New Methods for the Integrative Dynamic Modeling of Biomolecular Structures

Présentée le 11 juin 2021

Faculté des sciences de la vie
Unité du Prof. Dal Peraro
Programme doctoral en biologie computationnelle et quantitative

pour l’obtention du grade de Docteur ès Sciences

par

Sylvain TRÄGER

Acceptée sur proposition du jury

Prof. P. De Los Rios, président du jury
Prof. M. Dal Peraro, directeur de thèse
Prof. F. Gervasio, rapporteur
Prof. S. Vanni, rapporteur
Prof. B. Correia, rapporteur
Acknowledgements

A PhD thesis is far from being a small endeavour. It is an opportunity to grow as a scientist and researcher but also as a person. I believe that the people surrounding us indirectly yet significantly contribute to the quality of our work, and for the most part, I consider myself lucky in that regard.

For the freedom to chose my projects, their directions and to explore almost any ideas, I want to thank my professor, Matteo Dal Peraro. His trust, enthusiasm and friendliness greatly contributed to the quality of the time I spent in his group. I want to thank all the LBM lab members that I crossed path with during these years, although COVID-19 limited the time spent with the latest arrivals. Particularly, I want to thank Dr. Giorgio Tamò for our enthusiastic discussions inside and outside the office and for being a great friend. I want to thank Dr. Deniz Aydin, Dr. Alessio Prunotto, Dr. Chan Cao and Dr. Giulia Fonti who created an extremely friendly and positive atmosphere. I want to thank Lucien Krapp for being an incredible office mate and friend through helpful and enthusiastic discussions. I want to thank Dr. Luciano Abriata, whom I consider to be the second PI of the lab, for his contagious scientific passion, mentoring, significant help throughout my thesis and great sense of humour. Not thanked enough for their help and support, I want to thank our secretaries Julia Prébandier and Magali Masson. I would like to thank all the different collaborators I had the chance to work with: Prof. Maia Topf, Dr. Sony Malhotra, Prof. Wah Chiu, Dr. Yanyan Zhao and Nandan Haloi. I also want to thank Melivoia Rapti for showing interest in my work and being my best student as well as all the others I mentored in a way or the other during my PhD. I want to thank Romain, Colas and Thomas for their humour, positivity and for the countless opportunities to relieve stressful moments around coffees and beers.
Acknowledgements

Plus que les personnes que j’ai rencontrées pendant mon doctorat, il en existe beaucoup d’autres que je tiens à remercier.

Tout d’abord, je souhaite remercier Marc Moser et Thiago Biolley pour être la crème de la crème en terme d’amis. À travers leur humour sans fond, leur présence, soutien, patience et compréhension, ils méritent une place d’honneur dans cette section.

In più, tengo a ringraziare la mia fidanzata, Giulia. Mi hai insegnato l’amore, la tua lingua, la vita in due e mi hai mostrato come diventare una persona migliore. Il tuo aiuto, supporto, amore, pazienza e presenza hanno fatto una grandissima differenza per il meglio, anche durante i momenti peggiori. Mi hai dato un’ altra prospettiva della vita. Inoltre, ci tengo a ringraziare Antonia, che attraverso la sua bontà ci ha dato un supporto senza fine. Sei una persona meravigliosa.

Finalement et par dessus tout, je tiens à remercier profondément ma famille. Mon frère Julien, pour être un fantastique grand frère et ami. Ton humour à tomber par terre et ton soutien en tout temps ont fait une incroyable différence tout au long de ma vie. Tous les moments passés ensemble et nos mémorables voyages sont gravés dans ma mémoire. Maman et papa, je vous réserve la place d’honneur finale, pour votre dédication, patience, confiance, compréhension, amour, présence et tout ce que vous avez contribué pour faire de moi la personne que je suis aujourd’hui. Un grand merci à vous deux.

Morges, April 11, 2021

Sylvain Träger
Abstract

The true understanding of most cellular functions is only really achievable through the structural determination of their underlying macromolecular assemblies. However, their size, large number of individual components and metastable states make their elucidation by canonical structural biology techniques a vain effort in many situations. As a result, integrative approaches have been in high demand in the recent years. By combining any available data, both computational and experimental, integrative or hybrid modeling strategies have been able to tackle biomolecular structures otherwise intractable by any singular method alone.

Over the past decade, structural biology has undergone a significant revolution in the form of cryo-electron microscopy (cryo-EM). With more than 12’000 models of biomolecular structures under its name, cryo-EM however still struggles to approach larger, flexible assemblies at atomic resolution. Instead, such assemblies now routinely fall in an intermediate resolution range, opening new avenues for the development of hybrid approaches primarily based on cryo-EM data. However, the large degrees of flexibility of such assemblies imposes the inclusion of dynamics in the modeling process. Unfortunately, this only makes the already challenging task of defining a scoring function even more difficult. In general, current integrative strategies are thus unable to approach flexible assemblies in a reliable manner.

To tackle these issues, two new methods made available to the community are presented in this thesis in addition to their applications to a variety of systems. A new clustering analysis tool, CLoNe, is first introduced. As a significant upgrade to the recent Density Peaks algorithm, we show that CLoNe rivals and outperforms many state-of-the-art algorithms even beyond structural biology. Then, we show how CLoNe is able to extract a variety of information from structural ensembles in general or reduce large ensembles to key components, enabling their successful integration into hybrid modeling approaches. Second, the MaD
software is presented. Taking inspiration from traditional computer vision concepts and methods, MaD bypasses the need of a traditional scoring function through the generation of local macromolecular feature descriptors. MaD takes full advantage of the ongoing cryo-EM resolution revolution by integrating local structural information from both cryo-EM data and existing atomic structures. Specifically, MaD is able to predict the quaternary structure of large assemblies regardless of symmetry and conformational variability.

Finally, the MaD-CLoNe combination enabled the modeling of molecular chaperones with unprecedented ease. Chaperonins are of crucial importance to ensure proper protein folding and proteostasis. Perturbation of these processes are implicated in neurodegenerative diseases and cancer. In the specific case of the group II chaperonin Mm-Cpn, the use of MaD-CLoNe along with molecular dynamics simulations uncovered new structural models and insight into the chaperonin's functional pathway.

**Keywords:** integrative modeling, protein flexibility, molecular dynamics, cryo-EM, macromolecular assemblies, flexible assemblies, molecular chaperones, cluster analysis, density peaks, computer vision, feature descriptors, SIFT, MDFF, CLoNe, MaD
Résumé

Une véritable compréhension des fonctions cellulaires majeures est uniquement atteignable à travers l’étude des structures des complexes macromoléculaire sous-jacents. Cependant, leur taille, grand nombre de composants individuels et leurs états métastables souvent nombreux rendent l’obtention de leurs structures par les techniques canoniques de biologie structurale particulièrement difficile. Par conséquent, les approches dites intégratives sont en grande demande. En combinant toute information disponible, provenant tant de sources computationnelles ou expérimentales, ces stratégies intégratives ou hybrides ont été capables d’étudier des structures biomoléculaires autrement trop complexe pour être élucidée par une méthode unique.

Au fil de la dernière décennie, la biologie structurale a bénéficié d’une révolution significative en la forme de la microscopie cryo-électronique (cryo-EM). Avec plus de 12’000 modèles de structures biomoléculaires à son actif, la cryo-EM cependant rencontre des problèmes à obtenir des modèles atomiques d’assemblées macromoléculaires plus grandes et flexibles. Plutôt, ces assemblées sont limitées à des résolutions intermédiaires, ouvrant alors de nouvelles opportunités pour le développement d’approches intégratives focalisées sur les données de cryo-EM. Cependant, le caractère flexible de ces complexes macromoléculaires impose l’inclusion d’une forme de dynamique dans le processus de modélisation. Malheureusement, ceci ne fait que compliquer la définition de fonction d’évaluation de modèle intermédiaires et finaux, une tâche historiquement ardue. En général donc, les méthodes intégratives actuelles peinent à étudier les assemblées flexibles de manière régulière et uniforme.

