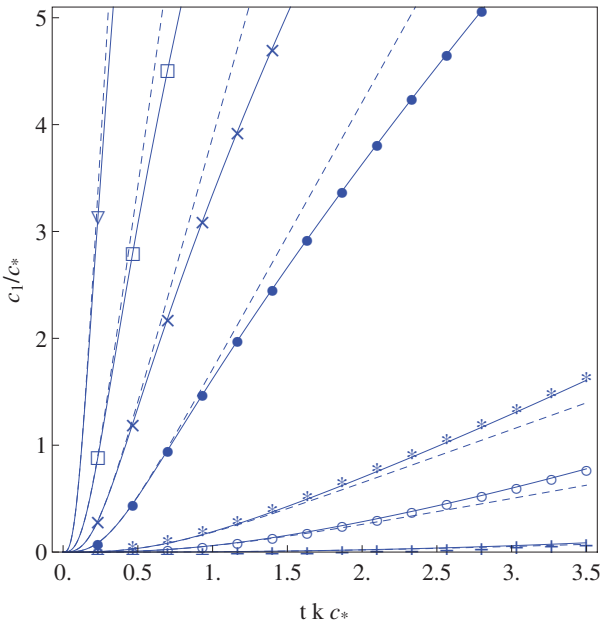
Figure 2.9 – Short time scale, overshoot



(a) Log-Log plot of the first peak of the monomer concentration as J_{on} is turned on for $\Omega = \{5, 10, 50, 10^2, 5.10^2, 10^3, 5.10^3, 10^4, 5.10^4, 10^5\}$ (from bottom to top) according to approximation from equation (2.70) to compare with the full resolutions figure 2.9. Scaling of the maximum position and time is obtained from equation (2.75) are shown (∘).

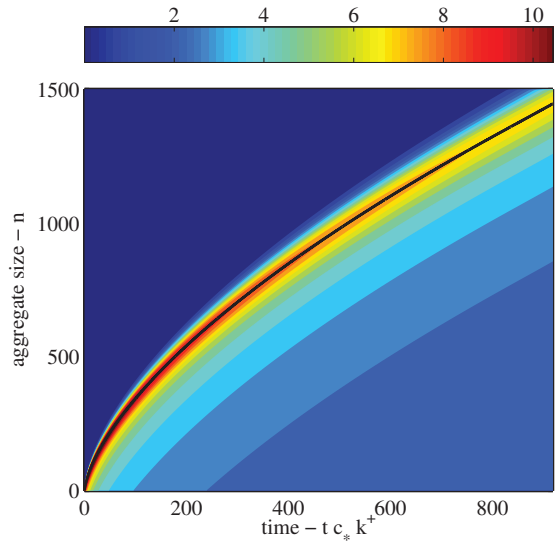


(b) Monomers concentration $c_1(t)$ and critical concentration (red line)

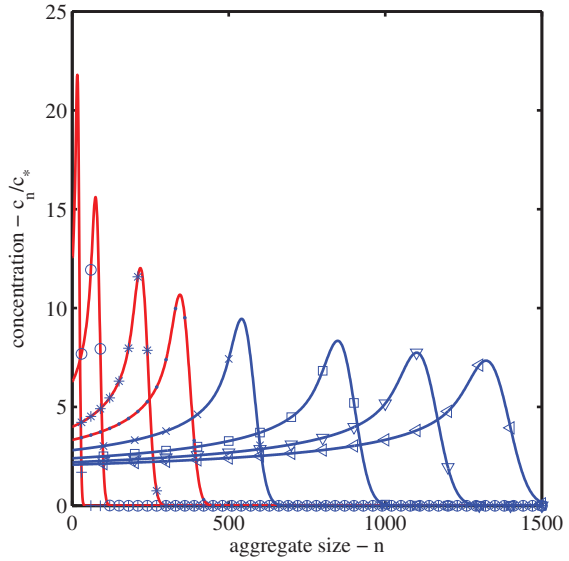


(c) Oligomers concentration $\Xi(t)$

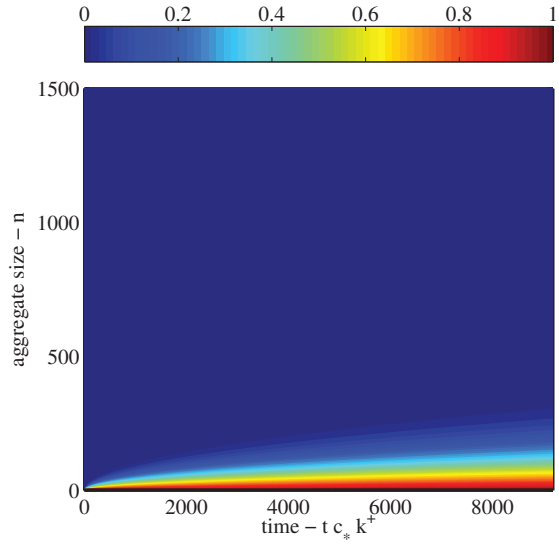
Figure 2.10 – Evolution of the species concentrations according to our approximation (equation (2.70) – plain lines) shortly after the input flux J_{on} as been turned on at $t = 0$ for various $\Omega \in \{0.1(+), 0.5(\circ), 1(*), 5(.), 10(\times), 20(\square), 50(\triangledown)\}$, are shown for the monomers concentration $c_1(t)$ (left) and aggregates concentration $\Xi(t)$(right)). Neglecting the dimers breaking makes no visible difference compared to the full calculation (see figure 2.8a for comparison). Approximation from equation (2.75) (dashed lines) works well for $c_1(t)$ until the inflection point at $k^+c_*t = (2\Omega)^{-\frac{1}{2}}$ where concentrations either drops (large $J_{\text{on}} \gg k^+c_*^2$) or increase further (small $J_{\text{on}} < k^+c_*^2$)



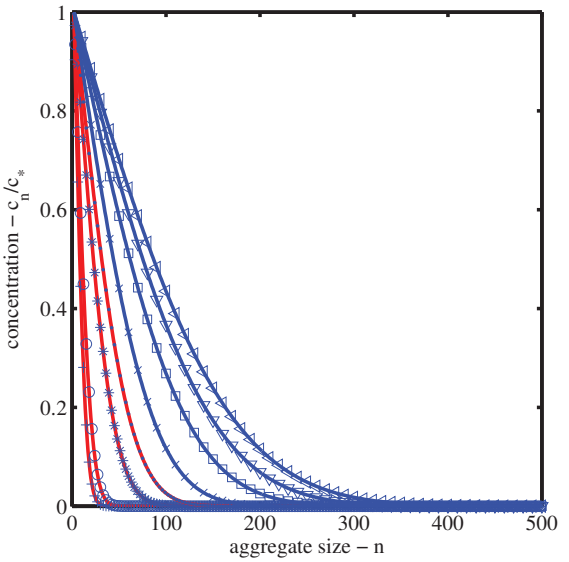
(a) Time evolution of the size distribution in the system for $\Omega = 5.10^3$ and $N = 1500$: Colors indicate concentrations $\frac{c_n}{c_*}$ for each aggregate sizes in time. The wave trajectory ($n(t) = \max_n c_n(t)$ – plain black line) shows a strong deceleration followed by a stabilization of the speed



(b) The propagating wave: Snapshots of the sizes distribution aside at various times ($t_i k^+ c_* \in \{1(+),10(\circ),50(*),100(.)\}$ – plain red) and ($t_i k^+ c_* \in \{200(\times),400(\square),600(\triangledown),800(\triangleleft)\}$ blue).



(c) Time evolution of the size distribution in the system for $\Omega = 0.999$ and $N = 1500$. The monomer concentration is always the highest ($n(t) = \max_n c_n(t)$ – plain black line) and asymptotically rises to its steady state value allowing the system to progressively fill.



(d) The diffusing front: Snapshots of the sizes distribution aside at various times ($t_i k^+ c_* \in \{50(+),100(\circ),500(*),1000(.)\}$ – plain red) and ($t_i k^+ c_* \in \{2.10^3(\times),4.10^3(\square),6.10^3(\triangledown),8.10^3(\triangleleft)\}$ blue).

Figure 2.11 – $\Omega = 5000 \gg 1$ assembly wave (top) and $\Omega = 0.999 \simeq 1$ diffusion front (below) propagating in the system. Beware the change in axis due to the different scales of those two process both in time and size.

state" with very low evolution, asymptotically approaching c_*. On one hand, should $c_1(t)$ become smaller than c_*, the aggregates would shrink and $c_1(t)$ would rise again. On the other hand for any $c_1 > c_*$, both aggregates nucleation and growth occur. As the number of aggregates only increases when $c_1 > c_*$, so does the total amount of monomers needed to maintain the growth. In a closed system with fixed amount of proteins Morozov et al.[98] concluded that similar propagating front would eventually stop after having covered a finite distance as c_1 reach c_*. With the fixed input flux sustaining the growth, the story might be somehow different, even if the velocity should eventually decrease towards 0. Indeed its decay is very low and in the simulations it rather looks like the speed does not evolve on the time scale needed to reach the virion state (full completion). Unlike the early peak, we have not been able to characterize this decay rate.

To illustrate the previous discussion, we will evaluate the asymptotic velocity evolution when growth is limited by the input flux. Let us consider the front propagation with no boundary (in an infinite system $N \to \infty$). As $c_1(t)$ quickly decreases towards c_*, the tail height is of the same order. We neglect the exact form of the propagating wave and take instead a simple step function from $x = 0$ to $x = n_{\max}(t)$ dominated by the tail contribution:

$$\begin{cases} c(x,t) = c_* \quad \text{for} \quad x \leq n_{\max}(t) \\ c(x,t) = 0 \quad \text{for} \quad x > n_{\max}(t) \\ v(t) \equiv \partial_t \, n_{\max}(t) \end{cases} \tag{2.77}$$

For the front to move forward to the next size ($\Delta n_{\max} = 1$), one has roughly to supply one monomer to each aggregate of each sizes in $[0, n_{\max}(t)]$, which takes Δt with an input flux J_{on}:

$$\begin{aligned} J_{\text{on}} \Delta t &= \sum_{n=1}^{n_{\max}(t)} c_* = n_{\max}(t).c_* \\ \Rightarrow v(n_{\max}(t)) &= \frac{\Delta n_{\max}}{\Delta t} \sim \frac{J_{\text{on}}}{n_{\max}(t) c_*} \end{aligned} \tag{2.78}$$

We now have to estimate the time t it takes to reach size $n_{\max}(t)$. Using a simple material conservation argument, this is equivalent to estimate the total number of monomers needed to build all the aggregates. By definition, an aggregate of size n contains n monomers:

$$J_{\text{on}}.t = \sum_{n=0}^{n_{\max}(t)} n.c_* \sim c_* \frac{n_{\max}^2}{2} \tag{2.79}$$

Which leads substituting $n_{\max}$ from equation (2.79) in equation (2.78):

$$\begin{aligned} v(t) = k^+(c_1(t) - c_*) &\sim \sqrt{\frac{2 J_{\text{on}}}{c_* t}} \\ \Rightarrow (c_1(t)/c_* - 1) &\sim \sqrt{\frac{2\Omega}{k^+ c_* t}} \\ \Rightarrow n_{\max}(t) &\sim \int_{\tau=0}^{t} v(\tau)\, \mathrm{d}\tau \sim 2\sqrt{2.\Omega.t.k^+ c_*} \end{aligned} \tag{2.80}$$

Unlike in a closed system, the front eventually moves as the square root of the time and does not stop. We note that this behavior is linked to "material conservation" in the step function front propagation rather than to the physics of diffusion. Monomer concentration should tend to c_* as $t^{-1/2}$ if our approximations are correct –this is indeed observed on figure 2.12.

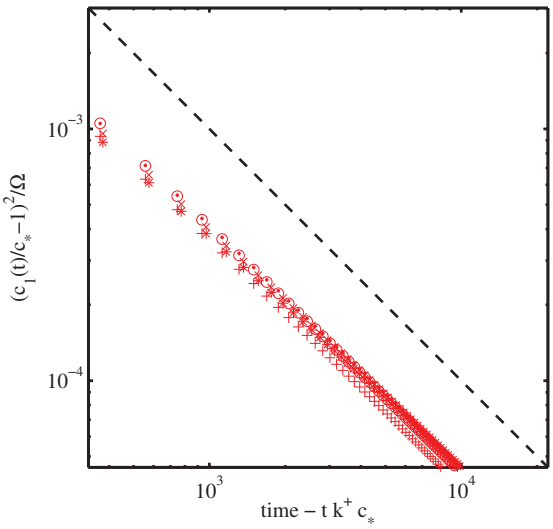


Figure 2.12 – Scaling of the asymptotic decrease of the velocity of the propagating wave in time for $\Omega \in \{5(+), 10(\circ), 20(*), 50(.), 100(\times)\}$ (red) in an assembly line long enough so that the absorbing boundary does not affect the evolution during running time ($N = 3000$). Dashed line shows $y = x^{-1}$. The expected power law and dependency in Ω are correct since all the curves collapse.

2.2.4.6 Long times: oscillations in finite systems

According to our model, the formed virion reaches the absorbing boundary of the system at size N. Or, in other words, when growing viruses reach their final size, they immediately escape and do not weight any more on the monomer population consumption. So in small enough system, the asymptotic flux limited evolution discussed in the previous section is not reached and long time is dominated by the boundary condition.

In $\Omega < 1$ subcritical regime, this takes place at a very slow rate due to the evanescent tail of the diffusion front. And concentrations eventually reach their steady states.

In $\Omega \gg 1$ overcritical regime, the change of the aggregate population is important since buds are grown in a wave. If the tail of the assembly wave is a small part of the whole aggregate population compared to the front peak (large overshoot $\Omega \gg 1$ and small system sizes N), the aggregate population rapidly vanishes altogether with the monomer consumption. Situation is then very much alike the early time described in section 2.2.4.3. The monomer population increases again and peaks a second time. Then a new assembly wave forms and propagates through the assembly line (see figure 2.14). However the amount of aggregates remaining in the system is frequently important and does not match the early evolution. Simulations results, show that monomers concentration oscillates around the steady state value c_1^∞ (see figure 2.13). Subsequently, the newly formed assembly wave propagates at fixed average speed as described in section 2.2.2 until all the populations reach their steady states. As c_1^∞ is an increasing function of Ω (equation (2.66)), the velocity at which the secondary waves evolves in the system follows the same tendency.

We have a qualitative understanding of the damping of the successive waves. Total input flux of monomer can be considered as partly directed to sustain the growth of aggregates

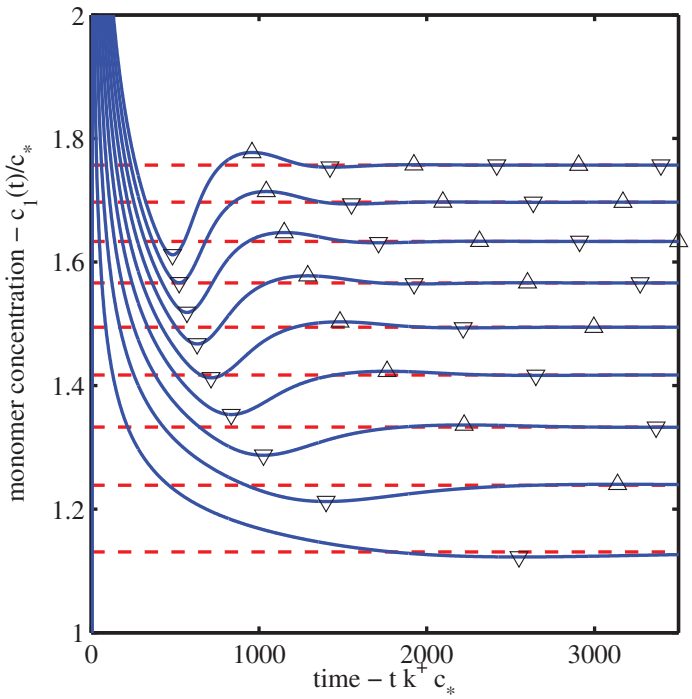


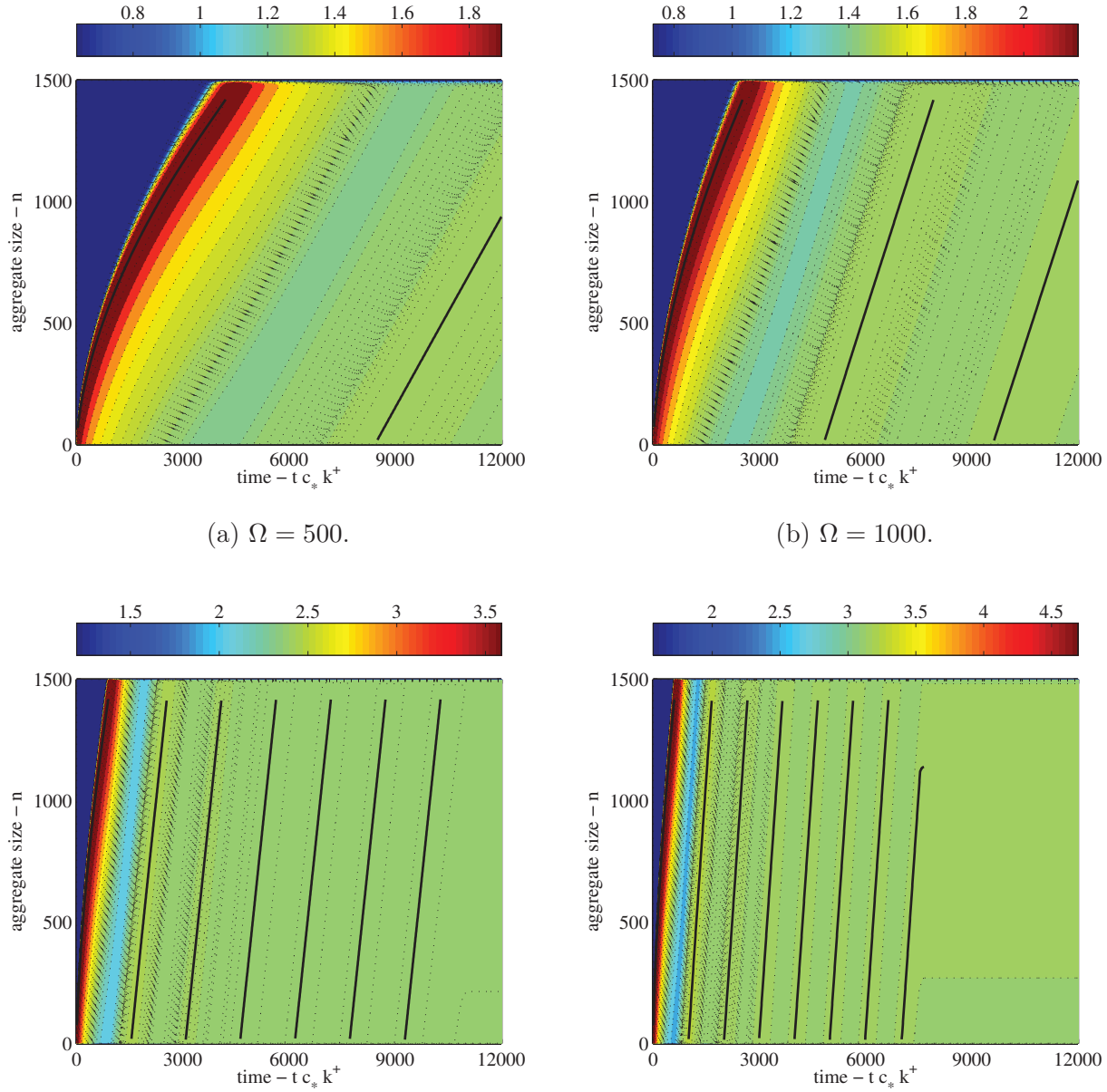
Figure 2.13 – Monomer concentrations in time for $\Omega \in \{74, 148, 222, 296, 369, 443, 517, 591, 665\}$ (from bottom to top – blue) and steady states values (dashed red). As the front reaches the absorbing boundary of the system at $N = 500$, formed virus stop consuming monomers which foster a second rise of the monomers concentration (minima indicated by $\triangledown$) until a new assembly wave is launched (maxima $\triangle$).

remaining in the system after the front wave exited. This results in a lower effective input flux to the monomer population. As the system progressively fills with more aggregates as diffusion spread successive waves, each overshoot of the monomer concentration is weaker than the previous one. With this in mind, dependency to the size and input flux can be forecast from basic considerations.

For instance, the greater the exit size N and the longer the tail of the size distribution, and thus the more remaining aggregates after the front reach the absorbing boundary. Furthermore, in long assembly line, the time to cover the distance at given Ω is longer and thus the front of the assembly wave is spread on a larger scale by diffusion. As it eventually reaches the end of the line, the number of completed virions that exits the system per unit of time is lowered. The assembly wave magnitude is therefore expected to be lower with increasing length N.

Stronger J_{on} leads to stronger overshoot, and then peaked waves. Also, the monomer concentration at a given time is constrained to higher values for higher J_{on}. With increased J_{on}, more peaked initial waves are moving at faster velocities throughout the system but face higher diffusion coefficient. As variation of diffusion coefficient and velocity in c_1 are the same, the diffusion length obtain for the typical time to go through the assembly line $l \propto \sqrt{\frac{D.N}{v}}$, should not change much. As a result, higher Ω should exhibit waves of stronger magnitude, consistent with what is observed on figure 2.13.

We observe furthermore that for higher Ω, first waves are initiated with higher frequencies



(a) $\Omega = 500$.

(b) $\Omega = 1000$.

(c) $\Omega = 5000$.

(d) $\Omega = 10\,000$.

Figure 2.14 – Long time size distribution evolutions for $N = 1500$ and $\Omega \in \{5.10^2, 10^3, 5.10^3, 10^4\}$. Isoconcentration curves (point lines) are drawn logarithmically spaced around the steady state monomer concentration c_1^∞ to emphasis small fluctuations and color scale is linear in concentration. Assembly wave trajectories are showed ($n(t) = \max_n c_n(t)$ – plain black line) and become straight lines once the velocities stabilize. The steady state can be considered reached on figure 2.14d at $tk^+c_* \simeq 8.10^3$ as concentrations are equal within calculation precision and maximum cannot be localized any more.

(figure 2.14). This second observation is consistent with a smaller travel time $T = \frac{N}{k^+c_1^\infty - k^-}$ (c_1^∞ being a increasing function of Ω). But it takes an extra time to form the following wave. This lag time is apparent on trajectories showed on figure 2.14 (black lines): the next trajectory is not initiated right after the previous ended but after an empty period that length for roughly the same duration and during which aggregates leave the system and simultaneously monomer concentration increases. From this observation, we expect the frequency to be of order $f = 1/2T$. In simulations, we found that the measured period was given by:

$$\mathrm{T}_{\triangle\triangle} = \frac{N}{k^+c_1^\infty - k^-}.\alpha(\Omega, N) \tag{2.81}$$

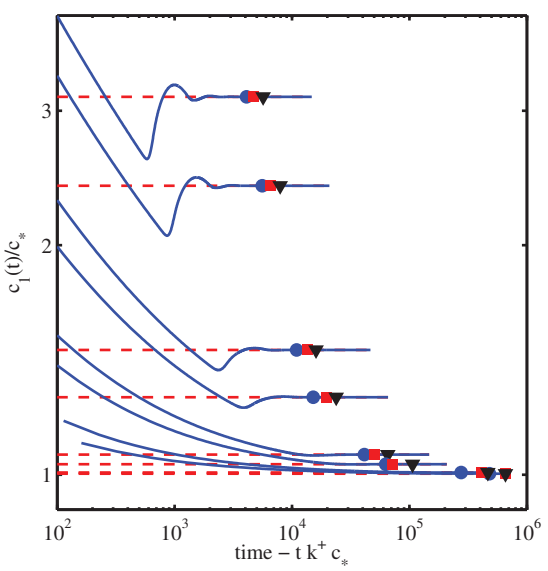
with α a coefficient with values of order 2 that decreases as Ω increases and increases with system size N. The values computed from the data showed on figures 2.13 and 2.15a are given in following table (empty value $\emptyset$ indicates that two consecutive extrema could not be found over the whole simulated time):

Ω ($N = 500$)	$\overline{\mathrm{T}_{\triangle\triangle}}k^+c_*$ ($\times 10^3$)	$\alpha = \overline{\mathrm{T}_{\triangle\triangle}}k^+c_*(c_1^\infty/c_* - 1)/N$ (± 0.05)
74	$\emptyset$	$\emptyset$
148	3.45	1,65
222	2.35	1,55
296	1.80	1,50
369	1.50	1,50
443	1.30	1,50
517	1.15	1,50
591	1.05	1,45
665	0.95	1,45
Ω ($N = 1500$)	$\overline{\mathrm{T}_{\triangle\triangle}}k^+c_*$ ($\times 10^3$)	$\alpha = \overline{\mathrm{T}_{\triangle\triangle}}k^+c_*(c_1^\infty/c_* - 1)/N$ (± 0.1)
5	$\emptyset$	$\emptyset$
10	$\emptyset$	$\emptyset$
50	98.7	2.1
100	43.6	1.8
500	8.8	1.6
1000	4.9	1.5
5000	1.5	1.4
10000	1.0	1.4

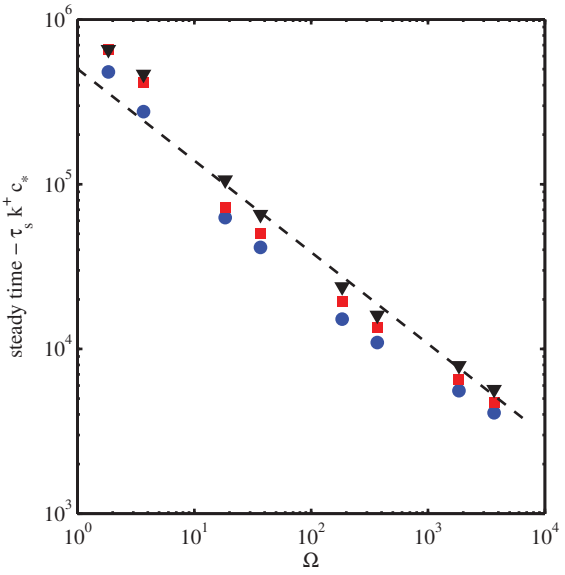
Time $\tau_s(\eta)$ to reach steady states as a function of dimensionless input flux Ω has been monitored on long time simulations for system characterized by a final size $N = 1500$. We defined it as the time we have to wait for the monomer concentration $c_1(t)/c_*$ to remain in a confidence interval of size 2η centered around its steady state concentration: $[c_1^\infty/c_* - \eta\,,\, c_1^\infty/c_* + \eta]$. Simulation results are shown on figures 2.14 and 2.15. Our simulations show that the time to reach steady state is shorter for stronger input flux Ω, and evolve approximately according to $1/\sqrt{\Omega}$.

2.2.4.7 Flux of virions released from the cell

As the system is opened, viruses eventually exit the cell. In the model presented here, there is no particular effect associated to this steps which is instantaneous. We showed in the section 2.2.3 that when a constant flux of monomer J_{on} is imposed at the entrance of the assembly



(a) log-log plot of the monomeric concentration evolution in a finite size system ($N = 1500$) for $\Omega \in \{5, 10, 50, 10^2, 5.10^2, 10^3, 5.10^3, 10^4\}$ (plain blue) and steady state concentrations numerically solved from equation (2.45) (red dashed).



(b) Log-log plot of the time $\tau_s(\eta)$ for the monomers to reach steady state concentration within η confidence bounds ($\eta \in \{10^{-5}(\circ), 10^{-6}(\square), 10^{-7}(\triangledown)\}$) depending on the intensity of the input flux (Ω). Steady states is reached faster for higher input flux. Black dashed line has slope $-1/2$.

Figure 2.15 – Long time evolution of the monomers concentration in a finite system depending on the input overcritical flux $\Omega > 1$. Time $\tau_s(\eta)$ to reach steady state value within η confidence bounds are indicated by symbols ($\eta \in \{10^{-5}(\circ), 10^{-6}(\square), 10^{-7}(\triangledown)\}$).

line, a steady state is eventually reached with a flux of virus given by the material balance between entry and exit of the system: $j_v = \frac{J_{on}}{N}$. In the following sections (section 2.2.4.2) we investigated the evolution of the growing virus inside the system and evaluated a few characteristics of the assembly line time scales in the simple scheme of constant rates for all sizes. We found that the dynamic towards the steady state is controlled by the ratio of incoming monomer flux to a critical flux $\Omega = \frac{J_{on}}{k^+ c_*^2}$ as well as the total number of components in an assembled virus N. Output flux monitored in simulation are displayed on figure 2.16. When the system is

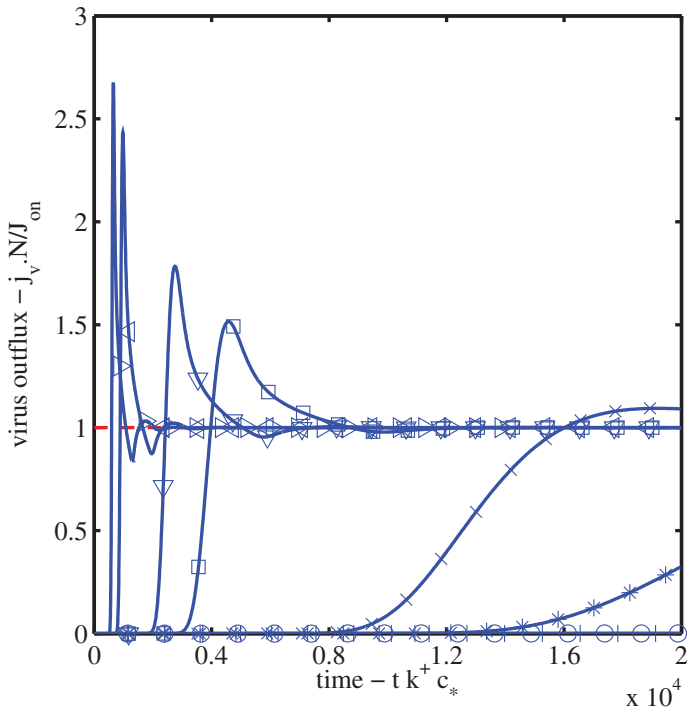


Figure 2.16 – Plot of the normalized flux of viruses exiting the system $\frac{Nj_v}{J_{on}}$ as J_{on} is turned on at $t = 0$ for $\Omega = 5(+), 10(\circ), 50(*), 10^2(\times), 5.10^2(\square), 10^3(\triangledown), 5.10^3(\triangleleft), 10^4(\triangleright)$ (from bottom to top) and $N = 1500$.

propagative, the shape of the virus flux is inherited from the shape of the monomeric evolution up to a broadening due to its diffusion part. As we noticed concerning the evolution of the monomers and intermediate species, the flux of virus overshoots the steady flux for $\Omega \gg 1$ but very rapidly comes back to stabilize after strongly damped oscillations around its steady state. Oscillations period is conserved and defined by the ratio of the size of the system to the steady state velocity (equation (2.81)).

The first viruses are observed after a lag time. This lag time accounts for the propagation time in the system from the time the input flux is switched on:

— When input flux is lower than the critical flux $\Omega \leq 1$, the system fills very slowly as it remains dominated by a negative drift. As the virus flux rises also at very low rate, lag time is not easy to define: since the derivative of the flux in time is low, a change in the arbitrary limit above which we consider the virus production to be noticeable implies important variation of the measured lag-time. For that reason we study instead the time for the outflux to reach half of its steady state value $j_v(\tau_0) = J_{on}/2$. This remarkable point

is easy to identify in all regime.

— When the input flux is overcritical $\Omega \gg 1$, the average velocity is positive and assembly is achieved by propagating waves. The lag time τ_0 is of the same order of the travel time T_1 of the first wave across the assembly line. Indeed travel time dominates the time to initiate the wave $\tau_i = (c_* k^+ \sqrt{2\Omega})^{-1}$ (see equation (2.75)). In small systems the wave velocity decreases rapidly during this early time as the first wave propagates, and we have not been able to solve the related integro-differential equation (2.71) for a proper calculation. For very large systems submitted to low input flux, however, this stiff deceleration phase can be neglected compared to the asymptotic regime, the lag time should obey the flux limited propagation equation (2.80):

$$\tau_0 \propto \frac{N^2}{\Omega c_* k^+} \tag{2.82}$$

We numerically studied the relationship between the size of the system, the input monomeric flux and τ_0. Figure 2.17 summarizes the observations.

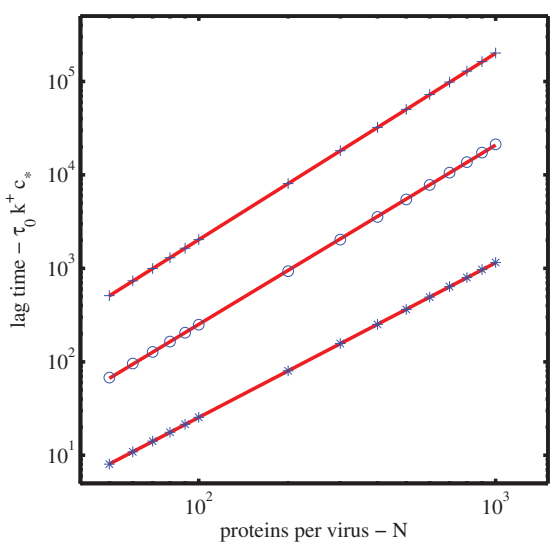
In both regimes, the lag time is an increasing function of the system size and strictly decreases with the input flux. This last remark underlines that the process is not subject to a kinetic trapping effect comparable to what Zlotnick[139] reported for closed system kinetics. Hence virus efficient strategies should always be oriented towards maximizing the production of monomers to ensure both faster assembly of the first viruses and steady state onset.

For input flux smaller than the critical one, $\Omega < 1$, the size of the system does not seem to affect the dependency of the waiting time in the input flux $\Omega^{-0.6}$ (figure 2.17b). Furthermore τ_0 appears to scale in L^2 for nearly critical regime ($\Omega = 0.5$ on figure 2.17a), a behavior similar to the asymptotic propagative regime but likely to have a very different physical source. Unlike the flux limiting scenario, the undercritical size distribution remains dominated by small sizes aggregates. The evolution is probably similar to a pure diffusive process (which have a similar scaling law) since even if the mean velocity in the system is negative, it decreases towards zero as $c_1(t)$ approaches c_* (equation (2.66): $c_1(t) < c_1^\infty \propto c_* \Omega^{1/N} \simeq c_*$ for $N \gg 1$).

For input flux larger than the critical one, $\Omega \gg 1$, the evolution is more complex as effect of size and input flux cannot be separated: a change in the size of the system affects the dependency in the input flux (figure 2.17b – table $\Omega > 10^2$). On this same figure, it is apparent that the linear behavior observed for under critical flux is not maintained at higher values. In fact, we never fully reached the asymptotic behavior in any of our simulations since it requires large system with low flux and results in very time consuming calculations. However this evolution can be foreseen in the data shown on figure 2.17a at fixed flux $\Omega = 20$ where the slope in log-log plot slightly diminish with respect to the total proteins numbers in a virus N, getting closer to 2 (linear adjustments on $N < 10^2$ and $N > 5\,10^2$ produces respectively slopes of 1.92 and 1.96). For increasing input flux Ω, the waiting time scales with a decreasing power in the size of the system (figure 2.17a, $\Omega = 1\,10^3$), further apart from the flux limiting regime as propagation time is more and more controlled by the early time regime.

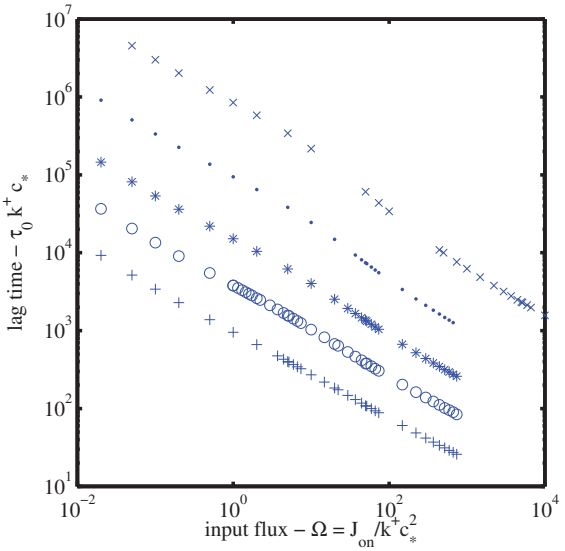
2.2.5 Line tension effect and fixed on-rate

In case of constant onward and backward aggregation rates, we showed that the steady state concentrations under a flux of monomer can be derived analytically in various regimes thanks to simplifications in equation (2.45) summations and following simplifications. We were also



(a) τ_0 as a function of the number of proteins in virus N for various input flux $\Omega \in \{0.5(+), 20(\circ), 10^3(*)\}$.

Ω	$\frac{\mathrm{d}\log\tau_0}{\mathrm{d}\log N}$
0.50	2.00
20	1.94
$1.0\,10^3$	1.66



(b) τ_0 as a function of the input flux Ω and for various number of proteins in virus $N \in \{50(+), 100(\circ), 200(*), 500(.), 1500(\times)\}$.

	$\frac{\mathrm{d}\log\tau_0}{\mathrm{d}\log\Omega}$	
N	$\Omega < 1$	$\Omega > 10^2$
50	-0.59	-0.53
100	-0.59	-0.54
200	-0.59	-0.59
500	-0.59	-0.65
1500	-0.57	-0.62

Figure 2.17 – Log-log plots of waiting time τ_0 for the virus flux to reach half of its steady state value after the input flux of monomer has been turned on (t=0). Slopes obtained from linear adjustment are given in the table below the plots

able to calculate the continuous limit of the equation and both the associated drift velocity and the diffusion coefficient. From those two quantities we have been able to derive various scaling limits.

But it is unlikely that, in a large proteins aggregate, the geometry of the assembly plays no role on the interactions between proteins and their neighbors that are described in the energy evolution. The next order in the model is to make the distinction between the mean energy of a protein with all its neighbors in the core of the aggregate, and a protein at the edge of the aggregate which interacts with significantly less other proteins, and cannot form as many bounds as the former.

Taking the line tension per capsomer at the periphery $\gamma \neq 0$, the internal free energy of the aggregates is given by equation (2.27):

$$f(n) = \epsilon_b.(n-1) + \gamma.\pi\sqrt{(n-1)\left(1-\frac{n}{N}\right)} \tag{2.83}$$

The resulting free energy formation of the aggregate from n proteins with chemical potential μ_1 is, with the same definition for the critical concentration: $c_* = e^{-\beta\epsilon_b}$:

$$\begin{aligned}\Delta f(n) &= \epsilon_b.(n-1) + \gamma.\pi\sqrt{(n-1)\left(1-\frac{n}{N}\right)} - n.\mu_1 \\ &= k_B T \ln(c_*/c_1).(n-1) + \gamma.\pi\sqrt{(n-1)\left(1-\frac{n}{N}\right)} - \mu_1\end{aligned} \tag{2.84}$$

Where we recast $\mu_1 = f(1) + k_B T \ln C_1$ from equation (2.49). Similar equation has already been introduced (see equation (2.22)) and plotted on figure 2.3. The aggregate formation free energy is non monotonic with maximum at position $n = n_b$. Barrier height and position of the maximum both diminish for increasing c_1/c_*.

Keeping k^+ constant for the purpose of simplicity, the backward rate k^- given by detailed balance is not constant any more as a nucleation barrier has been added the model. Noticing that equation (2.28) and equation (2.49) gives:

$$\begin{aligned}\frac{k^+ c_1}{k^-_{n+1}} &= e^{-\beta(f(n+1)-f(n)-f(1)+f(1)-\mu_1)} \\ &= e^{-\beta(\Delta f(n+1)-\Delta f(n))}\end{aligned} \tag{2.85}$$

We see that in the presence of the barrier, k^-_n is stronger than k^+c_1 for small values of $n < n_b$, it decreases to $k^+c_1 = k^-_{n_b}$ at the top of the barrier and then to much lower values as n increases further. Hence the drift velocity is negative before the barrier and positive afterwards.The barrier therefore fosters the splitting of small aggregates and opposes to the nucleation (it also restrains unbinding in larger aggregates). This is the key difference with the constant aggregation rate model, in which the sign of the drift is the same in the whole system.

2.2.5.1 Scaling still holds

Within the detailed balance framework, the scaling demonstrated in section 2.2.5 still holds when onward rates are not all equal but dependent on the aggregate size k^+_n. It remains valid as well when the formation free energy is not a pure linear function of the aggregates size, provided that only its linear part is modified. A change of the coefficient of the linear part in

the expression of the free energy of formation can be isolated as a simple multiplicative constant in the detailed balance between k_n^+ and the new backward rate $\widetilde{k_{n+1}^-}$:

$$\frac{k_n^+}{\widetilde{k_{n+1}^-}} = e^{-\beta(\epsilon_b+\delta\epsilon_b)} = c_* e^{-\beta\delta\epsilon_b} \tag{2.86}$$

As a matter of fact, each local flow J_n satisfies the scaling individually thanks to the detailed balance, so that the overall system of equation does –see equation (2.58):

$$\widetilde{J_n} = k_n^+(c_1 c_{n-1} - e^{-\beta\delta\epsilon_b} c_* c_n) \quad \text{for} \quad 1 < n < N \tag{2.87}$$

And as a consequence, a solution of the new system is equivalent to a solution of the previous, up to a scaling of times and concentrations, and corresponds to the previous flux times the scare of the scaling factor.

$$\begin{cases} t & \longrightarrow & t.e^{\beta\delta\epsilon_b} \\ c_n(t) & \longrightarrow & c_n(t.e^{\beta\delta\epsilon_b}).e^{-\beta\delta\epsilon_b} \\ J_{\text{on}} & \longrightarrow & J_{\text{on}}.e^{-2\beta\delta\epsilon_b} \end{cases} \tag{2.88}$$

In the steady state, that we study in the following section, time does not play any role. So a change of the free energy of bounding only alters the size distribution through a constant factor multiplication and the equivalent input flux is found multiplying by the square of the same constant. Two parameters are enough to describe the solutions: steady state capsomer concentration c_1^∞ and line tension per capsomer at the periphery γ.

2.2.5.2 Steady state

The steady state solution is given by equations (2.50) and (2.51). With a non zero line tension, the equations do not simplify as in the linear case. Assuming the barrier dominates the formation energy landscape, the sum in equation (2.50), can however be approximated using the Laplace's method (steepest descent):

$$\frac{J_{\text{on}}}{N} \simeq k^+ c_1^\infty e^{-\beta\Delta f(n_b)} \sqrt{-\frac{1}{2\pi}\left.\frac{\partial^2 \Delta f(n)}{\partial n^2}\right|_{n_b}} \tag{2.89}$$

consistently, $n_b \ll \frac{N}{2}$, altogether with $\left(\frac{\gamma\pi}{\sqrt{N}\Delta\mu}\right)^2 \ll 1$ and $\Delta\mu = -k_B T \ln(c_1^\infty/c_*)$, so equations (2.23) and (2.24) lead to the following using those approximations:

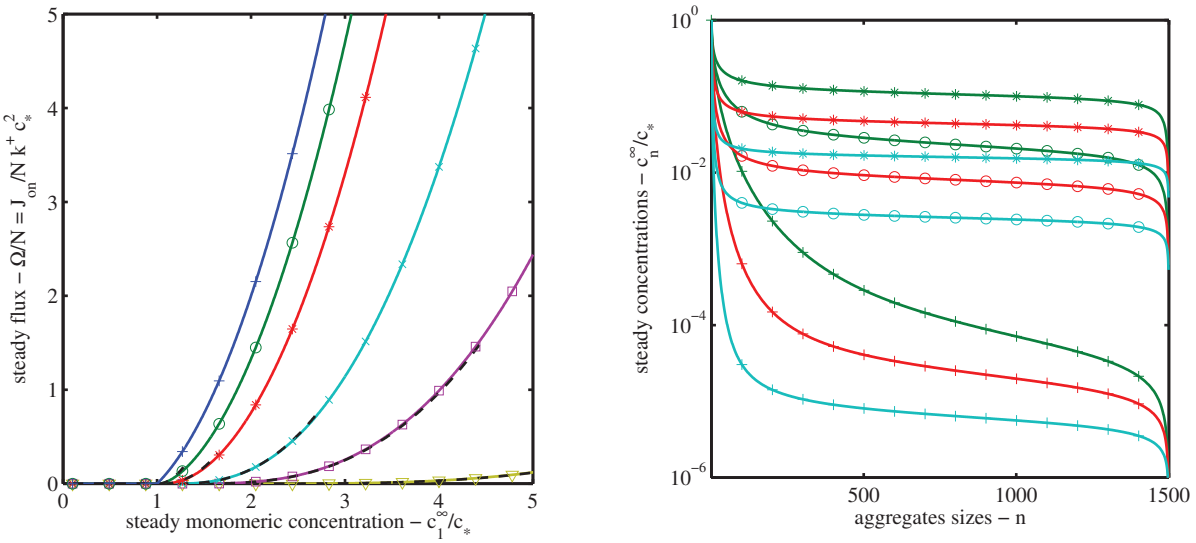
$$\begin{gathered} n_b \simeq 1 + \left(\frac{\pi\gamma}{2k_B T \ln(\frac{c_1^\infty}{c_*})}\right)^2, \qquad \Delta f(n_b) \simeq \frac{\pi^2\gamma^2}{4k_B T \ln(\frac{c_1^\infty}{c_*})} - \mu_1 \\ \left.\frac{\partial^2 \Delta f(n)}{\partial n^2}\right|_{n_b} \simeq -\frac{2\left(k_B T \ln(\frac{c_1^\infty}{c_*})\right)^3}{\pi^2\gamma^2} \end{gathered} \tag{2.90}$$

And thus,

$$\frac{J_{\text{on}}}{N} \simeq \frac{k^+ (c_1^\infty)^2 e^{-\beta^2 \frac{\pi^2\gamma^2}{4\ln(c_1^\infty/c_*)}} \left(k_B T \ln(\frac{c_1^\infty}{c_*})\right)^{\frac{3}{2}}}{\pi^{\frac{3}{2}}\gamma} \quad , \quad \frac{c_1^\infty}{c_*} \in]1, e^{\beta\frac{\pi\gamma}{2}}] \tag{2.91}$$

This last equation is valid as long as a pronounced barrier is present on a narrow interval of the system. This condition implies that the difference of the chemical potentials of bound and unbound proteins is such that it promotes binding $\Delta\mu < 0 \Leftrightarrow c_1^\infty \gg c_*$. In addition, if the input flux becomes too strong, free monomer concentration is high and so is their chemical potential. The line tension contribution is then very low compared to the free energy difference in binding a free monomer and the barrier eventually disappears. This is equivalent to the condition $n_b = 1$ in equation (2.90): $c_1^\infty < c_* e^{\pi\gamma/2k_BT}$. In such a scenario, the steady state regime tends to be the same as the linear free energy case described previously in section 2.2.4 and our approximation should also fail. This approximation of the steady flux in the system as function of the steady concentration works well on it validity domain as shown on figure 2.18a. Unfortunately, we have no analytical inverse expression of equation (2.91), to express $c_1^\infty(J_{\text{on}})$ in the narrow barrier regime.

Relation between monomer concentration and the flux imposed to the system computed from equation (2.50) is plotted on figure 2.18a and examples of the associated size distributions computed from equation (2.51) with a solution of the equation (2.50) found numerically are shown on figure 2.18b.



(a) relation between the steady state monomer concentration and the flux imposed in the system for different values of the line tension $\gamma \in \{0(+),\ 0.5(\circ),\ 1.0(*),\ 2.0(\times),\ 3.0(\square),\ 5.0(\triangledown)\}k_BT$. Dashed lines display relation 2.91 on its validity domain.

(b) Semi-log plot of the aggregates sizes distribution at steady state for different input flux $\Omega \in \{10^{-3}(+),\ 1.0(\circ),\ 10(*)\}$ and for various line tension values $\gamma \in \{0.5(\text{green}), 1.0(\text{red}), 2.0(\text{cyan})\}$.

Figure 2.18 – steady state flux and size distributions with line tension $\gamma \in \{0(\text{blue}),\ 0.5(\text{green}),\ 1.0(\text{red}),\ 2.0(\text{cyan}),\ 3.0(\text{purple}),\ 5.0(\text{yellow})\}k_BT$.

For a fixed line tension γ, the steady state regime establishes as follows. Capsomers flow into the system with the input flux J_{on}. If c_1 is low enough ($c_1^\infty < c_* e^{\pi\gamma/2k_BT}$), an energy barrier dominates the energy landscape. Only a portion of the capsomers are allowed to nucleate and pass the barrier, their flux decreasing exponentially with the barrier height (as in Kramers

theory). If the flux of capsomers in the forming capsids that passes the barrier is lower than the input flux, free capsomers accumulate in the system and the barrier is lowered. On the contrary, if the input flux is lower that the flux of capsomers in nucleated capsid, free capsomers concentration decreases so the barrier height increases. Steady state is reached once the two flux equilibrates, as it is stated by equation (2.91).

For increasing line tensions γ at a given concentration of capsomer c_1^∞, the formation free energy barrier is higher and so the steady flux Ω is lower as shown on figure 2.18a. Equivalently, at fixed input flux Ω, the free capsomer steady state concentration c_1^∞ is an increasing function of γ. Unlike the concentrations before the barrier, the concentrations of intermediates beyond the barrier diminish for increasing line tension and fixed input flux (figure 2.18b). Indeed, it requires lower intermediate concentrations c_n to maintain the same flux with higher c_1^∞ and thus with increased drift velocity. From an other point of view, as barrier height is maintained constant increasing both γ and c_1^∞, the slope of the energy landscape increases behind the barrier.

At fixed line tension, γ, the steady state concentrations c_n^∞ are all increasing with the input flux Ω.

