Principles of self-assembly for particles with simple geometries and complex interactions

Lara Koehler

université PARIS-SACLAY

Principles of self-assembly for particles with simple geometries and complex interactions

Principes d'auto-assemblage pour des particules avec des géométries simples et des interactions complexes

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°564, physique en Île-de-France (PIF)
Spécialité de doctorat: Physique
Graduate School : Physique. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **LPTMS** (Université Paris-Saclay, CNRS), sous la direction de **Martin LENZ**, directeur de recherche.

Thèse soutenue à Paris-Saclay, le 10 juillet 2023, par

Lara KOEHLER

Composition du jury

Membres du jury avec voix délibérative

Friedrich SIMMEL Professeur, Technical University Munich	Président
Michael BRENNER Professeur, Harvard University	Rapporteur & Examinateur
Ralf EVERAERS Professeur, ENS Lyon	Rapporteur & Examinateur
Zorana ZERAVCIC Maître de conférence, ESPCI	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASP070

université PARIS-SACLAY | ÉCOLE DOCTORALE Physique en Île-de-France (PIF)

Titre : Principes d'auto-assemblage pour des particules avec des géométries simples et des interactions complexes

Mots clés : Auto-assemblage, frustration géométrique, protéines, modèle de particules sur réseau

Résumé : Dans les cellules vivantes, les protéines s'auto-assemblent en agrégats de formes diverses pour réaliser des fonctions biologiques. Les interactions locales entre les protéines contrôlent la forme des agrégats, notamment les interactions attractives entre les résidus à la surface des protéines. Malgré la diversité de ces interactions, seules quelques catégories d'agrégats de protéines sont observées, tels que des oligomères, des fibres, des capsides virales et des micelles. Les protéines similaires provenant d'organismes différents peuvent former des agrégats différents. Les modèles de particules à patchs collants sont utilisés pour simuler l'auto-assemblage. Pourtant, ils ne tiennent pas compte des variations subtiles de l'attraction des patchs ni de la relation entre les propriétés des interactions et la forme de l'agrégat. Dans cette thèse, nous proposons l'hypothèse que les particules avec des interactions complexes peuvent présenter de la frustration géométrique, c'est-à-dire des interactions attractives incompatibles en raison de contraintes géométriques. Nous développons un modèle de particules sur un réseau à deux dimensions et étudions l'auto-assemblage en fonction des interactions locales. En pratique, nous considérons des particules hexagonales qui sont en contact par leurs faces et définissons une carte d'interaction. Nous étudions l'auto-assemblage de particules avec une carte d'interaction choisie avec un recuit simulé de type Monte-Carlo vers une température finie. Pour une particule avec un ensemble donné d'interactions locales, on peut alors déterminer la forme de l'agrégat résultant de l'auto-assemblage à l'équilibre des particules. Nous explorons également un grand nombre de cartes d'interactions aléatoires pour comprendre la relation entre les interactions locales et la forme de l'agrégat. Nous identifions huit catégories d'agrégats et utilisons l'apprentissage automatique pour classifier les résultats de l'auto-assemblage. Nous constatons que l'énergie de l'organisation périodique la plus stable des particules est un bon prédicteur de la forme de l'agrégat. Nous introduisons également une transformation numérique de renormalisation pour explorer les paramètres d'interactions et identifier les points fixes stables, qui garantit que le nombre d'occurrences de chaque paire de particules est conservé dans un réseau de maille plus grande. Nous n'utilisons pas la renormalisation comme outil pour mesurer les exposants critiques au voisinage d'une transition de phase. Nous constatons que la renormalisation permet de rationaliser l'existence de quelques catégories d'agrégats malgré la complexité des interactions. Ensuite, nous étudions un type spécifique d'interaction conduisant à un agrégat cristallin avec des lignes de défaut favorables, appelé agrégat camembert. Cet agrégat est frustré et peut avoir une taille finie contrôlée par la force des interactions. Nous établissons le diagramme de phases à température nulle et confirmons nos résultats par des simulations numériques à température finie. Ce mécanisme est complémentaire des mécanismes existants qui reposent sur le design individuel de chaque particule, sur son auto-fermeture, ou sur la déformabilité des particules. Nous expliquons des idées préliminaires pour tester ce design dans une réalisation expérimentale hors réseau à partir d'origami d'ADN. Enfin, nous proposons une méthode pour détecter l'auto-assemblage en fibres de protéines avec des interactions arbitraires en analysant les signaux de diffusion dans des expériences cristallographiques. Nous suggérons d'utiliser l'apprentissage automatique supervisé pour reconnaître les agrégats fibrillaires à partir des données expérimentales, et montrons qu'il est possible de tirer parti de la grande quantité de données disponibles, à condition que le réseau de neurones soit entraîné sur une grande variété d'agrégats protéiques de dimensionnalité connue, et dans différentes configurations expérimentales.

université PARIS-SACLAY | ÉCOLE DOCTORALE Physique en Île-de-France (PIF)

Title: Principles of self-assembly for particles with simple geometries and complex interactions
Keywords: Self-assembly, geometric frustration, proteins, lattice particle models

Abstract: In living cells, proteins self-assemble into aggregates of various shapes to perform biological functions. Local interactions between proteins control the shape of aggregates, including attractive interactions between residues on the surface of proteins. Despite the diversity of these interactions, only a few categories of protein aggregates are observed, such as oligomers, fibers, viral capsids and micelles. Similar proteins from different organisms may form different aggregates. Patchy particle models are used to simulate self-assembly, but they do not account for subtle variations in patch attraction or the relationship between the properties of interactions and the shape of the aggregate. In this thesis, we propose the hypothesis that particles with complex interactions can exhibit geometric frustration, that is, incompatible attractive interactions due to geometric constraints. We develop a model of particles on a two-dimensional lattice and study self-assembly as a function of local interactions. In practice, we consider hexagonal particles which are in contact by their faces and define an interaction map. We study the self-assembly of particles with a chosen interaction map with a simulated Monte-Carlo annealing towards a finite temperature. For a particle with a given set of local interactions, one can then determine the shape of the aggregate resulting from the equilibrium self-assembly of the particles.

We explore a large number of random interaction maps to understand the relationship between local interactions and the shape of the aggregate. We identify eight categories of aggregates and use machine learning to classify the results of self-assembly. We find that the energy of the most stable periodic organization of the particles is a good predictor of the shape of the aggregate. We also introduce a numerical renormalization transformation to explore the parameters of interactions and identify stable fixed points, which ensures that the number of occurrences of each pair of particles is conserved in a coarse-grained lattice. We do not use renormalization as a tool to measure critical exponents near a phase transition. We note that the renormalization allows to rationalize the existence of some categories of aggregates despite the complexity of the interactions. Next, we study a specific type of interaction leading to a crystalline aggregate with favorable disclination lines, called a camembert aggregate. This aggregate is frustrated and can have a finite size, controlled by the strength of the interactions. We establish the phase diagram at zero temperature and confirm our results by numerical simulations at finite temperature. This mechanism is complementary to existing mechanisms which rely on the individual design of each particle, on its self-closing, or on the deformability of the particles. We explain preliminary ideas for testing this design in an off-grid experimental realization from DNA origami. Finally, we propose a method to detect self-assembly into protein fibers with arbitrary interactions by analyzing scattering signals in crystallographic experiments. We suggest using supervised machine learning to recognize fibrillar aggregates from experimental data, and show that it is possible to take advantage of the large amount of data available, provided the neural network is trained on a wide variety of protein aggregates of known dimensionality, and in different experimental setups.

Contents

Remerciements

Cette thèse a été financée par le Corps des Ingénieurs des Ponts, des Eaux et des Forêts, et le ministère de la transition écologique. Je remercie l'ensemble du jury pour avoir accepté de revoir mon travail et de participer à ma soutenance.

Je suis extrêmement reconnaissante à Martin Lenz, de m'avoir donné l'opportunité de travailler sur un projet de recherche aussi motivant et passionnant. Son exigence sans relache, tant sur le contenu scientifique que sur la communication, m'ont donné des méthodes de travail que j'espère garder toute ma vie. Ces trois années de travail furent un vrai plaisir, notamment grâce à sa disponibilité, sa bienveillance, et ses conseils avisés. Si j'ai envie de continuer à faire de la recherche, je le dois en partie à l'exemple inspirant d'un scientifique équilibré, pédagogique et rigoureux qu'il donne au quotidien.

During my PhD, I had the great opportunity to have long stays in other institutes: the Frey group in Munich, the PCS in Daejeon, and the PMMH in Paris. I am extremely grateful to their members that welcomed and integrated me so well, leading to amazing encounters on top of fruitful scientific discussions. I thank their directors, Erwin Frey, Sergej Flach and Damien Vandembroucq for their hospitality.

I also thank the students and post-doc of the biophysics group with which I had the chance to work, discuss, learn, or share those long Friday afternoon journal clubs with: Hugo Le Roy, Felix Benoist, Mert Terzi, Valerio Sorichetti, Mayarani M., Lukas Kalvoda, Marianne Billoir, Martin Garic, Martin Flament, Clara Delahousse, and Pawat Akara. Many thanks to collaborators I had the chance to discuss and work with, in particular Pierre Ronceray, who was involved in most of the projects of this thesis, but also Monika Spano, William Shepard, Olivia du Roure and Julien Heuvingh. And thanks to my PhD referees, Aleksandra Walczak and Raoul Santachiara.

Je remercie l'ensemble des membres du LPTMS de contribuer à cet environnement de travail aussi positif, et en particulier ses directeurs successifs, Emmanuel Trizac et Alberto Rosso. Merci aussi à Claudine Le Vaou et Karolina Kolodziej pour leur aide et leur gestion remarquable. Ces trois années auraient aussi été bien moins joyeuses sans les innombrables pauses cafés, apéro, barbecues, et séances de course ou d'escalade, qui ont initié des amitiés qui, j'espère, dureront au-delà de la thèse. I am especially grateful to Sap, for his infinite generosity (and all the pain au chocolat), Jules pour son authenticité et son empathie, Ana for her constant curiosity and cheerfulness, Saverio for his motivation and cultivated conversation, Andrea for his enthusiasm, Vincenzo for his endless energy, Fabian for his federating musical skills, Mauro for his endearing cynicism, and also Federico, Louis, Benjamin, Marco, Charbel, Flavio and Lorenzo.

Cette thèse a également été supportée par des sponsors indirects. Un grand merci à Etienne, Marjorie, Cécile et Guillaume de m'avoir fourni des lieux de télétravail idylliques qui m'ont maintenu à flot pendant la pandémie. Merci à Mauro de m'avoir covoituré avec enthousiasme pendant un mois et demi de jambe platrée. Merci à ceux qui ont eu le malheur de proposer de relire cette thèse, ce que je me suis empressée d'accepter, Félix, Renaud, Inès, Pauline et Alban. Je remercie les proches avec qui j'ai pu partagé mon goût de la science très jeune, Jacques et Moana. Merci à celles qui m'ont apporté une oreille attentive, leur conseil et leur soutien, à l'heure des choix de postdoc, Ambre et Salambô. Enfin un grand merci à mes amis qui ont rendu le quotidien plus doux, et que je n'ai pas encore cité, Alix, Romane, Marine, Vérane, Mélisende, Philippine, Meriem, Mathilde, et Martin.

Finalement, je remercie mes parents, supporters infaillibles depuis toujours, de m'avoir transmis le gout du travail, mais aussi de l'intégrité et de la capacité à profiter de chaque instant. Un grand merci à ma soeur Alexa, pour sa confiance, son grain de folie, et ces centaines de fou rires. Et merci à Gabriel, d'avoir embelli la dernière année et demi, grâce à sa curiosité contagieuse, son humour et sa patience apaisante.

1 - Introduction: frustrated self-assembly of protein-like particles

Self-assembly is a process in which individual constituents come together in an aggregate with a specific geometry. The relationship between the shape of the individual constituents and the geometry of the aggregates they form is not well understood. For instance, proteins assemble into a large variety of functional biological aggregates, and there seem to be generic principles that dictate which shape they will form, beyond evolutionary pressure. We further detail this phenomenon in Sec. 1.1. The intrinsic dependency of the assembly shape on the interactions between its individual constituents can also be exploited beyond biological questions. Self-assembly is indeed a preferential building tool for small scale objects: the individual particles are engineered in such a way that they will assemble into the desired geometry. In Sec. 1.2, we show that individual particles are built following a few well-studied design principles. However, self-assembling constituents can have incompatible interactions due to geometric constraints. In that case, their interactions are defined as *frustrated*. The concept of frustration has only been partially studied in the context of self-assembly in dilute environments. In Sec. 1.3, we explain the implications of frustration for self-assembly. Frustration in the interactions could provide both new design principles for nano-materials, and a better understanding of the self-assembly of complex constituents such as proteins.

1.1 Self-assembly rules of proteins are not fully understood

To achieve various biological functions, cells rely on proteins that form large symmetric assemblies [1] of diverse shapes, sizes, and constituent organizations. In Sec. 1.1.1, we show how the functionality of a protein complex is dictated by its shape. There are also indications that the result of protein assembly is not just fine-tuned by evolution. Indeed, there is no unique outcome for the assembly of a protein (Sec. 1.1.2), and proteins can also assemble into pathological aggregates (Sec. 1.1.3).

1.1.1 Protein assemblies have biological functions

Here, we explain what a protein is and show that the functionality of a protein aggregate is often related to its geometry, suggesting that protein aggregates geometries are optimized by evolutionary processes.

A protein is a long polymer of amino-acids. Most of the time, the amino-acid chain folds and organizes in a three-dimensional well defined structure of a few nanometers, with some amino-acids at the surface of the protein, and some others buried inside the structure. The amino-acids on the surface of the protein can be seen as sticky spots, that will enable the protein to bind to its neighbors. An individual protein is a building subunit for the cell, that can assemble with others to perform specific functions.

Proteins often form *complexes* of a few subunits. We show with some examples the cooperative nature of those complexes. For instance, Piezo1 is a protein in the membrane of the cells composed of three subunits (Figure 1.1a, each subunits have a different color). Each subunit has a blade, that is thought to sense the mechanical distortion of the membrane. Upon mechanical change in the environment, the blades are deformed, and an ion channel will open in the middle of the three subunits to initiate the cell reaction to the changes [2]. Lactate dehydrogenase is composed of four subunits (Figure 1.1b), and it will eliminate lactic acid accumulated in the muscles after anaerobic exercise [3]. Cellu-

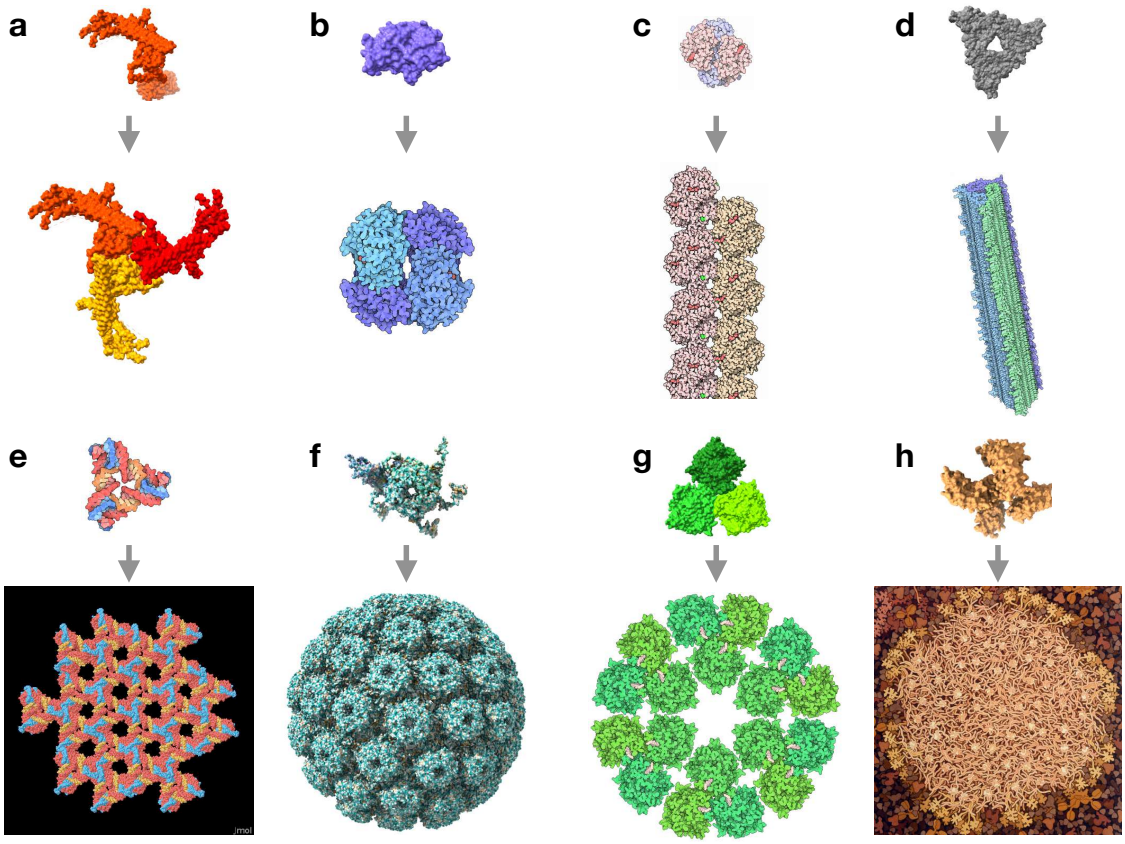

Figure 1.1: Biological assemblies have diverse shapes. a) Piezo1 is composed of three subunits. It is an ion channel that responds to pressure [2]. b) Lactate dehydrogenase is composed of four subunits. It is involved in the elimination of lactic acid after anaerobic exercise. [3]. c) Sickle cell hemoglobins form fibers [7] and are responsible for sickle-cell anemia. d) Beta-amyloid peptides form fibrils [8] and are responsible for Alzheimer's disease. e) Small pieces of DNA are engineered to form a lattice [9]. f) Simian virus capsids are composed of pentameric units [10]. g) Cellulose synthase is composed of six trimers of enzyme. Each enzyme synthesizes a cellulose fibril [4] h) Casein proteins form micelles in milk [11, 12].

lose synthase forms groups of six trimers (Figure 1.1h), and each of these 18 subunits will synthesize a cellulose filament. Those 18 filaments will then form a stiff fibril that gives structure to the cells of green plants [4].

Protein assemblies can also be larger than a few subunits, and have different shapes, such as capsids or fibers. The capsids of viruses result from the symmetric organization of a precise number of subunits [5], and enclose the genetic material of the virus. An example of viral capsids is shown in (Figure 1.1f). The mechanical resistance of the cells is controlled by the cytoskeleton filaments, which are linear assemblies of actin or microtubule subunits. These fibers form intertwined networks that control the rigidity of the cell.

Finally, protein assemblies are not necessarily organized. For instance, casein proteins, present in high concentration in milk, assemble into micelles of nanometric sizes, and are responsible for some of the milk properties [6]. Idealized image of such a casein micelle is shown in Figure 1.1h.

These examples show that proteins assemble into very diverse shapes. Because the assembly shape is related to the function, proteins have probably been optimized by evolution to form such assemblies.

1.1.2 Protein assemblies are modular

A protein can form different assemblies, such that there is no single functional complex or assembly for a given protein. We give some examples of proteins that form different complexes in the same organism, or in different species. It means that there is no clear rule to rationalize which parts of the protein surface are in contact with the neighboring protein in the complex. As a consequence, it is hard to predict what complex a protein will give.

In some cases, the same protein can alternatively assemble into one complex or another, within the same organism. Smad protein, involved in cell growth, usually forms a trimer, unless a phosphate group is attached to one of its surface residues, in which case it does not form a trimer, and does not perform its function [13]. Bacterial CTPS enzyme forms a tetramer, and it is inactivated through its polymerization into a filament of tetramer, providing a fast switching mechanism for the cell [14]. The activity of an HIV protein is inhibited when the tetrameric complex is made more stable than the dimeric one, a mechanism that has been explored as a drug design methodology [15].

Similar proteins also assemble into different complexes from one organism to the other. This has been observed for a specific enzyme that forms dimers: the orientations of the protein in the dimer are different between a virus and other species, and gave the virus an evolutionary advantage [16]. Similarly, plant lectins in different species have similar individual structures, but assemble into completely different oligomers from one species to the other [17]. Cytoplasmic enzymes CTPS also form different types of filaments in prokaryotes and eukaryotes [18].

As a consequence of this modularity, it is difficult to predict which amino-acids of a protein are involved in the contact with its neighbor in the complex. Then, it is difficult to predict the organization of a protein assembly from the 3D structure of the individual proteins it is made of. Comparing amino-acid compositions of many protein surfaces is not sufficient to discriminate between the parts of the protein surface involved in contacts, and the others [19]. For example, γD-crystallin protein crystallizes in two different polymorphs, and it was possible to identify which parts of the protein surface were involved in each of them only through detailed crystallographic analysis and modeling [20]. A whole field of research is dedicated to the determination of the residues of the protein surfaces that are in contact in a given complex (the protein *interfaces*) [21–23]. Computational tools have enabled some progress in this task [24], but correctly predicting the interface for protein complexes that are very dissimilar to previously identified complexes remains a challenge. In 2019, a panel of around 30 research group were not able to correctly predict complicated protein complexes [25].

The fact that there are several possibilities for the assembly of a protein suggests that a protein is not only fine-tuned by evolution to result in one specific assembly. In most cases, this modularity is beneficial for the cell because it provides switching mechanisms.

1.1.3 Unwanted protein aggregation is the cause of several diseases

Proteins sometimes form aggregates instead of remaining separated, and those aggregates are pathological. This can be illustrated by both specific examples of protein aggregation diseases, and systematic studies of protein mutations.

Pathological protein aggregates often have fibrillar shapes. This is the case with sickle cell hemoglobin, which aggregates into a stiff fiber after a mutation of an amino-acid on its surface. This fiber then deforms the red blood cell and causes sickle cell anemia [26]. The organization of such a fiber is shown in Figure 1.1c. Through a different mechanism, the amyloid beta precursor protein, when partially unfolded, aggregates in a fiber through the stacking of some of the amino-acids that are exposed to the solvent because of the unfold-

ing. An example of the organization of such a fiber obtained *in vitro* is shown in Figure1.1d. Those types of fibers are observed in patients suffering from Alzheimer's disease. Pathological amyloid fibrils are common in other diseases such as Parkinson's disease, related to α-synuclein aggregation, some cases of spongiform encephalopathies, related to prion aggregation, or type II diabetes [27, 28]. Pathological fiber aggregation in neurodegenerative diseases can also be caused by external factors, such as exposure to pesticides, that induce a change in the structure of the individual proteins [29, 30]. Modification of the structure of the protein can also lead to their aggregation in the eye crystallin, causing congenital cataracts: the protein aggregates scatter the light, and the crystallin loses its transparency [31]. Through very different mechanisms (misfolding, denaturation, mutation, and interactions with chemical substances), pathological protein aggregation results most of the time in the formation of fiber.

In most of the examples above, mutation or misfolding of the protein enables the emergence of new sticky spots at the surface of the protein, *i.e.* the amino-acids at the surface will interact with the surface of another protein. A systematic *in vivo* study of the emergence of new supramolecular assemblies upon mutation revealed that out of 73 mutated proteins, 30 aggregated into fibers or small size aggregates, even at low protein concentrations [32], while the mutations were not specifically meant to lead to this aggregation. The study suggests that proteins are in general likely to aggregate, and that evolution rather prevents those unwanted aggregations by adding on their surface some amino-acids that will not interact with others.

Examples of pathological protein aggregation confirm that protein assembly is not necessarily the result of evolutionary optimization to build functional objects. Moreover, it appears that fibrilar aggregates are often the result of those unwanted aggregations. This suggests that protein assembly might rely on physical rules that are common to any particle with complex shape and surfaces, beyond the specificities of each biological process.

1.2 Self-assembly is a design tool

Proteins assemblies do not require external action to bring the subunits together. They rather *self-assemble* because their subunits have attractive interactions. This principle can be used to build many types of assemblies at a very small scale, provided that the interactions between the subunits are designed correctly. We show how self-assembly can be used to build biological and non-biological objects with applications in different scientific fields (Sec. 1.2.1). The design of the interaction is the key for self-assembly. We illustrate how very specific interactions between the subunits is achieved experimentally (Sec. 1.2.2). We then explain how the interactions between the individual particles and the properties of the self-assembly are related by looking at how the size (Sec. 1.2.3) and the shape (Sec. 1.2.4) of the aggregates are controlled.

1.2.1 Self-assembly is useful in many scientific fields

Proteins can be self-assembled into crystals, which serve for imaging purposes. Self-assembly also enables the building of new materials with applications in biology, photonics, or soft-matter.

An important use of self-assembly for proteins is crystallization: if a large number of identical proteins crystallize, the individual structure of the protein can be resolved with X-ray imaging [33]. Finding the correct conditions to achieve protein crystallization is still very challenging [34]. Proteins can also be assembled into an array of periodic cages where smaller proteins are trapped, to be imaged by electron microscopy [35].

Self-assembly can also be used to build new materials. Self-assembly of biological

molecules (proteins or DNA) is very broad, from advanced medical tools (such as drug delivery [36, 37], synthetic bioreactors [38] or virus trapping [39]), to the construction of synthetic organelles [40]. With tuning of the local interactions of the subunits, the self-assembled objects can also be reconfigured, enabling the design of nano-robots with motile parts and several configurations [41–43].

The self-assembly of non-biological individual subunits into a large organized array is also useful in other fields. It is used to build nanophotonic materials with nanoscale control over the local organization of metallic particles [44–46]. Soft particles can also be placed at the interface between two liquids and act as stabilizers of that interface [47]. The capillary interactions between those particles can then be tuned to achieve specific patterns of the particles at the interface [48].

Self-assembly is a tool for different scientific communities, but the challenges they face are similar. The interactions between the individual particles have to be tuned to achieve the expected assembly, which either has a specific size and shape, or is a periodic array that should be as large as possible, with precise organization of its constituents.

1.2.2 Experimental design of specific interactions

It is often possible to achieve precise control over the interactions between the individual constituents to obtain the desired self-assembly. It is necessary for the interactions between the subunits to be *specific*: two particles, or portions of the particle, interact with a large binding energy, but they do not interact with other particles, or portions of the particle. We explain on which physical mechanisms these interactions rely with three types of materials that can be used for self-assembly, among others: inorganic colloids, DNA, and proteins.

Colloidal spheres of micrometer sizes have long been used to study packing [49] and crystallization problems [50]. More recently, patchy [51] or anisotropic [52] colloids were designed experimentally to self-assemble into predefined structures such as crystals in 3D [53] and 2D [54, 55], chains [56] or small clusters of a few particles [57]. Colloids interact through sticky patches on their surface [58], hydrophobic interactions [59], electric or magnetic fields [60]. The specificity of the interaction between the colloids can, for instance, be achieved by choosing an optimized pattern of magnetic patches on the surface of a cylindrical colloid [61]. Colloids can also interact through *depletion* forces, where smaller particles will push together the colloids to decrease their excluded volumes [62]. Based on this principle, the use of non-spherical colloids [63] increases the design possibilities and enables the assembly of clusters of a finite number of particles (up to a few monomers) through lock-and-key interactions [64]: if two colloids have complementary shapes, they will bind, like jigsaw pieces. 3D printed colloids also appear as a preferential tool to achieve highly specific directional interactions and complex self-assembled structures [65].

In the last decades, the self-assembly of DNA-based materials has been developed. DNA nanotechnology often relies on the Watson and Crick pairing between complementary nucleotides of the DNA polymer. In one of those nanotechnologies, a long *scaffold* strand and many *stables* strands, when designed in a complementary manner, will fold into a 2 or 3D shape called *DNA origami* [66]. Because the size of individual DNA origami blocks is limited by the size of the scaffold stand [67] that is typically 7 kilo-base, self-assembly of multiple origamis is necessary to design large structures. Interestingly, the interactions between two DNA origamis can also rely on base pairing of the nucleotides, offering numerous design possibilities for highly specific interactions. Those DNA *bricks* can be designed individually, and therefore assemble into large and finite structures. Distinct DNA bricks were, for example, assembled into any arbitrary pattern and rendered shapes such as a smiley face in 2D [68], or planes in 3D [69]. DNA bricks can also be designed to interact through shape complementarity: a non-specific weak binding between the nucleic acid molecules is

combined with a shape recognition mechanism to enable specific interactions between the building blocks [41]. Finally, recent techniques were developed to cheaply and massively produce DNA origamis [70], making potential technological applications implementable at large scales.

Proteins are also used to design artificial assemblies, beyond the physiological examples we introduced in Sec. 1.1. The physical mechanisms of their interactions are now well understood: they interact because of polar or electrostatic interactions between the residues, because of shape complementarity, or because the residues at their surface are hydrophobic and need to be buried from the solvent [71–74]. Again, oligomers, fibers, or 2D and 3D arrays have been designed from physiological proteins [75]. *De novo* proteins can also be synthesized and assembled into protein complexes, where the interactions between the subunits are not constrained by the specificities of existing proteins, enriching even further the design space for both the interactions [76, 77].

In all those examples, interactions between the subunits are made specific because of shape complementarity, chemical specificity, or both. Self-assembled structures of precise shape or size are then obtained by carefully designing those interactions.

1.2.3 Controlling the size of the assembly

So far, however, extensive technological applications of self-assembly have been limited by the lack of size and shape control over the assembled structure [52, 78]. The technological bottleneck in the size control of self-assembly lies not only in the experimental design of individual particles but also in the physical principles that drive the assembly of the individual constituents. Here, we explain why it is difficult to achieve self-assembly of large but finite sizes, and present some of the methods used in DNA-origami or colloid experiments to circumvent these limitations. They rely on the individual design of each of the building blocks (addressable assembly), the self-closing of the assembled structure, or mechanical constraints (geometrical frustration). The distinction between those three mechanisms was introduced in [79]. We describe each of them.

As soon as there is an attractive interaction between subunits, there is no simple thermodynamic way to stop the growth of the assembly, unless the environment ran out of constituents. An aggregate of large size will always be more favorable than two aggregates of smaller size, because particles in the large aggregates realize more attractive interactions on average. The growth of the aggregate can be limited for kinetic reasons [80], but we focus on the self-assembly of constituents at equilibrium. Indeed, if there is a way to control the size of the aggregate thermodynamically, it also enables us to ensure that all the aggregates have the same size. With specific anisotropic interactions, a narrow size distribution of the aggregates can be observed for oligomers composed of a few subunits [57]. Reaching aggregate sizes that go beyond a few subunits while remaining finite and controlled is, however, one of the key challenges of self-assembly [79].

We now describe the possible mechanisms of equilibrium self-limited assembly. In *addressable* assemblies, each building block is individually designed so that it has a unique set of neighbors in the final assembly. For this reason, its position in the assembly is also fixed. This is illustrated in Figure 1.2a, where jigsaw pieces stand for the individual particles. Here, they are all distinguishable. This distinguishability allows for precise position of each particles in the aggregate. For instance, the red jigsaw piece should only interact with the gray and blue pieces. Because of this, it will always be positioned in the corner of the assembly. To design a large assembly using these techniques, a large number of distinct subunits with specific interactions should be designed individually, which can be costly. The difficulty is then to reach the highest possible *yield*, *i.e.* the ratio of aggregates that have been completely assembled. In the jigsaw example of Figure 1.2a, the second

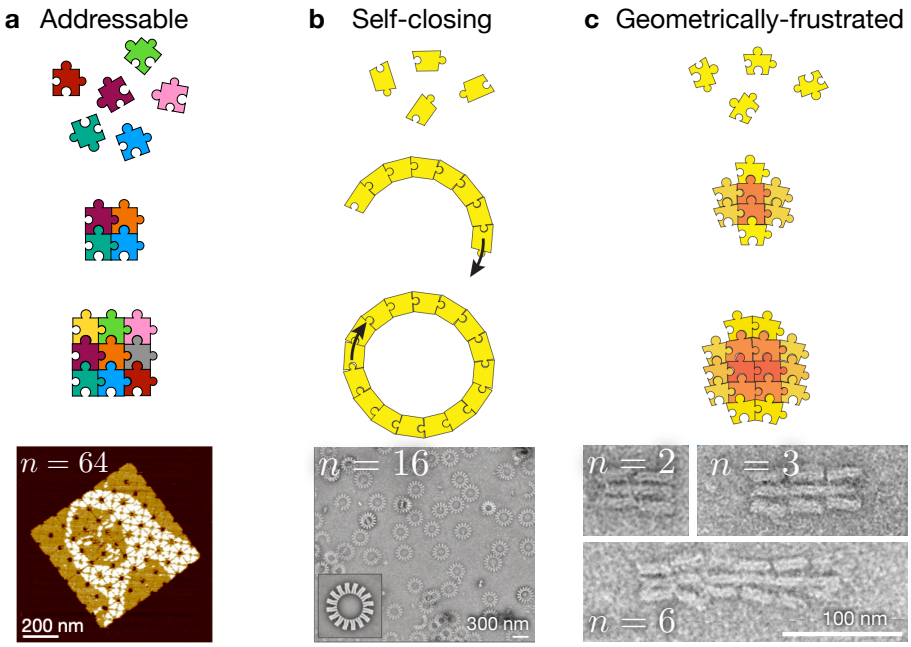

Figure 1.2: Self-limiting assembly at equilibrium relies on the specific design of the subunits interactions. Drawings are adapted from [79]. a) All the constituents are different (different colors) and have specific neighbors. b) The constituents have an intrinsic angle and form a ring of a fixed number of constituents. c) Constituents need to be deformed to be added to the assembly (the more orange, the more they are deformed). Images of experimental realization of each technique with DNA origami. a) AFM image of an array of DNA tiles representing the Mona-Lisa [81]. b) TEM micrograph of a self-closed ring of oligomers assembled from V-bricks [82]. c) TEM micrograph of a geometrically-frustrated polymer of deformable polybricks [83]. The number n of subunits per assembly is indicated in white on each image.

image is, for example, not fully assembled. The specificity of the interaction can then be quantified with information theory [84], and optimized to reach the highest possible yield [85, 86]. An example of a self-assembly of 64 DNA tiles, all patterned with a different portion of the Mona-Lisa is shown in Figure 1.2a. Because the yield decreases with the number of subunits [81], it is hard to build aggregates of very large size with this technique.

Self-limitation of the size is also achieved by *self-closing* of the assembly. Individual subunits can be geometrically designed such that the assembly will grow with an intrinsic angle. The final assembly is then a ring or a sphere. This is illustrated in Figure 1.2b, where all the jigsaw pieces have an angle and assemble into a ring. The radius of the self-assembly is then directly related to the angle in the shape of the individual particles. This is a possible mechanism for the self-assembly of viral capsids [87]. Assembling up to 30 identical building blocks in a predefined 3D-geometry has been demonstrated with DNA origamis, and a two-dimensional realization of the principle is illustrated in Figure 1.2b [82]. With this method, it is difficult to achieve assemblies that do not have spherical or cylindrical symmetry. This technique also requires fine tuning of the subunit geometry. Finally, if the subunits are slightly deformable, it can change the total number of subunits that can fit in a ring-like assembly such as the one shown in Figure 1.2b for instance, and its size. The minimum rigidity of the subunits increases with the desired size of the assembly, as demonstrated in a minimal model in [79].

As opposed to self-closing, an assembly has *open-boundaries* if some of the constituents have fewer neighbors than the others, *i.e.*, they are at the surface of the assembly. To achieve an assembly with an open boundary and limited size, the individual subunits can be designed such that they will be deformed when added to the assembly. Strain will then accumulate up to a limit where adding an extra monomer to the assembly is unfavored. The assembly is considered *frustrated* because if the particles are in their locally preferred configuration (undeformed), they cannot tile the plane [88]. This is illustrated in Figure1.2c: the jigsaw pieces have incompatible geometry, but if they are deformed (in orange), assemble. This mechanism was proposed as an explanation mechanism for the self-limited assembly of twisted tropocollagen filaments into collagen fibrils, where the fibril radius is controlled by the elastic properties of the individual filament [89]. Geometrical frustration has been used to control the length of the polymer of DNA deformable bricks [83], as illustrated in the experimental image in Figure 1.2c. A major requirement of this technique of self-limitation is the deformability of the individual particle. The mechanical properties of the individual particles need to be fine-tuned to ensure their self-assembly in the aggregate of the desired size, which can be challenging experimentally. Hagan and Grason [79] also suggest that there is a threshold to the size of an open-ended self-limited aggregate: if the frustration associated with a large aggregate is too important, the subunits will reorganize to avoid paying the energetic cost for frustration.

The main advances to achieve size control of self-assembled structures at equilibrium proposed so far rely on the particles being distinct, deformable, or on the self-closing of the assembly. This requires precisely controlling the shape, deformability, or interactions of the individual particles. In these methods, the shape of the aggregate cannot vary, and tuning the interactions can only control the size of the aggregates (the size of the self-closing ring in Figure 1.2b, or the size of the two dimensional isotropic bulk in 1.2c).

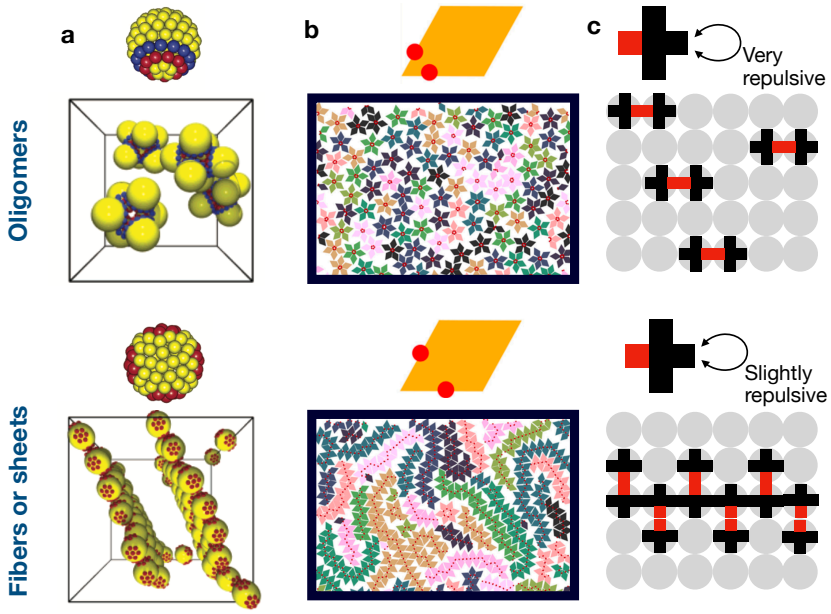

Figure 1.3: Small variations in the particle interactions lead to important changes for the aggregate. In all cases, the red patches are attractive. In a), the blue patches are also attractive. In c), the black-black and red-black interactions are repulsive, and the strength of the repulsive black-black interaction is tuned. The top images correspond to oligomers in 2D (b,c) or 3D (a), with only a few particles per cluster. The bottom images correspond to aggregates of infinite size but reduced dimensionality: the fibers (one-dimensional objects) in 2D for (b,c), and sheets (two-dimensional objects) in 3D for (a). Images taken from [51] (a), [90] (b) and reproduced from [91] (c).

1.2.4 Controlling the shape of the assembly

Beyond its size, it is also possible to control the shape of the aggregate. In Sec. 1.1, we introduced protein aggregates of very variable shapes, from small oligomers, to viral capsids, fibers, or crystals. Those aggregates vary not only in the number of their constituents, but also in their dimensionality (the oligomer is 0D, while the fiber is 1D). The position of the sticky regions of the particle, *i.e.* the *directionality* of the interaction, is responsible for controlling the shape of the aggregate. We introduce the concept of patchy particles and provide examples of the amount of possibilities it brings for designing the interaction of the individual particles. In particular, we show that minor changes in the interactions of the individual particles can lead to major changes in the shape of the aggregate.

A patchy particle has several sticky patches on its surface that will interact with the sticky patches of the neighboring particle through short range interactions. This was, for instance, introduced in [51] and is illustrated in Figure 1.3a: the yellow sphere has red and blue patches that bind and result in oligomers (top) or sheets (bottom), depending on the number and distribution of the patches on the surface of the particle. Those results are obtained with molecular dynamic simulations. In this study, Zhang and Glotzer also account for the self-assembly of fibers, sheets with different organization of the particle (into a square or a triangular lattice), and oligomers of 4, 6 or 12 particles. In this type of model, there are a lot of different qualitative ways to tune the interactions. One can vary the number of patches, their position (how dense and how isotropically distributed on the surface they are), the strength of the interactions, but also the type of interaction: patches of the same color can stick to themselves (homomeric interaction) or to other colors (heteromeric interaction) [52]. All those elements could be tuned independently, providing

a huge design space for the individual particles.

Changing only one feature of the interaction described above can already drastically change the result of the self-assembly. In [90], the position of two sticky patches on a rhombus of two-dimensional particles is systematically varied (see Figure 1.3b). The particles either remain as monomers (clusters have less than 3 particles), oligomers (such as the example of Figure 1.3b, top), or fibers (such as the example of Figure 1.3b, bottom). With Monte-Carlo simulation, Karner et al. systematically study what aggregates are observed for which patches positions. In both examples of Figure 1.3a and b, for instance, the sticky patches are not isotropically distributed on the surface of the particle. Then, depending on whether they are at the tip of the particle or in the middle, the assembly changes drastically, from a zero-dimensional oligomer to a one-dimensional fiber [51, 90].

The strength of the interaction can also be varied without changing the position of the sticky patches. A two-dimensional lattice gas model of cross-shape particles with directional and tunable interactions was introduced in [91], to study the assembly of anisotropic particles adsorbed on gold. In this model, a particle has four sides, one red and three black (see Figure 1.3c). There is three types of interactions: red-red, red-black, and black-black, which all have different interaction energies. A site of the lattice can be occupied by a particle or empty (the particles can be adsorbed on the substrate or not). They explore this parameter space and identify four types of aggregates: dimers, pentamers, fibers, and crystals. Again, we show in Figure 1.3 that a small change in the particle interactions (the repulsive black-black interaction is increased) leads to important changes in the equilibrium assembly, either dimers (top) or fibers (bottom).

Those studies provide specific examples of how one characteristic of the interactions (position or strength of the patches) influences the result of the self-assembly. To our knowledge, however, the relation between both is not understood systematically, and determining the equilibrium result of the self-assembly requires resorting to numerical simulation. Moreover, the particles in these examples are only designed such that all the favored interactions can be realized, without any competition between two pairs of attractive patches that cannot bind at the same time because of geometric constraints on the particles.

1.3 Frustration arises from incompatible interactions

Frustration arises in self-assembly when particles have incompatible favored interactions, or when they need to be deformed to take part in the assembly, as we mentioned in Sec. 1.2.3. We describe precisely the frustration of incompatible interactions in dense environments and lattice models (Sec. 1.3.1) and illustrate how only some of the concepts of frustration have been used so far as a design tool for self-assembly problems in dilute environments (Sec. **??**). We emphasize how frustration in the interaction can bring new insight into understanding the rules of protein self-assembly and can be used as a design principle for the individual interactions (Sec. **??**).

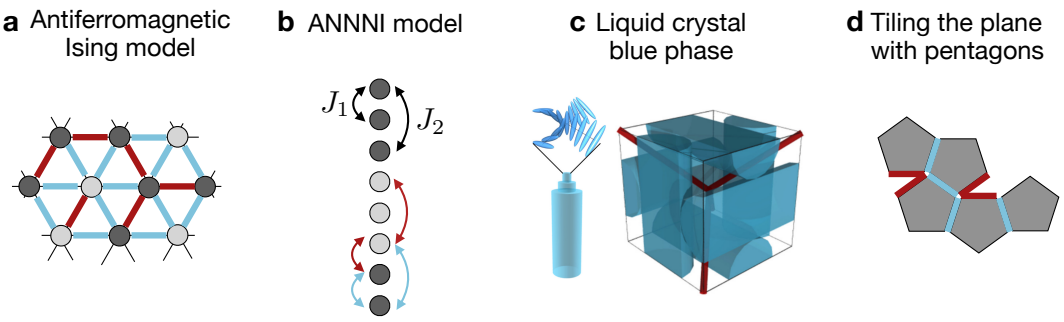

Figure 1.4: Because of frustration, there are some unfavored interactions between individual particles. Favored interactions are in blue, unfavored interactions or configurations in red. a) Lattice sites are up (light gray) or down (dark gray). There is at least one unfavored interactions (red) per triangle. b) Particles with ferromagnetic first-neighbor interactions and antiferromagnetic second-neighbor interactions can lead to patterns in the particle organization c) Rods organize into a double-twisted cylinder, and the cylinders form a lattice. Disclination lines (red) arise where the molecules alignment is unfavored. Drawing taken from [95] d) It is not possible to tile the plane with pentagons without deforming the particles, or leaving an unfavored space between them.

1.3.1 Frustration in a dense environment

We introduce the concept of geometrical frustration with two lattice models, and show how it can lead to non-trivial spatial organization of the particles.

The most simple example of frustrated interactions is that of an antiferromagnetic Ising model on a triangular lattice [92], illustrated in Figure 1.4a. The spin at position i is either up (dark grey, $s_i = 1$) or down (light grey, $s_i = -1$) and the coupling between two neighboring spins i and j is Js_is_j with $J > 0$, such that neighboring spins will try to anti-align. In the triangular lattice, it is not possible to satisfy the interactions of a group of three neighboring particles all at once, there is necessarily one of the two bonds that is unfavored. On Figure 1.4a, this results in necessarily having one unfavored (red) interaction per triangle, between pairs of particles that have the same orientation. In [93], Ronceray showed that the frustration in the antiferromagnetic Ising model is short-ranged: if we consider the state of the triangles of three spins, instead of the individual spins, as variables, then all the triangles are in the configuration of minimum energy (up-up-down or up-down-down). Frustration is not always that trivial, and in other cases, the extra energy of the ground state related to the impossibility to satisfy local constraints cannot be trivially removed by a change of the variables of the system [94].

Frustration can have effects on the way particles are organized, and lead to specific spatial patterns. The axial next-nearest neighbor Ising model (ANNNI) is an example of such effect [96]. It is an adaptation of the Ising model where the nearest neighbor has a ferromagnetic interaction $J_{i,i+1} = J_1 < 0$, and the next-nearest neighbor has an antiferromagnetic interaction $J_{i,i+2} = J_2 > 0$. This is shown in Figure 1.4c, where the nearest neighbors of the same orientation (same color) have favored interaction, but the next-nearest neighbor of the opposite orientation has favored interaction. The intrinsic competition between the couplings J_1 and J_2 can result in non-trivial spatial patterns at finite temperatures in two dimensions, and the relative strength of repulsive and attractive interactions can be varied to change the relative width of the domains of the same orientation [96].

Spatial patterns resulting from frustrated interactions have been observed experimen-

tally in the *blue phase* of liquid crystals [97, 98]. If the interaction energy between two chiral rods is minimal when they are slightly twisted with respect to one another, they organize into a double-twisted cylinder. The organization of the rods within one cylinder is shown in Figure 1.4c, left). However, the size of those cylinders is limited because it is impossible to extend this double twist organization to the whole space. For this reason, the most stable configuration is that of several cylinders intertwined. In this configuration, the relative orientations between most of the rods are those of minimal energy. There are, however, some disclination lines, *i.e.* regions where the interaction between two rods is unfavored. The cubic lattice organization of the cylinders is shown in Figure 1.4c, right, with the disclination lines in red. The period of the organization is of the order of the wavelength of blue light, hence the name blue phase, and it can be used for photonic applications.

When individual particles have competing interactions, they sometimes cannot organize without having some unfavored interactions, or defects. In those cases, minimization of the energy leads to periodic spatial patterns with length scales that are controlled by the interaction between individual particles.

2 - A model of lattice particles with arbitrary interactions

In this chapter, we propose a model of directional interactions, whose directionality and strength can be tuned between extreme cases: the particle can either be completely isotropic, or interact with its neighbor in very specific directions, and each interaction can be infinitely sticky, repulsive, or have arbitrary binding energies. Indeed, such model enables us to *explore* the *design space* of the particles interactions systematically: enumerate the different ways the particles can interact. Then, this systematic exploration enables to understand the relation between the interactions of the individual constituents, and the shape of the aggregate they form upon self-assembly. Here, we develop the tools to both design particles with very diverse interactions, and to study the shape of the aggregate they form at equilibrium. We also show how this model enables to consider frustrated interactions.

In Sec. 2.1, we introduce our model of directional lattice particles with arbitrary interactions. In Sec. 2.2, we explain how we use numerical simulation to determine the equilibrium configuration of the particles that self-assembled. In Sec. 2.3, we compare the different strategies to explore the parameter space of the interactions and show that this can be done by considering random interactions. In Sec. 2.4, we explain how we can characterize the geometry and the shape of the aggregates resulting from the numerical self-assembly of a particle with chosen interactions. Throughout the thesis, we mostly study the self-assembly of two-dimensional particles on a triangular lattice. In Sec. 2.5, we show how the model relies on concepts that are generic enough to be generalized towards other particle geometries.

2.1 Model of anisotropic particles

We propose a model which encompass most of the complexity of patchy particles introduced in Chapter 1 into the definition of an *interaction map*: for each possible relative orientations of neighboring particles, we define an interaction energy. Such an interaction map can then be used to model the self-assembly of patchy particles with variable number of patches, patches positions, or relative strength of the interaction. It also enables to take into account both the *homomeric* and *heteromeric* patches: a region of the particle can preferentially bind to the same region on the neighboring particle, or to another region. This generalization is made possible by considering lattice particles, that only have a finite number of relative orientations. We show that this model enables to observe a large variety of aggregates by varying quantities that are comparable (the interaction energies between two particles) as opposed to the patchy particles model where the design choices concerns quantities that are hard to compare, such as the number of patches on the particle and their number of colors. We introduce how the self-assembly of the particle depends on its interaction map that has tunable parameters in Sec. 2.1.1. In Sec. 2.1.2, we introduce state variables that enable to compute the energy of the system easily. We demonstrate in Sec. 2.1.3 that there are some invariances of the system that restrict the design space to 21 parameters. Finally, in Sec. 2.1.4, we illustrate the diversity of aggregates that are obtained from different interaction maps.

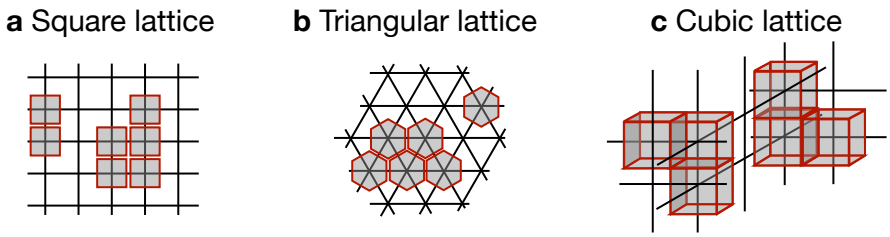

Figure 2.1: We define particles are Voronoï cells of the lattice. Lattice particles are for instance squares (a), hexagons (b) or cubes (c).

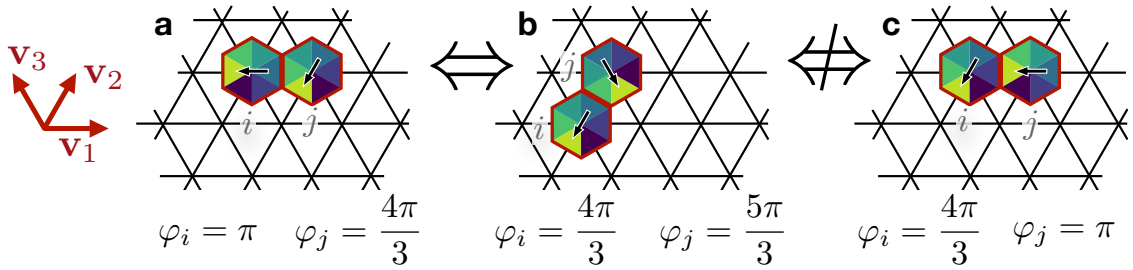

Figure 2.2: The interaction energy depends on the faces in contact. i and j are neighbor sites of the lattice. φ_i is the global orientation of particle at site i. In (a) and (b), the blue and light green faces of the particles are in contact: the interaction is the same. In (c), the light purple and yellow faces of the particle are in contact, the interaction is different.

2.1.1 Self-assembly of lattice particles with a chosen interaction map

As in the model of patchy rhombic dodecahedra introduced in Figure **??** of Chapter 1, we consider particle that occupy sites of a lattice. The maximum number of neighbors of a particle is then defined by the lattice: for instance, the maximum number of neighbors is four on a square lattice, six on a triangular lattice, and six on a cubic lattice, as shown in Figure 2.1. Then, the geometry of the particle is that of the Voronoï cells of a lattice. The number N_{faces} of faces of the particles is the maximum number of neighbors. In the following, we introduce notations that are valid for any lattice, but mostly show examples with the triangular lattice. We choose to focus our study on the self-assembly of hexagonal particles and not squares for instance, because as it was shown in Chapter 1, frustration due to incompatible interactions arises easily in triangular lattices: only three particles are required to make a loop.

The orientations of the individual particles is not the most convenient variable to describe their interactions, and the model we consider cannot be mapped to a Potts model [99]. For each particle, we can define an orientation φ, as in a Potts model, and the interaction depends on which faces of the two particles are in contact. This is exemplified in Figure 2.2. The pair of particles in the panel (a) and the panel (b) have the same interaction, because the same pair of faces are in contact (blue and light green in this case). However, the orientations of the particles are different, because the direction of the contact is different. We denote by $\mathbf{v}_{i\to j} = \mathbf{j} - \mathbf{i}$ the direction of the contact between particles on site i and j ($\mathbf{i}$ is the position of site i). The three possible directions of contact are shown in red in Figure 2.2. On the other hand, the pair of particles in panel (a) and panel (c) are not equivalent (blue and light green are in contact in (a), light purple and yellow are in contact in (b)), even if the global orientations are the same up to an inversion ($(\varphi_i, \varphi_j)_{\text{panel a}} = (\varphi_j, \varphi_i)_{\text{panel b}}$). In Potts models, the coupling energy between two particles is proportional to $\delta(\varphi_i - \varphi_j)$. In the Potts model, situations of panel (a) and (c) in Figure 2.2 would be associated with the same energy.

Instead of computing the system energy from the orientation of its particle, we describe it from the local configurations of the particles. We label the faces of the particles as $a, b, c, ..,$ (the colors in the previous discussion). We can define a mapping $\mathcal{M}$, between the orientation of two particles and the direction of their contact on the one hand $(\varphi_i, \varphi_j, \mathbf{v}_{i\to j})$, and the pair of faces that are in contact on the other hand (a, b). We only count the distinct pair of faces, such that $(a, b) = (b, a) = s$, where s is a reference integer for the situation where faces a and b of the particles are in contact.

$$\mathcal{M}(\varphi_i, \varphi_j, \mathbf{v}_{i\to j}) = (a, b) = s \tag{2.1}$$

For instance, both local configurations of the particles in Figure 2.2a and b are equivalent, such that $\mathcal{M}(\varphi_i = \pi, \varphi_j = 4\pi/3, \mathbf{v}_{i\to j} = \mathbf{v}_1) = (\text{blue} - \text{green}) = \mathcal{M}(\varphi_i = 4\pi/3, \varphi_j = 5\pi/3, \mathbf{v}_{i\to j} = \mathbf{v}_2)$. This formalism was initially introduced in [93] with the name *Local Energy Landscape*, to study the frustration in lattice spin systems. In that model, each local configuration of the lattice particles (labeled s), is assigned an energy ϵ_s.

Similarly, in our model, for each unique face pair (a, b), we define a binding energy that can be any arbitrary number. This energy should not directly depend on the value of the contact angle $\varphi_i - \varphi_j$, and should rather depend on the local properties of the surface (the patches of the colloid, or the amino-acid on the protein surface). To parameterize the interaction energies in our model, we define the interaction map $\{J_{ab}\}$, where J_{ab} is the energy associated with face a of a particle being in contact with face b of another particle.

For any pair of neighbor sites of the lattice $\langle i, j\rangle$, and direction of contact $\mathbf{v}_{i\to j}$, we can determine a_i and b_j, the faces of the particles at sites i and j involved in the contact.

The number of face pairs depends on the geometry of the particle. For two-dimensional particles, the number of distinct face pairs is simply

$$N_{\text{pairs}} = \frac{1}{2} N_{\text{faces}} \times (N_{\text{faces}} - 1) \tag{2.2}$$

N_{pairs} is not the square of the number of faces because a contact (a, b) is equivalent to a contact (b, a). Finally, we can represent the interaction map as a matrix: we order the entries of the matrix such that the column index refers to the face of the left particle, and the line index to the face of the right particle. In that convention, the matrix is symmetric. The interaction map of the hexagonal particle is shown in Figure 2.3b. Only the triangular inferior part of the matrix is independent, because (a, b) and (b, a) are equivalent. From eq. 2.2, it is straightforward that there are 21 independent coupling energies that can be defined for the hexagonal particles of 6 faces.

Along the thesis, we always represent the interaction map as a matrix. In the calculations, we either refer to its values by the label of the faces J_{ab} or by the label of the face pair, $J_s = J_{ab}$ if $s = (a, b)$ with the definition of eq. 2.1. Sometimes, we use the interaction vector $\mathbf{J}$ that is simply an array of all the interaction values in an arbitrary order.

2.1.2 The energy depends on the occurrence of the face pairs

Here, we describe the configuration of the system by counting the face pairs instead of counting the particles in each orientation, and show how this enables to compute the energy of the system.

We write the Hamiltonian of the system with the face pair variable. We first introduce the variable $\delta_{a_i b_j}$ that is 1 if the face a of the particle at site i and the face b of the particle at site j are in contact. With this notation, and given the interaction map introduced in 2.1.1, the Hamiltonian of the system reads

$$H = \sum_{\langle i,j\rangle} \delta_{a_i b_j} J_{ab} \tag{2.3}$$

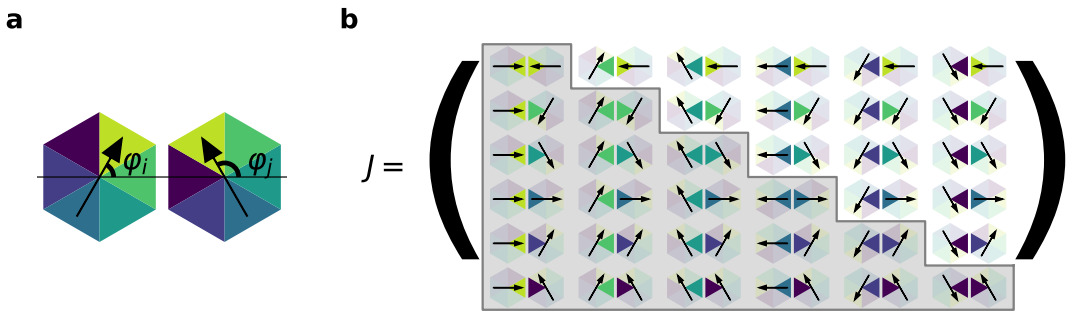

Figure 2.3: The interaction maps lists all the possible ways two particles can interact. It depends on the faces of the particles in contact. a) We determine the faces in contact from the relative orientations of two particles. b) We represent the interaction map as a symmetric matrix (only the entries colored in gray are chosen independently).

We then denote by N_{ab} the number of occurrence of each face pairs in the system. It verifies the following relation

$$N_{ab} = \sum_{\langle i,j \rangle} \delta_{a_i b_j} \tag{2.4}$$

With this simplified description of a system configuration, the Hamiltonian simply reads

$$H = \sum_{a \leq b} N_{ab} J_{ab} = \sum_{s} N_s J_s \tag{2.5}$$

In this formalism, the total energy of the system does not depend explicitly on the positions and the orientations of the particles. The occurrence of a face pair, N_{ab}, is a set of N_{pairs} number (21 for the hexagonal particle, as explained in Sec. 2.1.1). The measure of N_{ab} is then sufficient to determine the energy of a system.

2.1.3 Redundant surface interactions

In the interaction map shown in Figure 2.3, we only account for the interaction energy between the faces of two particles. We call *full* the sites of the lattice where there is a particle, and *empty* those where there are no particles. We adopt the following convention: if a site is empty, the face involved in the contacts with the neighbors is labeled 0. Then the *full-full interaction* is the interaction maps J_{ab} introduced before, and the *empty-full interactions*, J_{a0}, are the interaction energies between a face and an empty site. We also define an *empty-empty* interaction, J_{00}, which set the global level of the energies. In this section, we show that if the number of particles in the system is fixed, the empty-full and empty-empty interactions can always be set to zero, up to a shift of all the energies in the system.

In Sec. 2.1.2, we introduced eq. 2.5, which gives the total energy of the system for a given configuration, *i.e.* for a given set of $\{N_{ab}\}$, the count of the occurrences of a contact between face a and face b. We now expand this equation to distinguish between the full-full, empty-full and empty-empty interactions.

$$E = J_{00} N_{00} + \sum_{a=1}^{N_{\text{faces}}} J_{a0} N_{a0} + \sum_{a=1}^{N_{\text{faces}}} \sum_{b=1}^{a} J_{ab} N_{ab} \tag{2.6}$$

For the hexagonal particles, there are 7 conserved quantities in the system, that do not depend on the chosen interaction map, or of the configuration at a given time. Those quantities are the total number of bonds (1 conserved quantity, eq. 2.7) and the number of face a (the number of yellow face of the particle), because the number of particles is

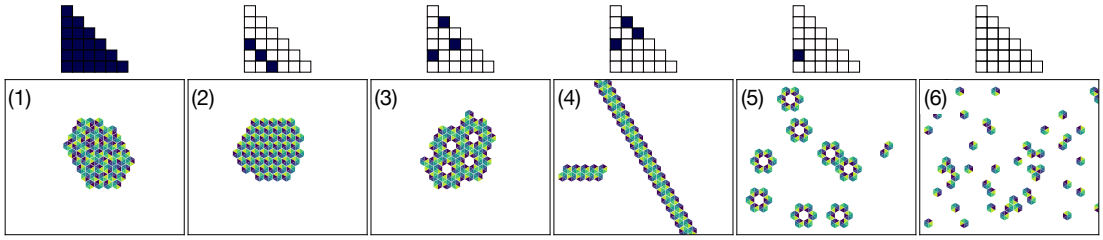

Figure 2.4: We obtain a large diversity of aggregates by changing the interaction map. We show the triangular inferior matrix for which the coordinates are defined in Figure 2.3 (top) and the corresponding equilibrium organization of the particles at temperature $kT = 1$ (bottom). The blue entries correspond to an interaction energy of $-10kT$ and the white to an interaction energy of $0kT$.

fixed (6 conserved quantities, one for each face, eq. 2.8). For a particle with arbitrary geometry, there are $1 + N_{\text{faces}}$ conserved quantities, and we derive the demonstration in the most generic situation. We denote by N_{bonds} the number of bonds (that can be either empty-empty, empty-full, and full-full) in a system. It is proportional to the size of the system. We also denote by $N_{\text{particles}}$, the number of particles (or occupied sites) in the system.

$$N_{00} + \sum_{a=1}^{N_{\text{faces}}} N_{a0} + \sum_{a=1}^{N_{\text{faces}}} \sum_{b=1}^{a} N_{ab} = N_{\text{bonds}} \tag{2.7}$$

$$\text{for } a \text{ in } [\![1, N_{\text{faces}}]\!],\ N_{a0} + \sum_{b \neq a} N_{ab} + 2N_{aa} = N_{\text{particles}} \tag{2.8}$$

The energy of the system is always defined up to a constant, and in eq. 2.9, we shift it by a constant that depends on the number of particles and the number of bonds in the system. We replace those values by their definition in the equations above (2.7 and 2.8). This shift is chosen such that the effective coupling for the empty-empty bonds and for the full-empty bonds are zero.

$$\begin{aligned} E' &= E - J_{00}N_{\text{bonds}} - \sum_{a=1}^{N_{\text{faces}}} (J_{a0} - J_{00})N_{\text{particles}} \\ &= \sum_{a=1}^{N_{\text{faces}}} \sum_{b=1}^{a} (J_{ab} - J_{00} - (J_{a0} - J_{00}) - (J_{b0} - J_{00}))N_{ab} \\ &= \sum_{a=1}^{N_{\text{faces}}} \sum_{b=1}^{a} (J_{ab} + J_{00} - J_{a0} - J_{b0})N_{ab} \\ &= \sum_{a=1}^{N_{\text{faces}}} \sum_{b=1}^{a} J'_{ab}N_{ab} \end{aligned} \tag{2.9}$$

We can therefore define a new interaction map J', with $J'_{ab} = J_{ab} + J_{00} - J_{a0} - J_{b0}$. The empty-empty and empty-full couplings of J' are zero, but it will result in the same equilibrium configuration of the particles as J.

In the rest of the thesis, we adopt the convention that the empty-full and empty-empty interactions (J_{00} and J_{a0}) are zero. Therefore, there are only 21 interactions to consider when exploring the design space of the individual particles.

2.1.4 Diversity of aggregates

We then look for the configuration of the system that minimizes the Hamiltonian of eq. 2.5. We give extensive details on how this is done in Sec. 2.2, but we start by showing

some examples of interaction maps and the corresponding equilibrium configuration of the system in Figure 2.3c. In this case, interaction maps are chosen in a simple way: some interactions have energy $J_0 = -10kT$ (the corresponding entries of the interaction map are colored in blue), while the other interactions have energy $0kT$ (colored in white). We see that this model enables to recover all the aggregates observed in the patchy particles models introduced in Sec. 1.2.4 of Chapter 1: fibers (4), oligomers (5), or monomers (6). The equivalent of a three-dimensional crystal is the aggregate shown in (2): it is a two-dimensional periodic aggregate in a two-dimensional space. Crystal can also have vacancies and form aggregate, which we call sponge (3). The aggregate in (1) corresponds to the situation where all faces of the particles are sticky, this is simply a lattice gas model of isotropic particles.

Because each interaction between faces is chosen independently, all the design possibilities of a patchy particles model (number of patches, number of different patches, strength, distribution of the patches, etc) are accounted for by the interaction map. The interaction map has numerous independent parameters (21), but those parameters are all binding energy values, and are comparable. The diversity of aggregates typically obtained with patchy particles were retrieved with our models, by changing the relative strength of the interactions. This is also an indication that we are not missing important phenomenology by considering lattice particles: we observe the same aggregates as in off-lattice molecular dynamics simulations of patchy particles [51].

2.2 Equilibrating with Monte-Carlo Metropolis-Hasting

For a given interaction map, we can change the relative levels of the interactions, which changes the shape of the aggregates, as shown in Figure 2.4. We can also change the global level of the interaction energies. Here, we show that we can determine the equilibrium of the system at $kT = 1$, such that changing the global level of the interaction amounts to changing the temperature of the system. In general, the Hamiltonian introduced in eq. 2.5 is not solvable analytically. We determine the equilibrium configuration of a system of identical particles associated with a given interaction map, by running Monte-Carlo simulated annealing that we coded in C++. In Sec. 2.2.1, we show that we can study the self-assembly by running simulations with a fixed-number of particles at low density. In Sec. 2.2.2, we choose elementary Monte-Carlo steps that ensure a fast equilibrating of the system. In Sec. 2.2.3, we show that progressively decreasing the temperature of the system towards a finite temperature enables both to equilibrate it and measure its fluctuations.

2.2.1 System definition

We fix the number of particles in the system, which we denote $N_{\text{particles}}$. The lattice has a size $N_{\text{sites}} = L_x \times L_y \times L_z$. For the two-dimensional case, $L_z = 1$. The number of bonds in the system, introduced in eq. 2.7, is simply the number of lattice sites, times the number of bonds per sites. In the triangular lattice, each particle makes 6 bonds, but a bond is shared between two sites, which makes $6/2 = 3$ bonds per particles. Then $N_{\text{bonds}} = 3N_{\text{sites}}$. We implement periodic boundary conditions. We choose a density of particles $N_{\text{particles}}/N_{\text{sites}}$ of the order of 0.1 to be in dilute condition. This enables to consider both dense and dilute organization of the particles: if the interaction map is such that the particles are attractive, the result of the equilibrating is a dense aggregate of the total number of particles (examples (1) of Figure 2.4). If the particles are not attractive, they will be distributed in random positions in the system, as the example (5) of Figure 2.4. The low density then ensures that the self-assembly driven by the interactions between the particles only.

2.2.2 Elementary steps

We explore the configurations of the system by flipping or displacing particles. At a given step t of the equilibrating, the configuration of the system is described by the positions $\{\mathbf{x}_i(t)\}$ and the orientations $\{\varphi_i(t)\}$ of all the particles labeled i. The energy of the system E_t, defined in eq. 2.5, is the sum of the interaction energy weighted by their occurrence in the system. We only change the configuration of one single particle per step t. At a given step, we draw with a uniform probability which particle will change configuration. With probability 0.5, the chosen particle changes orientation: $\varphi_i(t) \to \varphi'_i(t+1)$. $\varphi'_i(t+1)$ is chosen randomly among the orientations that are different from $\varphi_i(t)$. With probability 0.5, the particles changes position on the lattice $\mathbf{x}_i(t) \to \mathbf{x'}_i(t+1)$. The new position is chosen randomly among the empty sites of the lattice. Therefore, there are no correlations between $\mathbf{x}_i(t)$ and $\mathbf{x'}_i(t+1)$: the moves are delocalized. This ensures a faster exploration of the configuration space, but prevents us from studying kinetic considerations on the self-assembly. We do not perform collective moves of particles that belong to the same cluster [100].

Then we compute the new energy of the system E' and compare it to the old energy E. In practice, we only recompute the energy from the bonds of the particle i and its old and new neighbors. This elementary move is always accepted if $E' < E$. If $E' > E$, it is accepted with a probability $p = \exp(-(E'-E)/T_t)$, where T_t is the temperature at step t. This is the Metropolis-Hastings algorithm [101]. It ensures that the system converges to a steady state [102]: the possibility to make a step Monte-Carlo step towards a configuration of higher energy, while slowly decreasing the temperature, ensures that the optimization will not be trapped in a local minimum of the energies.

2.2.3 Annealing

Here, we show that decreasing the temperature of the system to $k_BT = 1$ enables to determine its equilibrium configuration at finite temperature, and to sample the fluctuations of the system. The level of the fluctuations is set by the strength of the attractive interactions in the interaction maps.

We start the simulation at high temperature and slowly decrease the temperature towards $kT = 1$, to ensure that the energy of the system at the end of the annealing corresponds to the minimization of the free-energy of the system. In the code, we implement the annealing such that there are N_T different values of temperatures, and at each temperature, there is N_{steps} Monte-Carlo steps. In Figure 2.5, we show the parallel evolution of the temperature and the energy per particle as a function of the number of Monte-Carlo steps. In this example, the number of temperatures is small ($N_T = 4$) for illustration purposes. We typically choose the number of temperatures to be large, to prevent the system from being trapped in a local minimum of the energy. The total number of annealing steps is simply $N_{\text{annealing}} = N_T \times N_{\text{steps}}$. At high temperature (left of Figure 2.5, image (a), the energy of the system is large, and the particles are in a gas phase. As the temperature decreases, the energy of the system decreases and the system reaches its equilibrium configuration, in which the particles are assembled (images (i) and (j)). We choose the annealing parameters (N_T, N_{steps}, and the initial temperature of the annealing T_0) such that increasing the annealing time does not decrease the energy of the system, which guarantees that the system is at equilibrium.

After the annealing, we keep performing Monte-Carlo steps at constant temperature $kT = 1$ to sample the fluctuations of the system. There is $N_{\text{statistics}}$ such extra steps. In Figure 2.5, this starts after point (j). In our simulation, the number of temperatures N_T, the number of steps per temperatures N_{steps}, the time of statistics collection $N_{\text{statistics}}$, and the initial and final temperatures are adjustable parameters. Typically, for one given

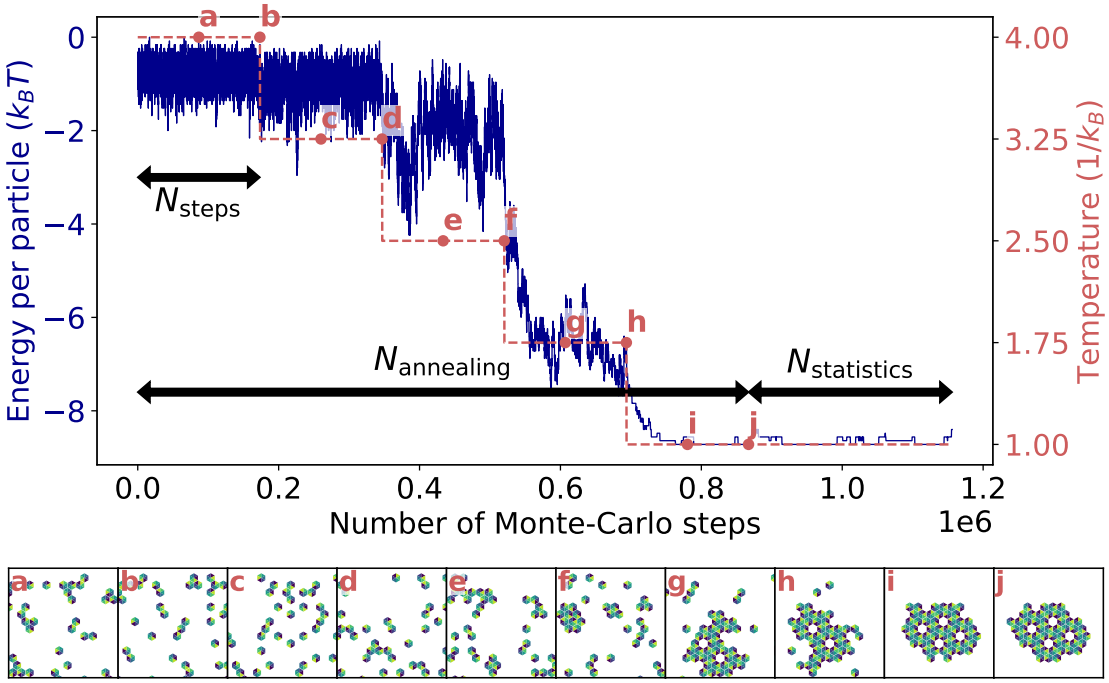

Figure 2.5: The system reaches its equilibrium configuration at $kT = 1$ after progressively decreasing the temperature. For the steps labelled with a letter, we show the configuration of the system at that step in the images below. As the temperature decreases, performing Monte-Carlo moves decrease the energy of the system. After running $N_{\text{annealing}}$ Monte-Carlo steps, (N_{steps} for each temperature) we compute statistics on the system during $N_{\text{statistics}}$ steps at temperature $kT = 1$ during which the energy slightly fluctuates.

interaction map, we also repeat the same annealing procedure several times, and we average every quantitative results over all the realizations of the system.