Afin de remédier à ces problèmes, deux nouvelles méthodes en libre-accès à la communauté sont présentées dans cette thèse en plus de leur application à des cas clés. Un nouvel outil pour l’analyse de regroupements, CLoNe, est d’abord introduit. Constituant une mise à
Résumé

jour significative du récent algorithme basé sur l’analyse de pics de densité, il est première-
ment démontré que CLoNe rivalise et surpasse de nombreux algorithmes actuels et populaires
mêmes en dehors de la biologie structurale. Puis, il est démontré que CLoNe est capable
d’extraire une variété d’information d’ensembles de structures biomoléculaires en général
ou de réduire d’autres grands ensembles en un set de constituants clés, permettant alors
leur intégration dans des méthodes hybrides de modélisation. Ensuite, le programme MaD
est présenté. En s’inspirant de concepts et méthodes typiques de vision computationnelle,
MaD échappe au besoin d’une fonction d’évaluation traditionnelle à travers la génération de
descripteurs de caractéristiques macromoléculaires locales. Notamment, MaD est capable
de prédire la structure quaternaire de nombreuses assemblées quels que soient leur degré de
variation conformationnelle ou leur symétrie.

Par ailleurs, la combinaison MaD-CLoNe ont permis la modélisation de chaperonnes
moléculaires avec une facilité sans précédent. Ces chaperonnes sont d’une importance ca-
pitale pour assurer le repliement des protéines et du maintien de la protéostase dans les
cellules. Des perturbations de ces phénomènes sont connus pour mener à des pathologies
neurodégénératives ou au cancer. Dans le cas précis de la chaperonine de groupe II Mm-Cpn,
MaD-CLoNe combiné avec des simulations de dynamique moléculaire classiques a permis la
découverte de nouveaux aspects et modèles structuraux de son mode de fonctionnement lors
du repliement de protéines.

**Mots-clés**: modélisation intégrative, flexibilité des protéines, dynamique moléculaire, cryo-
EM, assemblées macromoléculaires, assemblées flexibles, chaperonne moléculaire, analyse
de regroupements, pics de densité, vision computationnelle, descripteurs de caractéristiques,
SIFT, MDFF, CLoNe, MaD
## Contents

**Acknowledgements** i

**Abstract (English/Français)** iii

**List of Figures** xi

**List of Tables** xv

1 Introduction 1
   1.1 Paradigm shifts in structural biology 2
       1.1.1 High-resolution techniques 2
       1.1.2 Cryogenic revolution 3
       1.1.3 Into the era of (learned) dynamics 5
   1.2 Integrative modeling: strategies to enhance structural and dynamic resolution 6
   1.3 Objectives of the thesis 10
       1.3.1 Exploring structural landscapes efficiently with clustering 10
       1.3.2 Integrative modeling of macromolecular assemblies with computer vision 11
       1.3.3 Dynamic integrative modeling of the molecular chaperone Mm-Cpn 12

2 Methods 13
   2.1 Computational tools for cryo-electron microscopy 13
       2.1.1 Nature of cryo-EM density maps 13
       2.1.2 Simulation and comparison of electron density maps 15
       2.1.3 Rigid docking of macromolecular structures in cryo-EM density maps 17
   2.2 Molecular Dynamics 17
       2.2.1 Birth and basis of molecular dynamics 17
       2.2.2 General steps to setup a molecular dynamics simulation 23
Contents

2.2.3 Biased molecular dynamics and Molecular dynamics flexible fitting . . . 25
2.3 Cluster analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 27
  2.3.1 Center- and centroid-based clustering . . . . . . . . . . . . . . . . . . . . 28
  2.3.2 Hierarchical clustering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 28
  2.3.3 Density-based clustering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 30
  2.3.4 Density peaks algorithm . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 31
2.4 Computer vision: local feature descriptors . . . . . . . . . . . . . . . . . . . . . 33
  2.4.1 Motivation and relevance for thesis . . . . . . . . . . . . . . . . . . . . . . . 33
  2.4.2 Feature detectors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
  2.4.3 Feature descriptors . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 41

3 CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles 47
  3.1 Background . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 48
  3.2 Method development . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 50
    3.2.1 Nearest-neighbor-based density estimator . . . . . . . . . . . . . . . . . . . 50
    3.2.2 Automatic cluster center determination . . . . . . . . . . . . . . . . . . . . 50
    3.2.3 Merging subclusters . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52
    3.2.4 Outlier detection and removal . . . . . . . . . . . . . . . . . . . . . . . . . . . 53
  3.3 Application to toy datasets . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 55
  3.4 Application to structural ensembles . . . . . . . . . . . . . . . . . . . . . . . . . . 59
    3.4.1 Membrane binding of Coenzyme Q9 . . . . . . . . . . . . . . . . . . . . . . 59
    3.4.2 Allosteric pocket of TEM1 β-lactamase . . . . . . . . . . . . . . . . . . . . 60
    3.4.3 Oligomerisation of Amyloid Precursor Protein . . . . . . . . . . . . . . . . 61
  3.5 Case study: structural characterization of KAP1 . . . . . . . . . . . . . . . . . . . 64
  3.6 Comparison with other clustering algorithms . . . . . . . . . . . . . . . . . . . . 67
  3.7 Implementation and functionalities for structural ensembles . . . . . . . . . . 70
  3.8 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 70

4 MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies 75
  4.1 Solving biomolecular puzzles with feature descriptors . . . . . . . . . . . . . . . 75
  4.2 Anchor point detection in 3D biomolecular structures . . . . . . . . . . . . . . . 78
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.3</td>
<td>Orientation assignment</td>
<td>85</td>
</tr>
<tr>
<td>4.3.1</td>
<td>The sphere division problem</td>
<td>85</td>
</tr>
<tr>
<td>4.3.2</td>
<td>Using EQ spheres to orient anchors</td>
<td>88</td>
</tr>
<tr>
<td>4.3.3</td>
<td>Localizing an atomic structure within an electronic density map</td>
<td>90</td>
</tr>
<tr>
<td>4.3.4</td>
<td>Benchmark</td>
<td>90</td>
</tr>
<tr>
<td>4.4</td>
<td>Generating robust and discriminative descriptors</td>
<td>97</td>
</tr>
<tr>
<td>4.4.1</td>
<td>Benchmark</td>
<td>99</td>
</tr>
<tr>
<td>4.5</td>
<td>Descriptor matching, scoring and post-processing</td>
<td>100</td>
</tr>
<tr>
<td>4.6</td>
<td>Docking and refining macromolecular components in EM maps</td>
<td>103</td>
</tr>
<tr>
<td>4.7</td>
<td>Building assemblies from individually fitted components</td>
<td>109</td>
</tr>
<tr>
<td>4.8</td>
<td>Dealing with dynamics and large conformational changes</td>
<td>110</td>
</tr>
<tr>
<td>4.9</td>
<td>Conclusions</td>
<td>114</td>
</tr>
<tr>
<td>5</td>
<td>Modeling the group II chaperonin Mm-Cpn with MaD-CLoNe</td>
<td>119</td>
</tr>
<tr>
<td>5.1</td>
<td>Background</td>
<td>119</td>
</tr>
<tr>
<td>5.2</td>
<td>Molecular dynamics and docking with MaD-CLoNe</td>
<td>123</td>
</tr>
<tr>
<td>5.3</td>
<td>Modeling the mechanism of chamber closing of Mm-Cpn</td>
<td>124</td>
</tr>
<tr>
<td>5.4</td>
<td>Further insight from large-scale molecular dynamics</td>
<td>130</td>
</tr>
<tr>
<td>5.5</td>
<td>Conclusion</td>
<td>134</td>
</tr>
<tr>
<td>6</td>
<td>Conclusion and perspectives</td>
<td>137</td>
</tr>
<tr>
<td>Appendix</td>
<td></td>
<td>143</td>
</tr>
<tr>
<td>A</td>
<td>3D Hessian detector</td>
<td>143</td>
</tr>
<tr>
<td>B</td>
<td>3D Harris corner detector</td>
<td>146</td>
</tr>
<tr>
<td>References</td>
<td></td>
<td>149</td>
</tr>
</tbody>
</table>
## List of Figures

1.1 Number of solved structures and maps by canonical structural biology techniques 4
1.2 Determining structures of macromolecular complexes using cryo-EM/ET 7

2.1 Contour level of density maps 14
2.2 Simulation of density maps from atomic structures 16
2.3 Bonded and non-bonded terms of molecular force fields 20
2.4 Periodic boundary conditions 23
2.5 Different categories of clustering 29
2.6 Example of corner and blob features 34
2.7 Gaussian scale-space of an image 35
2.8 Difference of Gaussian and Laplacian of Gaussian applied to a signal at different scales 36
2.9 Gaussian partial derivatives, determinant of Hessian and box filters 39
2.10 Edge and corner detection with Harris and Shi-Tomasi 40
2.11 Orientation and description of a feature point 44
2.12 RANSAC 45
2.13 Example of SIFT 46