In the narrow barrier regime, linear term is the dominant contribution in the free energy of formation and therefore the numerator in equation (2.51) can be approximated to a geometric series whereas numerator is proportional to the flux in the system:

$$\begin{aligned} c_n &\simeq e^{-\beta(n-1)\Delta\mu}\frac{J_{\text{on}}}{c_1^\infty N k^+}\sum_{q=n}^{N-1}\left(e^{-\beta\Delta\mu}\right)^{q-1} \\ &\simeq \frac{J_{\text{on}}}{Nk^+(c_1^\infty - c_*)}\left(1-\left(\frac{c_*}{c_1^\infty}\right)^{N-n}\right) \\ &\simeq \frac{c_1^\infty e^{-\beta^2\frac{\pi^2\gamma^2}{4\ln(c_1^\infty/c_*)}}\left(k_BT\ln(\frac{c_1^\infty}{c_*})\right)^{\frac{3}{2}}}{\pi^{\frac{3}{2}}\gamma}\left(1-\left(\frac{c_*}{c_1^\infty}\right)^{N-n}\right) \end{aligned} \tag{2.92}$$

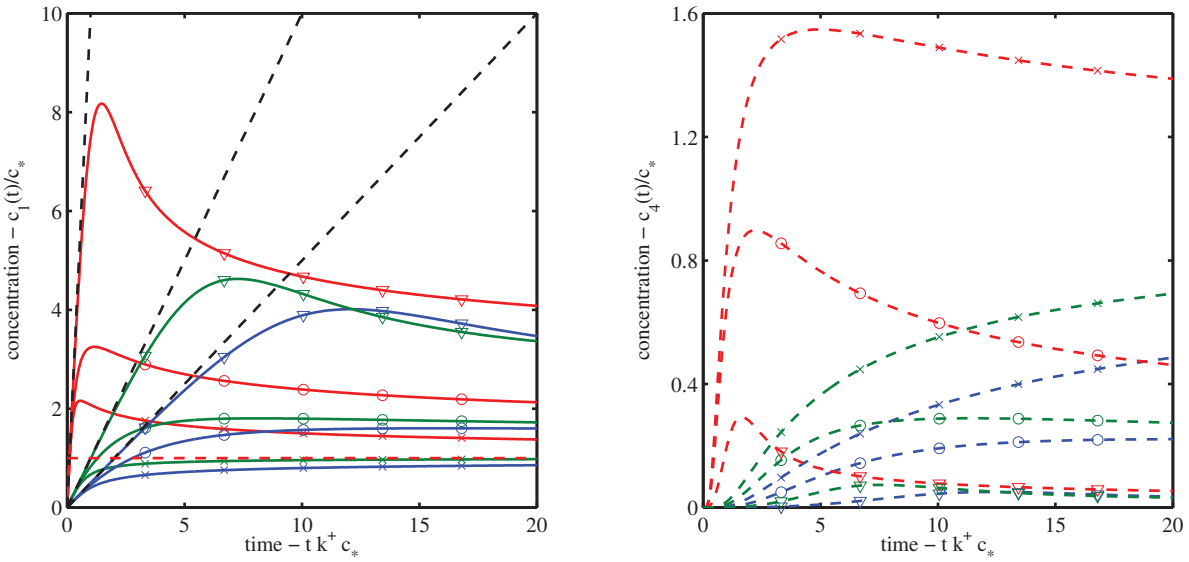
Which imposes a nearly constant concentration far from the system boundaries as $c_1^\infty > c_*$ (see figure 2.18b).

2.2.5.3 Dynamic

The dynamic of the line assembly model with a nucleation barrier is quite different from the dynamic with constant rates in the whole system. The drift velocity sign changes at the barrier crossing: negative drift before the barrier fosters the accumulation of monomer whereas positive drift beyond leads to faster assembly. The critical flux $\Omega = 1$ that we identified to separate low and high flux regime is no more of much interest. and regimes with $c_1^\infty < c_*$ are not found in practice. However, the onset of the dynamic as the flux is turned on follows a similar pattern at short times (figure 2.19).

At short time, the monomers accumulate to peak at a concentration higher than in the case $\gamma = 0$ consistently with a strong dissociation rate compare to the association rate figure 2.19a. The monomer concentration reaches its maximum value for the whole evolution. Despite a higher monomer concentration of monomers for increasing line tension at a fixed input flux Ω, the tetramers as well as greater sizes concentrations are lower due to the diminution of the

nucleation rate (curves of the same color are in reverse order in figures 2.19a and 2.19b). The evolution at short times is not described by equation (2.70) since most of the early nucleated dimers are immediately dissociated and shrink rather than grow. The return rate $k^- c_2(t)$ is likely to be of the same order of magnitude than $k^+ c_1^2(t)$ and cannot be neglected to produce a set of autonomous equations. Monomer concentration $c_1(t)$ increases until the barrier becomes low enough and the barrier position –which is the size on the assembly line where the drift velocity becomes positive– is close enough to the monomer size. Meanwhile, low sizes are progressively populated as the barrier height also lowers ($k_n^+ c_1$ increases compared to k_n^- in the whole system). As the barrier eventually reaches the neighborhood of the populated sizes, a pool of nucleated aggregates can pass in the part of the assembly line where the mean velocity is positive where they start to grow.



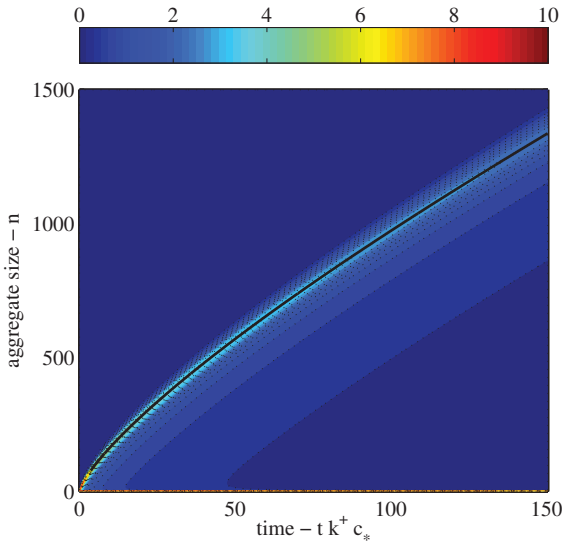
(a) Monomers concentration $c_1(t)$ compared to $\Omega.tk^+c_*$ (dashed lines).

(b) Tetramers concentration $c_4(t)$.

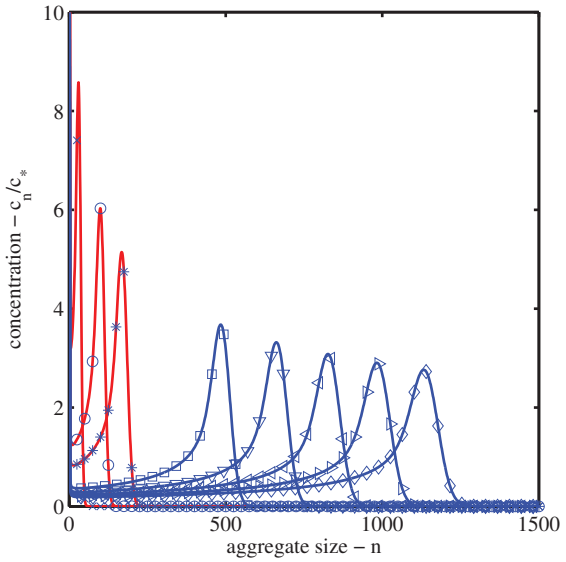
Figure 2.19 – short times: monomer and tetramers concentrations for $\gamma \in \{0.5(\times), 1(\circ), 5(\triangledown)\}k_BT$ and $\Omega \in \{0.5(blue), 1(green), 10(red)\}$

In the constant rate case the drift velocity remains positive in the whole system once this regime is reached ($c_1 > c_*$) and up to steady state. Otherwise the growth is stopped (as $c_1 \to c_*$) until more monomers are supplied. As long as the growing aggregate front propagates in the system, small aggregates are nucleated and the intermediate time distribution of sizes shows a propagative front with a long tail with a very slow decay (see figure 2.11b). On the opposite, when $\gamma \neq 0$ the monomer consumption of the growing aggregate and the consequent diminution of the monomer population impacts the drift velocity. As the barrier is displaced towards greater sizes, drift velocity turns back to negative values at the entrance of the assembly line. It follows that the nucleation of new aggregates is restricted as long as growing aggregates consumes monomers. As a consequence, the tail at the back of the propagating front decreases rapidly and the growth velocity is much higher than in the situation without line tension shown on figure 2.20.

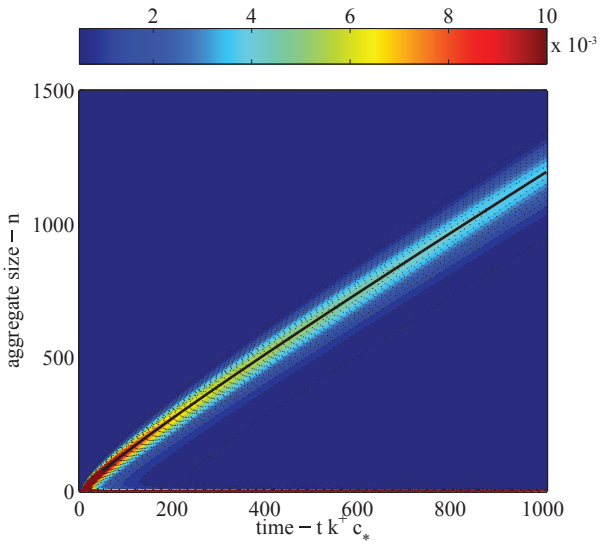
In very large systems, far from the boundaries, the linear part of $f(n)$ dominates. Hence, the



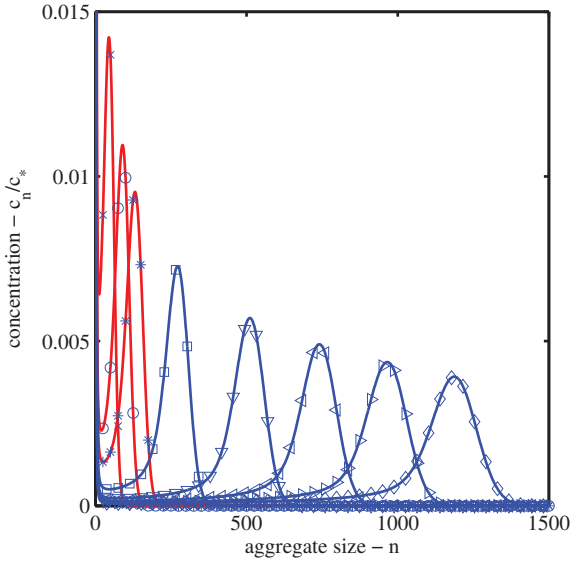
(a) Time evolution of the size distribution in the system: colors indicate concentrations $\frac{c_n}{c_*}$ for each aggregate sizes in time. The wave trajectory ($n(t) = \max_n c_n(t)$ – plain black line) shows a short deceleration phase followed by a stable speed.



(b) The propagating wave: Snapshots of the sizes distribution aside at various times ($t_i k^+ c_* \in \{1(\times),\ 5(\circ),\ 10(*)\}$ – plain red) and ($t_i k^+ c_* \in \{40(\square),\ 60(\triangledown),\ 80(\triangleleft),\ 100(\triangleright),120(\Diamond)\}$ plain blue).



(c) Time evolution of the size distribution in the system: colors indicate concentrations $\frac{c_n}{c_*}$ for each aggregate sizes in time. The wave trajectory ($n(t) = \max_n c_n(t)$ – plain black line) shows a stable speed.



(d) The propagating wave: Snapshots of the sizes distribution aside at various times ($t_i k^+ c_* \in \{30(\times),\ 60(\circ),\ 90(*)\}$ – plain red) and ($t_i k^+ c_* \in \{200(\square),\ 400(\triangledown),\ 600(\triangleleft),\ 800(\triangleright),1000(\Diamond)\}$ plain blue).

Figure 2.20 – Assembly wave propagating in the system $\Omega = 5\,10^3$ (top) and $\Omega = 0.9999$ (bottom), $\gamma = 5$ and $N = 1500$.

average drift velocity and diffusion are constant over the sizes and change only with monomeric concentrations. The propagation of a single peak rather than a front with a long tail changes the front asymptotic velocity in the flux limited regime. The number of monomers required to form the peak of width Δn and height c_p at size $n_{\max}$ (neglecting broadening) is:

$$\sum_{n=n_{\max}(t)-\Delta n}^{n_{\max}(t)} n.c_p \sim n_{\max}.\Delta n \tag{2.93}$$

If the monomer concentration is stable –assumption supported by the strong barrier that makes nucleation events extremely rare– the input flux sustains the progression of the pool of forming capsids:

$$n_{\max}(t) \sim \frac{J_{\text{on}}.t}{\Delta n} \tag{2.94}$$

Which implies a progression of the size of the growing aggregates linear with time and not slowing down like in the case previously studied without line tension.

2.2.5.4 Long time in finite size systems: oscillations

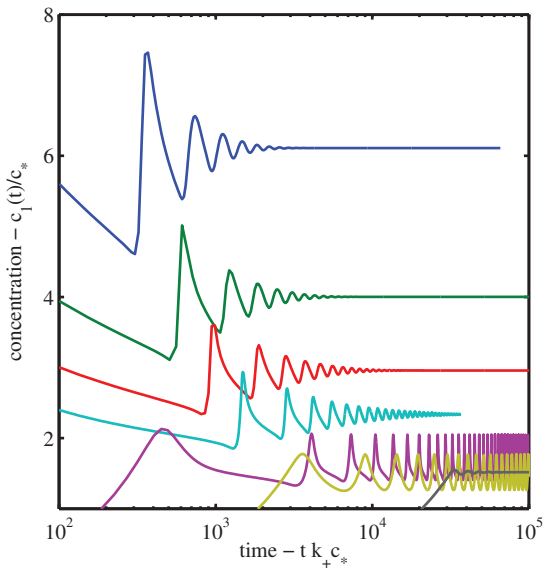


Figure 2.21 – Log-linear plot of the evolution in time of the monomeric concentration for different input flux $\Omega \in 5 \times \{\, 10^2$(blue), 10^1(green), 1(red), 10^{-1}(cyan), 10^{-2}(purple), 10^{-4}(yellow), 10^{-5}(black)$\}$. Starting from high Ω, damping decreases to 0 (oscillations with constant amplitudes –purple and yellow) and the steady state is not reached for $\Omega \in [\,10^{-2}\, 10^{-4}]$. It is only at very low flux that the damping is present once again ($\Omega = 5\, 10^{-5}$–black).

When the pool of growing capsid reaches the system boundary at $n = N$ they escape. This leads to oscillations of the concentration with the mechanism that we described in the case of constant rate and overcritical regime $\Omega > 1$. In presence of line tension, this phenomenon is intensified as the shape of the propagating wave is much more peaked. As shown on figure 2.21, many periods of oscillations are visible for any input flux Ω.

Unlike in the linear cases where oscillations are apparent only for very strong overshots consecutive to very high input flux, we observe that in the presence of line tension oscillations are visible even at low input flux. Furthermore, the damping of those oscillations is not monotonic with the input flux. There is indeed a range of input flux for which sustained oscillations are observed and steady state is never reached. The mechanism behind this "resonance" in the system is not elucidated. For input flux greater than the resonance range, the damping increases with the input flux Ω. It appears that greater flux leads indeed to heavier tails at the back of the propagating wave (compare for instance on figures 2.20b and 2.20d), which might be an important element to explain the increase of the damping with Ω.

The frequency of the oscillation is also an increasing function of Ω similarly with the linear case.

2.2.6 Estimations of the numerical values

In this section we have so-far used mainly dimensionless quantities and studied the behavior of the system in as much generality as possible. In this section we estimate the effective values of the parameters from those reported by the literature:

— s_0, mean surface per protein at close packing in the gag clusters. We estimate this value from the number of gag $N = 2500$ proteins Ganser-Pornillos et al.[50] covering 2/3 of the closed immature particle of radius $R = 70$ nm:

$$s_0 = \frac{\pi R^2}{N} \approx 16 \pm 5\,\mathrm{nm}^2 \tag{2.95}$$

— ϵ_b, average binding free energy of a free protein to the gag cluster. Protein-protein interaction must be stable enough for the structure to hold against thermal agitation but weak enough to allow for reorganization in order to find the most stable structure. This advocate for a protein-protein bound of the order of $k_B T$ ($1.3\,k_\mathrm{B}\,T$ is used in Kremer et al.[79] for instance) and therefore, as several neighbors are binded simultaneously, $\epsilon_b \simeq 5 - 10\,k_\mathrm{B}\,T$.

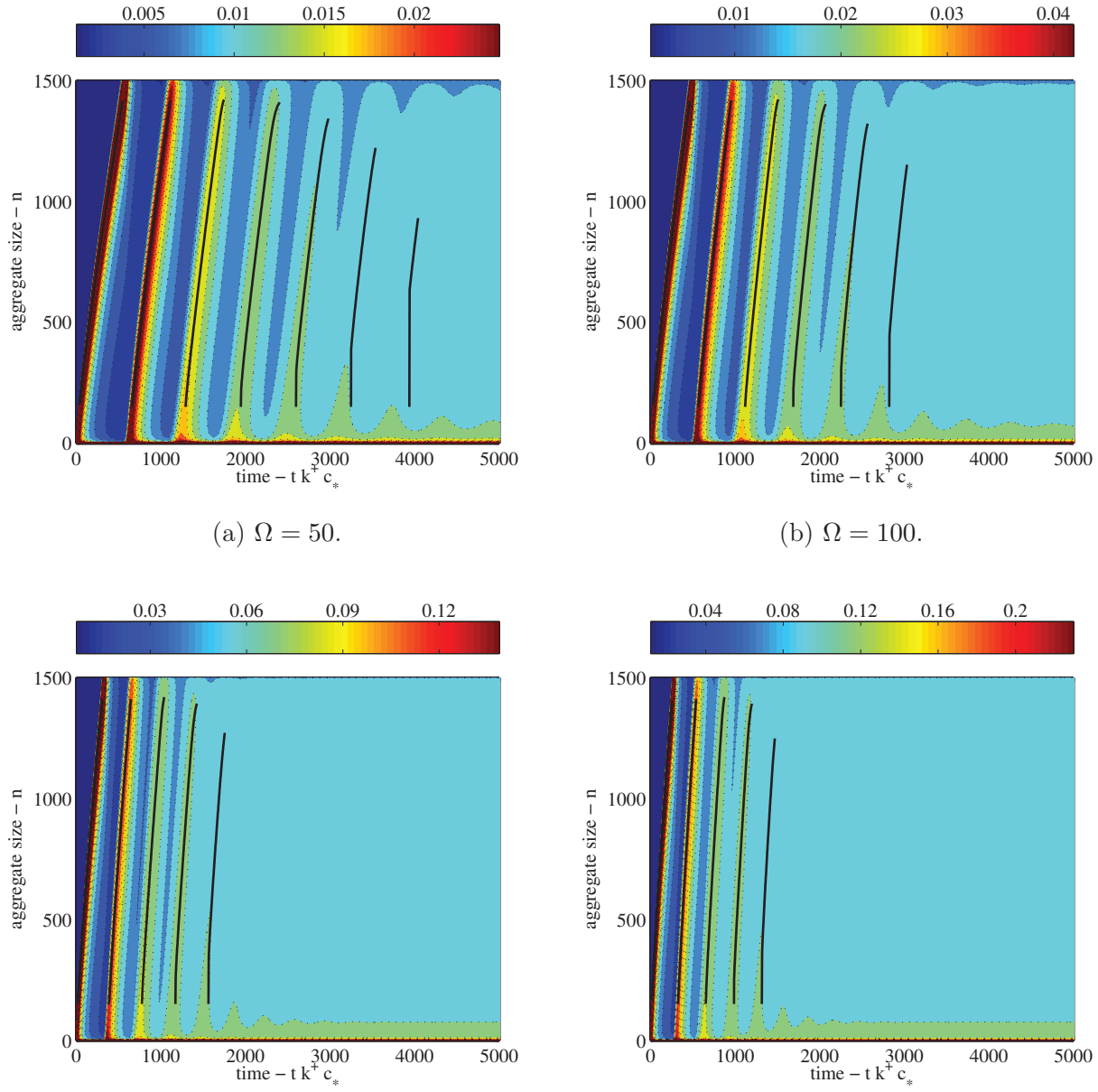
— γ, line tension per monomer : mean energetic loss at the gag cluster periphery due to unsatisfied bounds with missing neighbors. This is a fraction of ϵ_b (roughly $\frac{\epsilon_b}{2} \simeq 5\,k_\mathrm{B}\,T$).

— k^+, the diffusion limited clustering rate: the diffusion of gag clusters on the membrane has been measured by Ivanchenko et al.[69] during virus growth: $D_a = 1.8 \times 10^{-5}\,\mu\mathrm{m}^2 \cdot \mathrm{s}^{-1}$, whereas measurement have been performed on free gag proteins at the membrane by Manley et al.[89]: $D_\mathrm{gag} \lesssim 0.2\,\mu\mathrm{m}^2 \cdot \mathrm{s}^{-1}$ on short time scale (single particle tracking using PALM), coherent with the values given by Kenworthy[77] for a single protein bound to the cellular membrane $D_{\text{memb. prot}} = 0.1 - 0.5\,\mu\mathrm{m}^2 \cdot \mathrm{s}^{-1}$ .This allows to calculate an order of magnitude for the diffusion limited rate of gag aggregation to the clusters according to a 2D Schmoluchovsky theory[128] (using the nondimensionalization factor s_0 used for all the concentrations) :

$$k^+ \simeq 4\pi D_\mathrm{gag} n_{\text{binding sites}}/s_0 \simeq 10^6\,\mathrm{s}^{-1} \tag{2.96}$$

Where the order of magnitude of number of binding sites is estimated as $n_{\text{binding sites}} \simeq 10^2$.

— c_* the dimensionless critical concentration: since the dynamics studied in the previous section all exhibit concentration of free proteins of the same order of magnitude that c_*, we use $c_* \simeq \frac{s_0 N_{\text{free gag}}}{S}$. Therefore, c_* can be estimated by measuring the free units



(a) $\Omega = 50$.

(b) $\Omega = 100$.

(c) $\Omega = 500$.

(d) $\Omega = 1000$.

Figure 2.22 – Long time size distribution evolutions for $N = 1500$, $\gamma = 5k_BT$ and $\Omega \in \{50, 10^2, 5\,10^2, 10^3\}$. Isoconcentration curves (point lines) are linearly spaced around the steady state monomer concentration c_1^∞ and color scale is linear in concentration. Maximum of the size distribution on the central part of the assembly line are shown to emphasis assembly wave trajectories ($n(t) = \max_{5\%.N<n<95\%.N} c_n(t)$ – plain black line).

concentration on the cell membrane. A rough estimate is obtained by counting the number of isolated gag proteins localized in the superresolution image in Gunzenhäuser et al.[59] and give two free gag per square micrometer and $c_* \simeq 10^{-5}$ and is coherent with a corresponding value of $\epsilon_b \simeq 10\, k_\mathrm{B}\, T$.

— The order of magnitude for the time needed to build a full virus can be estimated from the average speed as:

$$\tau \simeq \frac{N}{k^+ c_*} \simeq 10^2\,\mathrm{s} \tag{2.97}$$

Consistent with the average time 10-20 min reported between first detection and saturation of the fluorescence signal in TIRF microscopy experiments (Ivanchenko et al.[69], Jouvenet et al.[73]).

— Ω, the dimensionless input flux: the critical flux $J_\mathrm{crit} = k^+ c_*^2 \simeq 10^{-4}\,\mathrm{s}^{-1}$ compared to the actual flux of proteins produced by the cell. Estimate of the flux of virus exiting the cell are not easy to find. De Boer et al.[34] reported that a single cell produce in the order of 10^4 a day. Output flux per surface unit is estimated using a sphere of diameter 10 µm for the typical sphere, leading to typical values of $1\,\mathrm{gag}\cdot\mathrm{s}^{-1}\cdot\mu\mathrm{m}^{-2}$ exiting the cell in formed virion and corresponding to flux $J_\mathrm{on} \simeq 10^{-5}\,\mathrm{s}^{-1}$ and $\Omega \simeq 10^{-1}$ in simulations.

2.2.7 Conclusion

In this section we studied the solutions of the line assembly model in an open system. According to the model, each bud grow by sequential additions of one subunit at a time. We made the simple assumption that the aggregation rate is proportional to the subunit concentration and equal for all the different aggregates and that the input flux of subunits was constant. The detailed balance imposes the backward rate of a reversible subunit addition from its onward rate and the work needed to add a new subunit to the growing capsid. We studied in details the scenario where this work was assumed independent of growing capsid size (linear case) and when a nucleation barrier was present in the system.

The steady state is defined by two coupled equations between the growing capsid flux in the system and the subunits concentrations equations (2.45) and (2.46). However it is not possible to find an exact solution of the system apart for the linear case scenario. In any case the size distribution when a constant flux flow throughout the system does not resemble the equilibrium distribution of the closed system. The most striking difference is that in all cases intermediates are as likely as full capsid to be observed on the membrane as those latter do not accumulate. The comparison between equilibrium distributions and steady state distribution when a flux is maintained throughout the system has already been published (Castelnovo, Verdier, and Foret[25]) and will be completed by a second on the transient dynamic.

In the linear case, two very different steady state distribution exist, depending on the input flux. This model of pure elongation is characterized by a critical input flux $J_\mathrm{on} = k^+ c_*^2$. The growth of virus is very unlikely below this flux that sustains the critical concentration of subunits leading to aggregation and the production of virions is minimal, but since steady state requires balance of input and output flux, virions are anyhow produced. This capsid production is sustained by aggregates size fluctuations comparable to capsid size balancing a situation were capsid shrinkage is favored. This flux becomes thus vanishingly small with the total number of proteins per complete virus, as the amplitude of the fluctuations needed to reach completion increases. So, it should not be observed in experimental conditions where

this number are typically large. The mean field description (rate equation) used here is also probably not adapted to describe these situations of small proteins number where fluctuations are expected to dominates the dynamic... The critical flux also controls the dynamic of the system:

— Below critical flux, the steady state distribution is reached very slowly. The subunits concentration is kept under the critical concentration and each size is progressively populated in the system until the whole system is filled.

— Above the critical flux, monomers accumulates and oscillations appears. It means that the cell releases virions in bursts in the early production time. As the relaxation to steady state is extremely fast, only one burst is noticeable in practice, sharper at higher input flux and for smaller complete capsid size.

In contrast with the closed system case, there is no kinetic trap in such dynamic as subunits are provided continuously, and the stronger the flux, the larger the output of virions. Steady state distribution is roughly constant over all the intermediate sizes for reasonably large number of subunits in the complete capsid. This linear case, however, is likely to be a too simplistic proposal to describe the aggregation of the capsid assembly since, to our knowledge, assemblies experiments all reports nucleation barrier effects.

The presence of a nucleation barrier is already a more physical situation. The barrier height is controlled by the subunits concentration. Subunits first accumulate irrespectively of the input flux until the barrier is low enough to be crossed. In presence of a nucleation barrier there is no critical input flux value anymore. The size distribution at a given time is bimodal, with either small aggregates and forming capsid with similar sizes. Completing capsids grow at constant speed in contrast with the linear case where the propagation slows down in time. The lag time between the onset of subunits production and the first formed capsid is then reduced. The nucleation barrier dramatically fosters the burst dynamic: many burst can be noticed and simulation even indicate a range of flux where the production does not relax to steady state.

This description is a macroscopic average over all the growing capsids evolution in the limit of a large system. At very low input flux, the fluctuations might be important compared to the averaged behavior. Also very large flux generate a strong free subunits concentration that might be incompatible with the single step mechanism proposed here.

We have not studied the membrane mechanics in the model at this point. Our model rely on the assumption that the protein coat imposes its preferred curvature to the surrounding membrane, and that this curvature is independent of the bud size. The curvature of the membrane contributes in the energy by a term linear with respect to the bud protein content (equation (2.26)) and can be absorbed in the average binding energy term. In contrast membrane tension generates a term that goes as the square of the aggregate size. Calculation from Deserno[36], Foret[46] show that the energy profile including this fluid membrane contribution also exhibit a barrier at high completion values which could give interesting behaviors such as kinetically trapped particles (metastable completion states) before the full completion is reached.

The released mechanism itself is not explicit in our model. We assume that there is an irreversible step at the end of the assembly process and that fully formed capsid do not anymore interact with the subunits. Human Immunodeficiency virus released mechanism is reported to rely on the recruitment of proteins able to induce scission by polymerization at the bud neck (Guizetti and Gerlich[57], Lenz et al.[84], Van Engelenburg et al.[130]). In such case, the bud is not anymore accessible to free gag proteins and our model is consistent. However a recruitment

rate may affect the dynamic. Furthermore, the virion are not immediately released and the observed size distribution on the membrane will show an accumulation of complete capsids.

Eventually, the size of the cluster could be taken into account in the onward aggregation rate to refine the model. Constant rate are maybe too simple to model the accretion on a ring in 2D. Possible refinements also include the effects of membrane curvature in the vicinity of the bud which might alter protein diffusion and mediate specific interactions (Matthews and Likos[92], Müller and Deserno[101]).

In the next section we look at a complementary approach focused on single growing capsid evolution at the membrane rather than measuring the size distribution of all the growing capsid at the membrane, which is still a complex task.

2.3 Model of single aggregate trajectories

In the previous section we studied the collective dynamic of the clusters using the framework of the rate equation. We studied the average evolution of the cluster populations in time disregarding their fluctuations. However, recent experiments managed to probe the time evolution of a single aggregate on the membrane inside the bath of capsomers.

The growth in time generate a trajectory in the sizes space which is inherently of stochastic nature. We are interested by the information it carries about the detail of the microscopic mechanism. At the population level the detail of each reaction event are averaged and it is difficult to "work backward" and extract the rates from the population evolution if we do not monitor the whole population evolution in time. Access to this information is not straightforward either in the single since one has to separate the random contribution of each single event from the conserved part that describes the physics of the accretion. The purpose of our work is to set on a simple model of the aggregation, comparable to chemical reaction, at the level of the single aggregate and offer a simple method to extract the model parameters from experimental data.

2.3.1 Generation of time trajectories on a modified Gillespie scheme

2.3.1.1 The direct Gillespie scheme

We consider the chemical composition in a vessel. The chemical composition is simply the number of molecule of each species present in the vessel. If chemical reaction are possible in the vessel, the chemical composition generally changes in time. In the course of its evolution, reactants transforms into products at random times depending on the probability that they collide and on the success rate of the collisions to lead to products. The evolution of the chemical composition in the vessel is random and will be different from one experiment to an other.

Let us consider that the vessel is filled with only these three species:

— aggregates of size 1 (monomers), noted (1)

— aggregates of size n, noted (n)

— aggregates of size $n+1$, noted $(n+1)$

And that the species can undergo only a single reversible aggregation step:

$$(n) + (1) \rightleftharpoons (n+1) \tag{2.98}$$

Our fundamental hypothesis in what follows, is that all the species are maintained at thermal equilibrium. Thus, on one hand the spatial probability distribution to find an aggregates is homogeneous on the surface, and on the other hand the velocity distribution is stable in time. Then, the probability for any given pair of one monomer and one (n)-aggregate to meet per unit time exists. The reaction occurrence is simply a successful collision between a pair. If we note N_1 and N_n the number of monomers and (n)-aggregate in the vessel at the time, reaction can then be properly characterized by its reaction probability that any pair of monomer and (n)-aggregate merges anywhere on the surface during the infinitesimal time $\mathrm{d}\tau$ (Gillespie[52]):

$$A(T, \rightharpoonup)\,\mathrm{d}\tau = N_1.N_n.\alpha(T, \rightharpoonup)\,\mathrm{d}\tau \tag{2.99}$$

This is the probability that any pair of (1) and $(n+1)$ aggregates merge during $\mathrm{d}\tau$. The factor $\alpha(T, \rightharpoonup)$ summarizes the thermal properties of the mixture on the membrane (the distribution of aggregates velocities) and the physical properties of the aggregates pair that affects the reaction "$\rightharpoonup$" (their probability of success per collision). The reverse reaction is also characterized by a different probability per unit time (as no collision is involved, it is describe as a thermally driven random event):

$$A(T, \leftharpoondown)\,\mathrm{d}\tau = N_{n+1}.\alpha(T, \leftharpoondown)\,\mathrm{d}\tau \tag{2.100}$$

In the following we will omit to remind the temperature dependence of our model as we consider it is fixed.

The probability density $P_\emptyset(\Delta t)$ that none of those two reaction has occurred during the time Δt can be calculated arguing that the reactions are independent events. The probability that any of those two reaction occurs is thus the sum of their probability, $N_1.N_n.\alpha(T, \rightharpoonup) + N_{n+1}.\alpha(T, \leftharpoondown)\,\mathrm{d}\tau$, hence:

$$\begin{aligned} P_\emptyset(\Delta t + \mathrm{d}\tau) &= (1 - (A_\rightharpoonup \mathrm{d}\tau + A_\leftharpoondown \mathrm{d}\tau))\, P_\emptyset(\Delta t) \\ P_\emptyset(\Delta t) &= (A_\rightharpoonup + A_\leftharpoondown)\, e^{-(A_\rightharpoonup + A_\leftharpoondown)\Delta t} \end{aligned} \tag{2.101}$$

The probability that reaction $(\rightharpoonup)$ occurs between Δt and $\Delta t + \mathrm{d}\tau$ has to be proportional to the probability given in equation (2.99):

$$P(\Delta t, \rightharpoonup)\,\mathrm{d}\tau \propto P_\emptyset(\Delta t).A_\rightharpoonup\,\mathrm{d}\tau \tag{2.102}$$

And the same is true for the reverse equation. Eventually the normalized join density probability to observe a given reaction occurring after Δt writes up:

$$\begin{aligned} P(\Delta t, \rightleftharpoons) &= A_\rightleftharpoons \times e^{-(A_\rightharpoonup + A_\leftharpoondown).\Delta t} \\ &= P_\rightleftharpoons \times P_\emptyset(\Delta t) \end{aligned} \tag{2.103}$$

Where we see that waiting time and reaction to occur afterwards can be chosen independently. The later being drawn according to the probability density:

$$P(\rightleftharpoons) = \frac{A_\rightleftharpoons}{A_\rightharpoonup + A_\leftharpoondown} \tag{2.104}$$

It is then possible to simulate successive reactions by starting from a known composition $\{N_1, N_n, N_{n+1}\}$ with probability of reactions per unit time $\{\alpha_\rightleftharpoons\}$ and iterate the following algorithm called the "direct Gillespie method":

- randomly choosing a time Δt for the next reaction to occur according to equation (2.101).
- randomly choosing the next reaction to occur according to equation (2.104).
- Update the time $t = t + \Delta t$ and the composition of the membrane at this time following the chosen reaction $\{N_1, N_n, N_{n+1}\} \rightharpoonup \{N_1 - 1, N_n - 1, N_{n+1} + 1\}$ or $\{N_1, N_n, N_{n+1}\} \leftharpoondown \{N_1 + 1, N_n + 1, N_{n+1} - 1\}$
- calculate the new probability of reaction per unit time $\{\alpha_\rightleftharpoons\}$ according to the new composition and update the density of probability accordingly for the next sampling.

This method naturally generalizes to any set of reactions involving the collision of at most [1] two "molecule" of the environment with reaction probability per unit time $\{A_r\}$ and the following probability density (resp. distribution) for the waiting time (resp. reaction to occur):

$$P_{\emptyset}(\Delta t) = \left(\sum_q A_q\right) e^{-\left(\sum_q A_q\right)\Delta t} \tag{2.105}$$

$$P(r) = \frac{A_r}{\sum_q A_q} \tag{2.106}$$

The advantage of the Gillespie method is to provide correct trajectories of the reaction as solutions of a Markov process in the condition of thermal equilibrium for all the components (equilibrium condition such as homogeneity of material and temperature in space are assumed independent of the chemical ongoing process). Unlike an equivalent coarse grain molecular dynamic simulation, it does not take into account the spatial position, orientations or speed of the component but rely on stochastic average provided in the effective probability of reaction per unit time. This is thus computationally much more efficient. It is however not a correct design to simulate very fast reaction that would rapidly induce spatial inhomogeneities and may not satisfy the requirement of thermal equilibration between each step. As it simulates one reaction at each step, it is obviously not designed to follow the evolution of vast amount of reactants in a long time which is left to kinetics rate equation in the thermodynamic limit where fluctuations are negligible. Last but not least, the trajectories generated here are those of the evolution of the broad chemical composition of a solution of indistinguishable particles in the vessel, and not trajectories of each single particle.

2.3.1.2 Link between the kinetics rate equations and Gillespie reaction per unit time

Gillespie[51] and Van Kampen[131] give the relation between the broadly used kinetics rates k_r of the "rate equations" proposed by Van't Hoff which describe the average evolution of the concentrations on an homogeneous medium and the probability of reaction per unit time α_r introduced in the previous section. Average is understood at a given time over an assembly of system all prepared in the same state at $t = 0$. k_r conventionally defined such that the average number of reaction per unit time and unit area is:

$$k_r \left\langle \frac{N_i}{S} \right\rangle \left\langle \frac{N_j}{S} \right\rangle \tag{2.107}$$

Whereas in the microscopic description it is given by:

$$\frac{1}{S}\langle \alpha_r N_i N_j \rangle \tag{2.108}$$

The Equivalence of this two definitions leads to:

$$k_r = S\alpha_r \frac{\langle N_i N_j \rangle}{\langle N_i \rangle \langle N_j \rangle} = S\alpha_r \left(1 - \frac{\mathrm{cov}(N_i, N_j)}{\langle N_i \rangle \langle N_j \rangle}\right) \tag{2.109}$$

1. reactions between more than two molecules at a time can be however included as composed of several two intermediates sub-reactions

The logic behind the rate equation and the use of equation (2.107) is that the fluctuations of the concentrations will be negligible compared to their mean values for a sufficiently large number of reacting molecules in the system apart for few trajectories with vanishing probability in the average. Hence the correlations (fluctuations) between the concentrations of each reactant being neglected, $\langle N_i N_j \rangle \simeq \langle N_i \rangle \langle N_j \rangle$, and those two rates are proportional up to the area of the system S:

$$k_r \simeq S.\alpha_r \tag{2.110}$$

as long as it makes sense to write down the rate equations.

2.3.1.3 Adaptation to single particle trajectories in a stationary monomers bath

Now we consider a singled out (n)-aggregate on the whole surface. Under the same hypothesis that formulated for the Gillespie model, the probability per unit time that our aggregate is involved in an occurring reaction knowing that one (n)-aggregate is undergoing this reaction is simply:

$$\frac{1}{\binom{N_n}{1}} = \frac{1}{N_n} \tag{2.111}$$

The probability that this (n)-aggregate undergoes a reaction during $\mathrm{d}\tau$ is then:

$$a_r \,\mathrm{d}\tau = \frac{A_r}{N_n}\,\mathrm{d}\tau = N_{i\neq n}.\alpha_r.\,\mathrm{d}\tau \tag{2.112}$$

Where α_r is the constant previously introduced that describes the average success rate of collisions of reactant and $N_{i\neq n}$ is the number of particle of the other specie involved, with convention $N_{i\neq n} = 1$ in the case of the dissociation of the aggregate.

In our playground, that is the line assembly model, an (n)-aggregate can be involved only in two reactions resulting in:

$$(n) \to (n+1) \tag{2.113}$$
$$(n) \to (n-1) \tag{2.114}$$

So that the master equation describing the probability $p_n(t)$ to find our single out aggregate with size n after a time t starting in n_0 is quite obvious:

$$\begin{cases} \dot{p}_n &= -\left(\alpha_n^+ N_1(t) + \alpha_n^-\right).p_n + \alpha_{n-1}^+ N_1(t).p_{n-1} + \alpha_{n+1}^-.p_{n+1} \\ \alpha_0^+ &= 0 \\ \alpha_N^- &= 0 \\ p(n, t=0) &= \delta_{n_0}^n \end{cases} \tag{2.115}$$

The associated Gillespie time of occurrence and occurring reaction probability depends on the evolution of $N_1(t)$ which is also a random variable... If the variation of the number of monomers $N_1(t)$ is slow compared to the aggregate evolution however, the distribution are given by equations (2.101) and (2.104):

$$P(\pm 1) = \left\{ \frac{\alpha_n^+ N_1}{\alpha_n^+ N_1 + \alpha_n^-}, \frac{\alpha_n^-}{\alpha_n^+ N_1 + \alpha_n^-} \right\} \tag{2.116}$$
$$P_\emptyset(\Delta t) = \left(\alpha_n^+ N_1 + \alpha_n^-\right) e^{-\left(\alpha_n^+ N_1 + \alpha_n^-\right)\Delta t} \tag{2.117}$$

The slow variation of N_1 compares to the reaction rate of one cluster is somehow a strong assumption since monomers are involved in all the reaction that takes place on the surface. It holds at least:

— in large systems at steady state, where the mean value of N_1^s does not evolve in time and when $N_1^s \gg 1$ fluctuations should be low compared to N_1^s (of order $\sqrt{N_1^s}$).

— as long as the total amount of monomers forming aggregates is low compared to the total number of free monomer in the system so that it can be considered as a reservoir.

We studied the somehow simplified case of the evolution of a singled out aggregate at the surface of a membrane where the aggregates populations have reached assembly steady state. We've seen in previous sections section 2.2.3 that there is a unique steady state solution in our model of line assembly under constant flux (flux over the system is uniquely defined by the input flux). This steady state point is generally reached and populations eventually stabilize at their steady values. In this context, the population of each aggregate size is stable since incoming flux balances the outgoing one at each size. However, at the individual level, aggregates keep growing or shrinking.

2.3.1.4 Microscopic rates from the trajectories

In most of the assembly experiments, the precise measurement of the size distribution is an hard task as the populations of intermediates are dominated by the assembled and monomers population by several orders of magnitude. This observation obviously includes the measure of its evolution on time.

Following a single particle size in time we access the information of its individual fate. Its size-time trajectory is one succession of events among the many possible realizations whose statistical properties rule the evolution of the whole system. Many trajectories must be collected and collectively treated to extract the information relevant at the global scale.

In our model, the succession of the aggregate sizes in time forms a Markov chain: the current state (size) of the aggregate is enough to fully determine the transition probability to future sizes as stated by the master equation (2.115). When the number of reachable sizes is finite, the matrix A whose element are $A_{n\to m}$ the transition rates from size n to size m, fully characterizes the Markov chain and is known as the generator of the time continuous Markov chain.

2.3.1.5 Elaborate methods for estimations of continuous time Markov chain generator

The estimation of the transitions rates of a time-homogeneous random walk (a random walk whose transition rates are time independent) from its trajectories are fairly easy when one has an infinite knowledge of the positions at any time during the process. To understand the construction of the estimation, we start with a discrete time random walk (a new jump occurs after a time Δt, so that the list of the successive occupied positions $\{n_v\}$ at each step v is enough to grant a full knowledge of the evolution). The maximum likelihood estimator of the transitions probability from any position i to a position j, $p_{i\to j}$ is constructed by maximizing

the probability of the observed sequence:

$$\begin{aligned} \mathrm{P}(\{n_v\}) &= \mathrm{P}(n_1, t=0) \prod_w \mathrm{P}(n_{w+1}|\{n_{v<w}\}) \\ &= \mathrm{P}(n_1, t=0) \prod_w p_{n_w \to n_{w+1}} \\ &= \mathrm{P}(n_1, t=0) \prod_{i,j} p_{i\to j}^{\mathrm{jumps}_{i\to j}} \end{aligned} \tag{2.118}$$

Where we first used the Markovian properties of the walk and then regroup altogether the $\mathrm{jumps}_{i\to j}$ occurrences of the jumps from site i to site j, $n_w = i, n_{w+1} = j$ in the sequences of positions. The transition probability $\mathrm{jumps}^*_{i\to j}$ that maximizes $\mathrm{P}(\{n_v\})$ under the constraint $\sum_j p_{i\to j} = 1$ is:

$$p^*_{i\to j} = \frac{\mathrm{jumps}_{i\to j}}{\sum_j \mathrm{jumps}_{i\to j}} \tag{2.119}$$

Which is the classic estimation of the transition probability by the statistical frequencies and converges for a sufficient number of observations. In the continuous time context, the transition probabilities become transitions rates (transition probabilities per unit time). For a complete observation over time $[0, t]$ the estimation is also analytical[96]:

$$A^*_{i\to j} = \frac{\mathrm{jumps}_{i\to j}(t)}{\mathrm{OT}_i(t)} \tag{2.120}$$

With $\mathrm{OT}_i(t)$ is the occupation time of the site: the portion of the total time t spent on the site i. However, the limitations of such method arise in more realistic cases that are continuous time, limited time sampling, and uncertainty on the position. When several transitions are missed or mistaken by others and that the residence time on a site is not well known, the characterization of the transition rates is harder...

Several works on the reconstruction of the Markov chain generator with finite sampling on time exist. The analysis of market or diseases evolution have prompted several works in economical and medical sciences[2,30].Metzner et al.[96] propose a comparison of classic iterative methods. When sampling in time is finite, none offers a direct determination of the transition rates but rely on subsidiary optimizations procedure. The common start of the methods is to use a first estimate of the generator exponential at sampling time Δt:

$$\left(e^{A\Delta t}\right)^* = \frac{\mathrm{jumps}_{i\to j}}{\sum_j \mathrm{jumps}_{i\to j}} \tag{2.121}$$

Where $\mathrm{jumps}_{i\to j}$ pool together the jumps observed between each observation separated by Δt (note the similarity with the discrete case, equation (2.119)). Unfortunately, such estimates of the exponential $e^{A\Delta t}$ does not lead to a good estimate of A. The calculation of the estimate logarithm does not forbid complex or negative off-diagonal entries (a theoretical issue known as the "embedding problem") incompatible with the physical meaning of a transition rate. Crommelin and Vanden-Eijnden[33] project the approximation obtained on the set of acceptable transition matrix defining a distance respectful of the left and right eigenvectors of the matrix which encodes its stochastic properties. Bladt and Sorensen[13] propose to find an optimum of likelihood function by iterations of the expectation maximization algorithm. Indeed, when no

analytical derivation of the maximum likelihood estimate is any more available, estimation of the rates relies on the iterations of expectation maximization algorithm which involves to solves systems of ordinary differential equations whose size is given by the total number of states in the system. Application of those methods in our interest case supposes a perfect knowledge of the occupied position at each measure, which is rather dubious concerning fluorescence intensity measurement. Furthermore their computation complexity quickly increases with the numbers of state considered. Eventually, they do not originally allow to restrict the form of the generator on a chosen family. For instance, concerning our aggregation process, we would like to forbid any transition involving more than one protein even when our resolution power restricts ourselves to the observation of size changes of several tens of proteins.

When a parametrization of the process exists, a third way, following the theory of the generalized moment method, consists in building a set of equations relating parameters values to the expectations of as many "test functions". The test functions expectations are evaluated with the empirical expectation of the observation, relying on the law of large numbers to obtain a consistent estimate. This is the spirit of the method proposed in the following.

2.3.1.6 A simple moment method to find microscopic rates from the trajectories

The master equation (2.115) can be seen as a parametrization of the random walk by $2 \times N$ parameters, namely $\alpha_n^{\pm}$. The determination of the random walk generator is equivalent to the determination of these parameters.

Averaging on all the possible trajectories using the probability of size given by the master equation (2.115) gives the average growth rate of a cluster after a time t:

$$\begin{aligned}
\sum_n n.\dot{p}_n &= \sum_n -\left(\alpha_n^+ N_1 + \alpha_n^-\right).p_n.n + \alpha_{n-1}^+ N_1.p_{n-1}.n + \alpha_{n+1}^-.p_{n+1}.n \\
\Leftrightarrow \partial_t \langle n \rangle &= \sum_n \left(\left(\alpha_n^+ N_1 + \alpha_n^-\right) + \alpha_n^+ N_1(n+1) + \alpha_n^-(n-1)\right) p_n(t) \\
&\quad + \alpha_0^+ N_1 p_0(t) + \alpha_N^-(N-1) p_N(t) \\
&= \langle \alpha_n^+ N_1 \rangle - \langle \alpha_n^- \rangle
\end{aligned} \tag{2.122}$$

Using the same method of shifting summation indices, we also find:

$$\begin{aligned}
\partial_t \langle (n - \langle n \rangle)^2 \rangle &= \langle \alpha_n^+ N_1 \rangle + \langle \alpha_n^- \rangle + 2\langle n(\alpha_n^+ N_1 - \alpha_n^-) \rangle - 2\langle n \rangle \partial_t \langle n \rangle \\
&= \langle \alpha_n^+ N_1 \rangle + \langle \alpha_n^- \rangle + 2\left(\langle n(\alpha_n^+ N_1 - \alpha_n^-) \rangle - \langle n \rangle \langle \alpha_n^+ N_1 - \alpha_n^- \rangle\right)
\end{aligned} \tag{2.123}$$

Conditioning the average over size n_0 a $t = 0$ for the aggregate, so that at $t = 0$, $p_n = \delta_{n_0}^n$ so $n(0) = n_0$, so we have:

$$\partial_t \langle n(t) \rangle_{n_0}\big|_{t=0} = \alpha_{n_0}^+ N_1 - \alpha_{n_0}^- \tag{2.124}$$

$$\partial_t \langle (n(t) - \langle n \rangle)^2 \rangle_{n_0}\big|_{t=0} = \alpha_{n_0}^+ N_1 + \alpha_{n_0}^- \tag{2.125}$$

Which gives us the difference of the local microscopic growing and shrinking rates from the time derivative of the mean displacement conditioned on the starting size. We can also obtain their sum using the time derivative of the variance of the positions at a given size. Hence they can be directly measured from the experiment to reconstruct the kinetic landscape of the aggregation. As times goes on, the time derivatives give those values averaged over all the sites that the particle can visits weighted by the probability to visit each site.

Concerning the mean squared displacement (equation (2.123)), we note an additional term that we link to spreading or squeezing contribution of inhomogeneous drift velocity in space. To understand its influence assuming a smooth variation of the microscopic transition with size, we use the limit of continuous space. We can write this term with $v_n = \alpha_n^+ N_1 + \alpha_n^-$ as:

$$\begin{aligned}(\langle n v_n\rangle - \langle n\rangle\langle v_n\rangle) &\equiv \int n v(n) p(n)\,\mathrm{d}n - \int n p(n)\,\mathrm{d}n \int v(n) p(n)\,\mathrm{d}n \\ &= \int n(v_0 + n\partial_n v|_0 + \ldots) p(n)\,\mathrm{d}n - \int n p(n)\,\mathrm{d}n \int (v_0 + n\partial_n v|_0 + \ldots) p(n)\,\mathrm{d}n \\ &\simeq \partial_n v|_0 \int n^2 p(n)\,\mathrm{d}n - \partial_n v|_0 \left(\int n p(n)\,\mathrm{d}n\right)^2 + \ldots \\ &\simeq (\langle n^2\rangle - \langle n\rangle^2).\partial_n v|_0 + \ldots\end{aligned} \tag{2.126}$$

In the case of inhomogeneous velocity in space, this leads to the set of relations between the two first moments of the position distributions at longer time scale :

$$\left\{\begin{array}{rl} \partial_t \langle n\rangle & = \langle \alpha_n^+ N_1\rangle - \langle \alpha_n^-\rangle \\ \partial_t \langle (n - \langle n\rangle)^2\rangle + \langle (n - \langle n\rangle)^2\rangle\, \partial_n \partial_t \langle n\rangle & \simeq \langle \alpha_n^+ N_1\rangle + \langle \alpha_n^-\rangle \end{array}\right. \tag{2.127}$$

Regions where velocity decreases in space tend to generate "traffic jams" in random walk (walkers are concentrated in lowest velocity regions) and oppositely. This contribution vanishes in regions with constant velocity and otherwise leads to non linear evolution of the mean square displacement. This correction becomes non negligible at longer time scale and leads to underestimating the diffusion in places where the velocity increases and overestimating it in the reverse case.

This general analysis holds as long as the evolution of the aggregates is simply determined by local microscopic rates (Markov linear random walk). In this condition, we can extract a simple estimation of the microscopic rates from the early linear behavior in time of the mean of displacement and its variance. In the following we will neglect the effect of the inhomogeneous velocity field and use the approximation:

$$\left\{\begin{array}{rl} \partial_t \langle n\rangle & = \langle \alpha_n^+ N_1\rangle - \langle \alpha_n^-\rangle \\ \partial_t \langle (n - \langle n\rangle)^2\rangle & \simeq \langle \alpha_n^+ N_1\rangle + \langle \alpha_n^-\rangle \end{array}\right. \tag{2.128}$$

This approximation becomes exact in the limit $t \to 0$ or of microscopic rates independent of the aggregate size.