We have developed numerical simulation that enable to reach the equilibrium configuration of the system for a given interaction map, and sample the fluctuation of the system at equilibrium.

2.3 Exploring a 21-dimensional space

We introduced this model to identify the relation between the interactions of individual particles and the shape of aggregate they form. In Sec. 2.1, we showed that the interaction maps corresponds to 21 interaction energies, and that choosing the favored and unfavored interactions lead to very diverse aggregates. We also showed in Sec. 2.2 that changing the global levels of the interactions was equivalent to changing the temperature of the system. We can now vary independently each interaction energy between two faces, and determine how it changes the shape of the equilibrium aggregate. For the hexagonal particle, there are 21 independent pairs of faces, which corresponds to a design space that is too large to be explored exhaustively ($\mathbb{R}^{21}$). Here, we compare the different strategies to explore this very large design space. In Sec. 2.3.1, we show that the vertices of the particles can be assigned some sticky patches. This design gives indication that incompatible interactions reduces the size of the equilibrium aggregates, but it is too simplistic to provide a systematic understanding of the relation between the interactions and the shape of the aggregate. In Sec. 2.3.2, we show that choosing the interactions to be either attractive or repulsive, with only two energy levels, results in equilibrium aggregates that are not well-defined. In Sec. 2.3.3, we show that assigning groups of face pairs with the same energy values is

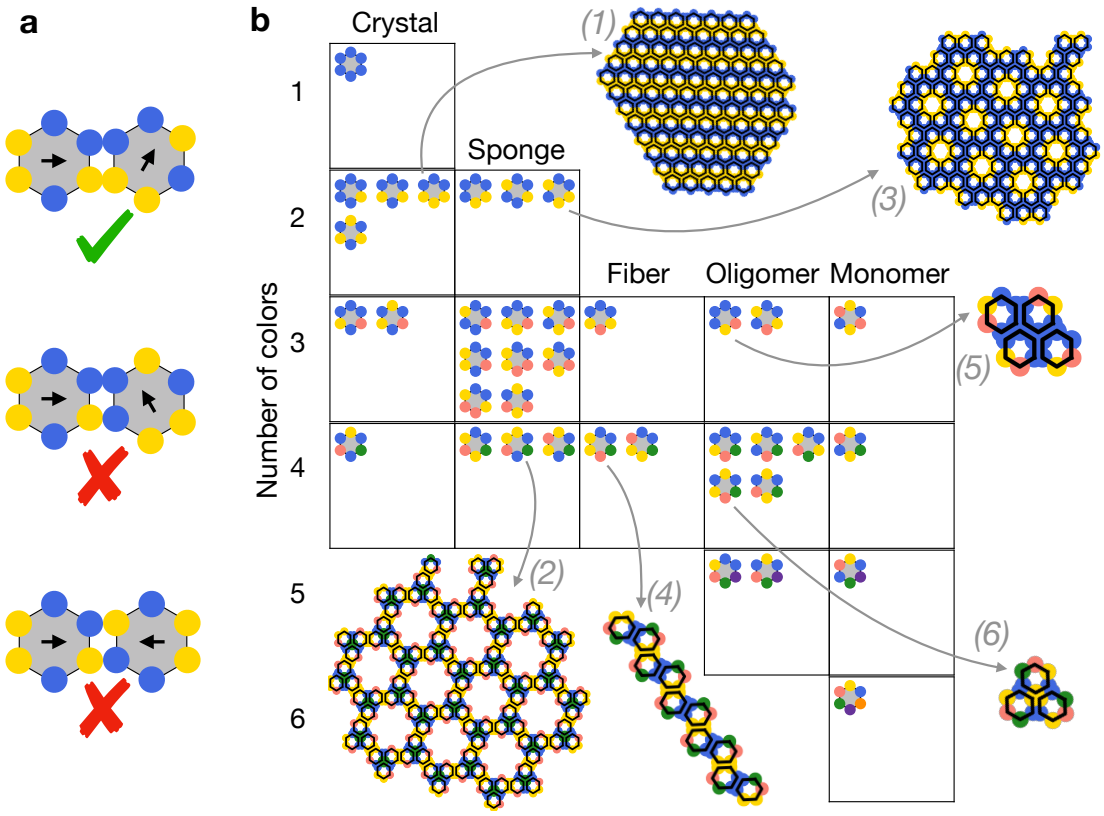

Figure 2.6: Aggregate sizes decrease when the number of incompatible interactions increases. a) Two faces of a particles interact if both vertices in contact are the same color. b) In the boxes, we enumerate all the different hexagons, classified by the number of colors of their vertices (vertical axis) and by the type of aggregate they form upon self-assembly. Equilibrium aggregates are shown for some particles. (1) is a crystal, (2) and (3) are sponge, with different size of vacancies, (4) is a fiber, and (5) and (6) are tetramers and trimers. Monomers are particles for which all interactions are repulsive.

a good design strategy for comparisons with experimental results. Finally, we suggest in Sec. 2.3.4 that choosing the interaction maps randomly enables to sample the design space systematically. In Sec. 2.3.5, we show that there is some redundancy in the 21-dimensional design space.

2.3.1 Particles with patchy vertices

Similarly to the model of patchy rhombus dodecahedron introduced in Fig. **??** of Chapter 1, we can color the vertices of the particle, such that two faces stick if all the colors of the vertices match. We exhaustively explore all the aggregates that can be created from those particles and show that the size of the aggregate decreases with the number of colors of its vertices.

We enumerate all the hexagonal particles with colored vertices. Each of the six vertices of the particle can be assigned one out of six colors, such that the vertices can be of identical colors, or all have a different color. An example of a particle with blue and yellow vertices is shown in Figure 2.6a. There are a lot of redundancies among those 6^6 particles. Indeed, if we label y and b the yellow and blue vertices, and enumerate in a cyclic order, it is straightforward that the particles *bbbbyy* (blue-blue-...) and *ybbbby* are identical: they are the same up to a cyclic permutation of the vertices. Similarly, *bbbbyy* and *yyyybb* are equivalent, up to an inversion of the colors. By removing those redundancies, we find only 38 non-equivalent hexagonal particles. All those particles are shown in Figure 2.6b. We order the particles by the number of different colors of the vertices: there is only 1 particle

with one color, 7 particles with two colors, etc.

Then, the interaction maps of those particles are defined as follows: if both pairs of vertices in contact match, the corresponding pair of faces is assigned an attractive interaction. If at least one of the two pairs of vertices in contact do not match, the corresponding pair of faces is assigned with repulsive interactions. This is illustrated in Figure 2.6a.

Then we determine the equilibrium state of numerous particles. Depending on the number of colors, they self-assemble into aggregates that are less and less dense. We show some examples of those aggregate in Figure 2.6b, labeled with a number. (1) is a crystal where the orientation of the particle alternate. It is of infinite size: if there were more particles in the system, the aggregate would be larger, and the size is only limited by the number of particles in the simulation. When the number of colors increases, more interactions are unfavored. As a consequence, particles can assemble into infinite aggregates with some vacancies, which we call sponge: no particle could be put in the vacancies without having an unfavored interaction with one of its neighbors. (2) and (3) are sponge with vacancy sizes of respectively one and seven particles. Increasing the number of colors also reduces the dimensionality of the aggregate: (4) is a fibrillar aggregate which can only grow in one direction. The number of incompatible interactions can be such that only aggregates of finite sizes can form, which we refer to as oligomers. (5) and (6) are oligomers of four or three particles. When the number of colors is maximal, the particle cannot bind to any other particle, and remains as a monomer.

In two dimension, the design of particles with colored vertices is useful to get a qualitative understanding of how the interaction between particles work, and the type of aggregate that can be achieved. We understand that the more complex the particles interactions are (the number of colors here), the more difficult it is to assemble into aggregates of infinite size. With this design, the parameter space is however very small (38 distinct particles). Indeed, because two neighboring faces of the particle share one vertex, all the entries of the matrix cannot be chosen independently. This reduces the amount of design possibilities for individual particles.

2.3.2 Two level interactions

Because the model of particles with colored vertices is constrained by the fact that two faces of the same particle share a vertex, we now consider particles for which each pair or face can be either sticky or repulsive, with no correlation between faces sharing a vertex. We show that this results in a broader diversity of aggregates, but that in some cases, we observe two distinct shapes of aggregates resulting from the self-assembly with particles of the same interaction map. Then, a systematic understanding of the relation between the interaction map and the shape of the aggregate is difficult.

With this design, the parameter space is immediately much larger than before: there are 2^{21} interaction maps ($2^{21} \approx 2 \times 10^6 \gg 6^6 \approx 4 \times 10^4$). As in the case of the patchy particles, there are many redundancies in this enumeration. After removing them (we give more details about how in Sec. 2.3.5), we find respectively 4, 26 and 134 non-equivalent interaction maps if there is one, two or three favored interactions.

Some examples of the matrices with three favored interactions are shown in Figure 2.7, where the favored (resp. repulsive) interaction have energy $-10kT$ (resp. $+10kT$) and the corresponding element in the interaction map is colored in blue (resp. red). The aggregates shown in (a) and (b) could not be obtained with any of the patchy particles introduced in Sec. 2.3.1. This confirms that the design space is increased by removing the constraint of the patchy vertices.

However, these ways of choosing the interaction map is not convenient, as we under-

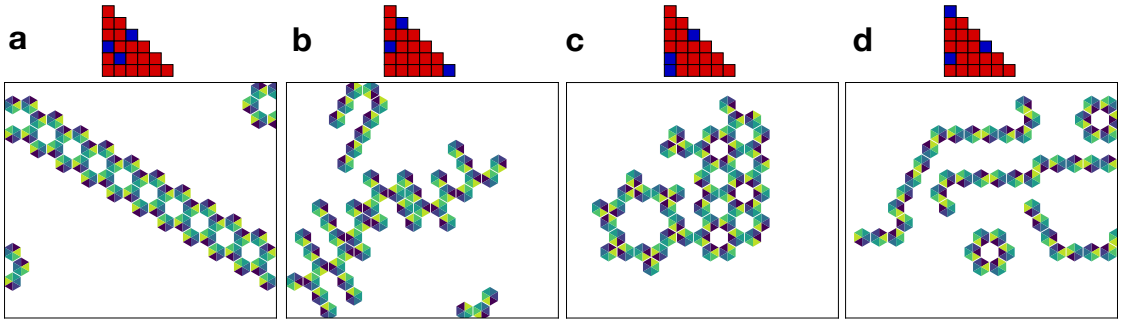

Figure 2.7: Examples of aggregates of particles with three favored interaction. Interaction maps are represented with the convention of Figure 2.3. Red interactions are unfavored ($+10kT$) and blue interactions are favored ($-10kT$). Aggregates in (a) and (b) could not have been obtained with particles of colored vertices. In (c) and (d), there is several possible organizations of the particles with same energy.

stand with the examples (c) and (d) of Figure 2.7. In both situations, we observe two different organizations of the particles within the same system: in (c), both sponges with small (on the right) and large (on the left) vacancies are observed. In (d), both fibers and hexamers are observed, and the curves on the fiber correspond to partially assembled hexamers. Because there are only two levels of energy for the interactions, two competing organization of the particles can have the same energy. However, this degeneracy would be solved if the energy levels were shifted of a very small quantity.

In the specific situation where there are few energy levels, it is not always possible to relate a given interaction map with a type of aggregate. This issue arises for three favored interaction, and could be more dramatic for higher number of favored interaction, where more competing organizations of the particles could be present at the same time in the system. We want to avoid those pathological situations, and for this reason we choose not to explore exhaustively the parameter space where the energy level can only take discrete values, and rather choose continuous scales for the interaction energies.

2.3.3 Physical particles

In Sec. 2.3.2, we showed that we could fix the energy levels (attractive and repulsive) and explore the design space by changing which interactions of the interaction maps are of those energy levels (change the positions of the blue and red squares in the matrices of Figure 2.14). Here, we consider an alternative exploration strategy: we fix which pair of faces should interact, and we tune their relative energy. We show that this type of model can be easily implemented in experimental studies.

We consider a colloidal particle with several lock-and-key mechanisms at its surface. For instance, a triangular-shaped key somewhere on its surface interacts with a triangular-shaped locks somewhere else, and a circular-shaped keys interacts with circular-circular shaped locks somewhere else. If those lock-and-key mechanisms are distributed in a regular way on the surface of the particle, we can predict its self-assembly with our model: each lock-and-key interaction involves a pair of faces, and the strength of the interaction is encoded as an interaction energy in our model.

In this example, the design space is much smaller, because there are only two attractive interactions. Then, we can exhaustively tune the strength of each interaction and determine the outcome of the self-assembly. This approach will be mostly tested in Chapter 5: we fix which faces of the particle should interact together, and we explore the parameter space corresponding to varying their relative energies.

2.3.4 Random interactions

Finally, to explore the space continuously in 21 dimension, one can use statistical approach and sample the parameter space, by drawing each of the 21 interactions randomly in a chosen distribution.

Here, we explain why we can draw the contact energies in a Gaussian distribution. We want to understand the interaction of particles with complex surfaces, such as proteins, where amino-acids of the surface can bind to other amino-acids with an arbitrary binding energy. The interaction between two faces is then the sum of the interaction between all individual pairs of amino-acids that are in contact. If there is a large number of terms in this sum (a large number of amino-acids in contact when two proteins are in contact), this sum can be approximated by a Gaussian variable according to the law of large numbers. Similar strategies to explore the parameter space has also been proposed in previous studies of self-assembly of distinct isotropic particles [85, 103].

In Chapter 3, we see that this sampling of the parameter space enables to identify a broad diversity of aggregates, and to relate specific characteristics of the interaction map (the local properties of the particle) to the type of aggregate they form (the macroscopic properties of the aggregate).

2.3.5 Permutation equivalence

In the example of the patchy hexagons, we saw that a 2D particle is defined up to a cyclic permutation of its vertices, or alternatively, a cyclic permutation of its faces. Here we generalize this principle. If we denote by a, b, c, d, e, f the six faces of the particles, a particle with faces a, b, c, d, e, f is equivalent to a particle with faces b, c, d, e, f, a (cyclic permutation). It is also equivalent to particle a, f, e, d, c, b (mirror permutation). These permutations correspond to the symmetries of the system. Here, we formally introduce how this permutation equivalence applies to the interaction maps. We write interaction maps as matrices and perform matrix operations. We denote as *equivalent*, two interactions matrices that will result in the same equilibrium configurations of the particles, after applying the same cyclic (or mirror) permutation to all the particles in the system.

Two interaction maps J and J' are equivalent up to a cyclic permutation of the faces of the particles if they verify

$$J' = P^k \cdot J \cdot P^{-k} \text{ with } P = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \text{ and } k \in [\![0, 6]\!] \tag{2.10}$$

Similarly, two interaction matrices J and J' are equivalent up to a mirror transformation of the particle if they verify

$$J' = M \cdot J \cdot M^{-1} \text{ with } M = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{2.11}$$

Therefore, for a given matrix J, there is a set $\mathcal{S}(J)$ of 12 equivalent permutations of J that give physically equivalent systems. We show an example of such 12 equivalent maps in Figure 2.8, together with the equilibrium configuration of the system. All those 12 aggregates are different. For instance, aggregates (a) and (g) both have the purple face

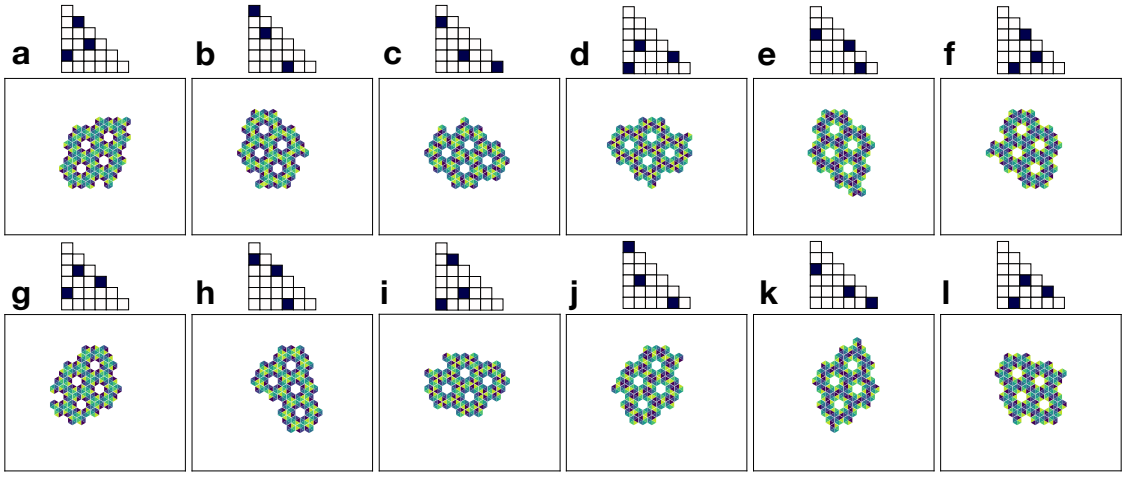

Figure 2.8: Equilibrium aggregates are equivalents up to a cyclic or mirror permutation of the faces of the particles. All the aggregates are similar (they form a sponge, and the position and number of the vacancies is the same in all aggregates), but the local organization of the particles (the color of the faces in contact) are all different.

of the particle facing the vacancies, but a closer look at the organization of the particles reveals that they are different.

In the rest of the thesis, we need to compare interaction maps. The permutation equivalence makes the element-wise comparison of interaction maps an ill-defined measure of their similarity. We will compare two interaction maps J and J' by exhaustively comparing all the elements of $\mathcal{S}(J)$ with J'.

2.4 Characterization of the aggregates at equilibrium

Our goal is to relate the interactions of individual particles with the shape of the aggregate upon self-assembly, by using the interaction maps framework. In Sec. 2.3, we showed how we could explore the parameter space of the interaction maps. Here, we show how to characterize the macroscopic properties of an aggregate resulting from self-assembly. In particular, we can look at the configuration of the system at a given time (Sec. 2.4.1), measure the averaged occurrence of each type of bonds (Sec. 2.4.2) and compute geometric descriptor to characterize the size and the shape of the aggregates (Sec. 2.4.3). All those descriptions are complementary methods we use to characterize the outcome of self-assembly in the rest of the thesis.

2.4.1 Visualization

We already made extensive use of the visualization of a system at equilibrium in the previous figures of this chapter. Although very qualitative, visualization of the system enables to get intuition of the characteristics of an aggregate, such as the fact that the particles are organized in a regular way within an aggregate, or the fact that the particles formed a dense aggregate or remained detached in a gas configuration. In practice, we use the positions $\{\mathbf{x}_i(t)\}$ and the orientations $\{\varphi_i(t)\}$ of all particles measured at a given time at the end of the simulated annealing described in Sec. 2.2.3. In the rest of the thesis, the representation of an aggregate will vary, from particles where the faces are colored differently, to particles with an arrow indicating its orientation. Because a visualization of the system configuration is measured at a given Monte-Carlo step, it does not allow estimating the fluctuation of the system.

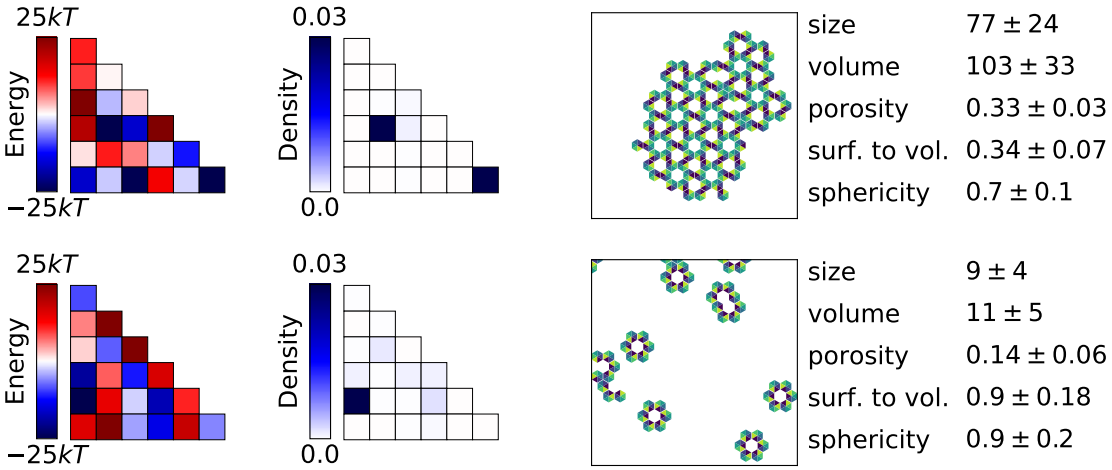

Figure 2.9: From a given interaction map, we characterize the equilibrium aggregate by its density map, image, and values of the geometric descriptors. (Top) the particles aggregate into a porous aggregate by only binding two pair of faces. (Bottom) the system forms oligomers of on average 9 particles. In both cases, the interactions that are frequent (blue in the density map) are also the one with low energy (blue in the interaction map).

2.4.2 Configuration averaging

In Sec. 2.1.2, we explained that the energy of the system is directly related to the number N_{ab} of occurrence of each pair of faces (a, b). Here, we show that the measure of N_{ab} in numerical simulations enables to measure the energy of the system, but also qualitatively evaluate the frustration of the interactions.

After the simulated annealing, we run N_{average} Monte-Carlo steps at finite temperature, as explained in Sec. 2.2.3. At each step t, we measure the number of each pair of faces in the system $N_{ab}(t)$. This number is then averaged for $N_{\text{statistics}}$ times step, and normalized by the total number of bonds in the system. We thus define the averaged density of a face pair $\langle c_{ab} \rangle$ as

$$\langle c_{ab} \rangle = \frac{\langle N_s(t) \rangle_{N_{\text{statistics}}}}{N_{\text{bonds}}} \tag{2.12}$$

There are as many values for the densities of full-full interactions as interaction energies. The averaged energy of the system over different configurations at finite temperature is then measured by the element wise product between $\langle N_{ab} \rangle$ (or $\langle c_{ab} \rangle$) and J_{ab}, as defined in eq. 2.5.

As we did for the interaction map, we represent the density map as a matrix, where the pair of faces corresponding to each of the entries are defined in Figure 2.3. We show examples of such interaction maps and density maps in Figure 2.9 for two system with random interaction maps. In this Figure, we verify that the pair of faces that are often observed in the system correspond to the one with low interaction energy: the blue entries in the density maps are also blue in the interaction map. This representation also shows that there are some favored interactions (blue in the interaction map) that are never observed (white in the density map). This suggests that the particles cannot satisfy all their favored interactions, because of geometric constraints, which is how we introduced frustration in Chapter 1. We introduce a quantitative measure of this frustration (the fact that there are favored but unrealized interactions in the system) in Chapter 3.

The density map is useful to measure the energy of the system and evaluate frustration, but it is not directly usable to characterize the shape and the size of the aggregates.

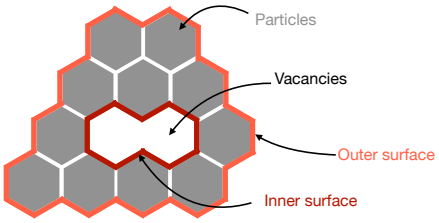

Figure 2.10: The size, sphericity and porosity (defined in the text) of an aggregate are measured by counting the number of outer surface bonds (pink) and the number of particles and vacancies in the aggregate. Here, the cluster has 11 particles, 1 vacancy of size 2, 10 inner surface bonds and 26 outer surface bonds.

2.4.3 Geometric properties of the cluster

Finally, we measure the geometric properties of the aggregates, to characterize its size, which can be infinite (all the particles in the system are within the same aggregate) or finite, its shape, which can be more or less spherical, and its porosity (the number of vacancies). We expect all those characteristics to be related to the interaction maps. Here, we explain which *geometric descriptors* we measure from the result of the self-assembly of a given interaction map.

In the simulations, we identify the particles in the same *cluster*, *i.e.* the particles that are all connected through full-full bonds. For each cluster, we count the number of particles n_p, the number of holes n_h, the number of inner empty-full bonds (or inner surface) n_{in} and the number of outer empty-full bonds (or outer surface) n_{out}. They are illustrated with an example in Figure 2.10. It is necessary to make the distinction between outer and inner surface to compute the aspect ratio of the aggregates.

From this, we can compute descriptors that are informative about the type of aggregates we observe. For a given cluster, we measure its

- size n_p,
- volume $v = n_p + n_h$,
- porosity $p = n_h/(n_p + n_h)$,
- surface to volume ratio $\frac{1}{2}n_{\text{out}}/(n_p + n_s)$,
- sphericity $n_{\text{out}}^{(\max)}(v)/n_{\text{out}}$.

The sphericity is the ratio between the number of surface of the aggregate, and a spherical aggregate of the same size. This measure is thus independent of the size of the cluster. An ideally spherical cluster thus has a sphericity equal to 1. A cluster of low sphericity is for example elongated, because it has more surface than if it were a sphere. This measure is more convenient than the surface to volume ratio, which scales as the size of the system. The average are done over all cluster within the all simulations, and weighted by the size of the cluster. Some of the descriptors can be seen in Figure 2.9. The top example emphasizes the importance of running several simulations: the averaged size of the cluster performed over 5 simulation is 77, while the total number of particles is 100. In the one image that we show, the cluster has size 100. The average size is then more informative than one picture.

We now have the tools to characterize the equilibrium aggregates, and to understand to which properties of the interaction map they are related.

2.5 Generalization beyond two-dimensional hexagonal particles

In the examples we showed so far, and throughout the thesis, we mostly consider hexagonal particles in two-dimensions. Here, we show that the framework of interaction and density maps describes situations that are more generic than the self-assembly of one type of two-dimensional particle that has the geometry of the Voronoï cell of a lattice. In Sec. 2.5.1, we show how three-dimensional examples increases the design space of the particle. In the simulation, the particles are the Voronoï cell of a lattice, which prevents the exploration of the self-assembly of other particles, as the triangle. In Sec. 2.5.2, we explain how to circumvent this limit and use the model also for triangular particles. Finally, we only studied the case of self-assembly of one particle type. In Sec. 2.5.3, we see how the framework of interaction map can also be generalized to more than one particle.

2.5.1 Generalization in 3D

The model we introduced for two-dimensional particles, like the square or the hexagon, is also valid for three-dimensional particles. We show how the dimension of the interaction maps, and therefore the design possibilities, are increased by considering 3D particles.

The orientation of a particle in 3D is not defined by one angle as before, but by three angles, because the particle can be rotated around a different axis. For a given 3D particle, we enumerate all the possible orientations. Each orientation is labelled with an integer φ. In 2D, we identified the energy associated with a contact between two particles by determining the faces (a, b) in contact. We introduced $(a, b) = \mathcal{M}(\varphi_i, \varphi_j, \mathbf{v}_{i\to j})$. In 3D, this is not sufficient, and a configuration of a pair of particles depends on the faces a and b that are in contact, but also on the relative orientations of the two faces. This is illustrated in Figure 2.11. In all four situations in the figure, the purple and green face of the particle are in contact. From situation (a) to situation (b) however, the particle i is rotated, around the axis of the contact direction $\mathbf{v}_3$, and the relative orientations of the green and purple faces changed. We assign a different energy to the interaction in both situations. From (a) to (c), both particles are rotated together. The orientations of each particle and the direction of contact change, but the way the two faces are in contact does not, and both (a) and (c) are associated with the same interaction energy. (c) and (d) are different for the same reason as (a) and (b). To take into account these differences, we introduce the angle of contact, ψ_{ij} which describes the relative orientations of two faces in contact. Then, the mapping to determine the configuration of the face pairs is contact is now

$$\mathcal{M}(\varphi_i, \varphi_j, \mathbf{v}_{i\to j}) = (a, b, \psi_{ij}) \tag{2.13}$$

For a regular polyhedron, there are $N_{\text{faces}} \times (N_{\text{faces}} + 1)/2$ couples of faces that can be in contact. We also define n_r, the number of relative orientations of two particles in contact, such that they still occupy the Voronoï cell of the lattice. The total number of ways two neighbor particles can be in contact is now

$$N_{\text{pairs}} = N_{\text{faces}} \times \frac{N_{\text{faces}} + 1}{2} \times n_r \tag{2.14}$$

In table 2.1, we reference for each lattice, the type of particle, the number of faces, the number of orientations, the number of relative rotation between two faces, and the total number of non-redundant pairs. The images corresponding to each geometry are also

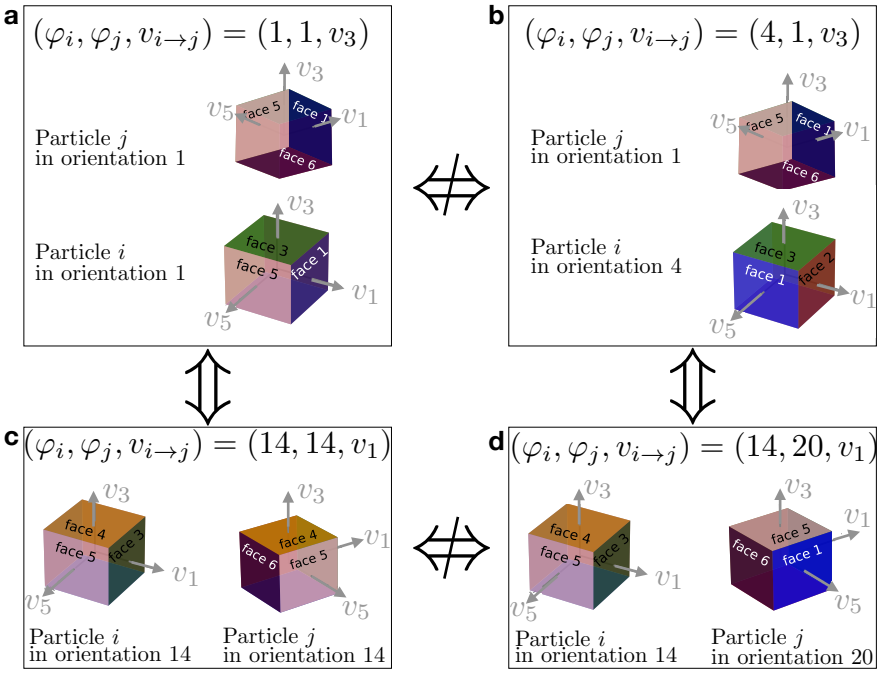

Figure 2.11: In 3D, the relative orientations of the faces in contact determines the energy of the pair configuration. ((a) and (c)) and ((b) and (d)) are equivalent configurations because the same faces are in contact with the same relative orientation. (a) and (b) are different because the same faces are in contact, but not in the same relative orientation.

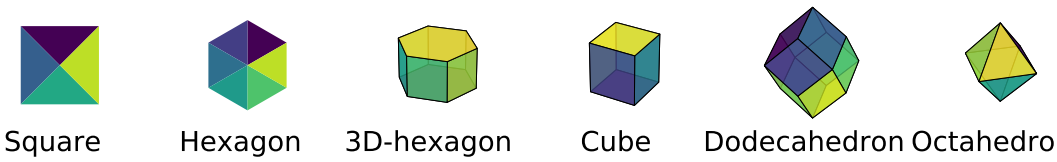

Figure 2.12: Lattice particles are Voronoï cells of a lattice. The square and the hexagon are two-dimensional and interact through their edges (that we call face throughout the thesis because we mostly consider 2D systems). The 3D-hexagon, cube, dodecahedron, and octahedron interact through their faces (each displayed a different color).

shown in Figure 2.12. This illustrates how large the dimension of the interaction map can get for three-dimensional particles. For instance, for the rhombic dodecahedron, there are 144 interaction energies to define.

It is not surprising that self-assembly of three-dimensional objects such as proteins can result in extremely broad diversity of aggregates, as we emphasized in the introduction of the thesis. Hexagonal particles provided rich enough behavior, and we did not explore the self-assembly of three-dimensional particles, but the tools to study such problem are implemented for further studies. In particular, dimensionality reduction of the aggregates could be investigated in more details. In 2D simulations, dimensionality reduction corresponds to the equilibrium aggregate being a fiber. 3D simulations enable to distinguish between a sheet (2D aggregate), and a crystal (3D aggregate), but also shapes like tubes or spherical shells. Sheet, tubes, or shells are examples of dimensionality reduction that are richer than the sole fiber example available in 2D simulations.

Lattice	Particle	N_{faces}	$N_{\text{orientations}}$	n_r	N_{pairs}
Square	Square	4	4	1	10
Triangular	Hexagon	6	6	1	21
Hexagonal	3D-hexagon	6+2	12	2 or 6	60
Cubic	Cube	6	24	4	84
FCC	Dodecahedron	12	48	2	144
BCC	Octahedron	8	24	1	36

Table 2.1: Increasing the number of faces of the particles increases the number of pair configurations, and the design space. BCC is Body Centered Cubic and FCC is Face Centered Cubic

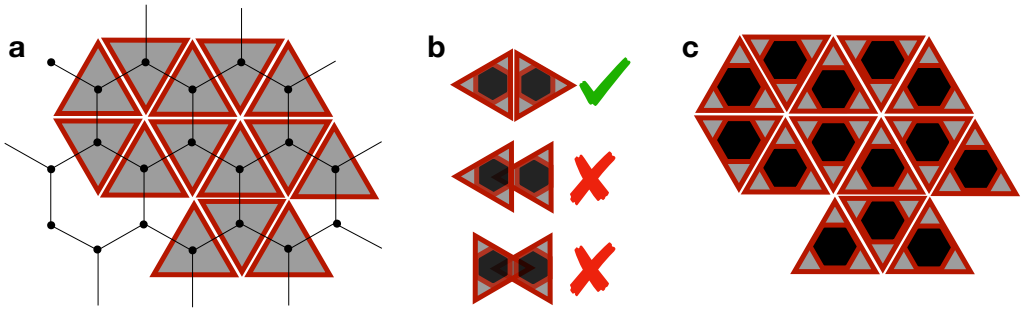

Figure 2.13: We use of implicit hexagons to simulate the self-assembly of lattice triangles. a) Closed packing of regular triangle on a honey-comb set, which is not a Bravais lattice. b) Implicit hexagons to model triangle interaction: some interactions between the hexagon correspond to the interaction between two faces of the triangles, and they are assigned with a physical energy. Other interactions between the hexagon correspond to interaction between corner of the triangle, and they are assigned infinite energy. c) Distribution of the implicit hexagons on the triangular lattice.

2.5.2 Self-assembly of triangles

Our numerical simulations could be used to model the self-assembly of colloidal particles on a 2D-substrate. These particles can be 3D-printed, and have arbitrary shape. The most simple polygon in 2D is the triangle, and we could study the self-assembly of triangular colloids. However, the triangle is not the Voronoï cell of a Bravais lattice, and it is not possible to implement this geometry in our code. Here, we show how to circumvent this limitation.

The triangle is the regular two-dimensional polygon with the least faces, and it is the Voronoï cell of the honeycomb set, that is not a Bravais lattice. Examples of honeycomb set and triangular particles are shown in Figure 2.13a. To study the self-assembly of triangles, we use the hexagonal particles as an implicit representation of the triangle. The hexagons have some constraints: some interactions are infinitely repulsive, because they correspond to interactions involving the corner of the explicit triangle. This is illustrated in Figure 2.13b, the explicit triangles are in gray, and the implicit hexagon in black. Some interactions between the hexagons are forbidden. The result of the packing of the triangle is shown in Figure 2.13c. A dense aggregate of triangular particles corresponds to a porous aggregate of the implicit aggregate of hexagonal particles.

We use this approximate representation of the problem to study the self-assembly of

triangular particles at the end of Chapter 3. We took advantage of the fact that only the lattice and the interactions between particles are implemented in the simulation, not the geometry of the particle. This enabled us to implicitly implement the self-assembly of triangular particle. It is not clear whether this trick could be used to study other particle geometries, like rhombi or pentagons. However, the possibility to assign infinite energy value to some interaction enables to take into account geometric specificities of the particle.

2.5.3 Several types of particles

We implemented a model of one particle with directional interactions. This model enables to explore the diversity of aggregate that can arise from directional interactions. Here, we show that we can also study the self-assembly of several particle types, each having different interaction maps, and show that the different types of particles can self-assemble in the same aggregate or not.

In the case of two particles, we need to define three interaction maps: that of particle A with itself, that of particle B with particle A, and that of particle B with itself. We show examples of the matrices and the simulation results in Figure 2.14. Particles A are represented with the color code used before, while particles B are represented in pink and orange colors. One should notice that the AA and BB interaction maps are symmetric as before, but the AB map is not: the interaction of face a of particle A with face b of particle B is not equivalent to that of face b of particle A with face a of particle B. For two hexagonal particles, there are $21 + 36 + 21$ interactions energies to define.

In Figure 2.14, we show two situations of self-assembly of 100 green and 100 pink particles. On top, the green and pink particles assemble in the same aggregates. On the bottom, the pink particles form fibers, and the green forms bulk. This can be observed in the green-pink density map: a contact between the green and the pink particle almost never occurs, and the density map is white. This suggests that directional interactions between more than one particle gives rise to very diverse aggregate shape, as before, but also a phenomenology associated to the ability of the two particles to mix or not.

This could also be generalized to a larger number of different particles. However, the total number of energy values in the interaction maps grows as the square of the number of particles type. This type of simulation would not be well-adapted to study the self-assembly of particles that are all different, for instance, because there are too many parameters to define.

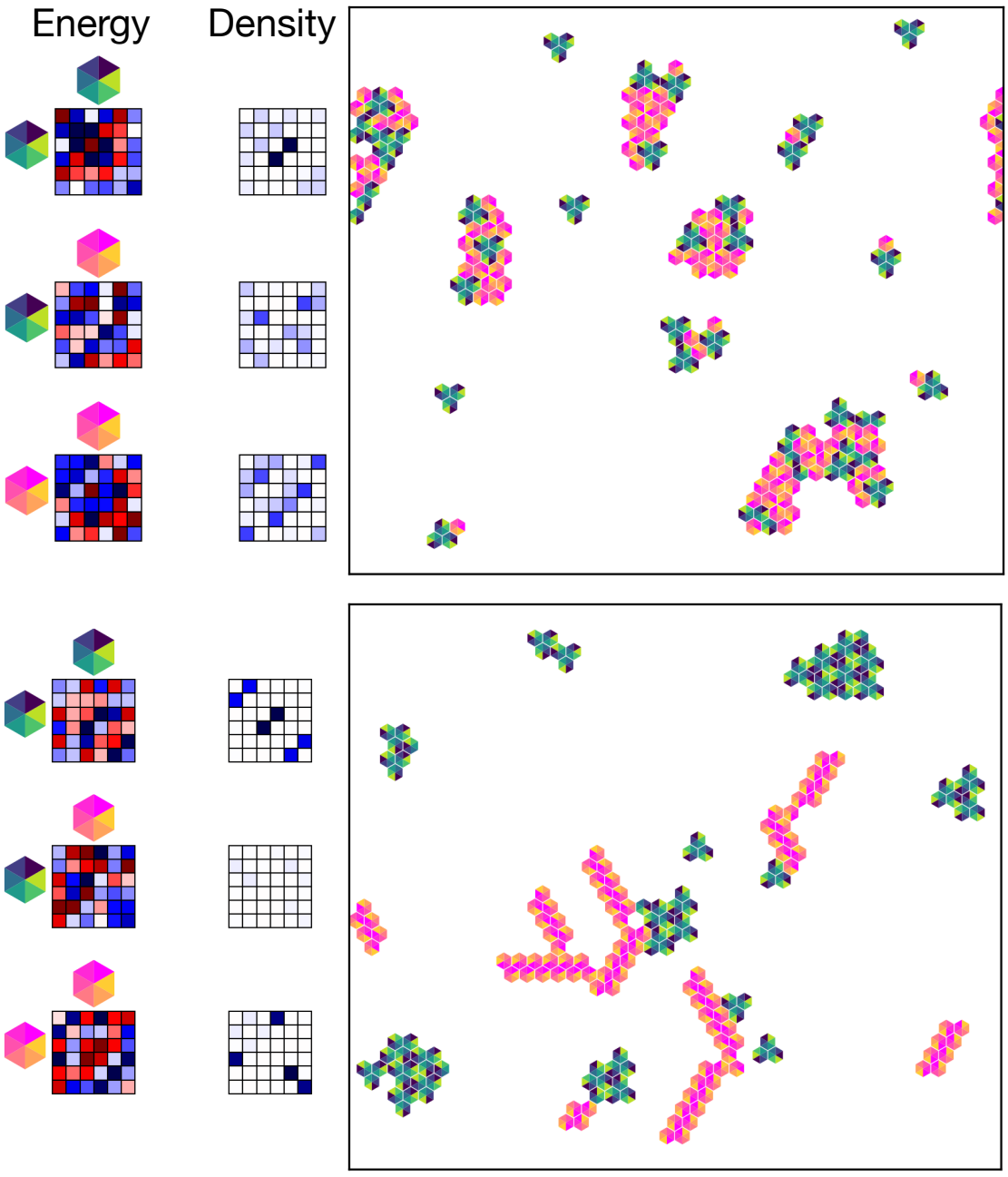

Figure 2.14: Two types of particles can either mix in the same aggregates (top) or form separated aggregates (bottom). The interaction maps and the density maps are shown, with the same conventions as in Figure 2.9 (we do not represent the color bar for clarity of the representation). The particles on the top and left of the matrix indicates which pair of particles the matrix corresponds to. We show the corresponding equilibrium configuration of the system.

3 - Anisotropic particles with random interactions form aggregates of reduced dimensionality because of frustration

We introduced a model of lattice particles with directional and arbitrary interactions. In this chapter, we understand how the local properties of the particle interactions are related to the type of aggregates they form upon self-assembly. We systematically explore the design space of those particles by studying a large number of different equilibrium configuration, resulting from the self-assembly of particles with random interactions. We show that most of the aggregates resulting from the self-assembly of anisotropic particles are frustrated, because there is no periodic organizations of the particles that tile the plane with low energy. Then, the aggregates *reduce their dimensionality* by forming fibers or micelles. We hypothesize that this reduced dimensionality is related to frustration.

In Sec. 3.1 we sample a large number of particles with a chosen affinity and anisotropy. They self-assemble in very diverse aggregates. We see how the affinity and anisotropy of the particle influence its self-assembly, on average. In Sec. 3.2, we provide an individual classification of all the aggregates and confirm that anisotropic particles form aggregates of reduced dimensionality. In Sec. 3.3, we show that the ability of the particles to assemble in periodic motifs of low energy, that determines the type of aggregates it forms. Finally, in Sec. 3.4, we propose preliminary interpretation of how these results extend when there are two types of particles in the system, and show how they could be tested experimentally.

3.1 Affinity and anisotropy as parameters

Here, we draw the interaction map of the particle randomly, while choosing the affinity and anisotropy of the particle and show that it leads to aggregates of very diverse shape, and frustrated aggregates. In Sec. 3.1.1, we explain how we explore the parameter space by drawing interaction maps in a Gaussian distribution. We then study the averaged properties of an aggregate of particles with similar affinity and anisotropy. In particular, we see that increasing particle anisotropy decreases the energy of the system (Sec. 3.1.2), results in aggregates of non-trivial shapes (3.1.3), but that those aggregates are not more frustrated than the aggregates of particles of low anisotropy (Sec. 3.1.4).

3.1.1 Random particles form aggregates of diverse shape

We recall that the particle is fully described by its interaction map, which, for the hexagonal particle, has 21 independent parameters. We show why the average and standard deviation of those parameters can be interpreted as the affinity and anisotropy of the particles. We detail how we choose those parameters to sample the design space of the particle. We show that the aggregates obtained by those design choices have very diverse shapes.

We draw those 21 parameters independently of a Gaussian distribution of average μ and standard deviation σ. μ determines the global *affinity* of the particle: if the interaction energies are on average negative, the particle will tend to form dense aggregates. If it is on average positive, the particle will be mostly repulsive. σ then describes the *anisotropy* of the particle. If σ is small, all the interaction energies are similar, and the particle is isotropic, it does not have preferred directions of binding. On the other hand, if σ is large, some interactions are repulsive, some are attractive.

We explore the space by varying μ and σ. We choose $\mu \in \{-4, -2, 0, 2, 4\}kT$ and $\sigma \in$

$\{0.1, 1, 3, 5, 7, 9, 11, 15\}kT$. For each condition, we draw $N_{\text{data}} = 200$ different interaction maps. This corresponds to a total of 90000 different systems. We refer to the distribution where the interactions are drawn as $\mathcal{N}(\mu, \sigma)$. For each of them, the equilibrium states and descriptors are computed according to Monte-Carlo simulation, as presented in Chapter 2. We consider hexagonal particles on a two-dimensional lattice with $L_x = L_y = 30$ ($L_z = 1$), $N_{\text{particles}} = 100$. For a given interaction map J, the annealing is performed between the temperatures $T_0 = \max(|J|)$ and $T_f = 1kT$. This choice of T_0 ensures that the largest energy barriers are being sampled at the beginning of the annealing. We choose $N_T = 100$ temperatures, and $N_{\text{steps}} = 100 \times N_{\text{sites}}$. The number of annealing Monte-Carlo steps is therefore $N_{\text{annealing}} = 9 \times 10^6$. The density of each structure is then averaged over $N_{\text{statistics}} = 1000 \times N_{\text{sites}}$ steps. For each interaction map, we perform $N_{\text{systems}} = 5$ different annealing and measurements.

For each value of affinity and anisotropy, we show the result of the simulation for one system in Figure 3.1. We also show the interaction map and density map. We can first study the low anisotropy limit in the lowest part of this figure. If the affinity is negative, the particles are isotopically attractive: on the bottom right, the interaction maps are completely blue, and the particles form dense aggregates and have random orientations. If the affinity is positive, the particles are repulsive: the interaction maps are red, and the particles are not in contact. If the affinity is zero, it is a gas of non-interacting particles. Upon increasing the anisotropy of the particles, the aggregates have less trivial shapes. We retrieve the typical aggregates introduced in the Chapter 2, such as the dense crystal (for instance at $(\mu = -4, \sigma = 5)$), the sponge $((\mu =, \sigma = 11))$, or the fibers $((\mu = 0, \sigma = 5))$, or the aggregates of small size $((\mu = 2, \sigma = 9))$. We also observe aggregates that were not observed with the two level interaction maps, or the particles with colored vertices introduced in the Sec. 2.3 of Chapter 2, such as branching fibers $((0, 7)$ or $(4, 0))$, dense aggregates of intertwined fibers $((-2, 7))$, or small aggregates with no clear motif $((0, 15)$, $(4, 11)$, $(-4, 9))$. Those aggregate images are not trivially related to the interaction maps. This is also understood by looking at the density map: some favored face pairs are present with high density in the system, but for almost all the examples of Figure 3.1, there are some favored interactions in the interaction map (blue entry in the left matrix) that are not observed (white in the right matrix).

From these examples, we conclude that the aggregates generated from particles with random interaction are very diverse, as were the protein aggregated presented in the introduction of the thesis. The exploration of the design space of the particles by choosing a large number of random interaction maps thus appear as a reasonable method to draw systematic conclusion on the relation between the interaction between the particles, and the shape of the aggregate.

3.1.2 The energy of a system is governed by the values of the best interactions

Here, we show that the aggregates are frustrated, because the system is not governed by the interaction of lowest energy only. The energy of a particle depends on the interaction it has with its neighbors, and is easily measured in the numerical simulation. The affinity of the particles sets the global level of those interactions, and should influence the energy of the system, but the influence of the anisotropy is not clear. Here, we determine this influence and show that the energy of the system is governed by some of the interactions of lowest energies, but not only the lowest one.

We recall that the energy of a particle is the sum of the interaction energies, weighted

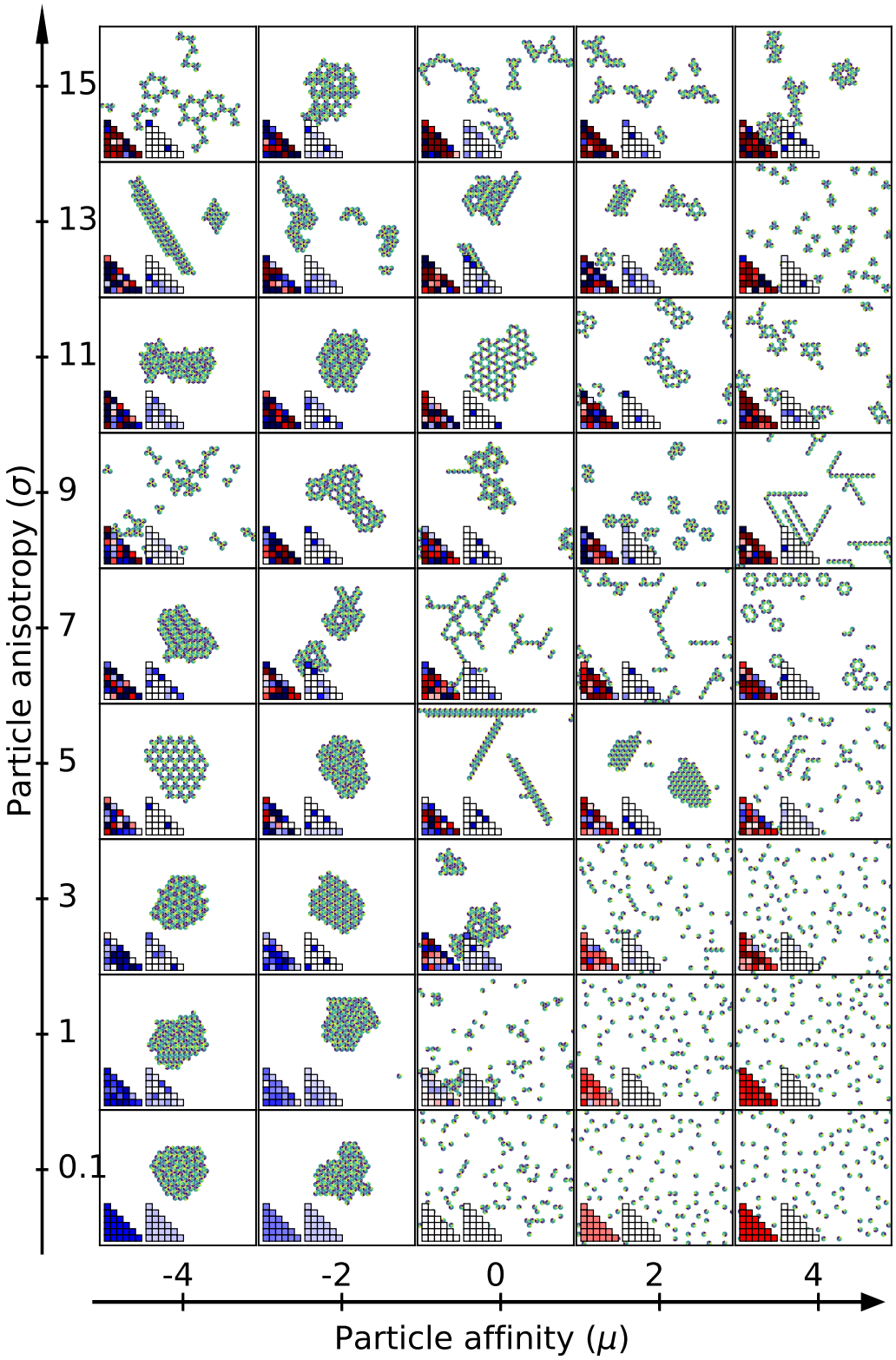

Figure 3.1: Particles with random interaction self-assemble in aggregates of diverse shape. For a given value of affinity μ and anisotropy σ, we show an image of the system, and the interaction (bottom left) and density maps (bottom middle). For the interaction map, red is positive and blue is negative (the darkest colors correspond to 15 and $-15kT$). For the density map, white correspond to $c = 0$ and the darkest blue corresponds to $c = 0.05$.

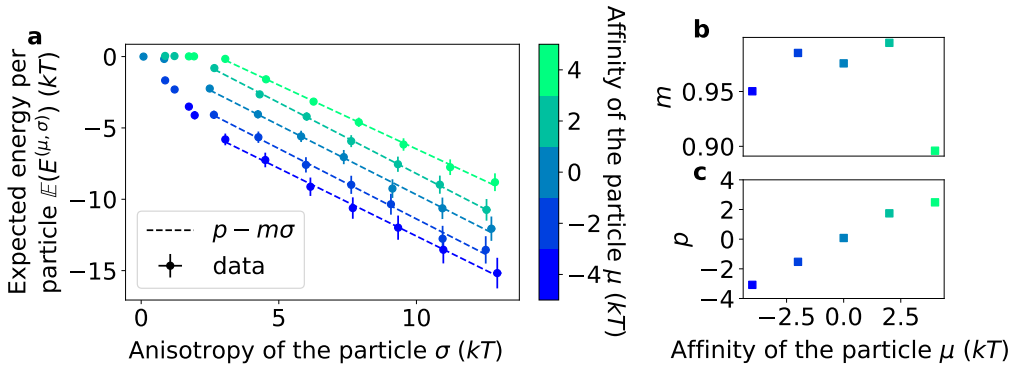

Figure 3.2: The energy per particle decreases linearly with the particle affinity and anisotropy, on average. a) We fit the linear dependence of the expected energy per particle with σ. We show that the slope does not depend on σ (b) and that the origin scales linearly like μ (c), which verifies equation 3.2. The vertical (resp. horizontal) error bars correspond to the standard error of the measured energy (resp. anisotropy) of the data (200 data per points).

by how often they are observed at equilibrium.

$$E = \frac{1}{N_{\text{particles}}} \sum_s N_s J_s \tag{3.1}$$

We expect this energy to be negative: if the particles are repulsive (positive mean of the interaction map), the system will remain in a gas configuration. Then, all the bonds in the system are empty-full or empty-empty, which were assigned energy zero.

We also expect the energy of the system to depend both on the mean and the standard deviation of the interaction map. To understand this dependence, let us consider two extreme situations: the situation where the occurrence of each type of bond is independent of their energy, and the situation where all the bonds are in the configuration of minimum energy. Neither of this extreme cases are realistic. Indeed, because the system is at equilibrium at $kT = 1$, the bonds with high energy will not be observed. Also, because one particle always has six faces, the equilibrium organization of the particle will involve a set of compatible favored faces, and not just the most favored one.

In the first situation, the occurrence of a face pair $(a, b) = s$ is independent of its energy, such that $N_s = n$. Then, the energy of a system would just proportional to $\mathbb{E}(J_s)$, the expected value for the bonds' energy. $\mathbb{E}(J_s) = \sum_s J_s/21$. There are 21 values for J_s in the case we study, and because of the law of large numbers applied to the values of J_s, we then expect $\mathbb{E}(J_s) \approx \mu$, with μ the average of the distribution where the interaction map was drawn. Then, the expected energy of a system with interaction map drawn in $\mathcal{N}(\mu, \sigma)$, which we denote as $\mathbb{E}(E^{(\mu,\sigma)})$ verifies $\mathbb{E}(E^{(\mu,\sigma)}) \approx \mu$.

In the second situation, all the full-full bonds in the system are in the configuration of the face pair of lowest energy, which we refer to as $s_m in$. Then $N_{s=s_{\min}} = n$ and $N_{s\neq s_{\min}} = 0$. The energy of a system would then be proportional to the expected value of the minimum of the interaction map $\mathbb{E}(\min J_s)$. The expected value of the minimum of 21 randomly Gaussian variables depends both on μ and σ. Its expression is $\mathbb{E}(\min J_s) = \mu - \psi_{21}\sigma$, with ψ_n is the expectation of the maximum of n random variable drawn in $\mathcal{N}(0, 1)$, and verifies $\psi_n \sim \sqrt{\log n}$ in the asymptotic limit [104]. For 21 variables, the expected value of the minimum is $\psi_{21} \approx 1.7$. Then, the expected energy of a system with interaction map drawn in $\mathcal{N}(\mu, \sigma)$ is $\mathbb{E}(E^{(\mu,\sigma)}) \approx \mu - 1.7\sigma$.

We expect the situation to be in between both extreme cases we just described:

$$\mathbb{E}(E^{(\mu,\sigma)}) \approx \mu - m\sigma \tag{3.2}$$

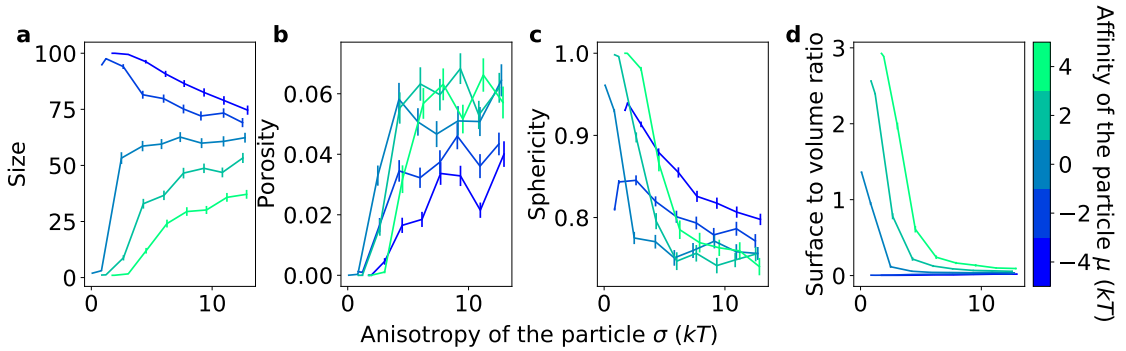

Figure 3.3: Aggregates have less trivial geometry when the anisotropy of the particle increases. (a) Increasing the anisotropy reduces (resp. increases) the size of aggregate of attractive particle (resp. repulsive) particles. It increases the porosity of all aggregates (b) and decreases their sphericity (c) and surface to volume ration (d), and reduces the aggregates size. The error bars are the standard error of the measure (200 data per points)

with m between 0 and ψ_{21}. For each interaction map drawn in $\mathcal{N}(\mu, \sigma)$, we compute the energy of the system with Eq. 3.1. We then average this quantity for all the $N_{\text{data}} = 200$ interaction maps drawn in $\mathcal{N}(\mu, \sigma)$. In Figure 3.2a, we plot $\mathbb{E}(E^{(\mu,\sigma)})$ as a function of $\mathbb{E}(\sigma)$ for all the values of μ (referenced in the color bar). We denote as m and p the slope and the origin measured by the linear fit. As expected, the energy of the system decreases when μ decreases and when σ increases. We fit the data with a linear function, for each value of μ, and plot the slope (resp. the origin) of the curves in Figure 3.2b (resp. c). Because the energy is necessarily lower than zero, the linear evolution of the energy is capped for the lowest values of μ and σ. We only fit the curve in the linear regime (see the beginning of the dashed lines in the figure). m appears independent of the value of μ in plot (b), and p appears to scale linearly with μ in plot (c). We verified the expected scaling of the equation 3.2. The measured value for the coefficient of the linear dependence of the energy m on σ is around $0.9 < \psi_{21}$. This means that the energy of the system is not on average governed by the energy of the lowest interaction, if the asymptotic limit of ψ is verified. There is however a linear dependence in σ, which means that the energy of the system is governed by some of the interaction of lowest energy. This measure however does not provide information about the shape of the aggregates and its dependency on the particle anisotropy.

3.1.3 Tendencies in the descriptors

We now show that increasing the anisotropy of the particles leads to less trivial aggregates, *i.e.* porous and non-spherical aggregates of intermediate sizes. We observed important variations of the aggregate shape in the examples shown in Figure 3.1: as the anisotropy of the particle increases, the aggregates have more complex shapes than the gas or the liquid. In Chapter 2, we introduced quantitative descriptors of the aggregate shapes: the average size of the cluster in the system, their porosity (number of holes per particles), their sphericity (between 0 and 1) and the surface to volume ratio. We compute those shape descriptors for all the systems and measure their dependency on the anisotropy of the particle.

For each interaction map, we measure the averaged size, porosity, sphericity and surface to volume ratio of all the aggregates measured at equilibrium. We then average those quantities for the $N_{\text{data}} = 200$ maps drawn in the same distribution $\mathcal{N}(\mu, \sigma)$. The average values of the geometric descriptors as a function of σ for each values of μ (indicated by the color bar) are plotted in Figure 3.3.

For quasi-isotropic particles ($\sigma = 0.1$), we recover the trivial lattice gas model: aggregates are either bulks when the particles are attractive, or monomers when they are repulsive. Such aggregates are either of size 1 or 100 (the total number of particles) (left points of panel a). They are never porous (left points of panel b). They are spherical (left points of panel c), except for the case $\mu = 0$ where small aggregates self-assemble for entropic reasons, and they are not necessarily spherical (such aggregates can be observed in Figure 3.1 for $\mu = 0$ and $\sigma = 0.1$). The surface to volume ratio (left points of panel d) is large for small aggregates (3 surface bonds per particle) and low for large aggregates (the number of surface bonds scales like the square of the aggregate size).

When the particle anisotropy increases, the influence of its affinity is less important, curves of different colors reach more similar values. Figure 3.3a shows that the aggregate size increases with anisotropy for repulsive particles and decreases for attractive particles: For highly anisotropic particles, there are too many constraints for the particles to form an aggregate of infinite size when they are sticky on average. When they are repulsive on average, the few attractive interactions enable the particles to form an aggregate of bigger sizes. Figure 3.3b shows that the porosity of the aggregates always increases with the anisotropy of the particle: dense aggregates cannot form when there are too many unfavored interactions. Figure 3.3c shows that the shape of the aggregates becomes less trivial: we recall that the sphericity decreases if the aggregate has more surface than a spherical aggregate of the same size. The sphericity is around 0.8 for largest values of anisotropy, which means that on average, the aggregates have 20% more surfaces than their spherical equivalent. The surface to volume ratio Figure 3.3d is less informative, because all its variations are governed by the size of the aggregate.

The anisotropy is a relevant measure of the particle interactions to characterize the complexity of the aggregate they form. Indeed, upon increasing anisotropy, the particles form aggregates that are more complex than dense aggregates or infinite sizes, or monomers. From the examples of Figure 3.1, it is however clear that the averages we computed in this section are not sufficient to characterize the diversity of aggregates observed with our sampling of the interaction maps, and for the same value of affinity and anisotropy, aggregates can still be very different.

3.1.4 Measure of frustration

Here, we show that aggregates are frustrated. A particle in an aggregate is in interaction with all its neighbors such that there are geometric constraints that do not allow all the interactions to occur: the system can be frustrated. We expect the aggregates of anisotropic particles to be frustrated, because the particles have more incompatible interactions. Therefore, we are mostly interested in measuring the frustration of anisotropic particles. By measuring the occurrence of each interaction (the density map introduced in Chapter 2), we determine a quantitative measure of this frustration (Sec. 3.1.4.1). We see that aggregates of anisotropic particle are not more frustrated than those of isotropic particles, according to that measure (Sec. 3.1.4.2).

3.1.4.1 Naive minimization of the energy

A system is frustrated when some favored interactions are unsatisfied, or when some unfavored interactions are satisfied. Here, we introduce a measure of relative frustration that does not depend on the averaged energy of the system. It measures how far the system is from the composition that minimizes its energy if they were no geometric constraints for the self-assembly of the particles.

The energy of the system is $E = \sum_s c_s J_s$. Given an interaction map J, a naive approach consists in finding the density map c that minimizes the energy: we expect that the face pairs of low energy are often observed, and that of high energy are not observed. However, the density cannot just be 1 for the most favored bond and 0 for all the others: we showed in Sec. 2.1.3 of Chapter 2 that the number of bonds in a given structure N_s respect some constrained because the number of particles (and therefore the number of faces) in the system is fixed. The fraction of bonds in structure $c_s = N_s/N_{\text{bonds}}$. We can therefore determine a *minimal density map*, $c^{(\min)}$ (or alternatively $\mathbf{c}^{(\min)}$ in the vectorized form), which is the density map that minimizes the energy, subject to the constraint of conservation of the number of particles. However, this measure does not take into account the geometric constraint of the system, as it is done by the Monte-Carlo simulation. Indeed, if there are loops of three particles involving contacts between the following faces (a, b), (b, c) and (a, c), and if contacts (a, b), (b, c) are very favorable, but contact (a, c) is repulsive, the idealized minimization will simply ensure that $c_{ab}^{(\min)}$ and $c_{bc}^{(\min)}$ are as large as possible, and that $c_{ac}^{(\min)}$ is zero, which is not possible geometrically.

Formally, we solve the following problem:

$$\mathbf{c}^{(\min)} = \min \mathbf{c} \cdot \mathbf{J} \tag{3.3}$$
$$\text{subject to} \begin{cases} c_s \in [0,1] \\ \sum_s c_s = 1 \\ c_{0a} + \sum_{b \neq a} c_{ab} + 2c_{aa} = \frac{N_{\text{particles}}}{N_{\text{bonds}}} \end{cases}$$

We write the third constraint of Eq. 3.3 with the reference of the faces a, b rather than the label of the face pair s, for clarity of the meaning of this equation (conservation of the faces), but each unique pair of face (a, b) corresponds to a face pair s. We solve Eq. 3.3 numerically for all interaction maps, and we can compare it to the measured density map that result from the numerical Monte-Carlo (MC) annealing, which we denote by $c^{(\text{MC})}$ to distinguish it from $c^{(\min)}$. In Figure 3.4, we show examples of such maps for three systems for which $(\mu, \sigma) = (0, 13)$. For system (a), the system composition computed from the naive minimization is very similar to that measured in the Monte-Carlo: the face pairs that are observed often (colored in blue) are the same for both $c^{(\text{MC})}$ and $c^{(\min)}$. In situations (b) and (c), they are different. Examples in panel (b) and (c) are frustrated in the sense that they do not reach the lowest possible energy, because of geometrical constraints.

We now introduce a quantitative measure of frustration. We can first calculate the difference between the energy of the system computed in the Monte-Carlo simulation, and that of the idealized system.

$$\Delta E^{\text{f}} = \mathbf{J} \cdot \mathbf{c}^{(\text{MC})} - \mathbf{J} \cdot \mathbf{c}^{(\min)} \tag{3.4}$$

This measure is however proportional to J and will directly scale like $\mu - m\sigma$ as was shown in Sec. 3.1.2. Therefore, we rescale this measure by the energy of the system $E^{(\mu,\sigma)}$. The relative frustration δE^{f} reads

$$\delta E^{\text{f}} = \frac{\Delta E^{\text{f}}}{|E^{(\mu,\sigma)}|} \tag{3.5}$$

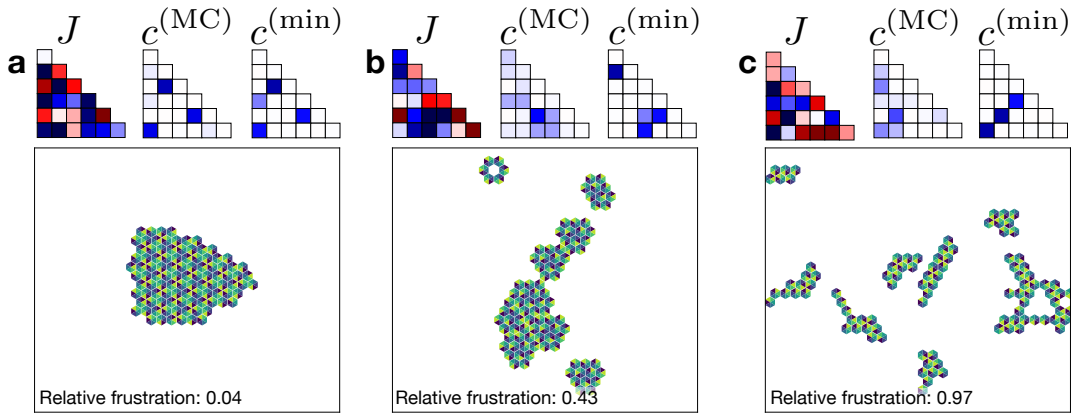

Figure 3.4: A system is frustrated when the density maps are not a trivial minimization of the interaction map. For each interaction map J, we show J, $c^{(\mathrm{MC})}$ and $c^{(\mathrm{min})}$, the density maps measured in the Monte-Carlo simulation, or derived by minimizing the energy without geometric constraints (equation 3.3) System (a) is not frustrated, and system (b) and (c) are. The colors scale of the matrices were detailed previously. Here, the darkest color correspond to -30 or $30kT$ for the interaction map, and 0.05 for the density. The value of relative frustration (equation 3.5) is shown on each image. For all those examples, the interactions were drawn in $\mathcal{N}(0kT, 13kT)$.

We only measure this quantity when $(E^{(\mu,\sigma)})$ is non-zero (see Figure 3.2), *i.e.* when $\sigma \geq 5$. For the examples of Figure 3.4, the computed relative frustration are, respectively 0.04, 0.43 and 0.97. This confirms the observation of the interaction maps, system (a) is not frustrated, and systems (b) and (c) are.

3.1.4.2 Aggregates of anisotropic particles are not more frustrated

We now look for a dependence of the relative frustration on the anisotropy of the particle: anisotropic particles have more unfavored interactions, and they might be more subject to geometric constraints.

We measure the relative frustration for each system, as introduced in Sec. 3.1.4.1 for systems where the energy per particle is non-zero (see Figure 3.2). For simplicity, we focus on the data for which $\sigma \geq 5$. Because within each group of similar σ and μ, the values of frustration are broadly distributed, we choose to represent the histogram of the values they take, rather than the average, as was done in previous plots of this section. The results are shown in Figure 3.5. We plot the histogram of the relative frustration for a given particle anisotropy (σ) in each subplot, and the color reference the affinity of the particle. The position of the histogram bars are identical in all plots. We do not observe a shift of the distribution towards higher values of frustration. The average of the distribution is always similar, and it even seems that the highest observed values of frustration (around $\delta E^{\mathrm{f}} = 2$ are more frequent for particles where the values of anisotropy are medium ($\sigma = 5$ or $\sigma = 7$). There is no influence of the affinity of the particle (no separation between the colors in a given histogram). The aggregates of particles with large anisotropy are not more frustrated than the others, which is not what we expected.

With the measure of frustration that we introduced, which quantifies the relative extra energy the system has compared to an equivalent system without geometric constraints, most of the aggregates are frustrated: for most of the aggregates, the density map is different from the result of the naive minimization, and there is an extra energy associated with this difference (up to around 200% extra energy). We do not observe differences between aggregates of medium and large anisotropy, which we interpret as follows: in our systems, the particles are in a dilute system (only 10% of the sites are occupied) and for

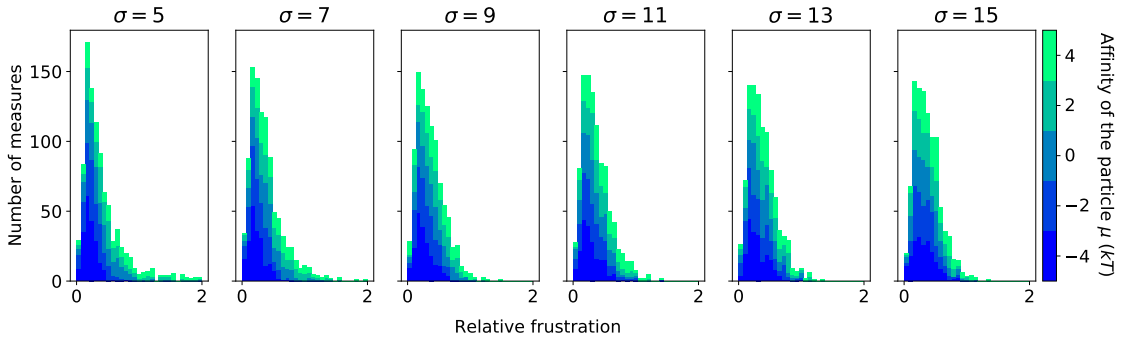

Figure 3.5: Distribution of relative frustration is independent of the affinity and anisotropy of the particle. Each graph corresponds to the data measured from a given value of the particle anisotropy σ. The colors correspond to the affinity of the particle μ. The distribution does not seem to depend on the values of μ and σ.

this reason, when the interactions are anisotropic and not compatible, the system will avoid frustration by adopting an aggregate morphology that is less densely packed (smaller, more porous, less spherical). For this reason, the energy of the system will not be much smaller than that of the idealized energy without geometric constraint.

In this section, we showed that anisotropic particle form aggregates with more complex shapes, and the reason might be that a dense aggregate would be too frustrated. Forming smaller aggregates, having vacancies, or being fibrillar would then be a way to escape frustration. However, the analysis we proposed here depends on averages over aggregates that are very different, even if the particles have the same values of affinity and anisotropy.

3.2 Classification of the aggregates

To understand the relation between the interaction map and the shape of the aggregate, we need to obtain an individual characterization of each system. Indeed, the examples of aggregates we gave in Figure 3.1 for a given affinity and anisotropy were one example among 200 results of equilibrating, and they were not representative of the other 199 aggregates of the same affinity and anisotropy. We observed in the images of aggregates that there seemed to be some stereotypical characteristics common to several aggregates that have different local organization of the particles. For instance, we identified as fibers any aggregates of width 1, 2 or 3 particles, regardless of the arrangement of the fibers in the aggregate, and of the fact that they are branching or not. Because the geometric descriptors introduced in Sec. 2.4.3 of Chapter 2 are not sufficient to systematically characterize individual aggregates, we will do it with supervised machine learning. In this section, we classify all the aggregates we obtained by the random sampling introduced in Sec. 3.1, and see that affinity and anisotropy are not determinant of a type of aggregate, but that fibers and small aggregates are more often formed by anisotropic particles. In Sec. 3.2.1, we introduce our categorization of the aggregates, and in Sec. 3.2.2, we explain which data we use to categorize the aggregates. In Sec. 3.2.3, we detail the machine learning method we use, and in Sec. 3.2.4, we give a phase diagram of which aggregates are observed for which values of affinity and anisotropy, and we justify *a posteriori* why this classification would not have been easily implemented without machine learning methods: there is no simple criteria that enables to retrieve the machine-learning categorization.