3.1 Clustering based on local density neighborhoods 51
3.2 General applicability of CLoNe on several datasets 55
3.3 Comparison between CLoNe and other clustering algorithms 57
3.4 Large range of acceptable values for CLoNe’s unique input parameter 58
3.5 Cluster centers as key biological states of the COQ9 membrane protein 60
3.6 Identification of different opening states of the allosteric cryptic pocket in TEM1 \( \beta \)-lactamase 62
## List of Figures

3.7 Classification and frequency estimation of APP dimerization motifs at the membrane .................................................. 63
3.8 Structural characterisation of KAP1 ........................................ 66
3.9 Comparison between CLoNe and other algorithms on all structural datasets .................................................. 69
3.10 Basic usage of CLoNe .......................................................... 71
4.1 MaD: Macromolecular Descriptors for density-guided quaternary structure prediction .................................................. 77
4.2 Systems used to test anchor detection methods ........................................ 79
4.3 Anchor detection with Laplacian of Gaussian .................................. 80
4.4 Multi-scale detection with Laplacian of Gaussian .................................. 83
4.5 Single-scale detection with Laplacian of Gaussian ............................... 84
4.6 Platonic solids ................................................................. 87
4.7 Zonal equal area partitioning of the sphere ....................................... 88
4.8 Orientation assignment and orientation-free referential of structures .......... 89
4.9 Estimating the position of a structure within a density map using a single pair of anchors .................................................. 91
4.10 Benchmark assemblies and relevant parameters for orientation assignment ....................................... 92
4.11 Bias from root-mean square deviation .......................................... 94
4.12 Benchmark results for orientation assignment .................................... 96
4.13 Benchmark results for orientation assignment on whole assemblies .......... 97
4.14 Descriptor generation .......................................................... 98
4.15 Benchmark for descriptor generation ............................................ 99
4.16 Local and global descriptor matching .......................................... 101
4.17 Local rigid refinement integrated in MaD ....................................... 104
4.18 Piece-by-piece docking with MaD .............................................. 106
4.19 Whole-structure docking with MaD ............................................. 107
4.20 Borderline cases in the simulated benchmark ................................... 111
4.21 Assembling GroEL with MaD .................................................. 112
4.22 Flexible modeling of GroEL with MaD ......................................... 113
5.1 Architecture and known structures of the Cpn complex from *Methanococcus maripaludis* .................................................. 120
5.2  Mm-Cpn subunit in 20 different states determined by cryo-EM . . . . . . . . . 122
5.3  Efficient sampling the dynamics of an Mm-Cpn subunit . . . . . . . . . . . . . 123
5.4  Docking of Mm-Cpn subunit ensemble with MaD-CLoNe . . . . . . . . . . . 125
5.5  Flexible refinement of docked Mm-Cpn subunits with MDFF . . . . . . . . . . . 126
5.6  Modeling the closing mechanism from subunit models . . . . . . . . . . . . . 127
5.7  Modeling the closing of the folding chamber of Mm-Cpn . . . . . . . . . . . . 129
5.8  Combined molecular dynamics of individual subunit models . . . . . . . . . . 131
5.9  Individual conformational sampling of subunit trajectories . . . . . . . . . . . 133

A.1  Single-scale detection with the determinant of the Hessian matrix in 3D . . . 145
B.1  Single-scale detection with Shi-Tomasi corner detector in 3D . . . . . . . . . . 147
B.2  Anchors detected by the Shi-Tomasi detector . . . . . . . . . . . . . . . . . . . . 148
List of Tables

3.1 CLoNe applied to high-dimensionality datasets ............................................. 56
3.2 Statistics for clusters and centers from structural ensembles and related features
obtained with CLoNe ......................................................................................... 68
4.1 Summary of MaD applied to a set of experimental cryo-EM maps. ................. 108
4.2 Summary of MaD applied to a set of simulated cryo-EM maps. ..................... 110
Large macromolecular complexes are at the core of many cellular functions. They are mostly made up of multiple proteinic components and intermittently include or interact with small molecules, lipids, nucleic acids and carbohydrates. They are involved in a wide variety of contexts, ranging from DNA remodelling and replication, gene expression, internal and external transport of nutrients, protein synthesis, folding and degradation, signal transmission and recognition to structural support and many others. The proteasome, nuclear pore complex, ATP synthase, ribosome, GroEL and other chaperones are all popular examples of these complex machineries and are comprised of dozens of individual components. Their elucidation, while daunting from both structural and dynamic perspectives and preferably *in situ*, constitutes some of the key achievements of structural biology to date and continuously provide a deeper understanding of their functional cycles.

Due to their involvement in these key processes, the structural disruption of these complexes often leads to diseases, or to curing them when they originate from infectious organisms. Thus, beyond fundamental knowledge, deciphering these assemblies and their mode of oligomerization at the highest resolution possible also provides opportunities for designing novel therapeutics, and often structure-based drug discovery has proved its worth in that regard. With the breakthroughs in cryogenic electronic microscopy, deep learning and the speed at which SARS-CoV-2-related protein structures have been solved, it is abundantly clear that structural biology has a well-deserved spot on the front line of scientific research.
### 1.1 Paradigm shifts in structural biology

Over the last century and half, we have acquired deeply complex methodologies and technologies to observe biological structures at the molecular level, shifting interests from simply solving structures to solving structures *in situ* and as part of interconnected, dynamic functional pathways. As testimony to these breakthroughs, several Nobel prizes were awarded over the last 60 years for the development of these techniques.

#### 1.1.1 High-resolution techniques

The molecular structures of haemoglobin\(^1\) and myoglobin\(^2\) through X-ray crystallography during the 1950s marked the birth of structural biology. Solving these structures lead to their respective authors, Perutz and Kendrew, receiving the Nobel prize of chemistry in 1962. At the time of writing, there are now more than 150’000 structures solved by X-ray crystallography in the Protein Data Bank (PDB\(^3\)) ([Figure 1.1a](#fig1.1a)). This method is however limited by the generation of diffractable and sizeable crystals. This process is often facilitated by chemically modifying proteins, which may however steer the result further from its native context on top of the crystallization process and packing. This method relies on complex sample preparation procedures and may fail for the larger, more flexible assemblies.

In 1985, Wüthrich laid out the key principles for the application of nuclear magnetic resonance to biomolecules in solution, for which he was awarded the Noble prize of chemistry in 2002. NMR enabled the study of protein structures in a medium closer to their physiological...
1.1. Paradigm shifts in structural biology

context and succeeded where other methods have failed. A key difference with X-ray crystallography is that NMR is applied to macromolecules in solution rather than crystallised, but is severely limited in size due to resonance spectra quickly overlapping as molecular weight increases. NMR however yields more than a structure, as it enables to check if a protein is folded, study internal dynamics, or measure water accessibility among others. Of note, its ability to detect interactions between protein and ligands makes it a common partner to X-ray crystallography. There currently exists close to 15,000 NMR ensembles in the Protein Data Bank (Figure 1.1b).

1.1.2 Cryogenic revolution

In 1981, Alasdair McDowall and Jacques Dubochet developed methods for the vitrification of pure water, paving the way to ground-breaking improvements to cryogenic electron microscopy (cryo-EM). The cryoprotection provided by vitrification enabled to increase exposure to electron beams while reducing sample damage due to radiation, resulting in higher signal-to-noise ratios. Jacques Dubochet, Joachim Frank and Richard Henderson were awarded the Nobel prize in Chemistry in 2017 for the development of the technique. Combined with the introduction of direct electron detectors, increasing the resolution of acquired micrographs, the Volta phase plate, filtering out non-elastic electrons, and new image processing techniques to get more from the same amount of images, cryo-EM is in the midst a resolution revolution. It is an attractive method compared to X-ray crystallography in that it is less stringent on sample purity and quantity. There are currently more than 12,000 maps obtained by either SPA or STA (Figure 1.1c) in the Electron Microscopy Data Bank (EMDB5,6), triple the amount of maps in 2016 when I started my thesis.