The average $\langle n(t)\rangle$ considered here supposes that we perfectly sample the probability distribution with an infinite number of trajectories which, as a matter of fact, will never be experimentally tractable. We will always consider instead a finite number T of trajectories over which we will average. Since each trajectory is an independent variable drawn from the same distribution, the best linear unbiased estimators of the moments that can be constructed[61] are the empirical mean displacement and the empirical variance of the positions:

$$\overline{n(t)}^T = \frac{1}{T}\sum_T n_T(t) \tag{2.129}$$

$$\overline{\overline{n(t)}}^T = \frac{1}{T-1}\sum_T (n_T(t) - \overline{n(t)}^T)^2 \tag{2.130}$$

And a central limit theorem can be applied, providing that the two first moments $m_1(t)$ and $m_2(t)$ of $p_n(t)$ are finite (to apply the theorem on $\overline{n(t)}^T$) as well as the fourth order moment $m_4(t)$ (to construct the variance of $n(t)^2$). Hence our estimations will asymptotically approach the true value with a variance shrinking with $T^{1/2}$. The calculation of the exact behavior of $\overline{n(t)}^T$ and $\overline{(n(t)-n_0)^2}^T$ would require however to properly solve the master equation to find $p_n(t)$, for the expectation and variance of $\overline{n(t)}^T$ and $\overline{(n(t)-\langle n\rangle)^2}^T$ gives for identically distributed variable:

$$\mathbb{E}\left(\overline{n(t)}^T\right) = m_1(t) = \langle n\rangle(t) \tag{2.131}$$

$$\text{var}\left(\overline{n(t)}^T\right) = \frac{1}{T}\langle(n-\langle n\rangle)^2\rangle(t) \tag{2.132}$$

$$\mathbb{E}\left(\overline{\overline{n(t)}^T}\right) = m_2(t) - m_1^2(t) = \langle(n-\langle n\rangle)^2\rangle(t) \tag{2.133}$$

$$\text{var}\left(\overline{\overline{n(t)}^T}\right) = \frac{1}{T}\langle(n-\langle n\rangle)^4\rangle(t) - \frac{T-3}{T(T-1)}\langle(n-\langle n\rangle)^2\rangle^2(t) \tag{2.134}$$

The slopes β_i of $\overline{n(t)}^T$ and $\overline{\overline{n(t)}}^T$ are calculated on their measured values sampled in time $\{t_i, \overline{n(t_i)}^T\}$ and $\{t_i, \overline{\overline{n(t_i)}}^T\}$ with an ordinary least square adjustment of the model $\overline{n(t)}^T = \beta_1.t$ and $\overline{\overline{n(t)}}^T = \beta_2.t$. The ordinary least square solution for the model $f(t) = \beta.t$ is straightforward (there is a unique normal equation see equation (A.7)):

$$\beta^* = \frac{\sum_i y_i t_i}{\sum_i t_i^2} \tag{2.135}$$

$$\sigma^2_{\beta^*} = \frac{\sum_i (y_i - \beta^* t_i)^2}{(\sum_i 1 - 1)\sum_i t_i^2} \tag{2.136}$$

In the following we investigate the use of $\overline{n(t)}^T$ and $\overline{\overline{n(t)}}^T$ to extract $\alpha_n^+ N_1$ and α_n^- from trajectories.

2.3.2 The asymmetric random walk

The simplest test that can be proposed to evaluate the construction of trajectories and subsequent reconstruction of the microscopic rates is the constant rates case :

$$\begin{aligned} \alpha_n^+.N_1 &= g \\ \alpha_n^- &= r \\ v &= g - r \\ D &= g + r \end{aligned} \tag{2.137}$$

The probability for a forward/backward step is independent of the position and biased in a direction. If the boundaries are far enough to be ignored, the master equation (2.115) can be solved by a discrete Laplace transform of p_n which turns it into a differential equation of the "probability generating function": $F(z,t) = \sum_{-\infty}^{\infty} z^n p_n$. Multiplying equation (2.115) on both side by z_n and summing yields:

$$\begin{cases} \partial_t F(z,t) &= (gz + \frac{r}{z} - (g+r))F(z,t) \\ F(z,0) &= \sum_{-\infty}^{\infty} z^n \delta_0^n = 1 \end{cases} \tag{2.138}$$

$$\Leftrightarrow F(z,t) = e^{(gz+\frac{r}{z}-(g+r))t} \tag{2.139}$$

The inverse transform (Van Kampen[131]) then gives the asymmetric random walk probability to find the walker on site n at time t:

$$p_n(t) = \left(\frac{g}{r}\right)^{\frac{n}{2}} e^{-(g+r)t} I_{|n|}(2\sqrt{rg}t) \tag{2.140}$$

Where we used I_k the k$^{\text{th}}$ modified Bessel function, and set the initial site to be $n = 0$. Either from this final expression or directly from the master equation, it appears that this random walk is controlled by the ratio g/r, and once fixed the solution are equivalent up to a proper scaling of the time $\tau = r.t$.

Successive differentiations of $F(z,t)$ with respect to z give the factorial moments of p_n as $z \to 1^-$:

$$\begin{aligned} \langle n \rangle(t) &= \partial_z F|_{z=1} = (g-r)t = v.t \\ \langle (n - \langle n \rangle)^2 \rangle(t) &= \partial_z^2 F|_{z=1} + \partial_z^1 F|_{z=1} - \left(\partial_z^1 F|_{z=1}\right)^2 = (g+r)t = D.t \\ \langle (n - \langle n \rangle)^4 \rangle(t) &= \ldots = (g+r)t(1+3(g+r)t) = Dt(1+3Dt) \end{aligned} \tag{2.141}$$

$$\tag{2.142}$$

The two first results were expected from equations (2.122) and (2.123) since the drift velocity is independent of the size. This equations set holds for an infinite space or at early times far from the boundaries as already mentioned.

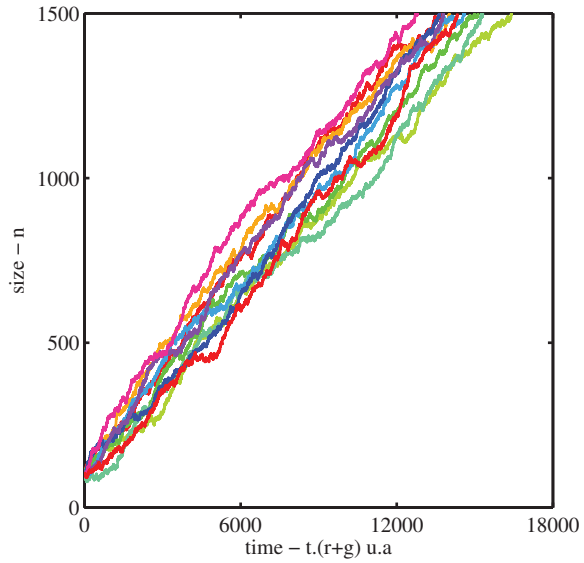
2.3.2.1 Simulations results:

The simplest stochastic simulation for the size evolution of a protein aggregates on time $n(t)$, both constant growth g and shrinking rates r, generates an asymmetric random walk. Ten different realizations of the asymmetric random walk, are displayed for two different set of transition rates on figure 2.23 with absorbing boundaries conditions set both at $n = 0$ and $n = N$. The sum and difference of the transitions rates g and r are shown below on the same figure. The estimations are distributed around the true values of the sum $D = r + g$ and difference $v = r - g$ as calculated (equations (2.122) and (2.123)) and according to the expected values (equations (2.131) and (2.133)). The variances of the empirical mean of the displacement equation (2.132) and empirical variance of the position equation (2.134) calculated from the exact results of equation (2.141) are:

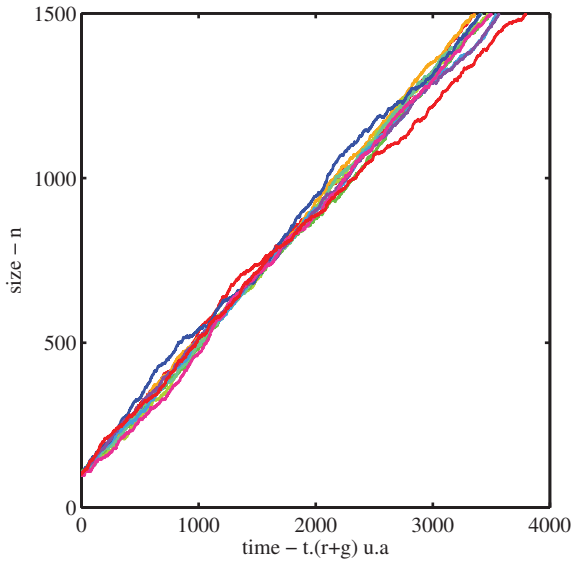
$$\text{var}\left(\overline{n(t)}^T\right) = \frac{D.t}{T} \tag{2.143}$$

$$\text{var}\left(\overline{\overline{n(t)}^T}\right) \simeq \frac{D.t}{T} + \frac{2D^2.t^2}{T}, \quad T \gg 1 \tag{2.144}$$

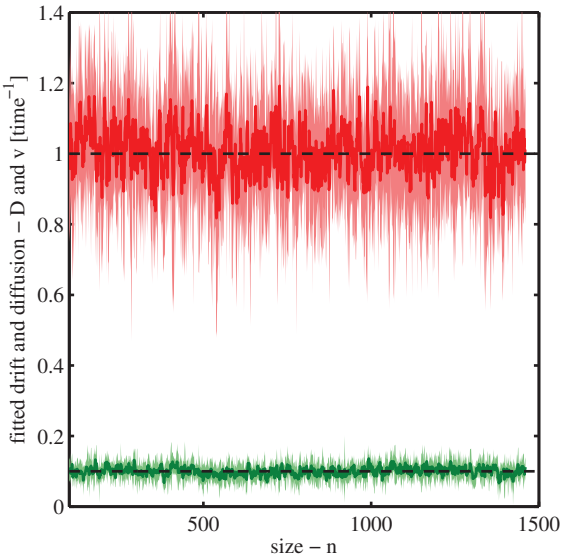
The stochastic nature of the trajectories is likely to induces variations of the mean position within an interval of typical length $\sqrt{\frac{Dt}{T}}$ around the expected value $v.t$ for an asymmetric random walk. As this typical length grows slower in time than the linear relation $v.t$ that we want to estimate, consequently a fit over a longer time should give better results in term of relative uncertainty on the slope. This consideration does not hold for the empirical variance of the positions whose variations are such that the typical interval grows at least linearly on time (see equation (2.144)) so that no gain in relative precision on the slope determination is expected.



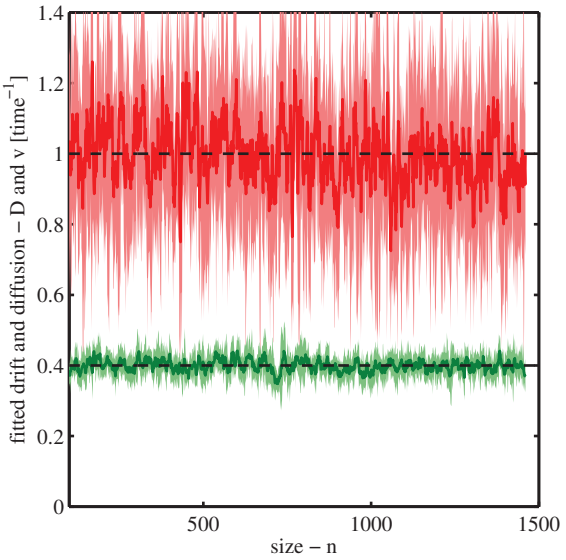
(a) $r = 0.45$ and $g = 0.55$



(b) $r = 0.3$ and $g = 0.7$



(c) theoretical values for the simulated trajectories (dashed lines) $D = g + r = 1$(red) and $v = g - r = 0.05$(green) and their estimated values (plain lines).



(d) theoretical values for the simulated trajectories (dashed lines) $D = g + r = 1$(red) and $v = g - r = 0.4$(green) and their estimated values (plain lines).

Figure 2.23 – 10 different realizations of the asymmetric random walk with different microscopic rates. The total number of proteins is $N = 1500$. Aggregates all start at $n = 100$ proteins at $t = 0$. Stronger asymmetry between the growth and shrinking rates generates stronger drift velocity: boundaries are met in less time and trajectories are indeed less dispersed since we work at constant D. Estimations of D and v are made based on a sample of 200 aggregates trajectories and confidence interval build from equation (2.145) are plotted in light colors

The fluctuations of $\overline{\overline{n(t)}}^T$ slope are also stronger with an increased drift velocity v (compare figures 2.23c and 2.23d). According to equations (2.143) and (2.144), the value of the drift velocity should not affect the variance of $\overline{\overline{n(t)}}^T$ in apparent contradictions with the simulations. This observation is an illustration of the indirect influence of v on the number T of trajectories actually averaged in our method. At fixed diffusion D, a smaller drift velocity v increases the probability for a particle to come back to a position it has already occupied (recurrent walk) probing another time this domain. Different portions from the same trajectories starting on the same point can thus provides several independent contributions in the sums so that a low drift virtually increase the number of trajectories analyzed. In figure 2.23c, $T \simeq 400$ portions of trajectories can on average be used to construct the empirical means because an aggregate shrinks back to earlier sizes many times. In contrast, on figure 2.23d, $T \simeq 200$ for very few aggregate shrinks so that each size is tested only once for each trajectory.

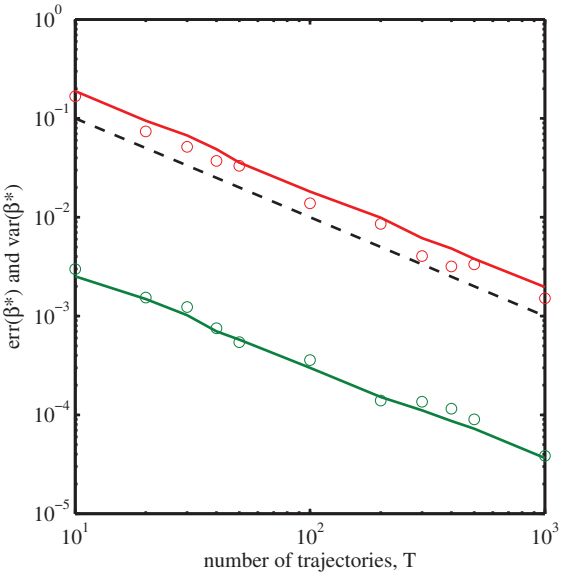
The ordinary least square method can certainly not be trusted to properly treats the stochastic fluctuations in estimating the slope of $\overline{\overline{n(t)}}^T$ and $\overline{n(t)}^T$. To illustrate this point, let us suppose that the recorded trajectories are not a good sample of the distribution so that we have a positive bias: the average position increases faster than the expectation on early sampled times. The Markovian property tells us that the evolution is independent of the early behavior. We thus expect the average position to increase according to equation (2.122) as times goes on and the position to remain biased until further strong deviations from the average adds to the previous either lower or increase the bias.

Consequently, deviations from the expected linear relation are certainly not homogeneous and uncorrelated on time as assumed by a least square estimation. The result does not change much when we take instead the average of the slope between consecutive data points. We propose to use least square not to correct fluctuations building up over time but to treat correctly additive fluctuations arising from the measurement noise and poor determination of the sizes in the experimental situations likely to fulfill the least square assumptions.

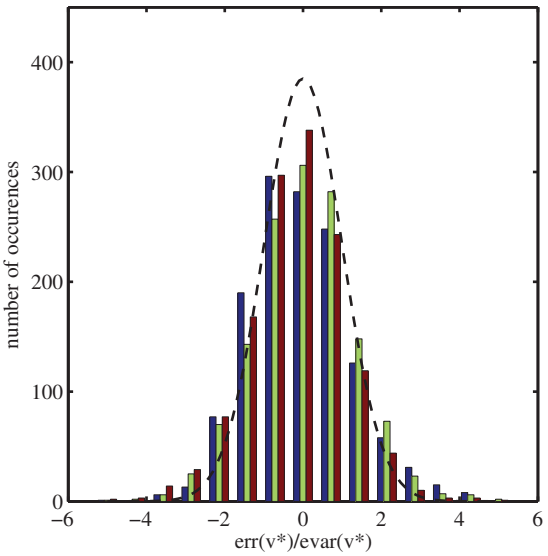
In those conditions, our confidence bounds for the estimated value of the slope are build on the wrong assumption of a normal distribution of error equation (2.136) and are several order of magnitude under the true amplitude of the fluctuations. They should not be used. Instead, we found in our simulations that the following ad-hoc estimation based on the dispersion of the sampled points around the fit gives a proper order of magnitude :

$$\mathrm{evar}\beta^* = \frac{\mathrm{var}(y_i - \beta^* t_i)}{\mathrm{var}\, t_i^2} \sim \mathrm{var}\ \beta^* \tag{2.145}$$

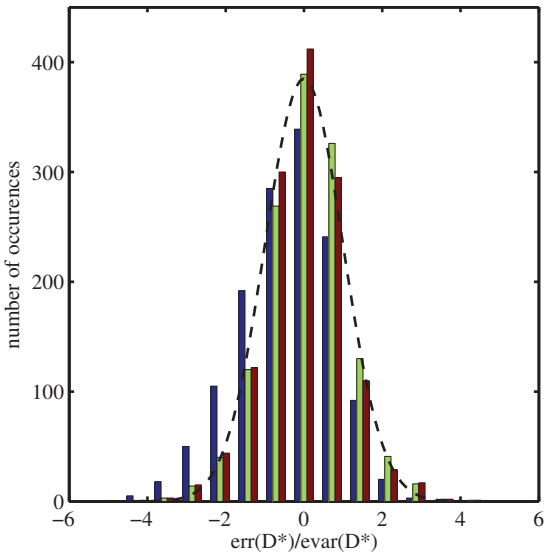
Where β is the estimated slope using equation (2.135). Assuming normal distribution, the 99%-confidence interval $\pm 3(\mathrm{evar}\beta^*)^{\frac{1}{2}}$ built on this estimation is displayed around the estimated value on figure 2.23. The ratio of the real error (the difference between the true value and the estimated one) over its estimated order of magnitude $\sigma_{\beta^*} = \sqrt{\mathrm{evar}\beta^*}$ is shown on figure 2.24. As stated by the central limit theorem, the estimators of the difference v and sum D of the microscopic rates are consistent and converge to the true values as the number of trajectories used increases (figure 2.24a). Furthermore they are unbiased in the case of asymmetric random walk studied here. Also the uncertainty on D dominate the uncertainty on v by one order of magnitude (equation (2.141)). The two variances estimated by equation (2.145) follow on average the real variance of the estimations $v*$ and D^* with a slight overestimation concerning D.



(a) Mean over all the aggregate sizes of the variance of the estimate ($\circ$) and of the estimated standard deviation using equation (2.145) (plain lines) of v (green) and D (red) as functions of the number T of trajectories used to estimate them. The black dashed line shows the power law $T^{\frac{1}{2}}$ predicted by the central limit theorem.



(b) Histogram of the ratio of the true error to the estimated standard deviation: $\frac{v^*-v}{\sqrt{\text{evar}v^*}}$, for $T \in \{10, 100, 1000\}$ (resp. blue, red, green). The probability density function of $\mathcal{N}(0,1)$ is shown for comparison (dashed black).



(c) Histogram of the ratio of the true error to estimated standard deviation: $\frac{D^*-D}{\sqrt{\text{evar}D^*}}$, for $T \in \{10, 100, 1000\}$ (resp. blue, red, green). The probability density function of $\mathcal{N}(0,1)$ is shown for comparison (dashed black).

Figure 2.24 – Statistical comparison of the true error $\text{err}(\beta^*) = \beta^* - \beta$ to the estimated variance $\text{evar}\beta^*$ (equation (2.145)). The data used were simulated in the same conditions than those displayed on figure 2.23b: $D = 1$, $v = 0.4$, each sample of T trajectories produces $N = 1500$ independent estimations of v and D for all the possible sizes).

Figures 2.24b and 2.24c show in more detail the distribution of the real errors $v^* - v$ and $D^* - D$ rescaled by σ_{v^*} and σ_{D^*}. It shows that what was established for the average holds at individual estimation level: equation (2.145) statistically gives the order of magnitude of the error. Note that the correlation between $(\mathrm{evar}\beta^*)^{\frac{1}{2}}$ and $|\beta^* - \beta|$ is almost null for any given $T \gg 1$ (correlation coefficient $r^2 \simeq 10^{-2}$), so that only the order of magnitude of the error can be inferred from $(\mathrm{evar}\beta^*)^{\frac{1}{2}}$. The normal distribution of the ratio on figures 2.24b and 2.24c directly comes from the normal distribution of the true error, that the central limit theorem guarantees asymptotically as $T \to \infty$. Convergence is already good with a few tens or trajectories averaged and justify a posteriori the use of the normal quantiles to deduce a α-confidence interval:

$$\beta^* \pm \phi^{-1}(\alpha)\sqrt{\mathrm{evar}\beta^*} \tag{2.146}$$

With ϕ^{-1} the inverse of the cumulative distribution function of the normal law, also called the probit function.

The asymmetric random walk correspond to the steady state of the macroscopic equations studied in the section 2.2.4 with a constant macroscopic on-rate corresponding to the diffusion limited hypothesis and a bud free energy linear or approximately linear in the number of proteins that determines off-rate through detail balance. Despite the strong impact of nucleation steps on the steady state distribution demonstrated by including the line tension in the droplet model, there is little hope that total internal reflection fluorescence single particle trajectories can shed light on it. The main reason is that the sizes that are directly concerned are typically of very few proteins and likely to lay beyond the resolution power of fluorescence among the subsequent enrichment of the system in small aggregates likely to generate a strong fluorescent background. Overall, the microscopic rates can be reconstituted with appreciable precision for middle sized particles. In this range the two models we studied are equivalent when monomers are abundant, as required for our stochastic description to make sense.

2.3.3 A non constant detailed balance case

A different model is proposed in the work of Ku et al.[80] where the difference between macroscopic kinetics on-rate and off-rate is assumed proportional to the rim of the growing protein cap in the bud and where the stochastic effects are neglected. At the microscopic level, and under detailed balance, both on-rate and off-rate are proportional and vary the same way with respect to the rim length. This is a model where the attachment rate is proportional to the numbers of sites available to anchor at the rim of the bud and the detachment rate depends on the strength of the binding at the edge of the protein cluster. We already derived the number of available binding sites at the rim of a spherical bud in the continuous limit equation (2.21):

$$n_L = \pi\sqrt{(n-1)\left(1 - \frac{n}{N}\right)} \tag{2.147}$$

This version differs from the previous constant rates case for which our estimations where exact. In this new model:

$$\begin{aligned} \alpha_n^+ &= \mathrm{C}^{\mathrm{te}} \times \pi\sqrt{(n-1)(1-\frac{n}{N})} \\ \alpha_n^- &= \mathrm{C}^{\mathrm{te}} \times \pi\sqrt{(n-2)(1-\frac{n-1}{N})}e^{-\frac{f(n)-f(n-1)-f(1)}{k_B T}} \end{aligned} \tag{2.148}$$

In order to be consistent with the model proposed by the article, we furthermore assume a linear relation between the size of the bud and the enthalpy of formation of the bud, $f(n)$. It follows:

$$\frac{\alpha_n^+.N_1/S}{\alpha_{n+1}^-} = e^{-\frac{\Delta f(n)-\Delta f(n-1)}{k_B T}} = e^{-\frac{\Delta \mu}{k_B T}} \tag{2.149}$$

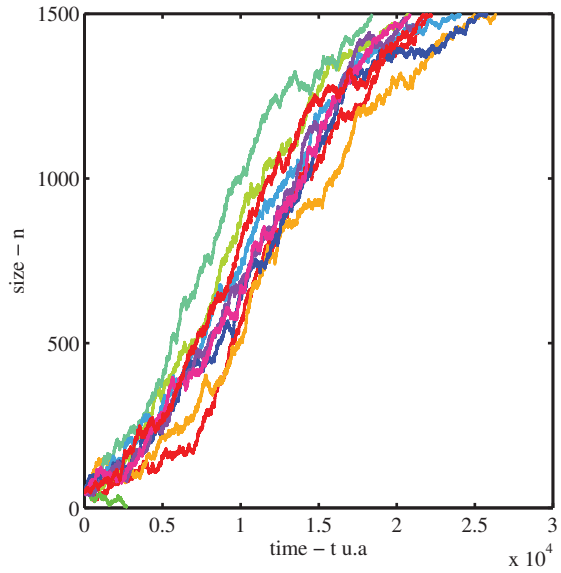
With $\Delta\mu$ the average free energy difference between a protein freely diffusing on the cell membrane or inside a bud. Using the modified Gillespie scheme, we obtained the trajectories shown on figure 2.25.

In contrast with the simple random walk model, the aggregate size evolution on time exhibits a sigmoid shape as reported by Jouvenet et al.[72], Ku et al.[80]. However the background noise hides in both cases the early evolution. A random walk with partial absorbing condition at $n = N$ (virus disappear at finite rate) would probably also accommodate the shape of the experimental data showed by both publications and in our opinion it is not clearly established that this model should be preferred without proceeding to a proper extraction of the microscopic rate from the trajectories.

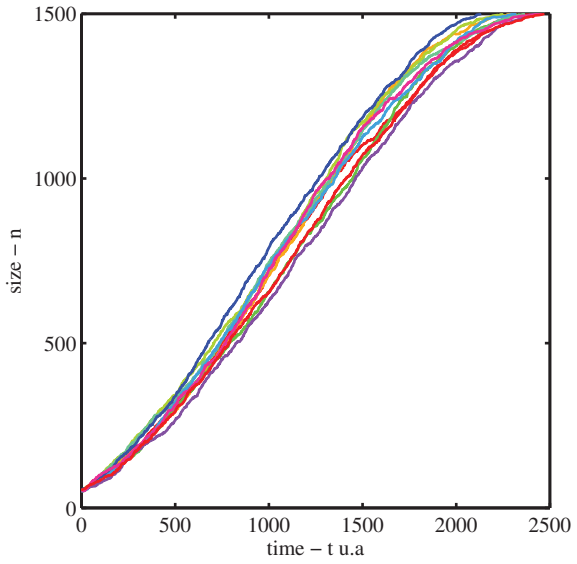
The microscopic rates can be found out of the time evolution of the mean displacement and the position variances using the estimations equation (2.128) which are not any more exact with size dependent rates. Results are shown on figures 2.25c and 2.25d. A experimentally unrealistic number of trajectories was used to calculate the rates figure 2.25d in order to reach a precision sufficient to see the limit of our method.

Both diffusion and drift velocity are overestimated for smallest sizes and underestimated for larger ones. Two effects can be invoked to explain this result. Firstly, we probe the average microscopic rates of the various sizes"visited" by the aggregate in the neighborhood of the stating position. As the drift is positive everywhere, larger sizes are most likely to be visited that smaller ones. Thus mostly probe the sizes ahead of the starting point and this is equivalent to switch both reconstructed curves to the left of the theoretical ones as observed. This effect is expected to be stronger at the center of the interval where the drift is stronger. Secondly, higher space derivative of the drift velocity are not null, and we neglected their contribution on the variance of the positions (equation (2.123)). As already discussed, this leads to an overestimation of the diffusion with increasing drift velocity and underestimation in the reverse case.It is expected to affect more the diffusion curves at the edges of the interval where the derivatives values are higher.

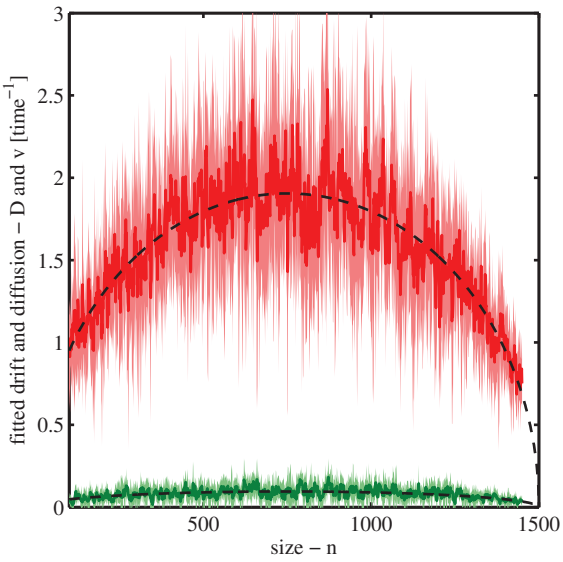
With lower drift velocity, the probability for an aggregate to visit many time a given size increases and a good enough sampling is reached with much less aggregates trajectories (figure 2.25c). In the condition of this simulation, the effects of our approximations are negligible compared to the error bars, hidden by the amplitude of the stochastic fluctuations. We noticed that our simulations produce estimations of the diffusion D_n with a relative error of constant magnitude $\frac{D_n^*-D_n}{D_n}$. According to the derivation of var $\overline{\overline{n(t)}^T}$ (equation (2.144)) for the asymmetric random walk, there is indeed a regime such as $\overline{\overline{n(t)}^T} = D.t \pm D.t\sqrt{\frac{2}{T}}$ when the quadratic term D^2t^2 dominates var $\overline{\overline{n(t)}^T}$. Then the standard deviation is proportional to Dt and we expect the relative error to be of order $\sqrt{\frac{2}{T}}$ (this leads to a 10% relative error on figure 2.23d which agrees reasonably with the observations). The features seen on the study of the asymmetric random walk seem general enough to hold locally in the more complicated landscape that we study here.



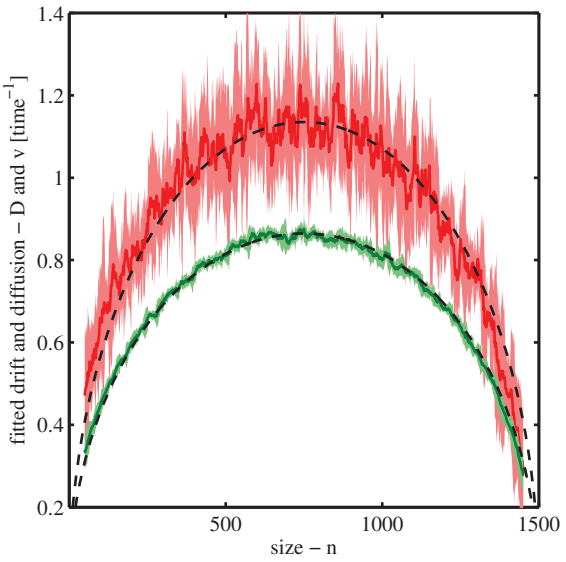
(a) 10 single aggregates size evolution for $\Delta\mu = 10^{-1}k_BT$



(b) 10 single aggregates size evolution for $\Delta\mu = 2k_BT$



(c) Values of $\alpha_n^+ \pm \alpha_n^-$ for the 50 simulated trajectories according to equation (2.148) with $\Delta\mu = 10^{-1}k_BT$ (dashed lines) and their estimated values (plain lines) for the drift (green) and diffusion (red)



(d) Values of $\alpha_n^+ \pm \alpha_n^-$ for the 1000 simulated trajectories according to equation (2.148) with $\Delta\mu = 2k_BT$ (dashed lines) and their estimated values (plain lines) for the drift (green) and diffusion (red)

Figure 2.25 – 10 different realizations of the aggregate evolution with microscopic rates proportional to the bud perimeter. The total number of proteins is $N = 1500$. Aggregates all starts at $n = 100$ proteins at $t = 0$. Confidence intervals build from equation (2.145) are plotted in light colors

Ku et al.[80] focused on the various "pauses" visible on the size evolution trajectories. The pauses are defined as events where a longer time is spent in a size interval compared to what would have been expected from the average sigmoid evolution. Or equivalently a pause is a plateau on the size-time trajectory curve. Similar events can be pointed out in the trajectories showed on figure 2.25a and are caused by mere fluctuations as the drift velocity is low compared to the diffusion. Such pauses in the course of a real biological process can either be interpreted as resulting of the stochastic fluctuations as in our model, or as the indirect signature of a biologically relevant event such as proposed by Ku et al.[80]: unavailability of the protein lattice to aggregation due to internal stress until relaxation by formation of a defect, or intervention of an external protein. This alternative can be decided comparing the observed statistic of the "pauses" to the expected statistic of the fluctuations.

The statistics of the time spent in a chosen interval due to the stochastic fluctuations alone are given by the first passage time distribution. The probability to find "pauses" that are pure fluctuations is given by the long time tail of the distribution for times larger than the mean. In the case of a simple asymmetric random walk, and considering an interval of length d "long enough", the first passage time distribution is approximately the first passage solution on a semi-infinite interval of a aggregate starting at distance d from the exit size. This distribution is obtained in the continuous approximation with a drift velocity v and a diffusion constant D by the renewal approach [2] using Laplace transform (see Redner[110]) and is sometimes called the "inverse Gaussian function":

$$\mathrm{P}(\tau, d) = \frac{d}{\sqrt{4\pi D\tau^3}} e^{-\frac{(d-v\tau)^2}{4D\tau}} \tag{2.150}$$

This distribution is however not easily comparable to the data provided by the article from Ku et al.[80] which are: the average growth rate of the buds B (completion is fulfilled when the zenithal angle reach π) and τ the time constant of the exponential tail of the pause length distribution.

$$B = 3 \pm 2\ 10^{-3}\ \mathrm{rad.s}^{-1} \tag{2.151}$$

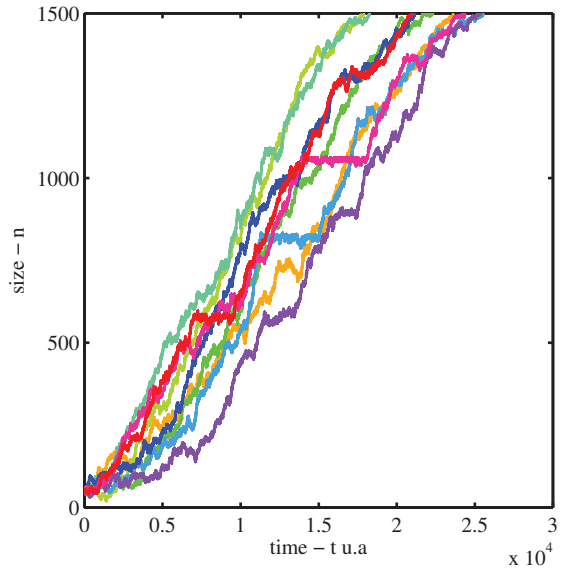
$$\tau = 4,7\ \mathrm{min} \tag{2.152}$$

As pauses length have been compiled irrespectively of the length of the "pause" interval or of the size in question and that only the large time tail have retained the attention of the authors the identification with our model parameter are not possible.

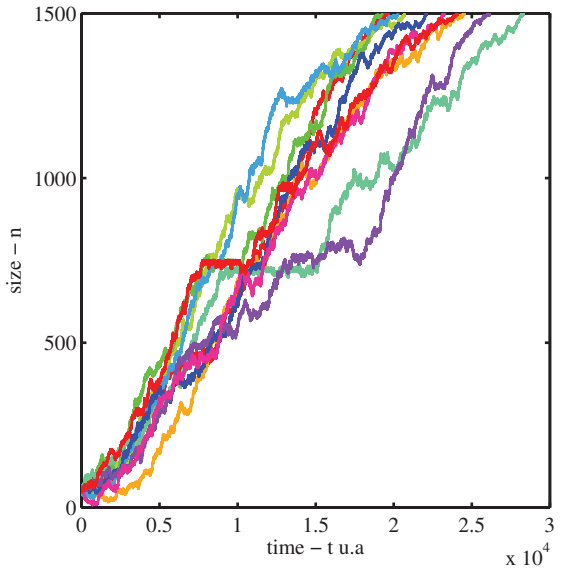
The inverse gaussian solution is probably not a safe guess if there is significant microscopic rates variation on the interval range or to evaluate the long time tail of the first passage time distribution on a finite interval which would decreases exponentially in time with a typical time constant given by the lowest eigenvalue of the Markov chain generator. More involved attempts to properly solve the first passage problem require the details of the microscopic rates.

Looking at the question from a different point of view, we might also ask whether such events are likely to leave their signature in the microscopic rates measured according to our

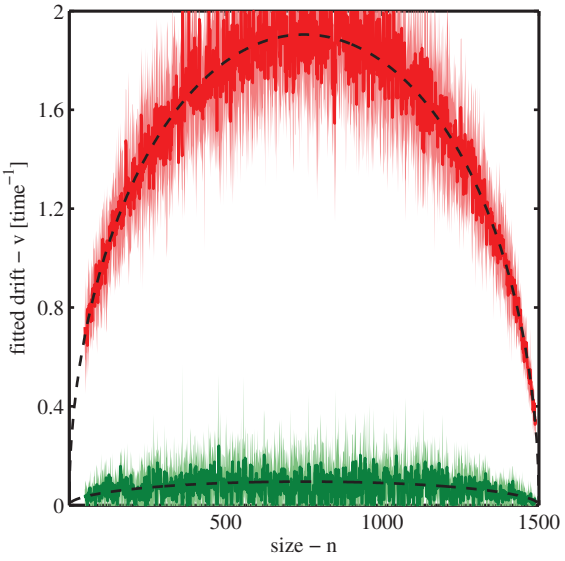
2. The renewal method requires to find the solution of the master equation considering infinite size interval which is often a difficult task with the additional difficulty to calculate the Laplace transform. Alternatively, it is also possible and sometimes easier to directly solve the adjoint equation on the interval of interest with absorbing boundary conditions. The first passage time distribution is the solution defined this way (see Redner[110], Van Kampen[131]).



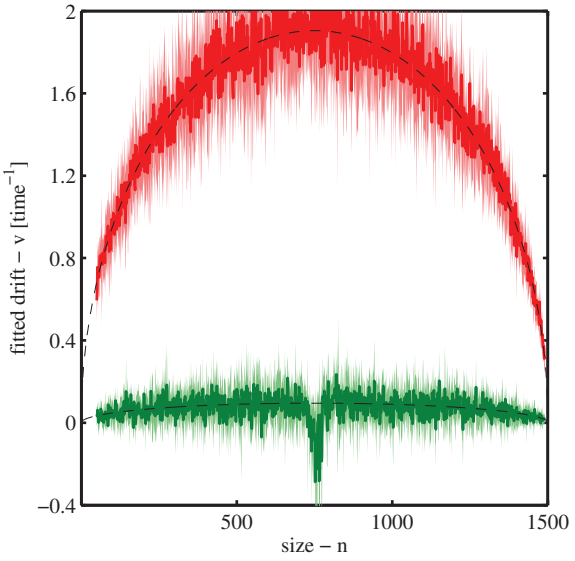
(a) 10 single aggregates size evolution for $\Delta\mu = 10^{-1}k_BT$ everywhere but at a random position drawn all over the available sizes where $\Delta\mu_b = 1k_BT$, simulated according to equation (2.148)



(b) 10 single aggregates size evolution for $\Delta\mu = 10^{-1}k_BT$ everywhere but at a random position where $\Delta\mu_b = 1k_BT$ drawn around half completion $N = 750 \pm 20$, simulated according to equation (2.148)



(c) Values of $\alpha_n^+ \pm \alpha_n^-$ for 100 simulated trajectories as shown above. Estimated values (plain lines) for the drift (green) and diffusion (red). Theoretical values in the absence of the random perturbations are shown (dashed lines)



(d) Values of $\alpha_n^+ \pm \alpha_n^-$ for 100 simulated trajectories as shown above. Estimated values (plain lines) for the drift (green) and diffusion (red). Theoretical values in the absence of the random perturbations (dashed lines)

Figure 2.26 – Effect of a randomly positioned barrier on the estimated microscopic rates

estimation. The answer depends much on the correlation between the event and the size. If pauses are triggered by a random event uncorrelated with the bud size, the perturbation effects are smoothed in the average over the different trajectories (figure 2.26c). The perturbation becomes noticeable only when pauses occurring on different trajectories overlap significantly. This is likely to occur when the bud size and the occurrence of pause are correlated. Equivalently, The problem can be stated considering the vector containing the attachment microscopic rates for all the positions. For each trajectories in the sample one element has been modified on a different position of this vectors. When all the trajectories are averaged together to build our estimation, the modification are visible either if they are large enough or when several smaller modifications happens on close positions (figure 2.26d). In any case the pauses will be easily identified on the time trajectories before they lead to a significant change of the reconstructed rates. However, an effect visible on the estimated rates is the signature of an event effecting the whole sample.

2.3.4 Conclusion

We adapted the classic Gillespie algorithm to follow the evolution of a single cluster in the steady state condition. This gives the possibility to generate the growth trajectory of an individual object which is not provided by the classic nucleation theory of the previous section.

Experiences following growth of the budding viruses in time have already been reported Jouvenet et al.[72], Ku et al.[80] and make possible the direct measurement of the reaction rates partly including the subunit concentration *in-vivo*. However extraction of the microscopic rates from the stochastic trajectories are not straightforward. We provide a simple treatment for the data with an estimation of the confidence interval of the estimation that was accurate in the conditions of our simulations. The method determine a quantitative kinetic landscape from the trajectories. the typical rate of convergence is set by the central limit theorem. In the simulations, the number of trajectories needed to differentiate the mean from the fluctuations appears to be of the order of a hundred.

The effects of the experimental limitations: the limit of size variation that can be detected in the fluorescence signal as well as the presence of additive experimental artifacts, are yet to be evaluated to set the proper precision limit of the estimation. The most prominent limitation is likely to be a rapid evolution of the free subunit concentration compared to the time a capsid needs to complete and make the averaging method meaningless.

Chapter 3

Superresolution image measurement

In this part we will address the issue of the measurement of a labeled structure imaged by superresolution microscopy technique. When viruses or any biological object labeled with fluorescent tags are imaged in superresolution microscopy, the label positions are recorded. Measurement in this situation means finding the best morphological parameters for a model that describes the observed positions. Regarding the immature HIV, we are willing to extract the radius and protein coverage of each particle. Our first step is to properly define the relationship between the underlying continuous structure that we want to characterize and the individual label positions recorded experimentally that are discrete. In a second step, we derive a method to center and orient the image in chosen references. Eventually we expose the method to find the best parameters from the image.

3.1 Superresolved microscopy images seen as a probability distribution

The quantity of interest in fluorescence microscopy is the spatial distribution of the labeled object in the field of view. Depending on the scale considered, this quantity can be treated in different ways that we will review here.

In the case of classical fluorescence and at a scale much higher than the diffraction limited resolution, the distribution of emitter is modeled as a classic density in space $D_{\text{classical}}(\vec{X})$. As the scale is orders of magnitude higher than the diffraction length, the intensity of light received by each pixel is assumed to be proportional to the number of emitters inside its area and the proportionality factor given by the photon yield n_i of the emitters:

$$I(\vec{X}) = D_{\text{classical}}(\vec{X}).\bar{n_i} \tag{3.1}$$

and the image is directly a representation of $D_{\text{classical}}$.

At smaller scales, close but larger than the diffraction limit, the continuous description $D_{\text{diffraction limited}}(\vec{X})$ is still valid as emitters typical size is much smaller than the diffraction length and each pixel receives the contribution from many of them. The blurring effect due to the diffraction in the imaging device adds and the photon from an point-like emitter are collected over all the pixel overlapping with its diffraction spot. The intensity of light received by each pixel is now assumed to be proportional to the number of emitters inside the imaged

area and an additive contribution of neighboring area whose diffraction spots overlaps. The resulting image is then the convolution of the spatial density of emitters by the point spread function introduced earlier in section 1.3.1:

$$I_{\text{diffraction limited}}(\vec{X}) = \iint D_{\text{diffraction limited}}(\vec{X} - \vec{Y}).\bar{n_i}.\,\text{psf}\,(\vec{Y})\,\mathrm{d}\vec{Y} \tag{3.2}$$

Closer again, at scales addressed by the diffraction limit, the discreteness of the emitter distribution is prevalent and each photon count is attached to an emitter whose position is estimated. But the emitter position is not determined with infinite precision. Instead superresolution imaging produce a density of probability for the emitter to be localized at a given position. Unlike previous description of the emitters by a density, which imply a spatial averaging over the emitter distribution, single molecules localization gives spatial information at each recorded emitter level. The final image can be defined as the sum of all the probability density for every localized molecule:

$$I_{\text{superresolved}}(\vec{X}) = \sum_i \text{pdf}_i(\vec{X}) \tag{3.3}$$

To relate the superresolution image to a emitter density as in precedent cases, we must consider the physics of the superresolution measurement which consists in randomly drawing a position from the emitter distribution known up to the localization precision. The emitter distribution can be modeled by a density as previously said or a discrete distribution depending on the level of description needed. Such description connects the image and the density taking into account both the stochastic nature of the emitter excitation and the precision of the localization procedure.

3.2 Modeling emitters physical distribution in the budded virus

The experimental data that we want to model are superresolution images of human immunodeficiency virus like particles budded from cells transfected with a labeled capsid proteins vector. The labels positions are recorded with a typical localization precision of $\sigma_{x,y} \simeq 15 - 20\text{nm}$. The typical structure of the gag protein shell reported in a review by Briggs and Kräusslich[15] (section 1.1.3) is an hexameric lattice partially coating two third of the inner surface of the lipid membrane folding the spherical particle. Distance between two hexamers centers observed in cryo-electron tomography is of the order of 8 nm, whereas radius are of the order of 65 nm. Regarding the uncertainty on precision at least twice as big as the typical distance between proteins, it seems illusory to access the information of the protein lattice organization. However the global features of the spatial distribution of proteins such as the spherical radius and the size of the uncoated membrane patch are of the same order of size as the positioning uncertainty and thus may be within the reach of the measure. With such a precision, we chose to model the distribution of gag proteins in the budded virus with a continuous spatial density consisting in a uniform distribution on a truncated sphere.

We parametrized the truncated sphere by its radius R and the maximal polar angle defining its completion θ in spherical coordinates. The uniform probability distribution in the spherical

coordinate system using the sphere center as origin and its symmetry axis aligned with $\vec{u}_z$ is written as:

$$D_{\text{model}} \equiv D^{\circ}_{R,\theta}(r,\vartheta,\varphi) = \frac{\delta(r-R)\mathcal{I}_{[0,\theta]}(\vartheta).r.\sin\vartheta}{2\pi R^2(1-\cos\theta)} \tag{3.4}$$

With $\mathcal{I}_{[0,\theta]}$ the indicator function of the domain $[0,\theta]$ and $\delta(r)$ the Dirac delta distribution.

Its position in space is given by the position of its center $\vec{X}_{\text{c}}$ and two Euler angle α and ϕ are enough to describe its orientation thanks to the cylindrical symmetry of the object (see figure 3.1-right). In the following, we note the probability density in 3D space with prescribed position and orientation $D_{\vec{\beta}}$, and $d_{\vec{\beta}}$ its projection in the plane (x,y) that we chose as the focal plane of the microscope in 2D imaging:

$$d_{\vec{\beta}}(x,y) = \int_{\mathcal{R}} D_{\vec{\beta}}(x,y,z)\,\mathrm{d}z \tag{3.5}$$

Where $\vec{\beta} = (R,\theta,\phi,\alpha,\vec{X_c}\}$ is the set that parametrize the spatial density of our model. Our goal is to identify the spatial density (equivalently the parameter set $\vec{\beta}$) which best describes the distribution of proteins imaged.

It is also necessary to take into account the effects of the environment and the measure artifacts that constitute the measure "background noise" and the distortions. Background noise adds to the object of interest on the image and distortions transform the image of the object compare to the object itself.

By background noise, we mean any aspect of the image that is not linked to the physical signal emitted by the structure that we are studying: positions of proteins that do not belong to the virus as well as any spurious signal generating artifact positions for instance CCD noise. We describe such background as a collection of outlier positions. In all this work, we used the minimal model of a constant density added to the model density to account for the background noise.

We use uncertainties of the measure or precision of the measure to speak about the distortions between the object and its image. The next section details our effective description of the superresolution positioning uncertainties. In order to be able to test our procedures, we will first precise how to generate superresolution simulated images from a known density.

3.3 Modeling superresolution imaging process

Starting from an given density of presence for the emitters distribution, how can we generate a superresolution image of it? Superresolution imaging process could be naively simulated by mimicking each of its step. First the position of an exited emitter is drawn from their distribution –for instance modeled as a continuous probability density function. Then a total number n_i of emitted photon randomly chosen from the emission statistic of the dies and as many pixels position are drawn on the CCD from the point spread function of the imaging device centered on the image of the exited emitter. Spurious photon detections must be added on the CCD to account for the background emission. In a last step, the standard centroid localization algorithm is used to obtain the probability density function for this emitter. And eventually those steps are iterated until the desired number of recorded positions is reached. However, all this cumbersome procedure can be simplified in a much basic one and details of the operations can be bypassed to obtain equivalent results.

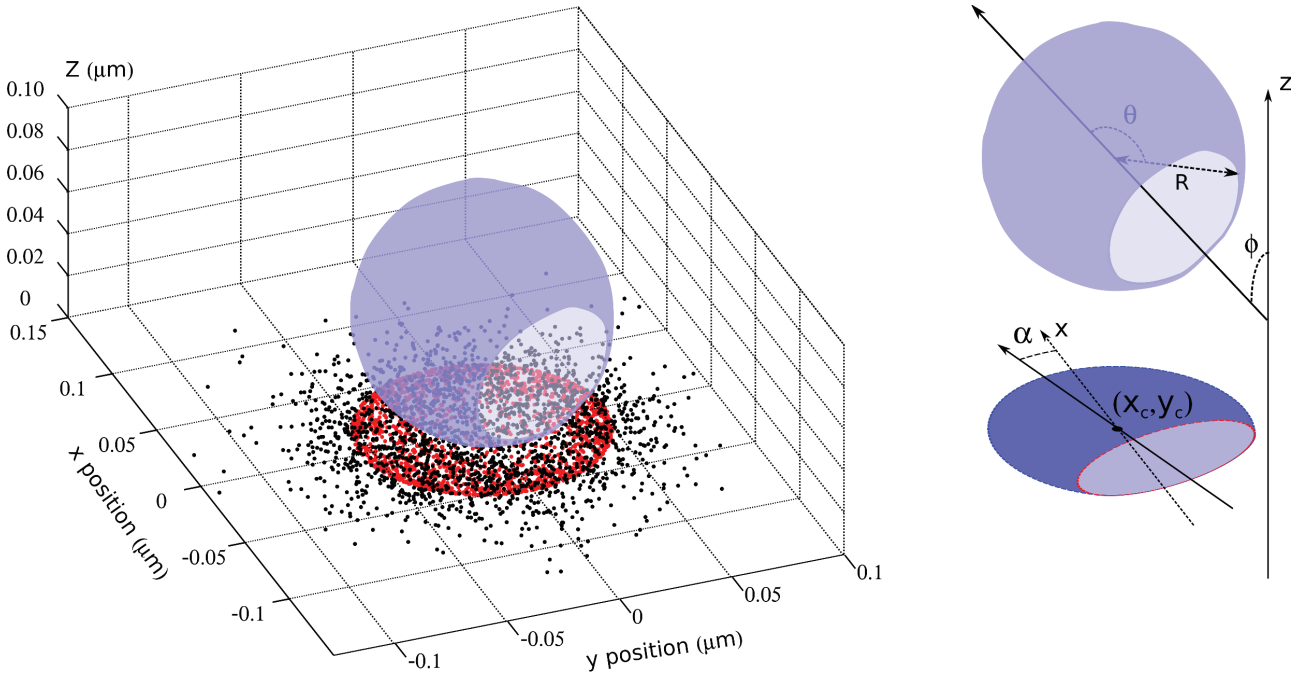


Figure 3.1 – Left: $5\,10^3$ positions uniformly sampled in the truncated sphere model (red dots) with $(R = 50\,\text{nm}, \theta = \frac{2\pi}{3}\text{rad}, \phi = \frac{\pi}{3}\text{rad})$ and the simulated measured positions with a precision $\bar{\sigma} = 20\,\text{nm}$ (black dots). Right: The 3D truncated sphere model and our parametrization (convention: z axis correspond to the optical axis of the microscope).