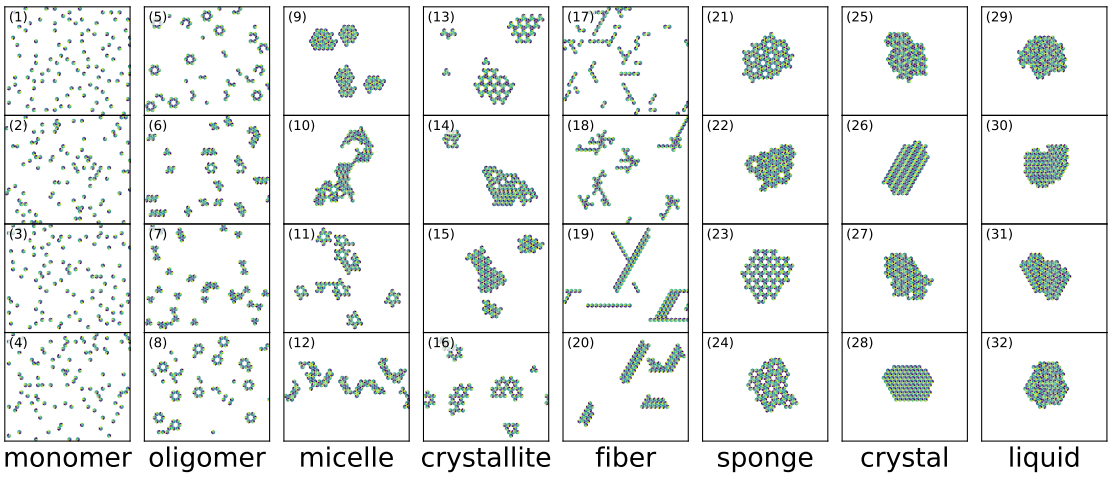

Figure 3.6: Labelling of the aggregates. Each column correspond to a label, and we give four examples of aggregate per label.

3.2.1 Labeling of the aggregates

We classify the observed aggregates according to their geometrical properties into 8 distinct categories: monomer, oligomer, micelle, crystallite, fiber, sponge, crystal and liquid. This is done manually by looking at the snapshot of the data. These labels will then be used as the categories of the supervised learning classification. Here, we explain the criterion that are used for the labelling, and give examples of aggregates in each categories in Figure 3.6.

We first distinguish the aggregates were all the particles have aggregated (the size of the clusters is typically close to 100, which we sometimes refer to as infinite size). This corresponds to the three columns on the right of Figure 3.6. In that case, the particles can be arranged in a periodic organization without vacancies, and it is a *crystal*, or with vacancies, and it is a *sponge*. The crystals can be patterned in different ways, all the particles are in the same orientation (example (25) in the Figure), or there is an alternation of the orientations $(26-28)$ in the figure. The vacancies in the sponge can be organized in different ways (see difference between ((23) and (24)). Sometimes, a fraction of the vacancies are filled with a particle (like in (22)), but we still label this as a sponge. The aggregates of size 100 that are not organized in a periodic way are labelled as *liquid*. In some cases there, is some parts of the aggregates that are organized, and some other that are not (like examples (30) and (31)), but as long as the pattern is not present in the whole aggregate, we still label it a liquid.

We can then distinguish between the aggregates of very small size. If there is an elementary motif repeated in each aggregate, it is an *oligomer*, otherwise, the aggregates are *monomers*. This corresponds to the two columns of the left of Figure 3.6. The monomer correspond to both repulsive particles (1) or non-interacting particles that can be next to each other because the dilution is not infinite (2). In cases like (4), where some particles are aggregated and some are not, we still label it a monomer. Oligomers correspond to cycles (8), trimers (7), dimers, tetramers, or objects or a mixture of those $(5-6)$.

Finally, we distinguish between the aggregates of large but finite size. This corresponds to the middle columns of Figure 3.6. When the aggregates are elongated in one dimension, we label them as *fibers*. Fibers can be of width one-particle (17), two-particles (19) or more (20). They can also branch, such as in (18), and be there can be an alternation of orientations within the fibers like in (17). When there is a crystalline pattern in the aggregates, but it was not sufficient for the aggregate to be of infinite size, we label it

crystallite. This encompasses porous (13) and dense (15) crystalline patterns. When portions of the small crystals did not arrange correctly (14), we also label it crystallite. Finally, the aggregates of finite sizes for which the size was limited because some faces of the particle were preferentially at the surface are called *micelle*. Those surface effects are clearly visible in situation (9) and (12). This last category also includes aggregates that could not be classified in other categories. For instance, aggregates in (11) are not identical enough to be classified as oligomers, and the local organization of the particles in (10) is not identifiable enough for the aggregates to be classified as crystallites.

With those criteria, we label 408 of the 9000 results of equilibrating presented in Sec. 3.2.2 manually. This corresponds to ≈ 7 data for all values of affinity and anisotropy. We also labelled data that were initially mis-predicted by the classification algorithm (for which we explain the method in Sec. 3.2.3). In the labelled data, there is 14% of monomers, 9% of oligomers, 20% of micelles, 13% of crystallites, 6% of fibers, 9% of sponge 16% of crystals and 13% of liquids.

There are some situations where the aggregate in the image are at the limit between two categories: some liquids are almost completely organized and could be classified as crystals, some crystallites are almost completely crystallized and could be classified as crystals or sponge, some micelles are a disordered arrangement of short fibers and could be classified as fiber, or there seem to be a pattern, and they could be classified as crystallite. However, the canonical examples of those categories, (such as those in the top line of Figure 3.6) appeared too different to merge the categories they belong to.

3.2.2 The classification relies on measure of density map and geometric descriptors

For a given interaction map, we measured the composition of the system, as well as geometric descriptors, which we introduced in Sec. 2.4.3 of Chapter 2. Here, we explain how those measures are used to create an *input vector* $\mathbf{X}$ for each interaction map, that can be used to classify the data.

For each data, we concatenate

- the interaction map J (21 values)
- the density map c, and the empty-full and empty-empty densities of bonds (a total of 28 values)
- the average size of the clusters, volume, porosity, sphericity and surface to volume ratio (details on how they are calculated were given in Chapter 2) (5 values)
- the total energy of the system (equation 3.1) and the measured average and standard deviation of the interaction map (3 values)

We call those values *features*. The result of the equilibrating for a given interaction map is therefore described by $21+28+5+7 = 57$ features. There are some redundancies in this information (the energy of the system is just the sum of the pairwise product between the interaction map and the density map), and the average and standard deviation of the interaction map are calculated from it. However, keeping those redundancies that correspond to physical description of the system (average and standard deviation of the interaction map correspond to the affinity and anisotropy of the particles, for instance) improved the performance of the machine-learning classification.

Each vector $\mathbf{X}$ is normalized by the features: for each feature, we calculate the norm of all the measures for all the data (labelled and unlabeled), and normalize each feature by this quantity.

In Sec. 2.3.5 of Chapter 2, we explained why some interaction maps are equivalent up to a cyclic permutation of the lines and the column. This is because there is an arbitrary convention that is chosen to order the faces of the particles. We want our classification to be independent of this arbitrary convention. For this reason, for one given simulation, there are 12 data in our dataset, which correspond to the 12 permutations of the interaction and density maps. All other features are left unchanged. This is similar to adding rotation and mirror images of cats and dogs in image recognition. This technique is called *data augmentation* [105], and it improves the performance of our machine-learning classification.

Finally, the categories in the labeled data are not equally distributed, which could make the algorithm learn better to recognize the most frequent categories. To avoid this, we use a technique called *up-sampling* that simply consists in duplicating some data in the least represented categories. We do this for the sponge and the fiber.

For each data, we determined in Sec. 3.2.1 a label that correspond to one of the eight categories of aggregate. We encode this labels as integer c between 0 and 7: monomer is $c = 0$, oligomer is $c = 1$, etc. We can now compute the vector $\mathbf{Y}_{\text{true}}$ of dimension, 8 where $\mathbf{Y}_{\text{true}}[c] = 1$ if c is the category of the corresponding data, and $\mathbf{Y}_{\text{true}}[c] = 0$ otherwise.

With all of those techniques, we now have an input vector $\{\mathbf{X}\}$ of dimension 5220×57 (the number of data times the number of features), and a vector $\{\mathbf{Y}_{\text{true}}\}$ of dimension 5220×8 (the number of data times the number of category), which corresponds to the labels of the data.

3.2.3 Method

We now build a *classifier* that predicts the aggregate category for a given data. We use a *feed-forward neural network classifier* to learn the aggregate categories from the labelled data [106]. This corresponds to measuring an output $\{\mathbf{Y}_{\text{pred}}\}$ from the input $\{\mathbf{X}\}$, such that $\{\mathbf{Y}_{\text{pred}}\}$ is as close as possible to $\{\mathbf{Y}_{\text{true}}\}$.

Classification is a widespread application of machine learning. Each data is assigned a label, and a neural network is fitted to correctly predict these labels. In our case, the data is a list of information about the energy and composition of the system, and the geometric features of the aggregates. The labels are the categories of aggregates described in Sec. 3.2.1. In practice, the network computes a probability for a data to correspond to each of the categories of aggregate. The output is just an array of dimension 8, the number of categories.

Detailed explanations about on neural network can be found in [105]. Here we present the main concepts. A network is composed of several *layers* that performs operations on the input vector. One layer will transform a vector $\mathbf{X_0}$ of dimension n_0 into a vector $\mathbf{X_1}$ of dimension n_1 with the relation

$$\mathbf{X_1} = \sigma^{(1)}(W^{(1)} \cdot \mathbf{X_0} + \mathbf{b}^{(1)}) \tag{3.6}$$

where $W^{(1)}$ and $\mathbf{b}^{(1)}$ are respectively the matrix of the weights (of dimension (n_0, n_1)) and the bias vector (of dimension n_1) of the first layer. σ is a non-linear function. A network is then composed of n layers, and for each layer k, there is a weight matrix $W^{(\mathrm{k})}$ and a bias vector $\mathbf{b}^{(\mathrm{k})}$. The initial input layers are passed through all the layers, by repeating the operation of equation 3.6.

$$\mathbf{x_0} \xrightarrow{\sigma^{(1)},W^{(1)},\mathbf{b}^{(1)}} \mathbf{x_1} \xrightarrow{\sigma^{(2)},W^{(2)},\mathbf{b}^{(2)}} \mathbf{x_2} \ldots \xrightarrow{\sigma^{(n)},W^{(n)},\mathbf{b}^{(n)}} \mathbf{x_n} \tag{3.7}$$

We choose $\sigma^{(k)}$ to be a *rectified linear unit* function defined as

$$\begin{aligned} \sigma(x) &= x \text{ if } x > 0 \\ &= 0 \text{ otherwise} \end{aligned}$$

For the last layer, the activation function is a *softmax* function: it rescales all the entries of the vector such that $\mathbf{Y}_{\text{pred}}[c]$ is the probability that the data belongs to the category c of aggregates.

$\mathbf{X}_0$ and $\mathbf{X}_n$ are the input and output vectors, that we call $\mathbf{X}$ and $\mathbf{Y}_{\text{pred}}$. The principle of machine learning is to determine the set of $\{W^{(k)}\}$ and $\{\mathbf{b}^{(k)}\}$ that will minimize the distance between $\mathbf{Y}_{\text{pred}}$ and $\mathbf{Y}_{\text{true}}$. This is measured with a loss function $\mathcal{L}$. We chose the cross entropy loss function, which is the usual choice for categorization problem [105].

$$\mathcal{L}(\mathbf{Y}_{\text{pred}}, \mathbf{Y}_{\text{true}}) = -\mathbf{Y}_{\text{true}} \cdot \log(\mathbf{Y}_{\text{pred}}) \tag{3.8}$$

To prevent the values of the weight to take too large values, we add a *regularization*. In practice, we add the term $l_1 \sum \left(W_{ij}^{(k)}\right)^2 + l_2 \sum |W_{ij}^{(k)}|$ to the loss function [105].

On a given dataset, we can also measure the accuracy $\mathcal{A}$ of the prediction: it is the number of correct prediction over the whole dataset

$$\mathcal{A} = \sum_{\text{data } i} \delta(\text{argmax}_c \mathbf{Y}^i_{\text{pred}} = \text{argmax}_c \mathbf{Y}^i_{\text{true}}) \tag{3.9}$$

where $\delta(x = 0) = 1$, and $\delta(x \neq 0) = 1$.

The minimization is done with several successive gradient descent, over a subset of the data called *minibatch*. Those successive minimizations are called *epoch*, and this technique is called *mini-batch gradient descent*. This fastens the training process.

This method is implemented in the python API Keras, and the Dense Layer class.

The dataset is then divided between a *training set* and a *test set*: the neural network is fitted to minimize the loss on the training set. We then evaluate the network by measuring the accuracy of the predictions on the test set. This ensures that the network learned general rules to relate the data and the labels, and that the classification is not specific to the data the network was trained on. If the difference between the accuracy on the training and test set is large, it means that the network *overfitted*: the network captured trends that are specific to the dataset it was trained on, but not generic. We chose the hyperparameters of the network (number of layers, size of each layer, learning rate, number of epochs, the values for l_1 and l_2) such that the test accuracy is large, and the difference between the train accuracy is small.

In our problem, we obtained 99% learning accuracy and 93% test accuracy with the following hyperparameters: there are 5 hidden layers of size $n_1 = 100$, $n_2 = 200$, $n_3 = 400$, $n_4 = 100$, $n_5 = 30$. There are 800 epochs for which the measure is performed on minibatches of 128 data. $l_1 = 10^{-4}$ and $l_2 = 10^{-5}$. In Figure 3.7 a and b, we show the accuracy and loss measured on the training set along the learning (one measure for each epoch). The accuracy progressively increases towards its maximum as the network is trained. We also count, for each category, the number of correct and mis-predictions on the data that we did not use to train the network. This is shown in Figure 3.7c, as a *confusion matrix*. The diagonal terms correspond to the correct prediction. All the categories are correctly predicted, but there are some mis-predictions between micelles and crystallites. This was expected since the limits between those two categories were difficult to identify, as explained in Sec. 3.2.1.

Because the mis-predictions correspond to distinctions that were hard to make also manually, we consider that the network now predicts with sufficient accuracy the category of any unlabeled data, and we use it as a classifier.

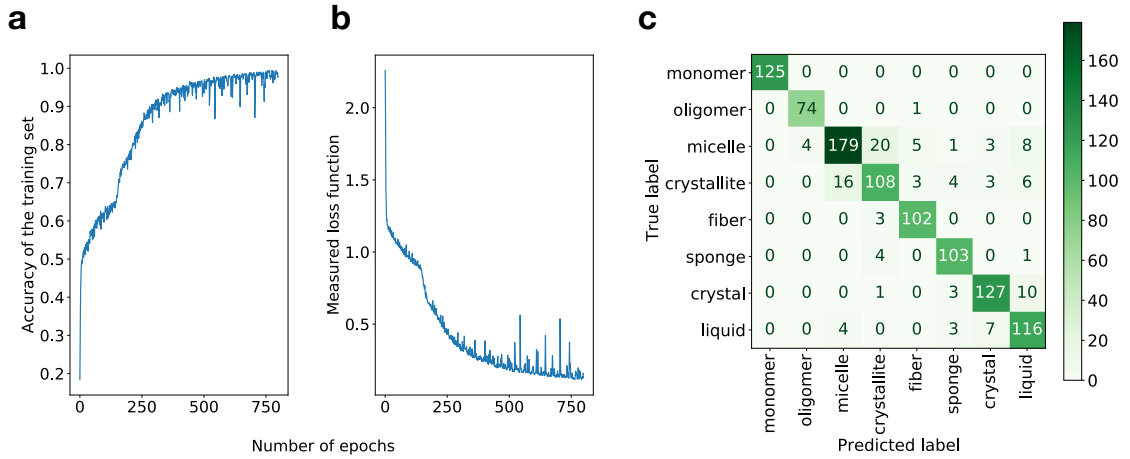

Figure 3.7: We train a neural network to classify the aggregates. (a) and (b) correspond to the evolution of the accuracy of the prediction on the training set, and on the measure of the loss, as a function of the number of epoch (the amount of training time). At the end of the training of the neural network, the accuracy (resp. loss) converged to a maximal (resp.) value. The loss and accuracy are defined in equations 3.8 and 3.9. c) Confusion matrix: for each true label, we show the number of data that were predicted with that label (in the diagonal) or with another label, for the data in the test set.

3.2.4 Phase diagram

With this classifier, we can determine more precisely the relation between the interaction maps and the category of aggregates. Here, we classify the whole dataset, which correspond to 90000 data, 200 for each value of affinity and anisotropy of the particle. We show that affinity and anisotropy are relevant parameters to predict the type of aggregates, but that they are not sufficient.

We regroup the particles for which the affinity and anisotropy is similar, and we show the occurrence of each category of aggregates within this subset in Figure 3.8. In this figure, each pie-chart plotted at coordinates (μ, σ) correspond to statistics of data for which the affinity (resp. anisotropy) of the particle, *i.e.* the measured average (resp. standard deviation) of the interaction map is in $[\mu - 0.5, \mu + 0.5]$ (resp. $[\sigma - 0.5, \sigma + 0.5]$). The pie-chart thus correspond to between 21 and 256 interaction maps. The colors then show how often each category is observed. We find again the isotropic limit (bottom of the diagram) for which the particles either aggregate in liquids when the affinity is negative, or stay as a gas if the affinity is positive. We can distinguish three main regions, delimited by the gray dotted line. In the lower right region, $\mu - \sigma > 2kT$, which suggests that the bond with the lower interaction energy typically is of the order $2kT$. The particles are mostly repulsive, and the observed aggregates are mostly monomer or oligomers (green). In the bottom right region, $\mu + \sigma < 2kT$, which means that the bond of highest energy is not repulsive. The particles are mostly assembling in liquids, crystals, or sponges (blue). The rest of the diagram is much more mixed, and almost all types of aggregates are observed for each couple (μ, σ). In the regions where the anisotropy is large, we mostly observe aggregates of lowered dimensionality (fiber, crystallite, micelles, and oligomers), or two-dimensional but porous aggregates (sponge). The anisotropy of the interaction prevents the aggregate from assembling into a trivial shape, like the liquid, but they still self-assemble (they do not remain as monomers): dark blue and dark green slices are scarce in the upper regions of the diagram. The effect of the particle affinity is also less and less important as the anisotropy increases: no clear difference between the top left and the top right of the diagram. Indeed, for high anisotropies, the self-assembly is governed by few of the most favored interactions

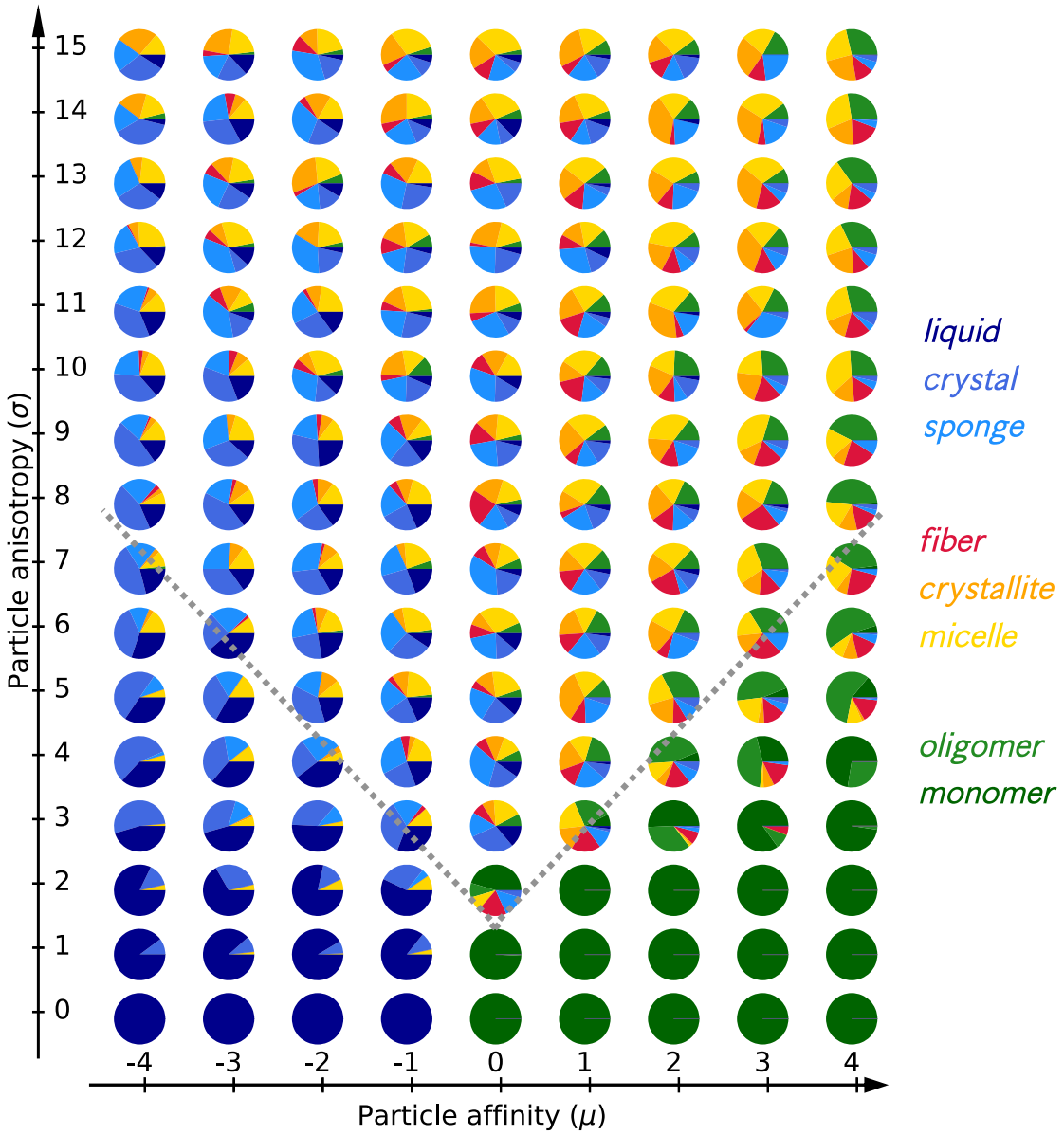

Figure 3.8: We observe aggregates of reduced dimensionality for anisotropic particles. Each pie-chart shows which categories of aggregates are observed for the aggregation of random particles of given affinity and anisotropy. Each pie-chart represent statistics for between 21 and 256 particles, and the whole diagram represents statistics for 9000 particles. Aggregation categories are assigned with machine learning.

only, as was suggested in Figure 3.5.

Now that each data is classified, we ask whether it would have been possible to achieve such classification without machine learning, simply by using simple criterion on the interaction map, the density map, or the geometric descriptors. For this, we use principal component analysis (PCA) to project the data in the space where the variance is maximal [107]. If there are some simple linear rules that could be applied to categorize the data, they will appear well separated according to their category in the projection. We do this for some features of the dataset $\{\mathbf{X}\}$ introduced in Sec. 3.2.2. We consider $\{\mathbf{X}_1\}$, for which the features correspond to the normalized values of the interaction map only (dimension 9000×21), $\{\mathbf{X}_2\}$, for which the features correspond to the normalized values of the densities only (dimension 9000×28), and $\{\mathbf{X}_3\}$, for which the features are the size, porosity, sphericity, and surface to volume ration of the aggregates (dimension 9000×4). For $\{\mathbf{X}_1\}$ and $\{\mathbf{X}_2\}$, we augment the data as was done in Sec. 3.2.2, to take into account the 12 cyclic permutations of the matrices. We do principal component analysis of $\{\mathbf{X}_1\}$,$\{\mathbf{X}_2\}$ and $\{\mathbf{X}_3\}$ and show the result in Figure 3.9 a, b and c. The colors correspond to the categories introduced above and are referenced in Figure 3.8. This diagram shows that we could not find a trivial projection of the interactions map that is related to the aggregate category. The projection of the data points of density maps and of the geometric descriptors are on the contrary well separated according to the aggregate categories, because those measures refer to the equilibrium configuration of the system. There is no clear separation of the points, however, and it is clear from this projection that the limits between two categories are not uniquely defined. There is no trivial criterion that we could have decided to define aggregate categories from the density map and the geometric descriptors.

We also verify that the aggregate categories are not related to their frustration. The initial hypothesis of this chapter was that the aggregates of lower dimension, such as fibers or micelles, emerge when there is frustration in the interactions, *i.e.*, if some favored bonds are not realized because of geometric constraints. We now have a tool to verify if there is a relation between frustration and dimensionality reduction. In Figure 3.9d, we plot the measure of relative frustration introduced in Sec. 3.1.4. We do not observe any correlation between the relative frustration and the aggregate category. This confirms the conclusion we draw in Sec. 3.1.4: particles may assemble into aggregates of lower dimensionality to avoid frustration, and this is the reason they form fibers, or micelles. As a consequence, these types of aggregates are not necessarily more frustrated.

In this section, we classified individually each of the 9000 aggregates of particles with random interactions, and observed that for non-isotropic particles, all aggregate categories could be observed. We also confirmed the tendency of anisotropic particles to self-assemble in aggregates of lower dimensions, like fiber, micelles, oligomers, or crystallites. The method we used to predict the category of aggregate did not however provide a rational understanding of what specificities of the interaction maps are responsible for the aggregation in one category or the other. Anisotropy and affinity are also not sufficient to discriminate between the aggregate categories.

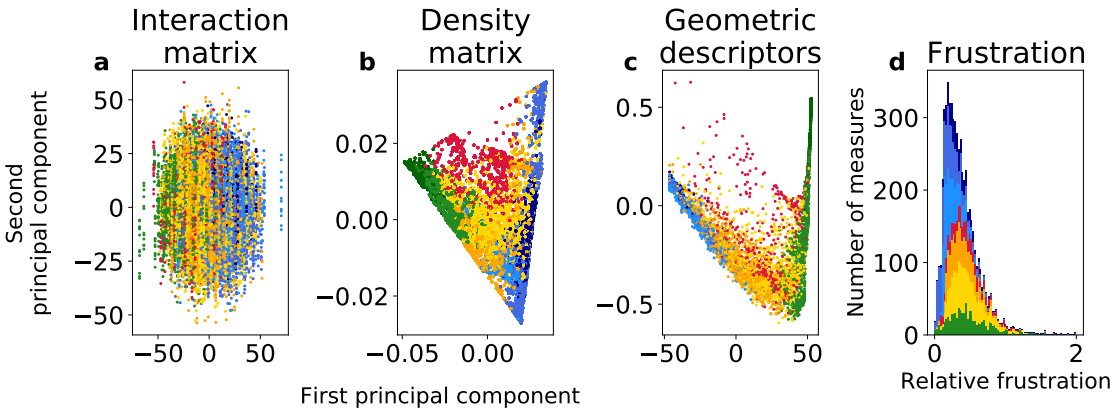

Figure 3.9: The determination of the aggregate categories is not trivial. Principal component analysis of the interaction maps (a), density maps (b) and geometric descriptors (c), with the labels learned with neural network. The colors are the same as in Figure 3.8. d) Histogram of the values of relative frustration measured with equation 3.5. We do not observe a correlation between frustration and aggregates category.

3.3 Relation between particles interactions and aggregates shapes

We now take advantage of the large dataset of interaction maps for which the aggregates have been assigned with a label to rationalize the relation between both. We want to determine how the properties of the local interactions between the particles are related to the result of their self-assembly. In more technical terms, which characteristics of the interaction maps are related to the aggregate category. We compute *predictors* from the interaction map, and introduce a method to quantitatively measure the quality of a predictor (Sec. 3.3.1). We then test this method on different predictors, such as the averaged value of some interactions (Sec. 3.3.2). This method also suggests that both the energy level of the interactions and the relative orientation of the particles they correspond to are necessary information to explain the equilibrium aggregates (Sec. 3.3.3). Finally, we introduce a predictor that describes the ability of the particle to form periodic motifs, and show that it is directly related to the shape of the aggregate (Sec. 3.3.4).

3.3.1 Test importance of the elements of the interaction map with machine learning

We show how neural networks can be used to determine which information in the interaction map is related to the aggregate category. We check that from the interaction map (without adding the density map and the geometric descriptor, as in Sec. 3.2.2), we can train a neural network to predict the category of the equilibrium organization of the particles that have this interaction map. We show that we can also train a neural network from partial information of the interaction map only, and compare how good the prediction of the category is. This will bring insight on how important this partial information is.

Here we compare how the aggregate categories are learned over three datasets $\{\mathbf{X}_1\}$, $\{\mathbf{X}_2\}$ and $\{\mathbf{X}_3\}$. The features of $\{\mathbf{X}_1\}$ are all the values in the interaction map, and the measured values of affinity and anisotropy ($21+2$ features). The features of $\{\mathbf{X}_2\}$ is the interaction map, from which we only keep the diagonal terms, and the measured values of affinity and anisotropy ($6+2$ features). The features of $\{\mathbf{X}_3\}$ are just the measured values of affinity and anisotropy (2 features).

For each of those partial datasets, we train the network on 1300 data, and predict the accuracy of the prediction on 300 other data. We repeat this process 20 times for

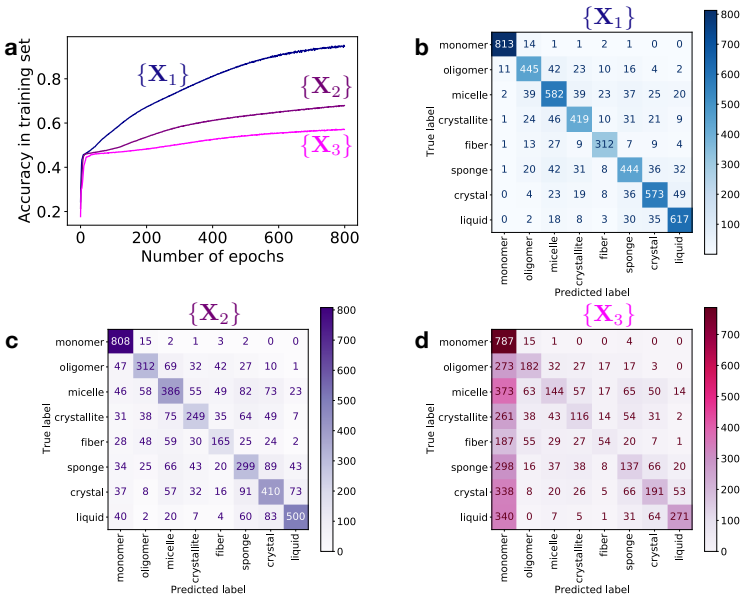

Figure 3.10: The learning is more difficult for partial dataset ($\{\mathbf{X_2}\}$, only the diagonal terms of the interaction map and $\{\mathbf{X_3}\}$, only the affinity and anisotropy), and there are more errors of the prediction (more off-diagonal terms in the confusion matrices). The information in $\{\mathbf{X_2}\}$ is more relevant than the information in $\{\mathbf{X_3}\}$ because the learning is better.

different random seeds. The data are normalized as described in Sec. 3.2.2. The network and hyperparameters are the same as described in Sec. 3.2.3. In Figure 3.10a we show the evolution of the averaged training accuracy during the fitting of the network. When the full interaction map is used for the training (blue curve), the maximal accuracy is reached. When the data used to train contains only partial information, the maximal accuracy is not reached (purple and pink curves). We see, however, that the learning is better when the diagonal terms of the interaction maps are left ($\mathbf{X_2}$) rather than when the only information is the average and standard deviation of the whole map ($\mathbf{X_3}$). This is also confirmed on the confusion matrices of the test set shown in Figure 3.10b,c and d. These matrices show how each category was predicted by the network on data for which it was not trained. The numbers are averaged over the different training. In all cases, the information is sufficient to achieve partial learning, and some aggregates are correctly categorized. However, the number of mis-prediction is more important for $\{\mathbf{X}_2\}$ than for $\{\mathbf{X}_1\}$ and more for $\{\mathbf{X}_3\}$ than for $\{\mathbf{X}_2\}$. This is intuitive: the more information was removed from the dataset, the more difficult it is to learn the aggregates categories.

The accuracy of the training of the neural network on partial data can be used as a quantifier of the importance of those features to explain the category of the aggregate [108]. The measured accuracy of the test set depends on several parameters of the learning, such as the duration of the training (number of epochs) or the architecture of the network. However, if the test accuracy is worse from training different datasets on identical networks, we conclude that those datasets are worse predictors of the aggregate category. We expect the neural network to learn better the aggregate categories from datasets where there is more features, because more information is available. However, if the measured accuracy on the test set is lower when the number of features used to train the network is lower, it means that a good predictor was found: the aggregate category is accurately learned with a small number of well-chosen features. We take as a reference point the learning accuracy for a dataset with all the features of the interaction map. This is the dark blue plus-shaped point in Figure 3.12. For this dataset, the accuracy of the prediction is 0.81.

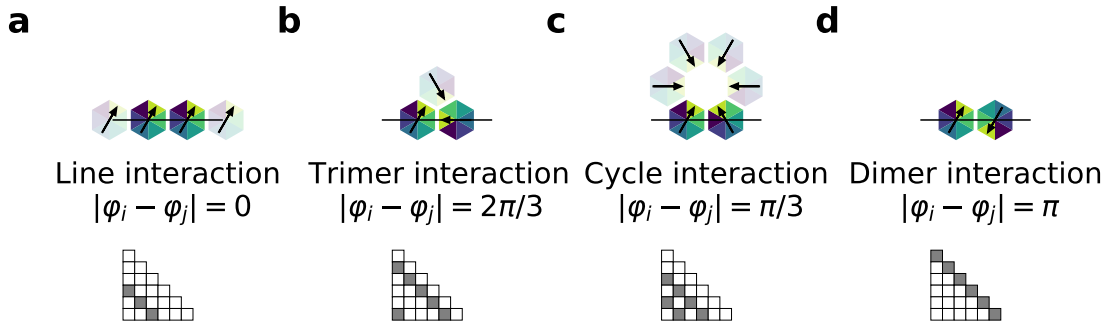

Figure 3.11: Depending on the angle of two neighboring particle orientations, the corresponding interaction leads to aggregate lines, trimer, cycle, or dimers. This angle corresponds to entries of the interaction map that are in the same diagonal (colored in grey).

3.3.2 Masking and averaging

Here, we compare the accuracy of the predictions on the modified interaction map to determine which part of the interaction map is important to predict the aggregate category. With the method in introduced in Sec. 3.3.1, we can test naive hypothesis, such as whether the interactions of the first face of the particles (first column in the interaction map) is more important than the interactions of its second face (second column). On this specific case, it is clear that both should be equally important, because of the particles symmetries. We introduce competing hypothesis on what is important in the interaction map, and measure the accuracy of the predictions on the corresponding reduced interaction matrix. Here, for instance, we test whether some groups of interactions in the interaction maps are more important than others. We explain in Sec. 3.3.2.1 how the different entries of the interaction map can be regrouped. We then train the network while masking some of the group of interactions Sec. 3.3.2.2, or by averaging together the energies within the same group of interaction Sec. 3.3.2.3. This will reveal that the ability of the particles to form lines is a very important predictor of the aggregate category.

3.3.2.1 Angle of interaction or faces of the particle

There are two ways to regroup the entries of the interaction maps. We can first regroup the entries that involve the same face of the particle, which are simply the lines (or columns) of the interaction map. We can also regroup the entries that correspond to the same *angle of interaction* that we define here.

In Figure 3.11, we show what are the possible angles between two neighboring particles, and to what entries of the interaction map they correspond (colored in gray in the matrix). The angle of the favored interaction is determinant for the type of aggregate it can lead to. Favored interaction that align the particle (*line* interactions), such as in panel (a), will favor the formation of fibers). Favored interactions that have angle $2\pi/3$ or $\pi/3$, such as in panel (b) and (c) will favor the formation of loops of three particles (*trimer* interactions) or six particles (*cycle* interactions). Finally, the interactions for which the angle in π, panel (d), can only favor an aggregate of two particles (*dimer* interactions).

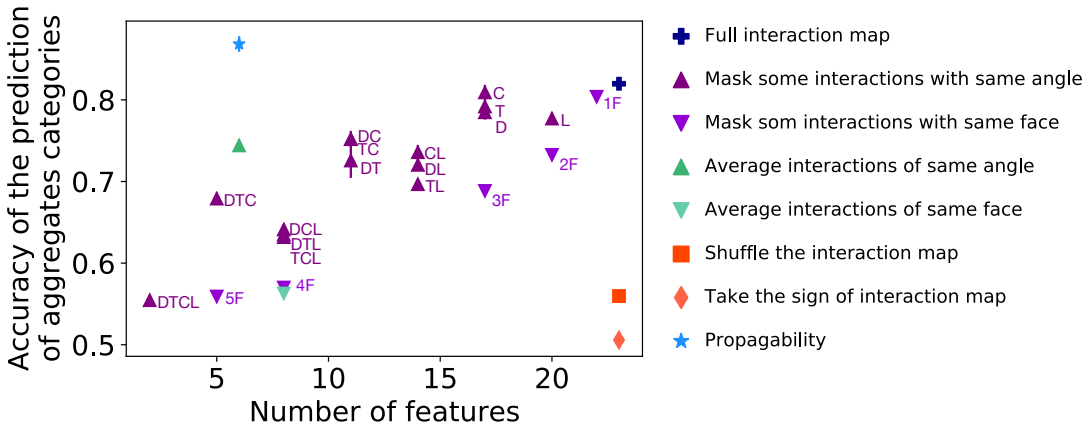

Figure 3.12: The prediction of the neural network decreases when some features are removed. The deviation to this linear evolution corresponds to good predictors of the aggregate category. The letters in purple indicate which portion of the aggregate category was masked: TL means that trimer and line interaction were masked, 4F means that the interaction energies of four faces were masked. The error bars correspond to standard errors over the training of 20 different random dataset (almost always invisible)

3.3.2.2 Masking

A simple way to reduce the information in the interaction map is to set to zero some of its entries. This technique is called *masking*, and it is the one we used in Sec. 3.3.1. We use masking of some values of the interaction map. We can for instance mask the value of the interaction that involve the face a of the particle (one column of the matrix), or the trimer interactions (some sub-diagonals of the matrix, as shown in Figure 3.11). We can mask one or several groups of interactions.

The learning accuracy as a function of the number of features of the dataset is shown in Figure 3.12 in purple triangles. The dark purple triangles pointing up correspond to dataset where some interactions with the same angle have been masked, and the light purple triangles pointing down to dataset where some faces of the particle have been masked. The label next to the point references which part has been masked: the point labeled DC corresponds to masking of the dimer and cycle interactions. The point label $3F$ correspond to masking the interactions of three faces. Because the interaction map is equivalent upon cyclic permutation, it is equivalent to mask the three first and three last faces, and we test only one of them.

We first recover the expected tendency for the prediction accuracy with the number of features: the more features are used to train the network, the more accurate the prediction is. There is however some deviation to this tendency. For instance, point L that correspond to the training accuracy without the line interaction is lower than points C, T and D, where one of the other group of interactions has been masked, even if there are more features. The ability of a particle to form lines is more important to predict the aggregate category than its ability to form dimers. Comparing points DTC and DTL leads to the same results: the accuracy is better from the sole line interaction (DTC) than from the sole trimer, dimer, or cycle interaction. There is no such non-monotonous effect observed upon masking the interactions of the same face: if the interaction energies of fewer faces are known, the accuracy is worse. In general, the prediction accuracy is better from masking some interaction of same angle than by masking some interaction of same face (dark purple triangles are above light purple triangles), which is hard to clearly relate to the importance of the face features compared to the interaction angle feature.

3.3.2.3 Averaging

Here, we test whether all the interaction energies are important, or if it would be sufficient to only take the average of a group of interactions to predict the aggregate category. This also enables to test the importance of a group of interaction: by averaging all the interactions within one group of interaction. For instance, we replace the interaction map by the six value of the averaged face interaction. We can also replace it by the four values of the average for each angle of interaction introduced in the previous subsection. This corresponds to the dark green (average of the angle) and light green (average of the faces) triangle in Figure 3.12. Interaction maps with averages over the interaction angle gives a good prediction of the aggregate category, while that with averages over the faces do not. Again, averaging together interactions of the same angle conserves the information about the ability of the particles to organize in large scale structure, while averaging together the interactions of the same face only conserves information about the local properties of the interactions.

By comparing the prediction accuracy with some partial interaction maps, we found indication that the ability of the particle to align with its neighbors is a good predictor of the shape of the aggregate it will form. The averaged ability of particles to form lines, trimer, cycle and dimers, which we denote as the *topology* of the interactions, corresponds to only four numbers (+2 with the averaged and standard deviation of the interaction map). This measure is a less complex information than the full interaction map, and it still enables to accurately predict the shape of the aggregate. This suggests that the category of an aggregate depends on the topology of the interactions between the particles

3.3.3 Shuffling and taking sign

With the method of Sec. 3.3.1, we can also test whether it is the values of the interactions that matter to predict the aggregate category, or their position in the interaction map.

In simple examples in Chapter 2, we saw that favoring three interactions could lead to very different aggregate, depending on the *position* of the favored interactions in the interaction map (which pair of faces it corresponds to). We want to test whether this generalizes to random interaction maps. For this, we shuffle the entries of each interaction map randomly. If only the values of the interaction map matter, but not the position, it should not lower the learning accuracy. Similarly, we can replace all the values of the interaction map by their sign, which reduces the matrix to a two-level interaction map. If the position of the favored and unfavored interaction in the matrix matter, but the values do not, this should not lower the learning accuracy. These two tests correspond to the orange (square and diamond points) in Figure 3.12. In both cases, the number of features has not been lowered, but the entries were either shuffled or replace by $+1$ or -1. In both cases, the algorithm is not able to predict the aggregate category after being trained on the partial information (the training accuracy is around 0.5).

This is an indication that both the level of energies and the type of interaction they correspond to are important. It suggests that designing the level of the interaction energy and their directionality are two complementary design strategies to obtain a large diversity of aggregates category.

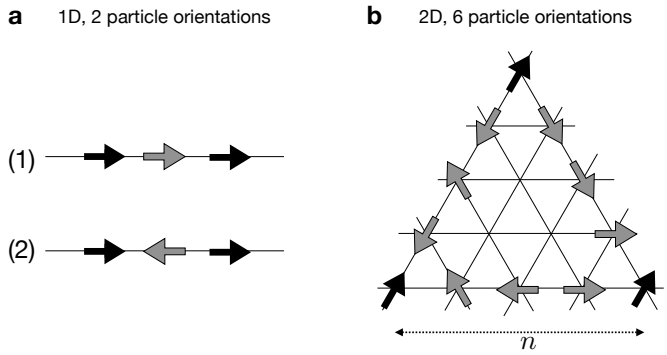

Figure 3.13: The best periodic motif is found by identifying the orientations of the gray arrow that minimizes the energy per bond in the line

3.3.4 Propagability: the ability to form periodic patterns

Large-size aggregate are often built from periodic organizations of the particles. This is also suggested by the fact that the ability of the particle to form lines is an accurate predictor for the aggregate category: a crystal is composed of periodic lines of particles in the three unitary directions of the triangular lattice, a fiber is a periodic line in one direction, and an oligomer is not composed of any periodic line. Here, we introduce a predictor computed from the interaction map that measures the ability to form periodic structures, and we call it *propagability*. A periodic organization of the particles can propagate and tile the plane, while if the energy associated to the formation of a periodic motif is high, the particles will rather assemble into micelles or oligomers. We first introduce the measure of propagability in one dimension (Sec. 3.3.4.1), then extend the definition to two dimensions (Sec. 3.3.4.2). We show that this measure is a good predictor of the aggregate category with the method introduced above (Sec. 3.3.4.3)

3.3.4.1 Propagability in 1D

Let us first consider particles with two orientations in one dimension. There are several sets of favored interactions that lead to periodic infinite aggregate: the sole ($\rightarrow\rightarrow$) interaction is sufficient, but the combination of two dimeric interactions ($\rightarrow\leftarrow$) and ($\leftarrow\rightarrow$) also works. To discriminate between those two organizations of infinite 1D aggregate ((1) $\rightarrow\rightarrow\rightarrow$... or (2) $\rightarrow\leftarrow\rightarrow$..., see Figure 3.13a), we compare the average energy per particle in the two organizations $J_1 = J_{\rightarrow\rightarrow}$ and $J_2 = \frac{1}{2}(J_{\rightarrow\leftarrow} + J_{\leftarrow\rightarrow})$. The effective coupling $J^{\text{eff}} = \min(J_1, J_2)$ will determine the best organization for an infinite 1D aggregate. If J^{eff} is positive, there is no way to assemble particles into infinite aggregates.

We generalize this concept in our problem with 6 particle orientations. For a given initial orientation φ_0, and a given periodicity n, we can consider all the possible set of $n-2$ orientations $\{\varphi_k\}$ of the particles such that a line is in the configuration $(\varphi_0, \varphi_1, ..\varphi_{n-1}, \varphi_0)$, which we refer to as a periodic motif. The values taken by φ_k are necessarily different from φ_0, because if it were not, the motif would be of periodicity lower than n. The effective coupling for a given initial orientation φ_0 and a given periodicity n, which we denote as $J^{\text{eff}}(n, \varphi_0)$ is the minimal possible energy for a periodic motif.

$$J^{\text{eff}}(n, \varphi_0) = \min_{\varphi_1, \cdots \varphi_{n-1}} \frac{1}{n-1}(J_{\varphi_0\varphi_1} + J_{\varphi_1\varphi_2} + ... + J_{\varphi_{n-1}\varphi_0}) \tag{3.10}$$

From a given interaction map, the computation of the values of $J^{\text{eff}}(n, \varphi_0)$ is a straightforward operation on the entries of the matrix. Because of the rotation invariance of the system, we only compute this value for $\varphi_0 = 0$, $\pi/3$ and $2\pi/3$. We also only compute this number for $n \leq 6$, because the particles only has six orientations, there should not be

periodic motifs of more than 6 particles. Computing one value for J^{eff} is at maximum a minimization over 6^4 configurations, which is accessible numerically.

3.3.4.2 Propagability in 2D

We generalize this concept in two dimensions. A periodic motif in 2D is a group of particles such as the one drawn in Figure 3.13b, where the edges of the motif (the black arrow) are fixed. This motif is then composed of three periodic lines, each with a different value for φ_0. For a given periodicity n, the organization of the other arrow (in gray) such that the energy of the whole motif is minimum is then simply computed from the effective interactions of equation 3.10:

$$\mathbf{J}^{\text{eff}}(n) = \left(J^{\text{eff}}(\varphi_0 = 0, n), J^{\text{eff}}(\varphi_0 = \pi/3, n), J^{\text{eff}}(\varphi_0 = 2\pi/3, n)\right) \tag{3.11}$$

Here, we do not count the interactions with the particles inside the motif, that are potentially empty. Indeed, for the largest possible cells, this would amount to considering the organizations of the 12 particles in the extremities, the 6 particles in the middle, that can all be in 7 configurations (6 orientation or empty site), which correspond to 7^{18} configurations, whereas we managed to reduce this computation to 3×6^4 in our case. we see that not taking into account the interior of the motif is already sufficient to predict aggregate categories.

We then choose the best periodicity n^* to be the one where the minimum of the three line energies is the lowest. We could also have chosen n^* to be the size of the cell of minimal energy, but the accuracy of the prediction was less good. The optimal effective coupling vector, $\mathbf{J}^{\text{eff}}(n^*)$ and the optimal periodicity n^*, which can be derived directly from the interaction map, are now used to predict the aggregation category, aside with the particle affinity and anisotropy. This is the propagability, and it has $4 + 2$ features.

3.3.4.3 The propagability is a good predictor of the aggregate shape

We now train the neural network to predict the aggregate category from the propagability measured for each interaction map.

With the chosen network and with the chosen training conditions, the prediction of the aggregate category from the propagability is 0.86, which is even better than the prediction from the full interaction map. The reason for this better prediction is the following: the propagability is independent of the permutation of the interaction map. For values computed on the interaction map, one need to add the 12 invariants in the dataset, as was explained in section 3.2.2. For the propagability, this is not necessary, and the algorithm is therefore easier to train. We expect that the difference between the accuracy of the prediction of the total interaction map and of the propagability will vanish by using a more complex architecture of the network, and training it longer. However, this was not feasible for us in reasonable computational time.

The quality of the prediction of the aggregate shape from the propagability is however interesting: it suggests that this descriptor correctly captures the ability of the particle to form large scale structures.

We introduced the propagability, that is directly computed from the interaction map, and that measures the energy of the best pattern of the particle. If this energy is low, the aggregates will form large scale structure. If only one periodic line is possible, the aggregate will be a fiber. Otherwise, it will be an aggregate of small scale. Because this descriptor also takes into account the possible holes in the structure, and quantitative information about the energy values, it is able to predict all the categories of aggregates, and to distinguish more subtle characteristics like the distinction between a crystal and a sponge, or an oligomer and a micelle. Here, we used machine learning as a tool to test the relevance of an indicator. The measured accuracy then depend on the architecture of the network and from the training protocol. Therefore, we emphasize that it is only relative comparison of the predictions that enabled us to characterize a good and a bad predictor.

3.4 Discussion and extension to two particle types

In the last section, we showed that the ability to form periodic patterns is what relates the local interactions of the particles and the result of its self-assembly. For this, we introduced a method to test what information in the interaction map is related to the macroscopic properties of the aggregate. Our results suggest that both the relative strength of the interactions, and their angle, is determinant to build an aggregate with a periodic pattern or not.

This finding also provides a better understanding of the consequences of frustration in models of particles with directional interactions: we suggest that a particle is frustrated if it cannot *propagate* a low-energy periodic organization of the particles in all the directions of the space, *i.e.*, if the particles cannot tile the plane.

We expect this results to be more generic than the self-assembly of two-dimensional lattice particles: non-lattice particles, and three-dimensional particles, often assemble into periodic aggregates, like the particles that form crystals. Then, the interactions are distributed as the elementary directions of a lattice. This is the case of proteins, for instance. The interactions between particles in a protein fiber or a protein crystal are distributed in a regular way. Then, the concept of propagability of the directional interactions holds: those aggregates can have very large sizes in some directions of the space because there are no geometric constraints that prevent their attractive interactions to be repeated periodically in the aggregate. However, our model does not account for the self-assembly of deformable particles: if the particles are deformable, the periodicity of the interactions is an ill-defined measure. Indeed, it is not possible to compare the energy of different organizations of the particles by enumerating how the particles can organize locally and through which interactions.

We also emphasized that both the strength and the directionality of the interactions are determinant for self-assembly. We also expect this result to hold in the case of non-lattice particles, as long as they are not deformable. The fact that two interactions can have different binding energies is indeed not specific to the study of lattice particles.

We only considered the self-assembly of one type of particles. In several cases, however, different types of particles self-assemble. In Sec. 3.4.1, we will show how the diversity of the self-assembly of several types of aggregate could be investigated. In particular, we show that we do not expect the complexity of the shapes of the aggregate to increase by adding another type of particles.

Finally, the generalization of our results to off-lattice particles could be tested experimentally with the self-assembly of colloids. In Sec. 3.4.2, we show how this could be implemented.

3.4.1 Self-assembly of two types of random particles

Here, we show examples of self-assembly of two types of particles with random interaction, and discuss qualitatively the extra complexity it brings to the one particle case.

If there is not one but two different types of particles within the system (A-B), we need to define three interaction maps, to describe the interaction between the faces of A and A (21 parameters), A and B (36 parameters), and B and B (21 parameters), as was introduced in Sec. 2.5.3. There is now 78 interaction, that we can again draw randomly and independently, with the same parameters as detailed in Sec. 3.1.1. The simulations are now done with a system of 100 particles of each type, and we keep the same total density of particles, which means the system is twice as large.

Examples of equilibrium configurations for each value of affinity and anisotropy are shown in Figure 3.14. The isotropic limit is the same as for one particle: attractive particles self-assemble in bulks and where they have random orientations, and repulsive particles do not self-assemble. We also recover stereotypical formed with one aggregates (liquids ($(\mu, \sigma) = (-4, 15)$, crystals $(-2, 7)$, sponge $(-4, 15)$, fibers $(2, 7)$, crystallites $(4, 15)$, micelles $(-2, 5)$, or oligomers $(0, 13)$). Having two types of particles however, introduces a major difference: the particle can either mix $(-4, 1)$, or phase separate $(-2, 7$. There is however no trivially binary distinction between those two extreme cases, and the situation in $(2, 11)$ is a good example: pink particle occupy the porosity of the sponge of green particles, but also form fibers outside those aggregates. Sometimes, the particles phase separate but still form one dense aggregate $(-2, 7)$, and sometimes each of them form aggregates of lower dimensions, like $(4, 13)$ where the green particles form fibers and the pink particles form trimer. Sometimes, both particles form an aggregate where each particle has an equivalent role (like the fibers of $(0, 15)$ or the crystal of $(-2, 9)$, and sometimes one particle acts as a surfactant to the other (like fibers of $(2, 7)$ or micelles of $(4, 11)$).

It would be hard to use the geometric descriptors we introduced in this chapter to study these problems: the average size or sphericity of all the aggregates in the system would not mean much when the particles form distinct types of aggregate, like in $(4, 13)$. It is also difficult to assign one category of aggregates to one system and do classification. One could use classification with multiple categories [105]. In future studies, we could however measure the propagability of the interactions. In particular, we could measure the propagability of the individual particles, and the effective propagability resulting of the two types of particles. It is not clear however, how those quantities could be compared to the measured propagability for the aggregates of one type of particles.

Apart from the 78 interaction energies of the particle, we could also tune the ratio between the number of both particle. This might influence the size of the aggregates when one particle acts as a surfactant.

From the examples of Figure 3.14, it is not clear however that adding another type of particles enables to introduce new categories of aggregates. The additional complexity arises from the combinatorics of the aggregates (several categories in the same system), but there is no aggregate shapes that were not already observed in Figure 3.1.

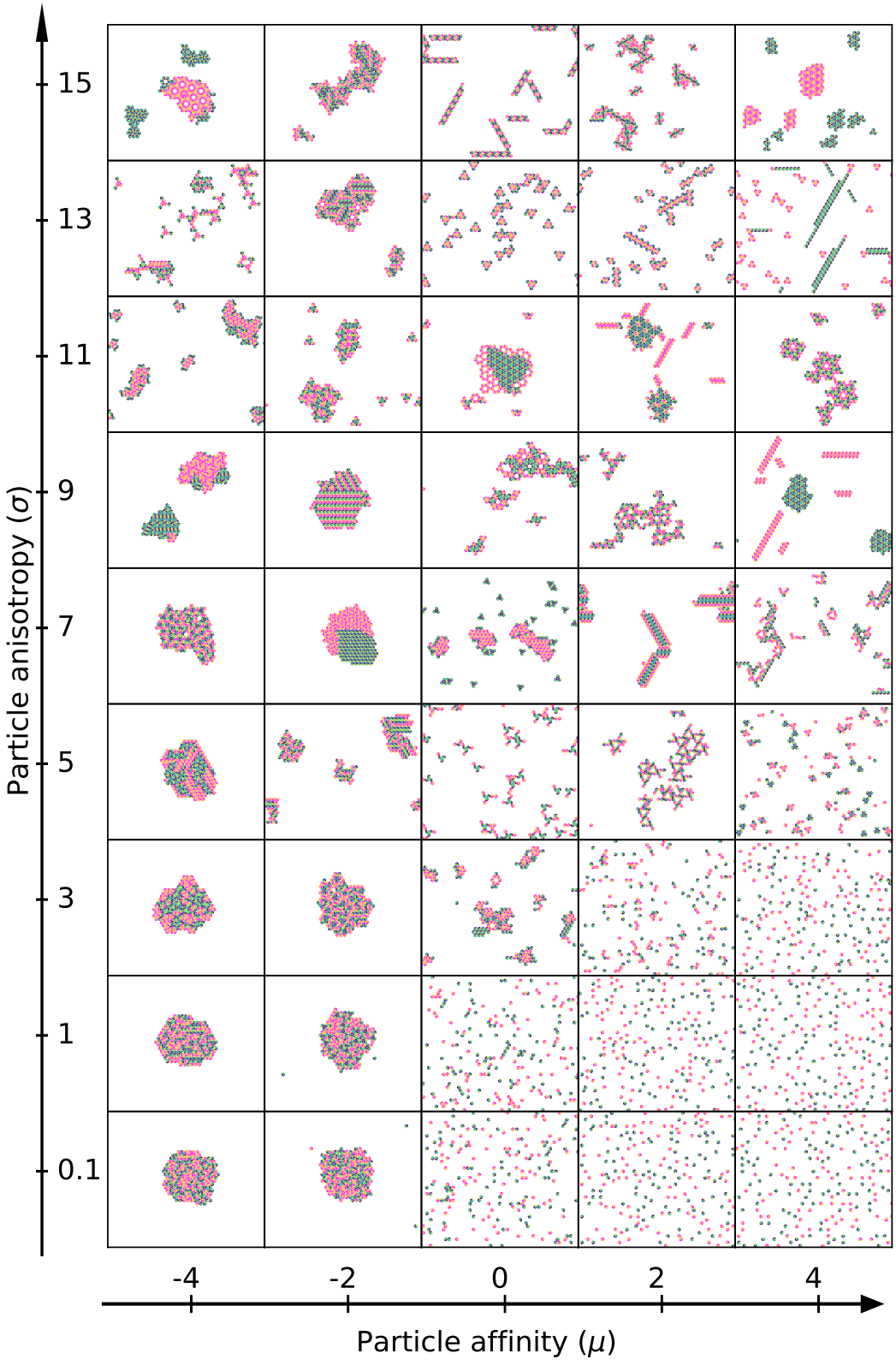

Figure 3.14: Diversity of aggregates formed by two types of particles. We show one snapshot per value of affinity and anisotropy

3.4.2 Experimental applications with triangular particles

It would be interesting to test whether the diversity of aggregate shapes observed numerically would also emerge in experimental system, with colloids for instance. Here, we explain how directional interactions can be set up for colloidal particles, and show how studying two types of particles could provide a large enough design space to observe how particles accommodate frustration.

A preferential set-up for this could be to study the self-assembly of 3D printed colloids on a substrate. This technique is developed by our collaborators at ESPCI, O. du Roure, J. Heuvingh and Mayarani M. The colloid are printed with a resolution of 150 nm, and interact through depletion interaction. In Figure 3.15a, we show electron microscope images of such 3D printed colloids, and in (b) we show images of two colloids interacting: on the left, the particles are far apart, then they interact by increasing the surface contact between their flat faces, and remain attached to reduce the excluded volumes of the *depletants* (small solutes) in the solution. This is the depletion interaction. It is then possible to modify the face of the colloids with complementary notches, to make the interaction specific. This is illustrated in Figure 3.15c and d: the particle has complementary notches on its faces, such that it will preferentially interact when their relative orientation is that of example (1). Examples (2), (3) would also interact, with lower interaction energy, because less surface are in contact. Interaction (4) would be forbidden, because of steric hindrance of the two particles.

There is then a size trade-off to consider: the smaller the particles, the faster they will diffuse on the substrate, and the faster they will self-assemble. On the other hand, there is a low limit to the size of the face of the particle for the notch to be printed with sufficient precision. For this reason, triangular particles are more promising than hexagonal particles, because they can be smaller with the same precision for the notch of the face. We showed in Chapter 2 how we could implement self-assembly of triangular particles, even if the triangular shape is not the dual of a regular lattice. Those simulations can then be used in parallel of experiments, to explore the design space and select examples worth testing in experiments.

Triangles have three faces, and only $3 \times 4/2 = 6$ distinct pairs of faces, *i.e.* 6 different values in the interaction map. This might not be enough to achieve complex enough aggregates, such as aggregates of reduced dimensionality. It is for this reason that we consider the self-assembly of two types of triangle together. This enables both to have small enough particle that will equilibrate in short times, and to reach a sufficient complexity in the interactions (21 pair of faces).

In Figure 3.15e and f, we show simulation results of two pairs of particles, where the colors for the faces of the triangle correspond to Figure 3.15b: yellow is the flat face that interact to itself, and light blue and dark blue are the *lock* and *key* faces that interact with each other. In (e), the particles have respectively flat-lock-key faces, and lock-lock-key faces. There is a dense organization of aggregate where both particles are together. In (f), the particles have flat-flat-lock and lock-lock-key faces. Then, there is no such dense packing of the particles anymore, the first particle form oligomers surrounded by the second particle, while the second particle mostly form aggregates alone.

Considering a pair of triangular particles with notches enables to explore an important design space (particles can have any combination of the three types of faces), and the strength of the interaction could be tuned by increasing or decreasing the size of the notch. The work is still in progress, but the goal is to examine how those triangular particles would avoid frustration by forming aggregates of lower dimensions, or porous aggregates, in experiments. In this chapter, we found that aggregates of particles with incompatible interactions avoid frustration by reducing their dimensionality. If such aggregates were

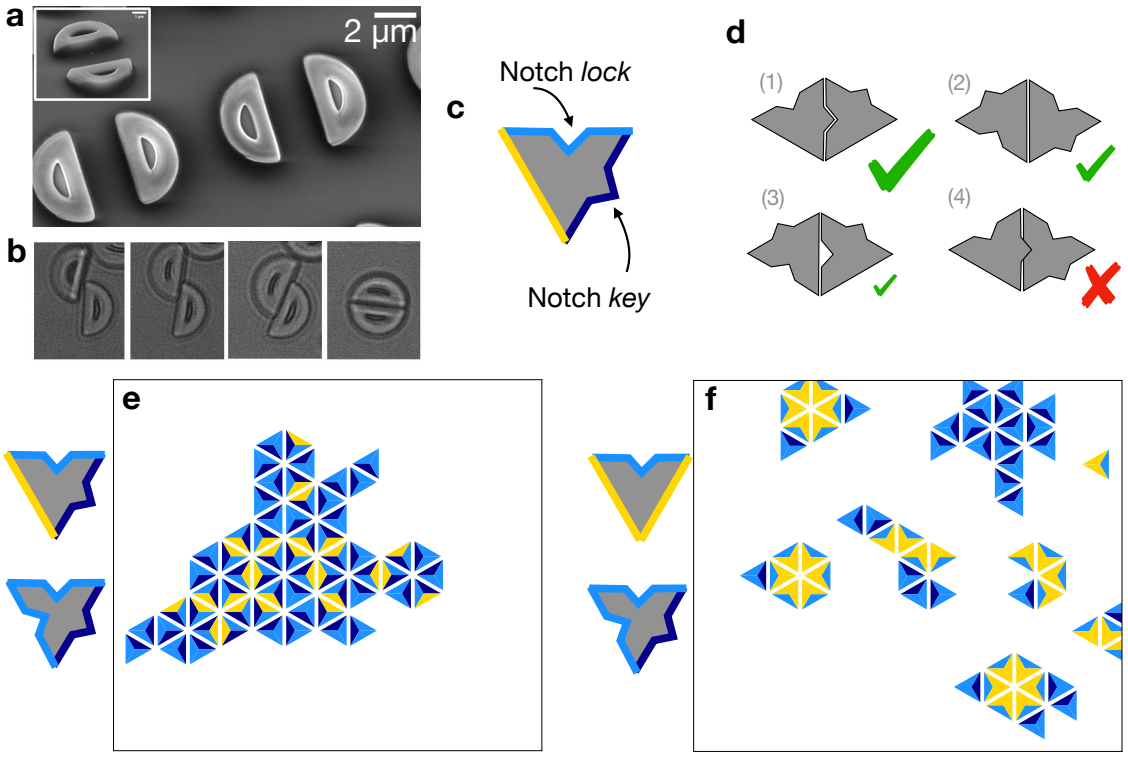

Figure 3.15: Self-assembly of two triangular particles, for experimental implementation. 3D printed colloids can interact through depletion interaction (a,b, courtesy of Mayarani M), and triangular particles could be designed with notches to enable directional interactions (c,d). We show simulation results for the self-assembly of two pairs of particles in (e) and (f). In (e), the two types of particles are mixed in a dense aggregate, and in (f) they are separated. For each simulation, there are 20 particles of each type in a 30×30 lattice, and the interactions energy are, in units of kT, $J_{\text{flat}-\text{flat}} = -8$, $J_{\text{lock}-\text{key}} = -10$, $J_{\text{flat}-\text{lock}} = 2$, $J_{\text{lock}-\text{lock}} = 2$, $J_{\text{key}-\text{key}} = 10$ and $J_{\text{flat}-\text{key}} = 10$.

observed experimentally for triangle colloids with incompatible interactions, it would be an indication that our findings describes a phenomenology that is more generic than the self-assembly of lattice particles in numerical simulation. There might be several limitations arising from these experiments, such as kinetic limitations of the assembly, or aggregates emerging from partial interactions between regions of the particles that were not designed to interact. If self-limitation of the assembly arises, it would require cautious verification that these limitations are not driven by kinetic effects.

4 - Renormalization of anisotropic particles self-assembly models

In Chapter 3, we explored the diversity of self-assembly resulting from a lattice model of anisotropic particles. We discovered the appearance of non-trivial structures such as the sponge or the micelles. The definition of categories of aggregates relied on their macroscopic characteristics, such as the size and dimensionality of the aggregates, and on the level of local order in the particles. We used machine-learning, which is a phenomenological method, to identify interaction maps that would lead to similar aggregates. In this chapter, we aim at understanding what are the typical aggregates with real-space renormalization group. We expect that some canonical aggregates described in Figure 2.4, such as the liquid or the crystal, are representative of broader universality classes, and that aggregates within this class have the same properties, from a coarse-grained view. We call *basin of attraction* of a fixed-points the ensemble of interaction maps that are renormalized to that fixed-point. In this chapter, we identify the fixed-points of the renormalization, and the common features of the interaction maps within the same basin of attraction. We use renormalization as an exploration tool of the 21-dimensional parameter space. This is illustrated in the schematic of Figure 4.1: the fixed-points are specific points in the parameter space, and all the interaction maps within the basin of attraction of a fixed-point are renormalized to this fixed-point. Here, we explain the principles of renormalization and show how we set a numerical decimation procedure that conserves the shape of an aggregate (Sec. 4.1). This renormalization rely on solving the inverse problem of determining the interaction map for a given density map numerically. We explain how this is implemented in Sec. 4.2. We then sample the parameters space with random interaction maps to identify 3 types of fixed points: the interaction maps of isotropic non-interacting particles, isotropic attractive particle, and particles that form crystals (Sec. 4.3). We show that the interaction maps in the basin of attraction of the gas lead to very diverse aggregates, while that in the basin of attraction of the liquid and the crystalline fixed-points all correspond to aggregates of infinite size. We also show that the fiber is an unstable fixed-point of our renormalization scheme (Sec. 4.4). Finally, we show that liquid, gas, and crystalline fixed-points are stable, by a measuring small deviations in the interaction map at the vicinity of the fixed-points (Sec. 4.5).

4.1 Renormalization: from Ising to anisotropic lattice models

Real space renormalization was originally introduced for lattice spin systems, with the aim to compute the critical exponents of a phase transition, using the system scale invariance. The objective of this study is different: we aim at exploring the phase diagram, relying on the same ideas of scale invariances. The model we study is less trivial than the Ising model, and we have very little prior knowledge about the critical points of the system. Therefore, we do not use real-space renormalization to study the behavior of the system near the critical point, but to discover what the critical points are, and which portions of the parameter space will renormalize to a given fixed-point. We recall the main concepts of real-space renormalization (Sec. 4.1.1), and show how our model of lattice particles with directional interactions can be renormalized according to the same ideas (Sec. 4.1.2).

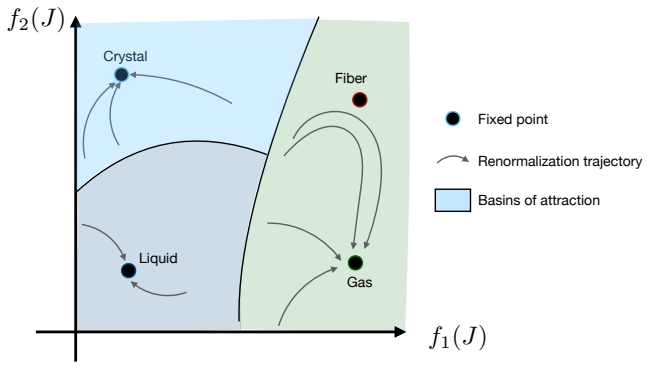

Figure 4.1: Renormalization can be a tool to explore complex phase diagram. f_1 and f_2 are functions of the interaction map. We identify the fixed-points and some features of their basins of attraction. We did not manage to identify a two-dimensional projection of the 21-dimensional where the boundaries of the different basin of attractions is identifiable. We expect the fiber to be an unstable fixed-point, such as the ferromagnetic phase in Ising model, in one dimension. Trajectories will be slowed down near this fixed-point, before renormalizing to the gas.

4.1.1 Traditional and modern use of the renormalization group

In the context of spin systems, the goal of renormalization was to derive critical exponents, *i.e.* to determine with which scaling of the temperature the macroscopic properties of the system vary near the phase transition. In Sec. 4.1.1.1, we recall which principles the renormalization procedure relies on, and emphasize why this cannot be trivially applied to the model we study in this thesis. It is possible to take advantage of numerical simulation to implement a renormalization transformation that cannot be solved analytically. We show examples of such study in Sec. 4.1.1.2. Finally, renormalization was used to study systems that are not Ising models, which requires determining precisely what quantities are expected to be conserved upon renormalization (Sec. 4.1.1.3).

4.1.1.1 Origins of the renormalization

The renormalization group is a statistical physics tool initially introduced to study the behavior of a system near its phase transition. It relies on scale invariance close to the critical point: some characteristics of the system are the same at different length scales. Because of this, it is possible to apply a series of transformation on the system that will integrate out degrees of freedom over the irrelevant short length scales, while conserving some statistical properties unchanged. Renormalization was first introduced by Wilson in 1975 [109], based, among others, on the ideas developed by Kadanoff on the Ising model [110].

If a system is described by its Hamiltonian H, the renormalization transformation is $H' = \mathcal{R}(H)$ where H' is applied on a new set of coarse-grained variables. For example, in the one-dimensional Ising model, the Hamiltonian reads $H = K \sum_{i=1}^{N} s_i s_{i+1}$, where the s_i describes the state of spin i (-1 or 1), and the coupling between the nearest neighbors depends on K. One can then average over every second site (the gray sites in Figure 4.2a) and describe the system with a new Hamiltonian $H' = K' \sum_{i=1}^{N/2} \sigma_i \sigma_{i+2}$. The new coupling K' is chosen such that the partition function $Z(K, N) = \mathrm{e}^{-H/kT}$ of the new system is proportional to that of the initial system: $Z(K, N) = A \times Z(K', N/2)$. We introduce the variables σ for the spins with pair indices that will be decimated (gray spins of Figure 4.2a).

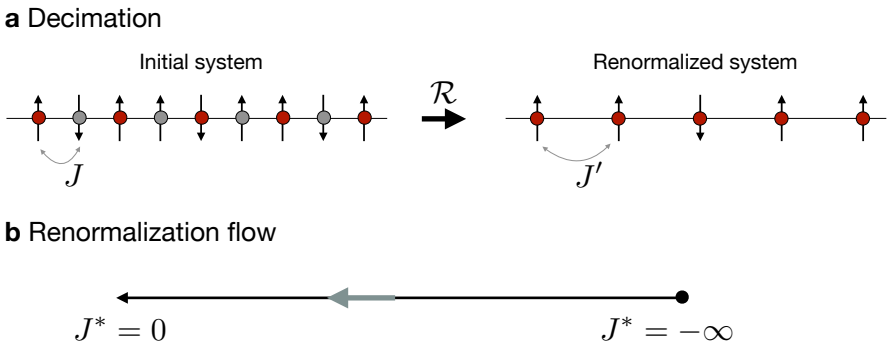

Figure 4.2: Renormalization scheme of the one-dimensional Ising model. a) Decimation procedure: the gray sites are decimated, and J' is the coupling of the new variables (which are in red). b) Renormalization flow, there are two fixed-points, $J^* = 0$ is unstable and $J^* = -\infty$ is stable. The gray arrow indicates the direction of the renormalization flow.

The partition function is computed by summing over all the possible spins configuration.

$$Z(K, N) = \sum_{\{s\}} \sum_{\{\sigma\}} \exp[-K(s_1\sigma_2 + \sigma_2 s_3)]... \tag{4.1}$$

$$Z(K, N) = \sum_{\{s\}} \exp[-K(s_1 + s_3)] + \exp[+K(s_1 + s_3)]...$$

$$AZ(K', N/2) = \sum_{\{s\}} \exp(-K' s_1 s_3)... \tag{4.2}$$

We then solve (4.1) = (4.2) for the possible values of the pairs (s_i, s_{i+2}) we get the following renormalization relations

$$K' = \frac{1}{2} \ln(\cosh(2K)) \text{ and } A = 4\cosh(2K) \tag{4.3}$$

The fixed-points of the renormalization are such that $K^* = \mathcal{R}(K^*)$, and they correspond to the physical critical point of the system (in that case, $K^* = 0$, or $K^* = -\infty$, see Figure 4.2b).