In cryo-EM, thousands of single particle images are picked from acquired micrographs. By sorting, averaging and classifying these images, it is possible to reconstruct 3D models of proteins through a process called single particle analysis (SPA). Electron cryotomography (cryo-ET) distinguishes itself from cryo-EM by observing samples in their native biological context rather than relying on vitrified samples in vacuum. Vitrified cells are milled, and thin cryo-lamella are tilted during acquisition to obtain a tilt-series. These are computationally reconstructed into 3D tomograms of the sample in its native context. Particles of interest may be isolated into sub-tomograms, which may be averaged in order to yield higher-resolution
Introduction

Figure 1.1 – Number of solved structures and maps by canonical structural biology techniques. Data for X-ray crystallography and NMR has been taken from RCSB.org. Data for single particle analysis and sub-tomogram averaging has been taken from the EMDB database, excluding cellular components in the case of sub-tomogram averaging. Bars in dark gray correspond to new structures and maps deposited in their respective database since the start of my thesis. (a) Total structures solved by X-ray crystallography since 2000. (b) Total structures solved by nuclear magnetic resonance (NMR) since 2000. (c) Total density maps, regardless of resolution, reconstructed using either single particle analysis (SPA) or sub-tomogram averaging (STA) over the years. (d) Evolution of the highest and average resolutions achieved every year by both SPA and STA. (e-f) Distribution of resolutions for SPA and STA, respectively. The inner circle highlights the growth of the technique over the course of this thesis.

3D reconstructions through sub-tomogram averaging (STA). When enough single particles may be retrieved from tomograms, it is possible to apply single particle analysis instead. While SPA remained a low-resolution technique for many years, cryo-EM is now rivalling X-ray crystallography in terms of potential, with the highest and average resolution of density maps becoming lower every year (Figure 1.1d). Starting from resolutions of 4 Å, it is possible to build atomic models directly within the density. Most recent cryo-EM maps are below 5 Å, but maps at intermediate resolutions (5-10 Å) are still common (Figure 1.1e). Of note, the atomic barrier has recently been crossed with maps at 1.2 Å, allowing for the direct visualization of atomic positions. High resolutions have been achieved with STA as well, but the technique remains mostly at an intermediate to low resolution range (Figure 1.1f), certainly due to the reduced amount of particles in tomograms and hardware differences with cryo-EM and SPA. Last, map resolution is, traditionally and now practically, reported as a single number and its
estimation has been subject of debate for many years\textsuperscript{9}. In reality, resolution varies throughout maps, and regions with lower local resolution are typically more flexible, as is often the case near the structure's surface. It is now common to complement the reported resolution with local resolution maps, for which several tools are available\textsuperscript{10,11,12}.

1.1.3 Into the era of (learned) dynamics

There are currently over 170,000 atomic structures at the time of writing - only 125,000 of which were available at the start of my thesis, which in itself is proof of how far structural biology has gone. More than singular structures, it has become abundantly clear that most proteins exist as a continuum of states, often with corresponding free-energy landscapes as complex as their functional pathways. For a long time, it was common to think that one protein had a single structure, a rigid mindset that probably stemmed from X-ray crystallography, where the single output structure corresponds to the energy minima within crystal packing\textsuperscript{13}. Different conformations were often limited to ligand-bound and ligand-unbound structures, as is the case in the Protein-Protein benchmark\textsuperscript{14}.

Sample heterogeneity has been the target of large efforts in the cryo-EM community, as the presence of multiple states within the same sample may impede both reconstruction and resolution. Through software improvements, this issue has now been turned into a means of extracting insights into a structure's internal dynamics while weeding out inconsistent particles and improving overall resolution\textsuperscript{12}. Some methods focus on multi-state reconstruction\textsuperscript{15}, while others might focus on structures whose energy landscapes cannot be easily discretized into clear-cut energy minima\textsuperscript{13,16}, such as the ribosome\textsuperscript{17} or more recently RyR1\textsuperscript{18}. This shows that the success of SPA and STA is not only due to experimental advances, but to computational ones as well, and that dynamic information is no longer exclusive to purely computational methods.

Molecular dynamics has grown tremendously since the first simulation in 1977 and has proven to be a powerful microscope into a protein's dynamics, now able to simulate molecular events occurring at unprecedented scales (see Chapter 3 for examples). Machine learning is rapidly taking hold of scientific areas and structural biology is no exception. Deeply learned methods allow researchers to bypass solving complex computations in favor of letting clusters...
of GPUs learn approximations from large amounts of data instead. Molecular simulations are in the midst of many learned upgrades and alternatives\textsuperscript{19,20}. To top it all off, the 50-year problem of protein folding has been practically solved through deep learning with unprecedented accuracy\textsuperscript{21}. Cryo-EM revolutionised experimental structural biology, sprouting atomic structures of many proteins previously out of reach from other techniques. Machine learning may be inbound to provide yet another gigantic breakthrough in the field, this time on the computational side.

1.2 Integrative modeling: strategies to enhance structural and dynamic resolution

Despite many claims that cryo-EM is now a high-resolution technique, it often requires the input of external methods and data in order to reach atomic models regardless of resolution. For the most part, cryo-EM thus still relies on hybrid, or integrative approaches to get the most out of its data. This remains true even with resolutions below 4 Å, as substantial portions of the map may be well above this value due to local variations\textsuperscript{12}. The workflow to adopt facing the modeling structures from density maps ultimately depends on their resolution. In collaboration with Dr. Sony Malhotra and Prof. Maya Topf, we recently wrote a review on the subject\textsuperscript{12} (Figure 1.2). When resolution is below 4 Å, methods from X-ray crystallography such as COOT\textsuperscript{22} or PHENIX\textsuperscript{23} may be used successfully to build the model \textit{de novo} within appropriate parts of the density. Above this threshold, the map itself lacks the necessary information for such direct model building as sidechain densities may not be entirely or at all resolved. This lower resolution range is typically split between resolutions below and above 10 Å, threshold where secondary structure starts to blend into blob-like structures.

In 2020, 876 maps between 4 and 10 Å have been released on EMDB using either SPA or STA, 468 of which have a corresponding fitted PDB structure. In contrast, only 103 maps between 10 and 20 Å were released in the same time frame, of which 22 have a corresponding PDB structure. The intermediate range of resolutions up to 10 Å is thus of crucial interest and typically involves large, multi-component macromolecular assemblies. Recent examples include the GluK3 receptor\textsuperscript{24} at 7.4 Å, four states of the MoxR AAA+ ATPase RavA\textsuperscript{25} between 6 and 7.8 Å by SPA, an \textit{in situ} structure of the SARS-CoV-2 spike protein in an open state\textsuperscript{26} at 5.4 Å by STA, CtBP2 by SPA at 5 Å\textsuperscript{27}, 6 states of the E. coli ATP synthase\textsuperscript{28} between 5 and 7.22 Å by...
1.2. Integrative modeling: strategies to enhance structural and dynamic resolution

Figure 1.2 – Determining structures of macromolecular complexes using cryo-EM/ET. Taken from our review in collaboration with Dr. Sony Malhotra and Prof. Maya Topf. The structure of the macromolecular assembly can be determined using cryo-EM microscopy or tomography (top-left box). Single particle analysis can ensue from both techniques, while subtomogram averaging is specific to latter. When the dataset is heterogeneous, the next step is to extract homogeneous subsets of single particles. The analysis of this heterogeneity can lead to the 3D reconstruction of multiple states or, when these cannot be discretised, a state continuum which may or may not be based on existing models. The atomic structure of the individual components or subcomplexes (top-right box) may have been previously solved experimentally or need to be modelled (with or without a template). Conformational ensembles of these components can be generated using a variety of techniques or readily be available in the case of an NMR structure. If available, spatial restraints (top-center box) in the form of inter-subunit or intra-subunit connectivity information can be added. The EM maps, the atomic component structures and any restraints are then integrated using a workflow that is generally dictated by the resolution of the EM map (bottom box), yielding structural models of macromolecular complexes. Acronyms used in the figure: EM, electron microscopy; ET, electron tomography; XRC, X-ray crystallography; NMR, nuclear magnetic resonance; MD, molecular dynamics; NMA, normal mode analysis and MC, Monte Carlo.
Introduction

SPA, to name a few from the last couple of years.