We start by drawing the positions $\{\vec{X}_i\}$ of the set of emitters that are to be imaged in the emitter distribution. We first draw positions uniformly on the unit sphere as follows:

— we generate a set of normally distributed values using Matlab® `randn` command:

$$\{x, y, z\} \sim \mathcal{N}(0, 1) \tag{3.6}$$

— we normalize, so that $\{x_1, y_1, z_1\}$ are uniformly distributed on the unit sphere

$$\{x_1, y_1, z_1\} = \{\frac{x}{\sqrt{x^2+y^2+z^2}}, \frac{y}{\sqrt{x^2+y^2+z^2}}, \frac{z}{\sqrt{x^2+y^2+z^2}}\} \tag{3.7}$$

We then reject the positions with height $z_1 < \cos\theta$, so that no positions are sampled on the surface patch $\vartheta > \theta$.

$$\{x_2, y_2, z_2\} = \{x_1, y_1, z_1 | z_1 < \cos\theta\} \tag{3.8}$$

following this procedure, we produced a set containing an arbitrary number of 3D positions $\vec{X}_2 = (x_2, y_2, z_2)$ uniformly drawn in the truncated unit sphere $\mathcal{S}_t$, according to the distribution $D^{\circ}_{R=1,\theta}$. This set is eventually transformed into a set of positions uniformly sampled on our target distribution $\vec{X}_3 \sim D_{\vec{\beta}}$ by two successive rotations of angle ϕ with axis $\vec{u}_x$ ($Rot_{\vec{u}_x,\phi}$) and α with axis $\vec{u}_z$ ($Rot_{\vec{u}_z,\alpha}$), a dilatation of factor R, and finally a translation of vector $\vec{X}_c$.

$$\begin{pmatrix} x_3 \\ y_3 \\ z_3 \end{pmatrix} = R. \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} . \begin{pmatrix} x_2 \\ y_2 \\ z_2 \end{pmatrix} + \vec{X}_c \tag{3.9}$$

As we work on 2D images, a projection along $\vec{u}_z$ is considered, and positions are given by $\vec{X}_i \sim d_{\vec{\beta}}$:

$$\{\vec{X}_i\} = \left\{ \begin{pmatrix} x_3 \\ y_3 \end{pmatrix} \right\} \tag{3.10}$$

In 2D imaging, the uncertainty measurement is expected isotropic and given for each protein by the variance σ_i^2 of position determination, that summarizes the results of the localization procedure including all that can affect its precision such as the photon yield, the stochastic nature of the photon distribution on the CCD, the background noise.... In the real measurement, each protein position is attached to a positioning uncertainty $\{\sigma_i^2\}$ reflecting the variation of the number of photons emitted and of the background intensity at the precise time of its detection (equation (1.2)). In our simulation, we randomly choose the positioning precision of each sampled position (equation (3.10)). The positioning precision distribution should correspond to the imaging conditions. Practically, we use a normal distribution with mean and variance identical to the experimental distribution of positioning precisions observed in an actual experiment in the conditions that we are interested to model.

$$\{\sigma_i^2\} \sim \mathcal{N}(\overline{\sigma^2}, \operatorname{var} \sigma^2) \tag{3.11}$$

Then the simulated localized position of each emitter $\vec{L}_i$ is obtained by adding a random displacement according to their localization uncertainty $\overrightarrow{\delta X_i}$. This is done by drawing the displacement $\overrightarrow{\delta X_i}$ from a normal distribution with standard deviation σ_i and zero mean.

$$\vec{L}_i = \vec{X}_i + \overrightarrow{\delta X_i} \quad | \quad \overrightarrow{\delta X_i} \sim \mathcal{N}(\vec{0}, \sigma_i^2 \mathbb{1}) \tag{3.12}$$

Where $\mathbb{1}$ is the 2×2 identity matrix. This procedure generates for each position a new one sampled from the convolution of the projected protein density and the bivariate normal distribution probability density function with mean $\vec{0}$, and variance $\Sigma = \sigma_i^2 \mathbb{1}$:

$$(x_i, y_i) \sim d^0_{\vec{\beta}} = d_{\vec{\beta}} * \operatorname{pdf}(\mathcal{N}(\vec{0}, \sigma_i^2 \mathbb{1})) \tag{3.13}$$

In 3D imaging, the situation is quite similar, but the positioning precision is different –generally much worse– along the optical axis in the z-direction (Huang et al.[68], Juette et al.[74], Shtengel et al.[116]). The localization uncertainty are of the form $\{\sigma_i|^2_{xy}, \sigma_i|^2_z\}$. We can adopt the same procedure as in 2D, with an anisotropic distribution of error:

$$\vec{L}_i = \vec{X}_i + \overrightarrow{\delta X_i} \quad | \quad \overrightarrow{\delta X_i} \sim \mathcal{N}\left(\vec{0}, \underline{\underline{\sigma_i}}\right) \quad \text{with} \quad \underline{\underline{\sigma_i}} = \begin{pmatrix} \sigma_i|^2_{xy} & 0 & 0 \\ 0 & \sigma_i|^2_{xy} & 0 \\ 0 & 0 & \sigma_i|^2_z \end{pmatrix} \tag{3.14}$$

To account for the presence of proteins that do not belong to the virus, outlier positions uniformly sampled in the neighboring of the generated positions are optionally added to the set. The interest of this method is to generate simulated superresolution images consistent with the model and the positioning uncertainty estimated on the experimental setup by the localization algorithm. A drawback of this non-physical direct sampling is that whenever a source of error is not taken into account in the positioning precisions calculation, it will be ignored as well in the production of simulated images.

3.4 A first attempt to locate particles centers and estimate their radii

A simple idea to estimate the size of the protein cluster is to calculate the mean distance of each one to the center of the cluster. The center of the cluster has to be defined from the collection of the measured protein positions $\vec{X}_i$ in the cluster. The natural point is the mass center of the cluster defined as:

$$\overrightarrow{CM} = \frac{1}{N}\sum_{i=1}^{N}\vec{X}_i \tag{3.15}$$

From this center, the mean radius can be defined as:

$$R_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N}\|\vec{X}_i - \overrightarrow{CM}\| \tag{3.16}$$

In the limit case of a full sphere with infinite sampling, the center of mass and the sphere center coincide, and the mean radius to the center of mass is proportional to the sphere radius:

$$\begin{aligned}\|\vec{X}_i - \overrightarrow{CM}\| &= R\sin\vartheta \quad \text{(spherical coordinates)}\\ R_{\text{mean}} &= \int_{\vartheta=0}^{\pi}\int_{\varphi=0}^{2\pi} R\sin\vartheta.\frac{\sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi}{4\pi} = \frac{\pi}{4}R\end{aligned} \tag{3.17}$$

Estimation of the radius by this method is further investigated by simulation whose typical result is shown in figure 3.2. As expected, estimated values are closer to the actual radius with a higher sampling leading to narrower histograms (compare blue and green curves).

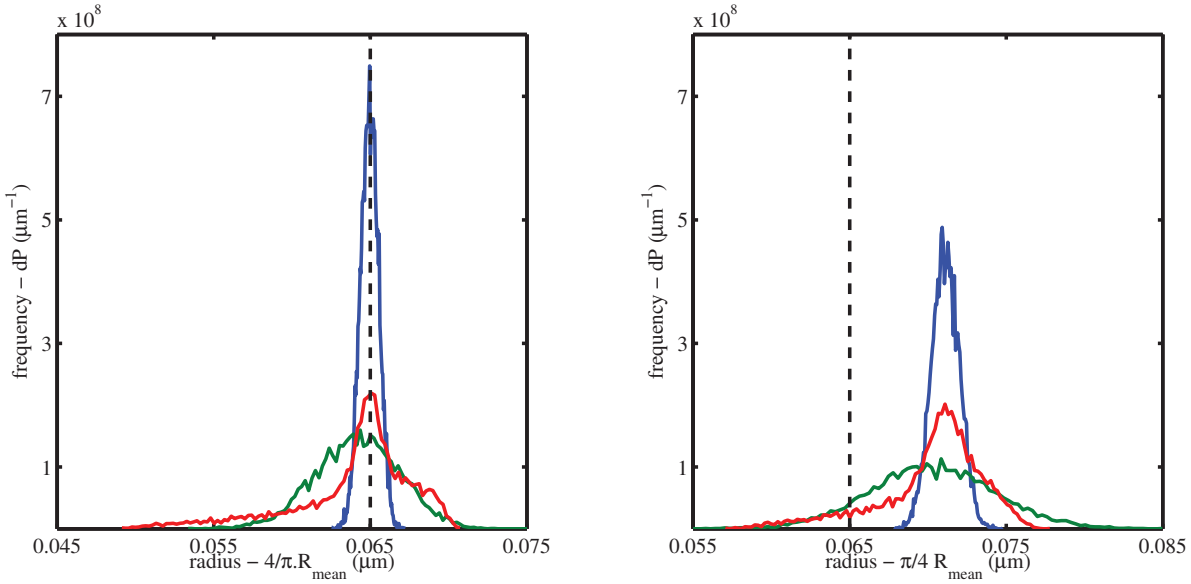
On the contrary when the sampled distribution is an incomplete sphere with a random orientation, the estimation is rather unreliable. The origin of the error is that the center of mass does not coincide any more with the sphere center, and that the projected positions are not any more isotropically distributed around the center of mass (red histograms on figure 3.2).

Including in the simulations a limit in the precision of the localized position of $\sigma \simeq 20\,\text{nm}$, leads to broader and shifted histogram. The bias (the shift) comes from the broadening of the point distribution convolved by a Gaussian function with standard deviation σ. It is neither simple to estimate as long as the typical size of the distribution is of the same order as σ, nor constant for different distributions sizes. This lack of robustness towards positioning precision is the strongest drawback of the naive method. A second limitation is that, being really simple, it is not able to take into account the geometry of the imaged particles and extract other parameter of interest than an effective radius.

However this method points out how to define a characteristic point directly from the measured positions. This reference point, the center of mass, is a natural choice of origin for the coordinates system. In the next section, we define the moments of the distribution, generalizing this idea. We also examine the information that we can extract from their calculation, and their behavior regarding positioning precision.

3.5 Identification of the distribution from its moments

Originally used in the physical study of mechanical systems, the first order moment of a material distribution is often known as the center of mass of the distribution. The mathematical



(a) sampled positions are determined with infinite precision

(b) sampled positions are determined with a standard deviation $\sigma \simeq 20\,\mathrm{nm}$

Figure 3.2 – distribution of radii estimated by the corrected mean radius ${}^4/_\pi \times R_{\mathrm{mean}}$ for various clusters of N positions sampled on spheres with actual radius $R = 56\,\mathrm{nm}$ (black dashed), either complete ($N = 50$ – green, $N = 10^3$ – blue) or truncated with random completions θ and random orientations ϕ uniformly drawn in $(\theta, \phi) \in [\frac{\pi}{2}, \pi] \times [0, \pi]$ (N$= 10^3$ –red). Each histogram is constructed from 10^4 clusters simulations and normalized to have the same unit area.

definition of moments extends to any integers. Moments are a useful mean to characterize a distribution even if, strictly speaking, the collection of all its moments is in general not enough to fully determine it. However when the choice is restricted to a family of distribution, it can be sufficient to identify the correct one. For instance, a Gaussian function is perfectly defined by its two first moments. Furthermore, Hu[67]'s theorem states that a piecewise continuous and compactly supported (non zero only in a finite region of the space) 2D distribution is fully determined by the collection of all its moments and vice versa. Physical structure are generally finite and are thus eligible for the theorem application when modeled by a continuous density. Moments have since then been proposed in image analysis and recognition to perform task such as defining symmetrical images orientation[54] or recognition in a library[45].

The r^{th} moment of the one dimensional distribution D is defined as:

$$M_r = \int x^r D(x)\,\mathrm{d}x = \mathbb{E}\left(x^r\right) \tag{3.18}$$

Central moments are calculated by setting the origin at the mass center and therefore constructed by subtraction of the 1^{st} order moment:

$$\mu_1 \equiv M_1 \tag{3.19}$$

$$\mu_r = \int (x - \mu_1)^r D(x)\,\mathrm{d}x = \mathbb{E}\left((x - \mu_1)^r\right) \tag{3.20}$$

These moments have the interesting feature to be translation invariants. Indeed, they remain unchanged after we apply a translation $T_t : d(x) \mapsto d(x+t)$ to the original distribution:

$$\begin{aligned}\mu_r(T_t \circ d) &= \mathbb{E}\left((x+\lambda-\mathbb{E}\left(x+\lambda\right))^r\right)\\ &= \mathbb{E}\left((x+\lambda-\mathbb{E}\left(x\right)-\lambda)^r\right)\\ &= \mathbb{E}\left((x-\mathbb{E}\left(x\right))^r\right)\\ &= \mu_r(d)\end{aligned} \tag{3.21}$$

The standard and central moments generalizes to random vectors in many dimensions:

$$\begin{aligned}M_{r_1,\ldots,r_n} &= \int x_1^{r_1}x_2^{r_2}\ldots x_n^{r_n}D(x_1,\ldots,x_n)\,\mathrm{d}x_1\ldots\,\mathrm{d}x_n\\ &= \mathbb{E}\left(x_1^{r_1}x_2^{r_2}\ldots x_n^{r_n}\right) \qquad (3.22)\\ \mu_{r_1,\ldots,r_n} &= \int (x_1-\mu_{1,0,\ldots,0})^{r_1}(x_2-\mu_{0,1,0,\ldots,0})^{r_2}\ldots(x_n-\mu_{0,\ldots,0,1})^{r_n}D(x_1,\ldots,x_n)\,\mathrm{d}x_1\ldots\,\mathrm{d}x_n\\ &= \mathbb{E}\left((x_1-\mu_{1,0,\ldots,0})^{r_1}(x_2-\mu_{0,1,0,\ldots,0})^{r_2}\ldots(x_n-\mu_{0,\ldots,0,1})^{r_n}\right) \qquad (3.23)\end{aligned}$$

The invariance of the central moments under any translation is trivially maintained. In the case of random vectors, the p^{th} order moment generally means the collection of all the (central) moments $(\mu_{r_1,\ldots,r_n})$ $M_{r_1,\ldots,r_n}$ verifying $\sum_i r_i = p$. Hence the "list" notations used in the following:

$$\vec{\mu_1} = \begin{pmatrix}\mu_{1,0,\ldots,0}\\ \mu_{0,1,\ldots,0}\\ \vdots\\ \mu_{0,\ldots,0,1}\end{pmatrix} \tag{3.24}$$

$$\underline{\underline{\mu_2}} = \begin{pmatrix}\mu_{2,0,0,\ldots,0} & \mu_{1,1,0,\ldots,0} & \cdots & \mu_{1,0,0,\ldots,1}\\ \mu_{1,1,0,\ldots,0} & \mu_{0,2,0,\ldots,0} & \cdots & \mu_{0,1,0,\ldots,1}\\ \vdots & \vdots & \ddots & \vdots\\ \mu_{1,0,\ldots,0,1} & \cdots & \cdots & \mu_{0,\ldots,0,0,2}\end{pmatrix} \tag{3.25}$$

3.6 Application to the identification of HIV particles

The localization of a protein on a structure is seen as drawing a point from a probability density describing both the original biological structure and the resolution of the measurement process. It is possible to identify the structure features from the measured spatial position of the proteins granted the knowledge of the measurement precision.

From a spatial positions set $\{\vec{X}_i\}$ randomly sampled from a distribution, the following quantities -called empirical estimators or symmetric unbiased estimators- can be defined:

Total number of positions: $$N = \sum_n 1 \tag{3.26}$$

center of mass: $$\overrightarrow{CM} = \tfrac{1}{N}\sum_n \vec{X}_i \tag{3.27}$$

Covariance matrix: $$\underline{\underline{COV}} = \tfrac{1}{N-1}\sum_n (\vec{X}_i - \overrightarrow{CM})(\vec{X}_i - \overrightarrow{CM})^t \tag{3.28}$$

In an experiment consisting of a series of N trials in which a fixed number of point are repetitively and independently drawn from a given probability density, the above quantities will be distributed around the original probability density moments. The name "unbiased empirical estimator" recalls that we expect to have an estimate of moments of the underlying density from the outcome of the measurement (see demonstration in appendix B.3):

$$\mathbb{E}\left(\overrightarrow{CM}\right) = \vec{\mu_1} \tag{3.29}$$

$$\mathbb{E}\left(\underline{\underline{COV}}\right) = \underline{\underline{\mu_2}} \tag{3.30}$$

And "symmetric" was coined to express the independence of the estimator in the order of the measurement. Under this additional property, Halmos[61] demonstrated existence and unicity of the above formula (equations (3.26) to (3.28)) at any order and proved that they are the best (regarding their minimal variance) compared to other unbiased estimator. They later received the name of "h-statistics".

As we manipulate mutually independent measurement, the single point probability density D is enough to fully define the join probability density function D_N for the outcome of any results of the N successive trials:

$$D_N(x_1 \dots x_N) \equiv \prod_{i=1}^{N} D(x_i) \tag{3.31}$$

Which enables us to determine the statistical behavior of the empirical estimators. Indeed empirical estimator for different samples can be seen as a random variable whom we already know the mean, and would like to know variance or other properties.

Limit and scaling with size of the sample N is given by the central limit theorem. When the distribution D has a defined mean and variance, the central limit theorem applies and the asymptotic law for $\overrightarrow{CM}$ is normal with mean given by equation (3.29) and variance scaling in $\propto \mathrm{var}(D)N^{-1}$. When the existence of higher moments is granted for the distribution of proteins in which position are measured, central limit theorem also applies for empirical estimator of higher order (equation (3.30)) with the same asymptotic scaling in N^{-1} for their variance, which means that as the number of measurement grow, the values of the empirical estimators are closer to the distribution moment.

In our case, the chosen model of the protein density inside a budded virus is a continuous distribution taking non zero values only over in a finite region of space. Geometric moments of the distribution are therefore defined at all orders. The image is a set of sampled positions in this distribution from which we can estimate the distribution thanks to their respective empirical estimators. The accuracy of the estimated moments is given by the inverse of their standard deviation and increases as $\sqrt{N}$. Since our family of distributions is parametrized by seven free parameters, namely $(R, \theta, \phi, \alpha, x_c, y_c)$(equation (3.10)), a set of seven moments would provide us all the equations to determine the best distribution. This approach is called the "moments method".

3.6.1 The moment method and its limitations

To illustrate this method, let us look at the first moments for a truncated sphere of radius R, completion θ with its symmetry axis aligned with $\vec{u}_z$ in euclidean coordinates (density given

by equation (3.4)) calculated as detailed in appendix B.1:

$$\vec{\mu_1^\circ} = R.\begin{pmatrix} 0 \\ 0 \\ \cos^2\left(\frac{\theta}{2}\right) \end{pmatrix} \tag{3.32}$$

$$\underline{\underline{\mu_2^\circ}} = R^2.\begin{pmatrix} \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} & 0 & 0 \\ 0 & \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} & 0 \\ 0 & 0 & \frac{\sin(\frac{\theta}{2})^4}{3} \end{pmatrix} \tag{3.33}$$

We note that redundancy in the expression of the second order moment, $\mu^\circ_{0,2,0} = \mu^\circ_{2,0,0}$, comes from the symmetries of the modeled distribution (x is equivalent to y in the truncated sphere) and the resulting parity of the density $D(-x,\ldots) = D(x,\ldots)$, gives the diagonal form.

If we were measuring budded viruses aligned perpendicularly to the microscope axis (tilt of $\phi = \frac{\pi}{2}$ or projection in the x direction), the first two central moments would provide us three expressions of R and θ:

$$\begin{cases} \mu^\circ_{0,0,1} &= R\cos^2(\frac{\theta}{2}) \\ \mu^\circ_{0,2,0} &= R^2\frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} \\ \mu^\circ_{0,0,2} &= R^2\frac{\sin^4(\frac{\theta}{2})}{3} \end{cases} \tag{3.34}$$

this overdetermined system of non linear equations has to be solved to give an expression of R and θ in terms of the distribution central moments. In this case the square of the first line and the last lines leads to:

$$\begin{cases} \theta_* &= 2\tan^{-1}\left(\sqrt{3\frac{\mu^\circ_{0,0,2}}{(\mu^\circ_{0,0,1})^2}}\right) \\ R_* &= \mu^\circ_{0,0,1}\sqrt{1+3\frac{\mu^\circ_{0,0,2}}{(\mu^\circ_{0,0,1})^2}} \end{cases} \tag{3.35}$$

It is clear that the task is not an easy one as the size of the system increases with the number of parameters. Furthermore, we do not have access to the moments of the distribution, but rather to their estimation from our sampling. We know that, due to sampling effect, the values obtained from the empirical estimators fluctuate around the real distribution moments. In a such overdetermined system, the fluctuations are likely to lead to incompatible equation.

The moments methods then turn to finding out the optimal parameter set (R_*, θ_*) for the sets moment equations regarding the observed value:

$$\begin{cases} z_{\overrightarrow{CM}} &= R_*\cos^2\left(\frac{\theta_*}{2}\right) \\ \underline{\underline{COV}}(y,y) &= (R_*)^2\frac{(2+\cos\theta_*)\sin^2(\frac{\theta_*}{2})}{3} \\ \underline{\underline{COV}}(z,z) &= (R_*)^2\frac{\sin^4\left(\frac{\theta_*}{2}\right)}{3} \end{cases} \tag{3.36}$$

In other words we have to fit the theoretical moments to the values measured with the empirical estimators. To define this optimum we need a measure of the fitness of a parameter set. The classic least squares method (see appendix A) is well suited in the case of normally distributed fluctuations as we expect from central limit theorem. We expect the observed moments values V^{obs} to be distributed around the true moments values V_0 according to:

$$P(V^{\text{obs}}) \propto e^{-\frac{(V^{\text{obs}}-V_0)^t.\Sigma_0^{-1}.(V^{\text{obs}}-V_0)}{2}} \tag{3.37}$$

With V_0 the true values and Σ the covariance matrix of the fluctuations. Then the Mahalanobis distance:

$$d(V^{\text{fit}}, V^{\text{obs}}) \equiv (V^{\text{obs}} - V^{\text{fit}})^t.(\Sigma^{\text{fit}})^{-1}.(V^{\text{obs}} - V^{\text{fit}}) \tag{3.38}$$

is appropriate to estimate how far are the values of V^{fit} given by the tested parameters from the observed values in V^{obs} regarding the expected fluctuations Σ^{fit}. And the least squares method consists in finding the optimal parameter that produces the values minimizing the Mahalanobis distance. We need to know the amplitude of the fluctuations and correlations –the theoretical "error bar" of our measurement. Those are the variances of the estimations given by equations (3.27) and (3.28). In the case of our example described by equation (3.36), direct calculations using the independence and identically distributed properties of the sampled positions gives:

$$\Sigma^{\text{fit}} = \text{var}\begin{pmatrix} z_{\overrightarrow{CM}} \\ \underline{\underline{COV}}(y,y) \\ \underline{\underline{COV}}(z,z) \end{pmatrix} \simeq \frac{1}{N}\begin{pmatrix} \mu^\circ_{0,0,2} & \mu^\circ_{0,2,1} & \mu^\circ_{0,0,3} \\ \mu^\circ_{0,2,1} & \mu^\circ_{0,4,0} - (\mu^\circ_{0,2,0})^2 & \mu^\circ_{0,2,2} - \mu^\circ_{0,2,0}\mu^\circ_{0,0,2} \\ \mu^\circ_{0,0,3} & \mu^\circ_{0,2,2} - \mu^\circ_{0,2,0}\mu^\circ_{0,0,2} & \mu^\circ_{0,0,4} - (\mu^\circ_{0,0,2})^2 \end{pmatrix} \tag{3.39}$$

Details as well as the expression of the coefficient as function of completion and radius are in appendix B. As shown by equation (3.39), fluctuations due to sampling of the empirical estimators of central moment are function of distribution higher central moments of the original distribution, and the asymptotic scaling of their covariance is $\propto N^{-1}$, thus consistent with the central limit theorem prediction.

Unfortunately, our distribution sampling is submitted to an additional noise since the positioning uncertainty imply a error of order σ_i on each x_i. In first approximation this error can be regarded as isotropic and identical for all positions, normally distributed with a mean standard deviation given by the mean variance:

$$\bar{\sigma} = \sqrt{\frac{1}{N}\sum_i \sigma_i^2} \tag{3.40}$$

The new random variable for the measure is the sum of the previous one and the error. Hence the new reference distribution probability density function is obtained by convolution of the previous one with the probability density function of the error. Equation (3.39) holds with the central moments of the new reference distribution [1]:

$$\mu^\circ_{0,2,0} \longrightarrow \mu^\circ_{0,2,0} + \bar{\sigma}^2 \tag{3.41}$$

$$\mu^\circ_{0,4,0} \longrightarrow \mu^\circ_{0,4,0} + 3\bar{\sigma}^4 + 6\bar{\sigma}^2\mu^\circ_{0,2,0} \tag{3.42}$$

And the moment method turns into a non linear optimization problem to find (R_*, θ_*) minimizing equation (3.38).

In the general case, we need the moment of the distribution of a truncated sphere with any orientation. Considering that any possible orientation (ϕ, α) is generated from the reference truncated sphere by a linear transformation as expressed in equation (3.9), we can derive the expression of the moments for any of those generated distributions using integration by

1. Note that variances add, but not fourth order moments. The cumulants of the function add at all order and can then be expressed in terms of central moments

substitution as showed in appendix B.2. If J is the Jacobian matrix of the transformation in euclidean coordinates:

$$J = \begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix} . \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \tag{3.43}$$

We have the following relation between the moments of the centered distribution after transformation:

$$\begin{aligned} \vec{\mu_1} &= J.\vec{\mu_1^\circ} \\ &= R. \begin{pmatrix} -\cos(\alpha)\cos(\theta/2)^2\sin(\phi) \\ -\cos(\theta/2)^2\sin(\alpha)\sin(\phi) \\ \cos(\theta/2)^2\cos(\phi) \end{pmatrix} \end{aligned} \tag{3.44}$$

As well as:

$$\begin{aligned} \underline{\underline{\mu_2}} &= J.\underline{\underline{\mu_2^\circ}}.J^t \\ &= R^2 Rot_{\vec{u}_z,\alpha} \begin{pmatrix} \mu_{2x}^\circ\cos^2\phi + \mu_{2z}^\circ\sin^2\phi & 0 & (\mu_{2x}^\circ - \mu_{2z}^\circ)\cos\phi\sin\phi \\ 0 & \mu_{2x}^\circ & 0 \\ (\mu_{2x}^\circ - \mu_{2z}^\circ)\cos\phi\sin\phi & 0 & \mu_{2z}^\circ\cos^2\phi + \mu_{2x}^\circ\sin^2\phi \end{pmatrix} Rot_{\vec{u}_z,\alpha}^t \\ \text{with: } \mu_{2x}^\circ &= \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} \quad \text{and} \quad \mu_{2z}^\circ = \frac{\sin^4(\frac{\theta}{2})}{3} \end{aligned} \tag{3.45}$$

Rules to calculate the moments for any possible orientations and radius of the truncated sphere from the moment of the original vertical unit sphere are akin the rules of tensors transformation.

Again, the last step is to find the minima of the non linear function of the problem parameter obtained in equation (3.38). This is a non trivial optimization problem given the number of parameters with no analytical solution available. As the measurement noise increases there is no guaranty that a well defined minimum is to be found. Teh and Chin[126] reported the general rule that the sensitivity to the noise of image moments increases with their orders. We will therefore limit ourselves to the use of the first moments and intent to limit the number of parameters to eventually optimize.

3.6.2 The first moments to center and orient the distribution

As we have seen, the position of the center of mass of the distribution can be localized from N measured positions (equation (3.29)) sampled with positioning uncertainty $\bar{\sigma}$ up to a precision given by:

$$\text{var}\left(\overrightarrow{CM}\right) = \frac{1}{N}.(\underline{\underline{\mu_2}} + \bar{\sigma}^2\mathbb{1}) \tag{3.46}$$

Where $\underline{\underline{\mu_2}}$ is the covariance matrix of the sampled distribution. This point is a natural reference and in the following we use it as the origin of the coordinate system instead of the geometric center of the sphere. This is a direct application of the previously introduced "moment method" to free ourself from the translation in equation (3.9) using the more robust moment. In a 3D (resp. 2D) super-resolution image this identification diminishes the number of degrees of freedom by 3 (resp. by 2).

A second parameter identification can be performed analytically using the already mentioned symmetries of the truncated sphere distribution. Prior to the rotations that describes a change of point of view, the second order central moment matrix is diagonal equation (3.33). If noise is added in measurement, the variances of the noise and of the distribution add so that this property is conserved as long as the noise covariance matrix is itself diagonal. We assume that the positioning uncertainty is isotropic in the x, y dimensions and so, that it preserves all the symmetries in this plane. In the absence of optical distortions, the noise on different spatial dimensions uncorrelated and the second order moments observed will remain diagonal.

Let us now consider a different coordinate system: the covariance matrix is symmetric with real coefficients. It can therefore be diagonalized using an orthogonal transformation. By identification with the transformation rule[2] (equation (3.45)), this orthogonal transformation corresponds to the inverse of the two successive rotations of angle ϕ and α in equation (3.9) and the eigenvector of the covariance matrix associated to the non degenerated eigenspace gives the direction of symmetry axis of the distribution.

In 3D imaging, it is possible to recover the initial orientation of the distribution by diagonalizing the covariance matrix. The covariance matrix of the distribution is real and positive, it defines the quadratic approximation of the distribution around its center of mass which is an ellipsoid. The axes of the ellipsoid are along or perpendicular to the symmetries axes of the distribution when they exist. With a higher degree of symmetry, the eigenstates are degenerated which means that several orthogonal bases are equivalent.

— In the case of a complete sphere all the directions are equivalent and we have neither lost nor gained information.

— In the case of a truncated sphere, the symmetry axis is given by the eigenvector associated to eigenspace of dimension one, and all the bases of the orthogonal plane are equivalent. However we have to choose between two opposite orientations for the z direction which are not equivalent: either toward the top of the sphere or towards the hole.

In 2D imaging, after the rotation, the projection in the focal plane removes the information we had about the z direction. The successive transformations in equation (3.45), leave the xy upper block of the covariance matrix diagonal until the α rotation. Several cases are possible:

— $\phi = 0$: the truncated sphere is viewed from the top and the cylindrical symmetry becomes a circular symmetry in the focal plan. The eigenvalues are identical and no information is obtained.

— $\phi \neq 0$: The eigenvectors of the 2D-covariance matrix are respectively aligned with and perpendicular to the projection of the 3D-axis of symmetry. The largest eigenvalue, corresponding to the greater spatial variance, is associated to the direction perpendicular to the symmetry axis whereas the smallest is associated with the direction of the projection of the 3D-symmetry axis. α is determined modulo π.

2. and regarding any rotation around the symmetry axis of the distribution as equivalent to the identity

We can look for the rotation that diagonalizes a real symmetric positive definite matrix:

$$\begin{aligned}\underline{\underline{COV}}|_{x,y} &= \begin{pmatrix} a & c \\ c & b \end{pmatrix} \\ &= \begin{pmatrix} v_+ & v_- \end{pmatrix} . \begin{pmatrix} \lambda_+ & 0 \\ 0 & \lambda_- \end{pmatrix} . \begin{pmatrix} v_+^t \\ v_-^t \end{pmatrix} \\ \lambda_\pm &= \frac{1}{2}\left(a+b \pm \sqrt{(a-b)^2+4c^2}\right) \\ v_\pm &\propto \begin{pmatrix} -2c \\ a-b \mp \sqrt{(a-b)^2+4c^2} \end{pmatrix} \quad (c \neq 0)\end{aligned} \tag{3.47}$$

The eigenvector associated to the smallest eigenvalue v_- makes an angle $\alpha[\pi]$ with $\vec{u}_x$:

$$\alpha(a,b,c)\,[\pi] = \begin{cases} \tan^{-1}(a-b+\sqrt{(a-b)^2+4c^2}\,,\,-2c) & c \neq 0 \\ \frac{\pi}{2} & c=0\,,\,a>b \\ 0 & c=0\,,\,a \leq b \end{cases} \tag{3.48}$$

The distribution can then be oriented with its symmetry axis along $\vec{u}_x$ by a rotation of angle $-\alpha$ around its mass center (the origin).

The precision of the orientation can be further investigated by assuming a small noise and uses the linear propagation of error around the estimated value. In this approximation, the fluctuations of the matrix parameter around the expected values are approximated by a normal law with covariance matrix $\Sigma_{a,b,c}$. Furthermore the function that associate α to (a,b,c) is differentiable around the evaluation point $J_\alpha = \vec{\nabla}\alpha|_{(a,b,c)}$. Under this conditions, we have (appendix A.4):

$$\operatorname{var}\alpha = J_\alpha . \Sigma_{a,b,c} . J_\alpha^t \tag{3.49}$$

This calculation can be restricted to a single case without loosing generality. Consider that we are given a set of sampled positions from the distribution and that we define the most probable orientation of the distribution symmetry axis using equation (3.48). A collective rotation of all the points together moves the actual symmetry axis and the estimation altogether, so that the angular difference is not affected. We can arbitrary set the symmetry axes of the distribution along $\vec{u}_x$ and look at the fluctuation of the estimated angle α around zero which is the limit $c \to 0$, $a \to \lambda_-$, $b \to \lambda_+$. This simplify the above equation into:

$$\operatorname{var}\alpha = \frac{\operatorname{var}(c)}{(\lambda_+ - \lambda_-)^2} \tag{3.50}$$

The direct calculation of $\Sigma_{a,b,c}$ would have required the explicit calculation of the second and fourth order central moments of the reference distribution (see equation (3.39)). The first is a 3×3 symmetric matrix with 6 coefficients and the second a 3^4 tensor with 21 free coefficients. Each coefficient calculated is a function of the parameters (R,θ,ϕ), still unknown as we try to orient the distribution.

According to the h-statistic, We have:

$$\begin{aligned}\operatorname{var}(c) &= \operatorname{var}\left(\underline{\underline{COV}}(x,y)\right) \\ &= \frac{1}{N}\left(\mu_{2,2,0} + \frac{N-2}{N-1}\mu_{1,1,0}^2 - \frac{1}{N-1}\mu_{0,2,0}\mu_{2,0,0}\right) \\ &\simeq \frac{1}{N}\left(\mu_{2,2,0} + \frac{N-2}{N-1}\mu_{1,1,0}^2\right)\end{aligned} \tag{3.51}$$

Where we have neglected a contribution of order N^{-2} in front of the other of order N^{-1}. Using the linear relations between the reference distribution and the one obtained through linear transformation (equation (3.45)), we carry on:

$$\begin{aligned}\operatorname{var}(c) &= \frac{1}{N}\left(\mu^\circ_{2,2,0}\cos^2\phi + \mu^\circ_{2,0,2}\sin^2\phi \right.\\ &\qquad \left. + \frac{N-2}{N-1}(\mu^\circ_{1,1,0}\cos\phi - \mu^\circ_{0,1,1}\sin\phi)^2\right)\\ &= \frac{1}{N}\left(\mu^\circ_{2,2,0}\cos^2\phi + \mu^\circ_{2,0,2}\sin^2\phi\right)\\ &\simeq \frac{\mu^\circ_{2,0,0}\left(\mu^\circ_{2,0,0}\cos^2\phi + \mu^\circ_{0,0,2}\sin^2\phi\right)}{N}\\ &= \frac{\lambda_+\lambda_-}{N}\end{aligned} \tag{3.52}$$

Where we have taken benefit of the axis symmetry of the reference distribution (equation (3.33)). Eventually we approximated the values of the fourth order moment coefficient by the square of the second order moment for which we have an estimation at hand: those are the product of the diagonal terms of $\mu_2|_{x,y}$ according to equation (3.45). This approximation is quite reasonable as shows the plot of their respective ratios figure 3.3.

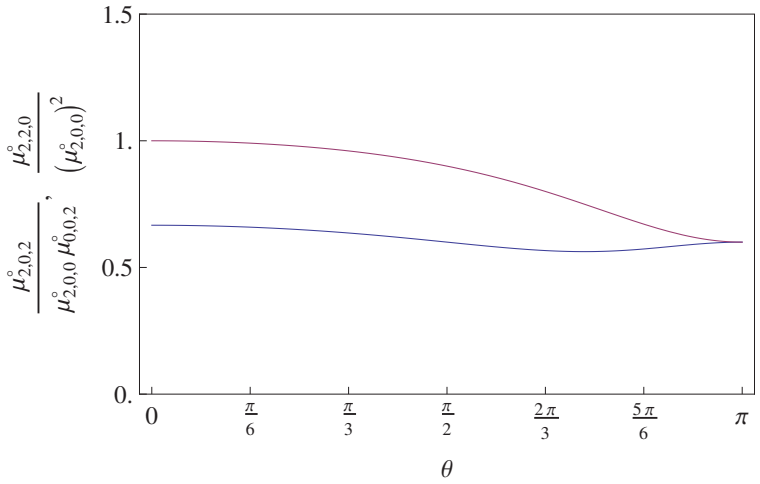


Figure 3.3 – ratios of the fourth and second order central moments of the truncated unit sphere reference distribution depending on the completion (maximum polar angle θ). $\frac{\mu^\circ_{2,2,0}}{(\mu^\circ_{2,0,0})^2}$ (blue) and $\frac{\mu^\circ_{2,0,2}}{\mu^\circ_{2,0,0}\mu^\circ_{0,0,2}}$ (red) are plot according to their calculated values in annexe. The ratios are of order 1 and do not vary much, which justifies the approximation used in equation (3.52)

This gives an estimation of the precision in the determination of the orientation α of the symmetry axis using the projection of a 3D object with cylindrical symmetry:

$$\operatorname{var}\alpha \simeq \frac{\lambda_+\lambda_-}{N(\lambda_+ - \lambda_-)^2}, \quad N \gg 1 \tag{3.53}$$

The strong advantage of this estimation is that it does not depends on the unknown parameters of the distribution, but only on the known eigenvalues of the covariance matrix of the sampled

position. Furthermore, adding a uncorrelated Gaussian noise to the distribution preserving the cylindrical symmetry in the (x, y) plane:

$$\Delta\vec{X} \sim \mathcal{N}(0, \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} \sigma_x^2 & 0 & 0 \\ 0 & \sigma_x^2 & 0 \\ 0 & 0 & \sigma_z^2 \end{pmatrix} \tag{3.54}$$

the expression for the second and fourth order moments transforms into:

$$\underline{\underline{\mu_2}} = \underline{\underline{\mu_2^\circ}} + \Sigma \tag{3.55}$$

$$\begin{aligned} \mu_{2,0,2} &= \mu_{2,0,2}^\circ + \sigma_x^2\sigma_z^2 + \sigma_x^2\mu_{0,0,2}^\circ + \mu_{2,0,0}^\circ\sigma_z^2 \\ &\simeq (\mu_{2,0,0}^\circ + \sigma_x^2)(\mu_{0,0,2}^\circ + \sigma_z^2) \\ &= \mu_{2,0,0}\mu_{0,0,2} \end{aligned} \tag{3.56}$$

$$\begin{aligned} \mu_{2,2,0} &= \mu_{2,2,0}^\circ + \sigma_x^2\sigma_x^2 + \sigma_x^2\mu_{2,0,0}^\circ + \mu_{0,2,0}^\circ\sigma_x^2 \\ &\simeq (\mu_{2,0,0}^\circ + \sigma_x^2)^2 \\ &= (\mu_{2,0,0})^2 \end{aligned} \tag{3.57}$$

So that the approximation holds and the noise contribution is also taken into account into the estimation of the uncertainty.

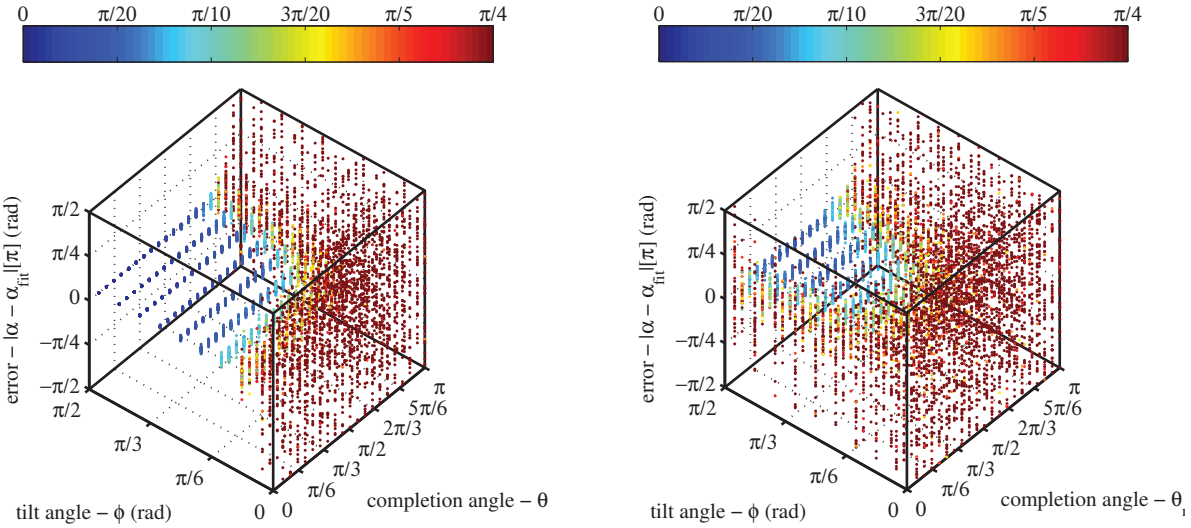
The variance diverges when the eigenvalues become identical indicating that the projection has a circular symmetry around its center of mass and all axis are equivalent. Figure 3.4 compares the real error to the estimated variance for various simulated geometries and additional noise. It shows that we have indeed a good statistical estimator of the orientation of a spherical particle in the focal plan. Furthermore we are able to estimate the statistical error including the effect of noise and sampling. The limitations of the sampling appears on figure 3.4a as the tilt angle decreases, the variance increases as the asymmetry of the particle becomes inappreciable with to few positions. The effect is of course worsen by increasing completion that lowers the actual asymmetry until it disappears (α does not makes sens anymore for $\theta = \pi$). As the strength of noise is increased, the error does too. The strongest impact is seen for low completions where the size of the particle is lower than the typical length of the spreading due to noise and that the organization of the sampling positions is completely shattered.

To fully determine the orientation angle $\alpha[2\pi]$ the two first order moments are not enough. The third order moment of the distribution has to be calculated because it is the smaller order at which the asymmetry due to the sphere incompleteness might appear in the centered and oriented distribution. Cylindrical symmetry of the distribution leads to a simple expression given in appendix B.1 with all non zero coefficients equal. Transformation through tilt rotation and $\alpha = 0$ leads to:

$$\mu_3|_{x,y}(i,j) = \begin{cases} \mu_3^\circ \sin\phi & i = 2 \text{ or } j = 2 \\ 3.\mu_3^\circ \sin\phi\cos^2\phi & i = 3 \\ 0 & \text{else} \end{cases} \tag{3.58}$$

The non-zero coefficients are plotted on figure 3.5

A rotation of angle $\alpha = \pi$ rad is equivalent here to the symmetry $x \to -x$, which turns each non zero term of $\mu_3|_{x,y}$ into its opposite ($x^3 \to -x^3$ and $xy^2 \to -xy^2$). We therefore use the sign of either $\mu_{3,0,0}$ or $\mu_{1,2,0}$ when the point distribution is oriented to fully determine



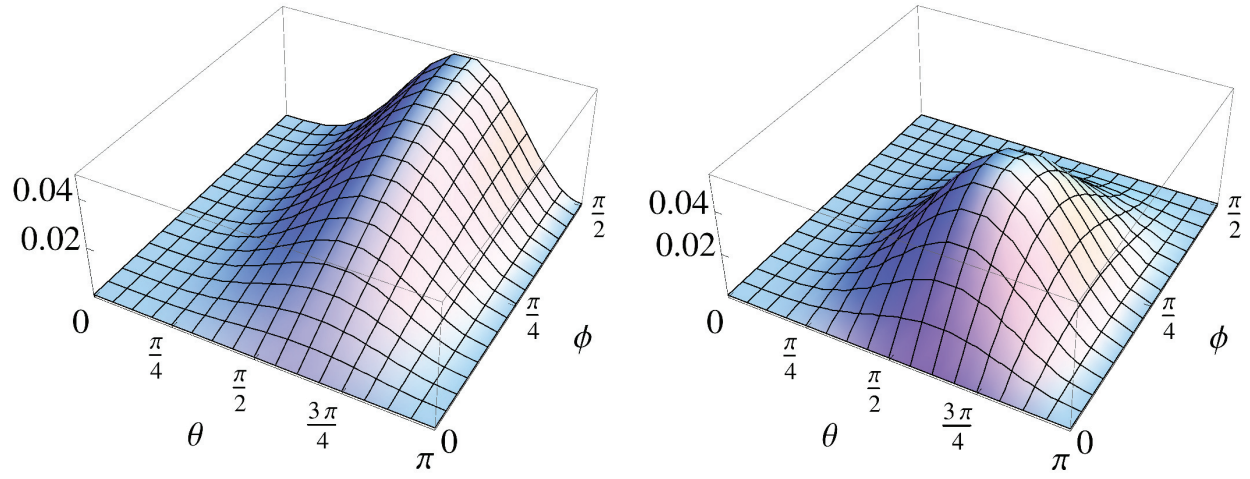
(a) infinite positioning precision

(b) additional isotropic Gaussian noise with a std. of half the sphere radius $\Sigma = (\frac{R}{2})^2 \mathbb{1}$

Figure 3.4 – Estimation of the orientation in focal plan α of truncated spheres depending on their 3D-orientations (each point is a different sampling, 40 per vertical line). The sphere are sampled with $N = 1500$ positions and planar angle is found using the second order moment according to equation (3.48). Colors show the estimations of the variance according to equation (3.53) in good agreement with the actual spread of the found values

α $[2\pi]$ rad. As sampling fluctuations and noise lead to fluctuations of the estimated values, the sign might be altered where the expectation is of the order of the variance, that is too close to zero. Both coefficients are close to zero either at small completion ($\theta < \frac{\pi}{2}$) or at nearly full completion ($\theta \simeq \pi$) (figure 3.5). However, they differs significantly regarding the tilt angle ϕ: $\mu_{1,2,0}$ takes larger values that $\mu_{3,0,0}$ for strong tilt ($\phi > \pi/4$) whereas $\mu_{3,0,0}$ dominates at low tilt angles. When the sign of both estimate are not identical, we select the sign associated to the higher absolute value $\max(|\mu_{3,0,0}|, |\mu_{1,2,0}|)$ –presumably $\mu_{3,0,0}$ when $\phi \ll \pi/2$ or $\mu_{1,2,0}$ in the other case. We limit this way the risk of error.

Using successively the three first moments of the 2D points distribution and analytical calculations, we are thus able to center and orient the image of an individual particle with cylindrical symmetry without setting up a complex optimization scheme as a native moments method would require. Furthermore we are able to estimate the precision of our estimation at each step. Doing so, we diminish the degrees of freedom of the single particle reconstruction problem by a factor 2. From the six free variables $\{R, \theta, \phi, \alpha, x_c, y_c\}$ of the initial parametrization of the particle density projection, those that correspond to positioning and orientation in the 2D plane have been extracted: we found out the reference position of pictures rotated and displaced. Only three parameters are left unknown at this point: $\{R, \theta, \phi\}$. They correspond to physical change of the scene imaged, either the point of view (ϕ) or the subject (R, θ). As already stated, we could use the moments method to find estimates of these parameters. We would face the dilemma of restricting ourself to the few first moments which leads to simpler optimization task but an important lost of the information contained in the positions distribu-



(a) $\mu_{1,2,0}$ (b) $\mu_{3,0,0}$

Figure 3.5 – Third order moment non zero coefficients for a sphere of completion θ, tilted with an angle ϕ, and symmetry axis aligned along $\vec{u}_x$. As both coefficients are positive, we deduce that the projection of the top of the sphere is in $x < 0$ which is the reference orientation we chose ($\alpha = 0$)

tion, or introducing higher moments rapidly increasing the amount of calculation of the target function to minimize and the assumption about the noise (for instance to generate the larger covariance matrix of the estimates Σ_{fit}). Or we could instead use a different method that best fit our needs.

3.6.3 Beyond the moments method: the maximum likelihood approach

The maximum likelihood approach, largely developed and popularized by R.A. Fisher[43], is a general method for estimating the probability distribution that best corresponds to the measured data. The choice criterion is the likelihood function that is the probability to obtain the measured data given the probability distribution. If each element of the ensemble of the probability distribution can be identified by a parameter $\vec{\beta}$, the likelihood function writes:

$$L(\vec{\beta}) = \text{P}(\text{measure}|\vec{\beta}) \tag{3.59}$$

or in case of a continuous model with a probability density function $p_{\vec{\beta}}$ (Pawitan[108]):

$$L(\vec{\beta}) = p_{\vec{\beta}}(\text{measure}) \tag{3.60}$$

And the maximum likelihood estimator of the parameter β is the maximizer of the likelihood function or, equivalently of the loglikelihood function that is $S(\beta) = \log L(\beta)$:

$$\beta^* = \underset{\beta}{\text{argmax}} \log L(\beta) \tag{3.61}$$

The attractiveness of the likelihood comes from the simplicity of the derivation of the likelihood function from the model and thus setting the optimization problem. For instance, let us consider the density of protein $D_{\vec{\beta}}$ that we adopted to model the viral structure. To build the likelihood function, all we need is the probability to detect a protein at position x_i, y_i. As described in section 3.3, the measurement process involved in superresolution imaging can be described by sampling position in the density of protein blurred by the uncertainty of the measure.

$$d^{\text{mes}}_{\vec{\beta}}(x_i, y_i) = \iint \frac{d_{\vec{\beta}}(x, y)}{2\pi\sigma_i^2} e^{-\frac{(x-x_i)^2+(y-y_i)^2}{2\sigma_i^2}} \,\mathrm{d}x\,\mathrm{d}y \tag{3.62}$$

Each localization event being independent of the previous ones, the density of probability for the whole set of measurements is obtained as a simple product of each individual probability density:

$$L(\vec{\beta}) = p_{\vec{\beta}}(\{\vec{X}_i\}) = \prod_i d^{\text{mes}}_{\vec{\beta}}(x_i, y_i) \tag{3.63}$$

$$\log L(\vec{\beta}) = \sum_i \log d^{\text{mes}}_{\vec{\beta}}(x_i, y_i) \tag{3.64}$$

The only required calculation is the convolution of the model density of emitters with the Gaussian. This corresponds to take into account the effect of the measurement process on the outcome of the experiment. The likelihood function otherwise only require the evaluation of the probability density at each measured position and multiplication. Even if the maximization of the likelihood function can reveal itself a much harder problem, its construction is straightforward.