For a two-dimensional Ising model, there is no exact renormalization transformation. For example, the new spins are chosen by a majority rule on a group of three spins called plaquette [111]. The new spins interact through two-body interactions, such as the initial spins, but also through interactions of a larger order, that need to be approximated. Setting a renormalization transformation then requires the following choices:

- identifying the variables to renormalize. In the spin systems, one can choose a model with energy assigned to single particles (K_1), a pair of neighbor particles (K_2, denoted K in the example above), groups of three interacting particles (K_3), etc. The variables to renormalize is then the set of coupling parameters K_p, with p the number of spins in interactions. In practical implementation, this is often limited to K_2 [111]. In our model on the triangular lattice with particles of $n = 6$ orientations, we limit our study to the interaction map K of the first-neighbors of pairs of particles, which already has $n(n+1)/2$ independent parameters
- expressing the new variables as a function of the old ones: in a 2D spin system, a block of $2n+1$ spins is chosen, and it is assigned with value ± 1 with a majority rule.

The renormalization transformation then results from the conservation of a chosen statistical property. Typically, the partition function Z is conserved. This leads to a

relation between the old and the new coupling, such as the one we derived in eq. 4.3 for the one-dimensional Ising model.

In our case, the second item is particularly difficult. We explained in Chapter 2 that the particles' orientation of the particles is not well-suited to study the model of anisotropic particles. For this reason, it is not clear that the new variables should be the orientations of the particles that were not decimated. We cannot use the majority rule on the orientations of a group of particles.

4.1.1.2 Monte-Carlo renormalization group

Because renormalization is in general not solvable in two dimensions, numerical sampling of the configurations of the system is a useful complementary tool to perform its renormalization transformation. We illustrate how it was beneficial for spin systems.

In the 70s, several studies were dedicated to the evaluation of critical exponents of lattice models with real-space renormalization ideas. The burning question was to establish a relation between $\mu = (K_1, K_2, K_3...)$ and $\mu' = \mathcal{R}(\mu) = (K'_1, K'_2, K'_3...)$, where K_n is the coupling between n spins. Along with analytical studies where the Hamiltonian were manipulated to isolate negligible quantities [111], some studies used a sampling of the configuration space with Monte-Carlo simulation to directly estimate the relation between μ and μ' [112, 113].

In [112], the states of both individual spins and blocks of spins are sampled, under a given set of couplings μ. μ is then evaluated from the sampling of the configuration of the individual spins, as a consistency check, while μ' is evaluated from the sampling of the configuration of the block spin. Those evaluations are performed very close to the fixed-point. The relation between μ and μ' then leads to an evaluation of the critical exponents, close to the exact solution derived by Onsager [114].

Three years later, Swendsen conducted similar numerical renormalization, but without running simulations precisely at the fixed-point [113]. Instead, the behavior at the fixed-point was linearized from measures done at its vicinity. Indeed, since the typical length diverges at the fixed-point, direct simulation of the fixed-point requires large systems, or a severe truncation of the number of coupling constants n.

Our renormalization transformation will be quite different from that of the Ising models that uses blocks of spins. Yet, we will follow the idea of estimating the new couplings from a numerical evaluation of the renormalized variable. In particular, we take advantage of the fact that numerical simulations allow to measure the correlations between particles that are very far, which is not easy in analytical study.

4.1.1.3 Examples of recent use of renormalization

The renormalization group concepts have reached fields beyond the usual statistical physics lattice models. In those systems, one difficulty is to identify how to regroup the new variables such that the properties of the system are conserved, but the local details are averaged. We show an example of those difficulties with the renormalization of complex networks. We also emphasize the difference between real space and momentum space renormalization.

Renormalization group transformation of complex real networks (such as internet, or the airports network) has been performed to identify interactions between different scales. This transformation requires regrouping the nodes of the network while conserving some of the graph's properties, such as the average number of neighbors of the nodes. In [115], placing the nodes of the network on an underlying geometry enabled the authors to introduce a measure of physical distance between the nodes, and use it to regroup the nodes that are close upon renormalization. An alternative approach to graph renormalization regroups the nodes between which the information travels fast [116]. The challenges of

applying real-space renormalization to a new type of problem thus does not only rely on the control of the approximations as before, but also on determining what properties of the system should be conserved, and how to transform the variables while ensuring this conservation.

Finally, it is important to mention that most of the current work using renormalization do the decimation in momentum space (as opposed to real space renormalization). Momentum space renormalization has found applications in recent physical problem that can be characterized with a field theory, such as the collective behavior of natural swarms [117], or epidemiological model [118]. Our lattice model cannot be trivially described by a field theory, and we only use real-space renormalization in this study.

4.1.2 Numerical implementation of the renormalization on lattice anisotropic particles

In this section, we propose a renormalization transformation for the model of directional interactions between lattice particles introduced in Chapter 2. In Sec. 4.1.2.1, we recall the main results derived on the renormalization of this model with analytical method by F. Benoist in his PhD thesis [119]. They suggest that particles with isotropic attractive interactions, and without interactions are expected to be fixed-points of the renormalization of the model, and that the shape of the aggregates should be conserved upon renormalization. In Sec. 4.1.2.2, we propose to choose the density map as the quantity to conserve upon renormalization. In Sec. 4.1.2.3 we give details on how our renormalization transformation will be implemented.

4.1.2.1 Analytical results and limit

In [119], F. Benoist studied the grand-canonical equivalent of the model of particle with directional interaction. In this case, the interactions between the face of a particle and an empty site (which we denoted by J_{a0}) cannot be chosen to be zero, as we showed in Sec. 2.1.3 of Chapter 2. Indeed, the number of particles in the system is not conserved. The one and two-dimensional analytical study of the model emphasized that the infinitely attractive particles and non-interactive particles are expected to be fixed-points of the renormalization, and that the choice of the decimation procedure can influence the result of the renormalization.

In the one-dimensional version of the model, the renormalization was performed analytically by decimating half of the sites, as described for the Ising model in Sec. 4.1.1.1. The particles have n possible orientations. Several types of fixed-points of the renormalization were found, with different values of the empty-full interaction. If the cost of all the empty-full interactions (J_{a0}) is infinite, it was shown that the system is either completely full (all the sites are occupied by a particle) or completely empty (none of the sites are occupied), depending on the full-full interactions values. The fixed-points for which the system is full corresponds to the situation where the full-full interactions are infinitely strong: $J_{ab} = -\infty$ for all values of a and b. In our numerical model, there is a fixed number of particles and the interactions take finite values. As a result, the situation described above corresponds to interactions that are strong enough for all the particles in the system to aggregate. Therefore, we expect that interactions maps leading to aggregates with the maximal number of particles are fixed-points of the renormalization. If the cost of all the empty-full interactions is finite, the only fixed-point is $J_{ab} = 0$, which corresponds to a gas of non-interacting particles. Therefore, we also expect that $J_{ab} = 0$ will be a fixed-point in the canonical ensemble.

In the two-dimensional version of the model, it was not possible to determine the fixed-points analytically. However, F. Benoist showed, using stereotypical aggregate geometries, that the choice of the renormalization procedure is decisive for the conservation of the ag-

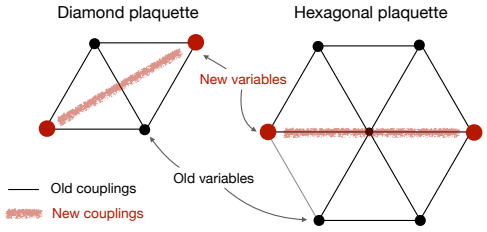

Figure 4.3: Possible choices of decimation on the triangular lattice. Old variables and couplings are represented in black, new variable and coupling in red. The new coupling is determined upon summing of the black and red variables. Figure inspired from [119]

gregate shape upon renormalization. More precisely, the new interaction between particles in orientation φ and ψ is determined as follows: particles in orientations φ and ψ occupy the red sites in one of the plaquettes shown in Figure 4.3. Then, the possible configurations of all the black sites are enumerated (a black site can be empty, or occupied with a particle in any orientation), and the corresponding energies of the plaquette (*i.e.* the sum of the energy of the bonds) is summed, and determines the energy of the new interaction. The choice of the plaquette can be decisive. Let us consider the interaction maps that leads to a sponge aggregate (such as the one presented in Figure 4.4b). When the plaquette is too small, like the diamond plaquette in 4.3, the renormalized interaction maps of the sponge is that of non-interacting particles $J = 0$ (gas configuration) within one step. On the contrary, with the hexagonal plaquette (also shown in Figure 4.3), the renormalized interaction maps of the sponge leads to the aggregation of the particles. Finally, for some specific aggregate geometry, as the fiber, the hexagonal plaquette is still not sufficient to conserve the geometry of the aggregate upon renormalization.

The analytical renormalization could not be performed exhaustively in two-dimensions, but its solving in 1D and the renormalization of some interaction maps in 2D sets some basic requirement for the renormalization procedure: we expect that non-interaction particles, and attractive isotropic particles will be fixed-points of the renormalization. We also expect the renormalization to conserve some geometric properties of the aggregate, if the plaquette on which we sum the configurations is large enough.

4.1.2.2 Decimation of the bonds and infinite plaquette

Here, we explain our chosen decimation procedure, and how the interaction map and density map variables introduced in Chapter 2 are well suited to perform this decimation. We also discuss the approximation that we make in this procedure.

We decimate 3/4 of the sites, and conserve the couplings shown on the hexagonal plaquette in Figure 4.3. In Figure 4.4, we show how this work for three examples of aggregates, the sponge, the fiber, and the micelle. We decimate the first neighbors of a particle, while conserving its relative orientation with its second neighbors. As a result of this decimation, the sponge is renormalized to a crystal where all the particles have the same orientation as their neighbors. A fiber of width 2 is renormalized to a shorter fiber of width 1. A micelle is renormalized to a smaller micelle. Those examples show that the geometric properties of the aggregates, are conserved upon renormalization. The exact organization of the particles is not conserved, but a periodic organization of the particles (such as for the sponge) will remain periodic after renormalization, while a non-periodic organization (such as the micelle), will remain non-periodic.

To perform this transformation, we rely on the density map introduced in Chapter 2, that counts the average occurrence of each pair of faces in the system. We can also count the average occurrence of each pair of second-neighbors. In the sponge example in Figure 4.4b, there are two pairs of faces that are in contact in the initial equilibrium aggregate: the

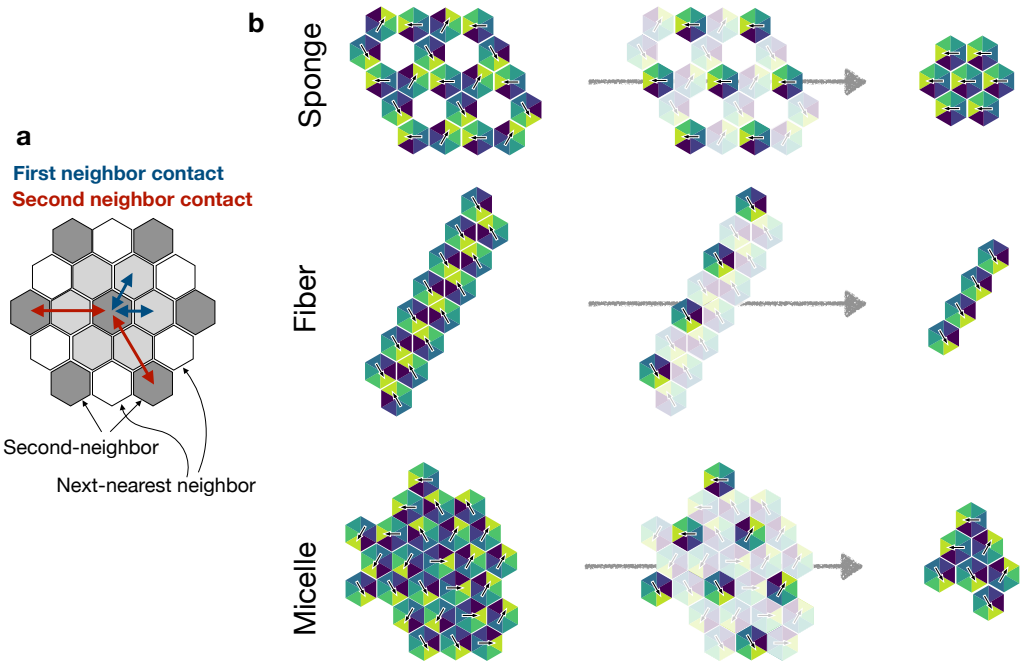

Figure 4.4: In one renormalization transformation, the second neighbors of the particles become the first neighbors. a) The second neighbor contacts are in the same direction as the first, but the distance between the particles is 2 lattice sites. b) We show an example of the renormalization for a sponge, a fiber of width 2, and a micelle. The initial aggregate (left) is renormalized to the new aggregates (right) by conserving the interactions between the second neighbors.

yellow is in contact with the green, and the blue is in contact with the light violet. We can also count the pair of faces that are in a second neighbor contact, *i.e.* for bonds in the lattice direction of length 2 (see the schematic in Figure 4.4a). There are three type of second-neighbor contact with the sponge: yellow with blue, green with light purple, and dark purple with dark green. We ensure that the renormalized aggregate has the same first neighbors contacts as the second neighbor contacts of the initial aggregate. For a given interaction map J, the first neighbor density map $\langle c \rangle_J$ counts the first neighbor contacts. We now introduce the second neighbor density map, $\langle d \rangle_J$ which counts the second neighbor contacts. The renormalization transformation consists in finding a new interaction map J' such that

$$\langle d \rangle_J = \langle c \rangle_{J'} \tag{4.4}$$

We show that unlike the Niemeijer-Van Leeuwen procedure [111], we do not need to make approximations on the size of the plaquette to evaluate the new variables. In Niemeijer-Van Leeuwen renormalization of an Ising model, the new variables are computed by defining a plaquette of for example three spins, and by determining a new variable associated to it, with a majority rule. This is illustrated on top of Figure 4.5a and b, the old spins are $+1$ or -1 (black or gray) and the new spins are the sign of the three spins in the plaquette (red or pink). In our model of particles with directional interaction, there is no meaning of taking the majority rules of the six possible orientations of the particles. Instead, we conserve the bonds' statistics. The old variables are the first neighbor contacts (in levels of gray in Figure 4.5c) and the new variables are the second neighbor contact (in levels of red in Figure 4.5d). The old and new variables are then a list of 28 numbers (we count the full-full (21), empty-full (6) and empty-empty (1) contacts). The Ising renormalization, performed analytically, requires defining a plaquette of finite size, and to neglect the interactions between spins that are not in the same plaquette. In our

renormalization procedure, which we perform numerically, we do not need to make such an approximation and to truncate the range over which the orientations can be correlated: the density map $\langle d \rangle_J$ measures the occurrence of each of the 28 possible configuration of a bond averaged on the whole system, and over a large number of Monte-Carlo steps. In this sense, the plaquette we consider is of the system size.

Similarly to the Niemeijer-Van Leeuwen procedure, we make the approximation that no new types of couplings that emerge after one renormalization step. In the Ising renormalization, Niemeijer and Van Leeuwen considers that both the old and the new variables are solely coupled through the same type of coupling as the old variables: both are coupled through a two-body first neighbor interactions. The other type of couplings, such as those involving more than one particle, or long-range couplings are neglected. Likewise, we assume that the interactions between the particles in the renormalized system only depends on the interaction map J', and that there is no new type of couplings that emerge.

Here and in the rest of the chapter, we call second neighbors of a particle the set of particles that are at a distance of two lattice sites, in the direction of the unit vectors of the lattice. They correspond to the dark gray particles in Figure 4.4a). Technically, those are the third-neighbors of the central particle, while the next-nearest neighbors are the white particles on Figure 4.4a. We choose to conserve the statistics of the contact between the dark gray particles for two reasons. First, the contact between the central particle and its next-nearest neighbors (white particle in the schematic) is ill-defined: they are in contact through a corner of the particle, not a face. Determining the configuration of this contact would require to perform a rotation of both particles, and make the visual interpretation less intuitive. Second, the white next-nearest neighbors are not in the direction of the lattice unit vectors, and because of that, a fiber will necessarily be renormalized to a monomer within one step of renormalization: in the example of the fiber of width 1 on the right of Figure 4.4b, we see that the next-nearest neighbors of a particle in the fiber are all empty site. Instead, our chosen decimation do conserve the fiber upon renormalization.

We defined a numerical renormalization transformation that seem to conserve the geometric properties of the aggregates, and does not require approximation on the size of the plaquette.

4.1.2.3 Renormalization procedure

We describe how the renormalization procedure will be implemented numerically, and show that it ensures that the infinite size of an aggregate will be conserved by the renormalization transformation.

To perform one renormalization transformation of on interaction map J, we follow two steps:

(i) Determine the equilibrium configuration of the system, using a Monte-Carlo simulation, as explained in Chapter 2. Measure the average second neighbor density $\langle d \rangle_J$ at equilibrium.

(ii) Determine the renormalized interaction map J' for which the average first neighbor density at equilibrium, $\langle c \rangle'_J$ is equal to $\langle d \rangle_J$, in a system four times smaller, with four times less particles.

This gives us the renormalization transformation $J' = \mathcal{R}(J)$, which we also illustrate in Figure 4.6 for the sponge aggregate. The initial system is shown in panel (a), and the renormalized system in panel (b). The second neighbor density map of the initial system is identical to the first neighbor density map of the renormalized system. The interaction map is chosen to verify this relation. The key challenge of this renormalization process is step (ii): finding the interaction map that results in a chosen density map.

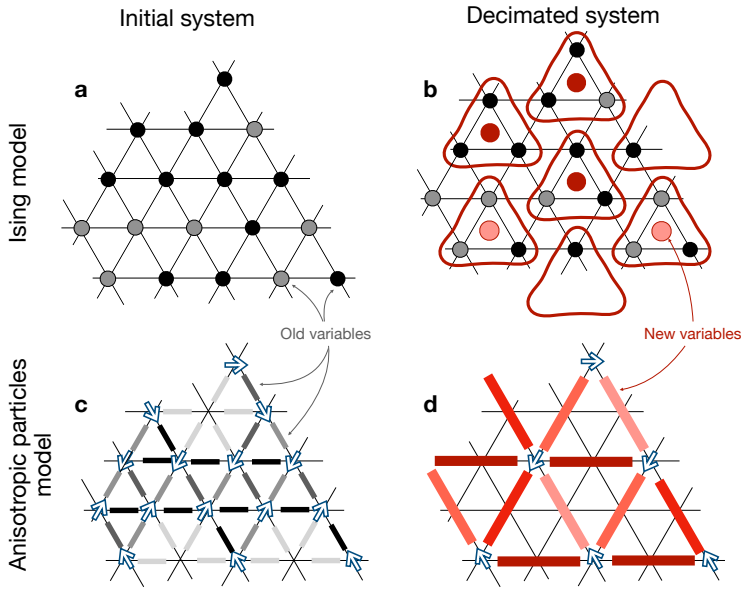

Figure 4.5: In our renormalization process, we do not need to define a truncation of the number of the correlations between two spins to evaluate the new couplings. a) Ising model where each site is either up (gray) or down (black). b) The new variables are the average orientation on a block of three spin, chosen by majority rule. They are either up (pink) or down (red). This is the decimation procedure of the Niemeijer van Leeuwen procedure [111]. c) Anisotropic particles models, particle have different orientation, and a bond is associated with a state that depends on the two particles composing it. Bonds with different levels of gray are in different states. d) The new variables are bonds between particles more distant in the lattice. They also depend on the relative orientations of the particles composing it.

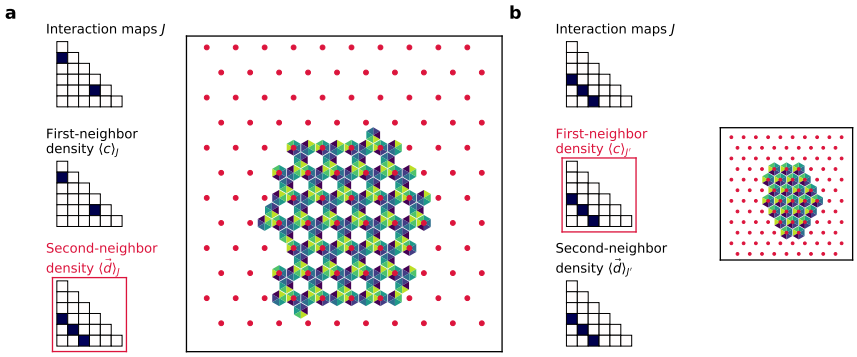

Figure 4.6: The second neighbor density of the initial system is equal to the first neighbor density of the renormalized system. In red, we show the lattice over which the statistics are collected, and the corresponding density matrix is framed in red.

Here, we explain why the density of particles is conserved upon renormalization, as it was verified analytically in the one-dimensional grand-canonical study derived in [119]. In Chapter 2, we showed that the density of particles is proportional to the sum of the empty-full contacts, because the number of particles is conserved. This is also true for the second neighbor density map d. Then, conserving the composition amounts to keeping a fixed density of particles.

We show why the size of the renormalized system should be four times smaller than that of the initial system. This is not intuitive: if the density of particles is conserved, the system size should not be important. Let us consider a system for which the attractive interactions are strong enough that there is only one aggregate in the system, such as the one in Figure 4.6. There, the number of particles in the bulk is $N_{\text{particles}}$, and the number of particles at the surface of the aggregate scales like $\sqrt{N_{\text{particles}}}$. The density map c counts the occurrence of each type of contact relative to the system size. For this reason, the number of full-full contacts scales like the density of particles $N_{\text{particles}}/N_{\text{sites}}$, (eq. 4.5), and the number of surfaces scales like $\sqrt{N_{\text{particles}}}/N_{\text{sites}}$ (eq. 4.6).

$$\sum_{0<a<b} c_{ab} \sim N_{\text{particles}}/N_{\text{sites}} \tag{4.5}$$

$$\sum_{a>0} c_{a0} \sim \sqrt{N_{\text{particles}}}/N_{\text{sites}} \tag{4.6}$$

With our chosen decimation, the number of second-neighbor surfaces is twice that of the number of surfaces: $\sum d_{a0} \sim 2\sum c_{a0}$. If we choose the number of particles in the decimated system as $N'_{\text{particles}} = N_{\text{particles}}/4$, and the system size to be $N'_{\text{sites}} = N_{\text{sites}}/4$ (conservation of the density of particles), we get

$$\sum_{a>0} d_{a0} \sim \sqrt{N'_{\text{particles}}}/N'_{\text{sites}} = (\sqrt{N_{\text{particles}}}/2)/(N_{\text{sites}}/4) \sim 2\sum_{a>0} c_{a0} \tag{4.7}$$

A system of infinite size (*i.e.* all the particles are in the aggregate), will conserve its number of surfaces upon renormalization, and will be renormalized to a system of infinite size. On the contrary, a system of finite size, such as a micelle, does not verify the scaling stated above, and the renormalization will increase the number of surfaces of the aggregate. For this reason, we expect the size of finite aggregates to decrease along the renormalization process.

In this section, we introduced a renormalization procedure that decimates a fraction of the particles, and thus their local organization (the faces in contact in the equilibrium aggregates), but conserves some large-scale properties of the aggregates, such as its geometry (spherical or fibrillar) and the fact that is of finite or infinite size. The renormalization is performed numerically, such that there is no need to make approximations on the length-scale over which the particles are correlated. However, it requires determining the interaction map that will lead to a chosen density map at equilibrium.

4.2 Determination of the renormalized interaction map with gradient descent

To perform one interaction step, we need to determine the interaction map J that gives a given density map c at equilibrium. J and c are of dimension 21. If the Monte-Carlo simulation is a function g such that $g(J) = c$, we need to solve the inverse problem and determine $g^{-1}(c)$. We showed in Chapter 2 that the relation between J and c is non-trivial, and non-linear. We cannot hope to solve this problem analytically. Each evaluation of the function g is also costly, because it requires to run a Monte-Carlo simulation. For those reasons, sophisticated optimization methods such as conjugate gradients are not well suited: they require knowing the full profile of the function in one direction to determine the optimal step size [120]. Instead, we solve the problem numerically with gradient descent, that is a well adapted algorithm for multi-variables non-linear optimization problem [121]. It requires evaluating the gradient of f at each optimization step. We show in Sec. 4.2.1 how we can take advantage of the fluctuation-dissipation theorem valid on the system to evaluate the gradient of g at a given value of J. In Sec. 4.2.2, we then explain how the gradient descent algorithm is implemented. In particular, it required to resort to several optimization tools to ensure the convergence of the algorithm in all cases.

4.2.1 We evaluate the gradient by measuring the fluctuations at equilibrium

We need to compute the multivariate gradient $\frac{\partial c_\alpha}{\partial J_\beta}$, which describes how a small change in the energy level of the face pair α affects the number of bonds in the face pair β. We recall that a face pair corresponds to one value in the interaction map. Instead of computing this number with finite differences, which can be very noisy, we use the fact that our system is at equilibrium at finite temperature, and that it verifies the fluctuation-dissipation theorem: the thermodynamic fluctuations predict the response to a change in the energies. This relation is generic to a large class of statistical physics problem [122] and reads as follows in our situation:

$$\frac{\partial \langle c_\alpha \rangle_{\mathbf{J}}}{\partial J_\beta} = -N_{\text{bonds}} \left\langle (c_\alpha - \langle c_\alpha \rangle_{\mathbf{J}})(c_\beta - \langle c_\beta \rangle_{\mathbf{J}}) \right\rangle_{\mathbf{J}} \tag{4.8}$$

The left-hand term corresponds to the dissipation, and the right-hand term to the fluctuations. The brackets stand for averages over the thermal fluctuations of the system. We showed in Chapter 2 how we could sample the fluctuations of the system at finite temperature. We can then easily measure the right-hand term in the numerical simulation, and evaluate the resulting gradient. We prove that this relation is verified in our model in Sec. 4.2.1.1 and verify that the measure of the fluctuations provide a better evaluation of the gradient than finite differences in Sec. 4.2.1.2.

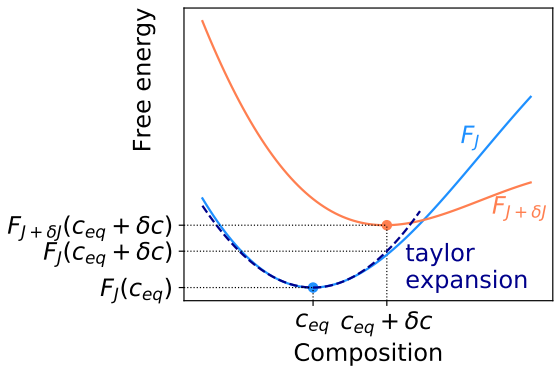

Figure 4.7: The difference between two equilibrium compositions is related both to the fluctuations and the dissipation of the system. We plot a 1D representation of the free-energy $F_{\mathbf{J}}$ and $F_{\mathbf{J}+\delta\mathbf{J}}$, and the Taylor expansion of $F_{\mathbf{J}}$ around its minimum $\mathbf{c}_{\text{eq}}$.

4.2.1.1 Proof

We prove eq. 4.8 by introducing the free energy of the system, and evaluating it for two interaction maps $\mathbf{J}$ and $\mathbf{J}+\delta\mathbf{J}$ around their equilibrium configuration. In this discussion, we only refer to interaction and density maps as vectors.

In Chapter 2, we showed that the energy of the system is simply the scalar product between the density map and the interaction map. We can now write the free energy, F which also depends on the entropy of the system $S(\mathbf{c})$. The entropy counts the configurations that are compatible with the occurrence of bonds measured in $\mathbf{c}$. Here, we are still using the formalism introduced in [93].

$$F_{\mathbf{J}}(\mathbf{c}) = N_{\text{bonds}}\left(\mathbf{J}\cdot\mathbf{c} - TS(\mathbf{c})\right) \tag{4.9}$$

where T is the temperature of the system. We not try to evaluate $S(\mathbf{c})$, which encompasses all the geometric constraints of the particles, but we use the fact that the dependency of F in $\mathbf{J}$ is linear.

We now relate the second derivative of the free-energy to small variations of the coupling and the composition. This is illustrated in Figure 4.7. The equilibrium composition of a system for an interaction vector $\mathbf{J}$ is $\mathbf{c}_{\text{eq}} = \langle\mathbf{c}\rangle_{\mathbf{J}}$. It is the value that minimizes $F_{\mathbf{J}}(\mathbf{c})$ (see blue curve in Figure 4.7). We can Taylor expand the expression of the free energy around this minimum:

$$F_{\mathbf{J}}(\mathbf{c}) = F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}) + \nabla_{\mathbf{c}}F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}).(\mathbf{c}-\mathbf{c}_{\text{eq}}) + \frac{1}{2}(\mathbf{c}-\mathbf{c}_{\text{eq}})^T.\hat{H}_F(\mathbf{c}_{\text{eq}}).(\mathbf{c}-\mathbf{c}_{\text{eq}}) + \mathcal{O}(\|\mathbf{c}-\mathbf{c}_{\text{eq}}\|^3) \tag{4.10}$$

where $\nabla_{\mathbf{c}}F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}) = 0$ by definition of $\mathbf{c}_{\text{eq}}$. $\hat{H}_F$ is the Hessian of F. Because the dependence of F on $\mathbf{J}$ is linear, the Hessian does not depend on $\mathbf{J}$. We introduce $\delta\mathbf{c} = \mathbf{c}-\mathbf{c}_{\text{eq}}$. From eq. 4.10, we obtain the quadratic expansion of $F_{\mathbf{J}}$ around its minimum, represented with the dashed dark blue line in Figure 4.7.

$$F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}+\delta\mathbf{c}) \approx F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}) + \frac{1}{2}\delta\mathbf{c}^T\cdot\hat{H}_F(\mathbf{c}_{\text{eq}})\cdot\delta\mathbf{c} \tag{4.11}$$

We now compute the free energy associated with an interaction vector $\mathbf{J}+\delta\mathbf{J}$ (pink curve in Figure 4.7). From eq. 4.9, it is straightforward that

$$F_{\mathbf{J}+\delta\mathbf{J}}(\mathbf{c}) = F_{\mathbf{J}}(\mathbf{c}) + N_{\text{bonds}}\delta\mathbf{J}\cdot\mathbf{c} \tag{4.12}$$

We evaluate this new free energy $F_{\mathbf{J}+\delta\mathbf{J}}$ around the equilibrium concentration for the initial free energy $F_{\mathbf{J}}$, $\mathbf{c}_{\text{eq}}(\mathbf{J})$ in eq. 4.13. We then replace in eq. 4.14) the evaluation of $F_{\mathbf{J}}$

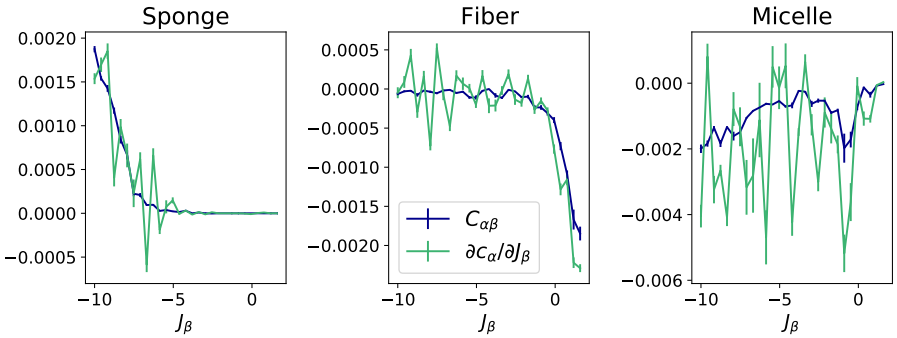

Figure 4.8: Fluctuations provide a more accurate evaluation of the gradient than finite differences. The fluctuation $C_{\alpha\beta}$ (blue) is measured directly in the simulation. The dissipation $\frac{\partial c_\alpha}{\partial J_{beta}}$ (green) is evaluated by finite differences. We compute the averaged value for c_α and $C_{\alpha\beta}$ over 5 systems and 10^4 Monte-Carlo steps per sites. The error bar correspond to the standard error. For the sponge, $\alpha = 11, \beta = 23$, for the fiber $\alpha = 21, \beta = 7$, for the micelle $\alpha = 12, \beta = 12$.

near its equilibrium value with the development introduced in eq. 4.11.

$$F_{\mathbf{J}+\delta\mathbf{J}}(\mathbf{c}_{\text{eq}} + \delta\mathbf{c}) = F_{\mathbf{J}}(\mathbf{c}_{\text{eq}} + \delta\mathbf{c}) + N_{\text{bonds}}\delta\mathbf{J} \cdot (\mathbf{c}_{\text{eq}} + \delta\mathbf{c}) \tag{4.13}$$

$$F_{\mathbf{J}+\delta\mathbf{J}}(\mathbf{c}_{\text{eq}} + \delta\mathbf{c}) \approx F_{\mathbf{J}}(\mathbf{c}_{\text{eq}}) + \frac{1}{2}\delta\mathbf{c}^T \cdot \hat{H}_F(\mathbf{c}_{\text{eq}}) \cdot \delta\mathbf{c} + N_{\text{bonds}}\delta\mathbf{J} \cdot \mathbf{c}_{\text{eq}} + N_{\text{bonds}}\delta\mathbf{J}.\delta\mathbf{c} \tag{4.14}$$

We differentiate eq. 4.14 with respect to $\mathbf{c}$ and evaluate it for $\delta\mathbf{c}$ such that $\mathbf{c}_{\text{eq}} + \delta\mathbf{c}$ is the equilibrium concentration for the new free energy $F_{\mathbf{J}+\delta\mathbf{J}}$. By definition of the equilibrium concentration, this vanishes.

$$\nabla_{\mathbf{c}} F_{\mathbf{J}+\delta\mathbf{J}}(\mathbf{c}_{\text{eq}} + \delta\mathbf{c}) = 0 = 0 + \hat{H}_F(\mathbf{c}_{\text{eq}}).\delta\mathbf{c} + 0 + N_{\text{bonds}}\delta\mathbf{J} \tag{4.15}$$

$$\frac{\delta\mathbf{c}}{\delta\mathbf{J}} = -N_{\text{bonds}}\hat{H}_F^{-1}(\mathbf{c}_{\text{eq}}) \tag{4.16}$$

Finally, we relate the Hessian of F to the measure of the fluctuations. We compute the average of the fluctuations around the equilibrium composition $\mathbf{c}_{\text{eq}}$: $\langle \delta c_\alpha \delta c_\beta \rangle$. Here, we drop the index $\mathbf{J}$ for F. We also replace F by its expansion around $\mathbf{c}_{\text{eq}}$, defined in eq. 4.11. We are left with a multivariate Gaussian integral, for which the solution is well known.

$$\langle \delta c_\alpha \delta c_\beta \rangle = \frac{1}{Z}\int_{\mathbf{c}} \delta c_\alpha \delta c_\beta \mathrm{e}^{-F(\mathbf{c})}\mathbf{dc} \text{ with } Z = \int_{\mathbf{c}} \mathrm{e}^{-F(\mathbf{c})}\mathbf{dc} \tag{4.17}$$

$$= \frac{1}{Z}\int_{\delta\mathbf{c}} \delta c_\alpha \delta c_\beta \mathrm{e}^{-F(\mathbf{c}_{\text{eq}}) - \frac{1}{2}\delta\mathbf{c}^T.\hat{H}(\mathbf{c}_{\text{eq}}).\delta\mathbf{c}} d\delta\mathbf{c} \tag{4.18}$$

$$= (H^{-1})_{\alpha\beta}(\mathbf{c}_{\text{eq}}) \tag{4.19}$$

We have a relation between the Hessian of the free energy at equilibrium and the dissipation on the one hand (eq. 4.16) and the fluctuation on the other hand (eq. 4.19). We proved relation 4.8.

4.2.1.2 Verification

We check that this relation is verified numerically. In particular, we want to verify that we sample fluctuations during a sufficient number of Monte-Carlo steps to obtain a reliable evaluation of the gradient. For this, we compare the gradient measured with the fluctuation with that evaluated using finite differences, and verify that the former is less noisy.

For three examples systems we considered in Figure 4.6 (sponge, micelle, and fiber) that are associated with a given interaction map $\mathbf{J}$, we choose a face pair β that had high concentration at equilibrium, and we vary J_β (Figure 4.8). For each value of J_β, we measure the composition c_α, and the covariance $C_{\alpha\beta} = \langle (c_\alpha - \langle c_\alpha \rangle_{\mathbf{J}})(c_\beta - \langle c_\beta \rangle_{\mathbf{J}}) \rangle_{\mathbf{J}}$ (blue curve on Figure 4.8). We then evaluate $\frac{\partial c_\alpha}{\partial J_\beta}$ by finite differences (green curve on Figure 4.8). We see that those curves match.

The measure of the gradient with fluctuation dissipation theorem is less noisy than that with finite differences. We use the former method to evaluate the gradient at each step of the gradient descent algorithm.

4.2.2 Gradient descent implementation

We want to determine the interaction map $\mathbf{J}$ that would give a chosen density map $\mathbf{d}$ at equilibrium. For any candidate interaction map $\mathbf{J}$, we can evaluate the gradient $\nabla_{\mathbf{J}}(\mathbf{c})$ numerically by measuring the fluctuation of the system. In Sec. 4.2.2.1, we present the generic ideas of the gradient descent algorithm that relies on the introduction of a cost function. We explain our implementation choices and modifications of the algorithm to ensure its convergence: choice of the learning rate (Sec. 4.2.2.2), addition of momentum (Sec. 4.2.2.3) and regulation with the rate with the second moment (Sec. 4.2.2.4). Along these subsections, we show how the cost function evolves along the optimization, for different sets of hyperparameters, and for different objective density maps. We determine the hyperparameters that allow convergence of the algorithm in all the examples situation, that will then be used in the rest of the study.

4.2.2.1 Method for gradient descent

The gradient descent method relies on the definition of a cost function. The gradient of that function is then used to gradually update the variable from which it is calculated (here $\mathbf{J}$). We denote by f the cost function. f is a function of the variable $\mathbf{J}$, it is positive, and it vanishes when the equilibrium density map $\mathbf{c}$ of an interaction map $\mathbf{J}$ is equal to the objective density map $\mathbf{d}$. In practice, we choose

$$f(\mathbf{J}) = \frac{1}{2} \left\| \frac{\langle \mathbf{c} \rangle_{\mathbf{J}} - \mathbf{d}}{\mathbf{c}_0} \right\|^2 \tag{4.20}$$

$\mathbf{c}_0 = \langle \mathbf{c} \rangle_{\mathbf{J}=\mathbf{0}}$ is the composition of a system of identical size and density with no interactions. The normalization by $\mathbf{c}_0$ ensures that we measure relative variations in the density map, and that f will not be dominated by variations in the face pair that are sampled the most. The gradient of f, $\nabla_{\mathbf{J}} f$ is then trivially computed from the multivariate gradient $\nabla_{\mathbf{J}}(\mathbf{c})$ introduced in the previous section.

In its simplest form, the gradient descent algorithm corresponds to an update of the parameter $\mathbf{J}$ at each optimization step by the opposite value of the gradient, multiplied by a coefficient η. In this section, the optimization steps are denoted with the letter t. We choose to initiate the gradient descent at $\mathbf{J} = \mathbf{0}$. The value of the interaction vector is then updated by the following recursion:

$$\mathbf{J}_{t+1} = \mathbf{J}_t - \eta \nabla_{\mathbf{J}} f(\mathbf{J}_t) \tag{4.21}$$

This technique is extensively used to train neural networks, in which case η is called the *learning rate*. As shown in a one-dimensional example in Figure 4.9, the choice of the learning rate highly influences the convergence and the speed of the gradient descent algorithm. If the function f is known and harmonic, an optimal learning rate η_{opt} can be computed such that the gradient descent will take one step to reach the minimum of f (Fig 4.9a). If the learning rate is too small, the algorithm will converge, but the number

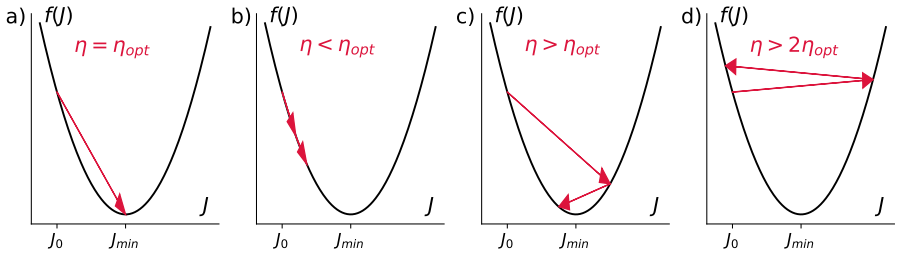

Figure 4.9: Gradient descent success and efficiency depends on the choice of the learning rate. Figure adapted from [105].

of optimization steps will be high (Fig 4.9b). If it is too high, it will oscillate around the minimum (Fig 4.9c). Above a certain limit for η_{opt}, the algorithm will never converge and escape away from the local minimum (Fig 4.9d). In the following subsections, we explain our method for choosing the learning rate.

4.2.2.2 Increasingly smaller learning rate

We choose a time dependent learning rate $\eta(t)$, that will decrease along the optimization process:

$$\eta(t) = \eta_0 \times \eta_1 \times \eta_2^t \tag{4.22}$$

Here, we explain this choice, and how we determine η_0, η_1 and η_2. In particular, η_0 is the order of magnitude of the gradient, while η_1 is a correction to this value. We explain our choice of η_0 and test different values of η_1 and η_2.

To evaluate η_0, we determine the typical scale of the fluctuations of the density map $\delta\mathbf{c}$, and deduce the corresponding scale of the fluctuation of the interaction map $\delta\mathbf{J}$. On the other hand, we determine the typical values of the gradient of f for a zero interaction map, $\nabla_{\mathbf{J}}(\mathbf{0})$. Fluctuations of the interaction map the gradient are related by eq. 4.21, which sets the typical scale of η_0: $\eta_0 = \|\delta\mathbf{J}\|/\|\nabla_{\mathbf{J}}(\mathbf{0})\|$. The learning rate should ensure that the steps to update the gradient are not much smaller or much larger than the fluctuations at zero temperature. This sets an order of magnitude for the learning rate. In practice, $\eta_0 \approx 10^{-2}$.

The exact value for the learning rate then depends on η_1, which is of order 1, and $\eta_2 < 1$, which ensures that the step size will decrease along the optimization process, such that the steps are small close to the minimum of f. The exploration of the hyperparameters η_1 and η_2 on three specific example is shown in Figure 4.10a and b. When the optimization process is easy, (panel (a) for the sponge), the algorithm converges faster for larger values of η_1, because the steps are large, as shown in the qualitative picture of Figure 4.9. However, for less trivial optimization, a small learning step is required to ensure the convergence of the algorithm (in panel (a) for the micelle, the algorithm does not converge for $\eta_1 = 4$). Additionally, when the learning rate is not decreased along the learning process ($\eta_2 = 1$), the algorithm does not converge for the micelle. From the comparison of the evolution of the cost function for different values of η_1 and η_2, we choose $\eta_1 = 0.5$ and $\eta_2 = 0.97$. This is not always the set of parameters that finds the most precise solution, but it guarantees that the total rate is small enough for the algorithm to converge given any target composition.

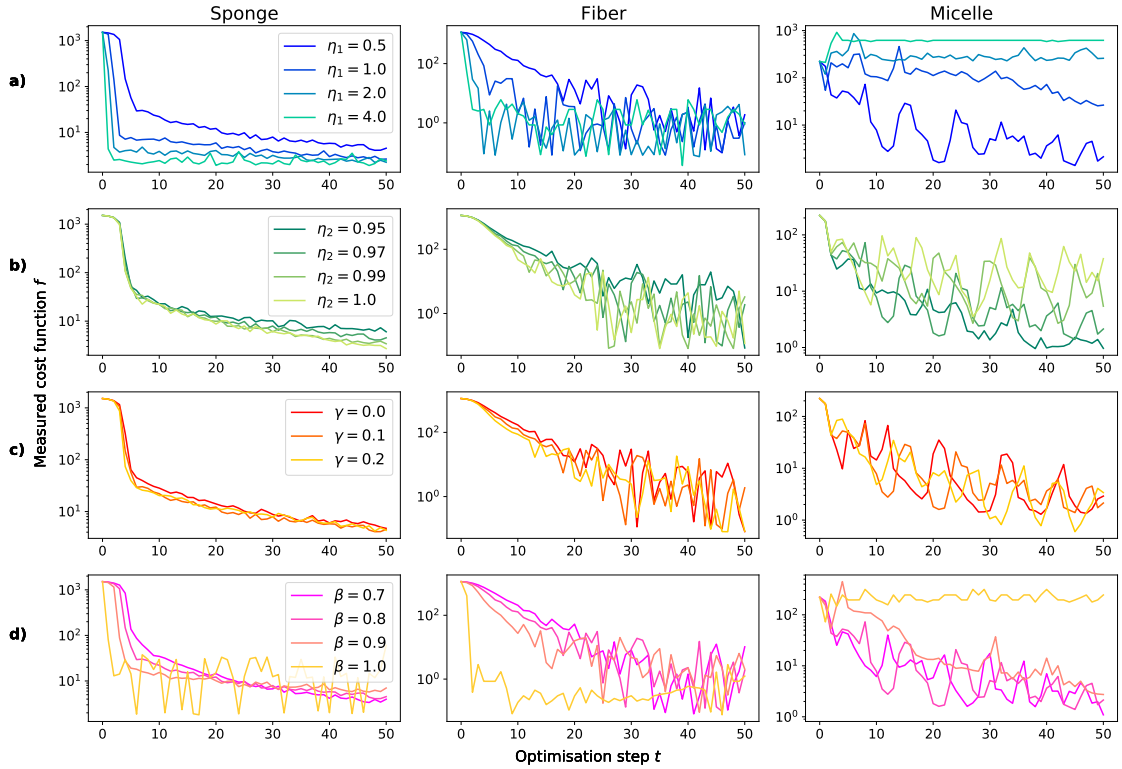

Figure 4.10: Influence of the learning rate on the optimization to solve the renormalization equation. We plot the cost function along the optimization process, to find the renormalized interaction vectors corresponding to sponge (left), fibers (left) and micelles (right). In all optimization process, the default parameters are $\eta_1 = 0.5$, $\eta_2 = 0.97$, $\gamma = 0.1$ and $\beta = 0.8$. a) We vary η_1. When η_1 is large, the initial steps of optimization are faster, but large oscillations of the cost function f can occur later. b) We vary η_2. When $\eta_2 < 1$, the learning rate decreases with optimization step, allowing for less fluctuations of the cost function f. $\eta_2 = 1$ corresponds to the limit case where the learning rate is independent of the optimization step. c) We vary γ. The influence of gamma is not very important. d) We very β. For the micelle, the optimization converges only for small enough values of β.

4.2.2.3 Gradient descent with momentum

In order to limit the time taken for the optimization process, we measure the equilibrium density map at a given optimization step $\langle \mathbf{c} \rangle_{\mathbf{J}}(t)$ with a precision that is smaller than the one we use to measure the objective density map $\mathbf{d}$. Typically, a measure at optimization step t is done on the result of the annealing of $N_{\text{systems}} = 2$ different random initial conditions. This causes the measures to be more affected by fluctuations. To address this limitation, a technique often used is to perform a running average on the value of the measured gradient. This corresponds to the step $\mathbf{v}_t$ used to update the interaction vector $\mathbf{J}_t$ [105].

$$\mathbf{v}_t = \gamma \mathbf{v}_{t-1} + \eta(t) \nabla_{\mathbf{J}} f(\mathbf{J}_t) \tag{4.23}$$

$$\mathbf{J}_{t+1} = \mathbf{J}_t - \mathbf{v_t} \tag{4.24}$$

The momentum parameter γ controls the characteristic time scale over which this average is done ($(1-\gamma)^{-1}$). We test the influence of γ on some specific optimization process as shown in Figure 4.10c. It seems that the choice of γ is not decisive. We choose $\gamma = 0.1$.

4.2.2.4 Regulation from the second moment of the gradient

The density and interaction maps have large dimensions. The behavior in each direction of this 21 dimensional space can be very different: some face pairs show very large density, which will lead to large variations during gradient descent (this is a steep region of the parameter space), while others have low densities (they are never observed) and their corresponding interaction do not vary too much (this is a flat region of the parameter space). To ensure the algorithm convergence, we want the gradient descent to perform large steps in the flat regions, and small steps in the steep regions.

This can be done by rescaling the gradient by its second moment in every direction. This technique is called Root Mean Squared Propagation (RMSprop) [123], and it is used to ensure convergence of the training of a neural network. In practice, the rule to update the interaction vector is now the following (the parameter γ was discussed in the previous subsection).

$$\mathbf{g}_t = \nabla_{\mathbf{J}} f(\mathbf{J}_t) \tag{4.25}$$

$$\mathbf{s}_t = \beta s_{t-1} + (1-\beta)\mathbf{g}_t^2 \tag{4.26}$$

$$\mathbf{v}_t = \gamma \mathbf{v}_{t-1} + \eta(t) \frac{\mathbf{g}_t}{\sqrt{\mathbf{s}_t + \epsilon}} \tag{4.27}$$

$$\mathbf{J}_{t+1} = \mathbf{J}_t - \mathbf{v}_t \tag{4.28}$$

In those equations, the vector division and square are taken element-wise.

We used this method because it leads to spectacular improvement of the convergence of our algorithm, as shown in Figure 4.10d: if $\beta = 1$, this is equivalent to not performing RMSprop. In that case, the gradient descent does not converge in the situation of the micelle. When β is lower than, 1 however, the algorithm converges. In the following, we choose $\beta = 0.8$.

The size of the learning steps (η_0 and η_1) their decrease along the learning process (η_2), their dependence on the value of previous learning steps (γ) and their dependence on a rescaling by the second moment of the learning step (β) were chosen such that the algorithm converges (the cost function is of order 1) in three typical situations.

In this section, we successfully implemented a gradient descent algorithm that, for a given averaged density of each 21 contact, recovers which set of interactions leads to this density at equilibrium. This was possible because the gradient of the density with the interactions was measured from the fluctuations of the system rather than for finite differences. The implementation of a gradient descent algorithm that converges in all cases required to resort to methods used in the machine learning community. For a given interaction map J, associated with an equilibrium density map of the second neighbors $\langle d \rangle_J$, we now determined a way to compute $J' = \mathcal{R}(J)$: J' is the minimum of the cost function f (eq. 4.20) which ensures that the equilibrium density map of the first neighbors of J', $\langle c \rangle_{J'}$ is approximately equal to $\langle d \rangle_J$. Because the initial candidate interaction map for the optimization is $J = 0$, and all the entries are increased or decreased of small quantities along the optimization, until the cost function is small enough, the energies in the renormalized interaction map will never be very large: if the strength of the interactions of J' are sufficient to ensure $\langle c \rangle_{J'} \approx \langle d \rangle_J$, the optimization steps and the energies of J' will not increase further. For this reason, the typical scale of the energies of an interaction maps chosen manually and of the interaction maps resulting from the optimization process can be different.

4.3 Fixed-points identification with random sampling

We introduced a renormalization procedure: for a given interaction map J, we compute a renormalized interaction map $J' = \mathcal{R}(J)$, such that the equilibrium configuration associated with J' is a coarse-grained version of the equilibrium configuration associated with J. We now identify the *fixed-points* of this renormalization, *i.e.* the interaction maps J^* such that $J^* = \mathcal{R}(J^*)$. In Chapter 3, we explored the 21-dimensional space of interaction map by sampling random interaction maps in a Gaussian distribution. Here, we use a set of random interaction maps as initial steps of the renormalization process. For a given interaction map $J^{(0)}$, we repeat the renormalization process and determine $J^{(1)}, J^{(2)}, ...$, such that $J^{(t+1)} = \mathcal{R}(J^{(t)})$. We call this set of $\{J^{(t)}\}$ a renormalization *trajectory*, and t is referred to as the renormalization *step*. We identify the fixed-points as the interaction maps towards which the renormalization trajectories converge. In Sec. 4.3.1, we explain how to study a renormalization trajectory, and its convergence. Then in Sec. 4.3.2, we show that a broad diversity of initial random interaction maps all converge towards only a few-fixed-points, that we characterize. We see that there are three types of fixed-points of the model: the non-interacting isotropic particle (which leads to a gas), the attractive isotropic particles (which leads to a liquid), and particles that assemble into periodic crystals.

4.3.1 Renormalization trajectory

Studying the convergence of the renormalization trajectories is difficult for two reasons. First, an interaction map is a 21-dimensional vector, and it is difficult to visualize. Second, the interaction maps are defined up to a cyclic permutation of the vertices of the particle it corresponds to, as was explained in Sec. 2.3.5 of Chapter 2. For this reason, the Euclidean distance is an ill-defined measure of the similarity between two interaction maps. Indeed, two interaction maps can correspond to equivalent equilibrium configuration, but the element-wise difference will be very large. In this section, we explain the choices we made to circumvent those limitations. We first illustrate the problematic of the permutation-equivalent interaction maps with one example in Sec. 4.3.1.1, and introduce a distance measurement between interaction maps that is independent of the permutation of the interaction map 4.3.1.2. Because this still does not solve the difficulty to visualize trajectories, we introduce a projection way of projecting interaction maps in a two-dimensional graph in 4.3.1.3.

4.3.1.1 We repeat the renormalization step until reaching a fixed-points

Here, we show an example of renormalization trajectory, and explain why some trajectories can be periodic, unless we apply a well-chosen permutation to the interaction maps.

In Figure 4.11, we show the interaction map at successive steps of the renormalization process. At each renormalization step t, we show the interaction map $J^{(t)}$, and the first-neighbor and second neighbor density map $\langle c \rangle_{J^{(t)}}$ and $\langle d \rangle_{J^{(t)}}$ a with the convention defined in Chapter 2. These maps are the equilibrium density map measured with Monte-Carlo simulation. We also plot one image of an equilibrium configuration at each step. We can first verify that the second-neighbor density map at step t, $\langle d \rangle_{J^{(t)}}$ is approximately equal to the first-neighbor density map at step $t+1$: on the Figure, the second matrix in a column is identical to the third matrix in the previous column.

The organization of the particles in the aggregate in each image appear similar, however these similarities are not easily identifiable in the interaction maps. Looking at the density maps, however, it seems that the first neighbor composition at step $t = 0$ is similar to the first neighbor composition at step $t = 3$, and similarly between $t = 1$ and $t = 4$. There seem to be a periodicity in the measured renormalization trajectory.

We consider a one-dimensional example to understand the origin of periodicity in the renormalization process. Consider a chain of particles that can be in three states, which we denote a, b and c. We renormalize it by conserving the statistics of the second neighbor. Initially, the particles can be arranged with the following motif *abcabcabc*. Upon renormalization, the organization of the particles will be *acbacbacb* (every second particle in the previous motif). After one extra renormalization step, the organization of the particles is again *abcabcabc*. In this situation, we measure a periodicity in the equilibrium motifs at successive steps of the renormalization, but the physical rules that dictate the equilibrium organization of the particles at each of the steps are similar.

In Chapter 2, we introduced a permutation matrix M such that, two interaction matrices J and $\tilde{J}$ are equivalent through a transformation of their lines and columns: J and $\tilde{J} = M^k J M^{-k}$. For the example of Figure 4.11, we perform such a permutation of the interaction, and first and second-neighbor density maps. For steps $t = 1, 4, 7$, we take $k = 2$ and for steps $t = 2, 5, 8$, we take $k = 4$. We plot the permuted interaction and density maps in Figure 4.12. Now, the interaction maps at different steps of the renormalization are comparable, and we can follow how each individual interaction is renormalized. For instance, the first diagonal term (top left entry) of the matrix corresponds to an interaction that is more and more repulsive along the renormalization process (from $t = 3$) and the

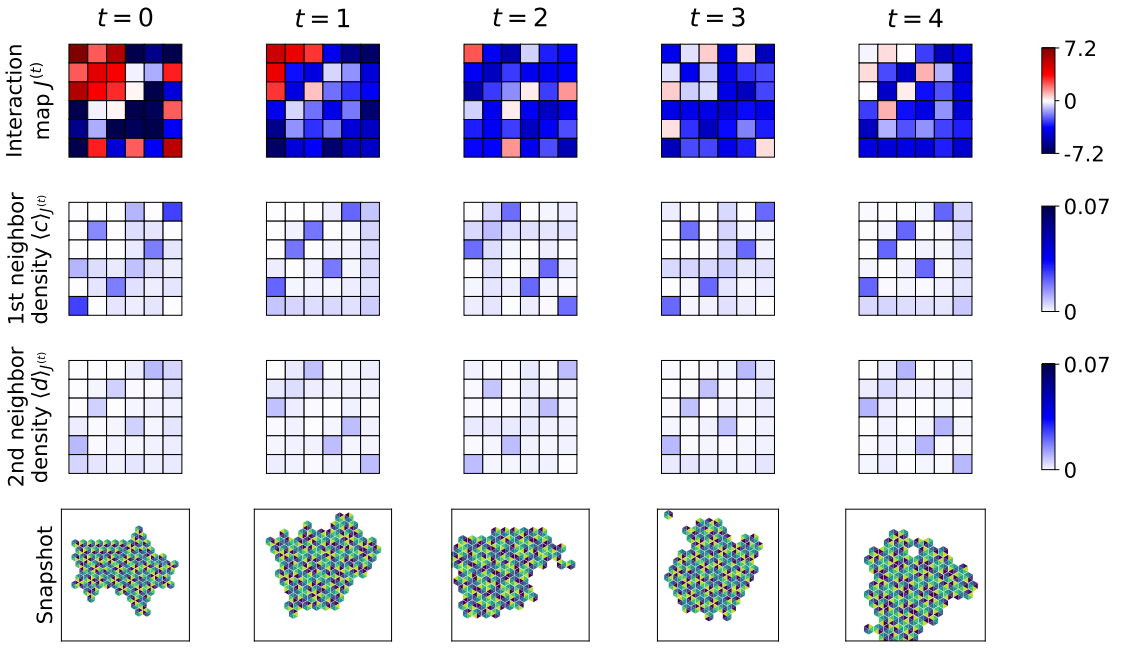

Figure 4.11: Similar pattern arise periodically along the renormalization trajectory. The organization of the particles, and the density map are identical at renormalization step $t = 0$ and $t = 3$ and $t = 1$ and $t = 4$. We verify that the density matrix of the second neighbors at step t is approximately equal to the density matrix of the first neighbors at step $t + 1$. The color scales for the interaction map (in unit of kT) and the density maps are plotted on the right. The initial interaction map is drawn in a Gaussian distribution of mean $0kT$ and standard deviation $3kT$. The snapshots are zoomed in on the aggregate, the systems density is $1/9$.

second diagonal term is a more and more attractive interaction.

In this example, we found manually which permutation should be applied at which step of the renormalization to make the renormalization trajectory understandable. However, such method cannot be used on a large number of renormalization trajectories.

4.3.1.2 Distance between two interaction maps

Here, we introduce a pseudo-distance between two interaction maps J and $\tilde{J}$. We explain why it serves our purposes of identifying fixed-points, but does not verify all the requirements for a distance.

We define the distance function $\mathcal{D}$ such that

$$\mathcal{D}(J, \tilde{J}) = \min_{k \in [\![-6,6]\!]} \|(M^k J M^{-k}) - \tilde{J}\| \tag{4.29}$$

We recall that M is the 6×6 matrix that performs one cyclic permutation of the lines and the columns of J. We compare all the permutations of J and choose the one such that J is closest to $\tilde{J}$. In this definition, we choose the matrix representation of the interaction maps: J and $\tilde{J}$ are matrices.

This distance is positive, the distance between two identical interaction maps is zero, and it is symmetric (it is equivalent to permute the entries of J or $\tilde{J}$). However, it does not verify the triangular inequality, which we show now. Consider three interaction maps J, L and N. We define k_l (respectively k_n) as the number of permutation of J to make its distance with L (respectively N) minimum. Similarly, k_0 is the number of permutation of L that minimizes the distance between L and N. We then write the distance between L and M, and introduce the term $M^{k_n} J M^{k_n}$ that is the permutation of J with minimal

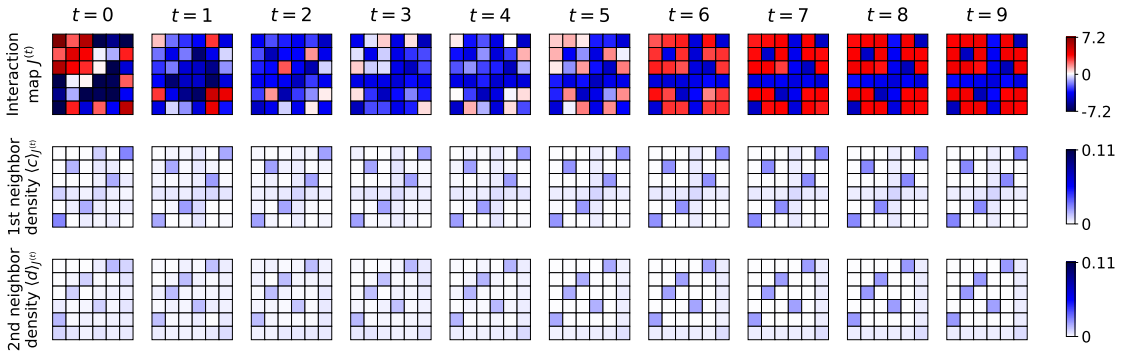

Figure 4.12: After doing a permutation of the interaction and density maps at certain steps, we observe the convergence of all the entries of the interaction map towards an attractive or a repulsive interaction. At steps $t = 3n + 1$ (respectively $t = 3n + 2$), we performed two (respectively, four) cyclic permutation of the interaction matrices of Figure 4.11. At the end of the renormalization process, all the interactions are either repulsive or attractive with the same strength (only two colors in the interaction map at $t = 9$).

distance to N.

$$\mathcal{D}(L, N) = \|M^{k_0} L M^{-k_0} - N\| \tag{4.30}$$

$$\mathcal{D}(L, N) = \|M^{k_0} L M^{-k_0} - M^{k_n} J M^{k_n} + M^{k_n} J M^{k_n} - N\| \tag{4.31}$$

We use the triangular inequality on the norm of matrices, and rewrite the first term of eq. 4.32 by permuting both matrix $-k_0$ times (in the other direction of cyclic permutation).

$$\mathcal{D}(L, N) \leq \|M^{k_0} L M^{-k_0} - M^{k_n} J M^{k_n}\| + \|M^{k_n} J M^{k_n} - N\| \tag{4.32}$$

$$\mathcal{D}(L, N) \leq \|M^{k_n - k_0} J M^{-(k_n - k_0)} - L\| + \mathcal{D}(J, N) \tag{4.33}$$

Here, we see how the triangular inequality could not be verified: there are no guarantees that $k_n - k_0 = k_l$. In words, the permutation of J that minimizes its distance to N is not the same as the permutation that minimizes its distance to L.

This measure is not relevant to compare interaction maps that are very different, however, it gives a good quantitative prediction of the proximity between similar interaction maps. Moreover, there is no ambiguity on the measure of a very small distance between J and $\tilde{J}$. A measured small distance means that both matrices are almost equal, up to a permutation. Upon studying convergence of the renormalization trajectories towards the fixed-points, we measure the evolution of the distance between the interaction map and the fixed-points along the successive steps of the renormalization.

4.3.1.3 Spectrum of the interaction maps

We also need an absolute representation of the renormalization trajectories, that does not require defining a reference point, as the distance introduced in 4.3.1.2. The eigenvalues of a matrix are independent of the permutation of its entries. Here, we show how the eigenvalues of a pseudo-transfer matrix computed from the interaction map can be interpreted. Therefore, their measure will provide a relevant space of lower dimensionality to plot the renormalization trajectories. We compute the spectrum of some canonical examples.

Here, we introduce the pseudo-transfer matrix T, because it is possible to give a physical interpretation of its eigenvalues and eigenvectors. Instead of comparing the spectrum of the matrix J_{ab} where a and b refer to the **faces** of the particle in contact, we study the spectrum of the transfer matrix $T_{\varphi\psi} = \mathrm{e}^{-J_{\varphi\psi}}$ where φ and ψ refer to the **orientation**

of the particle if the contact is in the horizontal direction. We also consider the case where there are no particles: T_{00} is associated with the energy of a contact between two empty sites, and $T_{\varphi 0}$ to a contact between a particle in orientation φ and an empty site. We also normalize T by the sum of its terms, such that $T_{\varphi\psi}$ can be interpreted as the probability of observing a particle in orientation φ next to a particle in orientation ψ. We call this matrix a pseudo-transfer matrix, because transfer matrices are usually only defined in one-dimension, though with the same conventions. We denote by $(p_m)^i_{m=0\ldots7}$ the vector of probabilities of each particle orientation at site i. Then $\mathbf{p}^j = \tilde{T}\mathbf{p}^i$ is the vector of probabilities of each orientation at site j knowing the orientation of a particle at a neighbor site i, if there is no third particles in contact with both i and j. We can now diagonalize the matrices T. Contrarily to the interaction matrix for the faces (J), the interaction matrix for the orientations are not symmetric, and they are diagonalized in $\mathbb{C}$. If a probability of state $\mathbf{p}$ is associated with zero eigenvalues, it means that if a site has probabilities $\mathbf{p}$ of having each orientation, its neighbor site cannot have the same probabilities. Inversely, a vector of probability associated with a non-zero eigenvalue can be observed for two neighboring particles.

We give examples of this interpretation for two simple examples, the isotropic particles without interaction, *i.e.* the gas configuration (all the entries of J_{gas} are zero) and the isotropic sticky particle, *i.e.* the liquid configuration (all the entries of J_{liquid} are $-\epsilon$) with ϵ the strength of the attractive interactions. We compute the corresponding pseudo-transfer matrices T_{gas} and T_{liquid}. T_{gas} has only one non-zero eigenvalue, associated with the probability vector, where all the configuration of the particle have the same probability (including the empty configuration). There is only one way the particles can be ordered locally, and it is the gas configuration. T_{liquid} has also one non-zero eigenvalues, which we interpret as follows: the sites are either all occupied or all empty, and therefore there is also only one way the particles can be organized. In the numerical simulation, because the number of particles is fixed, we observe both the dense and the empty phase in the same system.

In the following, we represent renormalization trajectories in the space of the norm of the eigenvalues of the pseudo-transfer matrix. For instance, we plot renormalization trajectories in the $(|\lambda_i|,|\lambda_j|)$ space with λ_i the norm of the i^{th} eigenvalue of the pseudo-transfer matrix. In this projection, the coordinates of a point will not depend on the permutation of the interaction map. It is also possible to give a partial interpretation of a coordinate being zero in these plots: an eigenvector associated with a zero eigenvalue is a set of probabilities of the orientations of the particles that two neighboring particles cannot have.

In this section, we emphasized the difficulties of comparing renormalization trajectories, and we identified two complementary strategies to tackle the question: we can project the renormalization trajectories in the space of the eigenvalues of the pseudo-transfer matrix, and we can measure distances between two interaction maps by enumerating all the permutations of a transfer matrix and choose the one that minimizes the Euclidean distance.

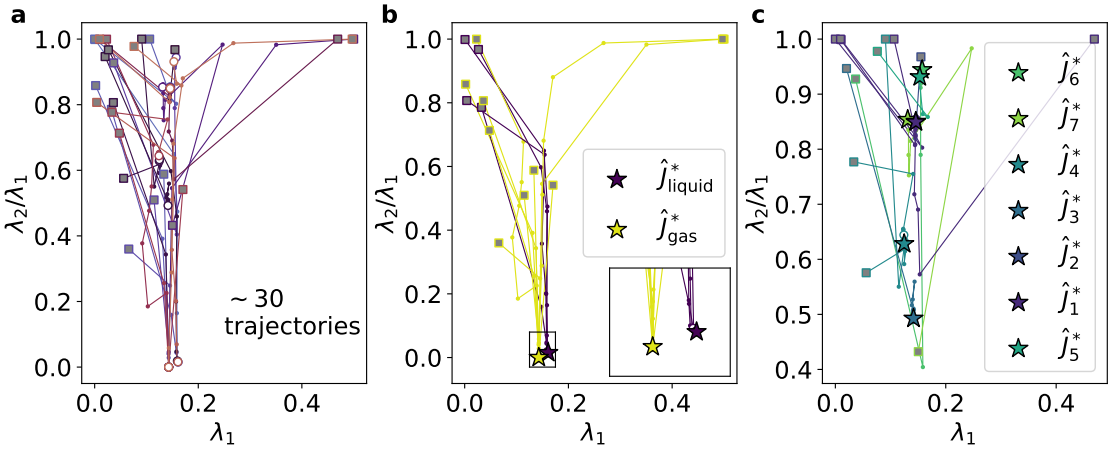

Figure 4.13: The renormalization trajectories seem to converge towards a few fixed-points when plotted in the projected space of the eigenvalues of the transfer matrix. The initial point of each trajectory is indicated with a gray square, and the final point with a white circle. a) All trajectories are plotted with a different color. b) and c) Colors correspond to a fixed-point that is assigned as described in eq. 4.34. b) We plot the trajectories renormalizing towards the J^*_{gas} and J^*_{liquid}. c) We plot the other trajectories. In this projection, initial interaction maps can be close and final interaction maps far, and vice-versa.

4.3.2 We identify fixed-points from statistical sampling

In this section, we identify stable fixed-points of the renormalization. We see that the renormalization trajectories starting from a large set of initial random interaction maps converge towards a few points J^*. These interaction maps correspond to aggregates with stereotypical shapes. In particular, the isotropic infinitely attractive particle and the isotropic non-interacting particle are fixed-points of the renormalization. The interaction map of some crystal aggregates with periodic organization of the particles are also fixed-points. With our statistical approach, we also show that some characteristics such as the affinity and anisotropy of the particles are partially conserved by the renormalization. In Sec. 4.3.2.1, we detail what are the stable fixed-points of the model and how we found them. In Sec. 4.3.2.2, we characterize the basin of attraction of those fixed-points with the measure of distance between interaction map introduced in Sec. 4.3.1.2. In Sec. 4.3.2.3, we show that the renormalization method is complementary with the machine learning method developed in Chapter 3 to classify the aggregates.

4.3.2.1 The trajectories converge towards a few fixed-points

We identify the stable fixed-points of the renormalization as the interaction maps towards which a large number of renormalization trajectories converge. Here, we explain what initial interaction maps we choose to renormalize, and show to which fixed-point they converge upon renormalization.

We sample 375 initial conditions by drawing the interaction map in a Gaussian distribution of average μ and standard deviation σ. We choose $\mu \in [-4, -2, 0, 2, 4](kT)$, $\sigma \in [3, 5, 7, 9, 11](kT)$, and we draw 15 interactions maps per couple (μ, σ). We consider systems of size 30×30 with 100 particles. The annealing protocol is similar to the one described in Chapter 3. We perform 10 successive numerical renormalization steps between $t_0 = 0$ and $t_f = 9$. If we find $J^{(t)} = J^{(t+1)}$, we stop the computation before reaching the 10^{th} step ($t_f < 9$). The hyperparameters of the optimization process are the one chosen in Sec. 4.2.2. We can plot the renormalization trajectories in the space of the eigenvalues of the pseudo-transfer matrix introduced in 4.3.1.3. In Figure 4.2, we showed the renormal-

ization trajectory of the 1D Ising model: it was converging towards the fixed-points $K = 0$. Here, we expect something similar: the trajectories in the parameter space should converge towards the fixed-point of the renormalization, even for the two-dimensional projection of a 21-dimensional space. Some examples of such trajectories are plotted in Figure 4.13a. The initial point of each trajectory is a gray square, and the final point is a white circle. We observe the expected convergence of the trajectories towards some fixed-points. In Figure 4.13b, we plot only the trajectories for which the final renormalization step verifies $\lambda_2 = 0$. The inset shows that they converge towards two well distinguishable points, the yellow and the purple star. Here, we only show the projection in the subspace $(\lambda_1, \lambda_2/\lambda_1)$. By looking at the projection of the trajectories along the other eigenvalues, we identify 9 fixed-points of the renormalization. This identification is done manually, but we justify it *a posteriori*. Some trajectories converging towards the other fixed-points are shown in Figure 4.13c, and the position of the identified fixed-points in this projection are blue and green stars.

We first characterize the identified fixed-points in Figure 4.14. For each of them, we show the interaction map, the first, and second neighbor density maps, the eigenvalues of the transfer matrix, and snapshots of an equilibrium configuration. We first see that there are two fixed-points for which the interactions take only one value, *i.e.* isotropic particles. When all the interaction are zero, the equilibrium configuration is a gas. We call this the gas fixed-point J^*_{gas}. When all the interaction are attractive, the equilibrium configuration is a dense aggregate in which particles have random interactions. We call these aggregates liquid in the previous chapter. This corresponds to the liquid fixed-point J^*_{liquid}. The seven other fixed-points correspond to aggregate with periodic organization of the particles. We recover some stereotypical aggregates of the Chapter 3, such as the crystal where all the particles are aligned (J^*_1) or the sponge (J^*_4). J^*_2 is the interaction map of particles that aggregate in a periodic crystal of period 3, and J^*_3 is a sponge for which the holes are filled with a particle of random orientation. The equilibrium configuration for J^*_5 and J^*_6 are not easily identified in the image, but from the interaction and density map, it is very clear that these correspond to particles with only two types of interactions, the favored and the repulsive ones. J^*_7 is a crystalline structure where there is a favored disclination line. This disclination line is sampled for both the first and second-neighbors density, and is conserved upon renormalization.

We check that those points verify $J^* = \mathcal{R}(J^*)$: the density matrix of the first and the second neighbor should be equal. We see in Figure 4.14 that this is true for the fixed-points we identified, up to a permutation of the second-neighbor density matrix corresponding to J^*_2, J^*_3 and J^*_4. This can also be understood from the snapshots: the orientations of the first and the second neighbors of a particle in the aggregates are on average identical (again, up to a rotation of the second neighbors in situation 2, 3 and 4).

The projection in the subspace of the eigenvalues of the transfer matrix enables us to identify the fixed-points. However, from that representation, it is not possible to identify which regions of the parameter space should converge to one fixed-point or the other. In particular, some trajectories start very close and converge towards different fixed-points, while some other start very far and converge towards the same fixed-point.