The integrative approach to undertake will depend mostly on which known related structures or models are available. If an existing model of the assembly is indeed available, then it is first rigidly docked within the map. This can be done using the coarse *Fit in map* tool within Chimera²⁹ or more recently, ChimeraX³⁰. Then, flexible fitting will ensure that the model adopts conformational states similar to those captured by cryo-EM. The latter involves biased deformation of the initial structure into the target density (see section 2.2.3). Flexible fitting approaches are numerous, with the most common being Molecular Dynamics Flexible Fitting (MDFF³¹, see also section 2.2.3) applied recently to the PAN-proteasome³². Rather than molecular dynamics, another class of flexible refinement protocols rely on normal mode analysis³³,³⁴,³⁵. Common to all these methods, an existed model pre-fitted within the map is a strict requirement as they typically rely on local density gradients in some way. Alternatively, a method within the Gorgon software utilizes α-helical correspondences between atomic structures and target density map, claiming to achieve docking and flexible fitting simultaneously after skeletizing the target density map³⁶. However, this method has yet to see any concrete applications beyond the data in their article. The most accurate method to sample dynamics while preserving the system's physico-chemical features is classical molecular dynamics simulations, which our group has utilized in combination with genetic algorithms³⁷,³⁸,³⁹.

In the cases where no pre-existing model exists, a different path must be taken. X-ray crystallography, NMR or comparative modelling structures are often available in the PDB³. The task essentially becomes solving a 3D puzzle, using cryo-EM density maps as guide to assemble structural pieces in their native configuration. The search space for each component is effectively 6-dimensional, including 3 angular and 3 translational coordinates. Methods aiming to fulfil this goal are numerous¹²,³⁹,⁴⁰, and generally aim at creating concise representations of both EM and atomic data and/or reducing the complexity of the search space through a variety of approaches.

The Integrative Modeling Platform (IMP⁴¹,⁴²) may be the first to explicitly tackle this problem, trying to include as much data as possible beyond structures, such as mutagenesis experiments or chemical cross-linking data. To tackle the the extreme dimensionality of the search space, heuristics were introduced. IMP's (CN)MultiFit⁴³ focuses on the simultaneous rigid
1.2. Integrative modeling: strategies to enhance structural and dynamic resolution

fitting of a set of atomic components within a cryo-EM map through an optimizer sampling a
discretized 6-dimensional search space. Recently, IMP was updated with a Bayesian inference
approach involving Gaussian mixture models, which properly weights the contribution of
multiple data sources, experimental and computational, including cryo-EM\textsuperscript{44}, although the
prediction themselves are still rigid. Zernike descriptors, which encode the shape of protein
surfaces, are combined with a genetic algorithm to predict the topology of multimeric protein
complexes but are by definition sensitive to large conformational changes\textsuperscript{45}.

Another set of strategies rely on a vector quantization approach\textsuperscript{46}, turning the problem of
docking high-resolution structures in low-resolution maps into a point-matching problem\textsuperscript{47,48}.\textcolor{red}{Going further, $\gamma$-TEMPy introduced the use of a genetic algorithm combined with a feature-
based approach for better initialization\textsuperscript{49}. This approach is purely rigid as well, relying on
Flex-EM\textsuperscript{50} for flexible refinement \textit{a posteriori}.

SITUS instead uses FFT acceleration to compute cross-correlation over the entire 6-
dimensional search space\textsuperscript{51} to individually or simultaneously fit components, again in a
rigid manner. Alternatively, PowerFit\textsuperscript{52} uses a core-weighted cross-correlation-based scoring
function to determine centroids corresponding to probable component locations in an EM
map. This effectively translates EM data to distance restraints, which are then used within
HADDOCK-EM\textsuperscript{53,54} to maximize cross-correlation with an input density map. Semiflexible
simulated annealing and restrained molecular dynamics finalize inter-subunit contacts,
yielding the final predictions. While dynamics are indeed included, they are in the form of
refinements rather than accounting for the component’s conformational landscape during
the fitting itself.

Thus, a significant caveat in all those methods is the absence of consideration of dynamics
during the fitting itself. Cases where conformational differences are too large for an initial
rigid-body fitting tend to be left aside. Indeed, it may be that isolated subunits adopt different
states upon oligomerization, or that the source methods captured the structure at different
steps of a functional pathway. Such cases include the aerolysin pore-forming toxin, the
hexameric HIV capsomer and the type III secretion system from \textit{Yersinia enterocolitica}. To
tackle such cases, our group designed \textit{pow}, relying on a custom Particle Swarm Optimization
(PSO) procedure or more recently, memetic Viability Evolution (mViE)\textsuperscript{55}. The advantage of
Introduction

\textit{pow} is that it relies on using conformations obtained through classical molecular dynamics simulations during the fitting process rather than \textit{a posteriori}. Moreover, such conformational ensembles may be the most biophysically plausible obtainable from computational sources. The type of structures tractable this way is however limited to symmetrical assemblies, where the correct symmetry may simply be applied to the single component during or after the fitting. Indeed, the search space for this problem being 6-dimensional in rigid form, additional degrees of freedom need to be considered in a dynamic context. Thankfully, atoms tend to move in a coordinated fashion within a structure, which allows us to consider only few dynamic coordinates to extend the search space. In the case of \textit{pow}, these coordinates were approximated using principal component analysis, typically adding 2 or 3 dimensions to the original search space. Still, this quickly becomes intractable as the number of unique components in a given assembly increases. An obvious solution would be the coarse-graining of the conformational ensembles. However, both PSO and mViE rely on continuous search space, invalidating this approach for \textit{pow}.

Guilty of code unavailability, difficulty of deployment, usage complexity or poor maintenance over time, the reality is that some of these methods see few uses outside of the group they originate from. Most external groups instead tend to perform manual fitting, sometimes aided by Chimera’s very local rigid-body fitting scheme \textit{Fit In Map}, before resorting to a flexible refinement method of choice, such as MDFF, iMODFIT or Flex-EM to remind some of the most popular ones.

1.3 Objectives of the thesis

1.3.1 Exploring structural landscapes efficiently with clustering

In the midst of the era of dynamics, structural ensembles grow in size and complexity. Tools for their analysis are in high demand for a variety of applications. The first goal of this thesis is to provide a general, easy to use and maintainable clustering method applicable to any kind of structural ensembles originating from molecular dynamics or structural biology techniques. Such a tool would enable any researcher to extract key information without tedious parametrization and fine-tuning steps. Additionally, the reduction of these large conformational ensembles to smaller sets of structures may enable the integration of key
1.3. Objectives of the thesis

dynamic information within hybrid methods beyond the work that has been done in and outside of our group. Clustering methods are plentiful, but most are both limited and sensitive to specific properties of the underlying data. In Chapter 3, a novel clustering package, named CLoNe and developed for application to structural ensembles of various sources will be presented.

1.3.2 Integrative modeling of macromolecular assemblies with computer vision

Large macromolecular assemblies are of crucial importance due to their central role in many intra- and extracellular processes. Due to their complexity, large number of both homogeneous and heterogeneous components and inherent flexibility, canonical structural determination techniques struggle in their elucidation. Hybrid strategies have been the go-to for such tasks for many years. Despite important breakthroughs, notably in regards to cryogenic electron microscopy, current approaches still lack in critical aspects. Namely, some methods are limited with respect to the number of components within the target assembly due to the resulting computational complexity or to symmetrical complexes. Furthermore, a smart inclusion of dynamics and robustness towards large conformational changes is sorely needed, as it often prevents modeling altogether and force researchers to rely on tedious manual data integration. The different origins of data to be integrated often include different biological states, but most current integrative methods are unable to account for dynamics effectively. Combined with the recent breakthrough in cryo-EM, integrative modeling is in need of a novel approach to tackle flexible cases at intermediate resolutions.

Taking a step back from current strategies, the task of assembly prediction using a set of individual components is reminiscent of typical computer vision tasks. Namely, object detection, localization and pose estimation. Relevant questions, such as Is that object in that scene? If yes, where is it located? In which position? in vision tasks may be translated to atomic structures within a target cryo-EM map. Through the use of local feature descriptors, computer vision achieves remarkable robustness to changes in shapes and partial occlusion. Chapter 4 covers the application of similar computer vision principles towards the development of Macromolecular Descriptors (MaD). The basis of the software makes the inclusion of dynamics through CLoNe easy and intuitive. The combination of both software is able to tackle large heteromultimeric assemblies while including dynamics on the sole basis of cryo-EM data,
Introduction

thus overcoming key limitations in the field.

1.3.3 Dynamic integrative modeling of the molecular chaperone Mm-Cpn

Proteins require folding into 3D structures after translation in order to be functionally active. Without specific conditions, aberrant folding may occur, in turn leading to dysfunctions within the cell and ultimately to a variety of conditions, including neurodegenerative diseases or cancer. Molecular chaperones are a class of proteins that provide ideal conditions to ensure proper protein (un)folding and thus, proteostasis.