Besides this simplicity, the loglikelihood estimator was designed to take benefit of all the informations about the parameters that are contained in the data, according to the "minimal sufficiency" property of the likelihood function put forward by R.A. Fisher[43]. Under sufficient (but not necessary) hypothesis formulated by Wald[132], the maximum likelihood estimation can be proven consistent, which means that the estimation is equal to the true value with probability one as the number of analyzed measures increase towards infinity. Those sufficient conditions ensure that for all possible sample but a subset with null probability, the likelihood function is continuous with respect to the parameters, has an unique and well identified maximum on the compact parameter space and does not diverge to $+\infty$ on the parameter space. Cases where the loglikelihood shows several local maxima are more complicated, the global maximum is not necessarily a consistent estimator.

When the maximum likelihood achieve consistency and if the likelihood function is regular enough around the maximum, it is asymptotically normally distributed with the minimal variance achievable for an unbiased estimator, the Cramer-Rao lower bound. The inverse of the asymptotic variance of the maximum likelihood estimation is given in this case by the mean second derivative of the likelihood function, called the Fisher information matrix (Pawitan[108]):

$$(\operatorname{var}\vec{\beta}^*)^{-1} \overset{N\to\infty}{\simeq} \mathcal{I}(\vec{\beta}_0) = \mathbb{E}_{\vec{\beta}_0}\left(I(\vec{\beta}^*)\right) = \mathbb{E}_{\vec{\beta}_0}\left(-\left.\frac{\partial^2 \log L(\vec{\beta})}{\partial \vec{\beta}^2}\right|_{\vec{\beta}^*}\right) \tag{3.65}$$

Where $\mathbb{E}_{\vec{\beta}_0}(.)$ stands for the conditional expectation on $\vec{\beta}_0$, that is, over all the different samplings that can be realized from the density parametrized by $\vec{\beta}_0$. $I(\vec{\beta}^*)$ is generally called the

observed Fisher information matrix as it is attached to the curvature of the loglikelihood function around its maximum for a given sampling. Efron and Hinkley[37] reported that the observed Fisher information matrix was a suited choice for accessing the precision of the estimation in the regular case.

The maximum likelihood might be used to determine, from the superresolution image, the most suited geometry for the protein coat among the truncated sphere. How does this results apply to our description of the measure of protein density using surperresolution microscopy? We used the normal law to model the measure precision. It is known among the regular distribution that satisfy the consistency conditions given by Wald[132]. We can expect that the distribution obtained by convolving the initial protein distribution with the Gaussian inherits the smooth and differentiable properties of this later. However, the existence of a unique maximum for the likelihood function is not granted. Indeed the parametrization we have chosen already contains degenerated cases as different parameters lead to the same density. For instance rotations do not affect the complete sphere, and for two different tilts ϕ_1 and ϕ_2 the observable density is the same: $D_{R,\theta=\pi,\phi_1} \equiv D_{R,\theta=\pi,\phi_2}$. This does not matter when the maxima correspond to physically equivalents distribution, but might call for additional assumptions if not.

3.6.4 Likelihood computation for the truncated sphere model

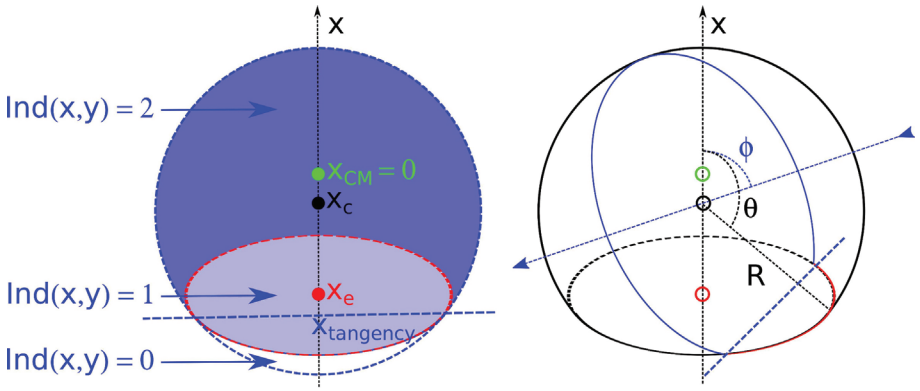


Figure 3.6 – The axial projection of uniform density on a truncated sphere. The three value domains of the indicator function (left) for the plane projection of a constant density laying on the incomplete spherical shell (parameters: R, θ – right in black) parallel to a given projection axis (parameter: ϕ – right in blue dashed arrow). Remarkable points (empty dots) and their projections (filled dots) are shown: mass center (green), sphere center (black), and border center (red). Note that the projection may change the distance measured between them. The expression of the indicator function is deduced by combining the indicator functions of the following regions: the circle that is the projected edge of the sphere (dashed blue – left) and the ellipse (dashed red – left) that is the projection of the border and the strait line that links their tangency points (dashed blue – left)

We computed the likelihood function of a measurement as follows. A mathematical difficulty arises from the $2D$ surface chosen to model the distribution. When a parametrization of the surface of the form $z = f(x, y)$ is available, the element of integration on the surface is:

$$\mathrm{d}S = \sqrt{1 + \|\vec{\nabla} f\|^2}\, \mathrm{d}x\, \mathrm{d}y \tag{3.66}$$

Using the obvious parametrization $z(x, y)$ of an hemisphere of radius R in Cartesian coordinates, we deduce its projected density:

$$\begin{aligned} z &= \sqrt{R^2 - (x^2 + y^2)} \\ \mathrm{d}S &= \sqrt{1 + \frac{x^2 + y^2}{R^2 - (x^2 + y^2)}}\,\mathrm{d}x\,\mathrm{d}y \\ &= \frac{R^2}{\sqrt{R^2 - (x^2 + y^2)}}\,\mathrm{d}x\,\mathrm{d}y \\ \Rightarrow d^\circ(\rho) &= \frac{1}{\sqrt{1 - (\rho/\mathrm{R})^2}} \end{aligned} \tag{3.67}$$

Where $\rho = \sqrt{x^2 + y^2}$ is the distance to origin in polar coordinates. It appears that the measure d° diverges close to the boundaries ($\rho = R$) where the surface normal is parallel to the (x, y) plane and the unit surface on the sphere is projected on a line.

An analytical expression of the planar projection of the constant density on the truncated sphere $\mathcal{S}_t$ with arbitrary orientation is obtained by combining various portions of the sphere of radius R which can be parametrized as shown below. The different portions can be encoded in a simple indicator function. This indicator function then traduces the completion: its value at a given point corresponds to how often the projection axis crosses the surface on which proteins are lying. Hence its values are in $\{0, 1, 2\}$ as shown on figure 3.6. And the domain of each of those values is given by the intersections of one circle, one ellipse and a strait line. Say, the projected contour of the sphere (Ind_1) centered in x_c, the one of the completion edges (Ind_2) centered in x_e and the line that joins their tangency points (Ind_3) of abscissa $x_{tangency}$. Those equation are easy to parametrize in the axis system described in the above section:

$$d_p[\vec{\beta}](\vec{r}) = \frac{1}{\sqrt{1 - (\rho/\mathrm{R})^2}} \times \mathrm{Ind}[\vec{\beta}](\vec{r}) \tag{3.68}$$

$$\mathrm{Ind}[\vec{\beta}](x, y) = 2.\mathrm{Ind}_1.\mathrm{Ind}_3.(1 - \mathrm{Ind}_2) + \mathrm{Ind}_2 \tag{3.69}$$

With the indicator functions given by:

$$\begin{aligned} \mathrm{Ind}_1[\vec{\beta}](x, y) &= \begin{cases} 1 & \text{if} \quad \|y\| \leq \mathrm{R}\sqrt{1 - \frac{(x - x_c)^2}{\mathrm{R}^2}} \\ 0 & \text{else.} \end{cases} \\ \mathrm{Ind}_2[\vec{\beta}](x, y) &= \begin{cases} 1 & \text{if} \quad \|y\| \leq \mathrm{R}\sin\theta\sqrt{1 - \frac{(x - x_e)^2}{\mathrm{R}^2 \sin^2\theta \cos^2\phi}} \\ 0 & \text{else.} \end{cases} \\ \mathrm{Ind}_3[\vec{\beta}](x, y) &= \begin{cases} 1 & \text{if} \quad x \geq x_{tangency} \\ 0 & \text{else.} \end{cases} \end{aligned} \tag{3.70}$$

And the following coordinate for the projected reference points:

$$\begin{aligned} x_c &= -\mathrm{R}\frac{1 + \cos\theta}{2} \\ x_e &= \mathrm{R}(\cos\theta - \frac{1 + \cos\theta}{2}) \\ x_{tangency} &= -\frac{\mathrm{R}}{8}(-2 + 6\cos\theta + \cos(\theta - 2\phi) + 2\cos 2\phi + \cos(\theta + 2\phi))\csc\phi \end{aligned} \tag{3.71}$$

The projected density obtained here $d_p[\vec{\beta}](x,y)$ is then numerically convolved to a Gaussian function of full width at half maximum (FWHM) $\bar{\sigma}$ in order to take into account the positioning imprecision of the super-resolution method. Eventually the value of the 2D-convolution at location (x_i, y_i) between the density and the Gaussian g_i of variance $\sigma_i^2 \mathbb{1}$ in equation (3.73) is evaluated numerically. As the projected density is null outside the disk of radius R, function is evaluated on a square grid of $R \times R$:

$$d_{\vec{\beta}}^{\text{mes}} \sim \sum_{y_{\text{grid}}} \sum_{x_{\text{grid}}} \frac{d_p[\vec{\beta}](x_{\text{grid}}, y_{\text{grid}})}{2\pi\sigma_i^2} e^{-\frac{(x_{\text{grid}}-x_i)^2+(y_{\text{grid}}-y_i)^2}{2\sigma_i^2}} \Delta x_{\text{grid}} \Delta y_{\text{grid}} \tag{3.72}$$

The divergence for $\rho \to \mathrm{R}$ in equation (3.68), albeit not problematic from a mathematical point of view –integral is defined and Gaussian convolution regulates it, is complicated to handle numerically on a basic grid evaluation scheme leading to strong numerical errors depending on the evaluation grid (evaluation close to $\rho = R$ produces Inf. values). In order to prevent problematic numerical behavior, we used a regulation procedure to ensure that it is properly treated. On a ring of size $\epsilon \ll \sigma_i$ around $R-\epsilon < \rho < R$ where the problem occurs, the projected density is not evaluated to its actual value but taken as its mean value over this interval instead:

$$d_p[\vec{\beta}](\vec{r}) \underset{\mathrm{R}-\epsilon<\rho<\mathrm{R}}{=} \frac{\int_{\mathrm{R}-\epsilon}^{\mathrm{R}} \frac{\rho \,\mathrm{d}\rho \,\mathrm{d}\varphi}{\sqrt{1-(\rho/\mathrm{R})^2}}}{\int_{\mathrm{R}-\epsilon}^{\mathrm{R}} \rho \,\mathrm{d}\rho \,\mathrm{d}\varphi} = \frac{\sqrt{(2\mathrm{R}-\epsilon)\epsilon}}{\mathrm{R}\epsilon - \epsilon^2/2} \sim \sqrt{\frac{2}{\epsilon \mathrm{R}}} \tag{3.73}$$

We have then fixed the grid step $\Delta x = \Delta y = 4\,10^{-4} \times R$ and $\epsilon = 2\Delta x$ in order to achieve a convergence of the calculation with a relative precision of 10^{-4} on the result. This is a trade-off with computation time and memory requirements. We used those parameters in all our calculation in the following. For a grid of $n \times n$, each position likelihood require $O(n^2)$ evaluations so the computational complexity is $O(P.n^2)$ for P positions. This ad-hoc calculation scheme generates rounding errors that may accumulates. However performing the integration in equation (3.68) using an iterative procedure with controlled error and fixed convergence tolerance as upper and lower integration (Since the integrand is analytical, the function `integral2` can be used for this purpose in Matlab®) is very consuming in time.

The convolution is to be repeated again and again to evaluate the likelihood of each recorded protein position $\vec{X}_i$ according to its known variance σ_i^2 issued by the surperresolution centroid finding algorithm.

We also considered the case of homogeneous precision $\sigma_i \simeq \bar{\sigma}^2$, for which the convolved density is identical for all sampled positions. The calculation was performed once for the whole grid followed by linear interpolation of the function on each sampled position. To speed up the calculation time we implemented Fast Fourier Transform (FFT) using the relation:

$$d_{\vec{\beta}}^{\text{mes}} = d_p * g = \mathrm{FFT}^{-1}(\mathrm{FFT}(d_p).\mathrm{FFT}(g)) \tag{3.74}$$

So that the evaluation grids can be set fine enough without slowing to much calculation time (classical convolution complexity is $\propto O(n^4)$ whereas the FFT is $\propto O(2n^2 \log n)$) which becomes competitive with the previous scheme if the number of sampled positions is of the same order as logarithm $\log n$ of the grid size. This provide a fast evaluation of the likelihood (use a fine grid for convolution) at the expense off assuming a constant precision for each measure point.

Once the projected measure probability density function is obtained, uniform background is added. This operation is equivalent to add a constant to the calculated density and diminish the

sensibility to outliers. We used a background probability equal to 10% of the object probability. The measure density of the object sums to $\simeq 1$ on the whole image, so we calculated the likelihood of parameter set given the position from:

$$d^{\text{mes}}_{\vec{\beta}} \to d^{\text{mes}}_{\vec{\beta}} + \frac{10}{100} A^{-1} \tag{3.75}$$

Where A is the area of the image, and normalization is not needed since the likelihood is defined up to a multiplicative constant.

3.6.5 Likelihood maximization

Once we were able to evaluate the likelihood function for any parameter set, the parameter estimation requires to find its maximizer. We used different strategies of maximization in order to favor control of the likelihood behavior or low computational cost and efficient search.

3.6.5.1 "Exhaustive" evaluation

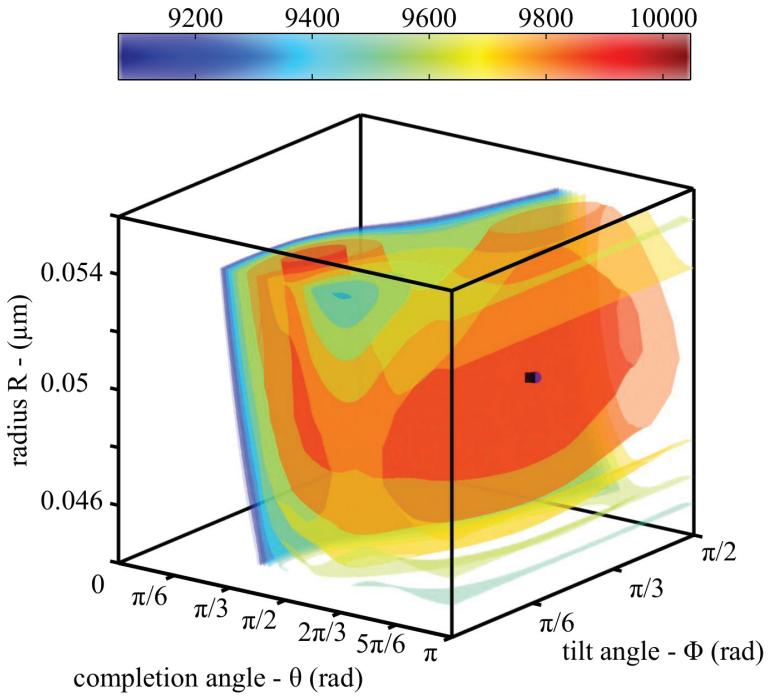


Figure 3.7 – loglikelihood "landscape": six linearly spaced isosurfaces of the loglikelihood (colors) are extrapolated from the its evaluation over a grid $(\theta, \phi, R) \in [\pi/6, \pi]$ rad $\times$ $[0, \pi/2]$ rad $\times$ $[44, 56]$ nm of $100 \times 50 \times 10$ points. The input value $(R_0, \theta_0, \phi_0) = (50\,\text{nm}, 3\pi/4\text{ rad}, \pi/3\text{ rad})$ (■) for the simulated the $N = 5000$ positions with a very low positioning uncertainty $\sigma \simeq 0.05$ nm coincides with the global maximum on the evaluation grid (◯ blue).

The most natural solution to check that the estimator corresponds to the requirement is the brute force evaluation of the loglikelihood function on the parameter space. The interest of this method is that, since our parameter space (R, θ, ϕ) is 3D, it is possible to visualize the loglikelihood "landscape" on the parameter space plotting the isovalues surfaces extrapolated

(linear extrapolation between neighbors) from the values calculated on the grid (see figure 3.7). The drawback is clearly that many evaluations are made in regions far from the maximizer and that the computational cost is enormous compared to the precision reached (which is of the order of the grid step). Simulated super resolution image whose loglikelihood landscape is displayed on figure 3.7 was generated using unrealistic positioning precision compared to superresolution microscopy power and very good sampling density compared to the viral content in order to emphasize the general features of the likelihood of the truncated sphere. On one hand, figure 3.7 shows that in this particular case, the maximum likelihood provides a good estimation of the distribution from which the sampled positions were drawn. On the other hand, the likelihood landscape is more complicated than what the perfect "regular" case should look like. Even if the evolution around the so found maximum seems indeed an acceptable candidate for a quadratic approximation, the function is not perfectly concave and for low tilts and completion $\phi < \pi/6$ and $\theta < \pi/2$ there is a "branch" region where the loglikelihood also takes stronger values than in its neighborhood indicating a second optimum of the function. Furthermore this tortured shape is not an easy task for optimization solvers. As the "branch" extend to higher radius, it is likely that an iterative solver initiated with a higher radius will be attracted towards this region and stuck in the vicinity far from the true optimum.

We studied in more details the maxima of these likelihood landscape. This study of the local maxima positions was done by extensive calculations of the function values on a grid around the initial value (R_0, θ_0, ϕ_0) used to simulate the various set of positions for which the "landscape" where calculated. No outliers were added and the set of point submitted for the calculation was already properly oriented. As the radius R_0 is only a scaling parameter for the density, changing it for a new value simply shift the landscape accordingly. We focused on the effect of the completion and orientation at fixed radius $R_0 = 50$ nm.

The conclusion holds when the input image is simulated from a distribution with more than half completed sphere figure 3.9. For an input completion of $\theta_0 = 2\pi/3$ the two branches of maxima overlap. The correct estimation lies this time in the horizontal branch. A unequivocal determination seems possible when the tilt angle is higher than the critical value $\phi_0 > \phi_c \simeq [\pi/6, \pi/5]$.

Figure 3.8 shows the regions of the parameter space where loglikelihood is maximum ($\log L \in [\max(\log L), \max(\log L) - 80]$) for $N = 5\,10^3$ input positions drawn from a spherical cap with small completion $\theta_0 = \pi/4$ viewed from different tilt angles ϕ_0. Low tilt angles $\phi_0 < \pi/6$ leads to a landscape where high loglikelihood values are distributed in two distinct regions of the space:

— The already mentioned vertical branch characterized by increasing radii R, low completions ($\theta < \pi/2$), low tilt angles. The input parameter of the actual density from which the image was simulated is localized in this region.

— An horizontal branch at constant radii R, high completions ($\theta > \pi/2$), over a large range of tilt angles ϕ. The typical radius of 40nm that maximizes the loglikelihood in the region is much smaller than the input radius of the simulation R_0.

The trends are either a low completion with a well determined tilt angle and poor precision on radius, or a high completion with poorly determined tilt angle but well known radius:

— In the vertical branch, proteins are assumed to lie on a small cap. So the projected density varies strongly with respect to angle of view (tilt angle ϕ). This explains why it is well determined: from the top it looks like a disk, whereas the side view only a small portion

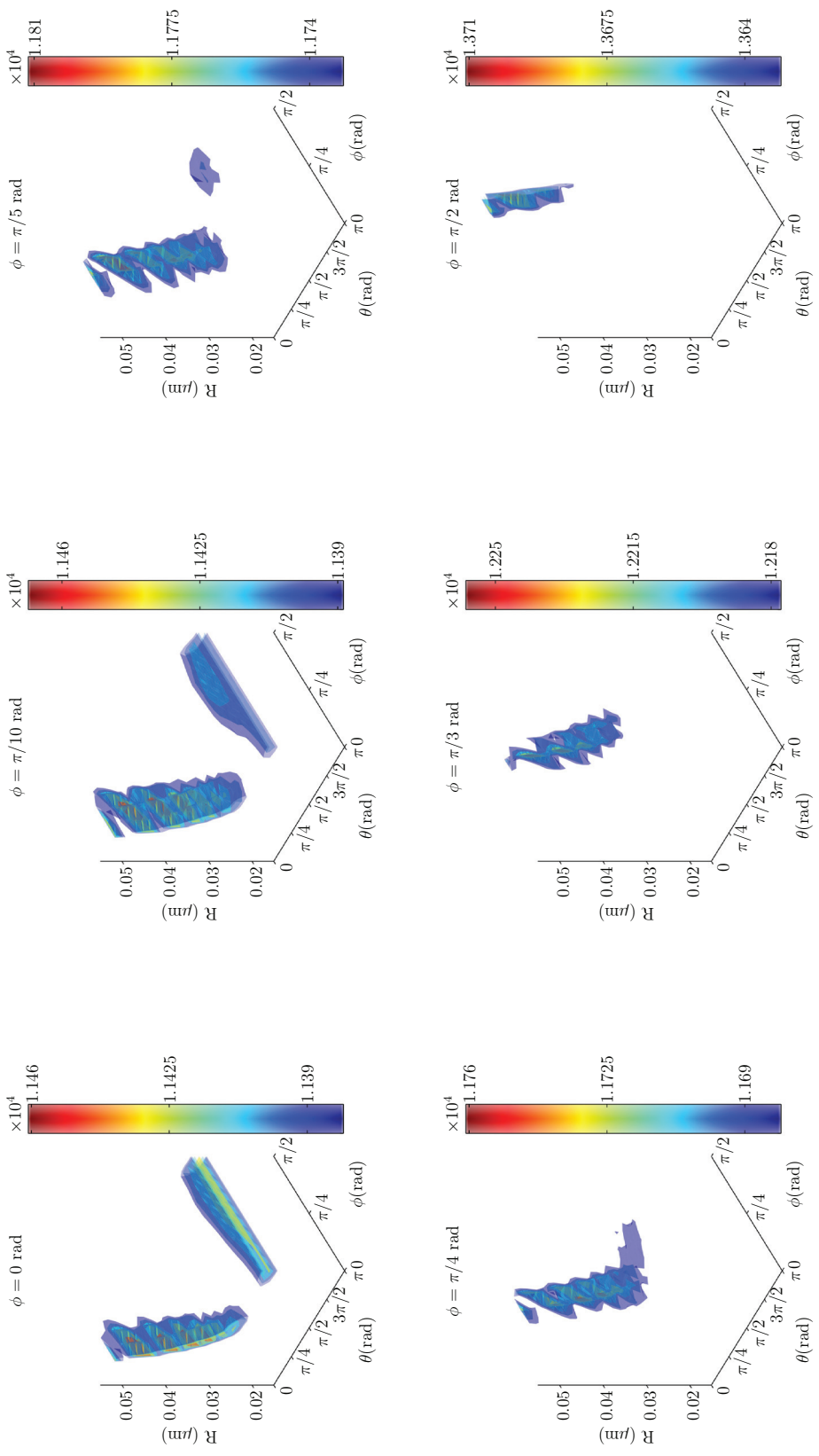


Figure 3.8 – loglikelihood landscapes for simulated images of $N = 5\,10^3$ positions drawn from density parametrized by fixed radius $R = 0.050$ µm, completion angle $\theta = \pi/4$ rad, positioning precision $\sigma = 5$ nm, and various tilt angles $\phi \in \{0, \pi/10, \pi/5, \pi/4, \pi/3, \pi/2\}$ rad.

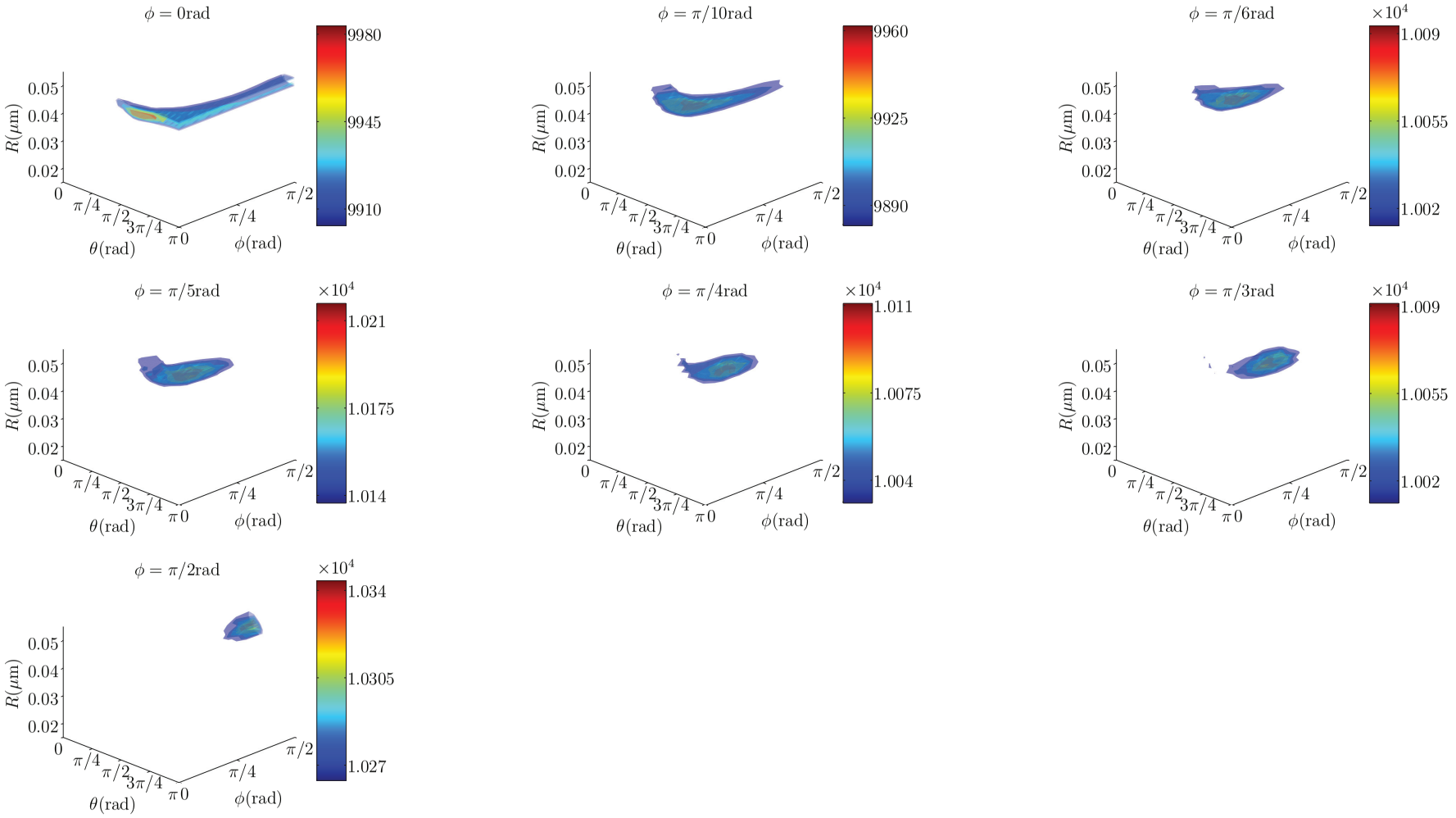


Figure 3.9 – loglikelihood landscapes for simulated images of $N = 5\,10^3$ positions drawn from density parametrized by fixed radius $R = 0.050\,\mu$m, completion angle $\theta = 2\pi/3$ rad, positioning precision $\sigma = 5$ nm, and various tilt angles $\phi \in \{0, \pi/10, \pi/5, \pi/4, \pi/3, \pi/2\}$.

of a disk appears. In the top view, the curvature seems hard to infer so that various trade-off between completion and radius are hard to distinguish: the compatible R and θ form the observed branch.

— In the horizontal branch, the protein support is close to a full sphere and its projection looks like a disk independently of tilt. However it is much more sensitive to a change in radius which is the well determined parameter.

As the input distribution tilt angle ϕ_0 is increased, the vertical branch is displaced to higher tilt values, remaining centered on the correct estimation. The horizontal branch remains only for high tilt angle and is moved towards smaller values of completions. Overall distribution with stronger asymmetry are put forward. The horizontal branch eventually vanishes for $\phi_0 > \pi/4$. The contour shape of the point distribution seems to be determining for the shape of the log-likelihood landscape. The more asymmetry in the input image and the easier the determination of completion by the loglikelihood. Under a critical tilt angle the determination between more or less than half completion is poor.

Investigating further the geometrical effect on the loglikelihood landscape, we hypothesize that the critical angle for a given input completion $\phi_c(\theta_0)$ was determined by a change in the shape of the projected protein density. Also, it should be a characteristic feature that explains similar loglikelihood value for each of the two branches of maxima observed.

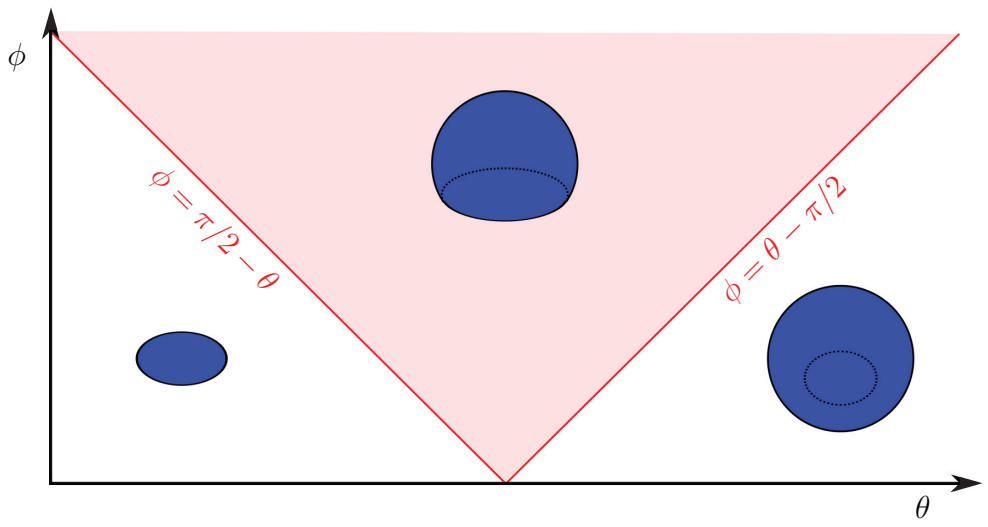


Figure 3.10 – the three different shapes identified for the projection of a truncated sphere depending on its completion θ and on the observer point of view ϕ: (from left to right) ellipse, portion of a disk and portion of ellipse (light red), full disk. The apparent area is the blue area. Dashed contours are an indication of original 3D object

We identified three domains of shapes in the tilt/completion space showed in figure 3.10. The projected truncated sphere leading either to an ellipse, a disc or to a mix of each. The boundaries between this regions are $\phi_c = \pi/2 - \theta_0$ and $\phi_c = \theta_0 - \pi/2$ so that we expect $\phi_c(\pi/4) = \pi/4$ and $\phi_c(2\pi/3) = \pi/6$ consistent with our previous observations. Under our hypothesis, the input densities from the colored region in figure 3.10 are easily discriminated

from the others. On the contrary, a parameter set in a white region is less easy to distinguish from other parameter sets in the other white region leading to distinct branches of maxima.

We refine further the description of the loglikelihood maxima branches. The dominant factor of discrimination between the possible proteins densities is the projected shape of the densities. This suggests that the area of the plane covered by their projections could be a simple criterion of equivalence for two truncated spheres. This area is the apparent area, for a full sphere it is the area of a disk $\mathcal{A}_{\text{apparent}} = \pi R^2$, whereas the projected area sums to 1 by definition (equation (3.68)).

To test this hypothesis we calculated the apparent area of the truncated sphere of radius R depending on its completion and orientation. As it appears on figure 3.10, the projection of any truncated sphere can be constructed as the union of several portions of ellipses.

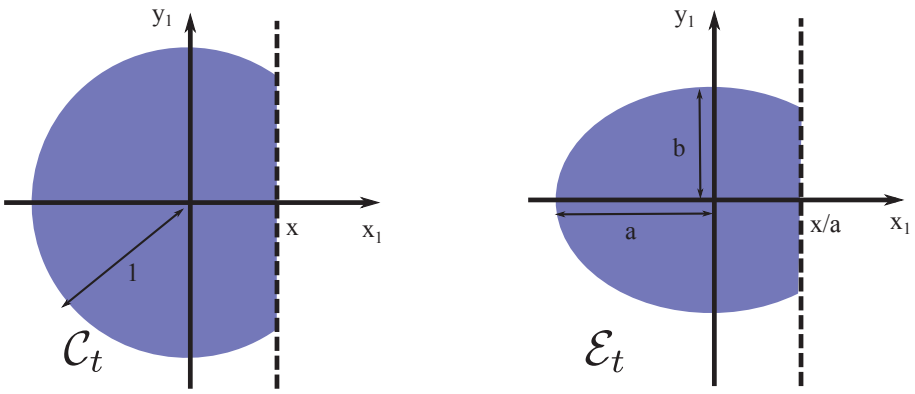


Figure 3.11 – portions of disk and ellipse used to calculate the apparent area of the projected truncated sphere

In order to the area of a portion of ellipse, we first calculate the area of $\mathcal{C}_t$ the unit circle centered on the origin truncated of the portion with abscissa $x_1 > x$ (see figure 3.11) is:

$$\begin{aligned} \mathcal{A}_{\mathcal{C}_t}(x) &= \iint_{\mathcal{C}_t} \mathrm{d}x\,\mathrm{d}y \\ &= 2\int_{-1}^{-1+x} \sqrt{1-x_1^2}\,\mathrm{d}x_1 \\ &= \frac{1}{2}\left(\sqrt{x(2-x)}(x-1) + 2\cos^{-1}(1-x)\right) \end{aligned} \tag{3.76}$$

The area of an ellipse portion $\mathcal{E}_t$, with arbitrary axis length (a, b) is simply deduced from the previous result using the 2D variable substitution $x \to \frac{x'}{a}$ and $y \to \frac{y'}{b}$:

$$\begin{aligned} A_{\mathcal{E}_t}(x, a, b) &= \iint_{\mathcal{E}_t} \mathrm{d}x\,\mathrm{d}y \\ &= ab A_{\mathcal{C}_t}\left(\frac{x}{a}\right) \end{aligned} \tag{3.77}$$

And eventually the apparent areas of the shapes showed on figure 3.10 writes (using equa-

tion (3.71) to position the truncature):

$$\begin{aligned}\mathcal{A}_{\text{apparent}}(R,\theta,\phi) = & \underset{x>0}{\text{Ind}}(\phi-(\theta-\pi/2)).\underset{x>0}{\text{Ind}}(\phi-(\pi/2-\theta))\times \\ & \Bigg(S_{\mathcal{E}_t}\left(R(1-\cos\theta/\sin\phi),R,R\right) \\ & + S_{\mathcal{E}_t}\left(R\left(\sin\theta+\frac{\cos\theta}{\tan\phi}\right)\cos\phi, R\sin\theta\cos\phi, R\sin\theta\right)\Bigg) \\ & + \underset{x>0}{\text{Ind}}(\pi/2-\theta-\phi).\pi R^2\sin^2\theta\cos\phi \\ & + \underset{x>0}{\text{Ind}}(\theta-\pi/2-\phi).\pi R^2\end{aligned} \tag{3.78}$$

Where the step function $\underset{x>0}{\text{Ind}}$ is used to generate the indicator function of the three different shape regions identified in figure 3.10:

$$\underset{x>0}{\text{Ind}}(y) = \begin{cases} 0 & y \leq 0 \\ 1 & y > 0 \end{cases} \tag{3.79}$$

Figure 3.12a shows the apparent area of the truncated unit sphere depending on its completion and orientation (θ,ϕ). Figure 3.12b is a plot of an isosurface the apparent area function, that is the parameters sets of truncated sphere with the same apparent area. As already told, R is a scaling parameter for the protein distribution, and thus for its projected area:

$$\mathcal{A}_{\text{apparent}}(R,\theta,\phi) = R^2\mathcal{A}_{\text{apparent}}(1,\theta,\phi) \tag{3.80}$$

Thus the knowledge of the values of $\mathcal{A}_{\text{apparent}}(1,\theta_{\text{iso}\mathcal{A}},\phi_{\text{iso}})$ is sufficient to deduce the isosurface passing by the point $\{R_0,\theta_0,\phi_0\}$ which can be parametrize as $R_{\text{iso}\mathcal{A}}(\theta,\phi)$:

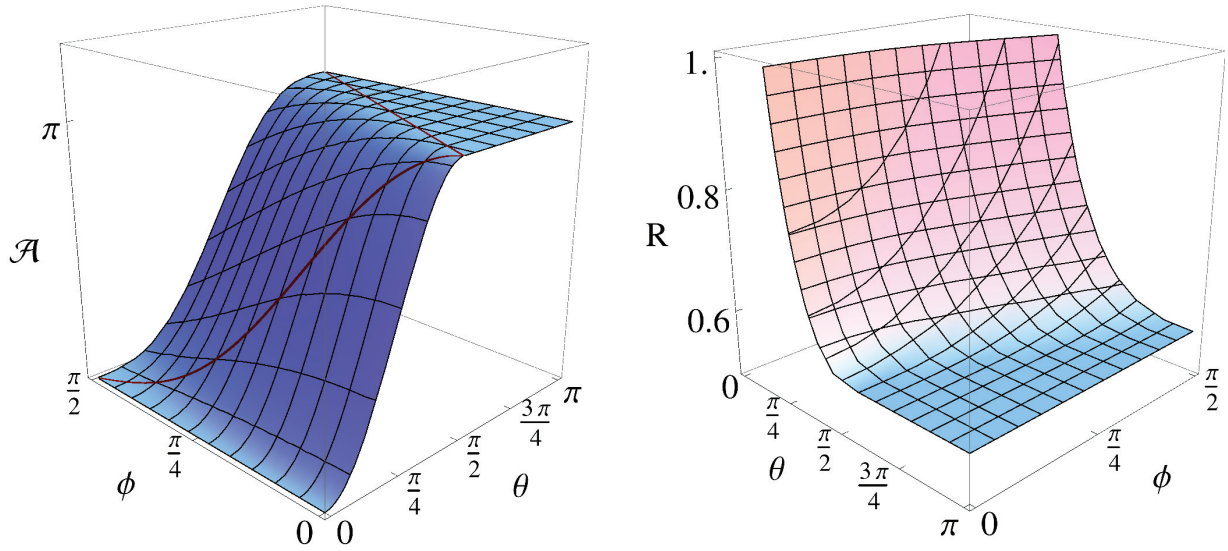
$$\mathcal{A}_{\text{apparent}}(R_{\text{iso}\mathcal{A}},\theta,\phi) = \mathcal{A}_{\text{apparent}}(R_0,\theta_0,\phi_0) \quad \Leftrightarrow \quad R_{\text{iso}\mathcal{A}}(\theta,\phi) = R_0.\sqrt{\frac{\mathcal{A}_{\text{apparent}}(1,\theta_0,\phi_0)}{\mathcal{A}_{\text{apparent}}(1,\theta,\phi)}} \tag{3.81}$$

It defines a natural change of coordinates $h:(R,\theta,\phi)\mapsto(R_{\text{iso}\mathcal{A}},\theta,\phi)$. In the new coordinates $R_{\text{iso}\mathcal{A}}$ measures the apparent area of the object and the shape is given by (θ,ϕ). The reverse transformation is:

$$\begin{pmatrix}\theta_{\text{iso}\mathcal{A}}\\ \phi_{\text{iso}}\\ R_{\text{iso}\mathcal{A}}\end{pmatrix} \overset{h^{-1}}{\mapsto} \begin{pmatrix}\theta_{\text{iso}\mathcal{A}}\\ \phi_{\text{iso}}\\ R_{\text{iso}\mathcal{A}}\sqrt{\frac{\pi}{A_{\text{apparent}}(1,\theta_{\text{iso}\mathcal{A}},\phi_{\text{iso}})}}\end{pmatrix} \tag{3.82}$$

The vertical and the horizontal branch of maxima follow an apparent area isosurface. The match is not perfect since a slightly different values of the apparent area parametrize each of the two branches of maxima (figure 3.13a).

The loglikelihood thus promotes the distributions showing the best overlap with sampled distribution. Regarding our model of image analysis, we see that the loglikelihood favours the extension of the distribution criterion at first (same apparent area, same overall shape), before it accounts for the fluctuations of density inside the shape contours (projected area). In particular, concerning the truncated sphere model, we noticed a low efficiency of the method to decide between the two white regions in figure 3.10 that exhibit similar disk shapes. This may lead to inconsistent results using the maximum likelihood estimator with the low sampling



(a) Apparent surface of a truncated sphere at fixed radius $A_{\text{apparent}}(R = 1, \theta, \phi)$. The red lines shows the three different shape regions determined in figure 3.10.

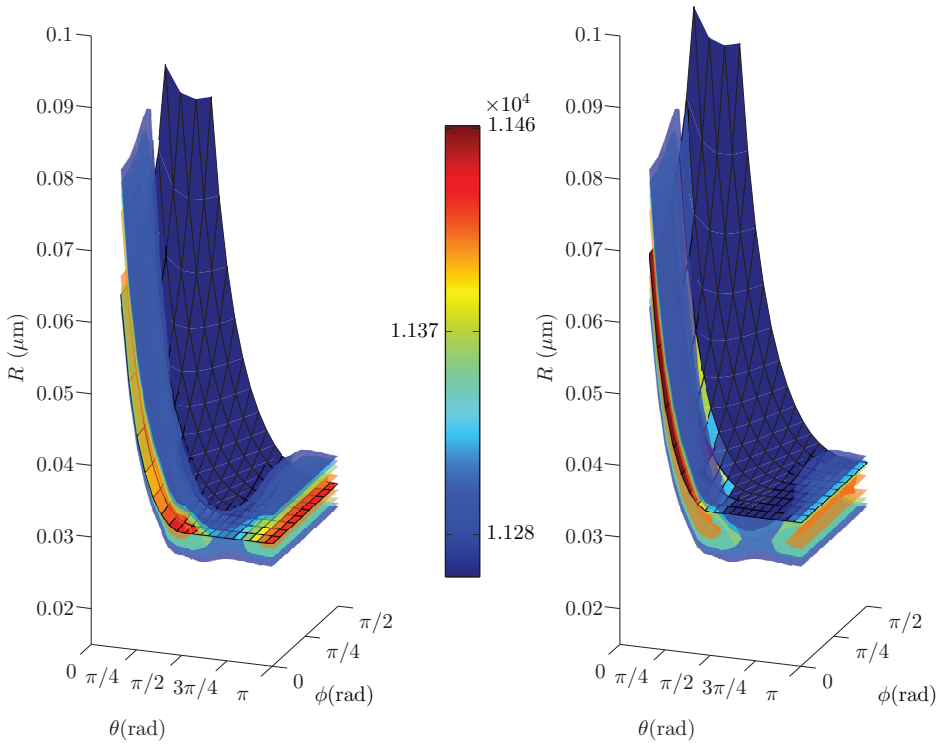
(b) $\mathcal{A}_{\text{apparent}}(R, \theta, \phi) = 1$ isosurface in the parameters space

Figure 3.12 – The apparent surface of a truncated sphere is the area seen from the observer point of view –the second plot is rotated by $\pi/2$ rad clockwise

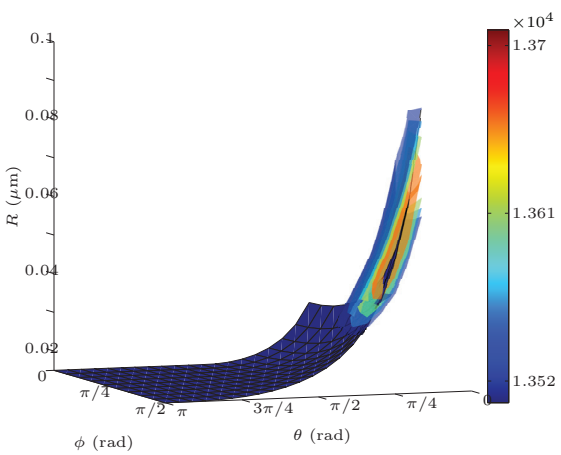
and higher positioning uncertainties that the experimental state of the art can provide. As such, the likelihood estimator is likely to be inconsistent for the model. The estimation would be greatly improved on a simplified case where the decision covers only a restricted parameter space excluding one of the two competing regions and would leads to consistent results.

Adding information to enhance the distinction between the alternatives is another possibility. A constant apparent area corresponds to surface in 3D that varies up to a factor 4 (comparing the disk and the sphere with identical apparent area and 3D area πR^2 and $4\pi R^2$ resp.). This implies the same variation of protein density for a fixed number of sampled positions. The mean protein density could be used as a constraint for the optimization, or added as a prior in a Bayesian approach. Concerning the HIV experimental study of budded particle, we assumed that the protein coverage was greater than half completion and that we could restrict the determination in the $\theta \in [\pi/2\pi]$ domain where a single maximum is found and where the maximum likelihood estimator is consistent.

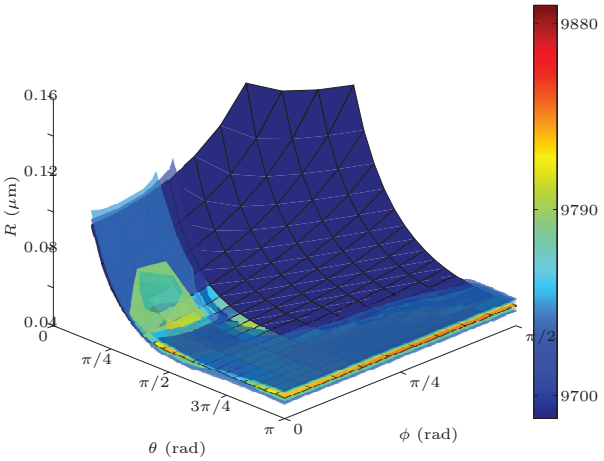
An other interesting outcome of the apparent area criterion, is the possibility to design the maximum search implementation to benefit from this result. Compared to exhaustive search, local maximum can be found in significantly fewer loglikelihood function evaluations using iterative optimization procedure such as simplex search, quasi-Newton method. On one hand it provides a constraint between the parameters to choose an appropriate starting point of the iteration. On the other hand it provides a natural coordinate substitution to map the space parameter in order to improve the convergence odd of the algorithm.



(a) $\theta_0 = \pi/4, \phi_0 = 0$
Apparent area $\mathcal{A} = \pi.r_{\mathcal{A}}^2$ – $r_{\mathcal{A}} = 35$nm (left), $r_{\mathcal{A}} = 32$nm – (right)



(b) $\theta_0 = \pi/4$ rad, $\phi_0 = \pi/2$ rad
$\mathcal{A} = \pi.r_{\mathcal{A}}^2$ – $r_{\mathcal{A}} = 15$ nm.



(c) $\theta_0 = 5\pi/6$ rad, $\phi_0 = \pi/4$ rad
$\mathcal{A} = \pi.r_{\mathcal{A}}^2$ – $r_{\mathcal{A}} = 50.7$ nm.

Figure 3.13 – The maxima of the likelihood function are lying on isovalue surfaces of the apparent area function. The original distribution radius used to simulate the image is $R = 50$ nm. Isolikelihood surfaces are evenly spaced for $\log L \in [\max(\log L), \max(\log L) - 200]$.

3.6.5.2 Quasi Newton-Algorithm optimization

The optimization problem was numerically solved using the "interior-point" algorithm of MATLAB*s*® "fmincon" solver. This algorithm is a member of the quasi-Newton optimization family. It looks for an extremum of the target function using successively its quadratic approximations around the iterations point. Rather than computing Hessian matrix and gradient at each iteration steps, the Hessian matrix is approximated at each step using the gradient values. Since no exact value is available for the gradient, it is evaluated numerically using finite difference. The "interior point" algorithm handles boundaries in the parameter space by introducing a effective function equal to the target function plus an extra barrier function that diverges approaching the interval limits penalizing any solution outside the boundaries. At each iteration, the proportion of the barrier function in the effective function is decreased so that the algorithm converges at an extremum of the target function.

We tested the optimization procedure on already known landscapes. In naive implementations, the solver almost already failed to converge to the expected extrema of the function and were either converging far from the optimum or did not converge at all. We identified two important limitations on our loglikelihood function that might lead to failure:

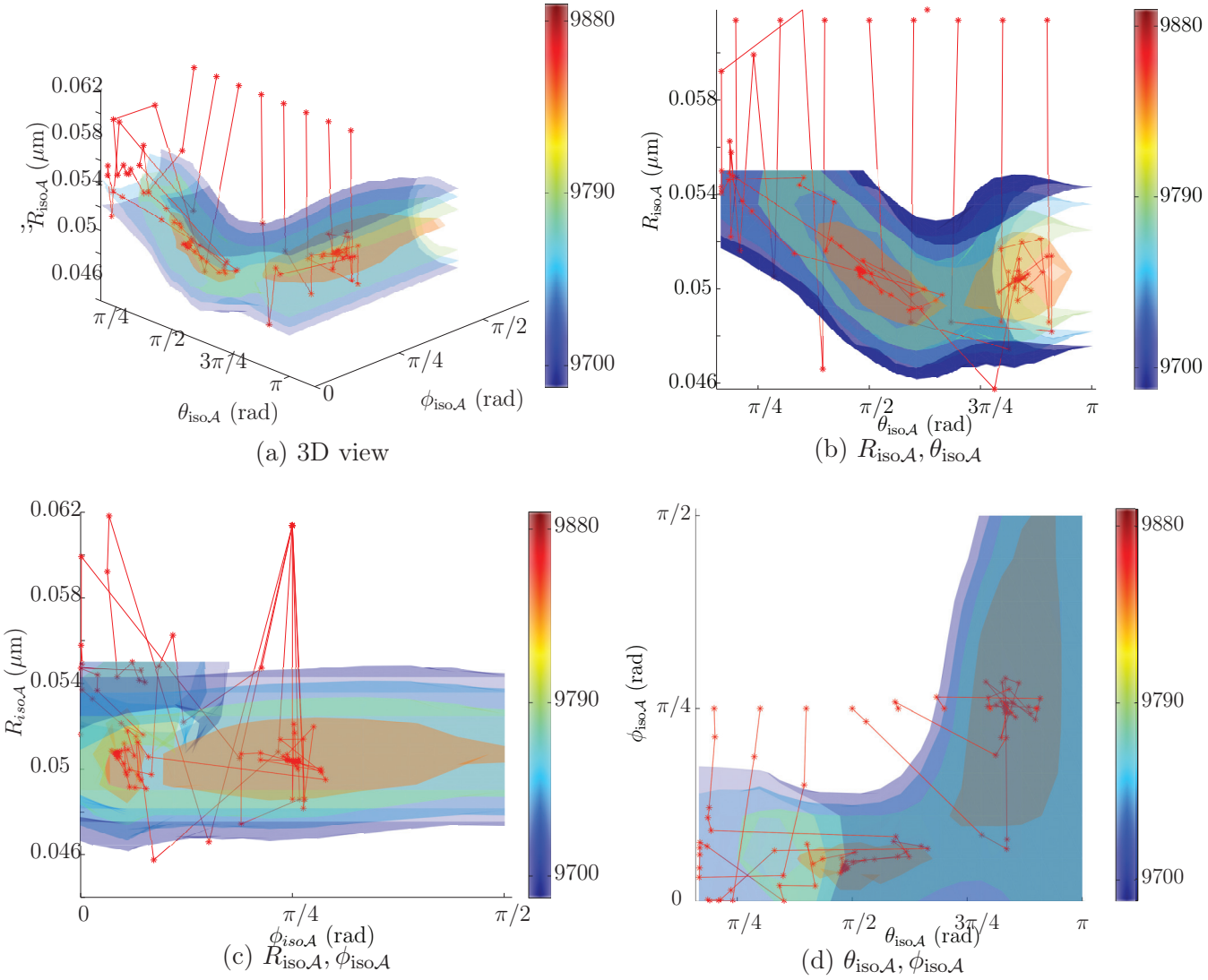
— numerical evaluation error: if the likelihood landscape looks smooth enough at large scales, at smaller scale the function is much more shaky since we only achieved a 10^{-3} relative precision. This impacts the solver both on the differentiation step leading to improper evaluation of the gradient and on the convergence as solver get stuck in the fluctuations (see figure 3.15).