This method enables to find some fixed-points, and to identify *a posteriori* that they have the correct behavior. However, it does not guarantee that we identified all the fixed-points of the model. Indeed, we identified only the one towards which the measured trajectories converge. We try to sample the initial interaction maps to renormalize as broadly as possible, by resorting to random interaction maps drawn in a Gaussian distribution. All the 375 trajectories converged towards only 9 fixed-points. We found that the fixed-points are either interaction maps of isotropic particles (gas and liquid) or anisotropic particles

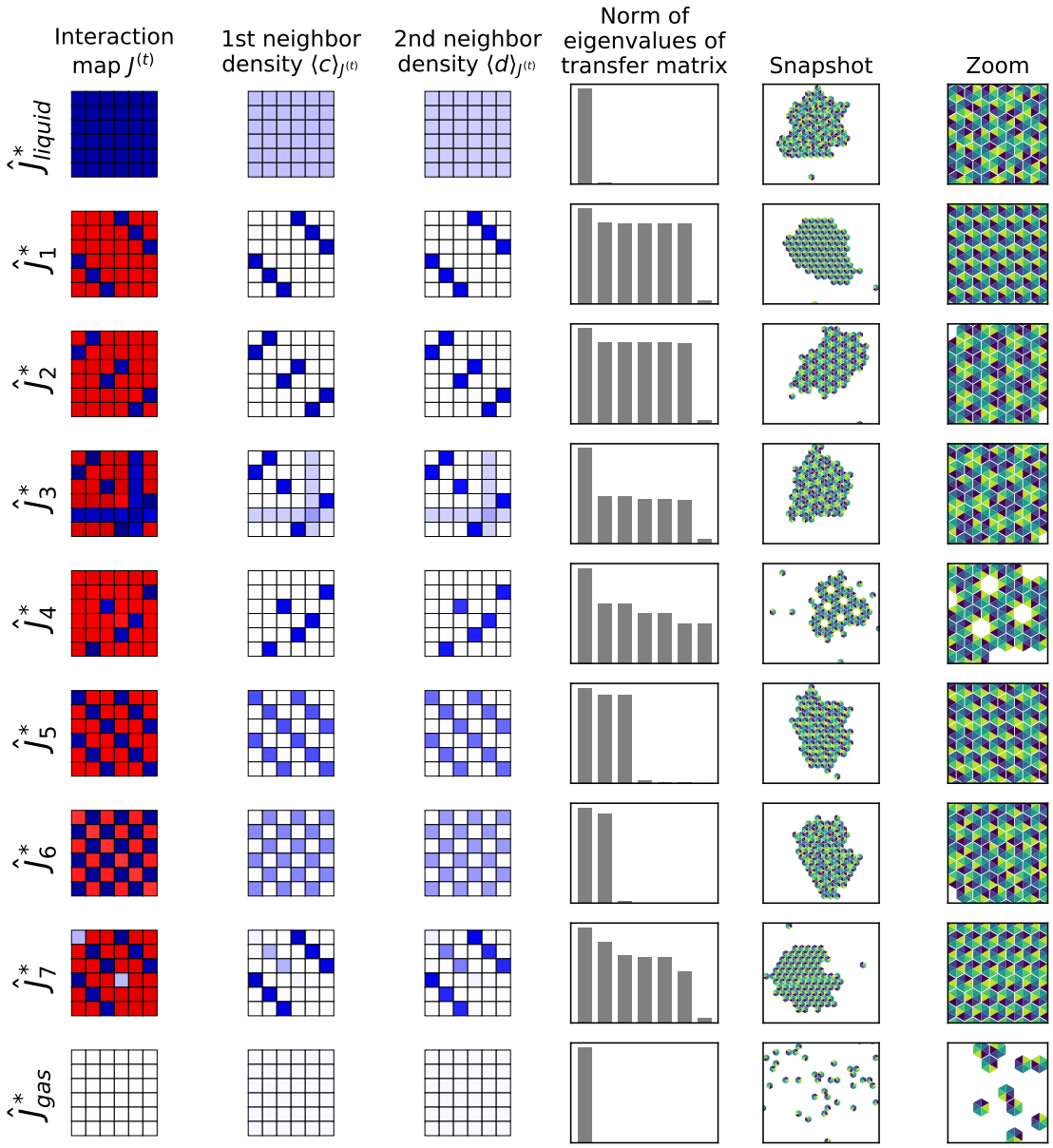

Figure 4.14: Fixed-points of the renormalization have identical first and second neighbor densities. Most of them correspond to periodic organization of the particles. The color scale for the interaction map is such that red corresponds to positive coupling, blue to negative coupling, and the darkest colors correspond to $\pm 6kT$. For the density matrix, blue correspond to non-zero measured densities, and darkest blue is 0.04. The scale of the eigenvalues is linear, with the highest eigenvalue scaling around 0.1 for all the fixed-point. For fixed-points 2,3 and 4, the second-neighbor density matrix has been permuted as described in 4.3.1.2 to be comparable to the first-neighbor density matrix

that tile the plane in a periodic way. In the case of the crystalline fixed-points, the interaction maps are such that there are only two or three levels in the interaction energies, even if the interaction energies at the beginning of the renormalization were random number drawn in a continuous distribution. Beyond those three types of fixed-points, we cannot think of other stereotypical behavior that should lead to an equality between the first and the second neighbor density, yet, we cannot prove it.

4.3.2.2 Basin of attraction

We identified fixed-points of the model. We now characterize their basins of attraction. We recall that the basin of attraction of a fixed-points J^* corresponds to a subspace of the parameter space in which all the points will renormalize to J^*. The projection of Figure 4.13 is not well adapted to characterize the basin of attraction of the fixed-points, because it did not show separations between the trajectories converging towards different fixed-points. Here, we show how we can use the measured renormalization trajectories to get a list of interaction maps that are in the basin of attraction of the fixed-points. We expect this list of interaction map to be representative of the basin of attraction, because they were sampled starting from a large number of different random interaction maps.

We recall that a trajectory in the renormalization space is a list of interaction maps $\{J^{(t)}\}$ such that $J^{(t+1)} = \mathcal{R}(J^{(t)})$. We assign all the interaction maps in the same trajectory to the fixed-point it converges to. Given a trajectory $\{J^{(t)}\}$, we assign it to the fixed-point J^*_k by determining to which fixed-point the final step of the trajectory is the closest. This also requires that this distance is below a threshold $\mathcal{D}_{\min}$. We choose to be $\mathcal{D}_{\min} = 1kT$, which we justify below.

$$J^*_k = \underset{J^*}{\operatorname{argmin}} \left(\mathcal{D}(J^{(t_f)}, J^*) \right) \text{ if } \mathcal{D}(J^{(t_f)}, J^*_k) < \mathcal{D}_{\min} \tag{4.34}$$

In Figure 4.13b and c, the trajectories are colored as the fixed-point they converge to, determined by eq. 4.34.

For each trajectory, we can now plot the evolution of the distance between the interaction map at each step t of the renormalization to the fixed-point: $\mathcal{D}(J^{(t)}, J^*)$. If the fixed-points we identified manually in Sec. 4.3.2.1 are correct, we expect the trajectory to converge to the fixed-point, *i.e.* we expect that the distance to the fixed-point will decrease with t and converge to zero. If the fixed-points were just arbitrary interaction maps identified because of an artifact of the projection, we would rather observe large measures of distance to the fixed-point along all the trajectory, except for the last step, which is below $\mathcal{D}_{\min}$ because of eq. 4.34. The evolution of the distances are plotted in Figure 4.15. We do not plot the evolution of the distance towards J^*_5, J^*_6 and J^*_7, because there is less than 5 trajectories attracted towards these fixed-point. In each panel, we show the evolution of the distance of the trajectories to the fixed-point they are attracted to. We do observe the expected convergence of the trajectory towards the fixed-point. We also observe that this convergence is often fast: after a few renormalization step, the interaction map is very similar to that of the fixed-point. This also confirms that the choice of $\mathcal{D}_{\min}$ is not a fine-tuned optimization: the measured distance is well below $\mathcal{D}_{\min} = 1kT$ for several steps before reaching the fixed-point.

With our method of numerical simulation, we cannot identify boundaries in the parameter space that separate the region that will converge to the same fixed-point, because this space is 21-dimensional. However, we can sample this space by drawing random interaction maps. For each trajectory, we are ensured that all the interaction maps $J^{(t)}$ within this trajectory are in the basin of attraction of the fixed-point J^* it converges to. Indeed, there is no memory involved in the computation of the renormalization process: if $J^{(t)}$ belongs to a trajectory that converges to J^*, it is in the basin of attraction of J^* independently

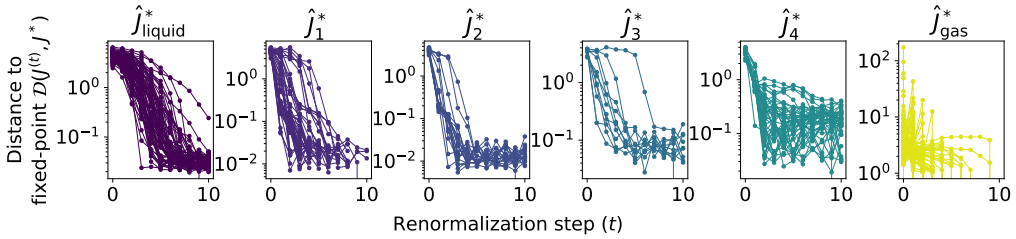

Figure 4.15: The distance to the fixed-point decreases along the renormalization trajectory. The distance is measured from the formula 4.29, and one panel shows the distance evolution of the trajectories that renormalize towards the same fixed-point.

of the initial point $J^{(t_0)}$. Therefore, for each fixed-point, we now have a list of interaction maps that are in their basin of attraction.

4.3.2.3 The renormalization partially conserves the affinity and anisotropy

We can now study the common characteristics of the interaction maps in the same basin of attraction. In Chapter 3, we introduced two ways to characterize an interaction map: the affinity and anisotropy of the particles, and the category of aggregates it belongs to (liquid, crystal, sponge, fiber, micelle, crystallite, oligomer, and monomer). Here, we show that affinity and anisotropy are not sufficient to discriminate in which basin of attraction a given interaction map is.

We first determine how often a random interaction map will renormalize to each fixed-point. We expect this to depend on the affinity and anisotropy of the initial random particle. The expectation in the low anisotropy case is simple: attractive particles aggregate in dense bulks with unordered orientations of the particle, and we expect them to renormalize to the liquid fixed-point J^*_{liquid} where all the interactions have the same energy. Similarly, repulsive particle do not aggregate, and we expect them to renormalize to the gas fixed-point J^*_{gas} where all the interactions are zero. We show the statistics on our dataset of 375 initial interaction maps in Figure 4.16. The gas is the fixed-point of 243 trajectories, the liquid of 73 trajectories, while the crystalline fixed-point together attract 221 trajectories. Only 11 trajectories were unclassified with our method, and they seem to correspond to crystalline structures, but there were not several trajectories converging towards them, and we did not include them in our list of fixed-points. The affinity and anisotropy projection was not sufficient to discriminate between aggregates of different categories in Chapter 3, and it is not sufficient either to discriminate between aggregates in different basin of attraction. Yet, it enables to verify the expected tendency for particles with low anisotropy (bottom of the diagram), that renormalize to a gas when repulsive, and to a liquid when attractive. From anisotropic particle (top of the diagram), the affinity of the particle seem to be important: anisotropic repulsive particles mostly renormalize to a gas, and anisotropic attractive particle either renormalize to crystals or liquid. When the particle is attractive (left of the diagram), the more anisotropic it is, the more it renormalizes to a crystalline structure instead of a gas (less and less dark violet while going up). Those observations suggest that the shape and size of the aggregate is decisive of the fixed-points it will renormalize to.

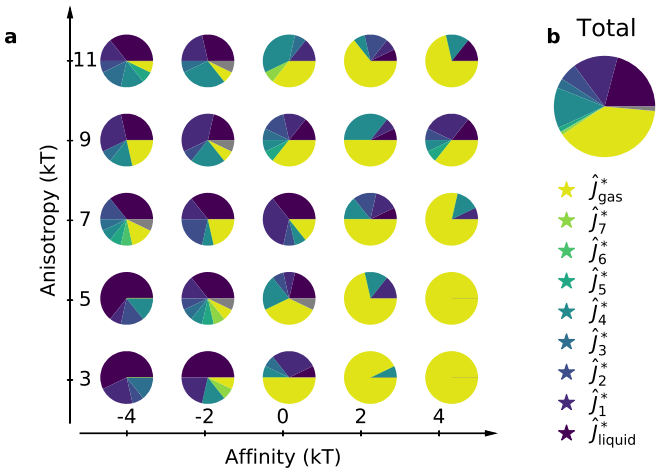

Figure 4.16: Repulsive particles are more renormalized as gas, attractive particles more as liquids. For a given affinity and anisotropy of a random particles, we plot how often the renormalization to each of the fixed-point was observed. Each pie in (a) corresponds to the initial step of 15 renormalization trajectory. The gray color corresponds to unidentified fixed-points. b) The total distribution of the fixed-points (375 trajectories).

4.3.2.4 The renormalized aggregates only fall into a few categories

Here, we see that aggregates in the same basin of attraction have common geometric properties. The renormalization provides a new classification procedure that we can compare to the machine-learning classification introduced in Chapter 3. We can use the classifier trained on the data of Chapter 3 for these new dataset of interaction maps, density maps, and geometrical descriptors. The interaction maps and descriptors measure at the initial step of the renormalization are exactly comparable: the interaction map was drawn in the same distributions (Gaussian) and the annealing protocol was the same, such that the density map and geometric descriptor measured on the equilibrium configuration should be comparable. It is less clear that we can use this classifier on the interaction maps computed as a result of the optimization process: mathematically, they are comparable objects, but we did not train the neural network on this type of interaction map.

We first discuss the classification of the random matrices that are the initial points of the renormalization trajectories. We expect dense aggregates such as crystals and liquids to renormalize towards the liquid fixed-point, or one of the crystal fixed-point. Indeed, if the particles can assemble in an aggregate of infinite size, the first, and second neighbors of a particle are occupied (unless they are at the boundary of the aggregates, which we discussed in section 4.1.2.3). Similarly, the aggregates that have finite size have too many boundaries, and they should renormalize to the gas fixed-point. More qualitatively, an aggregate of finite size, from a coarse-grained view, is a single particle without interaction. The fibers should also renormalize to a gas fixed-point. We show the occurrence of each category of aggregate for the initial step of renormalization in each basin of attraction in Figure 4.17, top. The main trend is the one we expected: liquid, crystal and sponge are mostly in the basin of attraction of the liquid and the crystalline fixed-points, while aggregates of small sizes, and fibers are mostly in the basin of attraction of the gas. The case for the micelle and the crystallite are less clear, and they are spread in the basins of attraction of all the fixed-points. This is because the aggregates in those categories are very heterogeneous, and sometimes have large sizes, but they are two different of crystals or liquids to be classified as such. Except for those aggregates that are hard to classify, the renormalization behaves as expected: infinite aggregates remain infinite along the renormalization, and small aggregates renormalize to aggregates of size 1. Most of

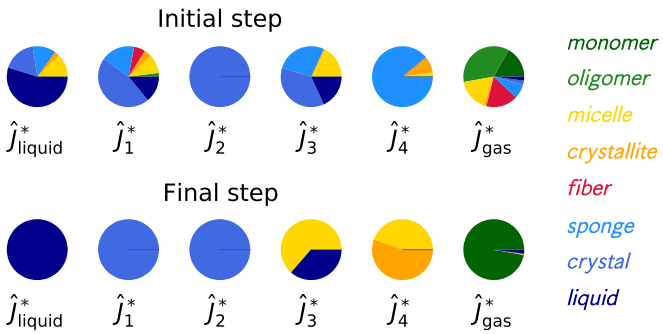

Figure 4.17: All the categories of aggregates are observed at the initial step of renormalization, but only a few of them are after 10 iterations of the renormalization process. The fixed-points J_l^*iquid, J_1^*, etc were introduced in Sec. 4.3.2.1, and the aggregate categories correspond to those introduced in Chapter 3.

the crystals conserve their crystalline organization upon renormalization, and the liquids mostly remain unorganized.

We can also determine the occurrence of the aggregate categories among the interaction maps that are the final step of the renormalization process. The results are plotted Figure 4.17, bottom. The aggregates within the same basin of attraction belong to the same category, which confirms that the renormalization selects features relative to the shape of the aggregates. All the final steps of renormalization for the trajectories in the basin of the liquids, are liquids, those in the basin of the crystalline fixed-points J_1^* and J_2^* are crystals, and those in the basin of the gas fixed-point are monomers. The results are unexpected for J_3^* and J_4^*: those are crystalline structures (with holes for J_4^*) and we would expect them to be classified as crystals or sponge.

We explain the unexpected classification of the interaction and density maps in the basin of attraction of J_3^* and J_4^*. As can be seen in the images of Figure 4.14, the aggregates at equilibrium are at the limit of crystallization, and the energy of their attractive interaction is just enough for the aggregate to form, but some monomers remain detached from the aggregates. This comes from the way we determine the renormalized interaction map with gradient descent. We change the interaction map by small quantities in the opposite direction until the cost function is low enough. For this reason, as soon as the interaction energy is strong enough for the particles to interact, the optimization process will stop. In that sense, the interaction maps are fine-tuned. The training of the neural network was not done on such fine-tuned interaction maps at the limit of the crystallization process. This is why those aggregates are classified as micelles, crystallite, or liquids: they do not correspond to stereotypical aggregates of the dataset, and are rather assigned labels of the categories where the aggregates are very heterogeneous.

The fixed-points that correspond to most of the sampled random trajectories, J^*_{gas}, J^*_{liquid}, J_1^* and J_2^*, correspond to stereotypical examples of the aggregate categories we determined in Chapter 3. However, the classifier we trained on the dataset of Chapter 3 is not well adapted to classify the interaction maps resulting from the optimization process, and the characterization of the interaction maps within the same basin of attraction is limited with this method.

In this section, we found that the isotropic attractive particle, the particle without interaction, and particles that organize in periodic patterns, are fixed-points of the renormalization. We could not identify which regions of the phase diagram correspond to the basin of attraction of each fixed-point, but we looked at interaction maps within the basins of attraction. From this statistical sampling, it seems that particles that form dense and unorganized aggregate will renormalize to the liquid fixed-point J^*_{liquid}, while the aggregates of small sizes will renormalize to the gas fixed-point J^*_{gas}. We identified 7 patterns of organization of the particles that correspond to a crystal fixed-point, but we have no guarantees that they are the only ones. Increasing the number of sampled trajectories will not guarantee the exhaustiveness of the fixed-point identification. This method only enables us to identify stable fixed-points: unstable fixed-points are repulsive in at least one dimension of the space, and we do not expect renormalization trajectories to converge towards it.

There seem to be important overlaps between the phenomenological classification of the aggregates we introduce in Chapter 3, that relied on a manual labelling of images of aggregates, and the renormalization classification, that rely on scale invariance properties of the particle organization. This is an indication that the classification we proposed in Chapter 3 relies on physical principles of the particle organization. The classification with renormalization however misses important features of the aggregates: all the aggregates of small sizes will renormalize to a gas, regardless of their dimensionality, or size, and it is not possible to introduce distinctions between them. Specifically, we can not systematically distinguish between the interaction maps leading to fibers or micelles, which have interesting properties (size and dimensionality reduction) with renormalization.

4.4 The fixed-points basin of attraction correspond to stereotypical aggregates

In this section, we go beyond the statistical characterization of the initial interaction maps within the same basins of attraction proposed in Sec. 4.3, and look in details at the common properties of the interaction maps within one basin of attraction. For the liquid and the gas fixed-points, there is a large amount of data within each basin of attraction, and the interaction maps are very diverse at the initial step of renormalization, and very homogeneous at the final step. We use this data to see how the affinity and anisotropy of the particles vary along the renormalization process. We show that the anisotropy of the particle progressively decreases for both, and that the affinity reaches a fixed value, zero for the trajectories in the basin of attraction of the gas, and around $-2kT$ for the basin of attraction of the liquid. The basin of attraction of the gas will be studied in Sec. 4.4.1, and that of the liquid in Sec. 4.4.2. We also study the interaction maps in the basin of attraction of the crystals, and show that each trajectory converges very fast to an interaction maps with only two type of interactions, one attractive and one repulsive, upon renormalization (Sec. 4.4.3). We show why this indicates that the gas, liquid and crystal fixed-points are stable. Finally, in Sec. 4.4.4, we show that the renormalization trajectories of stereotypical fibers suggest that the fiber is an unstable fixed-point of the model.

To study how some characteristics of the interaction maps of the aggregates are renormalized, we plot their measure at step $t+1$ of the renormalization as a function of their measure at step t, and compare this evolution with the first bisector. If f is a function that measures a characteristic of the renormalization map J, such as its average or standard deviation, we plot at each renormalization step of each category $f(J')$ as a function of $f(J)$. The position of the data compared to the first bisector is an indication of the stability of the fixed-point, as illustrated in Figure 4.18. If the values of $f(J)$ are below

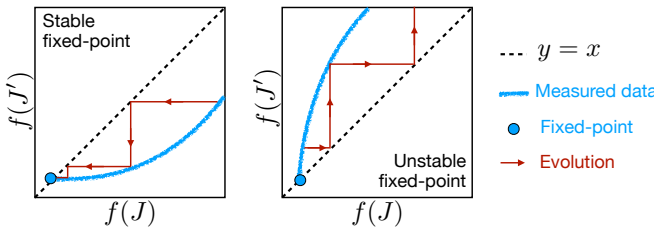

Figure 4.18: The position of a measure quantity along the renormalization with respect to the first bisector indicates whether the fixed-point is stable or not. This is a schematic of the expected behavior for any measure performed on the interaction map J, such as the affinity and anisotropy.

the first bisector, it means that the quantity of interest is supposed to converge towards its value for the fixed-point. On the left plot, the red arrows show how a measure at step t is renormalized at step $t+1$ closer to the fixed-point. On the contrary, if the data is above the first bisector, this quantity will increase away from the value of the fixed-point along the renormalization process. The slope of the data close to the fixed-points is also an indication of how fast this characteristic is renormalized. If the blue curve is completely flat, the system is renormalized in one step to the fixed-point value. If it is steep and close to the bisector, it is renormalized slowly. In this section, we choose specific functions f that are relevant characteristics of the interaction maps or the aggregate, and observe their evolution in this type of plot $(f(J'), f(J))$.

4.4.1 Diverse aggregates of finite size renormalized fast to a gas

Here, we determine common characteristics to the interaction maps in the basin of attraction of the gas fixed-point: they correspond to aggregate with possibly large anisotropy and large sizes, that will be renormalized in only one or two steps in small spherical aggregates, that are the equilibrium configuration of particles with both low affinity and anisotropy.

The interaction map of a gas has zero affinity and anisotropy, and the typical cluster size in a system with no interactions is around one. Here, we study how those characteristics (anisotropy, affinity, and size) behaves along renormalization for the random trajectories in the basin of attraction of the gas. At each renormalization step, we plot (Figure 4.19) $\mu' = \text{average}(J^{(t+1)})$ as a function of $\mu = \text{average}(J^{(t)})$ (panel a), $\sigma' = \text{std}(J^{(t+1)})$ as a function of $\sigma = \text{std}(J^{(t)})$ (panel b), and $\langle s \rangle (J^{(t+1)})$ as a function of $\langle s \rangle (J^{(t)})$, where $\langle s \rangle$ is the average size of the aggregate at equilibrium for a given interaction map J (panel c). We explained how the aggregate sizes were computed in Chapter 2. The measured data are mostly below the first bisector, and they cross the first bisector at $\mu = 0$, $\sigma = 0$ and $\langle s \rangle \approx 1$. This shows that the gas fixed-point is attractive, and that the affinity and anisotropy of the particle decreases along the renormalization, and so does the equilibrium size of the aggregate.

Our data suggest that the renormalization is very fast at the first step for most of the systems. The colors of the points in Figure 4.19 a and b indicates the relative step of renormalization: because some trajectories are shorter than other, the initial step of renormalization is colored in dark purple, the last one in yellow, and the intermediate steps in intermediate colors. At the initial step, the anisotropy of the particle can be large. After one step, however (blue and orange points), the anisotropy is low. The situation is not as clear as that of the schematic of Figure 4.18, but if we were to plot a line to explain the behavior of the data, its slope would be low close to the fixed-point, suggesting a fast renormalization.

We suppose that the points for which the anisotropy is large at the initial step concern

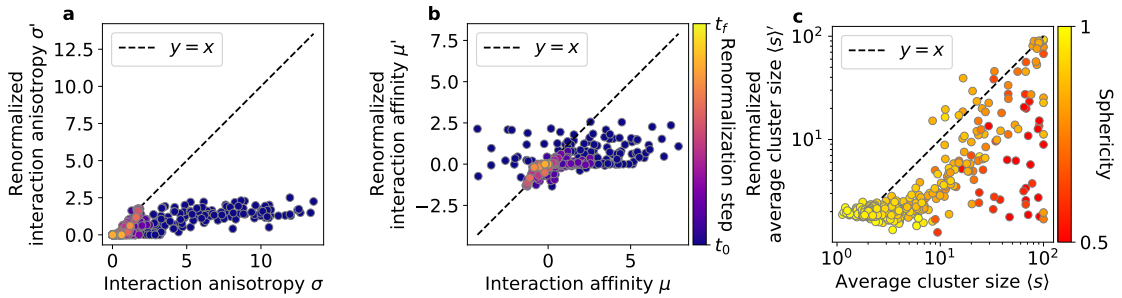

Figure 4.19: When interaction maps are renormalized to that of a gas, the anisotropy and affinity of the interactions, and the size of the aggregate, decrease fast. We plot the data for the trajectories introduced in Sec. 4.3 in the basin of attraction of the gas fixed-point. Panels (a) and (b) share the same color code that indicates the step of renormalization, and the color in (c) refers to the sphericity of the aggregate.

fibrillar aggregates. Indeed, such aggregates can have a well-defined dense organization, but the number of surface bonds per particle is such that this organization is lost in one renormalization step. This is confirmed in Figure 4.19c. For highly non-spherical aggregates (red points far below the bisector), the aggregate is diluted in one step, whereas spherical aggregates are renormalized slowly (orange points close below the bisector). The few data points above or on the bisector concern unreliable trajectories: at a given renormalization step, a dense aggregate renormalizes to an aggregate of larger sizes in a specific case where the optimization process did not converge.

Our data also suggest that repulsive particles renormalize to a gas very fast (Figure 4.19b, if $\mu > 0$, $\mu' \approx 0$). This was expected: we consider a model with short range interactions only, such that there is no cost for the repulsive particle to be a second neighbor. Repulsive particles on the renormalized lattice behave exactly like a gas.

Finally, we confirm what was observed in Figure 4.17: the basin of attraction of the gas is not only composed of repulsive particles or oligomers. Indeed, there is a significant amount of aggregate of sizes larger than 10 particles that renormalize to a gas: aggregates of large but finite size. This result, already derived in Chapter 3 with other method, emphasizes the rich phenomenology of self-assembly of particles with simple geometry and anisotropic interactions.

The projection of the data in the $(f(J'), f(J))$ space shows that the interaction maps leading to different phenomenology (fibrillar aggregate, oligomers, repulsive particles) all renormalize to the gas-fixed-point in a few renormalization steps.

4.4.2 Similar aggregates of infinite size, and no long-range order of the particles orientation, renormalize slowly to a liquid

The trajectories in the basin of attraction of the liquid correspond to aggregates of infinite size, for which the particle anisotropy can be initially large, but decreases slowly along the renormalization process.

The liquid fixed-point corresponds to attractive isotropic particles. However, from the classification of the initial points of Figure 4.17, we observed that the aggregates within the basin of attraction of the liquid were very diverse. We confirm this observation by plotting the data in the $((\mu', \mu)$, (σ', σ) and $(\langle s' \rangle, \langle s \rangle)$ projection in Figure 4.20, as was done for the gas fixed-point in Sec. 4.4.1. These plots confirm that the liquid fixed-point has zero anisotropy, the affinity is around $-2kT$, and the size is 100 particles, *i.e.* the maximal possible size. If the aggregate in the basin of attraction of the liquid are all large (panel c), the anisotropy of the interaction can be around $5kT$, which is not negligible. This means that the aggregate have all have infinite size, but the particles can be arranged in a non-

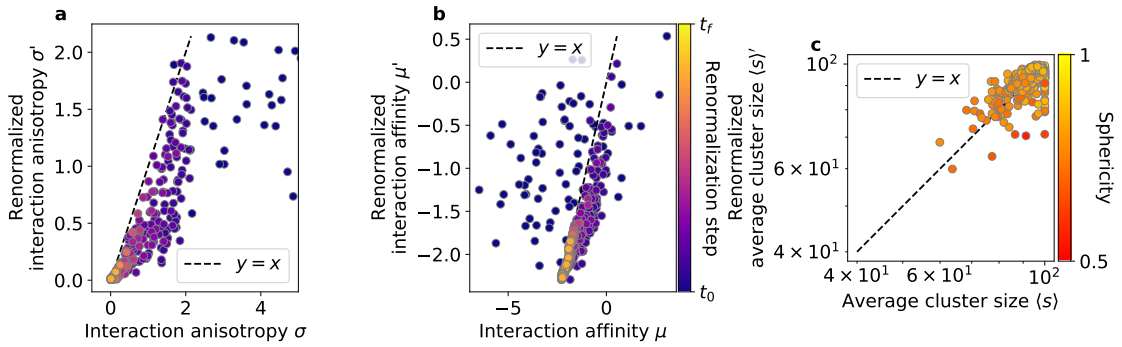

Figure 4.20: When interaction maps are renormalized to that of a liquid, the anisotropy and affinity of the interactions, decreases slowly and the aggregate remain spherical and of infinite size along the renormalization process. We plot the data for the trajectories introduced in Sec. 4.3 in the basin of attraction of the liquid fixed-point. Panels (a) and (b) share the same color code that indicates the step of renormalization, and the color in (c) refers to the sphericity of the aggregate.

random way in the aggregate. However, this non-randomness of the particle organization, which is measured with the particle anisotropy, will decrease to zero.

Contrarily to the gas fixed-point, the renormalization towards the liquid is slow because all the points are close to the first bisector, in the anisotropy and affinity projection. Within one renormalization step, the anisotropy of the interaction decreases of a small amount. We give the following interpretation to this observation: the aggregates in the basin of attraction of the liquid have some favored organization of the particles due to favored directional interactions. Those interactions are not strong enough for this organization to be long-ranged. After some renormalization step, the orientations of the particles in the aggregate will be completely random.

Finally, this plot enables us to evaluate the energy of the attractive interactions of the liquid fixed-point. We saw in Figure 4.14 that the density map of the liquid fixed-point is such that all the interactions are equal to a negative energy value. This is confirmed here by the fact that the anisotropy converges to zero. The value of all the attractive interactions then corresponds to the measured affinity, which is around $-2kT$.

We interpret the interaction anisotropy as an indicator of the long-range order of the particles orientations. This interpretation would need further verification, but it is not clear how one could measure this long range order from the 21-dimensional density map. We also determined an approximation of the exact interaction map of the liquid fixed points: J^*_{liquid} is such that all the interactions are around $-2kT$.

4.4.3 Two-level interactions allowing periodic organization of the particles renormalize to a crystal

We now study the renormalization trajectories for which the long-range organization of the particle was kept along the renormalization process, *i.e.* the crystalline fixed-points, J^*_1, J^*_2, ...,J^*_7. In the interaction maps plotted in Figure 4.14, we see that it is straightforward to identify the favored (blue) and unfavored (red) face pairs (or a pair of faces). Our goal is to study how each the strength of the favored and unfavored interaction evolve along the renormalization. We show that this evolution is very similar for all the crystalline fixed-points, and that the strength of the favored and unfavored interaction necessarily converge to well define values.

From one fixed-point to the other, the indices of the favored and unfavored face pairs are different. For each fixed-point k, we determine the list of indices $\{\alpha^{(k)}\}$ that correspond

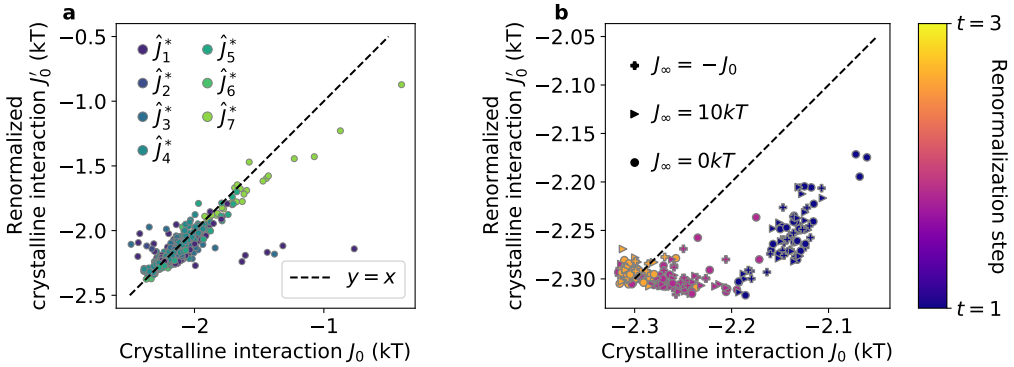

Figure 4.21: Renormalization of the crystalline structures. Averaged crystalline interaction is plotted as a function of its renormalized measure for each renormalization step above t, $t \geq 1$ a) Measures from the random trajectories of Sec. 4.3.2 in the basin of attraction of the crystalline fixed-points. The crystalline interaction is the average of the energy of the crystal contacts in the permuted interaction map. The color refers to the basin of attraction (as defined in Figure 4.14). b) Measures from idealized interaction map with crystal interactions of energy J_0 with J_0 between -2 and -10 kT, and all the other interactions of energy $J_\infty = -J_0$, $0kT$, or $10kT$. The color refers to the renormalization step.

to favored interactions in the interaction map, which we call *crystalline* interactions. There is N_k such favored interaction for the fixed-point k. From Figure 4.14, we see that $N_1 = 3$ and, $N_5 = 9$ for instance: there are respectively 3 and 9 blue squares in the triangular inferior matrix of J_1^* and J_5^*. We see how the strength of the crystalline interactions vary along the renormalization process. For an interaction map in the basin of attraction of J_k^*, we define the averaged crystalline interaction J_0 as,

$$J_0^{(t)} = \frac{\sum_{i=0}^{N_k} J^{(t)}_{\alpha_i^{(k)}}}{N_k} \tag{4.35}$$

Note that the set of crystalline interaction $\{\alpha\}^{(k)}$ is defined solely from the assigned fixed-point. $J_{\alpha_i^{(k)}}$ is not necessarily negative at all steps of the renormalization. For instance, in Figure 4.12, the interaction in the upper corner of the matrix is negative (red) at the end of the renormalization process, but is positive (blue) at $t = 2$.

It is expected that $J_0^{(t)}$ will converge towards $J_{0,k}^*$ for the trajectories in the basin of attraction of the k^{th} fixed-point. In Figure 4.21a, we plot $J_0^{(t+1)}$ for each trajectory as a function of $J_0^{(t)}$ for the random trajectories in the basin of attraction of the crystalline fixed-points. At the initial step of the renormalization, the interactions can have arbitrarily high values, because the interactions maps are drawn randomly, and not the result of the optimization process. The value for $J_0(t = 0)$ is then hard to compare with $J_0(t > 0)$. We only plot the data for renormalization steps larger than zero. We observe that all the values of J_0 fall in the same region: around $-2kT$ on the first bisector. The average crystalline interaction is stable along the renormalization, and it converges to the same for all the crystalline fixed-points. By introducing J_0, we identified a transformation of the interaction maps that enables to compare the trajectories in different basin of attraction, provided that the fixed-point is crystalline. This shows that all the crystalline fixed-points rely on the same physical principles, even if the organization of the particles is different: there is a few favored interaction that lead to a periodic organization of the particles in a dense aggregate.

A crystal has some favored interaction of strength J_0 and some repulsive interactions,

and the favored interaction lead to a periodic organization of the particles. We call J_∞ the strength of the unfavored interaction. We show that if J_0 is negative enough, all the interaction maps that verify this description will renormalize to the same fixed-point, where the repulsive and attractive interactions are of strength J_0^* and J_∞^*. For this, we consider interaction maps for which the favored interactions are at indices $\alpha^{(1)}$. Those matrices are similar to J_1^*, but we vary the strength of the attractive and repulsive interactions. We take J_0 in $[-1kT, -2kT, -3kT, ... - 10kT]$ and J_∞ between $0kT$, $10kT$ and $-J_0$. We perform the renormalization of those interaction maps with our numerical procedure. When $J_0 = -1kT$, the fixed-point is the gas. For all the other interaction maps, we plot the evolution of J_0 along the renormalization in the (J_0', J_0) space in Figure 4.21b. We plot the data from the second renormalization step only, because the initial values for J_0 take very different and potentially large value, and we want to observe the convergence of J_0 close to the fixed-point. The points for $t = 1$ are all gathered in the same region of the parameter space (dark purple points), and so are they after the second (pink point) and third step of renormalization (yellow points). The position of the points does not depend on the value of J_∞ (indicated by the symbols) or J_0 of the initial interaction map. The trajectories converge towards a crystal with $J_0^* = -2.294 \pm 0.007kT$ and $J_\infty^* = 1.823 \pm 0.008kT$. We conclude that the difference of interaction strength between matrices that correspond to a crystal vanishes after one step of renormalization.

Here, we introduced a partial definition of the basin of attraction of the crystal: it encompasses all the interaction maps with two levels of interaction energies, one attractive and one repulsive, for which the favored interactions are such that there is a periodic organization of the particles. Upon renormalization, the favored interactions will remain the same, and the strength of the interaction will renormalize to well-defined values J_0^* and J_∞^*. There are interaction maps that do not verify those criteria and still renormalize to a crystal fixed-point. We did not find an explanation for the measured values of J_0^* and J_∞^*.

4.4.4 Fibers are unstable fixed-points

Here, we show that fibers are unstable fixed-points of the renormalization.

We expect fibers to be fixed-point of the renormalization procedure. Indeed, the second neighbors of a particle within a fiber are the same as its first neighbors. This is illustrated in Figure 4.22a, the particle labeled 1 is first-neighbor with the particle labelled 2 and second neighbor with the particle labelled 3. The interaction between 1 and 2 and 1 and 3 are identical. Therefore, an infinitely long fiber will be identical after renormalization. An infinitely long fibers corresponds to an infinitely negative interaction energy. We call J_{fib} the interaction strength between particles along the fibers (dark blue interaction in the schematic). It seems that the interaction map for which $J_{\text{fib}} = -\infty$ and all the other interactions are zero is a fixed-point of the renormalization, which we call J_{fiber}^*.

We expect this fixed-point to be unstable. Indeed, if the fiber interaction is finite, the fiber is of finite length. As a consequence, it has extremities. Particles at the extremities of the fiber are neighbors with empty sites. If the fiber is of width n ($n = 2$ in the schematic), $2n$ particles have an empty first neighbor in the fiber direction, and $4n$ particles have an empty second neighbor in the fiber direction. Therefore, the number of fiber contacts will decrease as the system is renormalized. The interaction strength of the fiber J_{fib} is then expected to increase to zero along the renormalization process.

A stable direction dJ in the interaction space is attractive around a fixed-point J^* if an interaction map $J^* + \delta\ dJ$ is renormalized to $J^* + \delta'\ dJ$, with $|\delta'| < |\delta|$. A variation along this direction will vanish after a few renormalization step. Alternatively, an unstable direction around the fixed-point is such that $|\delta'| > |\delta|$. Here we study the renormalization of interaction maps around the fibrillar fixed-point J_{fiber}^* by considering two directions of

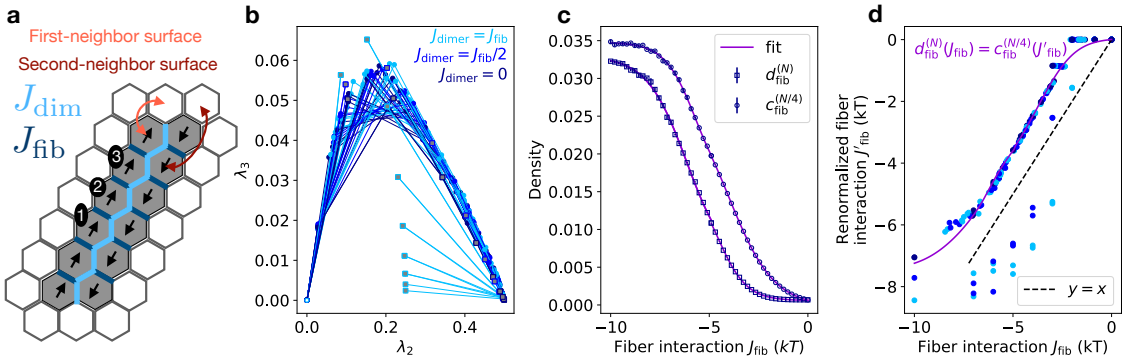

Figure 4.22: The direction of longitudinal interaction is attractive for fibers of width 1 and repulsive for fibers of width 2. a) Schematic of a double fiber, that depends on both the longitudinal interaction (dark blue, of energy J_{fib}) and transversal interaction (light blue, of energy J_{dim}). b) Renormalization of fibers all follow the same trajectories in the projection in a subspace of the eigenvalues. c) Fit (purple) of the dependence of the first or second neighbor density with the longitudinal interaction for a fiber, were $J_{\text{dim}} = 0$ measured on system of size 30×30 (squares) or 15×15 (circles). d) J'_{fib} as function of J_{fib} from the numerical renormalization (blue points) and the fit of first and second neighbor densities (purple line). The points are above the first bisector when $J_{\text{dim}} = 0$, the fixed-points is repulsive, and below when $J_{\text{dim}} < 0$, the fixed-point is attractive.

interaction, which we call longitudinal and transversal to the fiber, and which we describe below.

We perform renormalization of the interaction maps associated with fibers of width 1 and 2. Fibers of width 1 only rely on one head-to-tail interactions, whereas double fibers, rely on one head-to-tail interactions interaction (longitudinal to the fiber director) and two head-to-head interactions (transversal to the fiber director). The longitudinal interaction has strength J_{fib} as discussed above (darkblue on the schematic) and the transversal interaction has strength J_{dim} (light blue in the schematic). An increase or decrease of one of this parameters corresponds to a variation in the directions around the fiber fixed-point. We show that both interactions vanish upon renormalization, but that J_{fib} decreases slowly, and J_{dim} vanishes in only one renormalization step. We vary J_{fib} in $-10, -7, -6, ..., -1kT$ and we consider the case where $J_{\text{dim}} = 0, J_{\text{fib}}/2$ or J_{fib}. All the other interactions have energy zero. We compute two renormalization trajectories per initial condition that we plot in Figure 4.22b. Those renormalization flow are computed with more optimization steps than that of Sec. 4.2.2 (400 optimization step and $\eta_2 = 0.995$ to ensure a precise measure of the renormalized fiber interaction). We choose to plot the trajectories in the space (λ_2, λ_3), with λ the norm of the eigenvalue of the transfer matrix defined in Sec. 4.3.1.3. The trajectories start from different initial points but all follow the same evolution: λ_3 increases then decreases and λ_2 slowly decreases. Because λ_2 decreases slowly, we expect that it is associated with J_{fib}. The interpretation of λ_3 is less straightforward. This representation hints that the renormalization trajectories of fiber interaction maps are similar.

We now measure and predict the evolution of J_{fib} along the renormalization process, and show that it corresponds to the expectation: it slowly vanishes. We first predict the evolution of J_{fib} along the renormalization for a fiber of width 1. In that case, we can assume that $J_{\text{fib}}^{(t+1)}$ only depends on $J_{\text{fib}}^{(t)}$: if the system is diluted enough, the first and second neighbor density only depend on J_{fib}. Under this assumption, we can solve the renormalization relation (eq. 4.4) by measuring the dependency of $\langle c_{\text{fib}} \rangle_{J'_{\text{fib}}}(N/4)$ and $\langle d_{\text{fib}} \rangle_{J_{\text{fib}}}(N)$ with J_{fib} (N and $N/4$ refer to the system size). We show those dependencies in Figure 4.22b. The measured density saturates towards a finite value while $J_{\text{fib}} \to -\infty$

because the system is of finite size and fixed number of particles. We fit those curves with tanh functions (in purple in Figure 4.22b), and plot the dependency of J'fib with Jfib in Figure 4.22c. We compare this expected relation with the data of the numerical renormalization flows plotted in Fig. 4.22a). For fibers of width 1, the numerical renormalization is in agreement with the predicted evolution of J_{fib}. This means that the fibrillar interaction is the one that matter. We also see from this plot that the fiber is unstable, the points lie above the first bisector, and J_{fib} vanishes along the renormalization process. The points are however close to the first bisector, which means that the renormalization is slow.

We now look at the numerical renormalization of double fibers, *i.e.* the interaction maps for which $J_{\text{dim}} < 0$. This is the blue and light blue points in Figure 4.22c. After one step of renormalization, the relation between J_{fib} and J'_{fib} is the same for the fibers of width 1 and 2. During the first renormalization step, J_{fib} becomes more negative (points below the first bisector) and J_{dim} becomes zero: the fiber of width 2, is renormalized into a fiber of width 1 in one step. When there is another interaction ($J_{\text{dim}} < 0$) the direction of longitudinal interaction is attractive. Then, it becomes repulsive, and the strength of the longitudinal interaction decreases.

We could not sample the basin of attraction of the fiber because it is an unstable fixed-point. However, we studied the evolution of the fibrillar interaction by considering interaction maps with only two directions, that of the longitudinal and transversal interactions. The longitudinal direction is repulsive only when there is no interaction in the transversal direction, *i.e.* when the fiber has width 1.

In this section, we studied independently the four types of fixed-points. For each basin of attraction, we determined some observables $f(J)$ that can be measured on the interaction maps, and we studied their evolution in graphs of the type $(f(J'), f(J))$. With this method, we gained indication that the gas, the liquid, and the crystals fixed-points are stable fixed-points, while the fiber is an unstable fixed-point. This method also enabled to identify shared properties of the aggregates within the same basin of attraction. It revealed that the basins of attraction of the liquid and the crystal were homogeneous, both concerning aggregates of infinite size, that either have a periodic organization (crystal) or where the particles have random orientations (liquid). On the contrary, the aggregates in the basin of attraction of the gas are very different, because it concerns all the aggregates of finite size, including the fibers.

4.5 Fixed-points stability

We identified a set of fixed-point by randomly sampling the parameter space, and studied their stability by measuring the evolution of some observable along the renormalization trajectories. Here, we propose a method to study the stability of the fixed-point in a systematic way, by identifying the stable directions (Sec. 4.5.1). We then use this method to show that the gas, liquid and crystal fixed-points are stable fixed-points, and emphasize that we cannot use this method to study the fiber unstable fixed-point (Sec. 4.5.2).

4.5.1 The linearization of the density around the fixed-points enable its stability analysis

Here, we show how to study the stability of a fixed-point by linearizing the evolution of the density map at the vicinity of the fixed-point, and by computing a stability matrix at the vicinity of the fixed-point, of which the eigenvectors are the stable directions around the fixed-points, and the eigenvalues a measure of their stability.

We can perturb the interaction map of a fixed-point $\mathbf{J}^*$, by increasing or decreasing one of the direction of a quantity δ. We denote as $\mathbf{e}_i$ the i^{th} direction of the interaction map (i is between 1 and 21, and $\mathbf{e}_i$ is the 21 dimensional vector for which all values are 0, except the i^{th} coordinate.

$$c_j(\mathbf{J}^* + \delta^{(t+1)}\mathbf{e}_i) = c_j(\mathbf{J}^*) + M_{ij}\delta^{(t+1)} \tag{4.36}$$

$$d_j(\mathbf{J}^* + \delta^{(t)}\mathbf{e}_k\ ) = d_j(\mathbf{J}^*) + N_{jk}\delta^{(t)} \tag{4.37}$$

where the numbers M_{ij} and N_{jk} are the coefficient of the linearization. If δ is small, $M_{ij} = \frac{\partial c_j}{\partial J_i}$, and $M_{jk} = \frac{\partial d_j}{\partial J_k}$. By definition of the renormalization, $c_j^{(t+1)} = d_j^{(t)}$ ((4.36)=(4.37)). Because $c_j(\mathbf{J}^*) = d_j(\mathbf{J}^*)$, by definition, we get $M_{ij}\delta^{(t+1)} = N_{jk}\delta^{(t)}$. Moreover, by chain rule

$$\frac{\partial J'_i}{\partial J_k}\Big|_{J^*} = \frac{\partial J'_i}{\partial c_j}\frac{\partial d_j}{\partial J_k} = (\hat{M}^{-1}\hat{N})_{ik} \tag{4.38}$$

We call $\hat{\Delta}$ the stability matrix $\hat{M}^{-1}\hat{N}$. It is a matrix of dimension 21×21. $\hat{\Delta}$ contains all the information about the stability of the fixed-points. We diagonalize this matrix and study the norm of its eigenvalue. If an eigenvalue has its norm above one, it means that a deviation from the fixed-point in the corresponding eigen-direction of a value δ will be amplified. This is a repulsive direction. The eigenvalues below one correspond to directions for which a deviation will be renormalized to a smaller deviation: this is an attractive direction. A fixed-point is stable if all its eigen-directions are attractive.

4.5.2 The gas, liquid and crystal fixed-points are stable

We now explain how the values of the energies in the interaction maps of the fixed-points are approximated, show the results of the linearization at the vicinity of the gas, liquid and crystal fixed-point, and conclude that they are stable from the stability analysis.

Here, we explain why the values of J^* are only approximations. In the previous sections, we identified the interaction maps that were fixed-points of the renormalization. As can be seen in Figure 4.14, in most cases, a fixed-point is an interaction map with two levels of interactions, one attractive and one repulsive. We denote as J_0 and J_∞, the energy of the attractive and repulsive interactions. The set of face pairs corresponding to attractive interactions for a given fixed-point is $\{\alpha\}$, as was defined in Sec. 4.4.3. The numerical identification of the sampling enables to identify without ambiguity the values of $\{\alpha\}$ for each fixed-points. However, there are some uncertainties in the value of the exact strength of the attractive and repulsive interaction, J_0 and J_∞. We observed this along this chapter: the measure of the distance between the interaction maps and the fixed-points was of the order of $10^{-2}kT$ in Figure 4.15. In Figures 4.20 and 4.21, the averaged attractive interaction converges towards a zone around $-2kT$, but the points are scattered around that value. The numerical renormalization therefore only enables to make an estimation of the values for J_0 and J_∞.

We perform the linearization described above around the fixed-points J^*_{liquid}, J^*_{gas} and the crystal fixed-point J^*_1, in order to measure the stability matrix for those three fixed-points. The corresponding matrices are shown on top of Figure 4.23 with the usual color code: blue is attractive, red is repulsive, white is neutral. For both J^*_{liquid} and J^*_1, we set

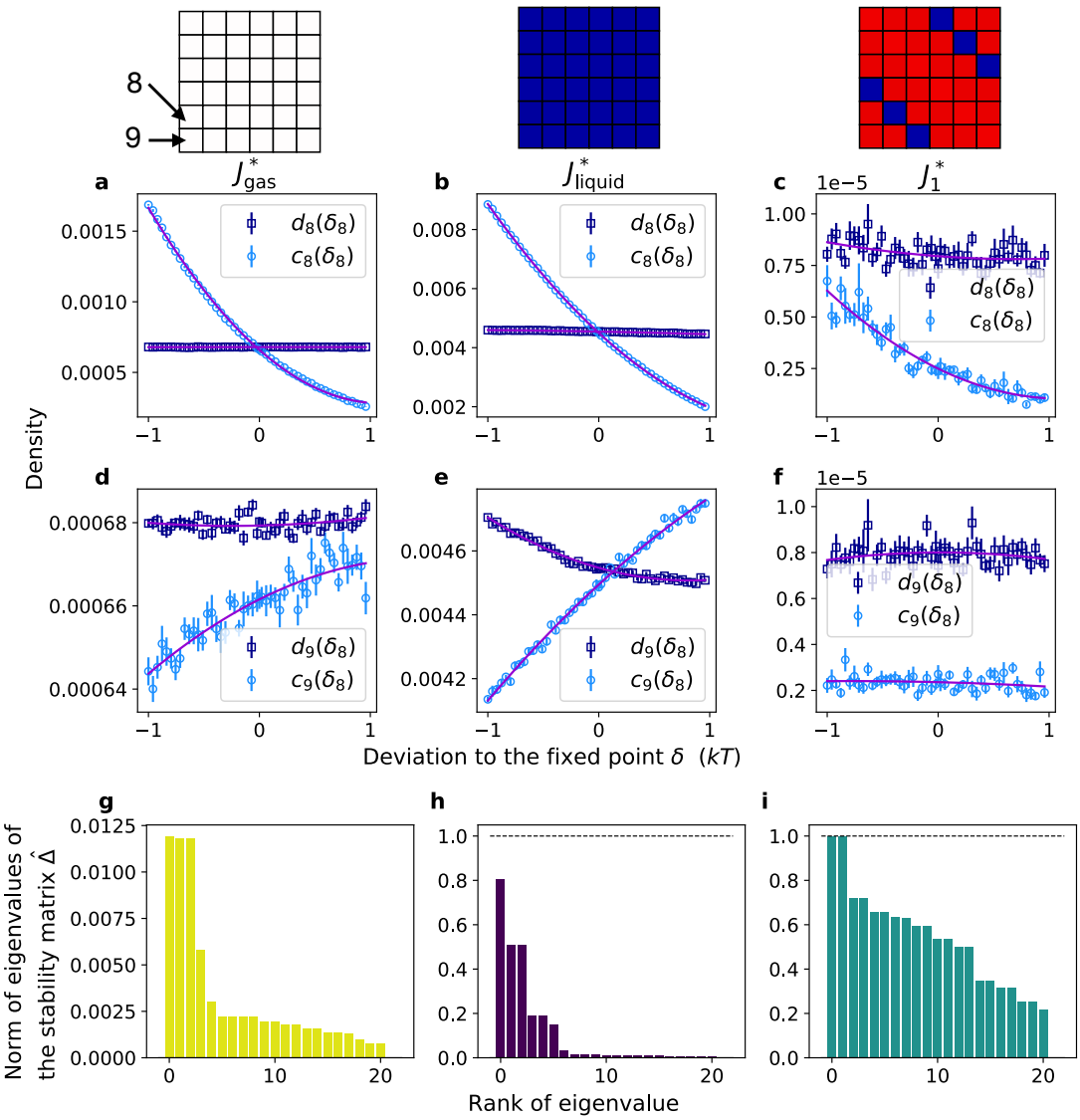

Figure 4.23: The gas, liquid and crystal fixed-points are stables, because the norm of the eigenvalues of the stability matrix measured by linearization are below one. (a-f) c_8, d_8, c_9 and d_9 depend on the extra energy δ of the interaction 8 (position of the 8^{th} and 9^{th} interactions are shown in the interaction map J^*_{gas} on top of the figure. We measure the slope of the curves around $\delta = 0$. The measures for the first-neighbor (light blue) and second-neighbor (dark blue) densities are respectively measured of systems of size 15×15 and 30×30. (g-i) the measured norm of the eigenvalues are less or equal to one.

the attractive interaction to be of value $J_0 = -2kT$. For J_1^*, we also set the value of the repulsive interaction to be $J_\infty = 1.8kT$. These values were determined in Sec. 4.4. J_{gas}^* is the interaction map where all the couplings are zero. We expect that upon an increase or a decrease of the i^{th} interaction of the matrix of a value δ, the first-neighbor density $c_i(J^* + \delta e_i)$ will decrease or increase. Indeed, if interaction i is more attractive, we expect more of the bonds to be in that configuration. The variation of the density of bonds in another face pair is not simple to predict, and so is the variation of the second neighbor densities d_i and d_j. We also expect that $c_i(\delta = 0) = d_i(\delta = 0)$, because this is how the fixed-point was defined.

In Figure 4.23, we show the evolution of c_8 and d_8 (panels a, b, c) and c_9 and d_9 (panels d, e, f) upon a variation δ of the 8^{th} interaction of the interaction map for the gas, liquid and crystal fixed-points. We observe the expected decrease of c_8 with δ. For the crystal, however, we do not observe $c_i(\delta = 0) = d_i(\delta = 0)$ for $i = 8$ and $i = 9$. This might be because the chosen value of J_0 and J_∞ are not the exact values of the fixed-point. The difference $c_i(\delta = 0) - d_i(\delta = 0)$ is of the order $5 \cdot 10^{-6}$ (measured on panels c and f) and the interactions 8 and 9 are not among the favored interactions of the crystal. For comparison, the measured density of the favored interactions in the crystal is 0.026. This difference is therefore negligible. $c_i(\delta = 0) = d_i(\delta = 0)$ is also not verified for the $i = 9$ for the gas fixed-point (panel d). In the case of J_{gas}^* however, there is no ambiguity on the value of the fixed-point. This is instead an artifact of the finite number of particles: when the particles do not have interactions, the probability of observing any pair of particles as neighbors is not $\left(\frac{N_{\text{particles}}}{N_{\text{sites}}}\right)^2$ but rather $\frac{N_{\text{particles}}}{N_{\text{sites}}} \times \frac{N_{\text{particles}}-1}{N_{\text{sites}}-1}$, which does not only depend on the density of particles but also on the system size.

We measure the slope of c and d around $\delta = 0$. For this, we fit the curve with a quadratic function (purple line on Figure 4.23, and we deduce the derivative at $\delta = 0$. We can then measure the stability matrix $\hat{\Delta}$ with eq. 4.38, diagonalize it and measure its eigenvalues. $\hat{\Delta}$ is not diagonalizable in general, and we therefore measure the norm of the complex eigenvalues. It is not clear for us what meaning we should give to complex eigenvalues in the space of interaction maps. The norm of the gas, liquid and crystal fixed-points are plotted in Figure 4.23(g-i). For the gas and the liquid, the fixed-point is attractive, which confirms the observation of the previous sections: those fixed-points are stable.

For the crystal, we measure two eigenvalues of norm one, and all the other of norm lower than one. The eigenvalues of norm one correspond to a deviation to the fixed-point, which will remain identical after one step of renormalization. The corresponding eigenvector are such that the crystalline interactions are 1, and the other interactions are zero. It means that if the crystalline interactions are $J_0 + \delta$, their renormalized strength will remain $J_0 + \delta$. This is in disagreement with what we measured in Sec. 4.4.3, where any initial crystalline interaction was renormalized to $J_0 = -2kT$. However, the measure we are making in this section is very local, and concerns variations of a few decimals of kT, while the observations of Sec. 4.4.3 concerned variations of the initial crystalline interaction of a few kT. This might also be due to the approximated evaluation of the crystalline interaction of the fixed-point.

From the linearization of the density maps at the vicinity of the fixed-points for which the interaction energy were determined numerically, we conclude that the gas, liquid, and fiber fixed-points are stable: all their stable directions are attractive (or neutral in the case of the crystal). It is not possible to use this technique for the fiber fixed-point. Indeed, we showed in 4.4.4 why this fixed-point is expected to have infinitely strong attractive interaction $J_0 = -\infty$. We cannot measure the evolution of the first and second-neighbor density maps closed to this value in the numerical simulation. This result also needs to be taken with caution, because the linearization was performed around a point that did not exactly verify $c = d$ for all the interactions, because of the approximated measure of the energies in the interaction maps of the fixed-points, and because of the fixed number of particles.

Discussion

In the Chapter 2 and 3, we introduced a model of lattice particles with directional interactions, for which the interactions energy depend on the relative orientations of the particles. This model shares similarities with the isotropic lattice gas model, where the only coupling parameter is the nearest neighbor coupling. With the real-space renormalization transformation that we designed in this chapter, we emphasized that this extension of the lattice gas model is non-trivial, because it gives rise to new fixed-points of the renormalization. Aside from the attractive isotropic particle (liquid fixed-point) and the non-interacting particle (gas), we showed here that there are several anisotropic fixed-points of the renormalization. In those cases, the anisotropy of the particles is conserved upon renormalization. It means that the anisotropy of the interactions is a long-range property of this class of parameters. Those new fixed-points, that we called crystalline fixed-points, correspond to aggregates where the local organization of the particles is identical at all length scales. Our renormalization transformation enabled us to identify 7 crystalline fixed-point, though this number depends on the geometry of the particle, and on the chosen decimation procedure. Despite the uncertainties in the evaluation of the number of crystalline fixed-point, the existence of such fixed-points suggest that the directionality of the interactions is a key feature of the model.

In Chapter 2, we showed that there are at least two ways to explore the parameter space of interaction maps: one can change which interactions vary together (the entries of identical colors in the representation of the interaction maps), and the energy of each of those group of interactions. Our results suggest that the fixed-points of the renormalization are mostly interaction maps with only two groups of interaction: the attractive (blue) and repulsive (red) ones. We showed that our numerical renormalization transformation is well suited to identify the groups of identical interactions in the density maps of the fixed-point. However, the evaluation of the value of the energy of the repulsive and attractive interactions is an approximation. Our results suggest that the energy of the attractive interaction is finite, while we would have expected them to be infinite, as in the analytical renormalization of the same model in the grand canonical ensemble. This is because the energy of interactions are the result of an optimization procedure. Combining a numerical identification of the groups of interactions of the fixed-points and an analytical renormalization of the interaction energies, such as the one in [119] would also enable to compensate the limits of both approaches.

Here, we explored the 21 dimensional space of the couplings by studying the renormalization trajectories of random initial couplings. This approach already used in Chapter 3 enabled us to identify a large diversity of behaviors. However, this is not a dense exploration of the space (in the mathematical term). The random sampling of the phase space is the best accessible method we could realistically implement, but it prevents us from drawing universal conclusions from our study. Because of the high dimension of the phase space, we were also not able to determine the boundaries between the basin of attractions of the fixed-points. In other terms, we did not find a projection of the space where the basin of attractions of the fixed-points are well separated. The eigenvalue of the transfer matrix were a reliable way to identify the fixed-points, but not the boundaries of their basins of attractions. However, we were able to determine common characteristics of the interaction maps within the same basin of attraction: the interaction maps that result in an aggregate of finite size are in the basin of attraction of the gas. The interaction maps resulting in aggregates of infinite size are either in the basin of attraction of the liquid or the crystal, depending on the local organization of the particles in the aggregate.

In this work, we made the assumption that the couplings of the decimated system are the same as the initial set of couplings, and we neglected the emergent couplings of higher orders. This approximation is not controlled. A possible way of testing it is would be to compare the results of a decimation that would conserve the statistics of the fourth neighbors, with one two successive decimation that conserve the statistics of the second neighbors (the one we implemented). This will be done in future studies. On the other hand, a great asset of the numerical renormalization is that it does not require to define a small plaquette on which the configuration should be summed. The size of the plaquette is the size of the simulated system, which in our case is 30×30 sites and 100 particles. This enabled us to successfully renormalize a dense aggregate with periodic organization of the particle, where the typical period was large, such as the sponge, or with effective surface tension, such as the micelle.

5 - Topological defects as a size-limiting mechanism for self-assembly

The exploration of the design parameters for lattice particles with directional interactions in the previous chapters suggested that aggregates of limited size can arise at equilibrium, because of frustration. In this Chapter, we examine a specific type of particle that has two favored and incompatible interactions. In the resulting aggregate, the organization of the particles maximizes the number of each type of interactions. Then, upon tuning the strength of those two interactions, the size of the aggregate can be controlled. This introduces a mechanism of self-limited assembly complementary to the ones introduced in Chapter 1, that rely on the individual design of each particle, on their deformability, or on the self-closing of the assembly. The fact that directional attractive interactions give rise to large aggregates of finite size at equilibrium is not-trivial, as shown in Figure 5.1 with the example of a particle (in gray) with attractive patches (in blue). The choice of distribution of the patches on the surface of the particle enables to form oligomers (panel a), fibers (panel), or aggregates of infinite size (panel c). It is not clear, however, how such direction interactions could give rise to aggregates of finite and controlled-size. Changing the temperature or the strength of the interaction could break an infinite aggregate in smaller aggregates, yet the size of such aggregate would not be controlled in that case.

In Sec. 5.1, we introduce the anisotropic particle that can lead to the limitation of the aggregate through the introduction of energetically favored disclination lines in a crystalline aggregate. The aggregate relies on two types of interaction, which we call *line* interaction and *crystal* interaction. In Sec. 5.2, we show how the energy of both interactions are expected to control the geometry and the size of the aggregate, and we derive a phase diagram at zero temperature. In Sec. 5.3, we retrieve the expected phase diagram with numerical simulation at finite temperature, but not the quantitative prediction of the equilibrium size of the aggregates. In Sec. 5.4 we show that kinetic and entropic effect also influence the equilibrium size of the aggregate. The proposed design of particles can also be adapted to self-assemble into fibrillar aggregates, which we show and test numerically in Sec. 5.5. In Sec. 5.6, we show how this mechanism could also be implemented experimentally with DNA-origami.

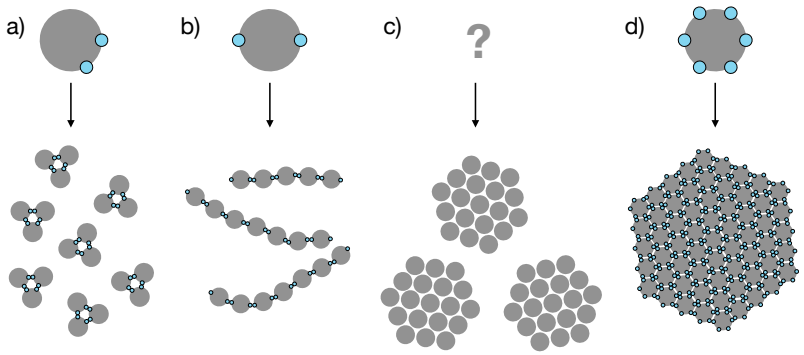

Figure 5.1: Self-limited assembly with directional interaction is non-trivial. gray particles have attractive interaction between the blue patches. Particles can be designed to assemble into oligomers (a), fibers (b), or infinite crystals (c). Self-assembly into large but finite aggregate (d) remains a difficult task.

5.1 Camembert aggregates have favored disclination lines

Here, we show that by designing a particle with incompatible favored interaction can lead to the formation of crystalline aggregates that are stabilized by disclination lines. In Sec. 5.1.1, we explain how to design such a particle and its expected self-assembly. In Sec. 5.1.2, we show why we expect the relative strength of the incompatible interactions to control the size of the aggregate.

5.1.1 Camembert geometry

Here, we introduce a design of a particle that has two incompatible favored interactions, and we explain that we expect it to self-assemble into a specific geometry which we call *camembert*, that has a finite number of disclination lines. We also explain the origin of this terminology.

We introduce a hexagonal particle that has different types of directional short-range interaction. We call *crystal* interaction an interaction between two faces of the particles such that two particles bind if they are in the same orientation. There can be three such interactions for an hexagonal particles, which is illustrated in Figure 5.2a: the three crystal interactions result from three lock-and-key mechanism, shown in light blue. There are three types of lock (triangular, square and round) that bind with shape complementarity. Particles with such interactions are expected to self-assemble into a crystal (Figure 5.2b). We now introduce two other types of interaction, that make the particle bind with a neighbor in a different orientation. These interactions are illustrated in dark blue in Figure 5.2 with another shape complementarity mechanism. Note that two patches bind if they are the same color and if their shape is complementary. If there are two pairs of patches in contact between two neighboring particles, we consider that they bind with an attractive interaction if at least one of the two pair of patches has complementary shape. There are no dense packing of particles, only relying on the dark blue interaction. If both types of interactions are favored, a possible self-assembly is that presented in Figure 5.2b: the particles mostly organize in a crystal configuration, but there are some disclination lines within the crystal where the particles are in contact through the dark blue interaction. For this reason, we call it the *line* interaction. Note that the disclination line is a topological defect of the crystalline organization of the particle, but it is favored energetically.

We call this aggregate *camembert*. Because of the hexagonal geometry of the particle, there cannot be more than 6 disclination lines within the same aggregate. This reminds of the Trivial Pursuit board game, where the round pieces are filled with 6 little triangles. Since the pieces of the trivial pursuit game are called camembert in French, we chose this terminology to refer to the geometry of that aggregate.

5.1.2 Size limitation

We show why the camembert aggregate can be of limited size at equilibrium. Camembert aggregates of small sizes can be more stable than those of large sizes if the line interaction is strong: there are more disclination lines per particle in a small aggregate than in a large one. This can be understood by writing the energy per particle in an aggregate as a function of its size. We denote by J_{crystal} the energy of the crystalline interaction and J_{line} that of the line interaction. In a camembert aggregate of size n, the energy per particle scales as $e(n) \sim (J_{\text{crystal}} n^2 + J_{\text{line}} n)/n^2$. If $J_{\text{line}} < 0$, the energy per particle has a minimum at finite size n^*. This design seems to fulfill the requirements that could not be achieved with the previously existing size-limitation mechanisms: the camembert particles are rigid, they interact with first neighbor local interactions, and a large number of copies of this particle will self-assemble into an open-ended shape.

For this discussion to be exact, we also need to compare the line and crystal interactions

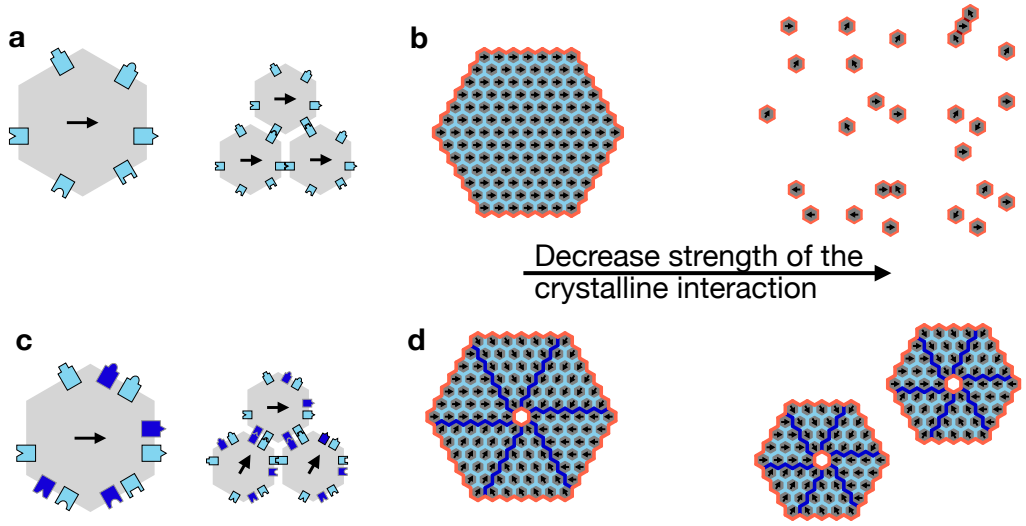

Figure 5.2: Directional interactions are designed such that the particle will form a cluster with disclination lines, that can have equilibrium finite size. a) If the local interactions are such that all particles are aligned, it will form a crystal. b) Decreasing the strength of this interaction will dissolve the crystal, without any regime where the crystal is of small sizes. c) On the contrary, combining competing crystal (light blue) and line (dark blue) interactions leads to competing interaction. d) This should result in a camembert geometry, which could reach finite size if the crystal interaction is low enough.

to the surface interaction of the particle, which we call σ. We also need to compare the energy per particle in a camembert geometry, with that of the energy per particle in a crystal and with that of a monomer, shown in Figure 5.2b.

5.2 Competing interactions control the aggregate size and stability

If the particle has the three types of interaction described above (crystal, line, and surface interaction), there are three possible geometries of aggregate: the particle self-assemble into a crystal, a camembert or remain as monomers, depending on the relative strength of each type of interaction. Here, we will predict the phase diagram, *i.e.* determine for which values of the interactions the camembert is the most stable geometry. For each type of aggregates, we will determine the energy per particle $e(r)$ in an aggregate of radius r. This is the right quantity to consider because it does not depend on the size of the system, and it enables to compare aggregates of different sizes: the crystal is more stable than the monomer if the energy per particles in the crystal is lower than the energy of a monomer. In this section, we only minimize the energy of the particle, not its free-energy, therefore deriving the ground-state of the system, at zero temperature. We will then minimize this energy with respect to r to determine the equilibrium size and energy of the aggregate in a given geometry. We do this in Sec. 5.2.1 for the crystal, in Sec. 5.2.2 for the camembert seed, and in Sec. 5.2.3 for the camembert. For a given set of interactions J_{line}, J_{crystal} and σ, the equilibrium aggregate will correspond to the aggregate with the lowest equilibrium energy. In Sec. 5.2.4, we compare the equilibrium energies of each aggregate and determine the phase diagram.

Along this section, we will mostly consider the rescaled line and crystal interaction

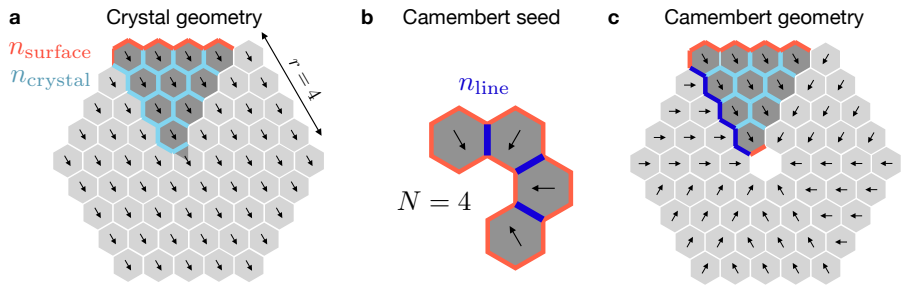

Figure 5.3: We compute the energy of the cluster by counting each the occurrence of each type of interactions for the crystal geometry of radius r (a), a camembert seed of N particles (b) and camembert seed of radius r (c).

energies, x_l and y_c, and the monomer energy, that has 6 surface bonds, e_1:

$$x_l = J_{\text{line}} - 2\sigma \tag{5.1}$$
$$y_c = J_{\text{crystal}} - 2\sigma \tag{5.2}$$
$$e_1 = 6\sigma \tag{5.3}$$

x_l is negative when it is more favored to bind two particles with a line interaction locally: one line interaction replaces two surface interactions. Similarly, y_c is negative when it is more favored to bind two particles with a crystal interaction locally. e_1 sets the level of the energies: an aggregate will be stable if the energy per particle is lower than the energy of a monomer e_1.