Among these chaperones is the group II chaperonin Mm-Cpn. This 16-mer assembly is a prime example of extremely dynamic macromolecular complexes requiring a specialized hybrid approach. Latest structural results date from the early 2010s and were limited to few structures of the closed and open states. While the former was solved at high resolution, the latter is really limited to lower resolutions due to the flexibility of all of its components. Any intermediate states, their link to Mm-Cpn-mediated folding and their mode of transition remain to be elucidated at high resolution. In chapter 5 and in collaboration with Prof. Wah Chiu and Dr. Yanyan Zhao from Stanford University, software developed in the previous aims were applied to the modeling of the internal dynamics of Mm-Cpn. Specifically, we generated atomistic models from a set of 20 cryo-EM maps of Mm-Cpn's monomer and, combined with large-scale molecular dynamics, we offer new insight on the mechanism behind the closing of Mm-Cpn's folding chamber.
2 Methods

This chapter will first explore the nature of cryo-EM density maps. Then, existing methods for the simulation of density maps from atomic structures as well as for the manipulation of cryo-EM density maps, which are relevant for Chapters 4 and 5 will be detailed. An overview of molecular dynamics simulation is shown next, which is relevant for all subsequent chapters. A section on cluster analysis will cover state-of-the-art algorithms in general before detailing the clustering by density peaks of Rodriguez and Laio. This is relevant for Chapter 3, where a new iteration of that approach will be presented and applied to a variety of biomolecular structural ensembles. Finally, details of local feature detectors and descriptors as they are known in computer vision are shown in order to approach Chapters 4 and 5 with all the necessary knowledge.

2.1 Computational tools for cryo-electron microscopy

This section covers existing tools to create, manipulate and optimize density data. They are of particular importance for the software described in Chapter 5, which predicts the assembly of high-resolution structures into their native complex as dictated by cryo-EM data. In order to obtain a stand-alone software without external software dependencies, these tools have been fully reimplemented.

2.1.1 Nature of cryo-EM density maps

Density maps are essentially a voxel grid, a voxel being a 3D pixel that refers the detected electron potential at that point in space. The voxel spacing is often referred to as pixel size
Chapter 2. Methods

Figure 2.1 – Contour level of density maps. The structure shown here is that of $\alpha$ and $\beta$ tubulin along with kinesin (PDB: 2P4N; density map: EMD-1340). The first isosurface shows the density data so that all three subunits are visible. The second isosurface corresponds to the threshold recommended by the authors, and the last one shows what remains of the tubulin densities when the kinesin is thresholded out.

In cryo-EM literature, as it depends on the effective magnification, which becomes a trade-off between resolution and the number of particles present in micrographs. Pixel size also depends on the microscope and electron detector pair. As a rule of thumb, resolution is limited to three times the pixel size\(^57\), but super-resolution detectors have pushed that limit further, with maps at resolution close to the pixel size\(^58\).

Header information also contains the origin of the grid, so that the structure can be accurately located in space and with respect to other structures when combined with the voxel spacing. There is no standard in the value range of density data, with some maps exhibiting value ranges in the scale of 0.1, 1 or even higher than 10. They often contain negatives values, which can generally be attributed to noise. A threshold, also called isovalue or contour level, can be set so that the data can be visualized as isosurfaces. Data deposited in the EMDB database is associated with an arbitrary contour level recommended by their authors, but no standard exists. Reasons for this include the disparity of value ranges, often within the same map, where specific domains may be better resolved than others. An example of such a density map is shown in Figure 2.1 along with the fitted atomic structure for reference. That structure contains tubulin-$\alpha$ and tubulin-$\beta$ protomers and kinesin. The recommended contour level disregards most of the density of the kinesin, which is otherwise visible when setting a lower threshold.
2.1.2 Simulation and comparison of electron density maps

In order to compare structures at atomic resolution with low-resolution maps, it may be useful to bring the former to a data structure and resolution similar to the latter. Such simulation of density data will be used extensively in Chapter 4. Simulating density grids from atomic coordinates is also used in flexible fitting schemes\(^{31,59,60,61}\) (see section 2.2.3 as well), but the intent is different and approximations are often made in favor of computing time, for instance by increasing the voxel spacing.

Opposite to electronic density maps, atomic structures only contain a set of atom coordinates, which do not depend on any kind of lattice. Thus, simulating data from atomic coordinates first requires their projection on a grid with the same voxel spacing as the experimental density map, followed by the blurring of the grid down to the target resolution. This procedure is available in similar form in the software Chimera and ChimeraX\(^ {29,30}\) and SITUS\(^ {62}\). The size of the projection grid is determined by taking the positional extrema of the atoms, plus a margin to account for the Gaussian blurring of the next step. The grid is created as to be in register with the origin. The projection itself is carried out by trilinear interpolation and individual atoms are weighted according to their mass. At this point, the grid contains atomic data at a theoretical atomic resolution, hindered by lattice projection artefacts and voxel spacing only. The second step involves a convolution with an appropriately scaled Gaussian kernel to blur the grid down to the target resolution while considering the voxel spacing of the grid. The resulting kernel is truncated to \(3\sigma\) and normalized so that the sum of its elements is 1:

\[
G(x, y, z) = \exp \left( -\frac{x^2 + y^2 + z^2}{2\sigma_F^2} \right) \quad \sigma_F = f(r) \frac{r}{\nu} \quad (2.1)
\]

Where \(r\) and \(\nu\) are the target resolution and voxel spacing, respectively. The normalisation factor \(f(r)\) in the expression of the kernel scale \(\sigma_F\) relates to the Fourier transform of the distribution and to the width of the Gaussian kernel applied to individual atom. Four factors are offered in Chimera\(^ {29}\):

\[
f(r) = \frac{1}{\pi \sqrt{2 \log 2}} \approx 0.187 \quad (2.2)
\]
Figure 2.2 - Simulation of density maps from atomic structures. In both panels, the first column depicts the experimental density map. Subsequent columns correspond to simulated density maps with normalization factors of 0.187, 0.225, 0.356 and 0.425, respectively, as suggested in Chimera’s *molmap* command. Each map is shown at the same contour level. Below each isosurface is shown a volume slice, color-coded from low density (blue) to high density (red). (a) Experimental map from EMD-4400 at 5.7 Å, and simulated density maps from its fitted PDB 6I2T at the same resolution. (b) Same process with the structure EMD-1340 at a resolution of 9 Å and its fitted PDB 2P4N.

\[
f(r) = \frac{1}{\pi^{\frac{1}{2}}} \approx 0.225 \tag{2.3}
\]

\[
f(r) = \frac{1}{2^{\frac{1}{2}}} \approx 0.356 \tag{2.4}
\]

\[
f(r) = \frac{1}{2^{\sqrt{2} \log 2}} \approx 0.425 \tag{2.5}
\]

Where Eq. 2.3 is the default value. We show in Figure 2.2 the effect of the different normalization factors on simulated data. The lowest factor tend to highlight details not necessarily...
resolved experimentally, while higher factors tend to over-blur the data. A sweet spot for all resolution points towards Eq. 2.3, which highlight similar details in both experimental and simulated data. Hence, Eq. 2.3 will be used throughout this thesis.

2.1.3 Rigid docking of macromolecular structures in cryo-EM density maps

When an atomic structure has been pre-docked, or pre-placed in a density map, it is usually good practice to refine this position as to obtain an optimal match between atomic coordinates and density data. The method presented here only applies rigid-body translation and rotation operations to the high-resolution structures to maximize a cross-correlation coefficient with an experimental density map. This procedure is originally part of Chimera and ChimeraX\textsuperscript{29,30}, where it is known as Fit-In-Map.

This method offers only a local optimization of the structure's alignment to the density. At each step, it will either slightly translate or rotate the structure, in the direction dictated by the sum of the gradient of each atom with respect to the input density. Every batch of $N$ steps, the step size for both translation and rotation is halved if the maximum distance between the atom positions computed after the last batch and the current atom positions is lower than the step size. When the step size is below a threshold or when a maximum number of steps is reached, the algorithm stops. As such, this optimization is simple but fast and will only drive the structure to the closest minima where atom gradients will be negligible, hence why this method is only realistically applicable when the structure is already close to the global minima. Pseudo-code of the method is shown in \textbf{Algorithm 1}.