— scaling of the problem: as the variations on one variable strongly dominates the loglikelihood evolution compared to the others (for instance in the horizontal branch, the loglikelihood evolve very slowly in θ and ϕ but extremely fast in R). The solver fails to find the narrow direction of optimization that follow the branch and stops once the dominant parameter only is optimized.

We solved the numerical precision issue by setting a lower limit to the step size of the finite difference calculations so that gradient was evaluated on a scale larger than the scale of the function fluctuation due to numerical error.

Scaling issue was more difficult to tackle. Simply rescaling the variables with a constant factor using the option provided by the solver did not yield to satisfactory enough ameliorations. As reported in the previous section, in first approximation the loglikelihood function maxima branches follows the surface of iso-apparent area. By substituting the natural coordinates of the physical model (R, θ, ϕ), to the iso-apparent surface parameters $(R_{\text{iso}\mathcal{A}}, \theta_{\text{iso}\mathcal{A}}, \phi_{\text{iso}})$ given in equation (3.81), we provide to the solver the natural direction for optimization along the iso-apparent surfaces and enable the solver to find the optimum in both maxima branches (compare). Maximization of $\log L \circ h^{-1}$ is numerically much easier (compare figure 3.13c and its transformed counterpart figure 3.14). Both maximum in the landscape are reached by the solver depending on the starting point of the iterations. The solver is therefore able to identify the two maxima as expected unlike in the original parameter space (R, θ, ϕ).

The solver produces an approximation of the Hessian matrix of the loglikelihood at the found optimum in the new coordinate system. This matrix is by definition $-I(\vec{\beta}_{\text{iso}})$, the opposite of the observed Fisher information matrix (equation (3.65)). However, due to the evaluation precision limit of our numerical loglikelihood, the finite difference matrix obtained this way had a very poor precision (only the order of magnitude can be trusted). We refine the result using



(a) 3D view

(b) $R_{\text{iso}\mathcal{A}}, \theta_{\text{iso}\mathcal{A}}$

(c) $R_{\text{iso}\mathcal{A}}, \phi_{\text{iso}\mathcal{A}}$

(d) $\theta_{\text{iso}\mathcal{A}}, \phi_{\text{iso}\mathcal{A}}$

Figure 3.14 – Loglikelihood landscape of an image of $N = 1500$ positions sampled from a proteins densities $(R, \theta, \phi) = (50, \frac{5\pi}{6}, \frac{\pi}{4})$(nm,rad,rad) and $\bar{\sigma} = 20$ nm. The successive positions (red $*$) at each solver iterations are shown for 8 different initial starting points $\vec{\beta}_{\text{iso}\mathcal{A}}|_{s=0} = (R_{\text{iso}\mathcal{A}}|_{s=0} = r_{CH}, \theta_{\text{iso}\mathcal{A}}|_{s=0}, \phi_{\text{iso}\mathcal{A}}|_{s=0})$. The apparent area of the positions distribution convex hull $\mathcal{A} = \pi r_{CH}^2$ is used to set the initial apparent surface parameter together with various completions $\theta_{\text{iso}\mathcal{A}}|_{s=0} \in [0, \pi]$ rad at fixed $\phi_{\text{iso}\mathcal{A}}|_{s=0} = \pi/4$ rad. In the simulation $r_{CH} = 64$ nm

$9 \times 9 \times 9$ evaluation points in the neighborhood of the optimal position $\vec{\beta}_{\text{iso}}$ found by the solver and a mean least square (see appendix A) fit of the quadratic form:

$$\log L \circ h^{-1}(\vec{\beta}_{\text{iso}}) = (\vec{\beta}_{\text{iso}} - \vec{\beta}^*_{\text{iso}})^t H_{\text{iso}} (\vec{\beta}_{\text{iso}} - \vec{\beta}^*_{\text{iso}}) \tag{3.83}$$

Where $-H_{\text{iso}}$ is the observed Fisher information matrix that we are willing to test as an estimation of the variance and $\vec{\beta}^*_{\text{iso}}$ is our optimal parameter in the modified coordinate system. We

come back to the natural parameter system using the inverse transformation (see appendix C).

$$\vec{\beta}^* = h^{-1}(\vec{\beta}^*_{\text{iso}}) \tag{3.84}$$

$$I^{-1}_{\vec{\beta}^*} = -J^t_h(\vec{\beta}^*).H^{-1}_{\text{iso}}.J_h(\vec{\beta}^*) \tag{3.85}$$

The whole process takes around 30 sec to localize the maximum and 3 min to provide the final estimation with error bars for a single particle.

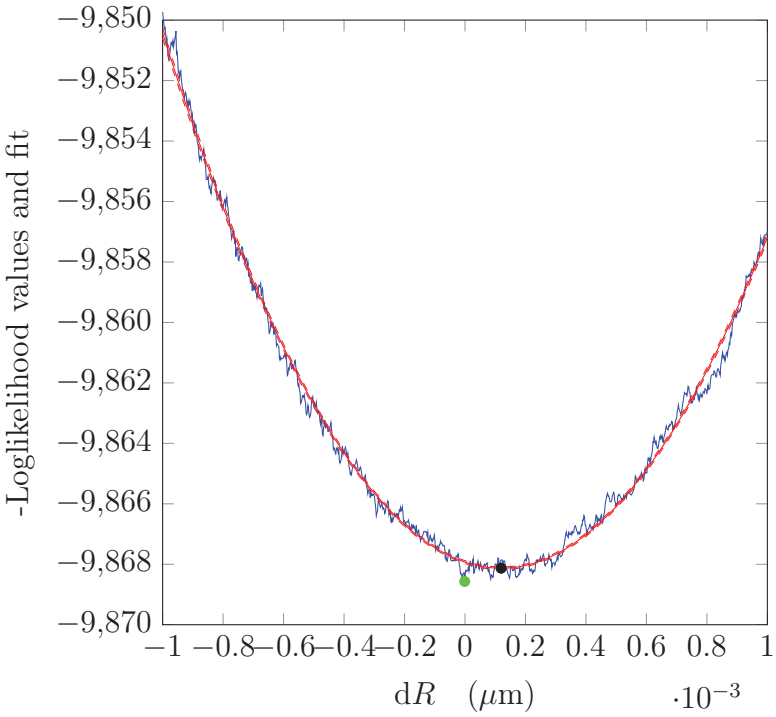


Figure 3.15 – fits of the minus loglikelihood along $\vec{u}_R$, $-\log L(\vec{\beta}^*_{\text{iso}\mathcal{A}} + \vec{u}_R.\,\mathrm{d}R)$ in the neighborhood of the minimum found by the solver (green dot) $-\frac{\partial^2 \log L}{\partial R^2_{\text{iso}\mathcal{A}}}$ to extract the Fisher information matrix first coefficient from a simulation ($N = 5000$ points $\theta = 5\pi/6$ rad, $\phi = \pi/4$ rad, $R = 50$nm). At low scales the shakiness of the function comes from numerical error (discretization approximations and rounding errors).

3.6.6 Likelihood test on simulated PALM images

To evaluate the performance of the maximum likelihood reconstruction procedure, we implemented it on simulated PALM images of truncated spheres of known geometrical parameters $\vec{\beta}_0 = (R, \theta, \phi)$. The simulated images where all produced according to section 3.3 (page 99) with parameters typical of the experimental data. We thus provided as an input 300 simulated datasets composed of $N = 1500$ points each with an uncertainty in position of 20 nm comparable to the experimental datasets (figure 3.1). To ensure consistency of the maximum likelihood method, we restricted the parameter space to completion $\theta \geq \pi/2$ rad which we assumed valid for the budded viruses according to what was reported from cryo-electron microscopy observations by Carlson et al.[21]. We could therefore compare the computed optimal

structure $\vec{\beta}^*$ to the known input structure $\vec{\beta}_0$ and estimate the efficiency of the maximum likelihood method from these differences. This also provides a test for the proposed confidence estimation calculated from the observed Fisher information matrix.

The input parameters set of the simulated data were randomly drawn with a uniform probability in the interval: $(\mathrm{R}, \theta, \phi) \in [30, 80]\text{nm} \times [\frac{\pi}{2}, \pi]$ rad $\times [0, \frac{\pi}{2}]$ rad and a mean positioning uncertainty of $\bar{\sigma} = 20$ nm was applied to simulate the image. The comparison between actual

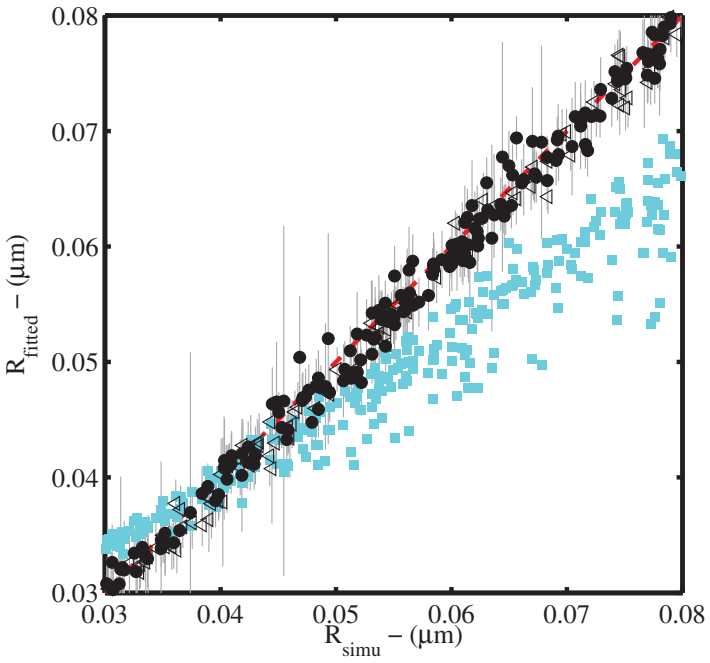


Figure 3.16 – Radii estimations R_* using maximum likelihood estimator on 300 simulated super-resolution images (• and ◁) and using the mean distance to center of mass R_{CM} (blue □) are plotted against the original ones R_0. Perfect estimation follows the red dashed line. Particles for which orientation was not properly determined (std.$\alpha > \frac{\pi}{10}$ in equation (3.53) are shown by ◁). Standard deviations values calculated from the Fisher information matrix are displayed ($\sqrt{I^{-1}_{RR}(\vec{\beta}^*)}$ grey error bars)

and estimated particle radii including the estimated error bars are shown in figure 3.16. We found that the particle radius estimation R_{fitted} is nicely distributed around the input value. The standard deviation of the difference $\Delta R = R_{\mathrm{fitted}} - R_0$ was $\mathrm{std}_{\Delta R} = 1.3$ nm while the average error, or bias $\overline{\Delta R}$ was under one nm). Thus 95% of the simulated particles have been reconstructed with an error on the radius smaller than 3 nm when the proteins where localized with a precision of 20 nm. Also, this error is independent from the original radius, unlike in simplest procedures based on averages over the distribution of emitter positions (naive implementation of the moment method). Such a geometry-dependent error clearly affects the basic 2D radius estimation using the mean distance to the center of mass discussed in section 3.4

(page 102) and shown for comparison.

$$R_{CM} = \frac{1}{N}\sum_j \sqrt{(x_j - \frac{1}{N}\sum_k x_k)^2 + (y_j - \frac{1}{N}\sum_k y_k)^2} \tag{3.86}$$

As already noticed simple estimator is biased with a strong variability since completion and tilt degrees of freedom are not taken into account. Besides it is not able to distinguish anything smaller or of the order of the positioning precision as does our procedure that takes this information into account. Remarkably, this result shows that the radius of the particle can be deduced at the single object level from a maximum likelihood estimation with an accuracy better than the microscope precision ($\bar{\sigma} \approx 20$ nm) by one order of magnitude in the conditions of our simulation.

Variance on each fitted parameter is estimated from the diagonal element of the observed Fisher information matrix given in equation (3.65). The average variance is equivalent to an estimated uncertainty (standard deviation) of 3.2 nm and constitutes an upper bound of the measured estimation error in good agreement with its order of magnitude. The details of the error distribution is given in figure 3.19a. At the level of individual object the same observation holds: error bar estimated from the observed Fisher matrix is an upper bound of the effective error, which is regularly distributed and centered.

We performed a similar analysis to quantify differences between the actual and estimated values of the Gag shell completion from the maximum likelihood estimation using the same data (figure 3.17). We observed a standard deviation $\sigma_{\Delta\theta} = 0.3$ rad. However, in contrast to the maximum likelihood estimation values for the radii, the standard deviation is not any more independent from the completion of the Gag shell. The best matches between simulated completions and the maximum likelihood results are obtained for θ between $\frac{2\pi}{3}$ rad and $\frac{5\pi}{6}$ rad. Values outside this range show both a larger systematic average bias and larger standard deviation. The effect is preponderant close to the boundaries, of the optimization domain, which prompted us to verify that this was not an artifact of the interior point algorithm. Exhaustive computation of the likelihood landscape were performed and we obtained the same maxima. Sampling effect has also been questioned: since the simulations are run with a fixed number of points, the sampling density decreases on the 3D surface when completion increases up to a factor 2. We found no evidence that such variation of the 3D sampling density is correlated to the precision of completion estimation. As a matter of fact, hemispheric densities, that have the strongest 3D sampling density can also exhibit a large estimation error, indicating that this is probably not the key parameter.

The observation of a stronger error close to the interval boundary is compatible with the equivalence between densities close to the two boundaries of the interval: an hemisphere ($\theta = \frac{\pi}{2}$ rad) viewed from the top ($\phi \sim 0$ rad) or a complete sphere ($\theta = \pi$ rad) cannot be distinguished from each other since they have the same density up to a multiplicative factor. This suggests a geometric effect. Furthermore large completion θ as well as low tilt angles ϕ results in low asymmetry of the position distribution. Combined with positioning error and small sampling the symmetry information might be insufficient to properly orient the point distribution in the plan. Indeed most of the nearly complete geometries are affected by a strong error whereas only part of the nearly hemispheric geometries are. And among the nearly hemispheric geometries affected, a large majority has uncertain orientation ($\triangleleft$) or in other word a nearly isotropic distribution. For low completions ($\theta \simeq \frac{\pi}{2}$), there is an obvious correlation between a strong

estimated variance of α (see $\triangleleft$ data in figure 3.17), a low tilt (cf. estimated orientation variance and tilt on figure 3.4b), and the failure in θ estimation (figure 3.17 also). Nevertheless since at low tilt the point distribution is close to isotropic, the orientation α is not a significant parameter. So the completion determination by the maximum likelihood estimation failure is rather conditioned by low tilt angle ϕ which also leads to high estimated variance of α (hence the correlation).

Our interpretation is that the cause of the estimation error is not a wrong orientation of the particle but that likelihood itself is not efficient to determine the completion when the distribution is isotropic. Again, it seems that the likelihood criterion is primarily sensible to deformation affecting the contours of the object, and less to its inner density fluctuations. Hence near complete spheres as well as hemispheres seen from the top are nearly equivalents "in the eyes" of the maximum likelihood estimator. This second interpretation advocates for a careful treatment of the completion results of dataset with unclear symmetry axes, for which only the fitted radius is not proven meaningful.

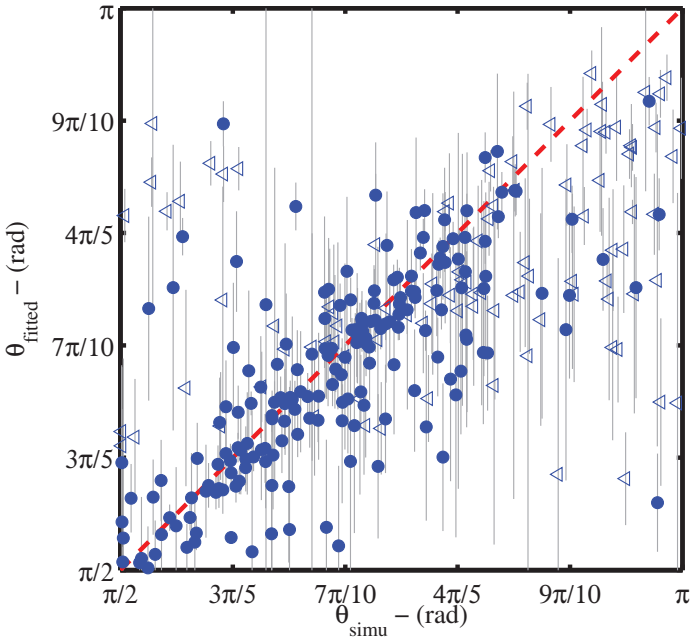


Figure 3.17 – Completion estimations θ_* using maximum likelihood estimator on 300 simulated super-resolution images ($\bullet$ and $\triangleleft$) are plotted against the original ones θ_0. Perfect estimation follows the red dashed line. Particle for which orientation was not properly determined (std.$\alpha > \frac{\pi}{10}$ in equation (3.53) are shown by $\triangleleft$). A third of the standard deviations values calculated from the Fisher information matrix is displayed for readability ($\frac{1}{3}\sqrt{I^{-1}_{\theta\theta}(\vec{\beta}^*)}$ grey error bars) since their values are very large.

Stronger sampling should enhance the precision on incomplete area detection associated to densities lower than expected. We tried to estimate an order of magnitude of the limiting resolution that can possibly be reached. Existing studies of this issue (Fitzgerald et al.[44], Mukamel and Schnitzer[99]) focused on the estimation of the resolution on the 2D plane image in term of

spatial frequency an Fourier analysis. Here we attempt to estimate the theoretical resolution limit on contour of an empty patch on the sphere set by the sampling density. Our basic idea is that the denser the points on the surface, the more suspect an empty space should be and that in return sampling fluctuations set the typical empty patch size indicating incompleteness. Given a set of N points uniformly sampled on a (possibly incomplete) sphere, what would be the minimal size of an observed empty area patch for which the hypothesis "the sphere is complete" should be rejected? To answer this question, we derived the probability distribution of distance to the nearest neighboring point (see detailed calculation in appendix D). Any void left between a point and the closest point can be suspected to identify an incomplete area and the distance distribution give their typical distribution. If we observe a void much larger than expected, then it is likely to be the incompleteness signature. To define "larger" we choose an acceptable error level ϵ. This defines the confidence interval that the observed distance between two neighbors lay within $[0, d_\epsilon]$ as the cumulated probability is equal to $1-\epsilon$ for d_ϵ and thus fix the resolution. Figure 3.18 show the first neighbor distribution for various sampling (inset) and approximated upper bounds of the confidence level ϵ –express in term of angle from the center of the sphere– as function of the sampling:

$$d = R.\theta \tag{3.87}$$

$$\theta_\epsilon = \cos^{-1}(1 - \frac{2}{N})\sqrt{-\pi \log \epsilon} \simeq \sqrt{\frac{-\pi \log \epsilon}{N}} = O(N^{\frac{1}{2}}) \quad N \gg 1 \tag{3.88}$$

In the absence of positioning uncertainty or projection, we see that the distance between nearest neighbors diminishes very slowly and so does the tail of the distribution which gives an idea of the large distances typically reached. Let us consider a sampling of $N = 10^3$ points. We choose $\epsilon \simeq \frac{1}{N} = 10^{-3}$ level leading to an estimated pure sampling resolution limit of $\simeq \frac{\pi}{20}$. This value is relatively constant when the sampling increases. In the dense sampling limit, the resolution limit is set by the Nyquist criterion as in a flat surface: $O(N^{\frac{1}{2}})$. This is still far from the limit observed in our simulations at least twice as large.

Variances calculated from the observed Fisher information matrix on the completion angle are very often much larger that the effective errors $\theta_0 - \theta_*$ and spread over the whole value interval. To enhance figure readability but yet deliver the information, the error bars shown on the figure 3.17 correspond to a third of the standard deviation calculated from the estimated variance. Again, the variance estimated from the Fisher Information matrix appears as an upper bound in the order of magnitude of the error, but is not of any interest as the size of the search interval is in the same range. The average variance is equivalent to an estimated uncertainty (standard deviation) of 0.57 rad whereas the effective error is 0.24 rad. But the variance of the actual error depends on the completion and on the tilt whereas this trend is not found in the observed Fisher information matrix coefficient. The details of the distribution of error normalized by this estimated variance is given in figure 3.19b.

The large variance and inaccuracy of the maximum likelihood completion estimator close to the interval boundaries limits its relevance for a single particle. We investigated further the distortions that it would induce at the scale of a larger population. We selected various subset of the simulated particles to see how localized distributions are distorted in the estimation process. Original subset and distorted distributions are shown on figure 3.20. Each estimate contributes as a Gaussian centered on the maximum likelihood value with standard deviation given by the inverse of observed Fisher information matrix $\sqrt{I^{-1}_{\theta\theta}(\vec{\beta}^*)}$. Each contribution is normalized over the allowed completion interval $[\pi/2, \pi]$ to guarantee that each particle has the

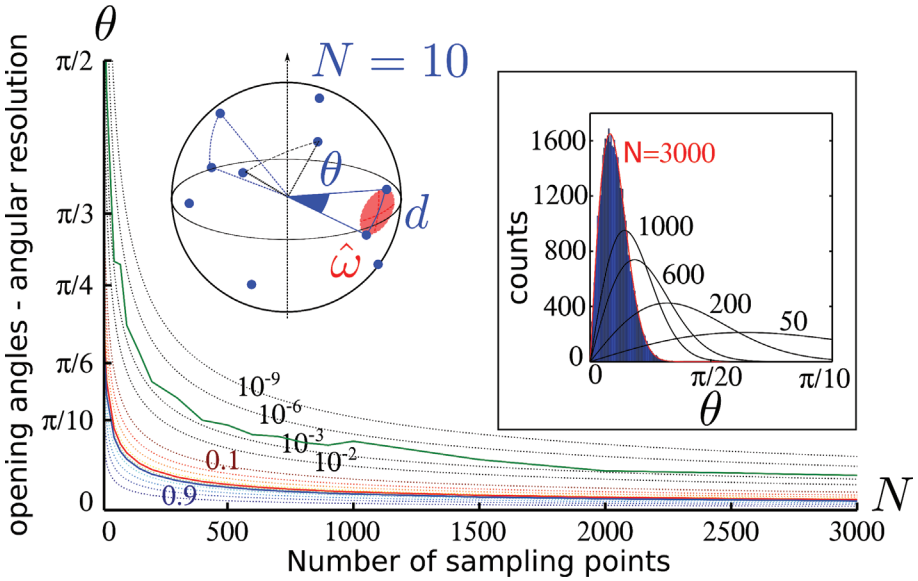


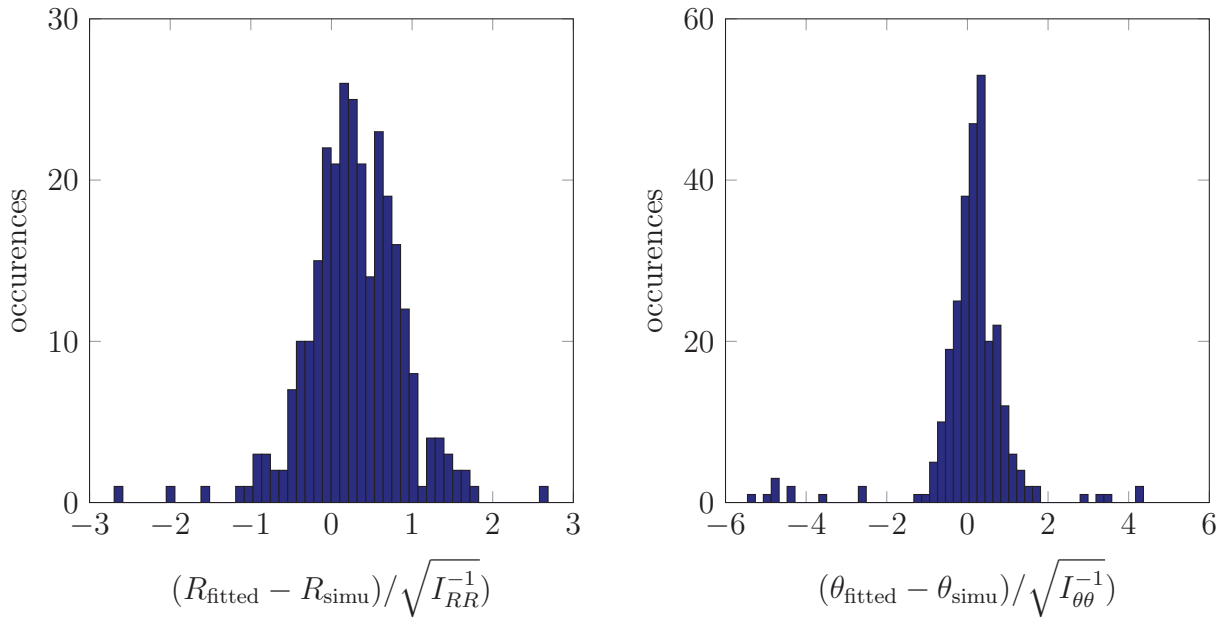
Figure 3.18 – Limiting resolution for completion estimation due to finite sampling effect: the dashed lines correspond to approximated confidence upper bound for various acceptable error $\epsilon \in \{0.9, 0.8, \ldots, 0.1,\ 10^{-2}, 10^{-3}, \ldots\}$. The average values are plot with solid lines: mean distance between points $\bar{\theta} = cos^{-1}(1 - \frac{2}{N})$ (blue) and its simple approximation $\bar{\theta}\ N^{-\frac{1}{2}}$ (red). The maximal distance between two points observed averaged on $5.\,10^3$ simulations for each sampling N is displayed (plain green). Inset: Several distribution of nearest neighbor distance expressed as angle at the sphere center for various number of sampled points N

same weight in the constructed distribution before. All contribution are eventually added to form the estimated distribution. Distributions are displayed with unit area.

As already pointed out the variances are overestimated and the spread of the distributions are largely exaggerated. The distortion of the uniform distribution is surprising low: the density at the interval center is slightly overestimated since wrong estimations at the boundaries are likely to largely redistribute density on the whole interval (including its center) whereas the accuracy is better at the center, creating a positive balance. Depletion is more sensible for complete particles whose projection are always isotropic, presumably a difficult case for the estimator. However, the estimated distribution of a population of complete particles remains unimodal figure 3.20b with a broad maximum in the vicinity of $\frac{9\pi}{10}$. A larger distribution leads to a mode displaced toward the interior of the interval ($\frac{4\pi}{5}$ in figure 3.20d).

3.7 Application: reconstruction from PALM images of budded HIV-1 virus like particles

We collaborate with Julia Gunzenhäuser and Suliana Manley from the Laboratory of Experimental Biophysics at the Ecole Polytechnique Federale de Lauzanne to apply the analysis on PALM images of purified immature HIV-1 Virus-Like Particles (virus like particles). Cell culture, transfection and imaging were made by Julia Gunzenhäuser according to the protocol described in the following.



(a) Distribution of the estimator error on radii normalized by the std. estimated using inverse observed Fisher information matrix $\frac{R_* - R_0}{\sqrt{I_{RR}^{-1}(\vec{\beta}^*)}}$

(b) Distribution of the estimator error on completion normalized by the std. estimated using inverse observed Fisher information matrix $\frac{\theta_* - \theta_0}{\sqrt{I_{\theta\theta}^{-1}(\vec{\beta}^*)}}$

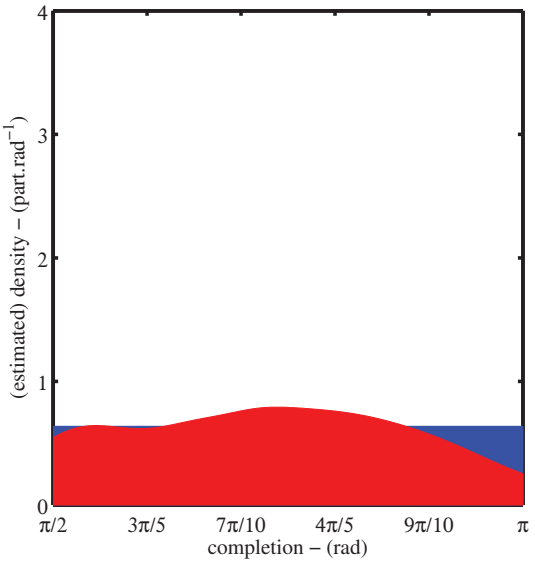
Figure 3.19 – Estimation error of the maximum likelihood standardized using the inverse Fisher information matrix.

3.7.0.1 Cell culture, transfection and virus like particles extraction

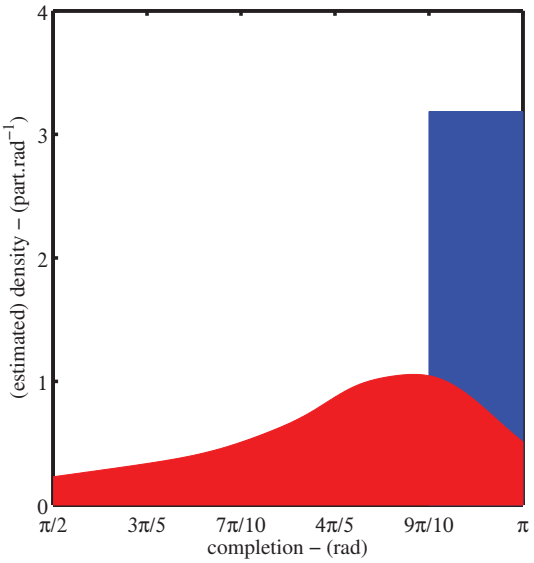
African green monkey kidney cells (Cos7) were cultured in DMEM supplemented with 10% FBS (Sigma Aldrich). For virus like particle production 600,000 cells were grown in T75 flasks and transfected with 34 μg of Gag-mEos2 plasmid (detailed description is available in Gunzenhäuser et al.[59] article) and 100 μl FuGene6 (Roche Diagnostics) in a total volume of 1 ml DMEM without FBS incubated for 15 min. 48 hours post transfection the supernatant was collected from the cells and filtered through 0.45 μm filters. For virus like particles extraction the supernatant was centrifuged over a 20% sucrose gradient at 27000 rpm for 2 hours at 4°C. The pellet was dissolved in filtered PBS and the virus like particles solution was directly used for imaging or stored for not more than 24 hours at 4°C prior to imaging. For imaging poly-L-lysine coated coverslips containing 100 nm Au fiducial markers were incubated with virus like particles for 1 hour at 4°C, rinsed with PBS and directly used.

3.7.0.2 Superresolution imaging

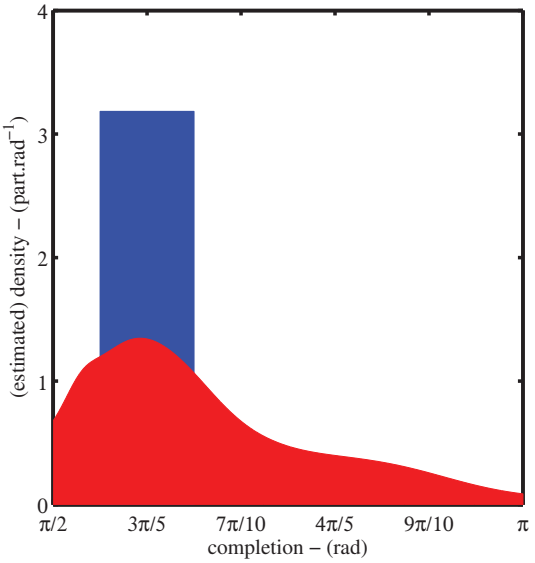
Virus like particles were imaged using a Zeiss Axio Observer D1 inverted microscope, equipped with a 100×, 1.49 NA objective (Zeiss). Activation and excitation lasers with wavelengths 405 nm (Coherent cube) and 561 nm (Crystal laser) illuminated the sample in total internal fluorescence (TIRF) mode. We used a four color dichroic 89100bs (Chroma), fluorescence emission was filtered with an emission filter ET605/70 (Chroma) and detected with an electron-multiplying CCD camera (iXon+, Andor Technology) with a resulting pixel size of



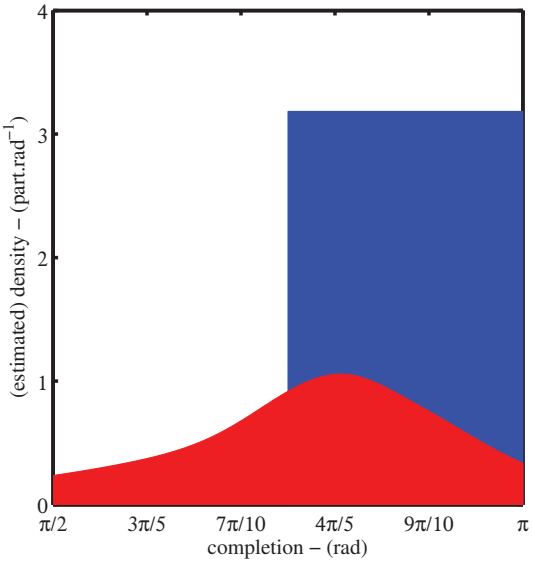
(a) uniformly spread on the whole completion interval $\theta_0 \sim \mathcal{U}_{[\pi/2,\pi]}$



(b) all particle are fully complete spheres $\theta_0 > \frac{9\pi}{10}$



(c) original particles distribution is essentially unimodal



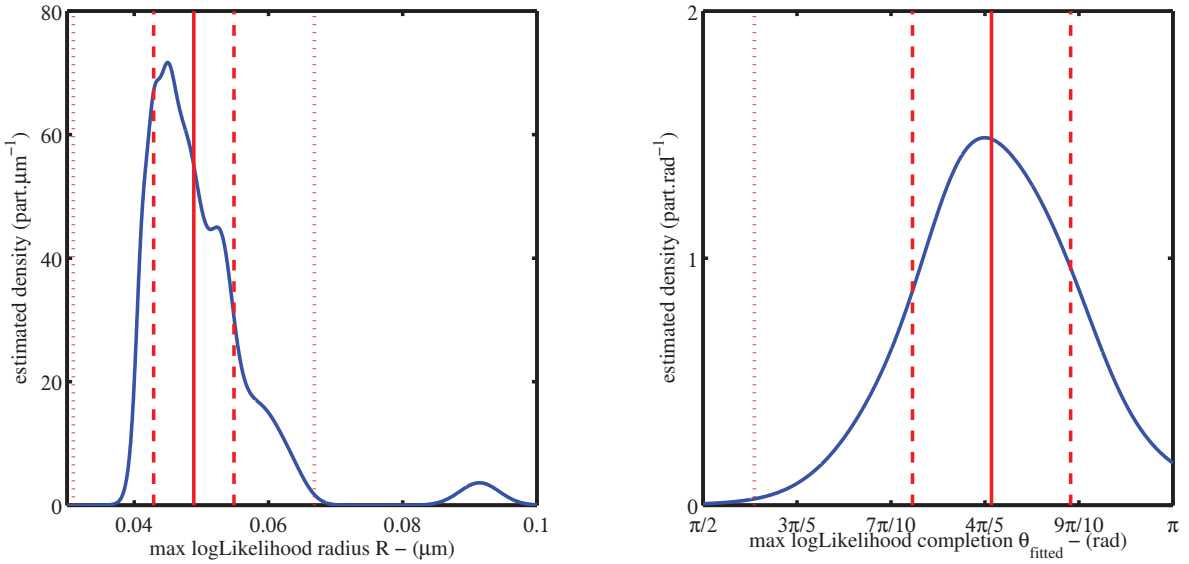
(d) particles are uniformly spread on half of the interval $\theta_0 \sim \mathcal{U}_{[3\pi/4,\pi]}$

Figure 3.20 – Deformation of the distribution by the estimation. The original distributions of completion in the simulated data are plot in blue and the maximum likelihood estimation of the distribution is plot in red.

160 nm. For each region of interest, typically 30000 to 40000 images of a 20.5 x 20.5 μm^2 area were collected with an exposure time of 30 ms. The irreversible photoactivatable protein mEos2 was activated with low continuous 405 nm laser intensity to guarantee very sparse activation and minimize blinking, and excited with 561 nm laser intensity of ≈ 1 kW.cm^{-2}. Molecules were localized using Peakselector (IDL, courtesy of Harald Hess). The single molecule localization procedure consisted of the following steps: a) fluorescent intensity peaks were detected on each image, b) each peak was fitted to a two-dimensional Gaussian by nonlinear least-square fitting to obtain x and y coordinates as well as the localization precision, c) images were "dedrifted" using Au fiducial markers, d) localizations detected within less than the measured mean localization precision (typically between 19 and 24 nm) in space and 300 ms in time were grouped to account for blinking of mEos2. One grouped molecular position is counted as one Gag-mEos2 protein (see Gunzenhäuser et al.[59]).

3.7.0.3 Superresolution analysis results

Dense clusters of molecules were chosen to be submitted to the maximum likelihood analysis described in the previous section. The data consisted of 33 clusters of N superresolved position triplets $\{x_j, y_j, \sigma_j\}$, N ranging from 714 to 3302 proteins and $\bar{\sigma}$ ranging from 15 nm to 21 nm. Results are shown on figure figures 3.21a and 3.21b.



(a) estimated particle radii distribution R_* (b) estimated particle completion distribution θ_*

Figure 3.21 – Sizes and completion distribution estimated using the maximum likelihood estimation based the truncated sphere model for the labeled proteins distribution (blue line). We emphasis the position of the mean (plain red line) and standard deviation (±1std. large red dashes, ±2std. small red dashes)

The distribution of estimated radii (figure 3.21a) is the sum of normalized Gaussians centered on each of the estimated radii with variance given by the first diagonal coefficient of the corresponding observed Fisher information matrix. Hence a radius estimated with a large error bar contributes on a large interval and each particle has the same weight in the total distri-

bution. The resulting distribution shows a strong peak at R=45 nm and an extended tail for larger radii. The major peak is characterized by a mean and standard deviation of 49 nm and 6 nm respectively. One object in the sample exhibits a radius much larger than the mode (up to twice the main peak radius). The outsider particle figure 3.22 exhibits an elongated shape and is likely to be an aggregate or an ill-formed particle for which the procedure produces abnormal result. Other apparent aggregates of virus like particles as well as clusters with abnormal shape, obviously not consistent with the truncated sphere model, were present in the field of view but not selected for the fit.

The fluorescent protein tags imaged are attached to the Gag N-terminus, located at the inner surface of the Gag layer in the virus like particle. Literature values from cryo-EM measurements give virus like particle sizes in terms of outer diameters. To take this into account for the purposes of comparison, we added the Gag length of 25 nm to the size of the measured particles reported by Wright et al.[134]. The adjusted mean radius of HIV-1 virus like particles in our measurement is on average 74 ± 6 nm, larger than values from cryo-EM of 66 ± 9 nm in Wilk et al.[133] work or 65 ± 17 nm according to Carlson et al.[20]. The size obtain is closer to the values reported by Briggs et al.[16] for mature particles 73 ± 12 nm.

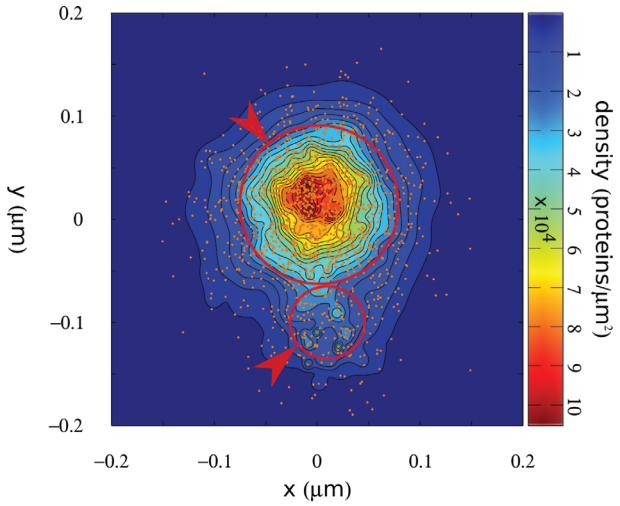


Figure 3.22 – Palm image of the protein cluster responsible for the peak at large sizes. Dots correspond to the most probable position for the detected labels whereas colors maps the density of proteins as when each detection contribution is a Gaussian with a width given by positioning uncertainty. Red circles show two putative aggregated particles

The distribution of completion angles estimated from the data is shown in figure 3.21b. This distribution is constructed using the protocol used for the radii with the difference due to the interval finite size normalization was performed on the portion of the Gaussian on the allowed completion interval $[\pi/2, \pi]$. This guarantees that each particle has the same weight in the constructed distribution. This distribution shows a single peak located at $\theta \approx \frac{4\pi}{5}$ rad, with a standard deviation of 0.3 rad. As discussed previously, the estimation of the completion is likely to produce inaccurate results and deform the distribution at the boundaries. The shape of the distribution obtained from real data is not qualitatively consistent with simulated scenarios the full completion of all the object neither with a uniform distribution on the interval figure 3.20a. The results might indicate that the method is sensitive to the 2/3 surface coverage reported

by Carlson et al.[20]. However the value of the peak is only indicative of a range of completion around this value for the imaged virus like particles and should be treated as such.

3.8 Conclusion

We have implemented and tested an estimation procedure to extract structural information from superresolution images from (f)PALM and STORM methods when a parametric structure is assumed for the labeled object. The method includes two successive steps.

In a first step the image is centered on its center of mass and oriented based on the moment method principles. At this stage, only the three first moments of the experimental position distribution are used which is the minimum requirement to achieve centering and orientation. This both limits the analytical and computational calculations while good robustness towards positioning error and outliers measurements is still guaranteed. In order to use the resulting centered and oriented image, one should know the model density values in the reference frame centered on its center of mass. The center of mass position is thus the only required knowledge. The orientation of the centered object used in this work apply to any object with symmetry axis and take the measured position as only input (no analytical derivation of the second order moment coefficients is necessary). We further precise how to control the relevance of the orientation by a direct calculation from the measured positions to estimate the orientation variance.

In a second step, the remaining model parameters are estimated by the mean of maximum likelihood method. We discussed the natural description of superresolution data in terms of statistical sampling in a projected probability density which brings us to this method of estimation. The well established maximum likelihood method is natural and simple to use when one needs to find the best match inside a family of parametric probability density function. The major difficulties are technical. First the evaluation of the probability density of measurement including the positioning uncertainty consists in a Gaussian convolution and a projection (the order is arbitrary and should be decided for convenience). Both are very classic mathematical steps, however greedy in computational resources when the required precision is high. This issue is compounded when the model density diverges at some points as for a projected surface density in our case. A bulk model, with protein distributed in volume, would be simpler to treat. We proposed a solution to regularize the density in the spherical case before the numerical Gaussian convolution. We detailed what we found to be the most efficient computation strategy in this scheme. Finally we set up the condition to use an optimization solver to efficiently find the maximizer of the loglikelihood function, limiting the number of evaluations.

The method was tested on simulated data to investigate which information could be extracted from a known content. We thus investigated the methods statistical performance on truncated spherical object with a sampling range compatible with the experiment.

As described in section 3.3, the modeling of the imaging process was simplified and translated in an effective procedure that would describe a theoretically perfect microscope. A discrepancies between the simulated precision and the experimental situation is expected whenever factor of error are not taken into account in the simulations. Concerning the physics of imaging the following optical aspect can be mentioned: in the TIRF configuration used to image the particle, the excitation field decreases with depth. As a consequence, the photon yield and the positioning precision also decreases with depth. We did not take this effect into account on the simulated the images. We basically took into account precision fluctuations by randomly

assigning the measurement precision from typical distribution to each sample position. In this work, the estimation method is typically designed for object whose size is of the order of the measure uncertainty up to a few times this value and so smaller or a worse comparable to the penetration depth (estimated at several hundreds of nanometers compared to HIV typical diameter of 140 nm). Modulation of the emission intensity by the excitation intensity decay should be typically, in the worse case, in $[1, \exp(1)]$ and, precision that goes as the square root of the photon number (of the intensity), in $[1, \exp(\frac{1}{2})]$. So relative fluctuations are of 30%-40% comparable with the stochastic fluctuations of photon yield that we simulated. Were the simulation method to be used on larger objects, the positioning precision would certainly convey information about the label depth and the shape of the image might also be altered to a larger extend. In this case, to obtain the precision of likelihood estimation procedure by simulations would require to generate more realistic images. In contrast, this does not affect the symmetry of the image in 2D that is crucial for our centering and orientation methods.

We found that, in the truncated sphere case simulated in the test, the projected area of the model was the dominant contribution to the likelihood, followed by the overall shape of the projection. However, in the tested conditions of shape and sampling, distinction between different parameters sets based on their inner variations of density seems a harder task. The study of the global variations of the loglikelihood function on various test case provided a good insight of the regions that conflict for determination. In such conditions we might need a restriction of the parameter space to achieve a consistent estimation. This restrictions are supplementary hypothesis on the model to make a decision between competitive possibilities. They must be representative of the physical knowledge of the data submitted. Restriction of the parameter space is the way to incorporate information on the data in addition to the model within the strict likelihood theory. Were the supplementary informations available in the form outcomes of penalization compared to other in a probabilistic way, the theory of Bayesian estimation using a prior probability could be considered but is not in the scope of our work.

We tested the performance of the method on simulated images generated out of a large population of truncated spheres with more than half completion. Accuracy of the method with a sampling of $N = 1500$ positions are of the order of the nanometer (≈ 1.3 nm), much smaller than the localization uncertainty of a single emitter (≈ 20 nm). The accuracy is constant over the tested range, even with radius as small as positioning precision itself.

Conversely, we found that the estimation of sphere completion using the reconstruction procedure is less successful. A confusion between different parameters sets is likely to happen when their related density have a similar projected contour even if the 3D reality is much different. This is a limitation probably contributing to this reduced efficiency. This is particularly clear for nearly complete spheres and hemispheres viewed from the top. The projection hardens the determination of the orientation of the particle in the focal plane which impacts the completion fit validity. Those effects accounts for the large spread at the extremes of the interval. Eventually, a part of the complete particle are falsely identified as half complete and vice-versa. Overall the balance of wrongly estimated particle is positive at the center of the interval which results in an overestimation of the number of particle with completion in $\left[\frac{2\pi}{3}, \frac{5\pi}{6}\right]$ and negative close to full completion.

The aforementioned reduced efficiency for the estimation of completion at the single object level only partially distorts an ensemble measurement. Nevertheless, we observed with the simulated data that for completion values uniformly sampled on an interval $[\theta_-, \theta_+]$, the maxi-

mum likelihood estimation provides completion values centered on this interval, and marginally spread values outside of this range. This information can be therefore used in order to roughly define the range of variation for the completion of the imaged objects (see figure 3.20). Yet, a better solution would be to use data with 3D information to avoid the loss of information inherent to projection. Depending on the solution chosen to implement the 3D imaging, the positioning precision is likely to be inhomogeneous with depth in addition to the effects originating from the TIRF configuration. This effect has been investigated theoretically by simulation of the optical devices in Badieirostami et al.[7]. In both bi-plane and astigmatic setup, the expected fluctuations of the positioning precisions are typically lower than the variation of photon yield by the inhomogeneous TIRF excitation field or the fluorophore stochastic emission. But in the most commonly used astigmatic setup, the deformation of the point spread function give rise to an anisotropic localization precision in the (x, y) plane (see figure 3.23), thus the overall shape of the image is deformed in the plane compare to the original object which impairs the orientation efficiency of the moments method. The bi-plane imaging method has the advantage to keep a theoretical precision isotropic in (x, y) and relatively constant with depth within the imaging volume of depth 600 nm chosen for the simulation.

However, in the reverse problem of estimating the best match for a given image, the positioning precision is directly estimated from the measurements. Spatially inhomogeneous positioning error should primarily impact the orientation of particle found by the moment method. Then the validity of the average positioning error approximation we used can be limited by the evolution of the positioning error with depth. An important variation of positioning error with depth would demand to take into account each point positioning error. This would require much longer computation but the likelihood method exposed in this work would remains valid. This is then on the side of the superresolution data treatment that limitations may arise preventing a correct estimation of the positioning uncertainty.

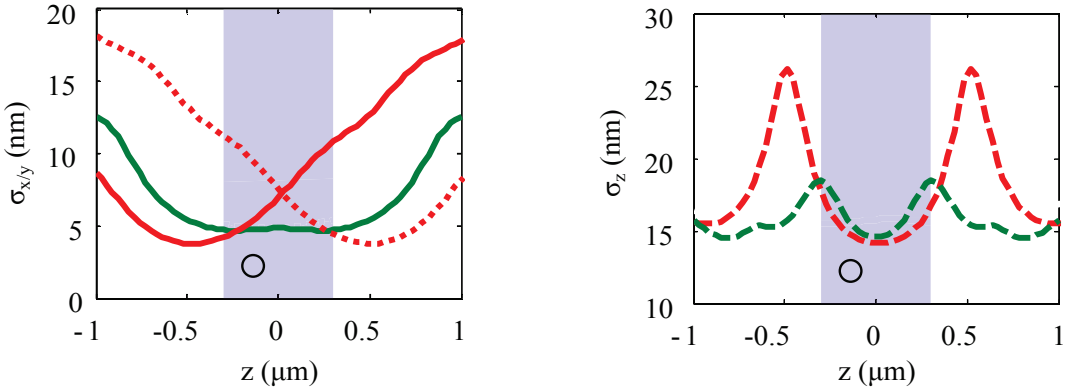


Figure 3.23 – Minimal theoretical positioning error evolution with emitter distance z to the focus plane for two 3D superresolution imaging setups: astigmatic setup (red) or two planes setup (green). Simulations were made using a fixed number of photons per emitter ($N = 10^3$) at wave length $\lambda = 630$ nm on a CCD of 25×25 pixels of effective width 160 nm in the focus plane, with a uniform background noise of 2 photons/pixel. Both simulated setups are designed to reconstruct the z position over a depth $\Delta z = 600$ nm (blue area). The typical size of the human immunodeficiency virus is show (black circle). The plot on the left shows the in plane positioning errors σ_x (dots) and σ_y (plain line), the plot on the right shows the error along optical axis σ_z (dashed). (adapted from Badieirostami et al.[7])

We analyzed PALM images obtained on immature HIV-1 virus like particle using the developed method in light of our simulations. The distribution of observed radii (figure 3.21a) is characterized by a mean and standard deviation of 49 nm and 6 nm respectively. The fluorescent protein tags that are imaged are attached to the Gag N-terminus, located at the inner surface of the Gag layer. Literature values from cryo-EM measurements give virus like particle sizes in terms of outer diameters. To take this into account for the purposes of comparison, we add the reported Gag length of 25 nm to the size of the measured particles (Wright et al.[134]). The adjusted mean radius of HIV-1 virus like particles in our measurement is on average 74 ± 6nm, larger than values measured in T-cells from cryo-EM of 66 ± 8 (Wilk et al.[133]) and 65 ± 17nm (Carlson et al.[20]). Many factor affect the measure. The fitted data were obtained in virus like particle grown from a different cell line that in our reference article and this is known to affect the virus formation. Furthermore, our estimation of the virion radius from the label measured radius take into account neither the label size, which is of the order of 4 nm (Bates et al.[11]), not the possible effect on the particle structure of the fused protein. On the other hand, the order of magnitude is credible and the approach allows to quantitatively estimate spatial feature of the imaged object with greater precision than the localization precision limit, providing insight of the variation in a sample for instance. In this sense, we note that the analysis was precisely able to point out an ill formed HIV-1 virus like particle in the submitted set of objects (see figure 3.22).