5.2.1 Crystal

Here, we compute the energy per particle in a crystal geometry, and we show that the crystal is only stable when $y_c < 0$, *i.e.*, when it is more favored for two particles to bind with a crystal interaction than to remain detached. When the crystal interaction is negative, the equilibrium radius of the aggregate is infinite. This is trivial: there is no mechanism limiting the size of aggregates formed by attractive particles. This simple demonstration will however enable us to introduce the framework that will be reused for the camembert geometry. Here, we do not consider the line interaction, because it is not observed in the crystal.

We first compute the energy per particle of the crystal. The energy of the total aggregate is determined by the number of crystal bonds and surface bonds in the aggregate, which we call n_{crystal} and n_{surface} and by their energy J_{crystal} and σ. Both n_{crystal} and n_{surface} depend on the radius of the aggregate r. Note that this notation will remain the same also for the camembert aggregates. We compute the energy of a *triangular slice*, which correspond to the particles in dark gray in Figure 5.3a. We also assume that the total number of particles in the crystal is such that the radius is an integer and the crystal is perfectly symmetric. The energy of a triangular slice of a crystal of radius r is $E^{(\text{cry})}(r)$, and the energy of the crystal is $6E^{(\text{cry})}(r)$.

$$E^{(\text{cry})}(r) = n_{\text{crystal}}(r)J_c + n_{\text{surface}}(r)\sigma \tag{5.4}$$

We determine $n_{\text{crystal}}(r)$ and $n_{\text{surface}}(r)$ from the drawing of Figure 5.3. The surface bonds are represented in pink and the crystal bonds in light blue. The number of crystal bonds scales like the volume of the aggregate, r^2 and the number of surface bonds like its radius r.

$$n_{\text{crystal}}(r) = 3r(r-1)/2 + 2r \tag{5.5}$$
$$n_{\text{surface}}(r) = 2r + 1 \tag{5.6}$$

The total number of particles in the crystal aggregate is

$$N^{(\mathrm{cry})}(r) = 3r(r+1) + 1 \tag{5.7}$$

We now compute the energy per particle in a crystal and its derivative, expressed with the rescaled variables y_c an e_1 defined in equations 5.2 and 5.3:

$$e^{(\mathrm{cry})}(r) = \frac{6E^{(\mathrm{cry})}(r)}{N^{(\mathrm{cry})}(r)} = \frac{(3y_c + e_1)r^2 + (y_c + e_1)r + e_1/3}{r(r+1) + 1/3} \tag{5.8}$$

$$\partial_r e^{(\mathrm{cry})}(r) = 3y_c \frac{6r^2 + 6r + 1}{(3r^2 + 3r + 1)^2} \tag{5.9}$$

For large values of r, $e(r)$ scales like $3y_c + e_1 = 3J_{\mathrm{crystal}}$: there are three crystal interaction per particles, and the surface energy is a linear correction to the quadratic dependence of the energy on the size of the aggregate. The derivative has the same sign as y_c for $r > 0$, which means that a crystal is of infinite size for $y_c < 0$, and of size 1 (the minimal possible size for an aggregate) for $y_c > 0$. It confirms that there cannot self-limited aggregate of regular particles in two-dimension.

5.2.2 Camembert seed

If the line interaction is favored, we expect the camembert to form. Before considering the total camembert geometry (which we will do in Sec. 5.2.3), we will compare the different stages of assembly of a hexamer where the particles interact through a line interaction, and show that there is no stable intermediate between the monomer and the hexamer, if the line interaction is stable. Here, we do not consider the crystal interaction, because it is not observed in the hexamer. We call *camembert seed* the geometry of an aggregate that is a partially assembled hexamer, as the one shown in Figure 5.3b.

The total hexamer has 6 line interactions, 6×4 surface interactions, and 6 particles. As a consequence

$$e^{(\mathrm{seed})}(N = 6) = J_l + 4\sigma = x_l + e_1 \tag{5.10}$$

We also compute the number of line and surface contacts in a camembert seed of size N. Here, N is between 1 and 5.

$$n_{\mathrm{line}} = N - 1,\ n_{\mathrm{surface}} = 4N + 2 \tag{5.11}$$

The energy per particles of a camembert seed for sizes smaller than 6 is

$$e^{(\mathrm{seed})}(N) = \frac{n_{\mathrm{line}} J_l + n_{\mathrm{surface}}\sigma}{N} = \frac{N-1}{N} x_l + e_1 \tag{5.12}$$

If the line interaction is more favored than the surface interaction, $x_l < 0$, and $e^{(\mathrm{seed})}(N = 6) < e^{(\mathrm{seed})}(1 < N < 6)$: the hexamer is more stable than any partially assembled hexamer. If $x_l > 0$, $e_1 < e^{(\mathrm{seed})}(1 < N < 6)$: the monomer is more stable than any partially assembled hexamer, and then the full hexamer.

5.2.3 Camembert

For the camembert, the determination of the stability is less trivial and the energy of the aggregate depends both on x_l and y_c. In Sec. 5.2.3.1, we first determine the energy per particle in a camembert, and the regions of the parameter space where it is stable. For a specific range of line and crystal interactions, the camembert is the most stable for a radius r_1 that is larger than one, and finite. We will show how r_1 depends on the interaction energies in Sec. 5.2.3.2.

5.2.3.1 Energy of the camembert

Here, we show that the energy per particle in the camembert has a non-trivial dependence in r, such that it can reach its minimum for values r_1 that are not 1 or $+\infty$, as the other geometries presented before.

We first write the energy of a triangular slice in a camembert aggregate as a function of its radius, as we did for the crystal geometry. The triangular slice is shown in dark gray in Figure 5.3c.

$$E^{(\text{cam})}(r) = n_{\text{crystal}}(r)\ J_c + n_{\text{line}}(r)\ J_l + n_{\text{surface}}(r)\ \sigma \tag{5.13}$$

$$N^{(\text{cam})}(r) = 3r(r+1) \tag{5.14}$$

We determine the number of crystal, line, and surface bonds in a triangular slice as before, with the example shown in Figure 5.3c. The number of crystal bonds (light blue) scales like r^2, and the number of line and surface bonds scales (dark blue and pink) scale like r.

$$n_{\text{crystal}}(r) = 3r(r-1)/2 \tag{5.15}$$

$$n_{\text{line}}(r) = 2r - 1 \tag{5.16}$$

$$n_{\text{surface}}(r) = 2r + 2 \tag{5.17}$$

We deduce the energy per particle in a camembert, and its derivative with r. The dependence of the energy in x_l and y_c is non-trivial.

$$e^{(\text{cam})}(r) = \frac{6E^{(\text{cam})}(r)}{N^{(\text{cam})}(r)} = \frac{(3y_c + e_1)r^2 + (-3y_c + 4x_l + e_1)r - 2x_l}{r(r+1)} \tag{5.18}$$

$$\partial_r e^{(\text{cam})}(r) = 2\frac{(3y_c - 2x_l)r^2 + x_l(2r+1)}{r^2(r+1)^2} \tag{5.19}$$

We verify that $e^{(\text{cam})}(r=1) = e^{(\text{seed})}(N=6) = x_l + e_1$. The energy per particle of the camembert is such that it has a finite minimum in some regions of the parameter space. The derivative of the energy is a quadratic function. If $3y_c - 2x_l > 0$, $\partial_r e^{(\text{cam})}(r)$ is negative when $r \to \infty$, and the energy will be increasing. Then, if the larger root of $\partial_r e^{(\text{cam})}(r)$ is larger than one, the energy has a minimum for a value of r that is finite and larger than one. It means that the camembert has a finite size. We call it *self-limited*: its size is limited for thermodynamic reasons. If $3y_c - 2x_l < 0$, the energy will be decreasing for $r \to \infty$ and the geometry is always stable for $r = \infty$. The size of the aggregate can be finite if the smaller root of $\partial_r e^{(\text{cam})}(r)$ is larger than one, and if the corresponding energy is lower than the energy at infinity.

We determine the sign of the roots of $\partial_r e(r)$ in all regions of the parameter space, to determine in which regions the energy has a positive local minimum. We distinguish between the case where this minimum is at infinity or not. The expression for the roots of $\partial_r e^{(\text{cam})}(r)$, which we call r_1 and r_2, is

$$r_{\substack{1\\2}} = \frac{-x_l \pm \sqrt{\Delta}}{3y_c - 2x_l} \text{ and } \Delta = -3x_l(y_c - x_l) \tag{5.20}$$

The stability of the camembert and the value of the minimum then depends on the sign of three quantities: $3y_c - 2x_l$, which determines the monotonicity for $r \to \infty$, and x_l and $y_c - x_l$, which determines the sign of the discriminant Δ. The sign study is summarized in table 5.1.

Therefore, the energy of a particle in the camembert geometry is minimum for a positive finite value of r if $x_l < 0$ or if $x_l > 0$ and $y_c < 2/3x_l$. Furthermore, there is only one region

$3y_c - 2x_l$	+	+	+	-	-	-
x_l	+	+	-	+	-	-
$y_c - x_l$	+	-	+	-	+	-
Δ	-	+	+	+	+	-
$-x_l + \sqrt{\Delta}$		-	+	+	+	
r_1		-	+	-	-	
$-x_l - \sqrt{\Delta}$		-	-	-	+	
r_2		-	-	+	-	
$\min_{r>0}(e(r))$	0	0	$r_1 > 0$	$+\infty$	$+\infty$	$+\infty$
Camembert stable	no	no	yes	yes	yes	yes

Table 5.1: Determination of the regions of camembert stability

of the parameter space where this minimum is reached for finite values of r, which is when $x_l < 0$ and $y_c > 2/3x_l$: it is more favored for two particles to bind with the line interaction than to stay apart, and the crystal coupling is weaker than the line coupling, of a factor 2/3.

5.2.3.2 Equilibrium size of the camembert

In the region where the camembert is of finite size, we can determine its equilibrium size from the expression of r_1 (eq. 5.20). In particular, we determine the regions of the parameter space for which the equilibrium radius of the camembert is an integer k. We solve $r_1 = k$. It corresponds to a straight line in the phase diagram, of equation

$$y_c = \frac{2k^2 - 2k - 1}{3k^2} x_l = a(k) x_l \tag{5.21}$$

If y_c and x_l are such that $y_c < a(k)$ and $y_c > a(k+1)$, it means that the equilibrium radius of the aggregate is between k and $k+1$. In Figure 5.4, we plot in the same color the regions where the rounding of the equilibrium radius is identical. Larger sizes of camembert are observed for decreasing values of y_c. All the lines $y_c = a(k)x_l$, which correspond to the radius being a finite number, cross at $x_l = y_c = 0$, and the slope increases with k (as $1/k$) to the limit $y_c = 2/3x_l$ which corresponds to the infinite size. It means that the lines where $r_1 = k$ in the phase diagram are closer and closer as k increases: the region of identical colors are more and more narrow in Figure 5.4. Therefore, the larger the size of the camembert, the finer the coupling have to be tuned to reach that size. This has also been observed for self-assembly of deformable tiles governed by shape geometrical frustration [124]. We showed in 5.2.2 that it is not necessary to consider the case where $r_1 < 1$, *i.e.* when the first hexamer is not assembled. Indeed, we showed that when $x_l < 0$, the full hexamer ($r = 1$) is more assembled than any partially assembled hexamer ($r < 1$). The approximate size of the camembert is then one above and below the line of equation $y_c = \alpha(1)x_l = -x_l/3$. The radius of the camembert is larger than 2 when $y_c = \alpha(2)x_l = x_l/4$. The region of the parameter space where the camembert is of radius larger than 2, and less than $+\infty$, which we consider as the non-trivial case, is between the boundaries $y_c = \alpha(2)x_l = x_l/4$ and $y_c = 2/3x_l$. The region in the parameter space where the size of the camembert is limited is not narrow.

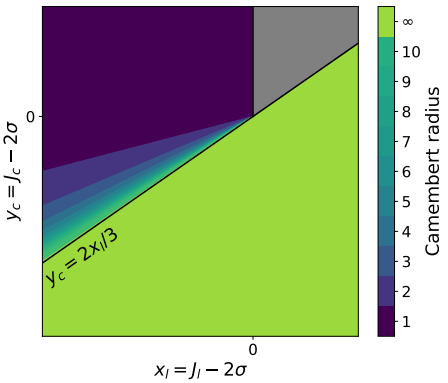

Figure 5.4: The size of the camembert finite and larger than one when $y_c > 2x_l/3$ and $x_l < 0$. The gray zone corresponds to the region where the camembert is unstable.

The energy of a camembert depends on both the energy of the line and the crystalline interaction. When the line interaction is larger than the crystalline interaction, the energy of an aggregate is smaller when its size is smaller, because there are more line interactions per particles in the whole aggregate. A fine-tuning of the relative strength of both interactions then enables to control the equilibrium size of the aggregate.

5.2.4 Phase diagram

We now determine the phase diagram of the camembert particle, *i.e.*, in which regions of the parameter space each of the three geometries described above is the most stable. In a system at zero temperature, the geometry of lowest energy is the ground state of the system. We compare the energy per particles in the camembert, the crystal, and the monomer, two by two. In Sec. 5.2.4.1, we compare the camembert and the crystal, in Sec. 5.2.4.2 the crystal and the monomer, and in Sec. 5.2.4.3 the camembert and the monomer.

5.2.4.1 Crystal-camembert boundary

For both the crystal and the camembert, the asymptotic energy for infinite size is $e(r = \infty) = 3J_c$. To determine which of the camembert or the crystal is most stable, we compare the correction to this limit, which scales as r^{-1}, as seen in equations 5.8 and 5.18.

$$e_{\min}^{(\mathrm{cry})} = 3J_c + \frac{1}{r}(J_c + 4\sigma) + \mathcal{O}(\frac{1}{r^2}) \tag{5.22}$$

$$e_{\min}^{(\mathrm{cam})} = 3J_c + \frac{1}{r}(-3J_c + 4J_l + 4\sigma) + \mathcal{O}(\frac{1}{r^2}) \tag{5.23}$$

The sign of $J_c - J_l$ is thus sufficient to determine the most stable geometry: it is the camembert for $J_l < J_c$ (or equivalently $x_l < y_c$) and the crystal otherwise. Near that boundary, the camembert is of infinite size, as was shown above (if is infinite when $y_c < 2x_l/2$). For this reason, it is sufficient to compare the crystal energy with that of the camembert at infinite r. We obtain the first boundary of the phase diagram, which is in the lower left part in Figure 5.5a. Above the line $y_c = x_l$, this corresponds to the set of parameter (3) plotted in (b): both the crystal and the camembert have decreasing energy as r increases, but the camembert energy (solid line) is below the crystal energy (dashed line). Below the line $y_c = x_l$, this corresponds to plot (4), and the crystal energy is below the camembert energy.

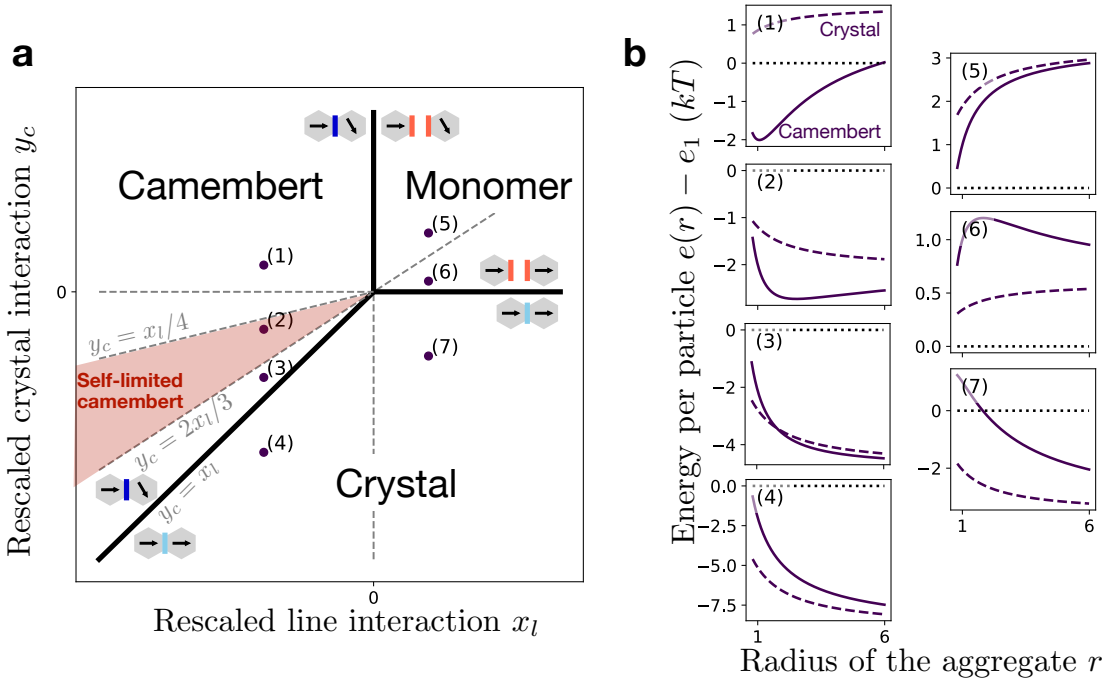

Figure 5.5: Camembert are stable if the crystal interaction is attractive ($x_l < 0$), and if the line interaction is stronger than the crystal interaction ($x_l < y_c$). They are self-limited for a large range of parameters. a) Phase diagram. The boundaries of the stability of each geometry is in black. The red zone corresponds to camembert of radius larger than one and smaller than ∞. b) For some specific set of parameters (purple points in (a), we plot the energy per particle for the camembert (solid line), crystal (dashed line) and monomer (dotted line). The most negative energy sets the stable geometry and the stable size. In (2), the camembert energy is minimum for a size larger than one.

5.2.4.2 Crystal-monomer boundary

The boundary between the crystal and the monomer geometry is computed easily. The minimum of the crystal energy is $e_{\min}^{(\text{cry})} = 3J_c$, and that of the monomer is $e_1 = 6\sigma$. We recover the fact that the crystal is more stable than the monomer if $J_c < 2\sigma$ ($y_c < 0$), if it is locally more favored to bind two particles with a crystalline interaction than leave them unbound. This can be verified by comparing plots (6) and (7) of Figure 5.5b, which corresponds to a set of parameter that are apart to the boundary between the monomer and the crystal geometry in Figure 5.5a. In (6), the energy of the crystal (solid line) is above the energy of the monomer (dotted line), as the opposite in (7).

5.2.4.3 Camembert-monomer boundary

We compare the energy of a monomer with that of a camembert. The camembert is unstable when $x_l > 0$ and $3y_c > 2_l$. In this region, the crystal is also less favored than the monomers. As a consequence, the most favored geometry is the monomer. We also determined in Sec. 5.2.2 that the camembert of radius 1 is always more stable than a partially assembled hexamer. The energy of the smallest camembert (when $r = 1$) is then $e(r = 1) = x_l + e_1$. We now show that for $x_l < 0$ the camembert is always more stable than the monomer.

We compute the minimal energy of a camembert in the region where its size is finite, $e^{(\text{cam})}(r_1)$, and compare it to the energy of a monomer, e_1. $e^{(\text{cam})}(r_1)$ is computed by injecting the value of r_1 (eq. 5.20) in the energy per particle of a camembert (eq. 5.18). To simplify the equations, we introduce the variable α (eq. 5.24), which simplifies the expression of r_1 (eq. 5.25).

$$\alpha = -x_l/(3y_c - 2x_l) > 0 \tag{5.24}$$

$$r_1 = \alpha(1 + \sqrt{1 + \alpha^{-1}}) \tag{5.25}$$

With this new variables, we can compute the numerator (eq. 5.26) and the denominator (eq. 5.27) of $e^{(\text{cam})}(r_1)$, and deduce a simplified equation of $e^{(\text{cam})}(r_1)$ (eq. 5.28 = (5.26)/(5.27)). We deduce the energy difference between the stable camembert and the monomer for any set of parameters where the size of the camembert is finite (eq. 5.29).

$$r_1(r_1 + 1) = 2\alpha^2(1 + \sqrt{1 + \alpha^{-1}}) + \alpha(2 + \sqrt{1 + \alpha^{-1}}) \tag{5.26}$$

$$(3y_c + e_1)r_1^2 + (-3y_c + 4x_l + e_1)r_1 - 2x_l = (3y_c + e_1)(r_1(r_1 + 1)) - 2(3y_c - 2x_l)r_1 \tag{5.27}$$

$$e(r_1) = (3y_c + e_1) - 2\frac{3y_c - 2x_l}{r_1 + 1} \tag{5.28}$$

$$e(r_1) - e_1 = 3y_c + -2\frac{3y_c - 2x_l}{r_1 + 1} \tag{5.29}$$

If $y_c < 0$, this is always negative and the camembert is always more stable than the monomer. If $y_c > 0$, the stable radius is also of value 1 as was shown in Sec. 5.2.3.2, and the camembert is then more stable than the monomer. Therefore, in the regions where the camembert is stable, and more stable than the crystal ($x_l < 0$, $y_c > x_l$), it is also always more stable than the monomer. We identified the last boundary of the phase diagram: when $x_l < 0$, the camembert is more stable than the monomer, which corresponds to the set of parameters at position (1) in the phase diagram of Figure 5.5a. The corresponding energies plotted in Figure 5.5b show that the camembert is more stable than the crystal (solid line below dotted line). When $x_l > 0$ (panel (5) or (6)), the monomer is more stable

than the camembert (solid line above dotted line). In Figure 5.5a, we also show in red the region where the camembert is self-limited, *i.e.* of radius larger than 2 and less than $+\infty$. This region of interest is within the boundaries where the camembert is the most stable geometry.

In panel (2) of Figure 5.5b, we see that the energy per particle in the camembert has a minimum, but that this minimum is very shallow. We remind that this is the pure energetic computation, and that there are no entropic effects taken into account in this discussion. We expect the free-energy to be slightly shifted from the energy when the temperature is above zero. Because the minimum of the energy is shallow, a small entropic effect might result in an important shift of the equilibrium size of the aggregate.

We identified the region in the parameter space of the particles with line and crystalline interactions where the camembert is the most stable: it requires that the line interaction is more favored than the surface interaction, and that the line interaction is more favored than the crystal interaction. We also showed that there is a large part of the stability region of the camembert where they are self-limited, *i.e.* when the equilibrium radius is larger than 2 and finite. This suggests that we identified a relevant mechanism for self-limiting assembly that only rely on directional short range interaction of the particles. However, we only compared the energies of three geometries in this section, because they were the only one that we could think of. We are not yet guaranteed that there is no alternative organization of the particles of lower energy than the camembert in the region of interest.

5.3 Lattice simulations validate the stability and size control of camembert aggregate

To verify that there is no alternative organization than the one we enumerated in the previous section for the anisotropic particles with competing line and crystal interactions, we use the interaction map model and numerical simulation introduced in Chapter 2. We show that for an interaction map that corresponds to the interaction of the particle introduced in Sec. 5.1, the equilibrium configuration determined numerically corresponds to the camembert aggregate, in the expected parameter regime. We run numerical simulation at temperature one, which enables to confirm that the camembert aggregate are not only observed at zero temperature. In Sec. 5.3.1, we explain our simulation choices to guarantee that the aggregates we observe in the simulations are at equilibrium. In Sec. 5.3.2, we explore the parameter space and show that the phase diagram determined in Figure 5.5 correctly predicts the equilibrium geometry of the aggregates. In Sec. 5.3.3, we quantitatively compare the prediction of the size and of the energy per particle in the aggregate, and show that while the energy is well predicted by the analytical computation of Sec. 5.2, it is not the case of the size of the aggregates.

5.3.1 Simulation method

In this section, we show how to implement the ideas of competing energy in the interaction map model of Chapter 2. Because we want to measure effects of size-limitation at equilibrium, we need to carefully choose the parameters of the simulated annealing to guarantee that the measured aggregates are equilibrated, and that their size-limitation is due to thermodynamic effects only and are independent of the annealing protocol.

From the design of the particle proposed in Sec. 5.1.1, which we show again in Figure 5.6a, it is straightforward to identify the corresponding interaction map. There is 3 pair of faces that correspond to the crystal interaction and two pair of faces that correspond to the line interaction. The interaction map is shown in Figure 5.6c. The crystal interactions are colored in light blue, and correspond to an interaction energy J_{crystal} and the

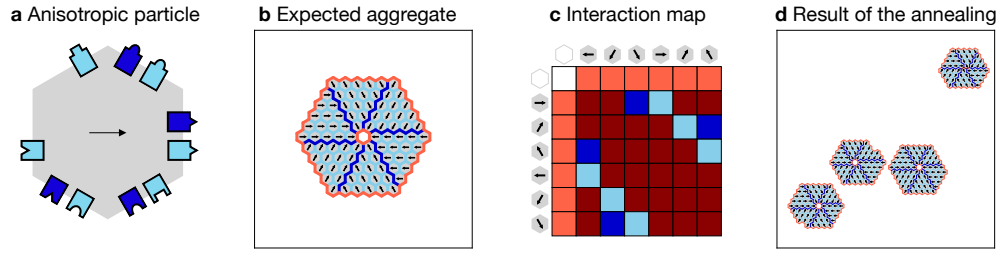

Figure 5.6: We define the interaction map of the anisotropic particle with directional interactions, and it assembles in the expected aggregate geometry in numerical simulations. a) The particle has two types of interactions, and b) is expected to aggregate in a camembert geometry. c) The interaction maps are chosen such that the dark blue entries have energy J_{line}, the light blue entries have energy J_{crystal}, the surface interactions (pink) have energy σ. All the other interactions (dark red) have energy J_{∞}. d) Result of numerical simulation where the interactions are shown with the same color code. $J_{\text{crystal}} = 0.5kT$, $J_{\text{line}} = -8kT$, $\sigma = 6kT$ and $J_{\infty} = 15kT$)

line interaction are shown in dark blue and correspond to an interaction energy J_{line}. The other possible interactions between two faces of the particles are repulsive, they are shown in dark red, and correspond to an interaction energy J_{∞}. In Chapter 2, we explained that the interaction of the particle could always be set to zero if the number of particle is fixed. Here, however, we set the interaction energy to be σ, because it makes the connection between the design of the particle and the interaction map easier to interpret. The line and crystal interactions (J_{line} and J_{crystal}) were always compared to 2σ in the analytical study of the section 1, and this will be similar in the numerical simulation during the energy comparison in the elementary steps of the Monte-Carlo simulation (see Sec. 2.2 of Chapter 2). Note that the colors in the interaction maps do not refer to the level of the energies as in previous chapters, but to the type of interaction.

We show the result of the simulated annealing for such an interaction map in Figure 5.6d. We give details about the parameters chosen for the simulation below. The equilibrium configuration of the particles are represented with a different convention than in previous chapters: the orientation of the particle is indicated with an arrow, and the bonds between two particles are shown with the same color code as that of the interaction map. We see that the aggregates simulated numerically is a camembert (Figure 5.4b).

Here, we explain our choices for the size of the system, number of particles, and annealing protocol. We run simulations of 270 particles on a triangular lattice of size 50×50. We choose the number of particles such that the radius of a camembert with that many particles is an integer: if $r = 9$, $N(r) = 270$ (eq. 5.14). This large number of particles also ensure that size limitation can be observed. For instance, if the equilibrium size of the aggregate is $r = 4$, there should be around 4 aggregates in the equilibrium configuration. The density of particles is low ($N_{\text{particles}}/N_{\text{sites}} = 0.108$) which guarantees that the system is diluted and that the particles will not aggregate because of crowding effects.

To ensure that the aggregates studied in the next sections are at equilibrium, we compare different temperature protocol (see Sec. 2.2 of Chapter 2), and we choose the initial temperature of the annealing and the number of Monte-Carlo steps such that the energy of the system at the end of the annealing is independent of the chosen protocol. We expect that if the number of annealing step is too low, the system is not well equilibrated, and there is a dependence of the energy in the number of step. Beyond a given number of steps, we expect the final energy to be independent of the number of annealing step: if the system reaches its equilibrium configuration, more Monte-Carlo steps should not make a difference. In Figure 5.7, we show the final energy of the system for different equilibrating

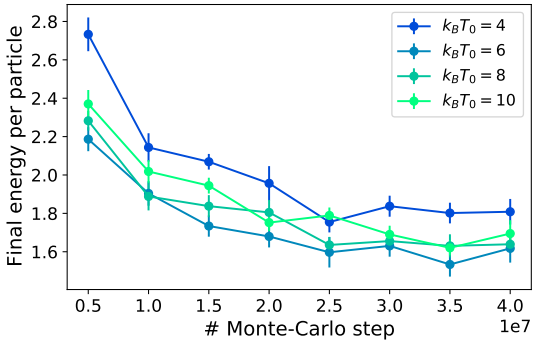

Figure 5.7: We choose the number of Monte-Carlo step and the initial temperature of the annealing such that an increase of the number of step and a change of intial temperature does not decrease the energy. We anneal a system of size 50×50, 270 particles, and binding energies $(J_{\text{crystal}}, J_{\text{line}}, \sigma, J_\infty = 0.5kT, -8kT, 6kT, 15kT)$

time and different initial temperature, and for a chosen interaction map, for an annealing with 100 temperature steps. The final energy decreases with the number of steps up to $\approx$ 20 millions of steps, and converges after. We choose to perform 25 millions steps for each annealing procedure. This number of steps can be performed in a reasonable computational time (one simulation is a few minutes). It corresponds to 1000 steps per temperature and per site of the system. There is also a small dependence in the initial temperature of the annealing, which is not monotonous. Indeed, for a fixed number of step, there might be more or less steps within the range of temperature where the particles aggregate. We choose $k_BT_0 = 6$, which leads to the minimum equilibrium energy for the chosen interaction map.

We choose interaction maps such that the equilibrium configuration leads to the camembert geometry predicted in Sec. 5.1.1. We determined the optimal temperature protocol for a given set of parameters. However, we did not perform such calibration of the annealing protocol for each interaction map, because it would require too much computational time.

5.3.2 Verification of the phase diagram

We now explore the parameter space of the interaction energies in the numerical simulation and compare it to the phase diagram predicted in Figure 5.5, and show that the camembert aggregate is the most stable geometry in the expected region. We however show that there are some fluctuations of the organization of the particle at the surface of the aggregates.

For different values of J_{line} and J_{crystal}, we perform the simulated annealing with the parameters described in Sec. 5.3.1 and compare the geometry of the aggregate (by looking at the snapshot) with the predicted geometry. We choose $J_\infty = 15kT$ and $\sigma = 6kT$. We choose the different values of J_{line} and J_{crystal} in order to collect several images in the region of interest colored in red in Figure 5.5a, for which the camembert are expected to have a finite size at equilibrium. We show the results in Figure 5.8. There is a direct correspondence between the analytical prediction and the simulation results. The points in orange correspond to trivial situations: (u) and (v) are such that the crystal interaction is unfavored, and the stable configuration is a hexamer (*i.e.* a camembert or radius $r = 1$). (x) and (y) are such that the line interaction is unfavored, and the stable configuration is crystal. (w) is such that both interactions are unfavored, and the stable configuration is a monomer.

Most of the green points are in the region where the camembert are expected. We do observe aggregates of finite but large size with disclination line in the region above the

$r = \infty$ line. Camembert aggregates of infinite size are observed below this line (e, j, m). The set of interactions (n) corresponds to a point very close to the boundary: there are only two disclination lines in the aggregate. For the points in the upper left region (a, f, k, p, b, g), the disclination lines are longer than the radius of the aggregate. We call those particles *extrusions*. We did not predict such extrusion for the camembert geometry in the analytical study.

As predicted, the size of the equilibrium aggregates also increases when the effective crystal energy decreases (from top to bottom), and when the line energy increases (from left to right). This suggests that the mechanism at stake is the decrease of the amount of defect lines per particle when the line interaction become less favored than the crystal interaction. This evaluation is not yet quantitative.

With the numerical simulation, we addressed the limitation of the analytical computation of the phase diagram, which was constructed only by comparing a few geometries. Here, the equilibrium configuration of lattice particles verifies the expected phase diagram. The non-zero temperature do not seem to shift the boundaries of the phase diagram predicted at zero temperature. We derived the phase diagram by considering symmetric aggregates, *i.e.* for which the radius of the aggregate is an integer. It did not lead to mispredictions of the regions of the phase diagram when the aggregates can be non-symetric.

5.3.3 Energy and size comparison

We predicted the equilibrium size of the camembert in the analytical derivations. This prediction resulted from the assumption that the aggregate are in the configuration for which the energy per particle is minimal. There was no consideration of entropic or kinetic effect in the determination of the equilibrium radius. Here, we show that the equilibrium energy of the aggregate measured for $k_B T = 1$ is well predicted by the analytical computations, but that there are important differences between the camembert sizes in the simulation and in the analytical computation.

We compute the averaged equilibrium size of the clusters in the system by performing a pondered average of the size of all the cluster within one system and within different simulated annealing for the same interaction map, as was explained in Chapter 2. If the aggregates of the simulation k are of sizes $\{N_i^{(k)}\}$, where $N_i^{(k)}$ is the size of the i^{th} aggregate in the k^{th} simulation, the average size of the aggregate is:

$$\langle N \rangle = \frac{\sum_k \sum_i \left(N_i^{(k)}\right)^2}{\sum_k \sum_i N_i^{(k)}} \tag{5.30}$$

The pondered average ensures that this is the average size of a cluster upon drawing one random cluster, and not upon drawing one random particle. This measure is equivalent to the weight average molecular weight used in polymer chemistry (and not the number average molecular weight). We also measure the standard deviation of the values of $\left(N_i^{(k)}\right)^2$, which is used to compute the error bars on the Figure shown below. We also compute the energy of the system, as explained in Chapter 2, by performing the scalar product between the density map and the interaction map. Here it corresponds to the following operation

$$e(J_{\text{crystal}}, J_{\text{line}}, J_\infty, \sigma) = \frac{1}{N_{\text{particles}}}(\langle n_{\text{crystal}} \rangle J_{\text{crystal}} + \langle n_{\text{line}} \rangle J_{\text{line}} + \langle n_{\text{surface}} \rangle \sigma + \langle n_\infty \rangle J_\infty) \tag{5.31}$$

where $N_{\text{particles}}$ is the total number of particles in the box, and $\langle n_{\text{line}} \rangle$ is the average number of interactions between two neighboring particles that are in a line configuration (and similarly for the crystal, surface, or forbidden configuration), measured during N_{average}

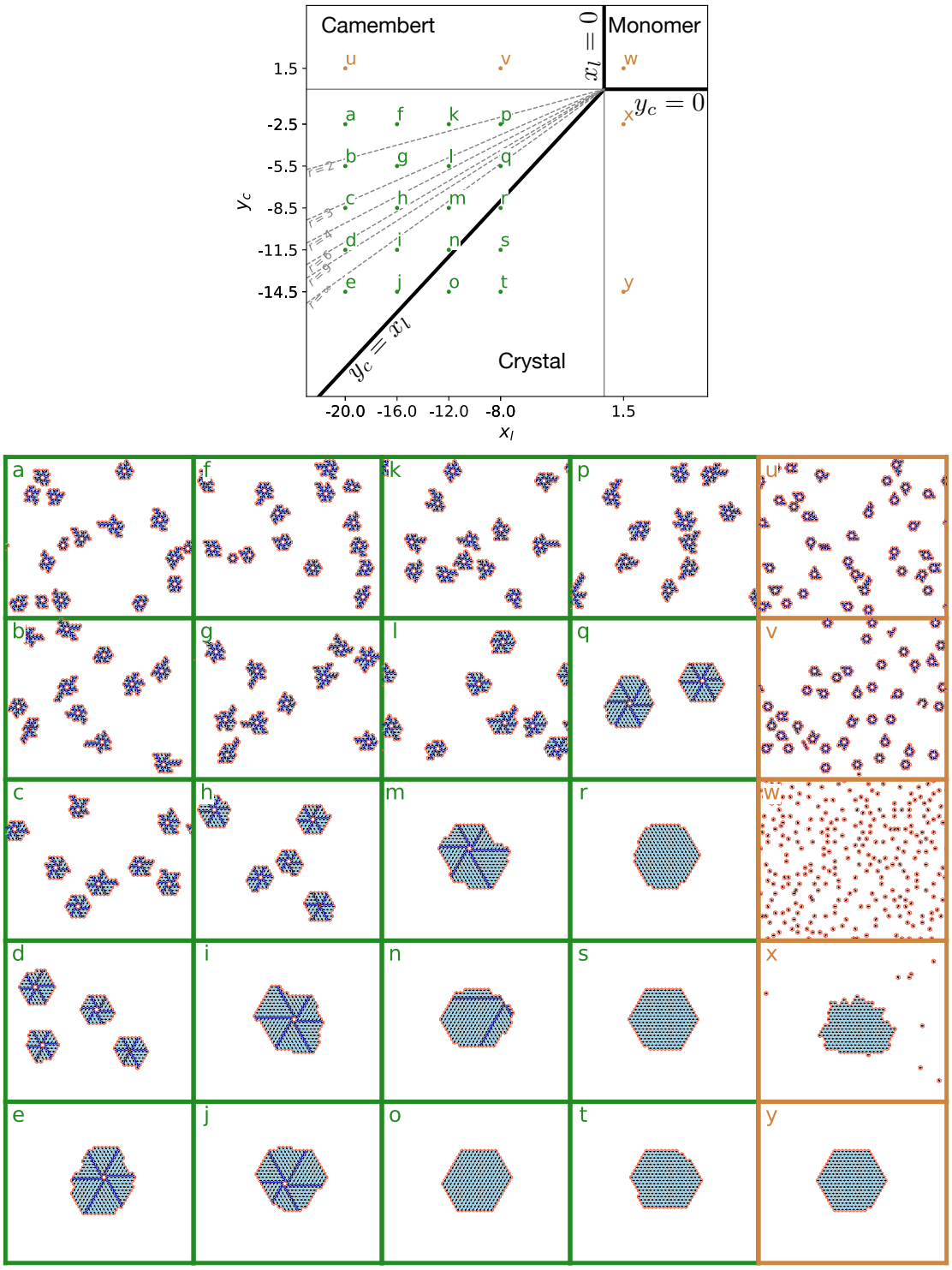

Figure 5.8: We observe camembert in numerical simulation at finite temperature within the parameter range where they were predicted to be stable at zero temperature. a) Theoretical phase diagram and positions of the parameters for which we run the simulation, b) Snapshot obtained in numerical simulation. Aggregates of finite size are observed above the limit $r = \infty$. $\sigma = 6kT$ and $J_\infty = 15kT$.

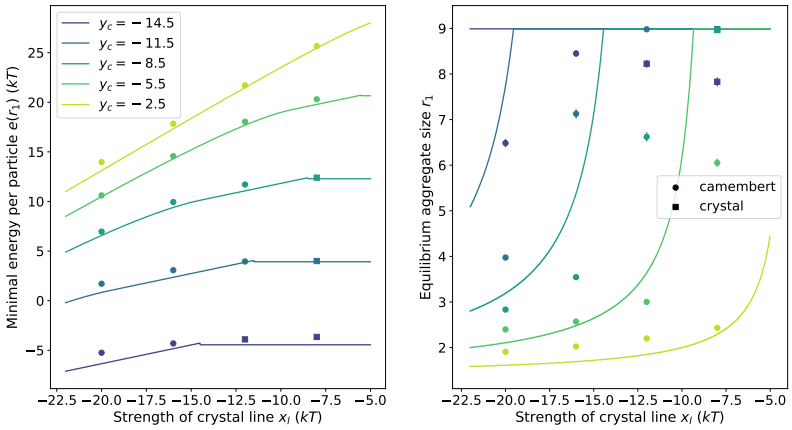

Figure 5.9: The analytical calculation (solid line) predicts the energy per particles in the simulation, but not the size of the aggregates (points). For each set of parameter, we compute the analytical minimal energy per particle and equilibrium radius, with formulas for the crystal (5.8) when $x_l > y_c$ or the camembert (5.18) when $x_l > y_c$. The points correspond to the simulation results of Figure 5.8.

Monte-Carlo steps. Here N_{average} is 1000 times the number of sites in the lattice. Both the size and the energy are averaged over 200 simulated annealing. We compare those quantities to the analytical energy of the camembert (or crystal, depending on the region of the diagram) predicted in Sec. 5.2.3.1 and 5.2.1 and to the equilibrium size of the camembert predicted in Sec. 5.2.3.2. The size is capped to the total number of particles in the system. Both comparisons are plotted in Figure 5.9. We plot the radius r corresponding to an aggregate size $N(r)$ (eq. 5.14). The interaction parameters for which we perform simulations correspond to the one of Figure 5.8.

The measure of the energy per particle is in agreement with the analytical result (left plot): the dots (measure in the simulation) are on the solid lines (analytical results). However, there is some quantitative disagreement between both (right plot): large aggregate tend to be smaller than predicted and vice versa.

There are two possible explanations for these discrepancies: the size of the aggregate was predicted from the minimization of the energy per particle, which assumes that the system is at zero temperature. Since the simulations are performed at temperature $k_B T = 1$, there might be a shift of the equilibrium size because of entropic effects. This would explain that the energies are in agreement, and not the sizes. Another possibility is that the numerical simulations are not well equilibrated, despite the precautions we took to ensure that the energy of the system did not depend on the annealing protocol.

5.4 Entropic and kinetic effects

The camembert geometry of the aggregate is observed in the numerical simulation when the analytical simulation predicted that it was the most stable geometry. However, the size of the aggregate is not correctly predicted. Here we show that both insufficient equilibration time and entropic effect could explain these differences. In Sec. 5.4.1, we show that increasing the annealing time changes the equilibrium size of the cluster without completely explaining the difference with the analytical prediction. In Sec. 5.4.2, we show taking into account the translational entropy of the aggregates in the analytical prediction of the size does not predict the aggregate size measured in the simulation. In Sec. 5.4.3, we also take into account the entropy due to fluctuations of the number of particles at the surface of the aggregate, and show that it only shifts the equilibrium size of small

aggregates.

5.4.1 Equilibrium is not completely reached in simulation

In this section, we choose a set of interaction energies leading to the camembert geometry for which the average size measured in the simulation is different from that of the analytical prediction. We vary the initial temperature of the annealing, the size, and number of particles in the system, and the number of annealing steps. The initial temperature and size of the system do not have a clear influence on the equilibrium size of the aggregate, but the size of the aggregate increases with the number of annealing step in this situation.

We consider the system labelled (d) in Figure 5.8 ($x_l = -20, y_c = -11.5$) for which the measured equilibrium radius is around 4, and the predicted equilibrium radius is around 7. We vary the initial temperature of the annealing while keeping the total number of Monte-Carlo step constant. If the initial temperature is too low, the system might be trapped in the local minimum of the configuration space, because the temperature is never sufficient to jump above energy barriers. If the initial temperature is too large, a large number of Monte-Carlo steps are performed at high temperature for which the particles do not aggregate because the system is dominated by entropic effect. Therefore, there should be an optimal intermediate temperature for which the energy is the lowest. We vary the initial temperature of the annealing between $k_BT = 4$ and $k_BT = 20$ and show the averaged equilibrium size of the aggregates in each case in the left panel of Figure 5.10: the size of the aggregate is maximal when the initial temperature of the annealing is around $k_BT = 6$, which we chose.

The numerical simulation are performed at finite sizes and finite number of particles. This might affect the equilibrium size of the aggregates as follows: if the equilibrium size of the cluster is 80 particles, and there is 100 particles in the system, the equilibrium configuration in the system should be such that there is 50 particles in per aggregate, which is below the equilibrium size. However, this effect should vanish upon varying the number of particles in the system. Here, we vary the number of particles such that the radius of the aggregate where all the particles are aggregated is between $r_{\max} = 4$ and $r\max = 12$. We remind that the relation between the number of particles in a cluster N and its radius r is $N(r) = 3r(r + 1)$. We also increase the number of sites in the system such that the density of particles is kept constant. We show the result in Figure 5.10 middle. Except for the small sizes, we do not observe an effect of the number of particles.

Finally, we vary the total number of Monte-Carlo steps for the annealing. As explained in Sec. 5.3.1, we expect the energy of the system (or the size of the aggregate) to reach a plateau after a certain number of steps. We show the results in Figure 5.10 right. While the number of steps did not seem to have an influence on the final energy of the system beyond 1000 steps per temperature and per lattice sites (Figure 5.7), the averaged size of the cluster keeps increasing after that limit. After 1600 Monte-Carlo steps per temperature and per lattice site, the clusters have an average size of 4.4 (which correspond to around 70 particles). This is still far below the expected equilibrium size which is around $r = 7$, *i.e.* 170 particles).

The size and initial temperature do not influence the averaged size of the aggregate for the set of parameters we chose, but the number of Monte-Carlo step does. With our code, we were not able to push the computation of Figure 5.10c further than 1600 Monte-Carlo steps per lattice site and per temperature. Indeed, each data point corresponds to 200 different annealing, each of them lasting around one hour. From this results, it is not clear whether the measured equilibrium radius would plateau at $r = 6$ in this case, as predicted by the analytical computation, or below.

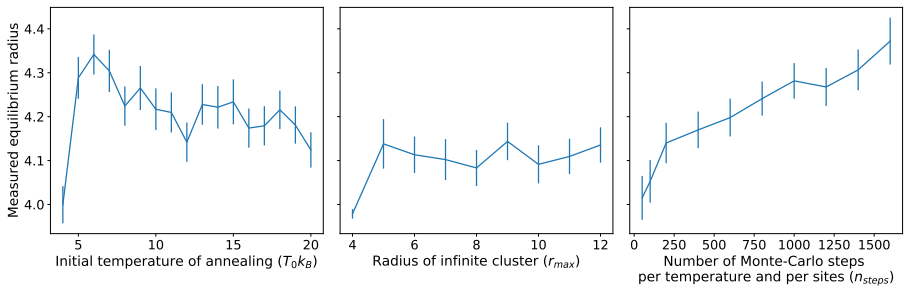

Figure 5.10: The averaged radius of the aggregate varies with the initial temperature of the annealing, the number of Monte-Carlo step, but not with the size of the system. It suggests that the system is not well equilibrated. We measure the equilibrium radius of the system for which $x_l = -20$, $y_c = -11.5$ (system 'd' of Figure 5.8) as a function of varying hyperparameters, and standard error. By default, $k_B T_0 = 7$, the radius of an infinite aggregate is $r_{max} = 9$ (which corresponds to 270 particles in a system with a number of sites $N_{sites} = 50 \times 50$), and we run $n_{steps} = 200$ Monte-Carlo steps per system site and per temperature. Each data point is an average over 200 realisations.

5.4.2 Size distribution

A possible explanation for the discrepancies between the theoretical equilibrium size of the camembert aggregate (at zero temperature) and the one measured in simulation (measured at temperature one) is some entropic contribution that could shift the equilibrium size. To test this hypothesis, we compute the free-energy per particle of the system, which is the sum of the energy per particle computed in Sec. 5.2.3, and of the entropic contribution. Here, we take into account the translational entropy of each cluster of size n, which scales as $\ln 1/n$ [79]. We then compute the expected size distribution of the aggregates, and show that taking into account this entropic effect does not explain the size of aggregates measured in the simulation for large but finite sizes.

We consider an ideal gas of aggregates where there are N_n aggregates of size n. Without interaction between the aggregates, the partition function of the total system Z is the product of the partition function of the individual aggregates z_n, weighted by their number of configurations.

$$Z = \prod_i \frac{1}{N_n!} (z_n)^{N_n} \tag{5.32}$$

The partition function of an individual aggregate counts all the possible positions of the aggregate in the system. There are N_{sites} possible positions. It reads $z_n = N_{\text{sites}} e^{-\beta n e^{(\text{cam})}(n)}$, where $e^{(\text{cam})}(n)$ is the energy per particle in a camembert aggregates, computed in eq. 5.18 ($\beta = k_B T$). With the Stirling approximation for the factorial, we obtain the free energy

$$F = -\frac{1}{\beta} \ln Z = \sum_n \left[-\frac{1}{\beta} N_n \ln \frac{\mathrm{e} N_{\text{sites}}}{N_n} + N_n n e^{(\text{cam})}(n) \right] \tag{5.33}$$

If we now introduce the concentration of cluster of size n, $c_n = N_n / N_{\text{sites}}$, we obtain for the free energy per unit volume

$$f = \frac{F}{N_{\text{sites}}} = \sum_n n c_n \left[\frac{1}{n\beta} (\ln c_n - 1) + e^{(\text{cam})}(n) \right]. \tag{5.34}$$

The total concentration of particles being fixed, we differentiate f with the constraint that the number of particle is fixed: $\sum_n n c_n = c^{(0)} = \frac{N_{\text{particles}}}{N_{\text{sites}}}$. The thermodynamic potential associated to this constraint is μ, and reads

$$\mu = \frac{\partial f}{\partial c_n} = \frac{1}{\beta} \left(\ln c_n + n e^{(\text{cam})}(n) \right) = e_1 + \ln c_1 \tag{5.35}$$

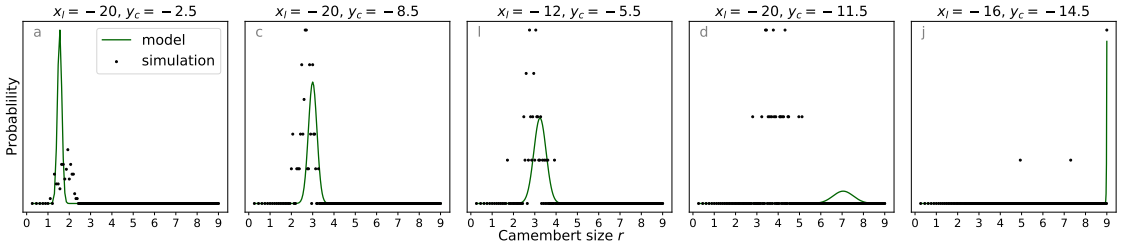

Figure 5.11: The size of aggregates of intermediate sizes in numerical simulation is not well predicted by the minimization of the free energy, taking into account the translational entropy of the aggregates (eq. 5.36). The top-left letters on each figure corresponds to the references of the images in Figure 5.8.

At equilibrium, $\frac{\partial f}{\partial c_n} = 0$. From this, we obtain the general expression for the equilibrium concentration of aggregates of size n:

$$c_n = \left(c_1 \mathrm{e}^{-\beta(e^{(\mathrm{cam})}(n) - e_1)} \right)^n \tag{5.36}$$

This expression depends on c_1 which can be determined by solving the equation on the particle concentration $\sum_n k c_n = c^{(0)}$.

For a given set of line, crystal and surface interaction (related to the rescaled line interaction x_l, the rescaled crystal interaction y_c and the energy of a monomer e_1), we derive the theoretical equilibrium concentrations of aggregates of size n, and compare it to the size distribution of the aggregates measured in the numerical simulations. With this measurement, we can compare the equilibrium average size of the aggregate, as we did in Figure 5.9, with the entropic correction. We can also compare the distribution of sizes around this average, between the numeric and analytic computation. We show the results in Figure 5.11, for some set of parameters chosen in Figure 5.8, and indicated with the same letters.

When the equilibrium size is small (the equilibrium radius is between $r = 1$ and $r = 3$), the numerical computation and theoretical prediction are in agreement (panels a, c and l of Figure 5.11). For a very large but finite aggregate (plot d of Figure 5.11), the measured size is far below the expected size. For the aggregates of infinite size (panel j), we see that the system is not completely equilibrated, because several smaller sizes of aggregates are observed, but peak of concentrations around a smaller size. For intermediate sizes of aggregate (panel d), it seems that the equilibrium is shifted towards intermediate size, with measured distribution that are clearly peaked around an average which is lower than the one expected.

For aggregates of intermediate sizes, we measure a shift in the size distribution. If the simulation is at equilibrium (which is not completely sure from the results of Sec. 5.4.1), there might be another entropic contribution than the translational entropy, which stabilizes the aggregate of smaller sizes at finite temperature.

5.4.3 Entropic contribution are negligible

Another hypothesis to explain the mis-prediction of the equilibrium radius is the entropic gain that arises from fluctuations of the surface of the aggregate. In some examples of camembert aggregates shown in Figure 5.15 (such as panels b and g), there are some extrusions at the surface of the aggregate: the defect line is longer than the radius of the bulk. The energetic cost of displacing a particle in one extrusion and placing it in another extrusion is zero, and the entropic cost is large (there are a lot of possible extrusions where to displace the particle). This effect might stabilize the smaller aggregates: in a system where the aggregates are smaller, there are more possible extrusions, and more entropic gain of displacing a particle from one extrusion to the other. This qualitative phenomenon is also true for the displacement of the particle from any spot of the surface to any other spots. There is an energetic price to pay to remove or add a particle at the surface of the aggregate, but also an entropic gain, that is increased if the number of surface particles is increased. Here we evaluate the free energy correction associated to the displacements of particles from one spot of the surface of the aggregate to the other, and show that this correction to the equilibrium distribution of the aggregate sizes is negligible in our range of parameters.

We call a *defect pair* a missing particle (hole) and an extra particle (bump) at the surface of the aggregate, as illustrated in Figure 5.12. If the number of defect pair is low compared to the number of particles at the surface of the aggregate, there are as many holes as bumps in the system. To enumerate analytically the number of configurations associated to a given number of defect pairs, and the corresponding energetic cost, we need to make some approximations. We first consider that the defects are not interacting, *i.e.* we neglect the energetic gain of having two holes or two bumps next to each other. We also neglect the fact energy gain associated to having a bump on the defect line is larger than that of having a bump elsewhere on the surface. We refer to this set of approximation as model A. We can then take into account the fact that a bump on the defect line can be energetically favored. Then, we assume that the bumps are necessarily on a defect line, and the holes necessarily elsewhere. This is model B. Finally, we will take into account the interactions between the defects: two bumps next to each other are less costly energetically than two bumps not in contact. This is model C. The distinctions between the three models is summarized in Figure 5.12.

We first evaluate the energetic cost of the defects. We compute the cost of having an extra particle or a hole in the bulk surface, η_+ and η_-, the cost of having a bump on the defect line ϵ_+. We also call $\eta_+^{(2)}$ and $\eta_-^{(2)}$ the cost of having a pair of bumps or holes. Finally, we call j, the cost of having a step of height 1 on the surface. Then $2j = \eta_+^{(2)} - 2\eta_+ = \eta_-^{(2)} - 2\eta_- = y_c$. All those coefficients depend on the values of the line, crystal and surface interactions (J_{line}, J_{crystal} and σ), and we compute them by counting the number of each types of interaction in the schematics of Figure 5.12. η_+ and η_- are computed by measuring the energy difference between situation A and the flat surface. ϵ_+ is computed by comparing situation B with the flat surface. $\eta_+^{(2)}$ and $\eta_-^{(2)}$ are computed by comparing the situation C and C' with the flat surface.

$$\eta_+ = 2J_c + 2\sigma = 2y_c + e_1 \tag{5.37}$$

$$\eta_- = -4J_c + 2\sigma = -4y_c - e_1 \tag{5.38}$$

$$\epsilon_+ = J_l + J_c + 2\sigma = x_l + y_c + e_1 \tag{5.39}$$

$$\eta_+^{(2)} = 5J_c + 2\sigma \tag{5.40}$$

$$\eta_-^{(2)} = -7J_c + 2\sigma \tag{5.41}$$

The number of particles at the surface of the aggregates is $M = 6r$, with r the radius of

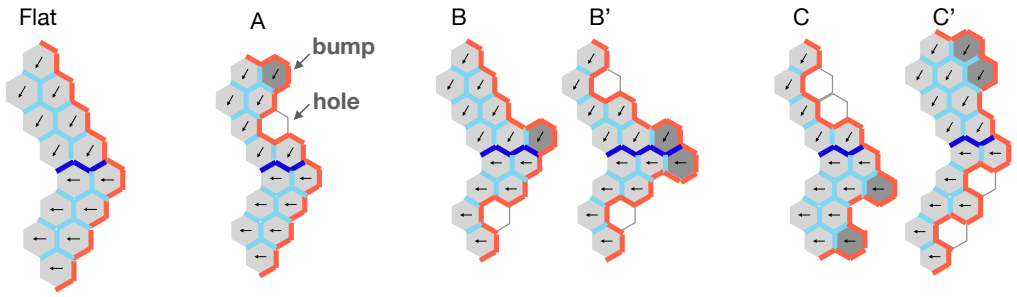

Figure 5.12: We calculate the energy associated with a defect by counting the difference in the number of each type of bonds between a surface with defects and a flat surface. The defects can be not on the disclination line (A), the bumps can be on the disclination lines (B and B') and the holes (C) and bumps (C') can interact.

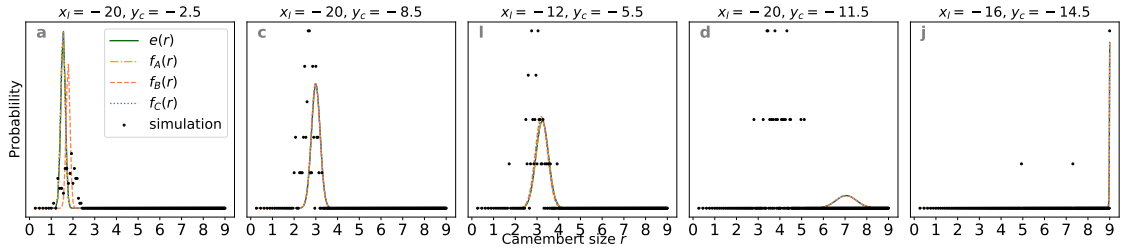

Figure 5.13: The correction to the equilibrium distribution of sizes taking into account the surface fluctuation of the aggregate does not predict the sizes measured numerically when the aggregate size is intermediate

the aggregate.

If the defect pairs are not on the blue lines, and do not interact (model A), we computed the associated shift in the free energy. The number of defect pair is n_0. Then

$$Z_A = \sum_{n_0=0}^{+\infty} M^{2n_0} e^{(-\eta_- - \eta_+)n_0} \tag{5.42}$$

$\eta_+ + \eta_- = -2y_c$. Then, if $M^2 \mathrm{e}^{2y_c} \ll 1$, we can simplify this sum.

$$\ln Z_A = -\ln(1 - M^2 \mathrm{e}^{2y_c}) \tag{5.43}$$

The free energy per unit volume is the one computed in Sec. 5.4.2, corrected with the entropic contribution of surface fluctuation per particle:

$$f_A(r) = e^{(\mathrm{cam})}(r) - \frac{\ln Z_A}{3r(r+1)} \tag{5.44}$$

We compute the equilibrium distribution associated with this free-energy from eq. 5.36, for the same set of parameters chosen in the plots of Figure 5.11, and show it in Figure 5.13 in yellow. The new distribution is indistinguishable from the one computed previous (in green) without the fluctuations: the energetic cost of forming defect pairs is too large.

We now consider the case where bumps are necessarily on a defect line, which is lower in energy (ϵ_+ is smaller than η_+ in the set of parameters we consider). This is model B. We treat each site on the surface independently, and count the number k of extra particle at this site. A site can have an extra particle on a defect line with cost ϵ_+ ($k = 1$) or a hole in the crystalline part with cost η_- ($k = -1$). We associate the conjugate variables λ to the created holes or particles. Then $\langle k \rangle = \frac{\partial \ln Z}{\lambda} = 0$, because the algebraic number of defects is 0 (there are as many holes as bumps). A site on a line is associated with the

partition function $z_{line} = \sum_{k=0}^{\infty} \mathrm{e}^{-k(\epsilon_+ + \lambda)}$ (the lines can grow indefinitely), and a site in the crystalline bulk with the partition function $z_{bulk} = 1 + e^{-\eta_- + \lambda}$ (there is maximum one hole per bulk-site). If there is L line spots, we obtain the following free energy, and after finding the value of λ we obtain the final expression.

$$Z_B = (z_{\text{line}})^L \, (z_{\text{bulk}})^{M-L} \tag{5.45}$$

$$\ln Z_B = -L \ln(1 - 2\frac{M-L}{M}(1+\sqrt{1+\alpha})^{-1}) + (M-L)\ln(1 + \frac{M}{2(M-L)} e^{3y_c - x_l}(1+\sqrt{1+\alpha})) \tag{5.46}$$

$$\text{with } \alpha = 4\frac{L(M-L)}{M^2}\mathrm{e}^{x_l - 3y_c} \tag{5.47}$$

We compute the free-energy $f_B(r) = e^{(\text{cam})}(r) - \frac{\ln Z_B}{3r(r+1)}$ for a given set of parameters, and the corresponding distribution of aggregate size. This is plotted in orange line in Figure 5.13. For aggregates of small sizes (panel a), there is a shift of the equilibrium distribution towards larger sizes, which seem to match better the numerical computation. For larger sizes, however, this correction is again negligible.

We now consider that the defects can interact (model C). As it has been studied for roughening problems, we introduce h_i, the absolute difference in the number of defects between particle at sites i and $i+1$, as it has been done to study the roughening transition of crystals (Chaikin Lubensky - chapter 10.6) [125]). If the number of bumps is larger on site i than on site $i+1$ of h_i particles, there are h_i particles with exposed surfaces. Then the partition function associated to a site is simply $z_i = \sum_{h=0}^{\infty} e^{-h_i j}$ with $j = -y_c/2$ the unit cost of the interface. With this notation, the sites can then be treated independently:

$$Z_C = (z_i)^M \tag{5.48}$$

$$\ln Z_C = -M \ln\left(1 - e^{y_c/2}\right) \tag{5.49}$$

We compute the free-energy $f_C(r) = e^{(\text{cam})}(r) - \frac{\ln Z_C}{3r(r+1)}$ for a given set of parameters, and the corresponding distribution of aggregate size. This is plotted in blue line in Figure 5.13. Once again, this correction of the free energy does not shift the equilibrium configuration.

The free energy correction due to surface fluctuation is negligible in the parameter regime where large aggregates are observed, and it does not explain the average size of aggregates measured in the simulation. The computation relied on the hypothesis that the number of defect at the surface of the particle is low. This approximation appears reasonable: the energetic cost of a defect is low compared to the entropic gain, and we do not expect our computations to be dependent on this approximation.

5.5 Design of fibrous aggregate of controlled width

In sections 5.2, 5.3 and 5.4, we discussed a specific particle design for which favored defect lines within the crystalline packing could limit the size of the assembly. This size-control results from the competition between interactions of an anisotropic particle. This concept could be used more broadly than the specific camembert design we described. Here, we introduce a minor change in the way the designed particle interact locally, so that they will self-assemble into fibers of controlled width which depend on the relative strength of the local interactions. We present the fiber design (5.5.1) and show that fibers of finite and large width are observed in simulation (5.5.2).

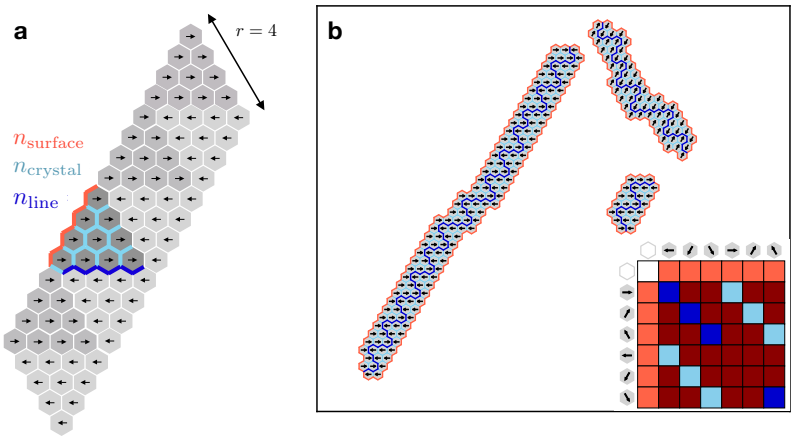

Figure 5.14: The line directional interaction is chosen such that the particle assemble into a fiber of finite width. a) The energy of an aggregate is measured by counting each type of interaction in a triangular slice (dark gray) b) The chosen interaction map leads to fibers of large width in the numerical simulation for $J_{\text{crystal}} = 2kT$, $J_{\text{line}} = -4.625kT$.

5.5.1 Design principles of a fiber stabilized by geometrical defects

In the camembert geometry, the size-limitation mechanisms come from the competition between the line and the bulk interaction within one triangular group of particles (in dark gray in Figure 5.3), with the number of line contacts scaling as the size of the triangle and the number of crystal contacts scaling as the square of its size. Here, we show that we can generalize this concept to another geometry.

We illustrate in Figure 5.14a how we can design an anisotropic particle that has both a crystal and a line interaction, but for which the relative orientation between the line interactions are such that the triangular portions of the crystal (dark gray in the figure) will alternate instead of forming a hexamer of triangle. We also show in 5.14b, the corresponding interaction map, and the result of a numerical simulation, which is in agreement with the predicted aggregate. The fiber is not infinitely long, because the energetic cost of having several smaller fibers instead does not scale like the number of particles in the aggregate.

The same calculations derived before holds, with slight changes in the coefficients in eq. 5.18. However, in that case we need to make the assumption that the fiber is of infinite size, and therefore we do not count the energy cost of the surface of the tips of the fiber.

The number of crystal, line, and surface interactions per gray triangle is

$$n_{\text{crystal}} = 3r(r-1)/2 + 1 \tag{5.50}$$

$$n_{\text{line}} = 2r - 1 \tag{5.51}$$

$$n_{surface} = 2r \tag{5.52}$$

The energy per particle is the total energy of a gray triangle, divided by the number of particles in the triangle. With the change of variable introduced in equations (5.1, 5.2, 5.3), we find

$$e^{(\text{fib})}(r) = \frac{(3y_c + e_1)r^2 + (-3y_c + 4x_l + e_1)r + 2(y_c - x_l)}{r(r+1)} \tag{5.53}$$

This expression is similar to that of the camembert, with a few changes in the coefficients. By minimizing this function with respect to r, we determine the stability region of the fiber, as we did for the camembert in Sec. 5.2.3. We find that it is the fiber is more stable than the crystal and of finite site in the same region as the camembert ($x_l < 0$,

$y_c > x_l$ for stable fibers, $y_c > 2/3x_l$ for fibers of finite size). The equilibrium size of the triangle is given by r_1, while the width of the fiber is $r_1 + 1$

$$r_1 = \frac{y_c - x_l + \sqrt{(y_c - x_l)(4y_c - 3x_l)}}{3y_c - 2x_l} \tag{5.54}$$

The lines where the stable size is an integer number k are again straight lines in the phase diagram, of equation. The larger the size of the desired aggregate, the narrower the region in the parameter space.

$$y_c = \frac{2k^2 - 2k - 1}{3k^2 - 2k - 1} x_l. \tag{5.55}$$

This computation suggests that it is possible to assemble fibers of finite width that depend on the value of the interaction energies of single particles, at zero temperature.

5.5.2 Phase diagram of zigzag fiber

We show that the predicted self-limitation of the width of zigzag fiber is confirmed in lattice simulations at temperature one. We choose a set of parameters such that fibers of variable width should be observed and show the result of the simulation in Figure 5.15. We do observe fibers of finite width above the $r = \infty$ dashed gray line (panels d, m, c, g, l, ...). In some cases, such as panel (c), the width of the fiber seem to be intermediate between two integer values, resulting in the surface of the fiber not being straight, but having steps of one particle (see the fiber in the bottom of panel (c)). When the individual triangular slices are more stable than the crystal configuration, but of infinite size, this cannot be considered a fiber (the width is infinite). The result is a bulky structure, with a zigzag defect line of arbitrary pattern in the middle (panels r, i, n, e, j). When both line and crystal interactions are repulsive, we recover the monomer geometry (w). When the line is repulsive, we recover the crystal geometry (x, y). The equivalent of the hexamers for the camembert is now a fiber of width one, where the particles alternate their orientation such that their interaction is always that of the line (u, v).

With very little adaptation of the initial camembert design, and following the same physical principles, we were able to design fibrillar aggregates for which the width is solely controlled by the local interactions between particles.

5.6 Perspective of experimental realization

We now present our perspectives of experimental verification of the mechanism of limited-assembly introduced above. Indeed, the numerical implementation is highly idealized, the particles are on a lattice, and each pair of interactions can be adjusted with high precision. The simulations also do not take into account the kinetic of the assembly (we remind that a particle can be moved to a very far position on the lattice within one Monte-Carlo move). We also do not account for displacement of a group of particles, which might lead to aggregation of the individual camemberts. An experimental realization of the self-assembly of this type of particle would enable to address these limitations. Here, we show that the mechanism of camembert formation could be tested experimentally with DNA-origami.

In Chapter 1, we described two widespread building blocks for self-assembly: colloids and DNA origamis. The particle we imagined assembling into the camembert geometry needs 5 types of specific interactions: 3 crystalline interactions and 2 line interactions that would work as lock-and-keys (see Figure 5.2c). Coating the colloid surface with specific chemical interactions is a difficult task [63], whereas DNA-origamis, because they interact with strands coming out of their core that bind through base-pair interactions, can be designed with highly specific and anisotropic interactions.

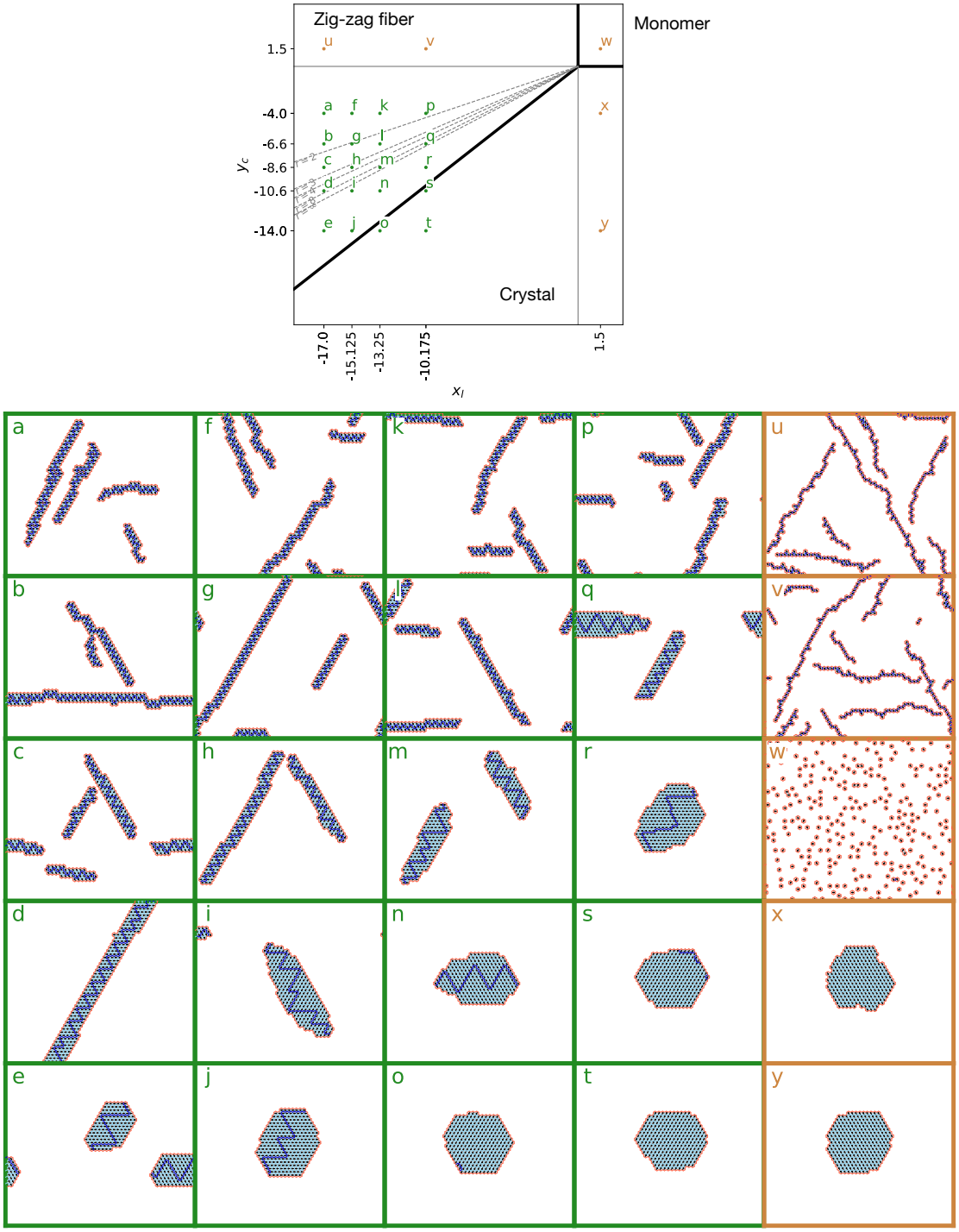

Figure 5.15: We observe fibers of finite width in numerical simulations at finite temperature in the region of the phase diagram where they were predicted analytically at zero temperature. a) Predicted phase diagram and chosen interactions. b) Snapshots of the system for $\sigma = 5kT$ and $J_\infty = 25kT$

In collaboration with the group of Pr. F. Simmel in TUM [42], we plan to adapt the barrel design of DNA origami proposed in [126] to fabricate camembert aggregates of controlled size (Figure 5.16). A schematic of the DNA barrel is shown in Figure 5.16a. Individual DNA strands can then be positioned on the surface of the barrel, at chosen vertical coordinates, and distributed around the barrel with a periodicity $2\pi/6$. We expect that such distribution will ensure that the particles assemble in a triangular lattice, even if the particle is not hexagonal. It is what is suggested by the preliminary crystalline assembly of the barrel particles shown in Figure 5.16c. The 3D version of the 2D particle that can be adapted to the barrel design is shown in Figure 5.16b. The particles are in solution and bind to a substrate. Therefore, it is relevant to compare the self-assembly process with the results of 2D simulations.