2.2 Molecular Dynamics

2.2.1 Birth and basis of molecular dynamics

Historical background

Molecular dynamics (MD) can be seen as a computational microscope into the internal dynamics of a system composed of atoms. Since the first simulation in 1957\textsuperscript{63}, the method has improved tremendously through both technological and algorithmic advances, allowing the simulation of atomic movements in many contexts in time scales ranging from the picosecond
Algorithm 1: Rigid local optimization of atomic coordinates in density maps

**Input:** densityMap, atomCoordsInitial, minStepSize

/* Center of mass of atoms */
atomCenter = mean(atomCoordsInitial)

/* Maximum distance from center to regulate rotations */
maxDistFromCenter ← maxDistance(atomCoordsInitial, atomCenter)

/* Initial translation vector and rotation matrix */
transVector ← [0,0,0]
rotMatrix ← identity(3,3,3)

/* Keep track of positions to adapt step size */
prevAtomPositions ← atomCoordsInitial

/* Compute gradient of density map and initialize interpolator */
mapGradient = computeGradient(densityMap)
mapGradientInterpolator ← GradientInterpolator(mapGradient)

while step ≤ nStep do

/* Transform structure with current translation and rotation */
atomCoords ← atomCoordsInitial
atomCoords ← transformAtoms(transVector, rotMatrix)
atomGradients ← mapGradientInterpolator(atomCoordinates)

/* Compute new transformation step */
if step % 2 == 0 then
    stepTrans ← sum(atomGradients) * stepSize
    atomCoords ← translate(atomCoords, stepTrans)
    transVector += stepTrans
else
    torqueAxis ← sum(atomGradients × AtomCoords)
    angle ← stepSize / maxDistFromCenter
    stepRot ← rotationMatrix(torqueAxis, angle)
    atomCoords ← rotate(atomCoords, stepRot)
    rotMatrix · stepRot
end

/* Adapt step size every N steps */
if step % N == 0 and step > 0 then
    if stepSize > maxDistance(atomCoords, prevAtomPositions) then
        stepSize *= 0.5
    prevAtomPositions ← atomCoords
end

/* Check for convergence */
if stepSize < minStepSize then
    break
end
at the time to milliseconds in more recent years. Molecular Dynamics was initially presented by Alder and Wainwright as a computational approach to circumvent the mathematical difficulties imposed by the many-body problem. This problem treats of the measuring of properties of microscopic systems composed of a number of interacting particles regulated by quantum laws. Measuring physico-chemical properties of these particles would require solving the Schrödinger equation for each of them, which quickly becomes intractable. Indeed, even with approximations, only a few hundred of atoms can be currently handled. For biomolecular systems consisting of proteins, biological membranes and/or DNA molecules, we can easily expect to exceed $10^5$ atoms including an aqueous medium. In 1927, Max Born and Robert Oppenheimer theorized an approximation that would allow atoms to be effectively uncoupled from quantum mechanics. This approximation stipulates that, when considering the mass difference between an atom nucleus and its electrons, it becomes evident that any change applied to a nucleus would be propagated almost instantly to its electrons thanks to their 2000-fold higher vibrational frequencies. Thinking of an atom as a single particle, it is thus acceptable to apply Newton's second law of motion rather than solving Schrödinger's equation, its quantum analog, ultimately reducing molecular dynamics to simulating a system of interacting spheres.

The first simulation of Alder and Wainwright was just that – a simulation of gas particles modelled as hard spheres, whose interactions were measured with a simplified Lennard-Jones potential. This is the blueprint of a molecular force field, which incorporates all forces and related parameters acting on the molecules of a system. These parameters can come from *ab initio*, semi-empirical or empirical means, although in the case of biomolecules, relevant force fields rely on empirical information as quantum calculations are often intractable for larger molecules. Almost too coincidentally, X-ray crystallography made a striking debut in the late 1950s as well, with the elucidation of DNA structure as well as 3D models of myoglobin and haemoglobin. Many parameters of molecular force fields are in fact derived from structures obtained by this method. Molecular force fields have been and are being improved on still today, with key contributions from Michael Levitt, Arieh Warshel and Martin Karplus that lead to their receiving the Nobel prize of chemistry in 2013. Popular force fields today include AMBER, CHARMM and GROMOS. Since the first simulation of a protein in 1977, molecular dynamics have continuously gained popularity and is often used...
Chapter 2. Methods

Figure 2.3 – Bonded and non-bonded terms of molecular force fields. (a) Bond length. (b) 3-body angle. (c) 4-body torsional (proper dihedral) angle. (d) 4-body improper dihedral angle. (e) Coulomb electrostatics. (f) Lennard-Jones potential displaying a strong repulsive force and a smoother attractive force between two atoms.

to complement experimental studies, or even provide significant results on its own\textsuperscript{73,74,75}. Moreover, the advances of X-ray crystallography and more recently of cryo-EM provides molecular dynamics with an ever-increasing number of starting structures, which can readily be used to simulate their dynamics, allowing to further comprehend their functions and mechanisms of actions in their respective contexts.

Calculating atomic movement

The motion of an atom $i$ can be calculated from its position at time $t$ using its mass and the forces acting on it. Typical timesteps $\delta t$ are in the femtosecond range and set to be faster than the fastest vibrational motion of the system. In typical atomistic simulations, this corresponds to 2 femtoseconds, shorter than the vibration frequency of a covalently bound hydrogen atom. Atomic positions over time is computed from its acceleration, which is obtained through numerical integration of Newton’s second law. This integration is most commonly done with the Verlet algorithm\textsuperscript{76}, which enables the computing of the position of atom $i$ from a time $t$ to a time $t + \delta t$ through the following equation:

$$r_i(t + \delta t) = 2r_i(t) - r_i(t - \delta t) + a_i(t)\delta t^2$$ \hspace{1cm} (2.6)
And the acceleration comes from Newton’s second law of motion:

\[ a_i = \frac{F_i}{m_i} = -\frac{\delta U(r)}{m_i \delta r_i(t)} \]  

(2.7)

Where \( m_i \) is the mass of atom and \( U \) is the atomistic potential, which is comprised of different terms:

\[ U(r) = U_{bonds} + U_{angles} + U_{torsion} + U_{improper} + U_{LJ} + U_{electrostatics} \]  

(2.8)

\[ U_{bonds} = \sum_{bonds} k_r (r - r_{eq})^2 \]  

(2.9)

\[ U_{angles} = \sum_{angles} k_\theta (\theta - \theta_{eq})^2 \]  

(2.10)

\[ U_{torsion} = \sum_{torsion} k_\psi [1 + \cos(n\psi - \psi_{eq})] \]  

(2.11)

\[ U_{improper} = \sum_{improper} k_\phi (\phi - \phi_{eq})^2 \]  

(2.12)

\[ U_{LJ} = \sum_{pairs} 4\varepsilon_{ij} \left[ \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left( \frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \]  

(2.13)

\[ U_{electrostatics} = \sum_{pairs} \frac{q_i q_j}{4\pi\varepsilon_0 r_{ij}} \]  

(2.14)

The atomistic potential hence includes four bonded (Eq. 2.9 - 2.12) and two non-bonded terms (Eq. 2.13 and 2.14). A visual aid in understanding each term is provided in Figure 2.3.

Non-bonded interactions are more complex and include the Lennard-Jones potential and Coulomb electrostatics. The Lennard-Jones potential models the London dispersion forces for weak, short-distance interactions between two atoms. It contains two terms, a strong repulsive force acting at short distances and another, smoother attracting force. The further
two atoms are, the closer to zero the potential will be. Theoretically, this means that all possible
atom pairs should be considered, but for practical purposes a cut-off is set above which the
contribution is zero, thereby saving considerable computing time and power. The Coulomb
electrostatics, conversely, measures long-range interactions and apply to all possible pairs of
charged atoms within a cut-off radius. Computation of this term is generally computed by
the Particle Mesh Ewald method\textsuperscript{77}. It separates short-range electrostatics from long-range
electrostatics, which are summed in Fourier space.

**Periodic boundary conditions**

Periodic boundary conditions constitute another staple of molecular dynamics. By setting
a unit cell and repeating it virtually infinitely, it prevents particles within the systems from
diffusing away from each other and keeps the amount of particles constant. This infinite
repetition however requires the system to have a total charge of zero. Importantly, when one
particle reaches one face of the box, it reappears on the opposite one (Figure 2.4). If the unit
cell is set too small, molecules within it may interact with themselves, generating significant
artefacts in the simulation. It is crucial to set the size of the unit cell according to the chosen
cut-offs for the non-bonded terms of the atomistic potential and size of the molecular entities
within the system.