We were also able to estimate the completion distribution. The main peak of the completion distribution is found at approximately $\theta = \dfrac{5\pi}{6}$ rad. As discussed previously, the value of the peak is only indicative of a range of completion around this value for the imaged virus like particles. Due to the increased uncertainty of the maximum likelihood estimation for closure, rather than giving a quantitative interpretation, we show that we can distinguish between complete versus incomplete closure. Our finding of incomplete closure is consistent with the 2/3 surface coverage reported in Carlson et al.[20]. These results are presented in an article submitted for review at Biophysical Journal.

The maximum likelihood reconstruction method proposed in this work can be broadly applied to analyze at the single object level other features imaged by superresolution-microscopy with known parameterizable shape. However, it is not possible to extract from the data more information than it carries, even though the optimization procedure always provides a result. It is therefore crucial to evaluate the amplitude of fluctuations in the estimated parameters for a fixed set of parameters using control tests as described in this article or theoretical calculations when available. For instance, we were also willing to reconstruct geometries for simulated budding sites. Due to the small number of sampling points, the strong positioning uncertainty, small observation angles ϕ, and closely related shapes, estimated parameters would probably not be meaningful. In such cases, improvements to the data such as 3D localizations or reduced positioning uncertainty $\bar{\sigma}$ would be required to make further progress. Brighter synthetic dyes used in STORM technique could potentially give a better positioning precision that the one reached in our data (see Malkusch et al.[88]). Fortunately, technological advances in the field continue to improve image quality, and will allow maximum likelihood to become a more powerful tool in the future. Overall, maximum likelihood estimation applied STORM and (f)PALM data appears a promising method to obtain quantitative measurements on structures, especially those showing variability.

Conclusion

In this work, we addressed the question of the self assembly of proteins during the formation of the viral bud at the cellular membrane. Using the framework that describe close system assembly as *in-vitro*, we studied the formation of the virus as a succession single protein additions which is known as the assembly line model. We extended the assembly line model to open systems and studied the response of the system to a constant production of proteins on time according to various scenarios of binding kinetics.

At steady state, the capsomers leaving the cell as complete virus balance the capsomer production and the flux of capsomers throughout the system maintain it out of equilibrium. This is the main difference with the closed system. In closed system the equilibrium size distribution shows either complete capsids or capsomers in solution (Zlotnick[139]). No intermediates can be seen. However, the capsids formed in the in-vitro experiment are more stable than the thermodynamical equilibrium would allow, suggesting a kinetic control of the reaction ([119]). The line assembly model including a nucleation step in a system containing an initially pure solution of capsomers, indeed shows a time evolution either leading to a quasi-equilibrium size distribution or a kinetically frozen system (Morozov et al.[98]).

In the open system, when the mean bending energy dominates the work to assemble a capsid and no nucleation effect is considered, we found that all the intermediates from capsomers to closed capsids, are likely to be present in the same proportions at steady state, which is not compatible with the experimental observations. If the nucleation barrier is taken into account, using the line tension model, all the intermediates populations are depleted accordingly compared to the capsomers, yet intermediates beyond the typical nucleus are found in the same amount resembling the distribution of clusters observed at the cell membrane (Gunzenhäuser et al.[59]).

We studied the dynamic of the aggregation. The time evolution of the system is controlled by the drift-diffusion properties of the equation and by the non linear coupling with the monomer consumption. Compared to a closed system, the size distribution of the open system also looks like a transportation front. However, this front does not stop to asymptotically tends to the equilibrium distribution since capsomers are constantly added in the system. So the front eventually reaches the absorbing boundary. **Our main finding is that nucleation barrier favors production of viruses by successive bursts in early production time and that, according to the model, the observed size distribution should be peaked with most of the growing viruses having the same sizes.** We have shown that spacing between successive bursts depends on the protein number in complete viruses. We also pointed out a domain of low flux where the system does not reach the steady state and oscillations are

maintained.

In future study, the effect of the membrane mechanical deformation on the dynamic is to be explored. It is expected that the bending change the dynamic and steady state by generating a second barrier in the free energy landscape. The constant input flux is a crude model of protein production in biological condition. More elaborated dynamic of the capsomers generation could be tested now that the behavior is clarified in the simple case.

We have also considered a description of the time evolution of a single virus. Under the assumption of a constant monomeric concentration we have proposed a analysis to experimentally estimate the microscopic rates of aggregations. This analysis is however limited to the case of a steady capsomer concentration. Would this requirement be satisfied in the experiment, the method is a simple answer to the stochastic nature of the data. The modified Gillespie scheme we built provides single particle trajectories that can be used as basis to simulate fluorescence signal including experimental artifacts and evaluate the performance of the average solution for real data treatment.

In addition of the model of virus growth, we have also developed a method to extract morphological measurement on the superresolution images. To our knowledge no parametric estimations had been previously attempted from superresolution point localization microscopy. Our first aim was to reconstruct the shape of the budding virus directly on the membrane from the superresolution data. We started however on the simpler analysis of complete virions released from the cell and purified. In this situation, the background noise is lower and the superresolution precision expected higher. The method we propose is based on maximum likelihood estimation theory. It requires a parametric model of the labels distribution on the object of interest and the superresolution data. The most likely parameters set given the data is then calculated. We designed a procedure to simulate the PALM image from ideal structures to test the procedure on known data. We set up a first procedure to orient the images along their symmetry axis using the moments of the image. **Our tests shown that it is indeed possible to extract geometric measurement with finner precision than the positioning measurement.** Our study suggests that the likelihood estimation performs better on distinguishing fluctuations of the contours and might need larger sampling density to be consistent at finner level. We were able to determine the radius of the simulated truncated sphere with an excellent accuracy (a precision finner that 3 nm for $\sim 10^3$ localized labels with an uncertainty in position on $\simeq 20$ nm) whereas their completion was not well determined for several projection angle. Test on real virus like particles gave realistic results compare to the literature. Our tests suggest that our method is not efficient enough for our primary purpose to study the shape of the budding viruses at the membrane cell, at least with the data at hand. In the future it would be interesting to challenge the method with 3D data and labeled structures of calibrated geometries.

Appendix A

Linear fit in the least square sense

A.1 Position of the problem

A polynomial relationship is expected between the variable x tuned during an experiment and the measured values y:

$$y = ax^2 + bx + c \tag{A.1}$$

We are interested in the values of the model parameters $\vec{\beta} = (a\,b\,c)^t$ that best explains the produced data.

In the case of an infinite precision in the determination of y, three measurement with different value of x are enough to fully determine the parameters by solving:

$$\left\{\begin{array}{ll} y_1 & = ax_1^2 + bx_1 + c \\ y_2 & = ax_2^2 + bx_2 + c \\ y_3 & = ax_3^2 + bx_3 + c \end{array}\right. \quad \equiv \quad \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ x_3^2 & x_3 & 1 \end{pmatrix} . \begin{pmatrix} a \\ b \\ c \end{pmatrix} \quad \equiv \quad \vec{Y} = X.\vec{\beta} \tag{A.2}$$

As the measure is not error-free, the values of y are not known with infinite precision, and we get instead:

$$z_i = y_i + \epsilon_i \tag{A.3}$$

Where ϵ_i is the measurement error. Thus the inversion of the system in equation (A.2) for three chosen values of x, $\{x_1, x_2, x_3\}$ corresponding to three rounds of the experiment:

$$\left\{\begin{array}{ll} z_1 & = ax_1^2 + bx_1 + c \\ z_2 & = ax_2^2 + bx_2 + c \\ z_3 & = ax_3^2 + bx_3 + c \end{array}\right. \tag{A.4}$$

will either lead to a result $\{a, b, c\}$ that depends on the choice of the selected rounds of measure, or be impossible (for instance two measures with the same values of x, yielding different y obviously generate an incompatible system).

A.2 Least square method and generalization

The least square method to estimate the optimal parameters of a linear model from the measured values was first published by Legendre[82] and which benefited from Friedrich Gauss

major treatment regarding it links to probabilities. Its core idea is that the best estimate of the parameters is the one that minimizes the square of the distance between the values predicted from the estimated parameters and the measured values:

$$\vec{\beta}^* = \underset{\vec{\beta}}{\operatorname{argmin}}\, \mathrm{d}(\vec{Z}, \vec{Y}(\vec{\beta}))^2 \tag{A.5}$$

With $S(\vec{\beta}) = \mathrm{d}(\vec{Z}, \vec{Y}(\vec{\beta}))^2$ often called the score function. "Best" as estimator property is understood as "with minimal variance" among the linear unbiased possible estimators. The Gauss-Markov theorem demonstrates this property in the case of measurements affected by uncorrelated errors of equal variance and the euclidean distance (neither the exact form of the distribution nor independence or identical distributions of the error is required). Generalization to correlated errors with known covariance matrix Ω was made by Aitken[3] giving the Mahalanobis distance $\mathrm{d}(\vec{Z}, \vec{Y})^2 = (\vec{Z} - \vec{Y})^t \Omega^{-1} (\vec{Z} - \vec{Y})$ as the proper one regarding this case. This is formally equivalent to apply a linear transformation on the variable to scale and decorrelate the error and then use the least square estimator on those modified variables. Linear model means that there is a linear relationship between the parameters and the observed values for given variables values. The model in our example below is linear as any polynomial relationship between variable and observation.

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} x_1^2 & x_1 & 1 \\ x_2^2 & x_2 & 1 \\ & \vdots & \\ x_n^2 & x_n & 1 \end{pmatrix} . \begin{pmatrix} a \\ b \\ c \end{pmatrix} \tag{A.6}$$

A.3 Normal equations

The normal equations correspond to the mathematical solution of the minimization defined in equation (A.5). We propose thereafter a derivation in the general case of correlated error with known covariance matrix Ω (symmetric and positive definite):

$$\begin{aligned}
S(\vec{\beta}) &= (\vec{Z} - \vec{Y})^t \Omega^{-1} (\vec{Z} - \vec{Y}) \\
&= \vec{Z}^t \Omega^{-1} \vec{Z} - 2\vec{Z}^t \Omega^{-1} \vec{Y} + \vec{Y}^t \Omega^{-1} \vec{Y} \\
&\quad \text{(we used the fact that a scalar is not affected by transposition: } \vec{Y}^t \Omega^{-1} \vec{Z} = \vec{Z}^t \Omega^{-1} \vec{Y}) \\
&= \vec{Z}^t \Omega^{-1} \vec{Z} - 2\vec{Z}^t \Omega^{-1} . X . \vec{\beta} + \vec{\beta}^t X^t \Omega^{-1} X \vec{\beta} \\
&\quad \text{differentiating with respect to } \vec{\beta} : \\
\partial_{\vec{\beta}} S &= -2\vec{Z}^t \Omega^{-1} X + 2\vec{\beta}^t X^t \Omega^{-1} X \\
&\quad \text{conditions for minimum: } \partial_{\vec{\beta}} S|_{\vec{\beta}^*} = 0 \quad \& \quad X^t \Omega^{-1} X > 0 \\
\vec{\beta}^* &= (X^t \Omega^{-1} X)^{-1} (\Omega^{-1} X)^t . \vec{Z}
\end{aligned} \tag{A.7}$$

Equation (A.7) are called the normal equations of the least square method and are theoretically sufficient to perfectly solve the generalized least squares problem (classic least square correspond to the particular case $\Omega = \mathbb{1}$ so that Mahalanobis distance is equivalent to the euclidean one). Uniqueness of the solution is equivalent to have linearly independent columns in the matrix X.

A.4 Error propagation - confidence interval

With the normal equation, we see that there is a linear dependency between the observation vector and the estimated parameters: $\vec{\beta}^* = J_X.\vec{Z}$ and that in absence of error we would get $\vec{\beta}_0 = J_X.\vec{Y}$. We know by hypothesis that the measurement is made with an error $\vec{\epsilon}$ with zero mean and variance Ω:

$$\vec{Z} = \vec{Y} + \vec{\epsilon} \quad | \quad \mathbb{E}\left(\vec{\epsilon}\right) = \vec{0} \quad \& \quad \mathbb{E}\left(\vec{\epsilon}.\vec{\epsilon}^t\right) = \Omega \tag{A.8}$$

Thus the covariance matrix of the estimated parameters is such that:

$$\begin{aligned} \mathbb{E}\left((\vec{\beta}^* - \vec{\beta}_0).(\vec{\beta}^* - \vec{\beta}_0)^t\right) &= \mathbb{E}\left(J_X.\vec{\epsilon}.\vec{\epsilon}^t.J_X^t\right) \\ &= J_X.\,\mathbb{E}\left(\vec{\epsilon}.\vec{\epsilon}^t\right).J_X^t \\ &= J_X.\Omega.J_X^t \end{aligned} \tag{A.9}$$

This formula is well known as the "propagation of error". Last transformation comes from the linearity of the expectation, A being a constant matrix: $\mathbb{E}\left(A.X\right) = A.\,\mathbb{E}\left(X\right)$. Eventually, replacing $J_x = (X^t\Omega^{-1}X)^{-1}(\Omega^{-1}X)^t$ leads to the final formula for the covariance matrix of the estimated parameters:

$$\begin{aligned} \mathbb{E}\left((\vec{\beta}^* - \vec{\beta}_0).(\vec{\beta}^* - \vec{\beta}_0)^t\right) &= (X^t\Omega^{-1}X)^{-1}(\Omega^{-1}X)^t.\Omega.\Omega^{-1}X(X^t\Omega^{-1}X)^{-t} \\ &= (X^t\Omega^{-1}X)^{-1}X^t.\Omega^{-t}.X.(X^t\Omega^{-1}X)^{-t} \\ &= (X^t\Omega^{-1}X)^{-1}X^t.\Omega^{-t}.X.(X^t\Omega^{-t}X)^{-1} \\ &= (X^t\Omega^{-1}X)^{-1} \end{aligned} \tag{A.10}$$

In the particular case of uncorrelated errors of variance σ^2 in the original least square method, it gives:

$$\mathbb{E}\left((\vec{\beta}^* - \vec{\beta}_0).(\vec{\beta}^* - \vec{\beta}_0)^t\right) = \sigma^2.(X^t.X)^{-1} \tag{A.11}$$

When the error is assumed normally distributed, the variance σ^2 can also be estimated from the residuals of the fit. $S(\vec{\beta}^*)$ is then the sum of the square of identically distributed independent normal random variables and thus $S(\vec{\beta}^*)/\sigma^2$ follows a χ^2 distribution with $n-m$ degree of freedom with expectation $n - m$. So:

$$(\sigma^*)^2 \simeq \frac{S(\vec{\beta}^*)}{n-m} \tag{A.12}$$

A rough derivation of the confidence region would consist in first estimating the unknown covariance of the error and then use this estimate to build the parameters covariance estimate to eventually assume a normal distribution. This would lead with a α confidence level:

$$(\vec{\beta}^* - \vec{\beta}_0)^t.X^tX.(\vec{\beta}^* - \vec{\beta}_0) \leq \frac{S(\vec{\beta}^*)}{n-m}(\sqrt{2}\,\mathrm{erf}^{-1}(1-\alpha))^2 \tag{A.13}$$

This procedure leads to a strong underestimation when the degree of freedom is low $n - m$ (few numbers of points compared to the umber of variables). The correct choice, asymptotically

equivalent as $n \gg m$, consists in constructing the ratio of the independent scalar variables following χ^2 distributions[121]:

$$\frac{(\vec{\beta^*} - \vec{\beta}_0)^t.X^tX.(\vec{\beta^*} - \vec{\beta}_0)}{\sigma^2} \sim \chi^2_m \tag{A.14}$$

$$\frac{S(\vec{\beta^*})}{\sigma^2} \sim \chi^2_{n-m} \tag{A.15}$$

The ratio is independent of the actual variance of the errors and follows the Fisher–Snedecor F-distribution:

$$\frac{(\vec{\beta^*} - \vec{\beta}_0)^t.X^tX.(\vec{\beta^*} - \vec{\beta}_0)}{S(\vec{\beta^*})} \sim \frac{m}{n-m}F(m, n-m) \tag{A.16}$$

The confidence region for the variable with a α confidence level is then obtain from the value of the Fisher–Snedecor F-distribution $1-\alpha$ quantile $Q^{1-\alpha}_{F(m,n-m)}$:

$$(\vec{\beta^*} - \vec{\beta}_0)^t.X^tX.(\vec{\beta^*} - \vec{\beta}_0) \leq \frac{mS(\vec{\beta^*})}{n-m}Q^{1-\alpha}_{F(m,n-m)} \tag{A.17}$$

The ensemble of points that satisfies the equation being an ellipsoid whose axes direction and length are given by the eigenvectors and square root of the eigenvalues of X^tX.

A.5 Further considerations about numerical stability

The inversion of $X^t\Omega^{-1}X$, as well defined as it may be mathematically, is not stable numerically (see Golub and Van Loan[53]). As each number is stored with a finite precision in a computer memory, the rounding leads to inaccurate results that can strongly diverge from the exact ones. Rather than using the normal equation, the orthogonal decomposition scheme is used. A simple description of the method follows in the simplest case.

First, we assume that Ω^{-1} is symmetric and positive definite[1]. As such, it can be Cholesky factorized as $\Omega^{-1} = C^t.C$ with C a real value upper triangular matrix with real positive diagonal entries. Secondly, any $n \times m$ real matrix with $n \geq m$ such as CX can be efficiently numerically factorized in the form $CX = Q.R$ where Q is an orthogonal matrix ($Q^tQ = \mathbb{1}$) of size n and R is an upper triangular matrix of size $n \times m$:

$$\begin{pmatrix} C_1 & C_2 \end{pmatrix}.X = \begin{pmatrix} Q_1 & Q_2 \end{pmatrix}.\begin{pmatrix} R_1 \\ 0 \end{pmatrix} \tag{A.18}$$

With those notations, we have:

$$Q^t.C(\vec{Z} - Y\vec{(\beta)}) = \begin{pmatrix} Q_1^t.C_1.\vec{z} - R_1.\vec{\beta} \\ Q_2^t.C_2.\vec{z} \end{pmatrix} \tag{A.19}$$

Which leads to the expression of the score function (inserting $Q^tQ = \mathbb{1}$):

$$\begin{aligned} S(\vec{\beta}) &= (\vec{Z} - \vec{Y})^t.\Omega.(\vec{Z} - \vec{Y}) \\ &= (\vec{Z} - \vec{Y})^t.C^t(QQ^t)C.(\vec{Z} - \vec{Y}) \\ &= (Q_1C_1\vec{z} - R_1\vec{\beta})^t.(Q_1C_1\vec{z} - R_1\vec{\beta}) + (Q_2C_2\vec{z})^t.Q_2C_2\vec{z} \end{aligned} \tag{A.20}$$

1. if Ω is only non-negative definite general least square method still makes sense defining Ω^{-1} as its pseudo inverse, the minimization problem is well defined. However the numerical solution are more elaborated (see Paige[107] for instance)

Only the first term depends on β. So, the minimization with respect to $\vec{\beta}$ is equivalent to cancel this term:

$$R_1.\vec{\beta} = Q_1 C_1 \vec{z} \tag{A.21}$$

This is easily solved thanks to the triangular form of R_1. As it avoid to form and invert the matrix product $X^t\Omega^{-1}X$, the QR factorization does not lead to a dramatic augmentation of the conditioning number of the problem and numerical instability of the solution.

A.6 Fit of a Hessian matrix / a quadratic form

Curvature around the maximum can be obtained fitting a second order polynomial on the values measured along all the directions and their combinations ($\vec{u}_x$,$\vec{u}_y$,$(\vec{u}_x + \vec{u}_y)/\sqrt{2}$, etc.). In 3D this would require six successive fits for each free parameter of the Hessian matrix. The correct fit is made out in one direct adjustment on the 3D function. Thus we take into account the correlations between all the degrees of freedom at once for all the direction. To go from $P(x) = ax^2 + bx + c$ to its 3D generalization $P(\vec{X}) = (\vec{X} - \vec{X}_0)^t H(\vec{X} - \vec{X}_0) + b$ where H is a 3×3 symmetric matrix and b a scalar does not add any complexity to the fitting procedure.

$$H = \begin{pmatrix} p_{xx} & 1/2.p_{xy} & 1/2.p_{xz} \\ 1/2.p_{xy} & p_{yy} & 1/2.p_{yz} \\ 1/2.p_{xz} & 1/2.p_{yz} & p_{zz} \end{pmatrix} \tag{A.22}$$

$$\begin{aligned} \log\mathrm{L}(\vec{X}) &\simeq \vec{X}^t H \vec{X} - 2\vec{X}_0^t H \vec{X} + \vec{X}_0^t H \vec{X}_0 + b & \text{(A.23)} \\ &= p_{xx}x^2 + p_{xy}xy + \ldots + \mathrm{C^{te}} & \text{(A.24)} \end{aligned}$$

The Hessian matrix coefficients are directly extracted as the second degrees coefficients of the polynomial. The maximum position estimation $\vec{X}_0$ has to be calculated by the inversion of the Hessian and first order coefficients ($-2\vec{X}_0^t H\vec{X}$ terms) and the constant part of the polynomial is left as a nuisance parameters.

At a given point there is still a linear relationship between fitting coefficients and function value. The classic least square estimation can be found using the above procedure:

$$\begin{pmatrix} \log\mathrm{L}_1^{\mathrm{theo}} \\ \vdots \\ \log\mathrm{L}_i^{\mathrm{theo}} \\ \vdots \\ \log\mathrm{L}_n^{\mathrm{theo}} \end{pmatrix} = \begin{pmatrix} x_1y_1 & z_1y_1 & \cdots & z_1 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_iy_i & z_iy_i & \cdots & z_i & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_ny_n & z_ny_n & \cdots & z_n & 1 \end{pmatrix} . \begin{pmatrix} p_{10} \\ \vdots \\ p_1 \end{pmatrix} \tag{A.25}$$

$$\vec{\log\mathrm{L}}^{\mathrm{theo}} = L.\vec{p} \tag{A.26}$$

One wants to find $\vec{p}^*$ such that:

$$p^* = \min_p \|\vec{\log\mathrm{L}}^{\mathrm{mes}} - L.\vec{p}\|^2 \tag{A.27}$$

Using the factorization $L = Q.R$ with $Q^t.Q = I$ an orthogonal $N \times N$ matrix, and R an upper triangular $N \times 10$ matrix with same rank as L. The equation becomes as shown in the previous

section:

$$p^*.R = Q^t|_{10\times N}.\vec{\log \mathrm{L}} \tag{A.28}$$

$$\text{norm of residuals} = \|Q^t|_{N-10\times N}.\vec{\log \mathrm{L}}\|^2 \tag{A.29}$$

Associated covariance for each of the Hessian matrix coefficients can be extracted from the fit. There is no constraints to assure that H will be negative definite and this relies on the shape of the function around its maximum.

Appendix B

Statistical estimation from point sampling in a distribution

We detail here several calculations about statistical estimation of the moments of a distribution sampled several times.

B.1 Uniform distribution on the truncated unit sphere

The unit sphere $\mathcal{S}$ is the surface formed by the ensemble of the point at unit distance from the origin. In Euclidean coordinates its equation writes $x^2+y^2+z^2=1$. Spherical coordinates (radius r, polar angle ϑ and azimuthal angle φ) are the natural coordinates to parametrize the sphere. With the transformation of coordinates:

$$\begin{cases} x = r\sin\vartheta\cos\varphi \\ y = r\sin\vartheta\sin\varphi \\ z = r\cos\vartheta \end{cases} \tag{B.1}$$

A parametrization is:

$$\begin{cases} r = 1 \\ \vartheta \in [0,\pi] \\ \varphi \in [0,2\pi] \end{cases} \tag{B.2}$$

We will in the following consider the unit sphere whom a cap has been removed at the bottom and that we will call the truncated sphere $\mathcal{S}_t$:

$$\begin{cases} r = 1 \\ \vartheta \in [0,\theta] \\ \varphi \in [0,2\pi] \end{cases} \tag{B.3}$$

From the elementary surface $\mathrm{d}S = \sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi$ in spherical coordinates, we deduce the uniform measure on the truncated sphere

$$\mathrm{dP} = \frac{\mathrm{d}S}{\iint_{\mathcal{S}_t}\mathrm{d}S} = \frac{\sin\vartheta\,\mathrm{d}\vartheta\,\mathrm{d}\varphi}{2\pi(1-\cos\theta)} \quad \vartheta\in[0,\theta],\ \varphi\in[0,2\pi] \tag{B.4}$$

So that the expectation of any function f defined on $\mathcal{S}_t$ is:

$$\mathbb{E}\left(f\right) \equiv \iint_{\mathcal{S}_t} f(\vartheta, \varphi)\,\mathrm{dP} \tag{B.5}$$

And the probability density function is:

$$d_{\mathcal{S}_t}(\vartheta, \varphi) = \frac{\sin\vartheta}{2\pi R^2(1-\cos\theta)} \tag{B.6}$$

Which can be rewritten as a surface density in $3D$:

$$D_{\mathcal{S}_t}(r, \vartheta, \varphi) = \frac{\delta(r-1)\mathcal{I}_{[0,\theta]}(\vartheta).r.\sin\vartheta}{2\pi R^2(1-\cos\theta)} \tag{B.7}$$

With $\mathcal{I}_{[0,\theta]}$ the indicator function of the domain $[0,\theta]$ and $\delta(r)$ the Dirac delta distribution.

In the Cartesian coordinates base but keeping the spherical parameters, central moments of the distribution write:

$$\begin{aligned}
\vec{\mu_1^\circ} &= \iiint_{\mathbb{R}^3} \begin{pmatrix} x \\ y \\ z \end{pmatrix} D_{\mathcal{S}_t}(r, \vartheta, \varphi)\,\mathrm{d}r\,\mathrm{d}\vartheta\,\mathrm{d}\varphi \\
&= \iiint_{\mathbb{R}^3} \begin{pmatrix} r\sin\vartheta\cos\varphi \\ r\sin\vartheta\sin\varphi \\ r\cos\vartheta \end{pmatrix} D_{\mathcal{S}_t}(r, \vartheta, \varphi)\,\mathrm{d}r\,\mathrm{d}\vartheta\,\mathrm{d}\varphi \\
&= \begin{pmatrix} 0 \\ 0 \\ \cos^2\frac{\theta}{2} \end{pmatrix}
\end{aligned} \tag{B.8}$$

$$\begin{aligned}
\underline{\underline{\mu_2^\circ}} &= \iiint_{\mathbb{R}^3} \begin{pmatrix} x^2 & yx & zx \\ xy & y^2 & zy \\ xz & yz & z^2 \end{pmatrix} D_{\mathcal{S}_t}(r, \vartheta, \varphi)\,\mathrm{d}r\,\mathrm{d}\vartheta\,\mathrm{d}\varphi - \vec{\mu_1^\circ} \otimes \vec{\mu_1^\circ} \\
&= \begin{pmatrix} \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} & 0 & 0 \\ 0 & \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} & 0 \\ 0 & 0 & \frac{\sin^4(\frac{\theta}{2})}{3} \end{pmatrix}
\end{aligned} \tag{B.9}$$

As x and y are identically distributed and uncorrelated, many terms are equal in the higher central moments, also the symmetries of the distribution lead to many null moments. We identify the different terms by the power exponent of each coordinates in $x^i y^j z^k$ such that the order is $p = i + j + k$. With those notations:

$$\mu_1^\circ = \begin{cases} \cos^2\frac{\theta}{2} & k=1 \\ 0 & \text{else} \end{cases} \tag{B.10}$$

$$\mu_2^\circ = \begin{cases} \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} & i=2 \text{ or } j=2 \\ \frac{\sin^4(\frac{\theta}{2})}{3} & k=2 \\ 0 & \text{else} \end{cases} \tag{B.11}$$

$$\mu_3^\circ = \begin{cases} -\frac{1}{12}\sin^2\left(\frac{\theta}{2}\right)\sin^2\theta & i=2 \text{ or } j=2 \\ 0 & \text{else} \end{cases} \tag{B.12}$$

$$\mu_4^\circ = \begin{cases} \frac{1}{20}(19+18\cos\theta+3\cos 2\theta)\sin^4\left(\frac{\theta}{2}\right) & i=4 \text{ or } j=4 \\ \frac{1}{60}(19+18\cos\theta+3\cos 2\theta)\sin^4\left(\frac{\theta}{2}\right) & i=2 \text{ and } j=2 \\ \frac{1}{15}(3+2\cos\theta)\sin^6\left(\frac{\theta}{2}\right) & k=2 \text{ and } (i=2 \text{ or } j=2) \\ \frac{\sin^8\left(\frac{\theta}{2}\right)}{5} & k=4 \\ 0 & \text{else} \end{cases} \tag{B.13}$$

B.2 Uniform distribution on a truncated sphere with random orientations

A linear transformation formed of two rotations $R_{\vec{u}_y,\phi}$, $R_{\vec{u}_z,\alpha}$ and a dilatation of H_R transforms the truncated unit sphere in a sphere of arbitrary radius and orientation. We can derive the expression of the moments for any of those generated distribution using integration by substitution. Let's introduce the Jacobian matrix $J = \vec{\nabla}_{\vec{X}}\vec{Y}$ of the transformation in euclidean coordinates:

$$J = R.\begin{pmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{pmatrix}.\begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \tag{B.14}$$

Then the probability density function in the new coordinates is obtained from the probability density function in the old ones so that:

$$P_{\vec{Y}}(\vec{Y}) \propto P_{\vec{X}}(J^{-1}.\vec{Y}) \tag{B.15}$$

The constant is determined by using the normalization conditions and substituting $\vec{X} = J^{-1}\vec{Y}$:

$$\begin{aligned} 1 &= \int P_{\vec{Y}}(\vec{Y})\,\mathrm{d}\vec{Y} \\ &= \int \mathrm{C}^{\mathrm{te}} P_{\vec{X}}(J^{-1}.\vec{Y})\,\mathrm{d}\vec{Y} \\ &= \int \mathrm{C}^{\mathrm{te}} P_{\vec{X}}(\vec{X})|\det J|\,\mathrm{d}\vec{X} \\ \Rightarrow \mathrm{C}^{\mathrm{te}} &= |\det J^{-1}| \end{aligned} \tag{B.16}$$

Eventually we can calculate the moment of the centered distribution as:

$$\begin{aligned}\vec{\mu}_1 &= \int \vec{Y} P_{\vec{Y}}(\vec{Y})\,\mathrm{d}\vec{Y} \\ &= \int J.\vec{X}.|\det J^{-1}|.P_{\vec{X}}(\vec{X}).|\det J|\,\mathrm{d}\vec{X} \\ &= J.\vec{\mu_1^\circ} \\ &= R\begin{pmatrix} -\cos(\alpha)\cos(\theta/2)^2\sin(\phi) \\ -\cos(\theta/2)^2\sin(\alpha)\sin(\phi) \\ \cos(\theta/2)^2\cos(\phi)\end{pmatrix}\end{aligned} \tag{B.17}$$

As well as:

$$\begin{aligned}\underline{\underline{\mu_2}} &= \int \vec{Y}.\vec{Y}^t P_{\vec{Y}}(\vec{Y})\,\mathrm{d}\vec{Y} \\ &= J.\underline{\underline{\mu_2^\circ}}.J^t \\ &= R^2 R_{\vec{u}_z,\alpha}\begin{pmatrix} \mu_{2x}^\circ\cos^2\phi + \mu_{2z}^\circ\sin^2\phi & 0 & (\mu_{2x}^\circ - \mu_{2z}^\circ)\cos\phi\sin\phi \\ 0 & \mu_{2x}^\circ & 0 \\ (\mu_{2x}^\circ - \mu_{2z}^\circ)\cos\phi\sin\phi & 0 & \mu_{2z}^\circ\cos^2\phi + \mu_{2x}^\circ\sin^2\phi\end{pmatrix} R^t_{\vec{u}_z,\alpha} \\ \text{with: } \mu_{2x}^\circ &= \frac{(2+\cos\theta)\sin^2(\frac{\theta}{2})}{3} \quad \text{and} \quad \mu_{2z}^\circ = \frac{\sin^4(\frac{\theta}{2})}{3}\end{aligned} \tag{B.18}$$

We note that those rules to calculate the moments for any possible orientations and radius of the truncated sphere from the moment of the original vertical unit sphere are akin the rules of tensors transformation.

B.3 The h-statistic

Given a set of N positions $\{\vec{X}_i\}$ independently and identically sampled in a 3D distribution D:

$$\{\vec{X}_i\} = \left\{\begin{pmatrix} x_i \\ y_i \\ z_i\end{pmatrix}\right\} \tag{B.19}$$

we examine the estimation of the underlying distribution properties. We focus on the best linear unbiased estimators described by Halmos[61]. The obvious sample moments of the distribution are estimates of the distribution moment around the origin:

$$\mathrm{em}_k = \frac{1}{N}\sum_i \vec{X}_i \otimes \ldots \otimes \vec{X}_i = \sum_i \vec{X}_i^{\otimes k} \tag{B.20}$$

where $\otimes$ is the natural tensor product so that:

$$\vec{U} \otimes \vec{V} = \vec{U}.\vec{V}^t = \begin{pmatrix} x_U x_V & x_U y_V & x_U z_V \\ y_U x_V & y_U y_V & y_U z_V \\ z_U x_V & z_U y_V & z_U z_V \end{pmatrix} \tag{B.21}$$

$$\vec{U} \otimes \vec{V} \otimes \vec{W} = \begin{pmatrix} z_W.\vec{U} \otimes \vec{V} \\ y_W.\vec{U} \otimes \vec{V} \\ x_W.\vec{U} \otimes \vec{V} \end{pmatrix} \tag{B.22}$$

Where equation (B.22) represents a cube of 3^3 coefficients. Although those moments are quite natural, the classical moments of the distribution have few interesting features for image analysis. Besides other drawbacks, they are not invariant to a simple translation. We are therefore more interested in an estimator of the central moments of the distribution.

The intuitive estimators for the central moment, derived from the estimation of the classical moments, are biased:

$$\mathrm{e}\mu_k \neq \frac{1}{N} \sum_i (\vec{X}_i - \mathrm{e}\vec{\mathrm{m}}_1)^{\otimes k} \tag{B.23}$$

The proper unbiased estimators of the central moments are known as "h-statistic" and are build from a linear combinations of the previous intuitive but biased estimators:

$$\mathrm{e}\mu_2 = \frac{1}{N-1} \sum_i (\vec{X}_i - \mathrm{e}\vec{\mathrm{m}}_1)^{\otimes 2} \tag{B.24}$$

$$\mathrm{e}\mu_3 = \frac{N}{(N-1)(N-2)} \sum_i (\vec{X}_i - \mathrm{e}\vec{\mathrm{m}}_1)^{\otimes 3} \tag{B.25}$$

$$\begin{aligned} \mathrm{e}\mu_4 = & \frac{N(N^2-2N+3)}{(N-1)(N-2)(N-3)} \sum_i (\vec{X}_i - \mathrm{e}\vec{\mathrm{m}}_1)^{\otimes 4} \\ & - \frac{3(2N-3)N}{(N-1)(N-2)(N-3)} \left(\sum_i (\vec{X}_i - \mathrm{e}\vec{\mathrm{m}}_1)^{\otimes 2} \right)^2 \end{aligned} \tag{B.26}$$

Since sampling are independently drawn with the same probability density function D, the probability density for the outcome of any results of the N successive trials is:

$$D_N(x_1 \dots x_N) \equiv \prod_{i=1}^{N} D(x_i) \tag{B.27}$$

Which enables us to determine the statistical behavior of the empirical estimators. We will note in the following:

$$\langle f(\vec{X}) \rangle = \int f(\vec{X}) \prod_{i=1}^{N} D(x_i) \tag{B.28}$$

Hence the classical results:

$$\langle \vec{X}_i \rangle = m_1 \tag{B.29}$$

$$\langle \vec{X}_i \otimes \vec{X}_j \rangle = \begin{cases} \langle \vec{X}_i \rangle \otimes \langle \vec{X}_i \rangle & i \neq j \\ m_2 & i = j \end{cases} \tag{B.30}$$

All the results come from those properties and linearity and different partition of the sum. The values of the terms we sums only depend on the equality or inequality of their indices (permutation included), hence:

$$\begin{aligned}\sum_{i,j} f_{i,j} &= \sum_{i=j} f_{=} + \sum_{i\neq j} f_{\neq} \\ &= N f_{=} + N(N-1) f_{\neq}\end{aligned} \tag{B.31}$$

$$\begin{aligned}\sum_{i,j,k} g_{i,j,k} &= \sum_{i=j=k} g_{=,=} + \binom{3}{2} \sum_{i=j\neq k} g_{=,\neq} + \sum_{i\neq j\neq k} g_{\neq,\neq} \\ &= N g_{=,=} + 3N(N-1) g_{=,\neq} + N(N-1)(N-2) g_{\neq,\neq}\end{aligned} \tag{B.32}$$

For instance, the expectation of the empirical mean is:

$$\begin{aligned}\langle \frac{1}{N} \sum_i \vec{X}_i \rangle &= \frac{1}{N} \sum_i \langle \vec{X}_i \rangle \\ &= \frac{1}{N} N m_1 \\ &= m_1\end{aligned} \tag{B.33}$$

and its variance matrix is given by:

$$\begin{aligned}\langle (\frac{1}{N} \sum_i \vec{X}_i)^{\otimes 2} - m_1^{\otimes 2} \rangle &= \frac{1}{N^2} (\sum_{i,j} \langle \vec{X}_i \otimes \vec{X}_j \rangle) - m_1^{\otimes 2} \\ &= \frac{1}{N^2} (N m_2 + N(N-1) m_1^{\otimes 2}) - m_1^{\otimes 2} \\ &= \frac{1}{N} (m_2 - m_1^{\otimes 2}) \\ &= \frac{\mu_2}{N}\end{aligned} \tag{B.34}$$

the expectation of the empirical covariance of the sample is

$$\begin{aligned}\langle \frac{1}{N-1} \sum_i (\vec{X}_i - \frac{1}{N} \sum_j \vec{X}_j)^{\otimes 2} \rangle &= \frac{1}{N-1} \langle \sum_i \vec{X_i}^{\otimes 2} - \frac{2}{N} \sum_{i,j} \vec{X}_j \otimes \vec{X}_i + \frac{1}{N^2} \sum_{i,k,l} \vec{X}_k \otimes \vec{X}_l \rangle \\ &= \frac{1}{N-1} \left(\sum_i \langle \vec{X_i}^{\otimes 2} \rangle - \frac{1}{N} \sum_{i,j} \langle \vec{X}_j \otimes \vec{X}_i \rangle \right) \\ &= \frac{1}{N-1} \left(N m_2 - \frac{1}{N} (N m_2 + N(N-1) m_1^{\otimes 2}) \right) \\ &= m_2 - m_1^{\otimes 2} \\ &= \mu_2\end{aligned} \tag{B.35}$$

The calculation of the variance matrix (fluctuations from one sampling to another) of the empirical covariance is a cumbersome calculation that follows the same principle. We find:

$$\begin{aligned}\mathrm{var}(\mathrm{e}\mu_2(i,j), \mathrm{e}\mu_2(k,l)) = &\frac{1}{N} \mu_4(i,j,k,l) - \frac{1}{N} \mu_2(i,j) \mu_2(k,l) \\ &+ \frac{1}{N(N-1)} (\mu_2(i,k) \mu_2(j,l) + \mu_2(i,l) \mu_2(j,k))\end{aligned} \tag{B.36}$$

Which reduces in one dimension to:

$$\mathrm{var}(\mathrm{e}\mu_2) = \frac{1}{N}(\mu_4 - \frac{N-3}{N-1}\mu_2^2) \tag{B.37}$$

Using the same method we can calculate the covariance matrix of the empirical mean and the empirical covariance:

$$\mathrm{var}(\mathrm{e}\mu_2(i,j), \mathrm{e}\mu_1(k)) = \frac{1}{N}\mu_3(i,j,k) \tag{B.38}$$

Which complete the calculation of the full 9×9 covariance matrix between each pair of elements in $\vec{\mu_1}$ and $\underline{\underline{\mu_2}}$.

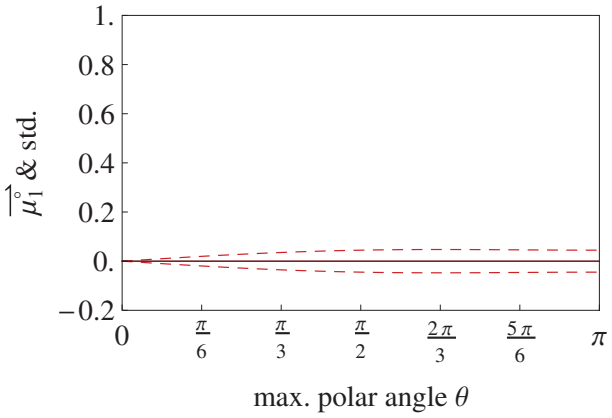
From the h-statistic equations (B.35) and (B.36) together with the reference distribution moments equations (B.10) to (B.13) associated to the rules of tensor transformations, we can calculate the mean and std. of point distributions sampled from the truncated sphere in any orientation. Figure B.1 shows the evolutions with growing sphere completion θ. As we only have access to the x and y coordinates of the sampled positions, only this part of the information is displayed. As x and y are equivalent in the absence of tilt, we give twice the same information (figures B.1a and B.1c). A tilt ($\phi \neq 0$) shows that the degeneracy can be removed observing from a different point of view in all the situations but full completion (figures B.1b and B.1d).

We semi-automated the calculation on Mathematica® by designing a set of rules to handle the sums and their partitions. The followings rules enable the software to treat symbolical sum expressions over various indices. All the sums in Mathematica® are of the form:

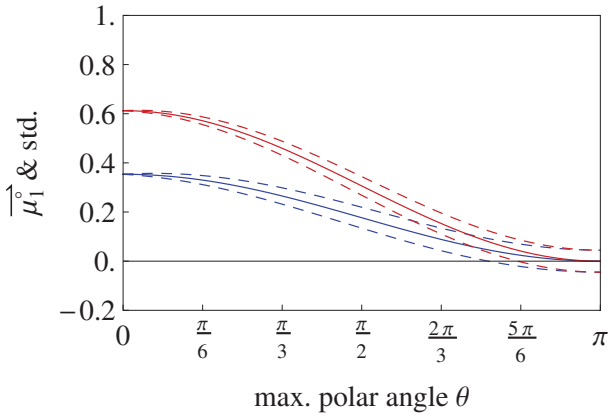
```
Sum[term[i,j] ,{i,imax},{j,jmax}]
```

With the explicit assumption that any sum start at 1. Symbolic manipulations such as expansion of the terms inside the symbolic sum, summation over dummy variables are not part of the default behavior but can be added as rules to partially automatize the calculation.

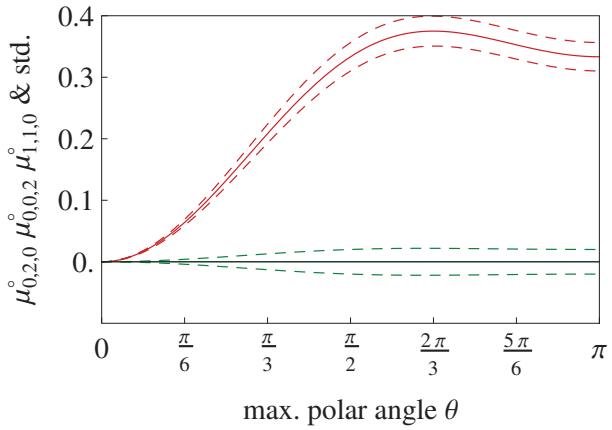
```
ExpandInSum=HoldPattern[Sum[a_Times ,y___List]]:>Sum[Expand[a],y];
OutnInSum=
  Sum[Times[a__,o:(Power[n,p_]|n|_Integer)],l__List]:>o Sum[Times[a],l];
SplitInSum=HoldPattern[Sum[a_+c_,y___]]:>Sum[a,y]+Sum[c,y];
DummiInSum=HoldPattern[Sum[x_,y___,{k_,m_},z___]/;FreeQ[x,k]]:>Sum[m x,y,z];
SumProduct=
  {HoldPattern[Times[Sum[a_ ,y___List],Sum[b_,z___List]]]:> Sum[a b,y,z],
   HoldPattern[Sum[a_ Sum[b_,z___List] ,y___List]]:> Sum[a b,y,z]};
```



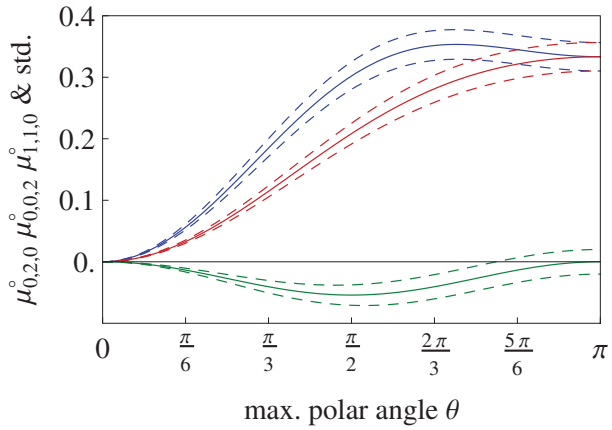
(a) $\mathcal{S}_t$ reference truncated unit sphere mass center $\vec{\mu}_1^{\circ}\Big|_{x,y}$: $\mu_{1,0,0} = \mu_{0,1,0}$ (red)



(b) Truncated unit sphere tilted by $\phi = \frac{\pi}{4}$ and rotated by $\alpha = \frac{\pi}{3}$ mass center $\vec{\mu_1}|_{x,y}$: $\mu_{1,0,0}$(blue), $\mu_{0,1,0}$ (red)



(c) $\mathcal{S}_t$ reference truncated unit sphere second order moment $\underline{\underline{\mu_2^{\circ}}}\Big|_{x,y}$ in the projection plan (x, y): $\mu_{2,0,0} = \mu_{0,2,0}$ (red), and $\mu_{1,1,0}$ (green)



(d) Truncated unit sphere tilted by $\phi = \frac{\pi}{4}$ and rotated by $\alpha = \frac{\pi}{3}$ mass center $\underline{\underline{\mu_2}}\Big|_{x,y}$ in the projection plan (x, y): $\mu_{2,0,0}$(blue), $\mu_{0,2,0}$ (red), and $\mu_{1,1,0}$ (green)

Figure B.1 – Moments of a $N = 1500$ positions sampling of truncated spheres calculated according to h-statistics, tensor transformation and reference truncated unit sphere moments depending on the completion angle θ. Expected values (plain lines) and ± 3 standard deviation (dashed lines –99% confidence bounds) are plotted for each coefficient to show their evolutions and dispersions.

Corresponding to the following operations:

$$\sum_i (a_i + b_i)^2 \overset{/.\text{ExpandInSum}}{\longrightarrow} \sum_i a_i^2 + b_i^2 + 2a_i b_i$$

$$\sum_i p^q . a_i \overset{/.\text{OutnInSum}}{\longrightarrow} p^q \sum_i a_i$$

$$\sum_i a_i + b_i \overset{/.\text{SplitInSum}}{\longrightarrow} \sum_i a_i + \sum_i b_i$$

$$\sum_{i,j}^{N} a_i \overset{/.\text{DummiInSum}}{\longrightarrow} N \sum_i^N a_i$$

$$\sum_i a_i \sum_j b_j \overset{/.\text{SumProduct}}{\longrightarrow} \sum_{i,j} a_i b_j$$

$$\sum_i \left(a_i \sum_j b_j \right) \overset{/.\text{SumProduct}}{\longrightarrow} \sum_{i,j} a_i b_j$$

The partitions of the indices of the sum depending on the equality of subset of the sum indices can also be treated by rules thanks to the `SetPartitions` function of the module `Needs["Combinatorica‘"]`. The sum $\sum_{i,j,k} a_i b_j c_k$ is coded as the set $\{a, b, c\}$ where each element is implicitly summed over an independent variable. We want all the possible subsets:

```
SetPartitions[a,b,c]
          ={{{a,b,c}},{{a},{b,c}},{{a,b},{c}},{{a,c},{b}},{{a},{b},{c}}}
```

Then we multiply the element of the inner lists and apply $\langle . \rangle \equiv$ esp to the product. The multiplicity of each new term is $N(N-1)\ldots = N!/(N-L)!$, L being the number of elements of the second level list obtain by the function `Length`. All terms and their multiplicity are eventually added:

```
moment[list_] := esp[Apply[Times, list]]
moments[list_] := Map[moment, list, 2]
multiplicity[L_] := N!/(N - L)!
momentsum[indicesSet_]:=
Apply[Plus,Map[multiplicity,Map[Length,SetPartitions[indicesSet]]]
*Apply[Times,moments[SetPartitions[indicesSet]],1],0]

momentsum[a,b,c]
        =(-2+N)(-1+N)N esp[a]esp[b]esp[c]
          +(-1+N)N esp[a b]esp[c]
          +(-1+N)N esp[b]esp[a c]
          +(-1+N)N esp[a]esp[b c]
          +N esp[a b c]
```

Which is equivalent to equation (B.32) in the case $a \equiv b \equiv c$.

Appendix C

Hessian matrix in a change of coordinates system

We want here to determine the relationship between the Hessian matrix around an extremum point in a new coordinate system and the Hessian of the function in the prim coordinate system.

Given a function $\mathcal{R}^3 \mapsto \mathcal{R}$, we define a change of coordinates as the composition with a continuous and twice differentiable function $\mathcal{R}^3 \mapsto \mathcal{R}^3$ (a diffeomorphisme) h:

$$F : \vec{X_1} \mapsto F[\vec{X_1}] \tag{C.1}$$
$$h : \vec{X_2} \mapsto \vec{X_1} \tag{C.2}$$
$$G : \vec{X_2} \mapsto F \circ h[\vec{X_2}] \tag{C.3}$$

There is no change of the bases vectors that remains constant (Cartesian system).

In all the following the operation (mainly derivation) regarding the coordinates system will be labeled for clarification: ∇_2 is thus the derivation operator on each of the coordinates of the second system $\nabla_{\vec{X_2}}$...

The Hessian matrix of a real-valued function is defined as

$$\begin{aligned} H_F[\vec{X}] &= \nabla_{\vec{X}}(\nabla_{\vec{X}}F)^t \\ &= \begin{pmatrix} F^{(2,0,0)} & F^{(1,1,0)} & F^{(1,0,1)} \\ F^{(1,1,0)} & F^{(0,2,0)} & F^{(0,1,1)} \\ F^{(1,0,1)} & F^{(0,1,1)} & F^{(0,0,2)} \end{pmatrix} [x, y, z] \end{aligned} \tag{C.4}$$

where:

$$F^{(i,j,k)} = \frac{\partial^{i+j+k} F}{\partial x^i \partial y^j \partial z^k} \tag{C.5}$$

Hessian matrix is real-valued and Schwartz theorem ensures that for all $C^2(\mathcal{R})$ the Hessian matrix is symmetric. Consequently, the matrix is diagonalizable with real eigenvalues (Spectral Theorem). The eigenvalues of the Hessian matrix determines the principal curvatures around the position, and its eigenvectors gives their directions. At points where the function F reaches an extremum (i.d. the gradient of the function is null),the principal curvatures informs us about the type of extremum:

— if all eigenvalues are positive (resp. negatives), this is a local minimum (resp. maximum) of the function

— if one of the value is equal to zero, the extremum is called degenerated.