If the particles are 3D, they can be in both vertical orientations (see the bottom of Figure 5.16b). Here, we show that it should not affect too much their self-assembly into camembert geometry. We run numerical simulation of the 3D-hexagonal particle (introduced in Chapter 2), on a lattice of size $50 \times 50 \times 1$. As a consequence, the particles can adopt each of the possible vertical orientation by flipping their orientation, but their position is always at $z = 0$. In Figure 5.16c, we show the result of the simulation, with the following color code: the particles are in green when they are in the upward vertical orientation, and in gray when in the downward vertical orientation. This preliminary result suggest that there is no camembert that can arise from a mixture of upward and downward particles, and that the particles of similar vertical orientations will aggregate together in the camembert geometry. However, in some cases aggregates of two orientations are merged (the aggregate on the right in the Figure).

We can experimentally fabricate a particle that has the directional interactions that we considered in this Chapter. The work is in progress, and will reveal whether the predicted camembert geometry will arise in from the self-assembly of such particles.

Discussion

In Chapter 1, we showed that the existing mechanisms of self-limiting assembly relied on the independent design of each particle, on the self-closing of the assembly, or on their deformability. Here, we showed that the competition between two incompatible favored interactions, could lead to the size control of open-ended assembly of identical rigid particles. We demonstrated the idea for two specific geometries in two-dimension : a two-dimensional aggregate of controlled radius and a fiber of controlled width, where a disclination line relied on an interaction more favored than the bulk interaction, but it was not geometrically accessible for all the bonds between particles.

This mechanism relies on the directionality of the interactions between the particles. Therefore, we expect that it could be implemented for other particle geometries, as long as two favored directional interactions cannot be realized at the same time in a dense aggregate. For instance, we expect this mechanism to be applicable to square particles, for which the equivalent of the camembert aggregate would be a square aggregate with four disclination lines. The particle also does not need to be hexagonal or square, as long as the interactions are directional. We also expect this principle to work in three dimensions, with defect planes instead of defect lines. However, we did not test it in numerical simulation.

The experimental implementation of such principle requires designing particles with directional independent interactions. DNA-origamis appears well adapted to this requirement, because the DNA-strands can be chosen to bind only to another specific DNA strand on the particle. This may enable to achieve the self-assembly of individual particles in aggregate of several tens of particles. However, we showed that large sizes of aggregate

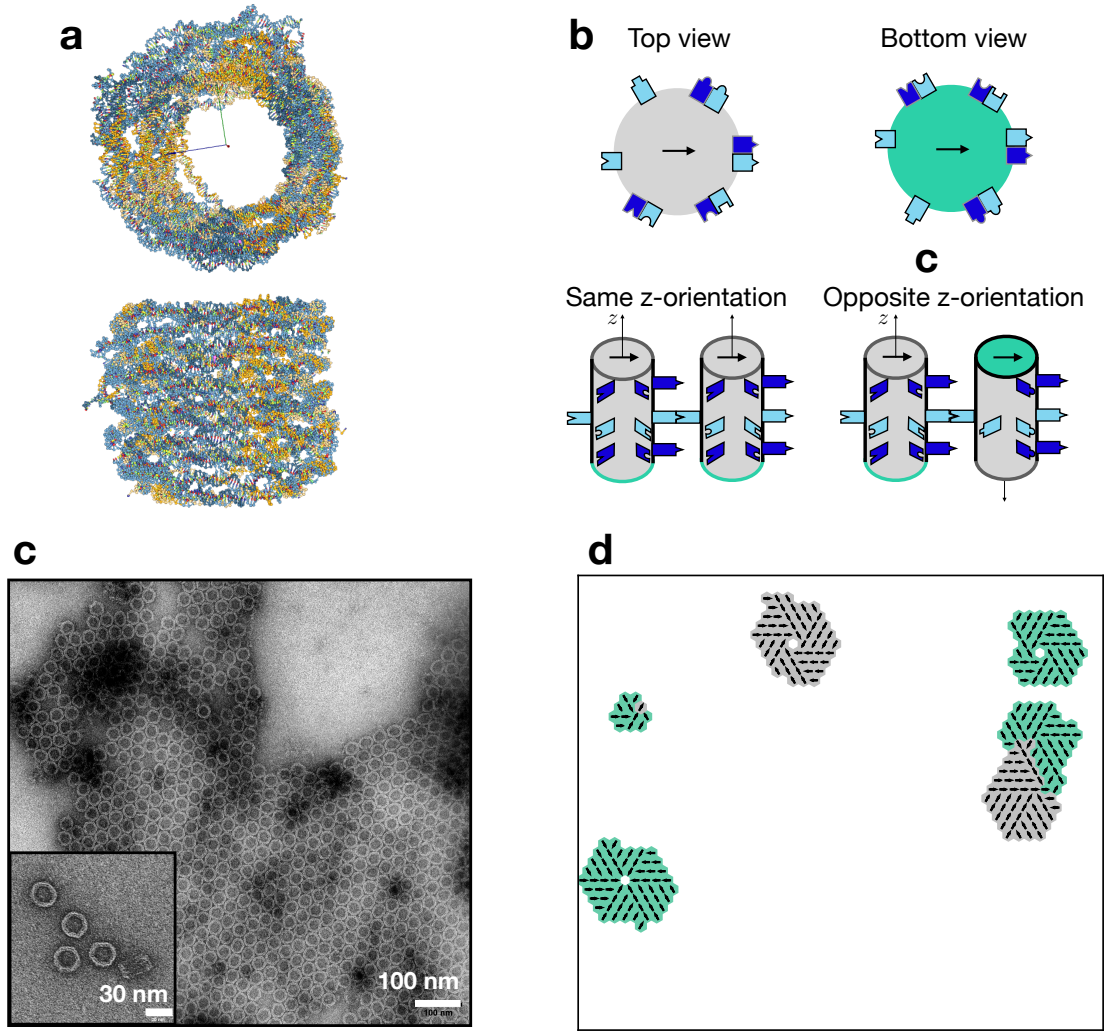

Figure 5.16: We implement the camembert design in DNA origami experiments. a) We use DNA-origami with a barrel shape. The surface can be patterned with strands that interact. b) The three-dimensional cylindrical particle can be in two vertical orientations. If the interactions are symmetric in z, it means that there are two ways to realize each crystalline contacts: particle can have same or opposite vertical orientation. c) Preliminary images of experimental self-assembly of DNA barrels imaged with electron microscopy. Courtesy of Christoph Karfusehr. d) Result of simulated annealing in two-dimension space where the three-dimensional particle can flip vertical orientation, while keeping the annealing parameters presented above ($J_c = 0.5kT$, $J_l = -8kT$, $\sigma = 6kT$, $e_\infty = 15kT$).

required a fine-tuning of the strength of the interaction: the range of parameters where an aggregate of radius r is most stable is more and more narrow as r increases. It was already difficult to achieve aggregate sizes of finite radius larger than 5 in the numerical simulation, and we expect that the same kind of difficulties could arise in experimental implementation.

If the camembert geometry mostly serves as a proof of principles that directional interactions can limit the size of the aggregate formed upon self-assembly of the individual particles, the fiber geometry could be used for biomedical implication. Indeed, in [38], the author showed how to build bioreactors from tubes of controlled radius. Those tubes result from the folding of two-dimensional fibers of controlled width. Here, we introduced a novel mechanism to self-assemble fibers of controlled width.

In this study, we applied the concepts of frustration to short-range directional interactions. The frustration arising from the fact that both the line and the crystal are not compatible introduces non-local effects and enables to control the size of the aggregate.

6 - Systematic identification of protein aggregate dimensionality in crystallographic data: methods and preliminary results

This thesis and previous work suggested that self-assembly of objects with complex shapes or interaction often leads to aggregates of *self-limited size* (like micelles) or reduced *dimensionality* (like fibers that are one-dimensional objects, and sheets that are two-dimensional objects in three dimensions). Proteins are canonical examples of particles with complex interactions. Specifically, when in non-physiological condition, or after mutation, they can exhibit aggregate of diverse shapes. We could test the hypothesis that dimensional and size reductions are generic phenomenon experimentally if

- we study a large amount of protein samples, in diverse physicochemical conditions,
- and if we are able to detect the presence of fibers or small size aggregate within such samples

Light interactions with matter are preferential tools to study matter at small scales. Particularly, **small angle X-ray scattering** is a technique used to investigate shape and size of macromolecules, typically on the length scale of interest for us (between 10 and 1000 Å), and is a solution to the second requirement. On the other hand, **crystallography** uses the periodicity of the atoms within a material to measure diffraction peaks, and identify the crystal structures. While we are not interested in measuring crystal diffraction, the approach of protein crystallographers for sample preparation is interesting: obtaining a protein crystal is hard, and proteins are aggregated in a large variety of physicochemical conditions, and analyzed with X-rays, before a crystal is finally found. There is therefore a large amount of scattering signals of protein aggregates available, including proteins that did not crystallize, while the approach of SAXS experiments does not provide a systematic exploration of the physicochemical conditions. In this chapter, we will ask whether it is possible to measure the dimensionality of protein aggregates, and detect dimensionality reductions, in the data collected by crystallographers, by using the analysis methods of SAXS. In particular, there is a small range of length scale for which crystallographic signals are collected, and for which information about the aggregate dimensionality can be measured. In Sec. 6.1, we describe the principles of SAXS and crystallography, and show that it is in principle possible to detect dimensionality reduction in crystallographic experiments. We will mostly attempt to detect fibers, that are easily formed from protein self-assembly, as was emphasized in Chapter 1. In Sec. 6.2, we analyze scattering signal from protein aggregate constructed numerically and show that we can systematically detect the dimensionality of the aggregate in those numerical data. This was the project of M. Billoir master internship, which we co-advised. Finally, in Sec. 6.3, we analyze experimental crystallographic data collected with the help of our collaborators W. Shepard (Synchrotron Soleil, Saclay), and M. Spano (Institute of Structural Biology, Grenoble). We took part in the experiments by collecting the scattering signals in the synchrotron beamline. This was the project of M. Garic master internship, which we co-advised. The main challenge with experimental data was to isolate the scattering signal of the protein from its background. For this reason, we suggest that the aggregate dimensionality can be identified with statistical methods like machine learning, provided that enough data of protein aggregates of known dimensionality are collected.

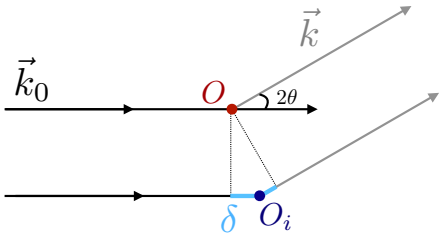

Figure 6.1: The scattering angle of light is related to the organization of the atoms in a sample. The incident light of wave vector $\mathbf{k}_0$ is scattered by atoms at positions O and O_i. The light intensity measured at the scattering angle $\mathbf{k}$ results from the interference between the two light rays. The phase difference is proportional to the path difference δ. Figure adapted from [127]

6.1 Identifying dimensionality reduction in scattering signals of crystallographic experiments

Both small angle X-ray scattering (SAXS) and crystallography rely on similar principles, which we describe in Sec. 6.1.1: information about the spatial distribution of the atoms is measured from the scattering of X-ray beams in the sample. We explain in Sec. 6.1.2, that SAXS is a particularly well suited technique to study the size and the dimensionality of protein aggregates. On the other hand, the sample preparation techniques in crystallography enable the collection of a large amount of scattering signals, which we describe in Sec. 6.1.3. The use of SAXS methods on crystallographic data could then be used to systematically identify dimensionality reduction and size reduction in protein aggregates.

6.1.1 X-ray scattering gives information on the material

X-ray scattering is used to determine information about materials at very small scales [127]. Here, we show how the intensity of the scattering signal is related to the composition of the sample it went through.

In most material, an X-ray beam is scattered by atoms and electrons in the sample, but it conserves its frequency. For this reason, all the waves measured in direction $\mathbf{k}$ that have been scattered for their initial direction $\mathbf{k}_0$ will interfere if they are coherent, or be added otherwise (see Figure 6.1). The phase difference ϕ between the signals of two scatterers is related to the distance between them OO_i through the path difference δ: $\phi = 2\pi\delta/\lambda$ and $\delta = -\mathbf{OO_i} \cdot (\mathbf{k} - \mathbf{k}_0)$. Thus, the measured signal will carry information about the material on a length scale of the order of its wavelength, *i.e.*, between 1 pm and 1 nm.

The initial amplitude A_0 of the light scattered by an atom at the position O_i in direction $\mathbf{k}$ is decreased by a factor f_i, called the scattering factor, such that its amplitude is $A_0 f_i$. It was shown that amplitude of the light in the direction $\mathbf{k}$ resulting of the contribution of atoms at positions O_j [128]

$$A(\mathbf{k}, t) = A_0 \sum_j f_j e^{i\phi_j} \tag{6.1}$$

We define the scattering vector as $\mathbf{q} = 2\pi(\mathbf{k} - \mathbf{k}_0)/\lambda$. The phase difference due to the interaction between the scatterer at positions O_i and a reference O is $\phi_i = -\mathbf{r}_i.\mathbf{q}$. If all scatterers are identical, so are the scattering factors ($f_i = f$). Then, the light intensity at angle q is

$$I(q) = AA^* = A_0^2 f^2 \sum_{i,j} e^{i(\phi_j - \phi_i)} = A_0^2 f^2 \sum_{i,j} e^{i(\mathbf{r}_i - \mathbf{r}_j)\cdot\mathbf{q}} \tag{6.2}$$

From this equation, both diffraction and scattering experiments can be understood. In

diffraction experiments, the periodic distribution of the scatterer leads to constructive and destructive interference, that will be detected as Bragg peaks [129]. This is the principle used in crystallography experiments. When the scatterers are not periodically arranged, waves of identical phases are regrouped, and eq. 6.2 can be written as a function of the scattering length of the atoms per unit volume $\rho(\mathbf{r})$. More precisely, it can be written as a function of the excess scattering length density compared to the solvent $\Delta\rho(\mathbf{r}) = \rho(\mathbf{r}) - \rho_s$ [128, 130].

$$I(q) = \int_V \Delta\rho(\mathbf{r}) \mathrm{e}^{i\mathbf{q}\cdot\mathbf{r}} d\mathbf{r} \tag{6.3}$$

Whether the scatterers are arranged in a periodic way or not, the scattered intensity for a given scattering vector q carries information on the spatial organization of the atoms and electrons in a sample. For this reason, we want to use measures performed during crystallographic experiments, for which no crystal was observed, and analyze it with tools of scattering techniques.

6.1.2 SAXS methods are adapted to describe the shape of the aggregates

The intensity of the scattered light as function of the scattering angle gives information about the spatial distribution of the molecules in the sample. Here, we show how scattering signals are analyzed in different regime of q. For low q, the Guinier's law enables to measure the gyration radius of a particle. For large q, the Porod's law gives a universal scaling of the intensity, provided that the surface of the aggregate is flat. In intermediate regimes, the dimensionality of the aggregate can be measured. This is the regime we are interested in. In the following, we only give a summary of the scaling of the scattered intensity in different ranges of the scattering vector. More details can be found in dedicated reviews [130, 131]. We also treat aggregates of proteins as the individual particle.

At low values of q, *i.e.* when measuring correlations on length scales that are larger than the typical size of the aggregate, it is possible to neglect the geometric specificities of the aggregate. The scattered intensity then only depends on the gyration radius. This is the Guinier approximation [128].

$$I(q) \approx I(0) \exp\left(-\frac{1}{3} R_g^2 q^2\right) \tag{6.4}$$

for values of q below $1/R_g$. Therefore, $\log I(q)$ can be fitted in the low q regime as a function of q^2 to obtain both the intensity at the origin and the gyration radius of the aggregates. This is the region where the values of q in the phase space corresponds to scales larger than the particle size in the real space, the left region in Figure 6.2.

At large values of q, it was shown by Porod [132] that

$$I(q) \sim q^{-4} \tag{6.5}$$

This is the region where the values of q in the phase space corresponds to scales much smaller than the individual particle size in the real space, and information about the surface of the particles are measured. A protein however has a rough surface, and it is not clear that Porod's law is valid. This corresponds to the right region in Figure 6.2.

We show that in the intermediate regime, the scattering intensity is related to distribution of distances between the electrons within one aggregate. Eq. 6.3 can be simplified knowing that $\langle \mathrm{e}^{i\mathbf{q}\cdot\mathbf{r}} \rangle_\Omega = \sin(qr)/(qr)$. The average over Ω stands for an average over all

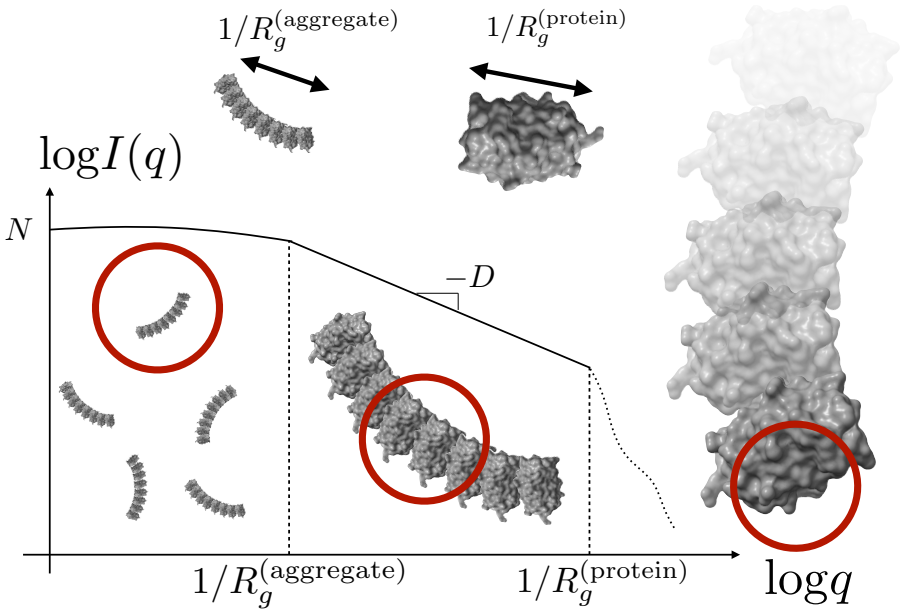

Figure 6.2: Depending on the range of scattering vector (q), information on different length-scales of the aggregates are measured. This is a highly idealized schematic. At low q (left), variations at the scale of the whole aggregate are measured (red circle is larger than the aggregate). $I(q)$ follows Guinier's law (eq. 6.4). At intermediate q, the dimensionality D of the aggregate can be measured (the red circle is larger than a single protein but smaller than the typical size of the aggregate, $I(q)$ follows eq. 6.8. At large q, only variations at the surface of the aggregate can be observed.

possible orientations of a particle.

$$I(q) = 4\pi \int r^2 \gamma(r) \frac{\sin(qr)}{qr} \tag{6.6}$$

$$\text{with } \gamma(r) = \left\langle \Delta\rho(\mathbf{r}_0)\Delta\rho(\mathbf{r}_0 + \mathbf{r}) d\mathbf{u} \right\rangle_\Omega \tag{6.7}$$

$\gamma(r)$ describes how the excess scattering densities are correlated in a given particle. In practice, $p(r) = r^2\gamma(r)$, which corresponds to the distance distribution in the particle, is used to deduce information about geometrical properties of the scattering particle. At intermediate values of q, *i.e.* in between the Guinier's and the Porod's law, this distribution scales like $p(r) \sim r^D$, where D is the dimensionality of the aggregate. Then the integration of eq. 6.6 is simply:

$$I(q) \sim q^{-D} \tag{6.8}$$

This regime (in the middle in Figure 6.2), is the one of interest for us. If we are able to measure scattering signal in that region, we can compute the aggregate dimensionality and identify which protein self-assemble into aggregates of reduced dimensionality (sheets for $D = 2$ or fibers for $D = 1$).

In those three regimes, we presented the scattered intensity for one particle. If the solution of aggregate is diluted enough, we can assume that the aggregates do not interact. Then the total scattered intensity is $N_{\text{aggregates}} \times I(q)$, with $N_{\text{aggregates}}$ the number of aggregates, and the scaling of eq. 6.4, 6.5 and 6.8 are still valid.

In the following study, we consider proteins that have gyration radius between 10Å for insulin or 20Å for actin, which means that the upper bound to observe the scaling of eq. 6.8 is around 0.1Å^{-1}.

It is possible to determine the exact dependence of $I(q)$ with the dimensionality D, for which an approximate scaling is eq. 6.8. In particular, we can also consider an aggregate

of proteins as the scattering signal of individual proteins in interaction. We then define, the *form factor* $P(q)$ as the scattering signal of individual proteins. The *structure factor* $S(q)$ describes the interactions of the N_{protein} individual proteins in one aggregate, such that the total scattering intensity is $I(q) = N_{\text{protein}} P(q) S(q)$. This is the notations we will use in the rest of the chapter.

This approach was extensively developed by supposing that a protein aggregates can be modeled as a fractal aggregates of nanoparticles [133]. In particular, it was derived that

$$S(q) = 1 + (N_{\text{protein}} - 1) \left(1 + \frac{2 \left(R_g^{(\text{protein})} q \right)^2}{3D} \right)^{-D/2} \tag{6.9}$$

At small values of q, $S(q) = N_{\text{protein}}$: at very large scales, there are no interactions between the proteins and the scattering signal is the sum of the scattering signal of each protein. In the large q limit, $S(q) = 1$, the scale is too small to distinguish between the different proteins, and the scattering intensity is that of individual proteins.

Being able to fit $S(q)$ for any protein aggregate with such formula would then directly lead to the information of interest: the dimensionality of the proteins and its size in terms of number of particle. We could then detect the presence of self-assembled fibers of micelle of proteins.

In the following, we will always call *dimensionality* the dimension of an aggregate (fiber has dimension 1, sheet has dimension 2, and bulk has dimension 3), and *size* the length dimensions of the aggregates. The term *dimension* will be kept for the dimension of an array of numbers.

6.1.3 Crystallographic data are widely available

We explained how the scattering signal is used to determine aggregates size and dimensionality, specifically in SAXS. We now explain why crystallographic experiments are more interesting for the purpose of this project in terms of sample preparations, but less suitable in terms of the range of scattering vectors for which a scattering intensity is measured.

The 3D structure of a protein can be determined by crystallography. If the protein has crystallized, all its atoms are repeated in a periodic way and the light they scatter will interfere. The spatial details of the protein can then be determined, as was explained in Sec. 6.1.1. When an electron density map of a protein, and its 3D structure, is determined experimentally, it is added to the Protein Data Bank (PDB) [134]. 85% of the protein structures deposited in the PDB were determined by X-ray crystallography [135]. The bottleneck of this method is the crystallization of the protein. The last decades witnessed the development of *in situ* crystal harvesting: in each well of a crystallization plate (see Figure 6.3), a protein droplet is mixed with a buffer solution containing different type of precipitants. For each plate, a different buffer solution is chosen. From one well to the other, the concentration of the buffer solution can also vary. X-ray is shone on all of those samples, directly on the plate where they were prepared. The scattering signals exhibiting Bragg peaks contain crystals, as explained in Sec. 6.1.1, and they are selected for further study [136, 137].

These experiments are typically done in synchrotron facilities. In the synchrotrons of the European Union [138], data of the experiments are made publicly available, and stored, three years after it was collected. For this reason, there is a large amount of publicly available data of X-ray experiments. Those data include the scattering signal of protein solutions in variable conditions, because crystallization does not always occur. The scattering signals of solutions where the proteins did not crystallize could then be analyzed to determine the shape and size of the protein aggregates that might have self-assembled

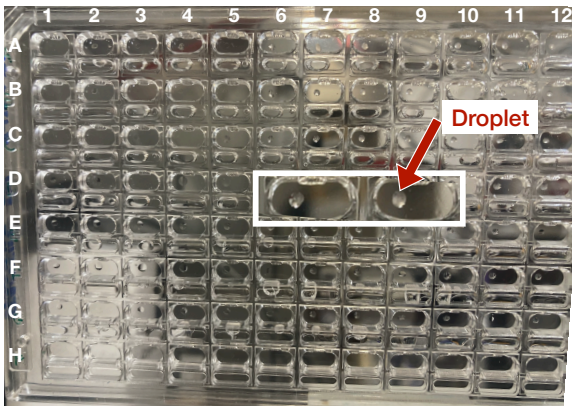

Figure 6.3: A crystallization plates enables to test a large variety of experimental conditions. In each well, a droplet of protein is deposited. Each well has a size of around 0.5 cm.

in each sample. This would provide a systematic identification of protein aggregates with reduced size or dimensionality, without having to perform those experiments for scratch, which would be extremely costly and long.

In crystallography experiments, the range of scattering vector q is typically between 5×10^{-2} and 5×10^{1} Å^{-1}, corresponding to sub-nanometric length scales. In the previous section, we determined that we should focus on scattering vectors below 0.1Å^{-1}. There is therefore a small range of q where (i) a large amount of data is available, and (ii) the information about the aggregate size and dimensionality can be measured.

6.2 Dimensionality identification in scattering of numerical aggregates

We estimated that the range of scattering vector where we could measure information of aggregate shape is below 10^{-1}Å^{-1}, and that crystallographic data measures intensity for scattering vectors above $5 \times 10^{-2}\text{Å}^{-1}$. Also, in the intermediate range of scattering vector, the scaling of the intensity is directly related to the aggregate dimension. Here, we go beyond those initial estimates and verify that an analysis of the scattering signal of very diverse protein aggregate in this range is sufficient to deduce its dimensionality on idealized data: we generate scattering signals of protein aggregates built numerically, and provide methods to systematically identify their dimensionality. This study is therefore a proof of principle of the idea of the broad project described above. Since analyzing experimental data can be challenging for many other reasons, it is useful to develop an analysis method on idealized data, which can then be adapted to account for the additional difficulties of real data. We create a database of scattering signals of numerical protein aggregates of known dimensionality in Sec. 6.2.1. We then show that the dimensionality of these idealized aggregates can be automatically identified from the scattering signals with machine learning methods, in the range of scattering vectors of crystallographic experiments (Sec. 6.2.2). All this work was done together with M. Billoir during her master internship [139].

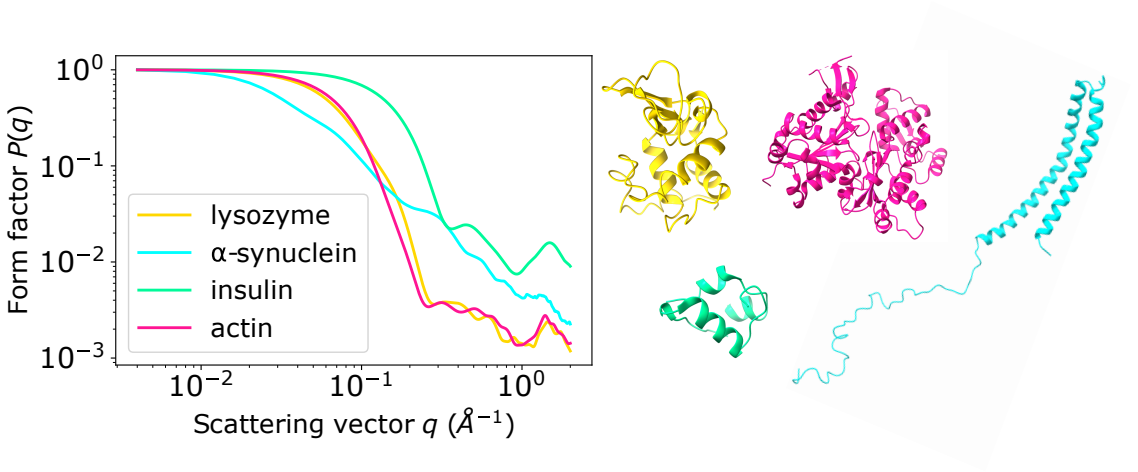

Figure 6.4: The software Crysol models the scattering signals of density maps of proteins from the PDB. The proteins we show here have the following reference in the PDB: lysozyme(6lyz), α-synuclein(1xq8), insulin (2C8R), actin (2hf4). The proteins are visualized with ChimeraX.

6.2.1 Build a database of protein aggregates numerically

Here, we will build a database of around 250 of scattering signals corresponding to protein aggregates with variable characteristics. Our objective is to create scattering signals of aggregates with predetermined dimensionality (0,1,2 or 3) that are as different as possible in other aspect: the local organization of the proteins in the aggregate, the size, the protein it is built of, and the geometry. We will present the methods and software we used in Sec. 6.2.1.1. In Sec. 6.2.1.2, we build numerical protein aggregates from the density maps in the protein data bank. In Sec. 6.2.1.3, we compute the corresponding scattering signals and show that while the fibers (dimensionality 1) seem to be easily identifiable, there is no trivial way to classify all the aggregates by their dimensionality.

6.2.1.1 Approach

We can use existing software used to analyze scattering data for our purpose of dimensionality detection.

We use the software Crysol [140] to compute the scattering signal of a protein aggregate in solution, from its electron density map. Examples of such signals are shown in Figure 6.4. For each protein, we extract the protein density map from the PDB [134]. The program then averages the scattered intensity over all orientations of the protein

We are interested in the scattering signal of aggregates of proteins, rather than single proteins. To get the density map of gas, fibers, sheets or crystals of proteins, we *concatenate* the density maps of the individual protein, *i.e.* we place each density map next to the other. This is done with the software ChimeraX [141], and images of examples of such aggregates are shown in Figure 6.5a. Each artificial aggregate is built from one of the four proteins of Figure 6.4: insulin, actin, lysozyme and α-synuclein. We choose this set of proteins because they have variable sizes and 3D structures.

To build each aggregate, we will thus follow these steps:

- collect the electron density map of the protein subunit in the PDB
- concatenate several density maps and observe the resulting protein aggregate with ChimeraX
- compute the scattering signal of the designed aggregate with Crysol

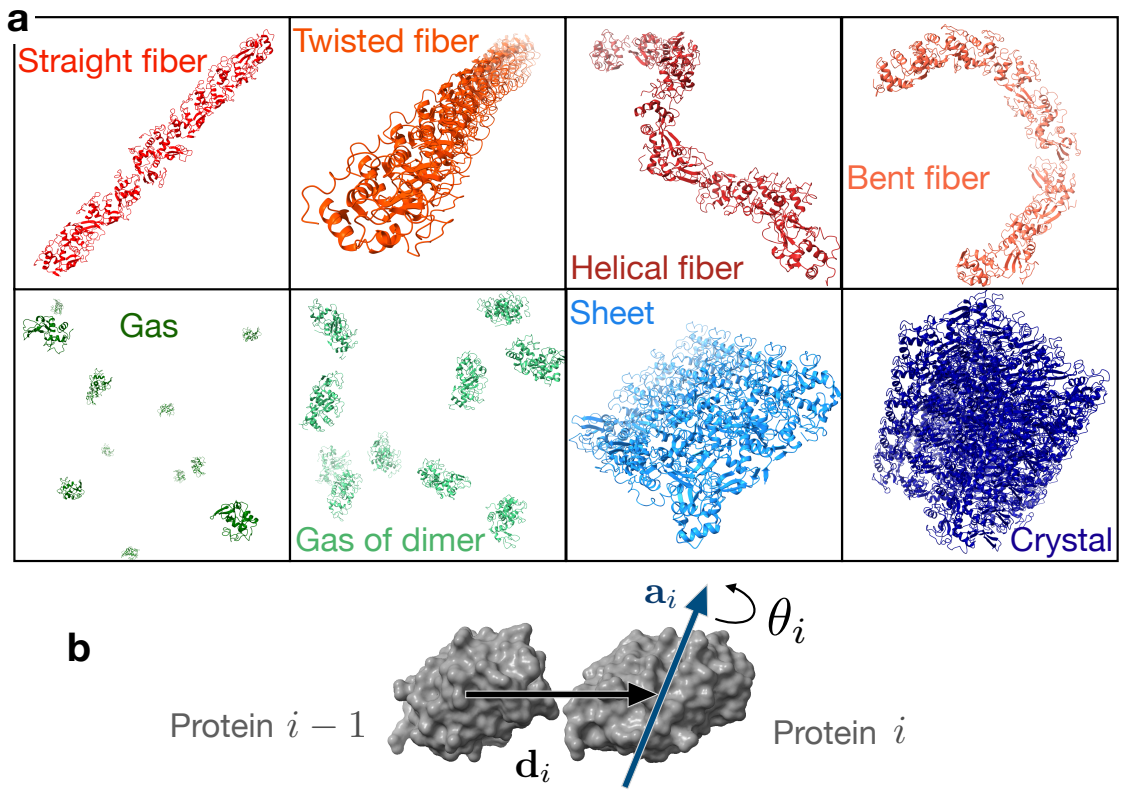

Figure 6.5: We build artificial aggregates of proteins of variable dimensionality by stacking individual proteins in the software ChimeraX. a) Artificial aggregates of 0, 1, 2 and 3D. Aggregates are built by stacking lysozyme proteins. Implementation and visualization are done with ChimeraX [141]. b) A protein in the aggregate is translated and rotated from the position of its neighbor.

6.2.1.2 Aggregate characteristics

We build aggregates of dimensionality 0, 1, 2 and 3, as shown in Figure 6.5a. For a given dimensionality, we build aggregates that are as different as possible from the others, by varying other characteristics of the aggregate, which we explain here.

Each monomer i is added at the position $\mathbf{x}_i$. In practice, the position is chosen relatively to the position of the previous monomer. A translation vector $\mathbf{d}_i$, an axis of rotation $\mathbf{a}_i$, and an angle θ_i are chosen such that $\mathbf{x}_i = \mathbf{x}_{i-1} + \mathbf{d}_i$, the first monomer being at position $\mathbf{0}$. The monomer can then be rotated of θ_i around the axis $\mathbf{a}_i$. This is illustrated in Figure 6.5b. We define four *category of aggregates* which share the same dimensionality:

- The gas is such that $\mathbf{d}_i$, θ_i and $\mathbf{a}_i$ are chosen randomly for each monomer. For the gas of dimer, the translations, and rotation are applied to previously built dimers. A gas is zero-dimensional (green in the following)
- The fibers are such that all the monomers are aligned: $\mathbf{d}_i = \mathbf{d}_1$, with $\mathbf{d_1}$ the translation vector. For the straight fiber, the orientation does not change (θ_i is always zero). For the other fibers, we introduce a constant rotation $\theta_i = i\theta_0$ between two monomers. If the axis of rotation $\mathbf{a}_i$ is collinear to the direction of the translation $\mathbf{d}_1$, the fiber is twisted. If the axis of rotation $\mathbf{a}_i$ is orthogonal to the direction of the translation $\mathbf{d}_1$, the fiber is bent. In the most generic case, the fiber is helical (see images in Figure 6.5a). Fibers are one-dimensional (red in the following).
- The sheets correspond to monomers translated from $\mathbf{d}_1$ or $\mathbf{d}_2$ from their neighbors, with $\mathbf{d}_1$ and $\mathbf{d}_2$ being orthogonal. Sheets are two-dimensional (light blue in the following)
- Crystals corresponds to monomers translated from $\mathbf{d}_1$, $\mathbf{d}_2$, or $\mathbf{d}_3$ from their neighbors, with $\mathbf{d}_1$, $\mathbf{d}_2$ and $\mathbf{d}_3$ being orthogonal. Crystals are three-dimensional (dark blue in the following)

The norm of the translation vector $\mathbf{d}$ is chosen such that the proteins are not intertwined ($\|\mathbf{d}\|$ is between 1 and 1.2 times the size of the protein). We choose not to optimize the relative positions of two monomers with the command implemented in ChimeraX (rigid body local optimization of two density maps). This would make the contact between two proteins more physical. However, it complexifies the building procedure, while introducing only minor changes in the scattering signal, for large values of q (only above $q = 30\text{Å}^{-1}$), which corresponds to scale that are not relevant to determine the aggregate shape.

To broaden the database of aggregates while keeping its dimensionality well-defined, we also build *random* versions of the fiber, sheets, and crystal: the positions of the subunits are implemented as above, but their orientations are chosen randomly. Finally, we vary the number of subunits in the aggregate. In practice, the side of each aggregate is between 3 and 40 proteins. We summarize the number of aggregates built from each category in table 6.1. For each category, we indicate two quantities, the number of organized aggregates and the number of random aggregates. For each protein, we build aggregates of the different categories. There is 67 α-synuclein aggregates, 42 actin aggregates, 103 lysozyme aggregates and 70 insulin aggregates.

We achieved to create density maps of a large variety of protein aggregates, that have a well-defined dimensionality, but vary in the other aspects, like the protein its built from, the size of the aggregate, the geometry of the aggregate and the local organizations of the particles.

Categories of aggregates							
Crystal	Sheet	Fiber				Gas	
		Straight	Twisted	Helical	Bent	Dimer	Monomer
3+19	18+10	26+10	0+9	37+27	43+25	0+10	0+45
22	28	177				55	

Table 6.1: For each category of aggregate, we collect data of organized and random aggregates. For instance, there is 3 organized and 19 random crystals in the dataset. The last line indicates the total in each category.

6.2.1.3 Computation of the scattering signal

From the density maps of the proteins of Sec. 6.2.1.2, we build a database of scattering signals for which the dimensionality of the aggregates is known. For each of the density maps created in 6.2.1.2, we compute the corresponding scattering signal with the software Crysol, introduced in 6.2.1.1. Here, we explain our choices of parameters for the use of Crysol, how we extract the relevant information from the computed scattering signal, and show that the dimensionality of the aggregate cannot be trivially identified from them.

We compute the scattering signal for scattering vectors in the interval $[0.03, 0.35]\text{Å}^{-1}$. The upper limit of q is chosen such that we do not measure signals for length scales below $2nm$, which is smaller than the gyration radius of the protein we consider. The upper bound is fixed by the typical range of scattering vectors in crystallographic experiments. We are taking some margin compared to the range $[0.05, 0.1]\text{Å}^{-1}$ identified in the introduction. Scattering intensities for 2500 values of q within this range are measured, which provides measurements as reliable as those with larger number of points, and reasonable computational times.

As explained in 6.1.2, the scattering signal of the aggregate $I(q)$ is then the product of the form factor of the individual proteins $P(q)$, the structure factor that describes the interactions between subunits $S(q)$, and of the number of proteins N_p [133].

$$S(q) = \frac{I(q)}{N_p P(q)} \tag{6.10}$$

The form factor was also computed with ChimeraX and shown in Figure 6.4. We show the computed signals of aggregates of chosen dimensionality, where the density map was built as explained in 6.2.1.2 in Figure 6.6. The orientations of the particles in the aggregate are regular (plots a-d) or random (plots e-h). The aggregates are of different sizes, and built for different proteins in different. We also show the gyration radius of the individual protein computed in Crysol. As explained in 6.1.2, we expect relevant information about the dimensionality of the aggregate to be below this limit.

We first notice that the expected limits at low and large q for the structure factor is not recovered from our measures with Crysol. Indeed, $S(q) \neq 1$ when $q \to 1$ [133]. This is the expected behavior, because at large q, the scattering signal does only depend on the number of constituents, and we expect $I(q) \sim N_p(q)P(q)$. Even if the large q limit does not converge to one, it does converge towards a finite positive value. A rescaling would therefore be necessary, but we could not determine it within the limited range of q (typically we do not measure scattering signal for q lower than 10^{-2}, and cannot recover the $I(q \to 0)$ limit). Thus, we could not find the correspondence between the normalization of structure factors computed from Crysol [140], and that expected from protein aggregates [133].

From this preliminary visualization, however, it seems that almost all fibers share the linear regime of relatively small slope (compared to other types of aggregates) as a common characteristic, despite their difference in number of particles and local organizations. This

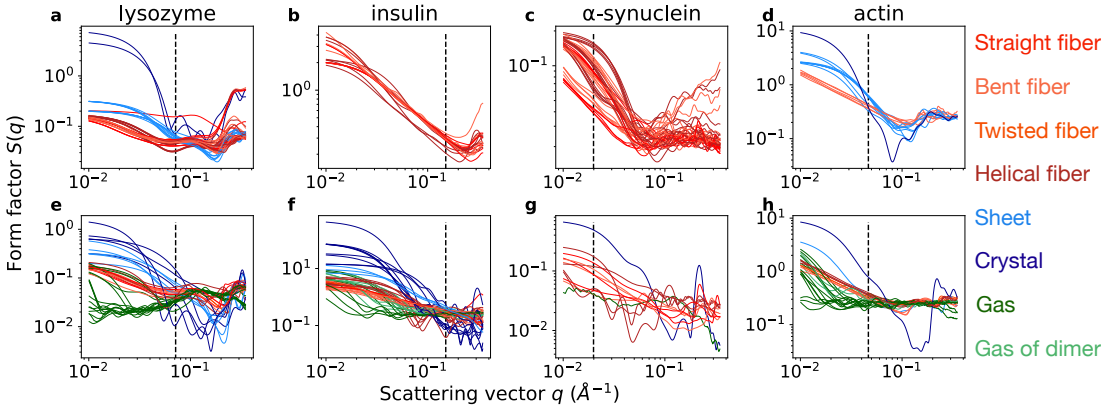

Figure 6.6: Aggregates of identical dimensionality built from the same protein have similar structure factors. The color represents the type of aggregate. (a-d) Subunits have regular orientations in the aggregate. (e-g) Subunits have random orientations in the aggregate. Aggregates are build from lysozyme (a, e), insulin (b, f), α-synuclein (c, g), and actin (d, h). The black dot line corresponds to $q = 1/R_g$, R_g is the gyration radius of one protein subunit.

suggests that the scaling of the scattering signal with the dimensionality of the aggregates is a feature common to fibers with very different properties. The identification of the other categories is not trivial.

6.2.2 Analysis of the numerical scattering signal

We computed the scattering signals of aggregates of well-defined dimensionality (gas in 0D, fibers in 1D, sheets in 2D, and crystals or bulks in 3D). The aggregates of identical dimensionality are very different in size, local organization, or built from different proteins. We now investigate whether the identification of a linear regime that was observed qualitatively in the numerical scattering signals (Sec. 6.2.1.3), and the measure of its slope, is enough to classify the aggregates according to their dimensionality. Because this method does not work systematically, and does not allow identifying the other types of aggregates (gas, sheets, and crystal) we propose alternative classification method based on the training of a neural network classifier and show that it enables to systematically identify all types of aggregate (Sec. 6.2.2.2).

6.2.2.1 Fit segment to measure dimensionality

We propose a method to systematically detect the scaling of the scattering signal with the dimensionality ($S(q) \sim q^{-D}$) introduced in Sec. 6.1.2. We show that this enables to detect fiber aggregates in most cases in our dataset, but not the other types of aggregates.

To analyze the scattering signals, the most intuitive approach would be to fit the signal with an expected structure factor like that of eq. 6.9, and deduce the dimension D, the number N of subunits and a gyration radius R_g. Such approach were adopted in [133]. However, because we could not identify the correct rescaling of the measured scattering signal, as explained in Sec. 6.2.1.3, we could not fit systematically fit them with eq. 6.9.

Here, we adopt a simpler approach, that consists in identifying a linear regime in the scattering signal, and deduce the dimension of the aggregates. We expect the scaling $S(q) \sim q^{-D}$, to be observed on a range of scattering vectors q that depend on the individual proteins gyration radius, and on the aggregates gyration radii. Thus, we fit all the portion of each signals in a log-log scale, and identify which will give the best linear fit. The slope of the linear regime is then an approximation of $-D$. We fit the signal to a linear curve in all portions of $[\log q_0, \log q_0 + \Delta(\log q))]$ with $\Delta(\log q) \approx 0.6$, *i.e.*, on a bit more than half a

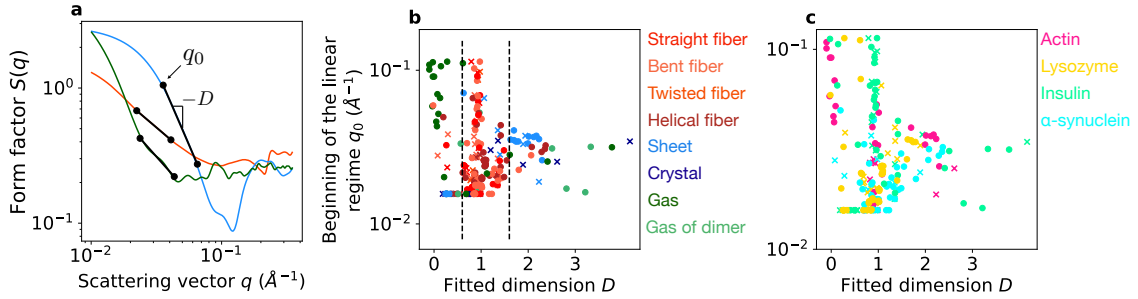

Figure 6.7: We detect a linear regime of slope -1 for the fibrillar aggregates, independently of the protein it is built from. (a) For each form factor, we choose q_0 such that the linear fit is the best, and deduce the dimensionality D. (b-c), q_0 as a function of D for the organized (round dot) and random (cross dot) aggregates. We measure a dimensionality of 1 for the fibers (red in b), regardless of the protein (other colors in Figure c). We show only the data points for which the best linear fit has $\chi^2 < 1$.

decade. We choose the value of q_0 such that the residual $\log I(q) + D \log q$ is minimal. We show the example of such a fit in Figure 6.7a on three scattering signals. The determination of q_0 is therefore simplistic, and rely on an arbitrary criterion, which could be refined in further analysis. We expect the dimensionality D measured with this method to be the dimensionality of the aggregates, *i.e.* 1 for the fibers, 2 for the sheets, etc.

The measured dimensionality D is plotted as a function of q_0 in Figure 6.7b and c. We confirm that we can measure the scaling of the signal with the dimensionality for most of the fibers within the chosen range of scattering vectors q: the red points are all at the abscissa $D = 1$. We count that 82% of the scattering signals of fibrillar aggregate 19% of the scattering signal of non-fibrillar aggregates are fitted with an exponent between -0.6 and -1.6 (dashed line in the Figure). If we set being in this range as a criterion to recognize fibers, this simplistic method enables to identify a majority of fibers. The fit is less successful for the other types of aggregates. But we still measure dimensionality 2 for some sheets, and dimensionality 3 for some crystals. Moreover, it is important to notice that this criterion to detect fiber is independent of the individual proteins of the aggregates: signals of different proteins (points of different colors in Figure 6.6c) are detected with the same criterion in Figure 6.6b). It is also independent of the local organization of the proteins within the aggregate (round dots correspond to regular orientations, and cross dots to random orientations of the proteins within the aggregate).

The approach of identifying the slope of the best linear regime, while not entirely satisfactory, indicates that most of the fibers could be detected from the scattering signal, independently of the aggregate size, protein building block, and local organization. With this method, the gas aggregates are not well characterized. This is because Crysol is not well adapted to deal with several aggregates within the same density map. Moreover, the approach proposed in [133] was to average the scattering signal over aggregates of different sizes that share the same characteristics, to better reproduce experimental scattering signals of a solution of proteins. This might also explain our difficulties to fit the data with eq. 6.9, additionally to the normalization issues detailed in Sec. 6.2.1.3. It is also not clear that this method could be reproduced on experimental data: without prior knowledge on the size of the protein, the estimate of D strongly depends on q_0 and Δq, which are determined arbitrarily. However, if we know the range of q where the linear regime is expected, and if this range is within the range measured in crystallographic experiments, this results suggest that the dimensionality could be deduced from the experimental signals of protein aggregates with very diverse characteristics.

6.2.2.2 Statistical methods to learn dimensionality

Because the fit in the linear regime did not provide a systematic classification of the aggregate according to their dimensionality, we now take advantage of the large amount of data to train a neural network to classify the aggregates. Indeed, in Sec. 6.2.1, we computed a dataset $\{X_k^{(i)}\}$ where $X_k^{(i)}$ refers to the scattered intensity measured at the k^{th} value of q and for the i^{th} aggregate. Each aggregate i is also associated with a category (fiber, gas, crystal, or sponge). This is a well-defined problem to use a *feed-forward neural network classifier*, for which we introduced the basic principles in Sec. 3.2.3 of Chapter 3: the network performs a series of non-linear transformation on each input data X_k which has a large number of *features*: it is an array of dimension N_{features}. This operation results into an output of dimension 4, which gives the probability for this data to correspond to each category of aggregate [105]. The network is optimized such that the predicted categories are similar to the true categories, on a set of training data. Then, the quality of the training is and evaluated on a new set of data, the test set. We consider that the classifier works if the true and predicted categories on the test set are similar.

Here, the number of features is $N_{\text{features}} = 2500$, which is the number of values of $I(q)$ measured values per signal. We also know the labels $D^{(i)}$ for each aggregate, which refers to its dimensionality (between 0 and 3): this is directly deduced from the way each protein aggregate (from which we computed the scattering signal) is constructed, as explained in Sec. 6.2.1.2 and 6.2.1.3. There is 248 data of scattering signals. The input layer of the neural network is of size N_{features} and the output layer of dimension 4 (the probability of the aggregate being of each dimension). We choose a network architecture with two hidden layers, of dimensions 25 and 12. We divide the dataset into a training set (89% of the data) and a test set, while ensuring that the training and test set have the same distribution of each type of aggregates. The distribution of each category was shown in Table 6.1. This distribution is uneven, because the database was initially built with the intention to characterize the fibers. Yet, we will see that the amount of data in the other categories is sufficient for the network to learn their common features.

The results of the prediction on the training and test set are given in Table 6.2a and b. The diagonal terms of each matrices count the data that were predicted correctly in each category. We see that most of the data are correctly predicted (92%) on the training set, which means that the network was correctly trained. Most of the data are also correctly predicted on the test set (90%), which mean that the network learned to recognize characteristics of the signals within a category that are not specific to the signals in the training set. We also tested aggregates of a protein that was not used to build the aggregates and the signals used in the training set (tubulin), for which the dimension was predicted correctly. Most of the incorrect predictions in the test set concern the gas. We suggest that this is because we built the scattering signals of gas aggregates from density maps containing several aggregates, which is not adapted for Crysol.

These results suggest that the dimensionality of a protein aggregate is a characteristic that can be identified in the scattering signal, despite the fact that the protein aggregates are different in many other aspects, like the protein it is composed of, its size, the organization of the aggregate, and the orientations of the proteins in the aggregate. The amount of data used in this study is low, and these results would need confirmation on larger dataset. Yet, despite the small amount of data, a very simple network reached 90% good predictions on the test set, which suggest that identifying the dimensionality is not a complex task. Beyond the specificities of this dataset, this results suggest that the machine-learning approach is efficient to identify the dimensionality of a protein aggregate from its scattering signal, and that we could also use it on experimental data. Indeed, it probably learns in which regions of the scattering vector q the linear regime which scales

		Predicted dimension			
		Gas	Fiber	Sheet	Crystal
True Dimension	Gas	39	10	0	0
	Fiber	0	157	0	0
	Sheet	0	3	22	0
	Crystal	0	0	5	15

a Training set

		Predicted dimension			
		Gas	Fiber	Sheet	Crystal
True Dimension	Gas	4	2	0	0
	Fiber	0	19	1	0
	Sheet	0	0	3	0
	Crystal	0	0	0	2

b Test set.

Table 6.2: The neural network learns to classify the aggregate category. We observe mostly correct predictions in both the training (a) and test (b) sets. Results taken from [139].

.

as the dimensionality of the aggregate is predicted, which we could not do in Sec. 6.2.2.1.

Here, we attempted to make the aggregates within the same category as different as possible. In experiments, besides the differences between the aggregates, the experimental conditions might also vary from one measurement to the other for aggregates with the same dimensionality. This adds an additional complexity to the identification of the dimensionality. Yet, we used a network with a very simple architecture, and it is possible that a more complex architecture could help to classify more complex data.

6.3 Attempt of dimensionality identification on crystallographic data

The preliminary numerical analysis confirmed that the signature of the dimensionality of aggregates with different characteristics could be measured within the range of crystallographic data. This suggests that it is possible to test the hypothesis that dimensionality reduction is a generic phenomenon arising from the self-assembly of complex particles that protein, by analyzing scattering signals collected by crystallographers for other purposes. We now investigate the challenges inherent to the experimental measurements. For this, we collected scattering signal of protein aggregates in different physicochemical conditions, to identify the dimensionality of the protein aggregates in each sample. This was done in collaboration with M. Spano (ISB, Grenoble) and W. Shepard (Synchrotron Soleil, Saclay). In Sec. 6.3.1, we explain how we collected scattering signals of different protein aggregates in different physicochemical conditions. The expected dimensionality of the aggregate was known only in some cases. In Sec. 6.3.2 we show that we could not separate the signals

of proteins from its background, which prevented us from detecting the aggregate dimensionality with standard methods, such as the measure of the slope of the linear regime. In Sec. 6.3.3, we use statistical methods such as machine learning or principal component analysis. We show that when the expected aggregate dimensionality of a fraction of the data is known, we can identify the dimensionality in other signals collected within the same experimental conditions, but that it cannot be generalized to data collected in different experimental conditions. Some experiments and analysis were done together with M. Garic during his master thesis [142].

6.3.1 Presentation of experiments

Here, we show how we measured the scattering signals of different proteins in a large variety of physicochemical conditions. As in Sec. 6.2, we aim at collecting data where the aggregates have the same dimensionality, but they are different in other aspects. In Sec. 6.3.1.1, we present the diversity of protein samples and experimental conditions for which we collected a scattering signal. In Sec. 6.3.1.2, we show how the experimental set-up enables a precise control of the part of the sample that scatters the X-ray. In Sec. 6.3.1.3, we explain how the scattering signals are extracted from the scattering image measured during the experiment, and show that the scattering signal highly depends on the experimental set-up.

6.3.1.1 Protein samples

Here, we highlight the differences between the protein samples we analyzed. We performed two series of experiments. In the first, we carefully chose the proteins and the experimental conditions such that the expected protein aggregate was known. In the following, we refer to this as the *labelled data*. In the second series of experiments, we used proteins samples already available because they were used for other projects by M. Spano. In this case, the expected aggregates was not known, and we refer to this as the *unlabeled data*.

The first experiments was conducted on actin, α-synuclein, tubulin and tau proteins, because those proteins are stable in both fibrillar aggregate (dimensionality 1) or monomeric state (dimensionality 0), depending on the drugs added in the solution. Actin proteins in physiological condition self-assemble into fibers. Upon addition of a toxin called latrunculin, it does not assemble [143]. Upon addition of fascin, the actin filament form bundles [144]. α-synuclein forms fibrils or ribbons upon addition of Tris-HCl [145, 146]. It remains in a monomeric state in regular physiological conditions. Tubulin forms microtubule upon addition of taxol [147]) and is in a monomeric state if nocodazole is added to the solution [148]. Tau protein is also tubular or monomeric depending on the experimental conditions. Our collaborators prepared solutions of proteins in each of these conditions (3 for actin, 3 for α-synuclein, 2 for tubulin and 2 for tau). The scattering signal was collected for each of them, for different concentration of proteins. This corresponds to data of scattering signals of proteins aggregates for which we know the dimensionality.

In the second series of experiments, we collected scattering signals of proteins samples that were prepared by M. Spano to try to achieve crystallization. Because crystallization is hard to achieve, the proteins were prepared in a large variety of physiological conditions. Therefore, we did not have control over the expected aggregate dimensionality. The proteins in these unlabelled aggregates are ispE, thaumatin and a protein for which we do not give the name, as a request of M. Spano. In the following, we call it "NATA".

The maximum concentration of proteins is typically between 10 and 20 mg/mL. All initial solutions of proteins are also diluted by a factor 2 or 3, depending on the sample. For the unlabeled samples, the concentration of protein and of the precipitants in the buffer was varied. The proteins we studied, and the corresponding aggregates, are listed in

Date	Protein	Types of aggregates	Buffer	# plates	# data
02-2021	actin	Filament, bundles, monomer	Yes	2	74
02-2021	α-synuclein	Filament, ribbons, monomer	Yes	2	76
02-2021	tau	Tubules, monomer	Yes	1	36
02-2021	tubulin	Tubules, monomer	Yes	1	36
03-2022	ispE	unknown	No	1	132
03-2022	thaumatin	unknown	No	1	93
03-2022	NATA	unknown	No	2	678
04-2022	ispE	unknown	Yes	2	209
04-2022	thaumatin	unknown	Yes	1	168

Table 6.3: We collect experimental data of different types of aggregates from different proteins in different experimental condition. The column Buffer indicates whether a droplet of buffer alone was measured for each droplet of buffer+proteins. The column # plates indicates in how many crystallization plates identical samples were measured. The last column indicates the number of scattering signals measured for each protein.

Table 6.3. We therefore have a set of protein samples containing monomers, fibers, tubes, or unknown aggregates.

6.3.1.2 Set-up

A droplet of each of the proteins samples described above are deposited on a 96-well plate. Here, we show how the experimental set-up of the Proxima-2 synchrotron beam line then enables to collect the scattering signal on precise positions of the droplet.

The droplets of protein solutions are deposited in each well of the plate with a robot that precisely control the volume of the droplet. In the schematic of Figure 6.8, we represent the plate in blue and the droplets of protein sample in yellow. For some data, a droplet of the buffer without the protein were also deposited in the same well, next to the solution of protein droplet. In Table 6.3, we summarize for which samples the buffer was collected in the column called Buffer. The plate is then sealed and left for equilibrating. Several droplets of the same protein samples were also deposited on different crystallization plates of the same kind. In Table 6.3, we show, in the column called # plates, on how many different plates a given sample was deposited. We therefore have data of proteins samples in different experimental conditions.

The plate is placed under the X-ray beam as shown in Figure 6.8. The position of the plate is controlled by a robotic arm. Live microscope imaging of the droplet enable to determine with precision the desired position of the shot in the droplet. Once the shot position has been determined, the microscope is displaced away from the beam path, and the X-ray

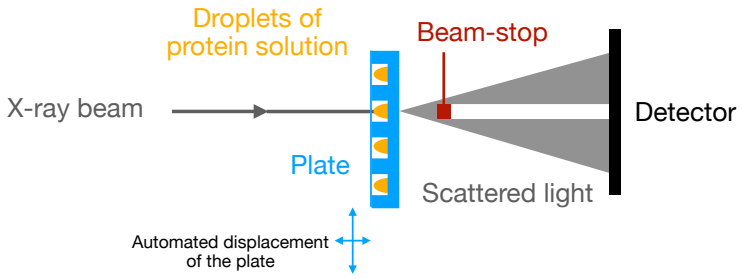

Figure 6.8: In the Proxima-2 beam line, the scattering signals are measured at precise position of the droplet on the plate, and the scattering signal of all the droplets in the plate is made faster with the automatic displacement of the plates.

is shot. All of these steps are controlled remotely with the software CRIBLEUR [149]. The scattering signal is then measured on a *detector*, *i.e.* a screen that detects the intensity of the light as a function of the positions, shown in black in Figure 6.8. The light that is not scattered (in the center of the scattering image) is stopped by a beam-stop (in red in Figure 6.8), because its intensity is much larger than the intensity of the scattered light, and would damage the detector. The CRIBLEUR allows collecting several scattering signals on the same droplets, at precise positions. It is also possible to measure the scattering signal in positions on the plate where there is no droplet. This enables to measure the scattering signal of the plate only. This partially automated set-up enables the collection of a large amount of data in a limited period of time. The combination of microscope and scattering images also ensures control over the position of the shot.

6.3.1.3 Signal analysis

Here, we explain how the image measured on the detector is converted to a scattering signal $I(q)$ and we show that the scattering signal depends on the protein but also on the buffer and the plate.

The intensity of the light on the detector depends on the angle it was scattered with. This is shown in Figure 6.9a, where the levels of gray indicate different level of intensities. In this image, we also see a white region in the center where no photons are detected: this is the beam-stop. The black grid corresponds to the limits between different portions of the detector. The scattering does only depend on the angle q, which is why we observe rings around the central point. We average the signals at identical values of q. This is called azimuthal integration and is done automatically from the detector properties with the python library PyFAI [150].

The result of this integration is shown for some actin scattering signal in Figure 6.9b. We plot measurements of solutions with actin in a monomeric state, or forming filaments, on two different plates, that are referred to as (p1) and (p2). We also plot in dashed line the corresponding buffer measurements. We notice that scattering signals of the same protein solution collected on different plates are slightly different. Those signals are very different from those of Figure 6.6, which we computed from numeric protein aggregates. Only the left most part of panel (b) corresponds to the range of q used to compute the numeric protein aggregates. The rest of the variations in the scattering intensity are due to the measure of the background, and not of the protein: the dashed line and solid line have the same variations. We also notice that for some values of scattering vector ($q \approx 3\text{Å}^{-1}$ for instance), the intensity of the scattering light of the buffer (plate+buffer), is larger than that of the scattering light of the protein solution (plate+buffer+proteins). This makes it difficult to directly extract the protein scattering signal by subtracting the buffer scattering signal. Despite the uncertainties due to the plate variability, the scattering signals of monomeric

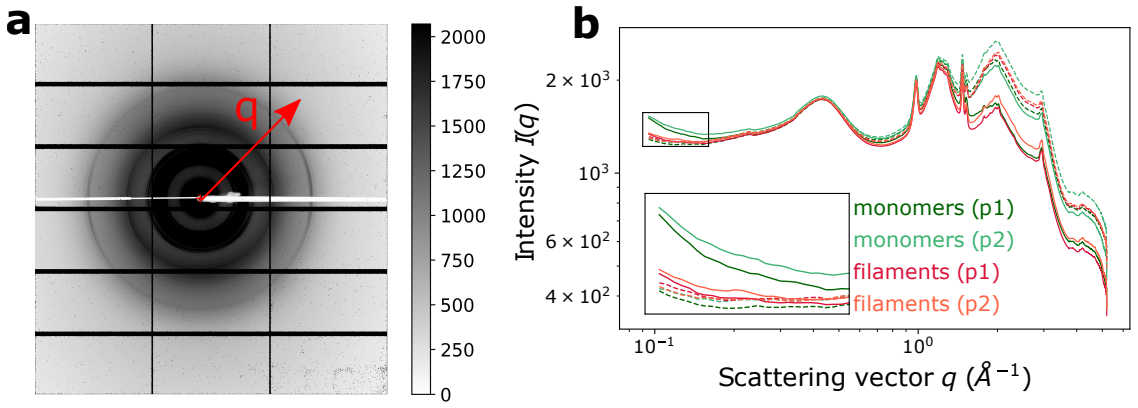

Figure 6.9: The scattering image is integrated, and the scattering signals mostly measure the scattering of the plate. a) Scattering image measured on the detector. The vertical black lines correspond to junctions between portions of the detector where no photons are detected. The white vertical line in the center is the beam-stop. The scattering vector $\vec{q}$ corresponds to a given radius b) Result of the azimuthal integration at constant values of the scattering vector $\vec{q}$. Measures on two different plates (p1 and p2). We show scattering signal of actin in monomeric (green) and filament (red) state. The dashed lines correspond to scattering signals of the buffer of the protein of the same color.

proteins can be distinguished from that of filaments of protein, because they have different variations in the low q regions: in the inset, the red signals are distinguishable from the green signals.

6.3.2 Extraction of the signal of the protein is difficult in crystallographic experiment

From the preliminary analysis of the scattering signals, it seems that the plate introduces significant variabilities in the measures, and that the intensity of scattering signal of the buffer is sometimes larger than the corresponding signal for the buffer and the protein. This makes it challenging to extract the signal of the proteins itself, which is the quantity of interest. In Sec. 6.3.2.1, we explain why the scattering intensity of the buffer can be larger than that of the protein solution in crystallographic experiments. In Sec. 6.3.2.2 we propose a method to subtract the buffer and plate scattering signal systematically with statistical methods, and show that it works to subtract the signal of the plate, but not the signal of the plate and the buffer.

6.3.2.1 Direct subtraction of the background is challenging

The measured scattering signal depends on the proteins in the solution, but also on the buffer, and on the plate. Here, we explain why the scattering signal of the protein is easily isolated in SAXS experiments, but not in crystallographic experiments. These differences were not known in the onset of this project, and their consequences are one of the most important challenges to the systematic analysis of scattering signal of protein aggregates.

Every molecule is a scatterer, this means also the plate, the air in the beam path, and the solvent with the precipitant agents (which we refer as the *salt*). We can decompose the total measured intensity as a sum of the individual intensities of the scatterers. The scattering signal is proportional to the number of scatterers on the beam path. For these reasons, the scattering signal of the air can be written as $h_{\text{plate}} I_{\text{plate}}(q)$, where h_{plate} is the height of the plate at the position where the X-ray was shot, and $I_{\text{plate}}(q)$ is the measure of the scattering signal of the plate of unitary height. We similarly define h_{air}, $I_{\text{air}}(q)$, h_{solvent} and $I_{\text{solvent}}(q)$, for the scattering by the air and the solvent. Then the total scattering signal I_{tot} depends on those quantities, and on $I_{\text{proteins}}(q)$, the scattering signal of the

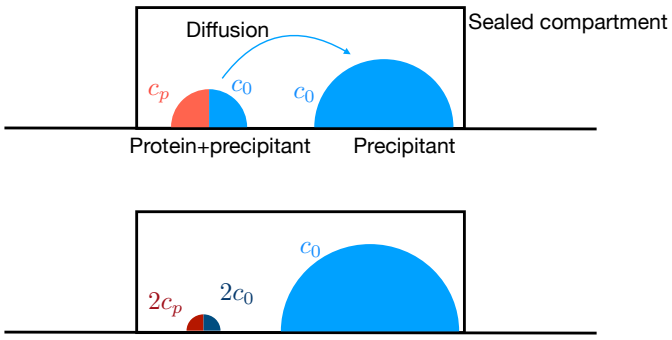

Figure 6.10: The volume of the buffer droplet and protein droplet are different in crystallographic experiments, because this enables to obtain larger concentrations of protein, and achieve crystallization.

protein, which contains the information about the aggregate dimensionality (the one we are interested in).

$$I_{\text{tot}}(q) = h_{\text{air}} I_{\text{air}}(q) + h_{\text{plate}} I_{\text{plate}}(q) + h_{\text{solvent}} I_{\text{solvent}}(q) + I_{\text{proteins}}(q) \tag{6.11}$$

In SAXS experiments, the scattering signal of the solute is evaluated with precision: the scattering of a solution containing all molecules, but the solute of interest is measured, and subtracted to the total signal [151]. This enables to subtract the solvent, the plate, and the air on the beam path. The volume of buffer needs to be exactly the same as that of the solution of interest, and the identical plate is reused for both measures.

In crystallographic experiment, the objective is to achieve crystallization and detect Bragg peaks in the scattering image. Protein often crystallizes when the concentration of protein in the solution is large. We describe how the protein droplets are prepared in such experiments, and show that it complexifies the extraction of the protein scattering signal. To increase the concentration of proteins in the sample *a posteriori*, two droplets are deposited in a sealed compartment, as illustrated in Figure 6.10: the first droplet contains protein and buffer (blue and red) and the second contains buffer only. Diffusion of buffer solution from the protein droplet to the buffer droplet then enable the increase of the concentration in the protein droplet. This process is detailed in [136, 152]. It is used by the Proxima-2 beam-line in synchrotron. Because the two droplets do not have the same volume, it is not possible to consistently subtract the scattering signal of the solvent and the plate by measuring the scattering signal of the buffer.

The dependence of the scattering signal of the buffer on the volume of the droplet could of course be measured independently, but the main idea of this project, which is to use existing data, because they are numerous, would then be lost. Moreover, a precise control over the buffer volume would not be sufficient, because the scattering signal of the plate alone displays some variability, from a position to another on the plate.

6.3.2.2 Attempt to subtract background with signal processing

The direct subtraction of the buffer signal is not possible in our data. Here, we suggest taking advantage of the large amount of data collected on the plate alone, and on the buffer, to overcome this challenge. We propose a method to determine the separate contribution of each element in the measured scattering signal that rely on an algorithm called *non-negative matrix factorization*, which we will explain. We show that it enables to extract the plate scattering signal from the buffer+plate scattering signal. We then show that this technique is not sufficient to extract the buffer+plate scattering signal from the protein+buffer+plate signal.

The scattering signals of the plate alone measured at different plate positions are not equal. These differences are explained by small variations of the height of the plate, which can vary from one position on the plate to the other. These variations could also be due to difference of the volume of air in the beam path, which can vary with time, because of temperature differences for instance. Based on eq. 6.11, we make this hypothesis that these dependencies are linear. To remain generic, we decompose the signal into two *components*, $I_0^{(\text{plate})}(q)$ and $I_1^{(\text{plate})}(q)$, which do not depend on the position of the measure. The dependency of the position and time on the measure are encompassed in coefficients, which we call a_k and b_k, that do not depend on the scattering vector q. Then a signal measured on the plate $I_k^{(\text{plate})}(q)$ is decomposed as

$$I_k^{(\text{plate})}(q) = a_k I_0^{(\text{plate})}(q) + b_k I_1^{(\text{plate})}(q) + \epsilon_k(q) \tag{6.12}$$

$\epsilon_k(q)$ is a correction to this decomposition, which should be as small as possible. The values of $I_0^{(\text{plate})}(q)$, $I_1^{(\text{plate})}(q)$, a_k, b_k and $\epsilon_k(q)$ are determined with non-negative matrix factorization [153] of the plate data, such that the norm of ϵ_k is minimum. This method ensures that I_0, I_1, a_k and b_k are all positive, which enables to interpret the two first terms in eq. 6.12 as physical contributions of the scatterers to the scattering intensities.

Then, any scattering signal $I_p(q)$ collected on a droplet of solution measures both the scattering of the plate and the solution:

$$I_p(q) = I_p^{(\text{plate})}(q) + I_p^{(\text{sol})}(q) \tag{6.13}$$

$$= a_p I_0^{(\text{plate})}(q) + b_p I_1^{(\text{plate})}(q) + I_p^{(\text{sol})}(q) \tag{6.14}$$

By measuring a_p and b_p, we can separate the contribution of the plate and of the solution in the measure of $I_p(q)$, and isolate the scattering signal of the solution only. We will use this method to isolate the buffer solution from the measures of plate+buffer, and the protein solution on the measure of plate+buffer+protein. This is possible because a large amount of data were collected on the same plate. Indeed, during one of series of experiments, we collected 87 scattering signals of the plate alone, at different location on the same plate. We also collected 88 (respectively 62) scattering signals of the buffer solution for ispE (respectively thaumatin) proteins, at different concentrations and location on the same plate. We consider a decomposition to be valid if the measure of $\|\epsilon_k\|/\|I_k\|$ is low, which means that we did isolate the components that varied from one measure to the other, up to a small correction.

We decompose the 87 plate data with this method. The relative error of this fit $\|\epsilon_k\|/\|I_k\|$, is between 0.001 and 0.04. Components $I_0^{(\text{plate})}(q)$ and $I_1^{(\text{plate})}(q)$ are shown in Figure 6.11a. $I_1^{(\text{plate})}$ is the most important contribution to the scattering signals, and we recognize the shape of signal also observed in Figure 6.9, which confirms that most of the scattering intensity is due to scattering of the plate. We similarly decompose the 68 and 62 signals of the plate+buffer for the ispE and Thaumatin bueffer. Here, we assume that there are three components that explain the variations between the signals: the plate, the air, and the volume of buffer. The 0^{th} components for both decomposition are denoted $I_0^{(\text{ISPE})}$ and $I_0^{(\text{Thaum.})}$. These components are shown in Figure 6.11b and c. The decomposition was not as good for the buffer signals as for the plate signals: relative errors of the fit could reach 0.19 (resp. 0.12) for the ispE (resp. thaumatin) buffers. As a consequence, the components shown in the Figure do not seem to correspond to the scattering signal of individual components, because they are very discontinuous.

We then extract the buffer signal from the decomposition of the plate signals computed above and the measure of the plate+buffer signal, according to eq. 6.13. We plot the extracted signals of the buffer solution of thaumatin and ispE proteins in Figure 6.12(a-b).

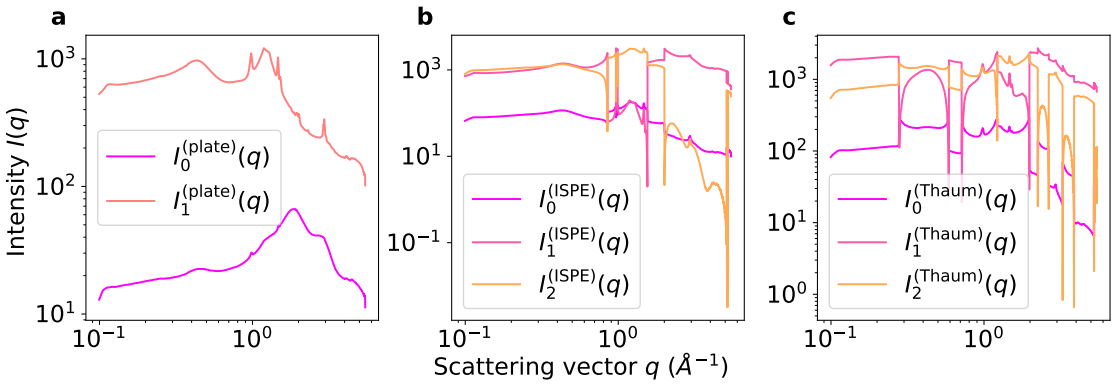

Figure 6.11: The decomposition of the plate signals is physical, but not the decomposition of the buffers signals. Results of the decomposition of eq. 6.12. a) Decomposition of the plate alone signals in two components. (b-c) Decomposition of the plate+buffer signals in three components for the ispE buffer (b) and thaumatin buffer (c)

Despite the fact that the extracted signal is not always positive, the intensity of the signal at a given value of q seem to vary monotonously with the concentration of salt in the buffer (coded with the colors in panels (a) and (b)). This is an indication that the signals we extracted with this method correspond to the scattering of the buffer.