— if the eigenvalues have opposite signs, this is a saddle point.

We use the rule of compound functions derivation and derivation of a product to calculate the Hessian matrix of the new function G in the new coordinates:

$$\begin{aligned} H_G &= \nabla_2(\nabla_2(F \circ h))^t \\ &= \nabla_2(\nabla_2 h.\nabla_1(F) \circ h)^t \\ &= \nabla_2(J_h^t.\nabla_1(F) \circ h)^t \\ &= \nabla_2((\nabla_1(F) \circ h)^t.J_h) \\ &= J_h^t.\nabla_1(\nabla_1(F))^t \circ h.J_h + (\nabla_1(F) \circ h)^t.\nabla_2(J_h) \\ &= J_h^t.(H_F \circ h).J_h + (\nabla_1(F) \circ h)^t.\nabla_2(J_h) \end{aligned} \tag{C.6}$$

Where J_h stands for the Jacobian matrix of the coordinate transformation:

$$J_h = \left(\frac{\partial\, h_i}{\partial x_j}\right)_{ij} \qquad J_h^t = \nabla_2(h) \tag{C.7}$$

Let's note that we also introduced a third order tensor: $\nabla_2(J_h)$. The signification of the calculation results is that the new curvature decomposes into the former curvature transported at the image point and modified according to the variations of the coordinates transformation $J_h^t.(H_F \circ h).J_h$, and a correction from the curvature of the coordinates transformation appropriately transformed by the variation of the function $(\nabla_1(F) \circ h)^t.\nabla_2(J_h)$.

C.1 (Inverse) Hessian matrix at extrema

A continuous change of coordinates affects the space so that the neighboring points of the image of a given point are image of its neighboring points before the change. When one point is a local maximum (resp. minimum) in a given system of coordinate –that is to say value of the function is greater (resp. smaller) that all neighboring values–, this properties is conserved in such a coordinate change even if the shape of the space is altered (thus the higher derivatives values). Therefore the extremum points are localized at the image of the former extremum points and gradient of the compound function remains zero $\nabla_1(F) \circ h|_{\text{extrema}} = 0$. We now look for the modification of the curvature around the image extremum.

The second term however is zero at the extrema points since as we have seen before $\nabla_1(F) \circ h|_{\text{extrema}} = 0$. And eventually the new Hessian matrix is:

$$H_G^* = J_h^t.H_F^*.J_h \tag{C.8}$$

thanks to which it is obvious that the new form does have the symmetry properties of a Hessian matrix. Furthermore, as the change of coordinate is continuous, a former local maximum remains a local maximum in the new system of coordinate.

And we can turn back to the Hessian matrix of F in its original coordinates. This latter can be expressed using the Jacobian matrix of the coordinates change at the extremum or using the

inverse function Jacobian matrix at the transformed extremum thanks to the inverse function derivation theorem ($J_{h^{-1}}(\vec{X_1}) = J_h^{-1}(\vec{X_2})$):

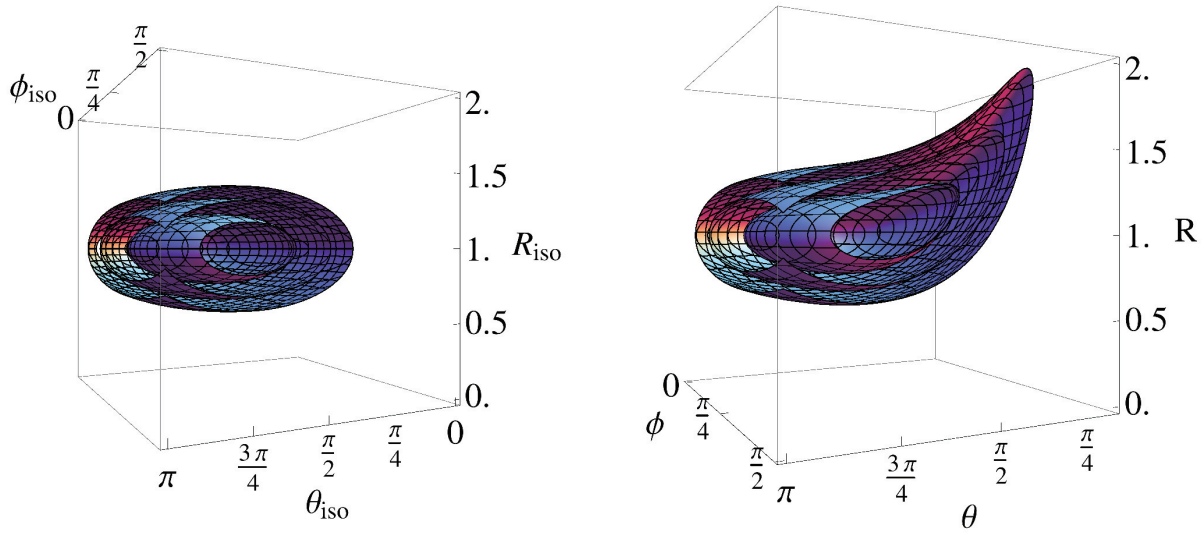
$$H_F^* = J_{h^{-1}}^t . H_G^* . J_{h^{-1}} = J_h^{-t} . H_G^* . J_h^{-1} \tag{C.9}$$

Sometimes we rather want the inverse of the Hessian matrix than the Hessian matrix itself –for instance to deduce a confidence interval. It makes mathematically no difference to inverse the matrix we know and proceed with the previous formula, but in the computer world, we only store rounded numbers in our finite size memory (see "float numbers") and inversion is an unstable operation, that we might be happy to skip. The formula at the extrema points is obvious to invert as each elements can be inverted:

$$(H_G^*)^{-1} = (J_h)^{-1} . (H_F^*)^{-1} . J_h^{-t} \tag{C.10}$$

And the invert of the Jacobian matrix is evaluated analytically to ensure stability.

C.2 Application to the truncated sphere iso apparent surface transformation



(a) Initial quadratic approximation around the optimum in the modified coordinate system

(b) Deformed shape around the transformed optimum in the initial coordinate system

Figure C.1 – Effect of the change of coordinate h on the likelihood function

In our work only one of the coordinates is affected by a change: $R \rightarrow R_{\text{iso}\mathcal{A}}(\theta, \phi)$. The

precious calculation can be explicitly written using this form:

$$\log L(\theta, \phi, \mathrm{R}) \quad \Leftrightarrow \quad (\log L \circ h^{-1})(\theta_{\mathrm{iso}\mathcal{A}}, \phi_{\mathrm{iso}\mathcal{A}}, R_{\mathrm{iso}\mathcal{A}}) \tag{C.11}$$

$$h^{-1}\begin{pmatrix}\theta_{\mathrm{iso}\mathcal{A}}\\ \phi_{\mathrm{iso}\mathcal{A}}\\ R_{\mathrm{iso}\mathcal{A}}\end{pmatrix} \equiv \begin{pmatrix}\theta_{\mathrm{iso}\mathcal{A}}\\ \phi_{\mathrm{iso}\mathcal{A}}\\ \frac{R_{\mathrm{iso}\mathcal{A}}}{g[\theta_{\mathrm{iso}\mathcal{A}}, \phi_{\mathrm{iso}\mathcal{A}}]}\end{pmatrix} \qquad h\begin{pmatrix}\theta\\ \phi\\ \mathrm{R}\end{pmatrix} \equiv \begin{pmatrix}\theta\\ \phi\\ Rg[\theta_{\mathrm{iso}\mathcal{A}}, \phi_{\mathrm{iso}\mathcal{A}}]\end{pmatrix} \tag{C.12}$$

$$g[\theta_{\mathrm{iso}\mathcal{A}}, \phi_{\mathrm{iso}\mathcal{A}}] = \sqrt{\frac{\mathcal{A}_{\mathrm{apparent}}(\theta_{\mathrm{iso}\mathcal{A}}, \phi_{\mathrm{iso}\mathcal{A}}, 1)}{\pi}} \tag{C.13}$$

$$J_h = \begin{pmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ R.g^{(1,0)}[\theta, \phi] & R.g^{(0,1)}[\theta, \phi] & g[\theta, \phi]\end{pmatrix} \tag{C.14}$$

Appendix D

Sampling resolution on a sphere

We consider a uniform density covering a truncated spherical shell. We sample this density of presence by several consecutive identical measurement, i.e. dots are laid on the shell surface. How to identify the smallest truncated area that one is able to discriminate with a given number of dots on such a sphere ? Looking at the measure, any area empty of dots can be assumed to be a void patch, but one has to verify the statistical significance of the observation. Indeed sampling a perfectly covered sphere would also lead to empty areas due to the finite number of sampling points. A simple, yet robust, test for the statistical significance consists in verifying that the size of the presumed empty patch is greater that the one that can be expected by the purely statistical fluctuations of the sampling. The size of the truncated section is bounded by the smallest distance along the sphere surface between the points adjacent to this area. We thus want to monitor the distribution of orthodromic distances[1] between each point and its closest neighbor. Such distance d is the length of the arc linking the two points on the circle centered on the center of the sphere. It is linked to the opening angle between the point by:

$$d = R\theta \tag{D.1}$$

As for the opening angle between two points A and B on the unit sphere of center O, the relation is:

$$\theta_{AB} = \tan^{-1}\left(\frac{\overrightarrow{OA} \times \overrightarrow{OB}}{\overrightarrow{OA}.\overrightarrow{OB}}\right) \tag{D.2}$$

For a uniform sampling on a sphere, one expects the points to be homogeneously distributed over the area. This means that, on average, the solid angle of 4π sr for the total sphere, is divided between the N points $\hat{\omega} = 4\pi/N$ sr. Note that a solid angle $\hat{\omega}$ is related to the opening angle $\bar{\theta}$ by the relation:

$$\hat{\omega} = 2\pi(1 - \cos\bar{\theta}) \tag{D.3}$$

Then, on a complete sphere, the area (solid angle) between neighboring points actually fluctuates around this value, which leads for N points to the expected mean apparent opening angle:

$$\bar{\theta}(N) = \cos^{-1}(1 - \frac{2}{N}) \underset{N\to\infty}{\sim} N^{-1/2} \tag{D.4}$$

This mean opening angle gives the mean to voids length between sampled points. We ran a series of simulations of sampling N points uniformly on a sphere to check this expression and

1. also called "great circle distance" this is smallest distance between two points on the sphere surface

evaluate the fluctuations around the mean value. The results are plotted in figure D.1-inset, in terms of opening angle distribution.

We calculated the exact form of the closest neighbor distribution on a sphere following the spirit of a demonstration reported by Chandrasekhar[28] due to Hertz for the volume case. Without loss of generality, we choose $R = 1$ (equivalent to divide all distances by R), so that the distance $d = R.\theta$ on the sphere simply writes $d = \theta$. We write the probability to find the first neighbor between the distance θ and $\theta + \mathrm{d}\theta$ as the simultaneous realization of the two following independent events:

— we did not find any neighbors in $[0, \theta]$

— we found a point in $[\theta, \theta + \mathrm{d}\theta]$

$$
\begin{aligned}
D(\theta,1)\,\mathrm{d}\theta &= \left(1 - \int_0^{\theta} D(\theta_1,1)\,\mathrm{d}\theta_1\right) \int_{\theta}^{\theta+\mathrm{d}\theta} \rho\ \mathrm{d}S \\
D(\theta,1)\,\mathrm{d}\theta &= \left(1 - \int_0^{\theta} D(\theta_1,1)\,\mathrm{d}\theta_1\right) 2\pi\rho\sin\theta\,\mathrm{d}\theta \\
\int_0^{\theta} D(\theta_1,1)\,\mathrm{d}\theta_1 &= 1 - \frac{D(\theta_1,1)}{2\pi\rho\sin\theta} \\
D(\theta,1) &= -\frac{\dot{D}(\theta,1)2\pi\rho\sin\theta - D(\theta,1)2\pi\rho\cos\theta}{4\pi^2\rho^2\sin^2\theta} \\
D(\theta,1) &= C^{te}\sin\theta\, e^{-2\pi\rho\cos\theta} \\
D(\theta,1) &= \frac{\pi\rho\sin\theta}{\sinh(2\pi\rho)} e^{-2\pi\rho\cos\theta}
\end{aligned}
\tag{D.5}
$$

And so:

$$
D(d,1) = \frac{\pi R\rho\sin(\frac{d}{R})}{\sinh(2\pi R^2\rho)} e^{-2\pi R^2\rho\cos(\frac{d}{R})} \tag{D.6}
$$

When the sampling density on the sphere is strong enough, so distance to first neighbor is low ($d \ll 1$), this distribution is extremely well fitted by a Rayleigh distribution which is the distribution of nearest neighbor distances in the plane case[2] with mean $\bar{\theta}$ following the above relation (equation (D.4)). The Rayleigh distribution has only one degree of freedom:

$$
\begin{aligned}
f_\lambda(\theta) &= \frac{\theta}{\lambda^2} e^{-\theta^2/2\lambda^2} && \text{density function} \\
\theta &\in [0,\infty[&& \text{support} \\
\bar{\theta} &= \lambda\sqrt{\frac{\pi}{2}} && \text{mean} \\
\mathrm{Var}\theta &= \frac{4-\pi}{2}\lambda^2 && \text{variance} \\
\theta_{\mathrm{rand}}[\epsilon](N) &= \bar{\theta}(N)\sqrt{-\pi\ln\epsilon} && \text{upper bound of confidence interval with error level } \epsilon \in [0,1]
\end{aligned}
\tag{D.7}
$$

The upper value of the confidence interval, $\theta_{\mathrm{rand}}[\epsilon](N)$, is the minimum measurable opening angle on sampled sphere that makes sense regarding fluctuations of the sampling given an arbitrary acceptable probability of error ϵ. An empty patch with opening angle outside of

2. the supports of those two distributions are never the less somehow different ($[0,\infty[$ v.s. $[0,\pi]$)! However we notice that the density of probability is vanishingly small on $[\pi,\infty[$ even for very low sampling densities.

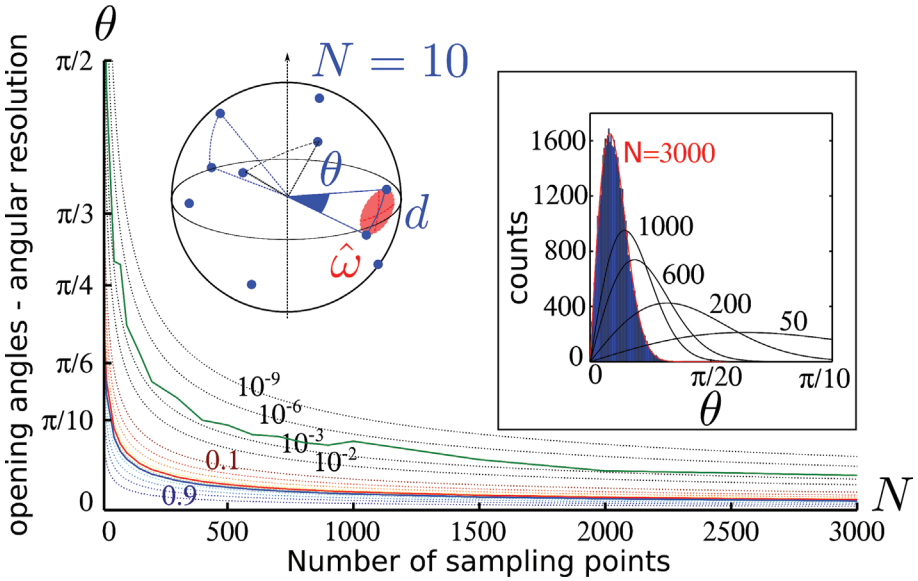


Figure D.1 – Limiting resolution for completion estimation due to finite sampling effect. The dashed lines correspond to approximated confidence upper bound for various acceptable error $\epsilon \in \{0.9, 0.8, \ldots, 0.1,\ 10^{-2},\ 10^{-3}, \ldots\}$. The expected distance between points $\bar{\theta}$ equation (D.4) (solid blue) and its simple approximation $\bar{\theta} \sim N^{-\frac{1}{2}}$ (solid red). The average maximal distance between two points observed on $5\,10^3$ simulation for each sampling N (plain green). Inset: Several calculated distribution of nearest neighbor distance probability expressed as angle at the sphere center for various number of sampled points N and approximated by the Rayleigh distribution. Blue histogram shows a simulated distribution.

the confidence interval $[0, \theta_{\text{rand}}]$ is likely to indicate a void in the density of presence (with probability $\mathrm{P} = 1 - \epsilon$). This theoretical limit is nonetheless far from being reached in the case described in the main text where a Gaussian convolution adds to the fluctuations and a planar projection withdraw information.

Bibliography

[1] E. Abbe, "Beitrüge zur theorie des mikroskops und der mikroskopischen wahrnehmung," *Archiv für mikroskopische Anatomie*, vol. 9, no. 1, pp. 413–418, 1873. [Online]. Available: http://dx.doi.org/10.1007/BF02956173

[2] Y. Aït-Sahalia, "Maximum likelihood estimation of discretly sampled diffusion: a closed-form approximation approach," *Econometrica*, vol. 70, no. 1, pp. 223–262, 2002.

[3] A. C. Aitken, "On least squares and linear combination of observations," *Proceedings of the Royal Society of Edinburgh*, vol. 55, pp. 42–48, 1934.

[4] S. Andrade, F. Pinho, A. M. Ribeiro, M. Carreira, C. F. F., P. Roy, and M. P. Monteiro, "Immunization against active ghrelin using virus-like particles for obesity treatment," *Current Pharmaceutical Design*, vol. 19, no. 36, pp. 6551–6558, 2013.

[5] D. G. Angelescu and P. Linse, "Viruses as supramolecular self-assemblies: modelling of capsid formation and genome packaging," *Soft Matter*, vol. 4, no. 10, p. 1981, 2008.

[6] C. E. Ashley, E. C. Carnes, G. K. Phillips, P. N. Durfee, M. D. Buley, C. A. Lino, D. P. Padilla, B. Phillips, M. B. Carter, C. L. Willman, C. J. Brinker, J. do Carmo Caldeira, B. Chackerian, W. Wharton, and D. S. Peabody, "Cell-specific delivery of diverse cargos by bacteriophage ms2 virus-like particles," *ACS Nano*, vol. 5, no. 7, pp. 5729–5745, 2011, pMID: 21615170. [Online]. Available: http://dx.doi.org/10.1021/nn201397z

[7] M. Badieirostami, M. D. Lew, M. a. Thompson, and W. E. Moerner, "Three-dimensional localization precision of the double-helix point spread function versus astigmatism and biplane." *Applied physics letters*, vol. 97, no. 16, p. 161103, 2010.

[8] J. B. Bancroft, "The Self-Assembly Of Spherical Plant Viruses," in *Advances in Virus Research*, ser. Advances in Virus Research, M. A. L. Kenneth M. Smith and F. B. Bang, Eds. Academic Press, 1970, vol. 16, pp. 99–134.

[9] J. B. Bancroft, G. J. Hills, and R. Markham, "A study of the self-assembly process in a small spherical virus. Formation of organized structures from protein subunits in vitro." *Virology*, vol. 31, no. 2, pp. 354–379, 1967.

[10] M. Bates, B. Huang, G. T. Dempsey, and X. Zhuang, "Multicolor super-resolution imaging with photo-switchable fluorescent probes." *Science (New York, N.Y.)*, vol. 317, no. 5845, pp. 1749–1753, 2007.

[11] M. Bates, B. Huang, and X. Zhuang, "Super-resolution microscopy by nanoscale localization of photo-switchable fluorescent probes." *Current opinion in chemical biology*, vol. 12, no. 5, pp. 505–14, Oct. 2008.

[12] E. Betzig, G. H. Patterson, R. Sougrat, O. W. Lindwasser, S. Olenych, J. S. Bonifacino, M. W. Davidson, J. Lippincott-Schwartz, and H. F. Hess, "Imaging intracellular fluorescent proteins at nanometer resolution." *Science (New York, N.Y.)*, vol. 313, no. 5793, pp. 1642–5, Sep. 2006.

[13] M. Bladt and M. Sorensen, "Statistical inference for discretely observed Markov jump processes," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 3, pp. 395–410, Jun. 2005.

[14] M. Bossi, J. Fölling, V. N. Belov, V. P. Boyarskiy, R. Medda, A. Egner, C. Eggeling, A. Schönle, and S. W. Hell, "Multicolor far-field fluorescence nanoscopy through isolated detection of distinct molecular species," *Nano Letters*, vol. 8, no. 8, pp. 2463–2468, 2008, pMID: 18642961. [Online]. Available: http://dx.doi.org/10.1021/nl801471d

[15] J. a. G. Briggs and H.-G. Kräusslich, "The molecular architecture of HIV." *Journal of molecular biology*, vol. 410, no. 4, pp. 491–500, Jul. 2011.

[16] J. a. G. Briggs, T. Wilk, R. Welker, H.-G. Kräusslich, and S. D. Fuller, "Structural organization of authentic, mature HIV-1 virions and cores." *The EMBO Journal*, vol. 22, no. 7, pp. 1707–15, Apr. 2003.

[17] J. a. G. Briggs, M. N. Simon, I. Gross, H.-G. Kräusslich, S. D. Fuller, V. M. Vogt, and M. C. Johnson, "The stoichiometry of Gag protein in HIV-1." *Nature structural & molecular biology*, vol. 11, no. 7, pp. 672–5, Jul. 2004.

[18] J. a. G. Briggs, M. C. Johnson, M. N. Simon, S. D. Fuller, and V. M. Vogt, "Cryo-electron Microscopy Reveals Conserved and Divergent Features of Gag Packing in Immature Particles of Rous Sarcoma Virus and Human Immunodeficiency Virus," *Journal of Molecular Biology*, vol. 355, no. 1, pp. 157–168, 2006.

[19] M. A. Bruckman, J. Liu, G. Koley, Y. Li, B. Benicewicz, Z. Niu, and Q. Wang, "Tobacco mosaic virus based thin film sensor for detection of volatile organic compounds," *J. Mater. Chem.*, vol. 20, pp. 5715–5719, 2010. [Online]. Available: http://dx.doi.org/10.1039/C0JM00634C

[20] L.-A. Carlson, J. a. G. Briggs, B. Glass, J. D. Riches, M. N. Simon, M. C. Johnson, B. Müller, K. Grünewald, and H.-G. Kräusslich, "Three-dimensional analysis of budding sites and released virus suggests a revised model for HIV-1 morphogenesis." *Cell host & microbe*, vol. 4, no. 6, pp. 592–9, Dec. 2008.

[21] L.-A. Carlson, A. de Marco, H. Oberwinkler, A. Habermann, J. a. G. Briggs, H.-G. Kräusslich, and K. Grünewald, "Cryo electron tomography of native HIV-1 budding sites." *PLoS pathogens*, vol. 6, no. 11, p. e1001173, Jan. 2010.

[22] M. Carrillo-Tripp, C. M. Shepherd, I. A. Borelli, S. Venkataraman, G. Lander, P. Natarajan, J. E. Johnson, C. L. Brooks, and V. S. Reddy, "VIPERdb2: an

enhanced and web API enabled relational database for structural virology," *Nucleic Acids Research*, vol. 37, no. Database, pp. D436–D442, Jan. 2009, copyright 1998-2004 by TSRI. [Online]. Available: http://viperdb.scripps.edu/

[23] D. Caspar and A. Klug, "Physical Principles in the Construction of Regular Viruses," *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 27, pp. 1–24, Jan. 1962.

[24] M. Castellanos, R. Pérez, C. Carrasco, M. Hernando-Pérez, J. Gómez-Herrero, P. J. de Pablo, and M. G. Mateu, "Mechanical elasticity as a physical signature of conformational dynamics in a virus particle." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 30, pp. 12 028–33, Jul. 2012.

[25] M. Castelnovo, T. Verdier, and L. Foret, "Comparing open and closed molecular self-assembly," *EPL (Europhysics Letters)*, vol. 105, 2014.

[26] J. Castón, "Conventional electron microscopy, cryo-electron microscopy and cryo-electron tomography of viruses," in *Structure and Physics of Viruses*, ser. Subcellular Biochemistry, M. G. Mateu, Ed. Springer Netherlands, 2013, vol. 68, pp. 79–115. [Online]. Available: http://dx.doi.org/10.1007/978-94-007-6552-8_3

[27] J. Cavanagh, W. J. Fairbrother, A. G. P. III, M. Rance, and N. J. Skelton, *Protein {NMR} Spectroscopy (Second Edition)*, 2nd ed., J. Cavanagh, W. J. Fairbrother, A. G. Palmer, M. Rance, and N. J. Skelton, Eds. Burlington: Academic Press, 2007.

[28] S. Chandrasekhar, "Stochastic Problems in Physics and Astronomy," *Reviews of Modern Physics*, vol. 15, no. 1, pp. 1–89, Jan. 1943.

[29] C. Chen, C. C. Kao, and B. Dragnea, "Self-assembly of brome mosaic virus capsids: insights from shorter time-scale experiments." *The journal of physical chemistry. A*, vol. 112, no. 39, pp. 9405–12, Oct. 2008.

[30] B. A. Craig and P. P. Sendi, "estimating the transition matrix of an homogeneous markov chain," Purdue University, Medizinishe Universitäts-Polikinik Basel, Tech. Rep., 1998.

[31] F. H. Crick and J. D. Watson, "Structure of small viruses." *Nature*, vol. 177, no. 4506, pp. 473–475, 1956.

[32] ——, "Virus Structure: General Principles," in *Ciba Foundation Symposium - The Nature of Viruses*, ser. Novartis Foundation Symposia, G. E. W. Wolstenholme and E. C. P. Millar, Eds. Chichester, UK: John Wiley & Sons, Ltd., Jan. 1957.

[33] D. Crommelin and E. Vanden-Eijnden, "Fitting timeseries by continuous-time Markov chains: a quadratic programming approach," *Journal of Computational Physics*, vol. 217, no. 2, pp. 782–805, Sep. 2006.

[34] R. J. De Boer, R. M. Ribeiro, and A. S. Perelson, "Current estimates for hiv-1 production imply rapid viral clearance in lymphoid tissues," *PLoS Comput Biol*, vol. 6, no. 9, p. e1000906, 09 2010. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.pcbi.1000906

[35] P. Dedecker, G. C. H. Mo, T. Dertinger, and J. Zhang, "Widely accessible method for superresolution fluorescence imaging of living systems," *Proceedings of the National Academy of Sciences*, vol. 109, no. 27, pp. 10 909–10 914, Jul. 2012.

[36] M. Deserno, "Elastic deformation of a fluid membrane upon colloid binding," *Physical Review E*, vol. 69, no. 3, p. 031903, Mar. 2004.

[37] B. Efron and D. V. Hinkley, "Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information," *Biometrika*, vol. 65, no. 3, pp. 457–483, 1978.

[38] N. Elia, R. Sougrat, T. A. Spurlin, J. H. Hurley, and J. Lippincott-Schwartz, "Dynamics of endosomal sorting complex required for transport (ESCRT) machinery during cytokinesis and its role in abscission." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 12, pp. 4846–51, Mar. 2011.

[39] O. M. Elrad and M. F. Hagan, "Encapsulation of a polymer by an icosahedral virus," *Physical biology*, vol. 7, no. 4, p. 45003, Dec. 2010.

[40] D. Endres and A. Zlotnick, "Model-based analysis of assembly kinetics for virus capsids or other spherical polymers." *Biophysical journal*, vol. 83, no. 2, pp. 1217–30, Aug. 2002.

[41] D. Endres, M. Miyahara, P. Moisant, and A. Zlotnick, "A reaction landscape identifies the intermediates critical for self-assembly of virus capsids and other polyhedral structures." *Protein science : a publication of the Protein Society*, vol. 14, no. 6, pp. 1518–25, Jun. 2005.

[42] J. Finch and A. Klug, "Structure of poliomyelitis virus." *Nature*, vol. 183, no. 4677, pp. 1709–1714, 1959.

[43] R. A. Fisher, "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 222, no. 594-604, pp. 309–368, Jan. 1922.

[44] J. E. Fitzgerald, J. Lu, and M. J. Schnitzer, "Estimation Theoretic Measure of Resolution for Stochastic Localization Microscopy," *Physical Review Letters*, vol. 109, no. 4, p. 048102, Jul. 2012.

[45] J. Flusser and T. Suk, "Rotation Moment Invariants for Recognition of Symmetric Objects," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3784–3790, Dec. 2006.

[46] L. Foret, "Shape and energy of a membrane bud induced by protein coats or viral protein assembly," *The European Physical Journal E*, vol. 37, no. 5, p. 42, May 2014.

[47] L. Foret and P. Sens, "Kinetic regulation of coated vesicle secretion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 39, pp. 14 763–8, Sep. 2008.

[48] H. Fraenkel-Conrat and R. C.Williams, "Reconstitution of Active Tobacco Mosaic Virus From Its Inactive Protein and Nucleic Acid Components," *Proceedings of the National Academy of Sciences*, vol. 41, no. 10, pp. 690–698, 1955.

[49] B. K. B. Ganser-Pornillos, M. Yeager, and W. I. Sundquist, "The structural biology of HIV assembly," *... opinion in structural biology*, vol. 18, no. 2, pp. 1–24, Apr. 2008.

[50] B. Ganser-Pornillos, M. Yeager, and O. Pornillos, "Assembly and architecture of hiv," in *Viral Molecular Machines*, ser. Advances in Experimental Medicine and Biology, M. G. Rossmann and V. B. Rao, Eds. Springer US, 2012, vol. 726, pp. 441–465. [Online]. Available: http://dx.doi.org/10.1007/978-1-4614-0980-9_20

[51] D. T. Gillespie, "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions," *Journal of Computational Physics*, vol. 22, no. 4, pp. 403–434, Dec. 1976.

[52] D. Gillespie, "Exact stochastic simulation of coupled chemical reactions," *The journal of physical chemistry*, vol. 93555, no. 1, pp. 2340–2361, 1977.

[53] G. H. Golub and C. F. Van Loan, *Matrix Computations (3rd Ed.)*. Baltimore, MD, USA: Johns Hopkins University Press, 1996.

[54] a. B. Goncharov, "Integral geometry and three-dimensional reconstruction of randomly oriented identical particles from their electron microphotos," *Acta Applicandae Mathematicae*, vol. 11, no. 3, pp. 199–211, Mar. 1988.

[55] H. G. Göttlinger, T. Dorfman, J. G. Sodroski, and W. a. Haseltine, "Effect of mutations affecting the p6 gag protein on human immunodeficiency virus particle release." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 8, pp. 3195–3199, 1991.

[56] A. T. Gres, K. A. Kirby, V. N. KewalRamani, J. J. Tanner, O. Pornillos, and S. G. Sarafianos, "X-ray crystal structures of native HIV-1 capsid protein reveal conformational variability," *Science*, vol. 349, no. 6243, pp. 99–103, Jul. 2015.

[57] J. Guizetti and D. W. Gerlich, "ESCRT-III polymers in membrane neck constriction." *Trends in cell biology*, vol. 22, no. 3, pp. 133–40, Mar. 2012.

[58] J. Guizetti, L. Schermelleh, J. Mäntler, S. Maar, I. Poser, H. Leonhardt, T. Müller-Reichert, and D. W. Gerlich, "Cortical constriction during abscission involves helices of ESCRT-III-dependent filaments." *Science (New York, N.Y.)*, vol. 331, no. 6024, pp. 1616–20, Mar. 2011.

[59] J. Gunzenhäuser, N. Olivier, T. Pengo, and S. a. Manley, "Quantitative super-resolution imaging reveals protein stoichiometry and nanoscale morphology of assembling HIV-Gag virions," *Nano letters*, vol. 12, no. 9, pp. 1–8, Sep. 2012.

[60] M. F. Hagan, "Controlling Viral Capsid Assembly with Templating," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 77, no. 5 Pt 1, p. 51904, May 2008.

[61] P. R. Halmos, "The Theory of Unbiased Estimation," *The Annals of Mathematical Statistics*, vol. 17, no. 1, pp. 34–43, Mar. 1946.

[62] D. Harper, "vaccination," 2011. [Online]. Available: http://www.etymonline.com

[63] W. Helfrich, "Elastic properties of lipid bilayers: theory and possible experiments." *Zeitschrift fur Naturforschung. Teil C: Biochemie, Biophysik, Biologie, Virologie*, vol. 28, no. 11, pp. 693–703, 1973.

[64] S. T. Hess, T. P. K. Girirajan, and M. D. Mason, "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy." *Biophysical journal*, vol. 91, no. 11, pp. 4258–4272, Dec. 2006.

[65] S. T. Hess, T. J. Gould, M. V. Gudheti, S. a. Maas, K. D. Mills, and J. Zimmerberg, "Dynamic clustered distribution of hemagglutinin resolved at 40 nm in living cell membranes discriminates between raft theories." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 44, pp. 17 370–5, Oct. 2007.

[66] T. L. Hill, *An introduction to statistical thermodynamics*. New York: Dover Publications, 1986.

[67] Hu, "Visual pattern recognition by moment invariants," *IEEE Transactions on Information Theory*, vol. 8, no. 2, pp. 179–187, Feb. 1962.

[68] B. Huang, W. Wang, M. Bates, and X. Zhuang, "Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy." *Science*, vol. 319, no. 5864, pp. 810–813, Feb. 2008.

[69] S. Ivanchenko, W. J. Godinez, M. Lampe, H.-G. Kräusslich, R. Eils, K. Rohr, C. Bräuchle, B. Müller, and D. C. Lamb, "Dynamics of hiv-1 assembly and release," *PLoS Pathog*, vol. 5, no. 11, p. e1000652, 11 2009. [Online]. Available: http://dx.doi.org/10.1371%2Fjournal.ppat.1000652

[70] N. P. A. S. Johnson and J. Mueller, "Updating the Accounts: Global Mortality of the 1918-1920 "Spanish" Influenza Pandemic," *Bulletin of the History of Medicine*, vol. 76, no. 1, pp. 105–115, 2002.

[71] S. A. Jones, S.-H. Shim, J. He, and X. Zhuang, "Fast, three-dimensional super-resolution imaging of live cells," *Nature methods*, vol. 8, no. 6, pp. 499–508, Jun. 2011.

[72] N. Jouvenet, S. M. Simon, and P. D. Bieniasz, "Imaging the interaction of HIV-1 genomes and Gag during assembly of individual viral particles." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 45, pp. 19 114–9, Nov. 2009.

[73] N. Jouvenet, S. S. M. Simon, and P. P. D. Bieniasz, "Visualizing HIV-1 assembly," *Journal of molecular biology*, vol. 410, no. 4, pp. 501–11, Jul. 2011.

[74] M. F. Juette, T. J. Gould, M. D. Lessard, M. J. Mlodzianoski, B. S. Nagpure, B. T. Bennett, S. T. Hess, and J. Bewersdorf, "Three-dimensional sub-100 nm resolution fluorescence microscopy of thick samples." *Nature methods*, vol. 5, no. 6, pp. 527–9, Jun. 2008.

[75] W. K. Kegel and P. V. D. Schoot Pv, "Competing hydrophobic and screened-coulomb interactions in hepatitis B virus capsid assembly." *Biophysical journal*, vol. 86, no. 6, pp. 3905–3913, 2004.

[76] D. Keller, M. Orpinell, N. Olivier, M. Wachsmuth, R. Mahen, R. Wyss, V. Hachet, J. Ellenberg, S. Manley, and P. Gönczy, "Mechanisms of HsSAS-6 assembly promoting centriole formation in human cells." *The Journal of cell biology*, vol. 204, no. 5, pp. 697–712, Mar. 2014.

[77] A. K. Kenworthy, "Dynamics of putative raft-associated proteins at the cell surface," *The Journal of Cell Biology*, vol. 165, no. 5, pp. 735–746, jun 2004. [Online]. Available: http://www.jcb.org/cgi/doi/10.1083/jcb.200312170

[78] M. a. Kiskowski, J. F. Hancock, and A. K. Kenworthy, "On the use of Ripley's K-function and its derivatives to analyze domain size." *Biophysical journal*, vol. 97, no. 4, pp. 1095–103, Aug. 2009.

[79] K. Kremer, M. Deserno, B. J. Reynwar, G. Illya, V. A. Harmandaris, M. M. Mu, and M. M. Müller, "Aggregation and vesiculation of membrane proteins by curvature-mediated interactions," *Nature*, vol. 447, no. 7143, pp. 461–464, may 2007. [Online]. Available: http://www.nature.com/nature/journal/v447/n7143/full/nature05840.htmlhttp://www.nature.com/nature/journal/v447/n7143/pdf/nature05840.pdf

[80] P.-I. Ku, A. K. Miller, J. Ballew, V. Sandrin, F. R. Adler, and S. Saffarian, "Identification of pauses during formation of HIV-1 virus like particles." *Biophysical journal*, vol. 105, no. 10, pp. 2262–72, Nov. 2013.

[81] K. B. Laupland and L. Valiquette, "Ebola virus disease," *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 25, no. 3, pp. 128–129, 2014.

[82] A.-M. Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*, Didot, Ed. Paris: Firmin Didot, 1805.

[83] M. Lelek, F. Di Nunzio, R. Henriques, P. Charneau, N. Arhel, C. Zimmer, and F. D. Nunzio, "Superresolution imaging of HIV in infected cells with FlAsH-PALM." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 22, pp. 8564–9, May 2012.

[84] M. Lenz, D. Crow, and J. Joanny, "Membrane buckling induced by curved filaments," *Physical review letters*, vol. 103, no. 3, Jul. 2009.

[85] J. K. Lewis, B. Bothner, T. J. Smith, and G. Siuzdak, "Antiviral agent blocks breathing of the common cold virus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 12, pp. 6774–6778, Jun. 1998.

[86] A. Luque Santolaria, "Structure, Mechanical Properties, and Self-Assembly of Viral Capsids," Ph.D. dissertation, Universitat de Barcelona., 2011.

[87] J. P. Mahalik and M. Muthukumar, "Langevin dynamics simulation of polymer-assisted virus-like assembly." *The Journal of chemical physics*, vol. 136, no. 13, p. 135101, Apr. 2012.

[88] S. Malkusch, W. Muranyi, B. Müller, H.-G. Kräusslich, and M. Heilemann, "Single-molecule coordinate-based analysis of the morphology of HIV-1 assembly sites with near-molecular spatial resolution." *Histochemistry and cell biology*, vol. 139, no. 1, pp. 173–9, Jan. 2013.

[89] S. Manley, J. J. M. Gillette, G. H. G. G. H. G. Patterson, H. Shroff, H. F. Hess, E. Betzig, J. Lippincott-Schwartz, and S. Methods, "High-density mapping of single-molecule trajectories with photoactivated localization microscopy." *Nature methods*, vol. 5, no. 2, pp. 155–7, Feb. 2008.

[90] M. G. Mateu, "Mechanical properties of viruses analyzed by atomic force microscopy: a virological perspective." *Virus research*, vol. 168, no. 1-2, pp. 1–22, Sep. 2012.

[91] M. Mateu, *Structure and Physics of Viruses*, ser. Subcellular Biochemistry, M. G. Mateu, Ed. Dordrecht: Springer Netherlands, 2013, vol. 68.

[92] R. Matthews and C. Likos, "Influence of Fluctuating Membranes on Self-Assembly of Patchy Colloids," *Physical Review Letters*, vol. 109, no. 17, pp. 1–5, Oct. 2012.

[93] A. Mcleod, N. Morgan, A. Prakash, and J. Hinrichs, "Economic and social impaacts of avian influenza," FAO Emergency Centre for Transboundary Animal Diseases Operations (ECTAD), Tech. Rep. September, 2005.

[94] A. McPherson, "Micelle formation and crystallization as paradigms for virus assembly," *BioEssays*, vol. 27, no. 4, pp. 447–458, 2005.

[95] V. Mennella, B. Keszthelyi, K. L. McDonald, B. Chhun, F. Kan, G. C. Rogers, B. Huang, and D. A. Agard, "Subdiffraction-resolution fluorescence microscopy reveals a domain of the centrosome critical for pericentriolar material organization." *Nature cell biology*, vol. 14, no. 11, pp. 1159–68, Nov. 2012.

[96] P. Metzner, E. Dittmer, T. Jahnke, and C. Schütte, "Generator estimation of Markov jump processes," *Journal of Computational Physics*, vol. 227, no. 1, pp. 353–375, Nov. 2007.

[97] C. H. Moon, M. Zhang, N. V. Myung, and E. D. Haberer, "Highly sensitive hydrogen sulfide (h2s) gas sensors from viral-templated nanocrystalline gold nanowires," *Nanotechnology*, vol. 25, no. 13, p. 135205, 2014. [Online]. Available: http://stacks.iop.org/0957-4484/25/i=13/a=135205

[98] A. Y. Morozov, R. F. Bruinsma, and J. Rudnick, "Assembly of viruses and the pseudo-law of mass action." *The Journal of chemical physics*, vol. 131, no. 15, p. 155101, Oct. 2009.

[99] E. a. Mukamel and M. J. Schnitzer, "Unified performance limits for conventional and stochastic localization microscopy," *Physical Review Letters*, vol. 109, no. 16, pp. 1–19, Oct. 2012.

[100] E. a. Mukamel, H. Babcock, and X. Zhuang, "Statistical deconvolution for superresolution fluorescence microscopy." *Biophysical journal*, vol. 102, no. 10, pp. 2391–400, May 2012.

[101] M. Müller and M. Deserno, "Cell model approach to membrane mediated protein interactions," *Progress of Theoretical Physics ...*, no. 184, 2010.

[102] W. Muranyi, S. Malkusch, B. Müller, M. Heilemann, and H.-G. Kräusslich, "Super-resolution microscopy reveals specific recruitment of HIV-1 envelope proteins to viral assembly sites dependent on the envelope C-terminal tail." *PLoS pathogens*, vol. 9, no. 2, p. e1003198, Feb. 2013.

[103] M. Muthukumar, "Polymers under confinement," in *Advances in Chemical Physics*, ser. Advances in Chemical Physics, S. A. Rice and A. R. Dinner, Eds. Hoboken, NJ, USA: John Wiley & Sons, Inc., Feb. 2012, vol. 149.

[104] M. V. Nermut, D. J. Hockley, P. Bron, D. Thomas, W. H. Zhang, and I. M. Jones, "Further evidence for hexagonal organization of HIV gag protein in prebudding assemblies and immature virus-like particles." *Journal of structural biology*, vol. 123, no. 2, pp. 143–149, 1998.

[105] A. J. Olson, Y. H. E. Hu, and E. Keinan, "Chemical mimicry of viral capsid self-assembly." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 52, pp. 20 731–20 736, 2007.

[106] "virus," Oxford University Press. [Online]. Available: http://www.oxforddictionaries.com/definition/english/virus

[107] C. C. Paige, "Computer solution and perturbation analysis of generalized linear least squares problems," *Mathematics of Computation*, vol. 33, no. 145, pp. 171–183, 1979.

[108] Y. Pawitan, *In all likelihood statistical modelling and inference using likelihood*. Oxford: Clarendon Press Oxford University Press, 2013.

[109] V. S. Reddy, S. K. Natchiar, P. L. Stewart, and G. R. Nemerow, "Crystal Structure of Human Adenovirus at 3.5 A Resolution," *Science*, vol. 329, no. 5995, pp. 1071–1075, Aug. 2010.

[110] S. Redner, *A Guide To First-Passage Processes*. Cambridge: Cambridge University Press, 2001.

[111] R.J. Reynolds Tobacco Company, "R.J. Reynolds Tobacco Company Slide Set." [Online]. Available: bugwood.org

[112] M. J. Rust, M. Bates, X. Zhuang, and B. Communications, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM)," *Nature methods*, vol. 3, no. 10, pp. 793–795, 2006.

[113] C. Sachse, J. Z. Chen, P.-D. Coureux, M. E. Stroupe, M. Fändrich, and N. Grigorieff, "High-resolution electron microscopy of helical specimens: a fresh look at tobacco mosaic virus." *Journal of molecular biology*, vol. 371, no. 3, pp. 812–35, Aug. 2007.

[114] P. Sengupta and J. Lippincott-Schwartz, "Quantitative analysis of photoactivated localization microscopy (PALM) datasets using pair-correlation analysis." *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 34, no. 5, pp. 396–405, May 2012.

[115] H. Shroff, C. G. Galbraith, J. A. Galbraith, and E. Betzig, "Live-cell photoactivated localization microscopy of nanoscale adhesion dynamics," *Nature Methods*, vol. 5, no. 5, pp. 417–423, May 2008.

[116] G. Shtengel, J. a. Galbraith, C. G. Galbraith, J. Lippincott-Schwartz, J. M. Gillette, S. Manley, R. Sougrat, C. M. Waterman, P. Kanchanawong, M. W. Davidson, R. D. Fetter, and H. F. Hess, "Interferometric fluorescent super-resolution microscopy resolves 3D cellular ultrastructure." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 9, pp. 3125–30, Mar. 2009.

[117] A. Siber, A. L. Bozic, and R. Podgornik, "Energies and pressures in viruses: contribution of nonspecific electrostatic interactions," *Phys. Chem. Chem. Phys.*, vol. 14, pp. 3746–3765, 2012.

[118] J. E. Sillibourne, C. G. Specht, I. Izeddin, I. Hurbain, P. Tran, A. Triller, X. Darzacq, M. Dahan, and M. Bornens, "Assessing the localization of centrosomal proteins by PALM/STORM nanoscopy." *Cytoskeleton (Hoboken, N.J.)*, vol. 68, no. 11, pp. 619–27, Nov. 2011.

[119] S. Singh and A. Zlotnick, "Observed hysteresis of virus capsid disassembly is implicit in kinetic models of assembly." *The Journal of biological chemistry*, vol. 278, no. 20, pp. 18 249–55, May 2003.

[120] D. E. Smith, S. J. Tans, S. B. Smith, S. Grimes, D. L. Anderson, and C. Bustamante, "The bacteriophage [phis]29 portal motor can package DNA against a large internal force," *Nature*, vol. 413, no. 6857, pp. 748–752, Oct. 2001.

[121] H. Smith, *Applied regression analysis*, 3rd ed., ser. Wiley series in probability and statistics. New York: J. Wiley & Sons Interscience, 1998.

[122] T. Splettstoesser, "TMV structure full," licensed under CC BY-SA 3.0 via Commons. [Online]. Available: www.scistyle.com

[123] W. M. Stanley, "Isolation of a crystalline protein possessing the properties of the tobacco-mosaic virus." *Science (New York, N.Y.)*, vol. 81, no. 2113, pp. 644–645, 1935.

[124] N. F. Steinmetz, A. L. Ablack, J. L. Hickey, J. Ablack, B. Manocha, J. S. Mymryk, L. G. Luyt, and J. D. Lewis, "Viral nanoparticles: Intravital imaging of human prostate cancer using viral nanoparticles targeted to gastrin-releasing peptide receptors (small 12/2011)," *Small*, vol. 7, no. 12, pp. 1602–1602, 2011.

[125] A. Szymborska, A. de Marco, N. Daigle, V. C. Cordes, J. A. G. Briggs, and J. Ellenberg, "Nuclear pore scaffold structure analyzed by super-resolution microscopy and particle averaging." *Science (New York, N.Y.)*, vol. 341, no. 6146, pp. 655–8, Aug. 2013.

[126] C.-H. Teh and R. Chin, "On image analysis by the methods of moments," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 10, no. 4, pp. 496–513, Jul. 1988.

[127] R. E. Thompson, D. R. Larson, and W. W. Webb, "Precise nanometer localization analysis for individual fluorescent probes." *Biophysical journal*, vol. 82, no. 5, pp. 2775–2783, 2002.

[128] D. C. Torney and H. M. McConnell, "Diffusion-limited reaction rate theory for two-dimensional systems," *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 387, no. 1792, pp. 147–170, 1983. [Online]. Available: http://rspa.royalsocietypublishing.org/content/387/1792/147

[129] B. G. Turner and M. F. Summers, "Structural biology of HIV." *Journal of molecular biology*, vol. 285, no. 1, pp. 1–32, 1999.

[130] S. B. Van Engelenburg, G. Shtengel, P. Sengupta, K. Waki, M. Jarnik, S. D. Ablan, E. O. Freed, H. F. Hess, and J. Lippincott-Schwartz, "Distribution of ESCRT machinery at HIV assembly sites reveals virus scaffolding of ESCRT subunits." *Science (New York, N.Y.)*, vol. 343, no. 6171, pp. 653–6, Feb. 2014.

[131] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, 3rd ed., ser. North-Holland personal library. Amsterdam Boston (Mass.) Heidelberg [etc.]: Elsevier, 2007.

[132] A. Wald, "Note on the Consistency of the Maximum Likelihood Estimate," *The Annals of Mathematical Statistics*, vol. 20, no. 4, pp. 595–601, Dec. 1949.

[133] T. Wilk, I. Gross, B. Gowen, and T. Rutten, "Organization of immature human immunodeficiency virus type 1," *Journal of Virology*, vol. 72, no. 2, pp. 759–771, 2001.

[134] E. R. Wright, J. B. Schooler, H. J. Ding, C. Kieffer, C. Fillmore, W. I. Sundquist, and G. J. Jensen, "Electron cryotomography of immature HIV-1 virions reveals the structure of the CA and SP1 Gag shells." *The EMBO Journal*, vol. 26, no. 8, pp. 2218–26, Apr. 2007.

[135] M. Zaitlin, "The Discovery of the Causal Agent of the Tobacco Mosaic Disease," in *Discoveries in Plant Biology*. World Publishing Co., Ltd. Hong Kong., 1998, pp. 105–110.

[136] R. Zandi, P. van der Schoot, D. Reguera, W. Kegel, and H. Reiss, "Classical nucleation theory of virus capsids." *Biophysical journal*, vol. 90, no. 6, pp. 1939–48, Mar. 2006.

[137] T. Zhang and R. Schwartz, "Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics." *Biophysical journal*, vol. 90, no. 1, pp. 57–64, Jan. 2006.

[138] P. Zhu, E. Chertova, J. Bess, J. D. Lifson, L. O. Arthur, J. Liu, K. a. Taylor, and K. H. Roux, "Electron tomography analysis of envelope glycoprotein trimers on HIV and simian immunodeficiency virus virions." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 26, pp. 15 812–15 817, 2003.

[139] 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, Aug. 1994.

[140] a. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, and M. J. Young, "Mechanism of capsid assembly for an icosahedral plant virus." *Virology*, vol. 277, no. 2, pp. 450–6, Nov. 2000.

[141] A. Zlotnick and S. J. Stray, "How does your virus grow? Understanding and interfering with virus assembly." *Trends in biotechnology*, vol. 21, no. 12, pp. 536–42, Dec. 2003.

[142] A. Zlotnick, J. M. Johnson, P. W. Wingfield, S. J. Stahl, and D. Endres, "A Theoretical Model Successfully Identifies Features of Hepatitis B Virus Capsid Assembly †," *Biochemistry*, vol. 38, no. 44, pp. 14 644–14 652, Nov. 1999.