We also extract the protein signal from the decomposition of the plate+buffer (computed from an equation similar to eq. 6.13). We show the extracted in Figure 6.12c and d. Because the decomposition of the plate+buffer signal was not satisfactory, neither is the extraction of the protein signals: the signals are also discontinuous. However, if these signals only measure the contribution of the proteins, they should not depend on the concentration of the buffer where the proteins were. We do not observe a dependency of those signals on the concentration of salt in the buffer: the intensity of the signal at a given value of q is not proportionally related to the color coding for the salt concentration in panels (c) and (d). This is an indication that the signals we isolated only contain information relative to the proteins, and not to the buffer. Yet, the poor quality of the decomposition of the buffer signals, the discontinuities of the extracted protein signals, and the fact that they take negative values prevent us from attempting to analyze those signal further and to identify the dimensionality of the protein aggregates.

We tried to take advantage of the large amount of data available to get around the problem of the precise buffer measurements in crystallography experiments. The statistical decomposition we used (non-negative matrix factorization) enabled us to measure a proxy of the scattering signals of the protein buffers. Yet, we could not use this technique to extract the scattering signal of the proteins themselves. Therefore, we could not extract the dimensionality of the protein aggregates within the available data. These results suggest that resorting to statistical methods and comparing a large amount of data collected in variable experimental conditions (like different plates) could help distinguish the scattering signals concerning different types of protein aggregates

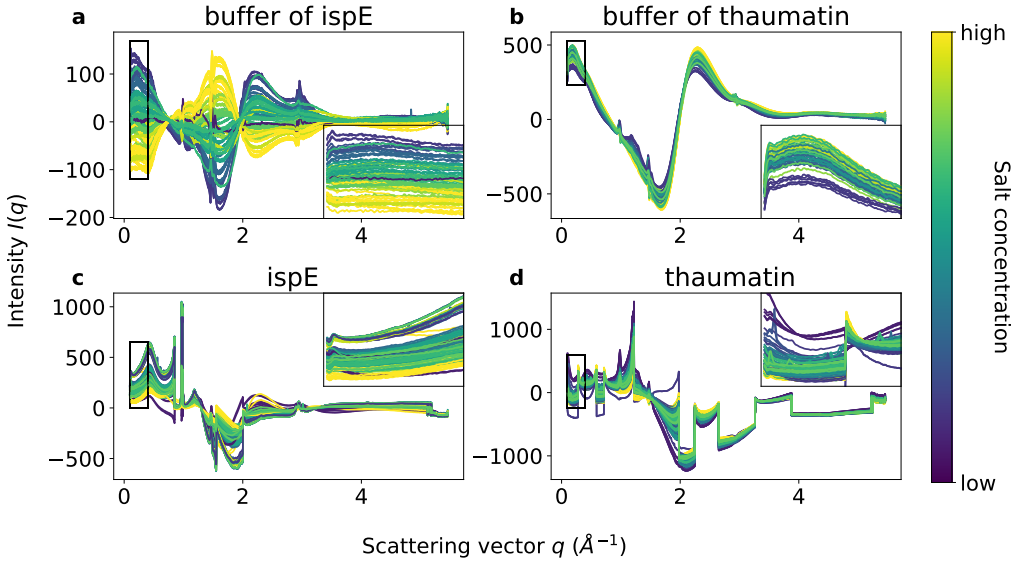

Figure 6.12: We can extract the buffer signal, but not the protein signals. We subtract the plate components from the scattering signals of the buffer of ispE (a) or thaumatin (b), according to eq. 6.13. We subtract the plate and buffer components from the scattering signals of ispE solution (c) and thaumatin solution (d). The extracted buffer signals vary monotonously with the concentration of salt in the buffer, but the extracted protein signals do not.

6.3.3 Statistical analysis

In this section, we propose two statistical methods to identify the dimensionality of aggregates within measured scattering signals for which the background (solvent, air, and plate) were not subtracted. We use principal component analysis [107] on the labelled dataset (for which the dimensionality of the aggregate is known) in Sec. 6.3.3.1, and show that we can discriminate experimental measure of scattering signals according to the dimensionality of the aggregate. In Sec. 6.3.3.2, we show that this method was however not sufficient on the unlabeled dataset, because most of the variability between the scattering signals is caused by the proteins in the aggregate, and not by the dimensionality of the aggregates they form. Because of this, we use machine-learning methods in Sec. 6.3.3.3 and show that it is partially sufficient to distinguish some macroscopic characteristics of the aggregates.

6.3.3.1 PCA on labelled data enables to identify the dimensionality

Here, we remind on which concepts principal component analysis rely, and show that it enables to distinguish scattering signal of aggregates of different proteins measured in different experimental conditions, by the dimensionality of their aggregate.

For each scattering signal k, we measure a list of numbers 2000 numbers $I(q)^{(k)}$, one for each value of q. We can project this high-dimensional lists on a two-dimensional space in which the variability between the data is maximal. If two data are similar, they should be closed to one another in this projection. We use this method on the labelled dataset introduced in Sec. 6.3.1.1. If the data for similar aggregate categories are close, despite differences in the protein it is composed of or in the crystallographic plate on which it was measured, it means that the aggregate category can be systematically distinguished.

From the numerical analysis of the previous section on artificial aggregates of protein, we identified that it was relevant to look at the signal for low values of q, in logarithmic scale. Thus, we only do the decomposition of the scattering intensity within the range $[0.1, 0.34]\text{Å}^{-1}$. We also standardize the data such that each signal is of zero mean and unit

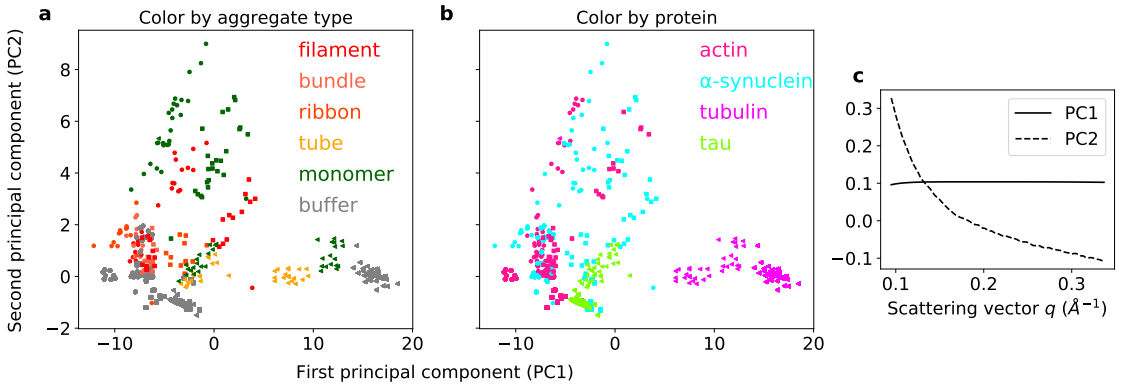

Figure 6.13: We distinguish the dimensionality of the aggregates in the labelled dataset with PCA. We plot the projection of the data in the plane of the first and second principal components. a) The data points are colored according to the type of aggregate. b) The data points are colored according to the protein. (a-b) data of same symbols (round, square, triangle) were collected on the same plate. c) First and second principal components as a function of the scattering vector.

variance. This ensures that the measured variability between signals is not due to a global shift in the intensity.

The result of the PCA on the data of the labelled proteins is shown in Figure 6.13. In panels (a) and (b), we show the same projection, but the data points are colored according to the aggregate category in (a), and to the protein in (b). For the measures on actin and α-synuclein, the data seem to be segregated according to the dimensionality of the aggregate, rather than according to the proteins. Indeed, most of the red points are close in the principal component projection (panel a), even if they correspond to signal of different proteins (panel b). On the other hand, data collected on tubulin and tau are clearly clustered according to the protein. However, within the data that correspond to tubulin (pink points on the lower right on Figure 6.13b), the clustering then happen according to the type of aggregate (see the same points on Figure 6.13a). Signals of the same solution measured on two different plates (different symbols) are still well distinguishable in this projection, which means that the plate still caries an important influence on the variability between signals.

These results suggest that there is a trace of the dimensionality of the protein aggregate within a raw scattered signal, but we cannot identify it systematically. Indeed, the differences between the plate, or the type of protein accounts for too much variability in the dataset. A possible solution to this problem would be to be more systematic in the data collection, such that all categories of aggregates for all proteins are collected for all plates. Then statistical techniques like PCA could isolate separately the influence of each, and there might be a subspace where only the difference between the aggregate category explain the differences between the signals. Another method is to train a neural network to only isolate the features of the characteristics of the signals explained by the aggregate category.

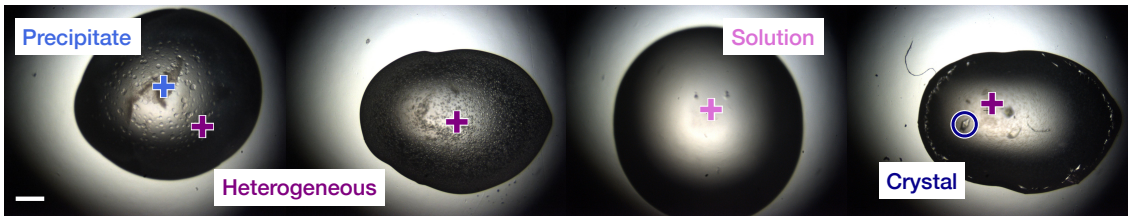

Figure 6.14: We label the scattering signals of the unlabeled dataset according to observations with the microscope. The cross (or circle) indicates where the X-ray beam was aimed, and the color indicates the label. One droplet can have different measures with different droplet.

6.3.3.2 Unlabeled data are mostly separated according to the protein with PCA

Here we use the same PCA decomposition as in Sec. 6.3.3.1, on the unlabeled data, *i.e.* those for which we do not know the dimensionality of the protein aggregate (introduced in Sec. 6.3.1.1). We introduce a classification based on visual characterization of the individual droplets, that may be related to the aggregates formed in the protein samples, and show that those visual differences do not explain variability between the scattering signals.

Even if we had no prior information on the dimensionality of the proteins aggregates, we were able to visualize the droplet with a microscope, as explained in Sec. 6.3.1.2. The droplets had visual characteristics that enabled us to classify them in a hopefully relevant way. In Figure 6.14, we show examples of microscope images of those categories. As explained in Sec. 6.3.1.2, we measure the scattering signals at different position on the droplet. For instance, if there is a crystal in the droplet (see the right most image in Figure 6.14), we can compute the scattering signal by shooting on the crystal, and next to the crystal. We distinguish between four categories relative to the macroscopic characteristics. The *crystals* are easily recognizable (dark blue on Figure 6.14). When the droplet is homogeneous, we label the corresponding signals as *solution* (third drop, pink on the Figure). When there are some heterogeneities, such as the ones on the two first droplets, we label the corresponding signals as *heterogeneous* (dark purple). Finally, we sometimes identify a *precipitate* on the droplet (first image, in blue), and label the scattered signal measured on these precipitates accordingly. Several measures on a single droplet can then have different labels. This categorization refers to macroscopic properties of the protein solutions, like the fact that it crystallizes, precipitates, leads to heterogeneity in the solution, or on the contrary, remains soluble. In the absence of further information on the dimensionality of the protein aggregates, we test whether the scattering signals are clustered on the PCA projection according to those categories, which we refer to as the *macroscopic properties* of the aggregates.

In Figure 6.15, we show the projection of the data in the principal component space, colored either by the characterization of the droplet (panel a), or by protein (panel b). Here, it is clear that the protein and the plate is what influences the most the variance of the data: we do not observe clustering according to the categories we defined (the points of different colors are separated in panel (b) and mixed in panel (a)). It is not clear, however, whether this is because the categories are ill-defined, because there is no information of the categories in the scattering signals, or whether principal component is not the correct tool to evaluate this contribution.

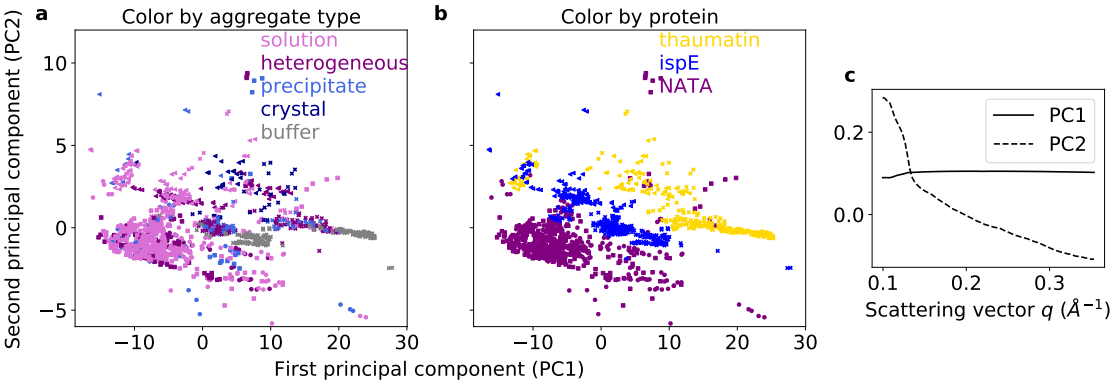

Figure 6.15: We cannot distinguish the scattering signals of identical macroscopic properties (introduced in Figure 6.14) with PCA, and the variability is only explained by the type of protein, or the plate. We plot the projection of the data in the plane of the first and second principal components. a) The data points are colored according to the type of aggregate. b) The data points are colored according to the protein. (a-b), the same symbol indicates that the measures were collected on the same plate. c) First and second principal components as a function of the scattering vector.

6.3.3.3 Machine learning to learn macroscopic characteristics of protein solutions

Because the principal component analysis only separated the data according to the plate of the measure or of the protein, on the unlabeled dataset, we train a neural network classifier to identify the macroscopic categories defined in Sec. 6.3.3.2 on this dataset. We show that it is partially possible to distinguish the scattering signals labeled as solution (no heterogeneities) from a scattering signal of a heterogeneous solutions, which might result from aggregations of the proteins in the sample.

We train a feed-forward neural network (which we defined in Sec.3.2.3 of Chapter 3 and used also in Sec. 6.2.2.2), to classify scattering signals according to their macroscopic properties. For the network to learn that a large part of the variability between signals is explained by the plate, we train it on data of all the signals, including those of plate only (labeled *plate*) and of buffer+plate (labeled *buffer*). We label the scattering signals of protein solutions with the categories introduces above, and we put under the same label the category the *heterogeneous* and *precipitate* categories defined above. Indeed, it is not clear from the observation of the microscope images of Figure 6.14 that heterogeneous solutions are not caused by small precipitates. Moreover, we initially trained the algorithm while keeping those categories separated, and the accuracy of the prediction between those two categories was not better than a random prediction. This might suggest that this categorization is physically irrelevant. Therefore, there are three categories for the signals of protein solutions: *solution*, *precipitate/heterogeneous* and *crystal*. We also train the algorithm on signals of the different proteins (thaumatin, ispE, and NATA), such that it will recognize characteristics of the signal that are independent of protein specificities. Finally, we either train the algorithm on the dataset of all the experimental results collected in 2022 (four plates, three proteins, 396 data, see table 6.3 of Sec. 6.3.1.1), or on the one plate where scattering signal of buffer solution was also collected (one plate, two proteins and 168 data). If the macroscopic properties are related to the microscopic organization of the proteins, they should have an influence on the scattering signals, and the neural network should be able to learn what this influence is, and classify the data accordingly.

The network is composed of four layers of 25 neurons. The analysis is done for values of the scattering vector in $[0.06, 5]\text{Å}^{-1}$, which corresponds to ≈ 1800 values of q. We first

		Predicted label				
		Plate	Buffer	Solution	Prec./Het.	Crystal
True label	Plate	20	0	0	0	0
	Buffer	0	56	0	0	0
	Solution	2	0	8	12	0
	Prec./Het.	0	0	1	56	0
	Crystal	0	0	0	8	5

a) Scattering signal collected on the same crystallization plate

		Predicted label				
		Plate	Buffer	Solution	Prec./Het.	Crystal
True label	Plate	25	0	0	0	0
	Buffer	0	52	1	4	0
	Solution	3	0	82	44	1
	Prec./Het.	2	0	33	129	2
	Crystal	0	0	1	0	45

b) Scattering signal collected on the different crystallization plates

Table 6.4: A neural network distinguishes the signals of buffer and plate from the signal of proteins solutions, but makes some mis-predictions on the macroscopic properties of the aggregates with proteins: a lot of *solutions* are labeled as *precipitate/heterogeneous*. Results taken from [142]

project those 1800-dimensional data in the 30 dimensional principal components space, and train the neural network on this data of reduced dimensionality, which is a usual and effective method for real-data classification with machine learning [154]. The dataset is divided between training and test set (80% and 20% of the data).

We show the prediction of the algorithm on the test sets, for a dataset containing only data on the same plate (table 6.4a), and of different plates (table 6.4b). For the data collected on one plate, the learning accuracy (the number of good prediction divided by number of data) is of 86% (table 6.4a), and for the total dataset, it is 77% (table 6.4b). In both cases, the buffer and plates signals are correctly classified: there are only diagonal terms in the table for those categories. The macroscopic properties of the data are however not correctly identified by the neural network in both cases: there are almost as many wrong predictions (solutions that are predicted as heterogeneous, for instance) than good predictions (solutions that are predicted as solutions). It might be because the macroscopic properties we identified by looking at the droplet in the microscope do not correspond to feature of the protein organization that are measured by the scattering signal. Another possibility is that the amount of data on different proteins and different plates was insufficient for the network to learn that the differences from one signal to the other because of those variations is irrelevant.

Discussion

In this chapter, we explored whether it is possible to take advantage of the large amount of data collected by crystallographers on protein solutions to measure occurrence of dimensionality reduction of the aggregate of particles with complex interactions, like proteins. To test this idea, we analyzed scattering signals of artificial and real protein aggregates in the range of crystallographic length scale. We expected difficulties due to the small range of length scale where information on aggregate dimensionality could be measured in crystallographic data. This did not appear too problematic, and we were able to detect aggregate dimensionality by fitting scattering signal of the protein.

However, in crystallographic experiments, the measured scattering signal is highly dependent on the background, which cannot be subtracted in a straightforward way. We could then take advantage of the large amount of data we collected to classify them according to the dimensionality of the aggregate. This requires to train a neural network on scattering signal of protein of known dimensionality. We showed that this approach enables to identify aggregate categories, provided that the data used to train the classifier were collected in different experimental conditions, and in particular, on different crystallization plates.

These preliminary results suggest that it is possible to identify aggregates of reduced dimensionality from crystallographic data with statistical methods. Yet, for this study to be performed systematically on scattering signals collected in different beam lines, on different plates, and for different protein, more scattering signals measured on aggregate of known dimensionality would be needed. In particular, a training database should be cautiously built such that protein aggregates of similar dimensionality are measured in experimental conditions that are as broad as possible.

7 - Synthèse en francais

Dans les cellules vivantes, les protéines s'*auto-assemblent* en agrégats de différentes formes pour réaliser des fonctions biologiques [1]. La forme des agrégats est contrôlée par les interactions locales entre les protéines individuelles. Ces interactions sont par exemple des interactions attractives entre les résidus à la surface de la protéine, ou reposent sur la complémentarité de forme entre les protéines. Malgré la diversité de ces interactions, il n'y a que quelques catégories typiques d'agrégats de protéines : des oligomères de quelques particules, des fibres, des capsides virales ou des micelles. Les protéines peuvent également former des agrégats cristallins dans des conditions très spécifiques. La relation entre les positions des résidus attractifs et la forme de l'agrégat est également non triviale : un contact entre deux protéines en interaction fait intervenir plusieurs résidus, mais une mutation d'un seul résidu peut modifier le résultat de l'auto-assemblage, d'un monomère à une fibre par exemple [32]. En outre, des protéines très similaires provenant d'organismes différents s'assemblent en différents agrégats [17, 18].

Les modèles de particules à patchs collants (patchy) sont utilisés pour comprendre les principes génériques de l'auto-assemblage : les particules colloïdales avec des patchs attractifs sont étudiées dans des simulations numériques et des expériences. Ces modèles permettent de retrouver des agrégats fibrillaires, cristallins ou oligomères [51]. Cependant, ils ne tiennent pas compte du fait que des variations subtiles de l'attraction des patchs peuvent modifier radicalement la forme de l'agrégat. Ils ne fournissent pas non plus une compréhension systématique de la relation entre les propriétés des interactions entre les particules individuelles et la forme de l'agrégat. Enfin, la plupart d'entre eux ne permettent pas des variations continues et indépendantes des différentes interactions attractives d'une particule. Dans cette thèse, nous émettons l'hypothèse que des particules aux interactions complexes, comme les protéines, peuvent avoir plusieurs interactions attractives qui sont incompatibles à cause de contraintes géométriques : il y a de la frustration géométrique. La frustration géométrique a été bien étudiée dans le contexte des systèmes de spin denses. Cependant, les conséquences des interactions incompatibles à courte portée sur le résultat de l'auto-assemblage ne semblent pas bien comprises.

Dans le Chapitre 2, nous introduisons un modèle de particules identiques, avec des interactions locales, sur un réseau à deux dimensions. En modifiant la force des interactions, nous retrouvons tous les agrégats stéréotypés issus de l'auto-assemblage des protéines. En

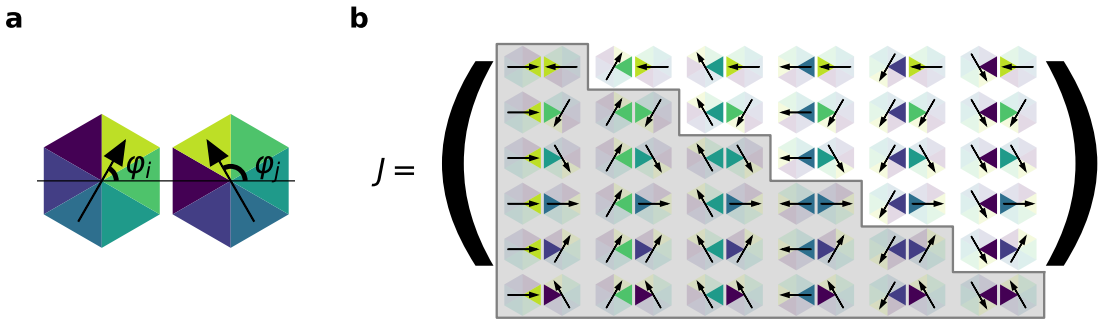

Figure 7.1: La carte d'interaction J énumère les différentes manières dont deux particules peuvent interagir. a) Les faces en contact dépendent de l'orientation des deux particules. b) La carte d'interaction est représentée par une matrice symétrique (seulement la partie grisée correspond aux interactions indépendantes)

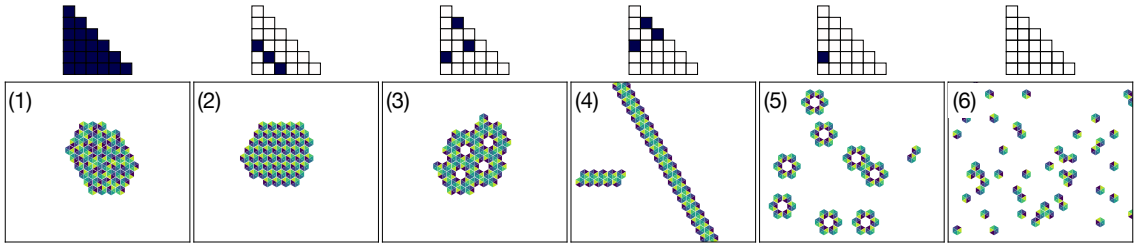

Figure 7.2: Des agrégats très divers sont obtenus en changeant la carte d'interaction. Pour chaque image, la carte d'interaction est présentée, les interactions attractives sont colorées en bleu ($-10kT$) et les interactions neutres en blanc ($0kT$). On observe des agrégats amorphes (1), des cristaux (2), des cristaux poreux, appelés *éponge* (3), des fibres (4), des oligomères (5) ou des monomères (7).

pratique, nous considérons des particules hexagonales qui sont en contact par leurs faces. Il y a 6×6 paires de faces et $6\times7/2 = 21$ paires de faces distinctes qui peuvent être en contact lorsque deux particules hexagonales occupent des sites voisins du réseau. Par conséquent, nous définissons une *carte d'interaction* avec 21 interactions indépendantes, pour lesquelles nous choisissons une force arbitraire, négative (interaction attractive) ou positive (interaction répulsive). Nous présentons cette matrice d'interaction (Figure 7.1) et des exemples d'agrégats stéréotypés dans la Figure 7.2. Nous étudions l'auto-assemblage de particules avec une carte d'interaction choisie avec un recuit simulé de type Monte-Carlo vers une température finie. Cela permet de faire varier la directionalité des interactions (lesquelles sont attractives ou répulsives). Changer la force des interactions revient aussi à changer la température du système. Pour une particule avec un ensemble donné d'interactions locales, on peut alors déterminer la forme de l'agrégat résultant de l'auto-assemblage à l'équilibre des particules.

Dans le Chapitre 3, nous comprenons la relation entre les interactions locales et la forme de l'agrégat en tirant un grand nombre de cartes d'interactions aléatoires dans une distribution gaussienne. La moyenne de la distribution correspond à l'affinité globale de la particule et l'écart type de la distribution correspond à l'anisotropie de la particule. Nous montrons que les particules avec des interactions anisotropes s'auto-assemblent en agrégats de formes non triviales, tels que des agrégats poreux ou des agrégats de tailles grandes, mais finies. Un exemple d'agrégat pour chaque valeur d'affinité et d'anisotropie est présenté en Figure 7.3. Nous introduisons également une mesure de frustration : la différence d'énergie entre la configuration d'équilibre du système, et une configuration du système sans contraintes géométriques des particules. Cette mesure révèle que la plupart des agrégats sont frustrés géométriquement. Nous introduisons une classification des agrégats en huit catégories (liquide, cristaux, éponge (agrégats poreux), fibres, cristallite (cristaux partiellement assemblés), micelles (agrégats de taille importante, mais limitée à cause de l'effet de surface), oligomères et monomères). Avec un algorithme d'apprentissage automatique supervisé, nous classifions le résultat de 9000 auto-assemblage de particules avec une carte d'interaction prédéfinie. La distribution de chaque catégorie d'agrégat pour une affinité et une anisotropie fixée de la particule sont présentées en Figure 7.4. Ceci confirme que les interactions anisotropes conduisent plus souvent à la formation de fibres, d'éponges et de micelles, qui sont des agrégats non-triviaux. Enfin, nous utilisons l'apprentissage automatique pour tester quelles quantités calculées à partir des cartes d'interaction permettent de prédire la catégorie de l'agrégat. Nous constatons que l'énergie de l'organisation périodique la plus stable des particules est un bon prédicteur de la forme de l'agrégat. Ces résultats suggèrent que les particules avec des interactions anisotropes sont sujettes à la frustration, et réduisent la taille de l'agrégat pour éviter cette frustration.

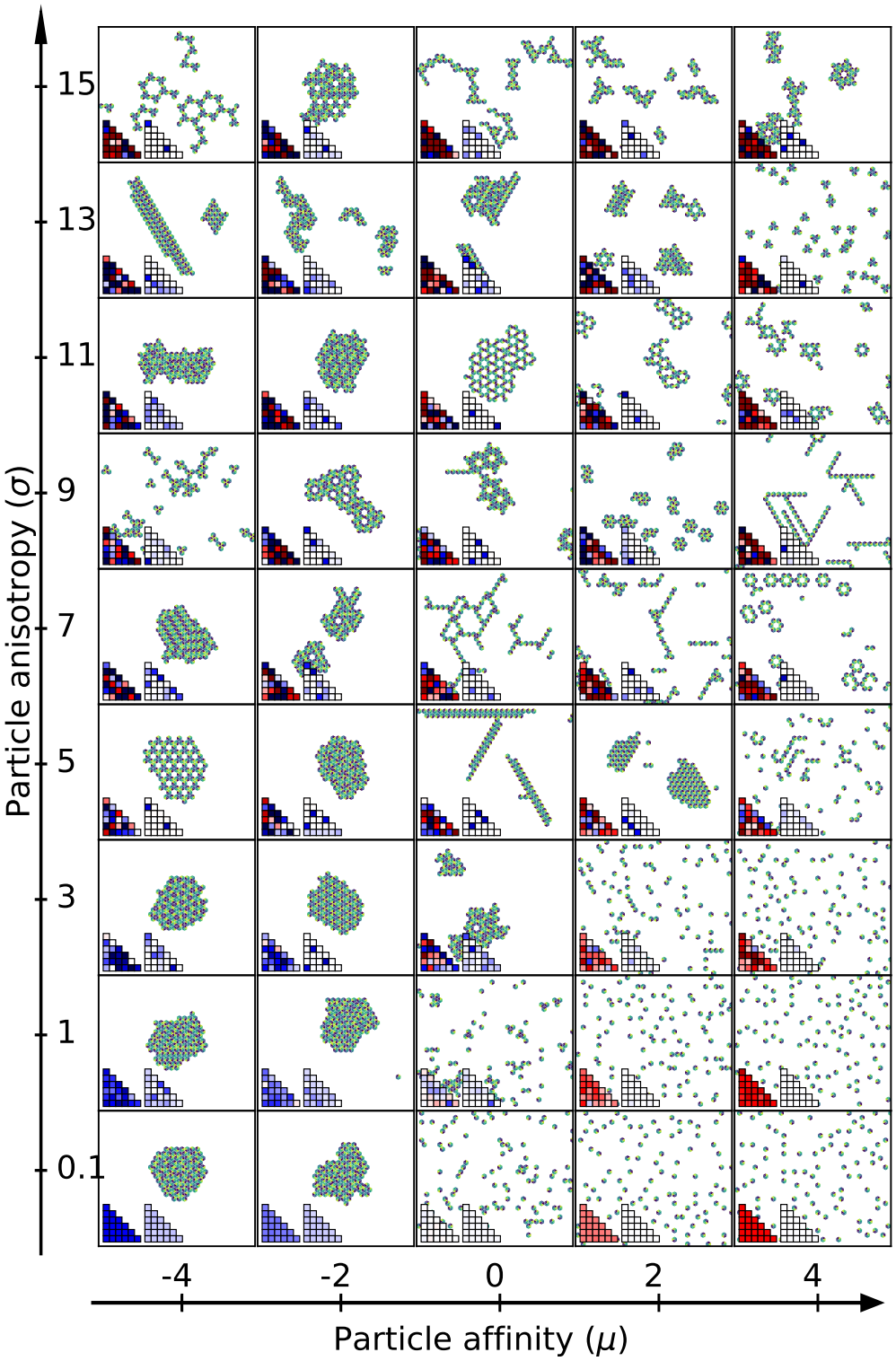

Figure 7.3: Des particules avec des cartes d'interaction aléatoires résultent en des agrégats de formes diverses. Pour chaque valeur d'affintité μ et d'anisotropi σ, on montre une image du système, la carte d'interaction (en bas à gauche) et la carte de densité (en bas à droite). Les énergies dans les cartes d'interactions sont répulsives (rouge) ou attractive (bleues). La densité de lien est codée en nuance de bleu, de 0% des liens en blanc, à 5% des liens en bleu foncé.

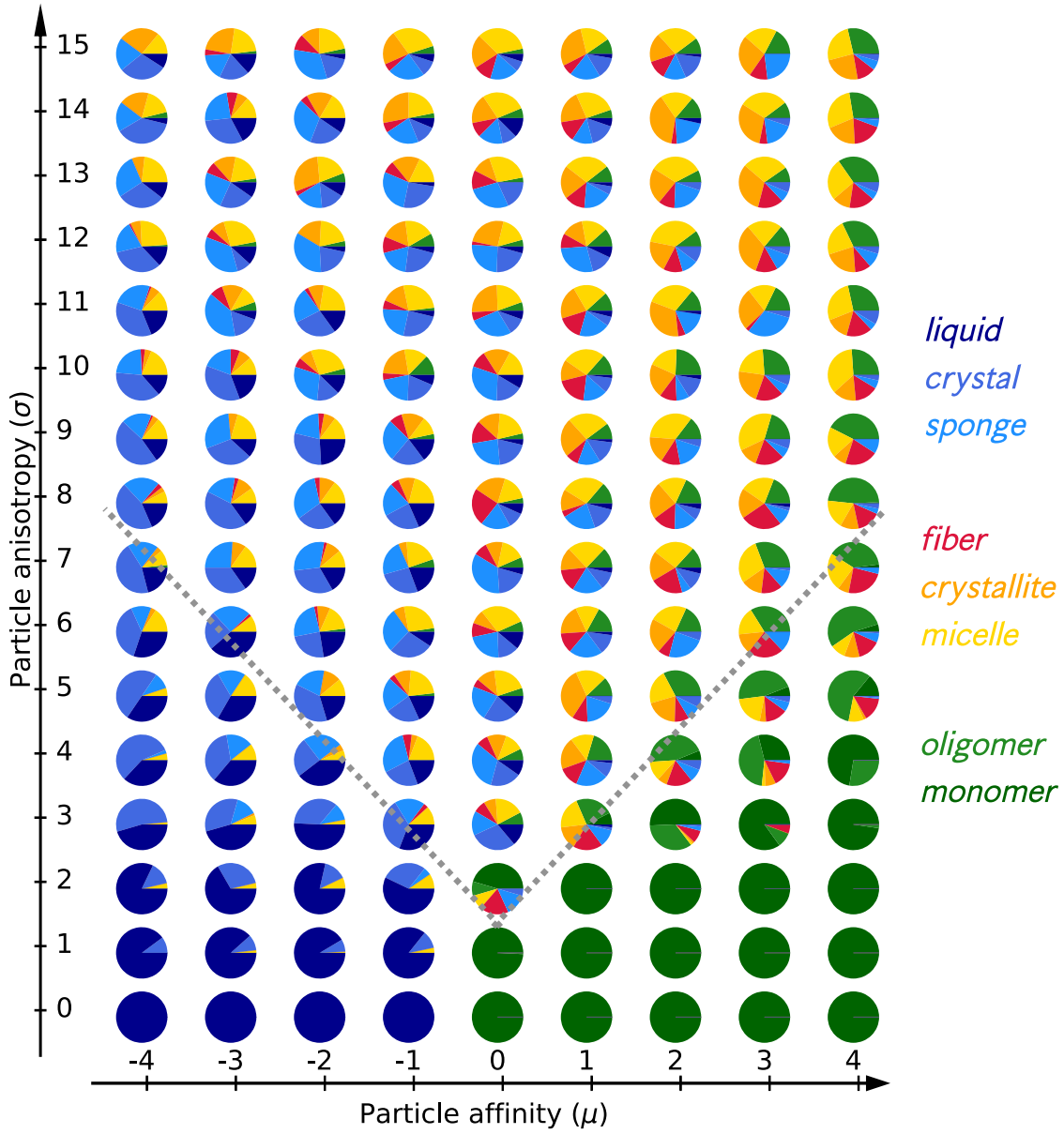

Figure 7.4: Les particules anisotropes forment des agrégats moins triviaux. Chaque diagramme indique la proportion de chaque catégories d'agrégats pour une valeur donnée d'affinité et d'anisotropie.

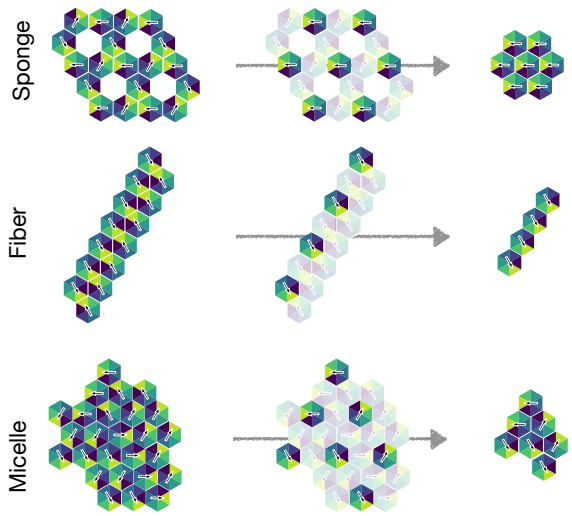

Figure 7.5: En un pas de renormalisation, les deuxièmes voisins de la particule deviennent les premiers voisins. On montre des exemples d'une transformation de renormalisation pour une éponge, une fibre et une micelle.

Dans le Chapitre 4, nous introduisons une transformation numérique de renormalisation dans l'espace réel de la carte d'interaction, qui garantit que le nombre d'occurrences de chaque paire de particules est conservé dans un réseau de maille plus grande. Un exemple de cette transformation pour quelques systèmes est montrée en Figure 7.5. Nous utilisons la renormalisation comme outil pour explorer l'espace des paramètres de dimension 21 des interactions des particules, plutôt que pour mesurer les exposants critiques au voisinage d'une transition de phase. Nous identifions trois types de points fixes stables de la procédure de renormalisation. Ces points fixes sont les cartes d'interactions vers lesquelles convergent les trajectoires de renormalisation à partir de cartes d'interaction aléatoires. Ils sont présentés en Figure 7.6. Ce sont les cartes d'interaction des particules attractives isotropes, celle des particules isotropes sans interaction et celle des particules conduisant à un motif cristallin périodique. Nous constatons également que la carte d'interaction d'une fibre est un point fixe instable. Les points-fixes de la renormalisation correspondent donc à des agrégats stéréotypés introduits au Chapitre 3. Nous montrons que le bassin d'attraction du point fixe isotrope sans interaction comprend la plupart des agrégats de tailles finies, tandis que les agrégats de taille infinie se renormalisent vers l'agrégat de particules isotropes, ou l'agrégat de particules cristallines, selon la périodicité de l'organisation des particules. La renormalisation permet donc de rationaliser l'existence de seulement quelques catégories de formes d'agrégats, malgré le grand espace des paramètres et la complexité de l'interaction des particules.

Dans le Chapitre 5, nous introduisons une carte d'interaction spécifique qui conduit à un agrégat cristallin avec des lignes de défaut favorables, que nous appelons un agrégat *camembert*. L'interaction cristalline et l'interaction conduisant à la ligne de défaut sont incompatibles. Pour cette raison, l'agrégat camembert est frustré et peut avoir une taille finie à l'équilibre. La taille de l'agrégat est contrôlée par la force relative de l'interaction cristalline et de l'interaction de la ligne de défaut. Ces deux interactions sont illustrées en bleu clair et bleu foncé sur la Figure 7.7, ainsi qu'un exemple d'agrégat camembert. Nous établissons analytiquement le diagramme de phases à température nulle, et vérifions que les agrégats de camembert sont observés dans les simulations numériques à température finie dans le régime des paramètres où ils sont les plus stables. Les résultats numériques dans différentes zones du diagramme de phase sont montrées en Figure 7.8. Nous montrons

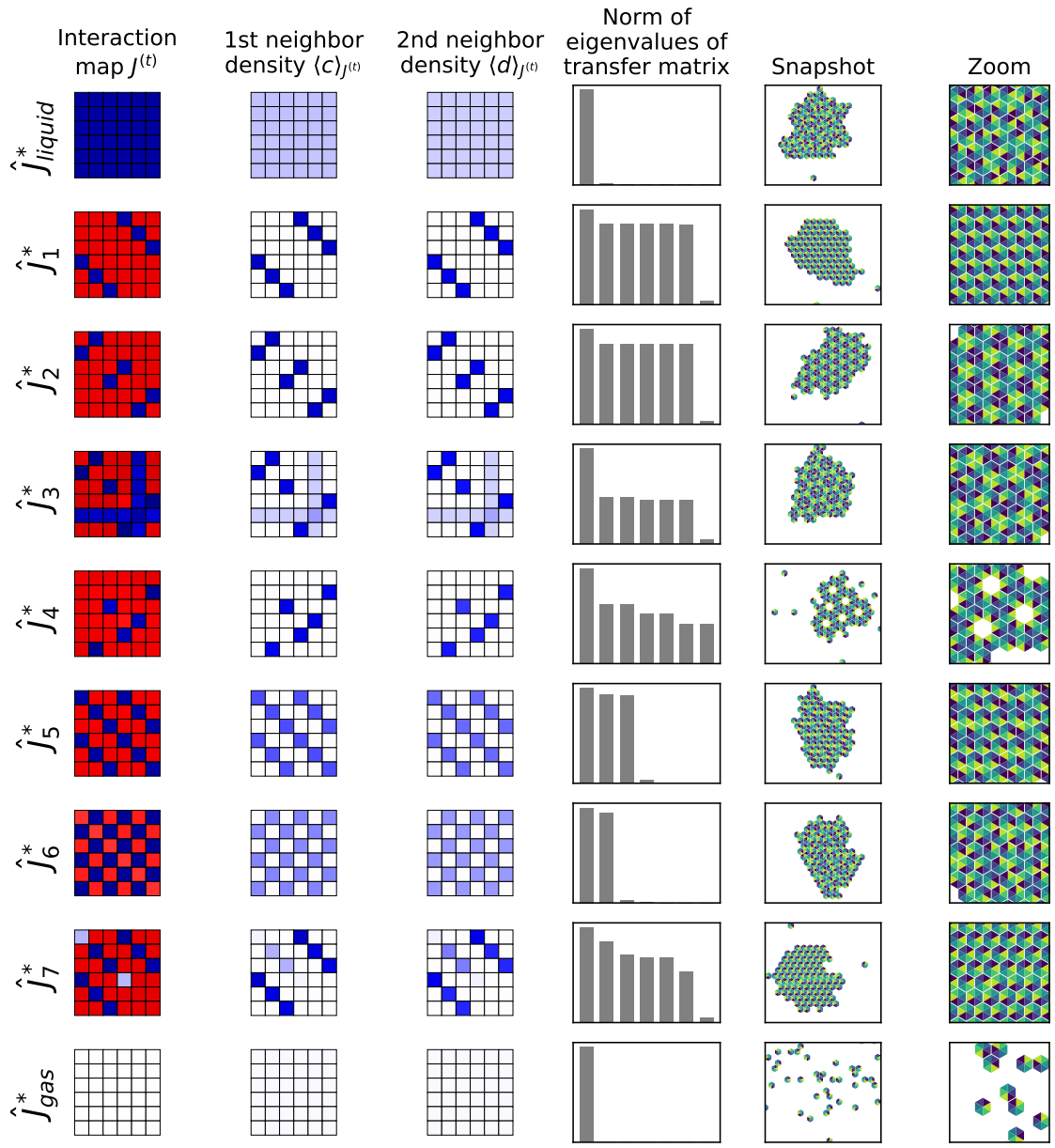

Figure 7.6: Les points fixes de la renormalisation ont les mêmes densités de premier et seconds voisins. La plupart correspondent à une organisation périodique des particules dans un agrégat dense.

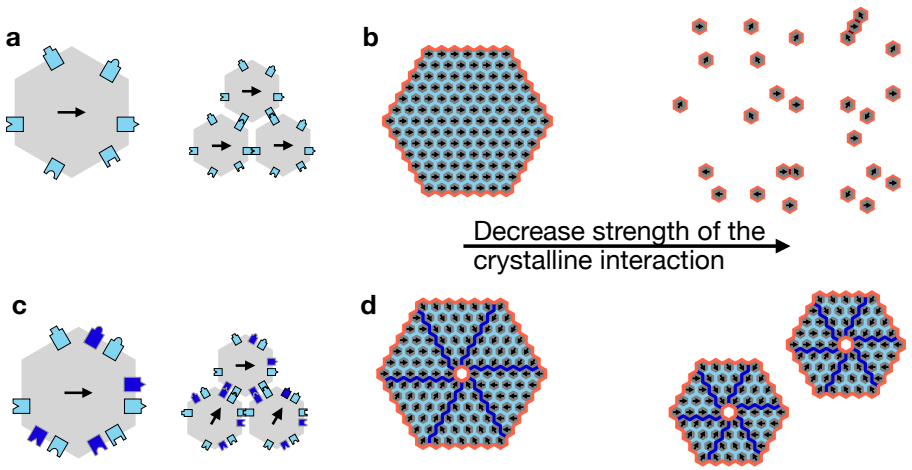

Figure 7.7: Les interactions directionnelles sont conçues pour que les particules forment un cristal avec des lignes de défaut (en bleu foncé). a) Si les interactions locales alignent les particules, elles forment un cristal. b) Diminuer la force de cette interaction dissout le cristal, sans obtenir de tailles intermédiaires. c) Au contraire, combiner les interactions cristallines (bleu clair) avec les interactions de ligne (bleu foncé), conduit à une compétition des interactions. d) Elles conduisent à la formation d'agrégat camembert, qui peuvent être de taille intermédiaire si l'interaction cristalline est faible

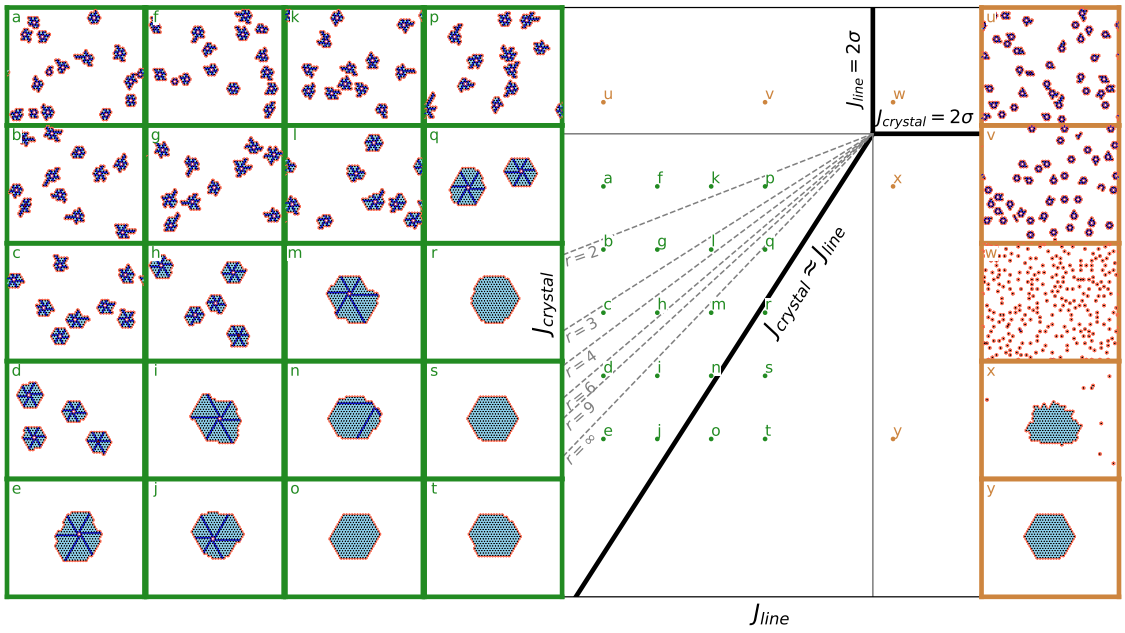

Figure 7.8: On observe des camemberts dans les simulations numériques dans la région du diagramme de phase où ils sont stables à température nulle. Pour un set de paramètres, indiqué sur le graphe par une lettre, on montre le résultat de la simulation numérique (à gauche ou à droite). Des camemberts de taille finie sont observés.

également que ce mécanisme permet d'auto-assembler des fibres de largeur contrôlée et finie. Ce mécanisme apparaît comme complémentaire des mécanismes d'assemblage auto-limité à l'équilibre qui reposent sur le design individuel de chaque particule, sur le fait que l'assemblage s'auto-ferme, ou sur la déformabilité des particules [79]. Les agrégats de camembert ont des frontières ouvertes, et toutes les particules sont indéformables et identiques. Nous expliquons des idées préliminaires pour tester ce design dans une réalisation expérimentale hors réseau à partir d'origami d'ADN [42].

Nos résultats et des études antérieures [155] suggèrent que l'une des caractéristiques de l'agrégation de particules avec des interactions complexes est la formation d'agrégats fibrillaires, qui ont une dimensionnalité réduite (ce sont des agrégats 1D dans un espace 2D ou 3D). Dans le chapitre 6, nous proposons une méthode pour tester cette hypothèse et détecter systématiquement l'auto-assemblage en fibre de protéines avec des interactions arbitraires. Nous proposons d'utiliser les signaux de diffusion collectés dans des expériences cristallographiques, où les protéines sont assemblées dans des conditions physico-chimiques variables, qui modifient les interactions entre les protéines. Seules certaines de ces conditions conduisent à la cristallisation, alors que la forme des agrégats dans les autres conditions physico-chimiques n'est pas étudiée. Nous collectons des signaux de diffusion d'agrégats de protéines, construits de manière numériques à partir des densités d'électrons des protéines individuelles et montrons que nous pouvons reconnaître des agrégats fibrillaires. Nous collectons également des signaux expérimentaux de diffusion de protéines dans des conditions variables, conduisant ou non à des agrégats fibrillaires. Bien qu'il soit difficile d'extraire le signal de diffusion de la protéine de son arrière-plan dans de telles expériences, il est possible de tirer parti de la grande quantité de données disponibles et d'utiliser l'apprentissage automatique supervisé pour reconnaître les agrégats fibrillaires. Cela suggère que la reconnaissance des fibres protéiques pourrait se faire parmi les signaux de diffusion collectés dans des expériences cristallographiques avec un algorithme d'apprentissage supervisé, à condition que le réseau de neurones soit entraîné sur une grande variété d'agrégats protéiques de dimensionnalité connue, et dans différentes configurations expérimentales.

Bibliography

1. Goodsell, D. S. & Olson, A. J. Structural symmetry and protein function. *Annual review of biophysics and biomolecular structure* **29,** 105 (2000).
2. Ge, J. *et al.* Architecture of the mammalian mechanosensitive Piezo1 channel. *Nature* **527,** 64–69 (2015).
3. White, J. L. *et al.* A comparison of the structures of apo dogfish M4 lactate dehydrogenase and its ternary complexes. *Journal of Molecular Biology* **102,** 759–779 (1976).
4. Purushotham, P., Ho, R. & Zimmer, J. Architecture of a catalytically active homotrimeric plant cellulose synthase complex. *Science* **369,** 1089–1094 (2020).
5. Caspar, D. L. & Klug, A. *Physical principles in the construction of regular viruses* in *Cold Spring Harbor symposia on quantitative biology* **27** (1962), 1–24.
6. Holt, C., Carver, J., Ecroyd, H. & Thorn, D. Invited review: Caseins and the casein micelle: Their biological functions, structures, and behavior in foods. *Journal of dairy science* **96,** 6127–6146 (2013).
7. Fermi, G., Perutz, M. F., Shaanan, B. & Fourme, R. The crystal structure of human deoxyhaemoglobin at 1.74 Å resolution. *Journal of molecular biology* **175,** 159–174 (1984).
8. Lu, J.-X. *et al.* Molecular structure of β-amyloid fibrils in Alzheimer's disease brain tissue. *Cell* **154,** 1257–1268 (2013).
9. Zheng, J. *et al.* The rational design and structural analysis of a self-assembled three-dimensional DNA crystal. *Nature* **461,** 74–77 (2009).
10. Stehle, T., Gamblin, S. J., Yan, Y. & Harrison, S. C. The structure of simian virus 40 refined at 3.1 Å resolution. *Structure* **4,** 165–182 (1996).
11. Gates, S. N. *et al.* Ratchet-like polypeptide translocation mechanism of the AAA+ disaggregase Hsp104. *Science* **357,** 273–279 (2017).
12. Goodsell, D. S. *Casein Micelle and Fat Globule in Milk* https://pdb101.rcsb.org/sci-art/goodsell-gallery/casein-micelle-and-fat-globule-in-milk.
13. Nishi, H., Hashimoto, K. & Panchenko, A. R. Phosphorylation in protein-protein binding: effect on stability and function. *Structure* **19,** 1807–1815 (2011).
14. Noree, C., Sato, B. K., Broyer, R. M. & Wilhelm, J. E. Identification of novel filament-forming proteins in Saccharomyces cerevisiae and Drosophila melanogaster. *Journal of Cell Biology* **190,** 541–551 (2010).
15. Hayouka, Z. *et al.* Inhibiting HIV-1 integrase by shifting its oligomerization equilibrium. *Proceedings of the National Academy of Sciences* **104,** 8316–8321 (2007).
16. Hakim, M. & Fass, D. Dimer interface migration in a viral sulfhydryl oxidase. *Journal of molecular biology* **391,** 758–768 (2009).

17. Sinha, S., Gupta, G., Vijayan, M. & Surolia, A. Subunit assembly of plant lectins. *Current opinion in structural biology* **17,** 498–505 (2007).

18. Lynch, E. M. *et al.* Human CTP synthase filament structure reveals the active enzyme conformation. *Nature structural & molecular biology* **24,** 507–514 (2017).

19. Levy, E. D. & Teichmann, S. A. Structural, evolutionary, and assembly principles of protein oligomerization. *Progress in molecular biology and translational science* **117,** 25–51 (2013).

20. Khan, A. R. *et al.* Temperature-dependent interactions explain normal and inverted solubility in a γD-Crystallin mutant. *Biophysical journal* **117,** 930–937 (2019).

21. Moal, I. H., Moretti, R., Baker, D. & Fernández-Recio, J. Scoring functions for protein–protein interactions. *Current opinion in structural biology* **23,** 862–867 (2013).

22. Keskin, O., Tuncbag, N. & Gursoy, A. Predicting protein–protein interactions from the molecular to the proteome level. *Chemical reviews* **116,** 4884–4909 (2016).

23. Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.-C. & Deane, C. M. Progress and challenges in predicting protein interfaces. *Briefings in bioinformatics* **17,** 117–131 (2016).

24. Das, S. & Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Scientific reports* **11,** 1–12 (2021).

25. Lensink, M. F. *et al.* Blind prediction of homo-and hetero-protein complexes: The CASP13-CAPRI experiment. *Proteins: Structure, Function, and Bioinformatics* **87,** 1200–1221 (2019).

26. Dykes, G. W., Crepeau, R. H. & Edelstein, S. J. Three-dimensional reconstruction of the 14-filament fibers of hemoglobin S. *Journal of molecular biology* **130,** 451–472 (1979).

27. Knowles, T. P., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nature reviews Molecular cell biology* **15,** 384–396 (2014).

28. Rambaran, R. N. & Serpell, L. C. Amyloid fibrils: abnormal protein assembly. *Prion* **2,** 112–117 (2008).

29. Lafon, P.-A. *et al.* Fungicide residues exposure and β-amyloid aggregation in a mouse model of Alzheimer's disease. *Environmental health perspectives* **128,** 017011 (2020).

30. Uversky, V. N., Li, J., Bower, K. & Fink, A. L. Synergistic effects of pesticides and metals on the fibrillation of α-synuclein: implications for Parkinson's disease. *Neurotoxicology* **23,** 527–536 (2002).

31. Shiels, A. & Hejtmancik, J. F. Biology of inherited cataracts and opportunities for treatment. *Annual review of vision science* **5,** 123–149 (2019).

32. Garcia-Seisdedos, H., Empereur-Mot, C., Elad, N. & Levy, E. D. Proteins evolve on the edge of supramolecular self-assembly. *Nature* **548,** 244–247 (2017).

33. McPherson, A. Introduction to protein crystallization. *Methods* **34,** 254–265 (2004).

34. McManus, J. J., Charbonneau, P., Zaccarelli, E. & Asherie, N. The physics of protein self-assembly. *Current opinion in colloid & interface science* **22,** 73–79 (2016).

35. Liu, Y., Gonen, S., Gonen, T. & Yeates, T. O. Near-atomic cryo-EM imaging of a small protein displayed on a designed scaffolding system. *Proceedings of the National Academy of Sciences* **115,** 3362–3367 (2018).

36. Zhang, Q. *et al.* DNA origami as an in vivo drug delivery vehicle for cancer therapy. *ACS nano* **8,** 6633–6643 (2014).

37. Cannon, K. A., Ochoa, J. M. & Yeates, T. O. High-symmetry protein assemblies: patterns and emerging applications. *Current opinion in structural biology* **55,** 77–84 (2019).

38. Fu, Y. *et al.* Single-step rapid assembly of DNA origami nanostructures for addressable nanoscale bioreactors. *Journal of the American Chemical Society* **135,** 696–702 (2013).

39. Sigl, C. *et al.* Programmable icosahedral shell system for virus trapping. *Nature materials* **20,** 1281–1289 (2021).

40. Chiesa, G., Kiriakov, S. & Khalil, A. S. Protein assembly systems in natural and synthetic biology. *BMC biology* **18,** 1–18 (2020).

41. Gerling, T., Wagenbauer, K. F., Neuner, A. M. & Dietz, H. Dynamic DNA devices and assemblies formed by shape-complementary, non–base pairing 3D components. *Science* **347,** 1446–1452 (2015).

42. Kopperger, E. *et al.* A self-assembled nanoscale robotic arm controlled by electric fields. *Science* **359,** 296–301 (2018).

43. Pumm, A.-K. *et al.* A DNA origami rotary ratchet motor. *Nature* **607,** 492–498 (2022).

44. Liu, N. & Liedl, T. DNA-assembled advanced plasmonic architectures. *Chemical reviews* **118,** 3032–3053 (2018).

45. Wang, M. *et al.* Programmable Assembly of Nano-architectures through Designing Anisotropic DNA Origami Patches. *Angewandte Chemie* **132,** 6451–6458 (2020).

46. Lin, Z. *et al.* Engineering organization of DNA nano-chambers through dimensionally controlled and multi-sequence encoded differentiated bonds. *Journal of the American Chemical Society* **142,** 17531–17542 (2020).

47. Style, R. W., Isa, L. & Dufresne, E. R. Adsorption of soft particles at fluid interfaces. *Soft Matter* **11,** 7412–7419 (2015).

48. Bae, J. *et al.* Programmable and reversible assembly of soft capillary multipoles. *Materials Horizons* **4,** 228–235 (2017).

49. Manoharan, V. N., Elsesser, M. T. & Pine, D. J. Dense packing and symmetry in small clusters of microspheres. *Science* **301,** 483–487 (2003).

50. Velikov, K. P., Christova, C. G., Dullens, R. P. & van Blaaderen, A. Layer-by-layer growth of binary colloidal crystals. *Science* **296,** 106–109 (2002).

51. Zhang, Z. & Glotzer, S. C. Self-assembly of patchy particles. *Nano letters* **4,** 1407–1413 (2004).

52. Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nature materials* **6,** 557–562 (2007).

53. Ding, T., Song, K., Clays, K. & Tung, C.-H. Fabrication of 3D photonic crystals of ellipsoids: convective self-assembly in magnetic field. *Advanced Materials* **21,** 1936–1940 (2009).

54. Wang, Z. L. *et al.* Superlattices of Self-Assembled Tetrahedral Ag Nanocrystals. *Advanced Materials* **10,** 808–812 (1998).

55. Chen, Q., Bae, S. C. & Granick, S. Directed self-assembly of a colloidal kagome lattice. *Nature* **469,** 381–384 (2011).

56. Tang, Z., Kotov, N. A. & Giersig, M. Spontaneous organization of single CdTe nanoparticles into luminescent nanowires. *Science* **297,** 237–240 (2002).

57. Wei, Y., Bishop, K. J., Kim, J., Soh, S. & Grzybowski, B. A. Making use of bond strength and steric hindrance in nanoscale "synthesis". *Angewandte Chemie International Edition* **48,** 9477–9480 (2009).

58. Bae, C., Moon, J., Shin, H., Kim, J. & Sung, M. M. Fabrication of monodisperse asymmetric colloidal clusters by using contact area lithography (CAL). *Journal of the American Chemical Society* **129,** 14232–14239 (2007).

59. Onoe, H., Matsumoto, K. & Shimoyama, I. Three-Dimensional Sequential Self-Assembly of Microscale Objects. *Small* **3,** 1383–1389 (2007).

60. Zerrouki, D., Baudry, J., Pine, D., Chaikin, P. & Bibette, J. Chiral colloidal clusters. *Nature* **455,** 380–382 (2008).

61. Du, C. X. *et al.* Programming interactions in magnetic handshake materials. *Soft Matter* **18,** 6404–6410 (2022).

62. Mao, Y., Cates, M. & Lekkerkerker, H. Depletion force in colloidal systems. *Physica A: Statistical Mechanics and its Applications* **222,** 10–24 (1995).

63. Sacanna, S. & Pine, D. J. Shape-anisotropic colloids: Building blocks for complex assemblies. *Current opinion in colloid & interface science* **16,** 96–105 (2011).

64. Sacanna, S., Irvine, W. T., Chaikin, P. M. & Pine, D. J. Lock and key colloids. *Nature* **464,** 575–578 (2010).

65. Mayarani, M., Heuvingh, J., du Roure, O. & Lenz, M. *3D printed colloids interacting through depletion interaction* 2023. In preparation.

66. Rothemund, P. W. Folding DNA to create nanoscale shapes and patterns. *Nature* **440,** 297–302 (2006).

67. Douglas, S. M. *et al.* Self-assembly of DNA into nanoscale three-dimensional shapes. *Nature* **459,** 414–418 (2009).

68. Wei, B., Dai, M. & Yin, P. Complex shapes self-assembled from single-stranded DNA tiles. *Nature* **485,** 623–626 (2012).

69. Ke, Y., Ong, L. L., Shih, W. M. & Yin, P. Three-dimensional structures self-assembled from DNA bricks. *science* **338,** 1177–1183 (2012).

70. Praetorius, F. *et al.* Biotechnological mass production of DNA origami. *Nature* **552,** 84–87 (2017).

71. Chothia, C. & Janin, J. Principles of protein–protein recognition. *Nature* **256,** 705–708 (1975).

72. Hu, Z., Ma, B., Wolfson, H. & Nussinov, R. Conservation of polar residues as hot spots at protein interfaces. *Proteins: Structure, Function, and Bioinformatics* **39,** 331–342 (2000).

73. McCoy, A. J., Epa, V. C. & Colman, P. M. Electrostatic complementarity at protein/protein interfaces. *Journal of molecular biology* **268,** 570–584 (1997).

74. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. Modelling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of molecular biology* **272,** 106–120 (1997).

75. Zhu, J. *et al.* Protein assembly by design. *Chemical reviews* **121,** 13701–13796 (2021).

76. Doyle, L. *et al.* Rational design of α-helical tandem repeat proteins with closed architectures. *Nature* **528,** 585–588 (2015).

77. Wicky, B. *et al.* Hallucinating symmetric protein assemblies. *Science* **378,** 56–61 (2022).

78. Pinheiro, A. V., Han, D., Shih, W. M. & Yan, H. Challenges and opportunities for structural DNA nanotechnology. *Nature nanotechnology* **6,** 763–772 (2011).

79. Hagan, M. F. & Grason, G. M. Equilibrium mechanisms of self-limiting assembly. *Reviews of modern physics* **93,** 025008 (2021).

80. Whitelam, S. & Jack, R. L. The statistical mechanics of dynamic pathways to self-assembly. *Annual review of physical chemistry* **66,** 143–163 (2015).

81. Tikhomirov, G., Petersen, P. & Qian, L. Fractal assembly of micrometre-scale DNA origami arrays with arbitrary patterns. *Nature* **552,** 67–71 (2017).

82. Wagenbauer, K. F., Sigl, C. & Dietz, H. Gigadalton-scale shape-programmable DNA assemblies. *Nature* **552,** 78–83 (2017).

83. Berengut, J. F. *et al.* Self-limiting polymerization of DNA origami subunits with strain accumulation. *ACS nano* **14,** 17428–17441 (2020).

84. Huntley, M. H., Murugan, A. & Brenner, M. P. Information capacity of specific interactions. *Proceedings of the National Academy of Sciences* **113,** 5841–5846 (2016).

85. Hormoz, S. & Brenner, M. P. Design principles for self-assembly with short-range interactions. *Proceedings of the National Academy of Sciences* **108,** 5193–5198 (2011).

86. Bohlin, J., Turberfield, A. J., Louis, A. A. & Sulc, P. Designing the self-assembly of arbitrary shapes using minimal complexity building blocks. *ACS nano* **17,** 5387–5398 (2023).

87. Zandi, R., Dragnea, B., Travesset, A. & Podgornik, R. On virus growth and form. *Physics Reports* **847,** 1–102 (2020).

88. Sadoc, J.-F. & Mosseri, R. *Geometrical frustration* (1999).

89. Brown, A. I., Kreplak, L. & Rutenberg, A. D. An equilibrium double-twist model for the radial structure of collagen fibrils. *Soft Matter* **10,** 8500–8511 (2014).

90. Karner, C., Dellago, C. & Bianchi, E. How patchiness controls the properties of chain-like assemblies of colloidal platelets. *Journal of Physics: Condensed Matter* **32,** 204001 (2020).

91. Akimenko, S., Gorbunov, V., Myshlyavtsev, A. & Fefelov, V. Self-organization of monodentate organic molecules on a solid surface—a Monte Carlo and transfer-matrix study. *Surface Science* **639,** 89–95 (2015).

92. Wannier, G. Antiferromagnetism. the triangular ising net. *Physical Review* **79,** 357 (1950).

93. Ronceray, P. & Le Floch, B. Range of geometrical frustration in lattice spin models. *Physical Review E* **100,** 052150 (2019).

94. Meiri, S. & Efrati, E. Cumulative geometric frustration and superextensive energy scaling in a nonlinear classical X Y-spin model. *Physical Review E* **105,** 024703 (2022).

95. Cho, S. & Ozaki, M. Blue Phase Liquid Crystals with Tailored Crystal Orientation for Photonic Applications. *Symmetry* **13,** 1584 (2021).

96. Selke, W. The ANNNI model—Theoretical analysis and experimental application. *Physics Reports* **170,** 213–264 (1988).

97. Wright, D. C. & Mermin, N. D. Crystalline liquids: the blue phases. *Reviews of Modern physics* **61,** 385 (1989).

98. Coles, H. J. & Pivnenko, M. N. Liquid crystal 'blue phases' with a wide temperature range. *Nature* **436,** 997–1000 (2005).

99. Potts, R. B. *Some generalized order-disorder transformations* in *Mathematical proceedings of the cambridge philosophical society* **48** (1952), 106–109.

100. Dress, C. & Krauth, W. Cluster algorithm for hard spheres and related systems. *Journal of Physics A: Mathematical and General* **28,** L597 (1995).

101. Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. *Equation of State Calculations by Fast Computing Machines* tech. rep. (1953).

102. Lundy, M. & Mees, A. Convergence of an annealing algorithm. *Mathematical programming* **34,** 111–124 (1986).

103. Zwicker, D. & Laan, L. Evolved interactions stabilize many coexisting phases in multicomponent liquids. *Proceedings of the National Academy of Sciences* **119,** e2201250119 (2022).

104. Gumbel, E. J. *Statistics of extremes* (Courier Corporation, 2004).

105. Mehta, P. *et al.* A high-bias, low-variance introduction to machine learning for physicists. *Physics reports* **810,** 1–124 (2019).

106. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural networks* **61,** 85–117 (2015).

107. Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science* **2,** 559–572 (1901).

108. Fisher, A., Rudin, C. & Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J. Mach. Learn. Res.* **20,** 1–81 (2019).

109. Wilson, K. G. The renormalization group: Critical phenomena and the Kondo problem. *Reviews of modern physics* **47,** 773 (1975).

110. Kadanoff, L. P. Scaling laws for Ising models near T c. *Physics Physique Fizika* **2,** 263 (1966).

111. Niemeijer, T. & Van Leeuwen, J. Wilson theory for spin systems on a triangular lattice. *Physical Review Letters* **31,** 1411 (1973).

112. Ma, S.-k. Renormalization group by Monte Carlo methods. *Physical Review Letters* **37,** 461 (1976).

113. Swendsen, R. H. Monte Carlo renormalization group. *Physical Review Letters* **42,** 859 (1979).

114. Onsager, L. Crystal statistics. I. A two-dimensional model with an order-disorder transition. *Physical Review* **65,** 117 (1944).

115. Garcıa-Pérez, G., Boguñá, M. & Serrano, M. Á. Multiscale unfolding of real networks by geometric renormalization. *Nature Physics* **14,** 583–589 (2018).

116. Villegas, P., Gili, T., Caldarelli, G. & Gabrielli, A. Laplacian renormalization group for heterogeneous networks. *Nature Physics,* 1–6 (2023).

117. Cavagna, A. *et al.* Dynamical renormalization group approach to the collective behavior of swarms. *Physical Review Letters* **123,** 268001 (2019).

118. Della Morte, M., Orlando, D. & Sannino, F. Renormalization group approach to pandemics: The COVID-19 case. *Frontiers in physics* **8,** 144 (2020).

119. Benoist, F. *Complex Materials: Frustrated Self-Assembly and Nonlinear Elasticity* PhD thesis (Universite Paris-Saclay, 2022).

120. Hestenes, M. R., Stiefel, E., *et al.* Methods of conjugate gradients for solving linear systems. *Journal of research of the National Bureau of Standards* **49,** 409–436 (1952).

121. Curry, H. B. The method of steepest descent for non-linear minimization problems. *Quarterly of Applied Mathematics* **2,** 258–261 (1944).

122. Callen, H. B. & Welton, T. A. Irreversibility and generalized noise. *Physical Review* **83,** 34 (1951).

123. Tieleman, T., Hinton, G., *et al.* Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* **4,** 26–31 (2012).

124. Hall, D. M., Stevens, M. J. & Grason, G. M. Building blocks of non-Euclidean ribbons: size-controlled self-assembly via discrete frustrated particles. *Soft Matter* **19,** 858–881 (2023).

125. Chaikin, P. M., Lubensky, T. C. & Witten, T. A. *Principles of condensed matter physics* (Cambridge university press Cambridge, 1995).

126. Wickham, S. F. *et al.* Complex multicomponent patterns rendered on a 3D DNA-barrel pegboard. *Nature Communications* **11,** 5768 (2020).

127. Guinier, A. *X-ray diffraction in crystals, imperfect crystals, and amorphous bodies* (Courier Corporation, 1994).

128. Guinier, A., Fournet, G. & Yudowitch, K. L. Small-angle scattering of X-rays (1955).

129. Bragg, W. H. & Bragg, W. L. The reflection of X-rays by crystals. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **88,** 428–438 (1913).

130. Svergun, D. I. & Koch, M. H. Small-angle scattering studies of biological macromolecules in solution. *Reports on Progress in Physics* **66,** 1735 (2003).

131. Gommes, C. J., Jaksch, S. & Frielinghaus, H. Small-angle scattering for beginners. *Journal of applied crystallography* **54,** 1832–1843 (2021).

132. Glatter, O., Kratky, O. & Kratky, H. *Small angle X-ray scattering* (Academic press, 1982).

133. Larsen, A. H., Pedersen, J. S. & Arleth, L. Assessment of structure factors for analysis of small-angle scattering data from desired or undesired aggregates. *Journal of Applied Crystallography* **53,** 991–1005 (2020).

134. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28,** 235–242 (2000).

135. RCSB. *PDB Data Distribution by Experimental Method and Molecular Type* https://www.rcsb.org/stats/summary (2023).

136. Jacquamet, L. *et al.* Automated analysis of vapor diffusion crystallization drops with an X-ray beam. *Structure* **12,** 1219–1225 (2004).

137. Cipriani, F. *et al.* CrystalDirect: a new method for automated crystal harvesting based on laser-induced photoablation of thin films. *Acta Crystallographica Section D: Biological Crystallography* **68,** 1393–1399 (2012).

138. ESRF. *ESRF Data Policy* https://www.esrf.fr/datapolicy (2023).

139. Billoir, M. *Characterizing the dimension of protein aggregates through simulated X-ray scattering data* (2021).

140. Svergun, D., Barberato, C. & Koch, M. H. CRYSOL–a program to evaluate X-ray solution scattering of biological macromolecules from atomic coordinates. *Journal of applied crystallography* **28,** 768–773 (1995).

141. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science* **27,** 14–25 (2018).

142. Garic, M. *Statistical identification of aggregates in protein crystallography datasets* (2022).

143. Fujiwara, I., Zweifel, M. E., Courtemanche, N. & Pollard, T. D. Latrunculin A accelerates actin filament depolymerization in addition to sequestering actin monomers. *Current Biology* **28,** 3183–3192 (2018).

144. Yamashiro, S., Yamakita, Y., Ono, S. & Matsumura, F. Fascin, an actin-bundling protein, induces membrane protrusions and increases cell motility of epithelial cells. *Molecular biology of the cell* **9,** 993–1006 (1998).

145. Guerrero-Ferreira, R. *et al.* Two new polymorphic structures of human full-length alpha-synuclein fibrils solved by cryo-electron microscopy. *Elife* **8,** e48907 (2019).

146. Bousset, L. *et al.* Structural and functional characterization of two alpha-synuclein strains. *Nature communications* **4,** 2575 (2013).

147. Parness, J. & Horwitz, S. B. Taxol binds to polymerized tubulin in vitro. *The Journal of cell biology* **91,** 479–487 (1981).

148. Vasquez, R. J., Howell, B., Yvon, A., Wadsworth, P. & Cassimeris, L. Nanomolar concentrations of nocodazole alter microtubule dynamic instability in vivo and in vitro. *Molecular biology of the cell* **8,** 973–985 (1997).

149. Jeangerard, D. *et al. An Any Format Screener for in situ X-ray Diffraction Experiments on PROXIMA 2A* 2021. Poster presentation.

150. Kieffer, J. & Karkoulis, D. *PyFAI, a versatile library for azimuthal regrouping* in *Journal of Physics: Conference Series* **425** (2013), 202012.

151. Jeffries, C. M. *et al.* Preparing monodisperse macromolecular samples for successful biological small-angle X-ray and neutron-scattering experiments. *Nature protocols* **11,** 2122–2153 (2016).

152. Dessau, M. A. & Modis, Y. Protein crystallization for X-ray crystallography. *JoVE (Journal of Visualized Experiments),* e2285 (2011).

153. Berry, M. W., Browne, M., Langville, A. N., Pauca, V. P. & Plemmons, R. J. Algorithms and applications for approximate nonnegative matrix factorization. *Computational statistics & data analysis* **52,** 155–173 (2007).

154. Howley, T., Madden, M. G., O'Connell, M.-L. & Ryder, A. G. *The effect of principal component analysis on machine learning accuracy with high dimensional spectral data* in *Applications and Innovations in Intelligent Systems XIII: Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, December 2005* (2006), 209–222.

155. Lenz, M. & Witten, T. A. Geometrical frustration yields fibre formation in self-assembly. *Nature physics* **13,** 1100–1104 (2017).