**Statistical mechanics and the ergodic hypothesis**

Over the course of a simulation in a box with periodic boundary conditions, the particles will
adopt a number of microstates. Statistical mechanics states that an ensemble of microstates
has identical macroscopic observables such as particle number $N$, pressure $P$, volume $V$, energy
$E$ and temperature $T$. All these observables can be represented as ensemble averages of the
particles’ microstates, and statistical mechanics provide tools to relate macro- and microscopic
states depending on the ensemble of choice. Ensembles relevant for molecular dynamics
include the microcanonical ensemble with a fixed $N$, $V$ and $E$, the canonical ensemble with
fixed $N$, $V$, and $T$ and finally the isobaric-isothermal ensemble with fixed $N$, $P$ and $T$.

There are two ways of computing averages: over the ensemble as allowed by statistical
mechanics, and over time. Hence, these observables are measured as an average of all mi-
icroscopic configurations of the system in the view of statistical mechanics. Conversely, these
2.2. Molecular Dynamics

Figure 2.4 – Periodic boundary conditions. The unit cell is shown in the center, and some virtual copies from periodic boundary conditions are shown around it. Molecules crossing a border appear on the opposite side. An example is shown for a green, positively-charged molecule whose movement follows the red arrow.

quantities are calculated over time in molecular dynamics. The ergodic hypothesis establishes a relationship between the time and ensemble averages, ultimately leading to the ability of computing thermodynamic quantities from long enough molecular dynamics simulations. Indeed, as simulation time increases, more microscopic configurations of the system will be sampled, and more accurate measures of thermodynamic quantities can hence be obtained through time averages. Under the ergodic hypothesis, accurate estimation of thermodynamic quantities as dictated by statistical mechanics can be made with time averages given a large enough simulation time.

2.2.2 General steps to setup a molecular dynamics simulation

The basic concepts behind molecular dynamics as it is known today have now been laid out; it is time to cover how to actually set up and run a molecular dynamics simulation. The first step is to obtain atom positions within the structure of interest. Nowadays and as outlined in section 1.1.1, X-ray crystallography, NMR and cryo-EM are the main providers of this information. The next crucial part is to reproduce an isolated system with macroscopic observables relevant to the physiological conditions of the biomolecule. Practically, this
Chapter 2. Methods

translates into a number of steps, each of which serves to ensure proper quantities are reached, maintained and monitored over the course of the simulation.

Solvation of the system

The first step is to solvate the box containing the target structure. As outlined in the previous section, the size of the box is crucial to avoid the structure to interact itself through periodic boundary conditions. A typical valid size can be obtained by padding the structure by 12 to 15 Å in each direction depending on the cut-off chosen for the computation of the atomistic potentials. While some simulations may be carried out in vacuum, an explicit solvation model is to be preferred to prevent the structure from collapsing. The choice of solvation method boils down to which water model to adopt, of which the TIP3P water model is the most widely used. Due to the periodic boundary conditions and the Particle Mesh Ewald method for computing Coulomb electrostatics, a neutral system charge is required. To ensure this, a number of counter ions are added to the solvent. Then, ionic salt is added to a widely common concentration of 0.15 M to mimic physiological conditions. To avoid steric clashes, non-physical solvent-solute interactions and to speed up the solvation process, solvent molecules are placed in the box less packed that they will be during the actual simulation. Volume and pressure within the solvated box will be equilibrated accordingly at a later step.

Energy minimization

Once the biomolecule has been solvated, it is still far from its native biological conditions. This is especially true if the starting structure has been solved by X-ray crystallography, which may alter the structure and native contacts due to the crystal packing or due to secondary effects from the crystallisation buffer. The solvation of the structure from the previous step may also introduce steric clashes with solvent molecules or ions. All these points may lead to drastic non-physical changes in the structure during the simulation itself. Hence, it is customary to proceed to a short energy minimization step to remove, or at least alleviate, such problems. In general, a few thousand steps are sufficient.

Equilibration

The next steps consist in bringing thermodynamic observables to their target values. Equilibration protocols can vary significantly and are ultimately a matter of preference, but the
key principles remain identical. While the particle number N remains obviously constant, the 
temperature of the system needs to be brought up and maintained to 300 K and pressure to 1 
atm. This means that the total energy E and volume V of the system will have to fluctuate, and 
fail-safes will need to be put in place to keep the structure from unfolding or collapsing on 
itself due to rapid changes in temperature and pressure. In this thesis and often in practice, this 
system equilibration step is divided in two series of two steps which will allow for macroscopic 
quantities to adjust to their target values. The first series restrains the movement of the amino 
acid backbone of the protein (but not residue side chains) while the second series does not 
apply any restraints. The first series of equilibration steps focus on the solvent molecules, 
while the second one allows the structure to adjust to the target biological conditions. Both 
series start with a step in the NVT ensemble followed by a step in the NPT ensemble. The NVT 
step brings the temperature of the system from 0 to 300 K while keeping the volume constant. 
The NPT step regulates the volume and pressure of the system following the temperature 
increase. Equilibration steps are in general short, e.g. 250'000 steps.

Production

The production phase is the simulation of atomistic movements as described in the previous 
section through the numerical integration of Newton’s second law of motion. This phase can 
be carried out in several ensembles, although the NVT ensemble is the most widely used for 
proteins and other macro-molecules. Simulation production is run on a cluster of CPUs or on 
GPU(s) due to the heavy weight of the computations required.

2.2.3 Biased molecular dynamics and Molecular dynamics flexible fitting

Non-classical forms of molecular dynamics

Classical molecular dynamics aims at replicating physics with high accuracy. However, the 
computational load of the calculations outlined in section 2.2.1 may prevent the simulation of 
biological events taking place on longer timescales, such as protein folding, certain interac-
tions, or landscapes with high energy barriers in general. To this end, many biased forms of 
molecular dynamics have been developed, aiming to maximize conformational sampling or to 
drive the simulation in a predetermined direction, often requiring the setup of fail-safes to pre-
vent biologically irrelevant states from occurring. Metadynamics\textsuperscript{78} and umbrella sampling\textsuperscript{79}
are alternate schemes aiming to sample across high energy barriers. Alternatively, coarse-grained molecular dynamics force fields\textsuperscript{80} enable the grouping of atoms within residues to diminish the number of particles to simulate and enabling a larger timestep, an approach particularly popular with membrane proteins\textsuperscript{73,74,75}. Replica exchange\textsuperscript{81,82} schemes aim at simultaneously simulating a system at different values of a variable, such as temperature, using a set of replicas that are swapped accordingly. Similar to Accelerated molecular dynamics\textsuperscript{83}, such schemes target more extensive conformational sampling in less time. Steered molecular dynamics (SMD)\textsuperscript{84} is used to study a variety of molecular phenomena, such as unfolding or ligand unbinding. SMD simulations typically considers external forces on a set of atoms on top of the existing internal molecular force field. While the next paragraph will cover a special case of SMD used in chapter 5, some examples of other non-classical MD will be covered in chapter 3.

**Molecular dynamics flexible fitting (MDFF)**

More relevant to the theme of the thesis, molecular dynamics flexible fitting\textsuperscript{31} aims at driving protein movement in order to maximise cross-correlation with a cryo-EM density map, i.e. flexing an atomic structure so that it best represents the conformational data captured by cryo-EM. This method is based on the observation that X-ray crystallography solves structures at resolutions inferior to 3 Å while cryo-EM still routinely yields resolutions from 5 Å to over 10 Å in spite of the recent resolution revolution. MDFF is not the only method aiming to bridge the resolution gap between the two methods\textsuperscript{59,60,61,85}, some not based on molecular dynamics, but it stands out by its ease of use and speedy yet accurate results. Examples of applications are shown later in chapter 5.

As a variant of steered molecular dynamics, MDFF succeeds at the task by adding an external potential related to the inverse of the cryo-EM map density to the potential from the molecular force field (Eq. 2.8). This has the net effect of applying an external force that will steer an atom $i$ to regions of high density and away from regions of lower density:

$$U_{EM}(r_i) = m_i \begin{cases} 
\zeta \left( \frac{\Phi(r) - \Phi_{thr}}{\Phi_{max} - \Phi_{thr}} \right) & \text{if } \Phi(r) \geq \Phi_{thr} \\
\zeta & \text{if } \Phi(r) < \Phi_{thr} 
\end{cases} \quad (2.15)$$

Where $\Phi_{max}$ is the maximum of the density of the sign-inverted density map $\Phi(r)$. A
density threshold $\Phi_{thr}$ can be set to avoid fitting the atomic structure to any background noise within $\Phi(r)$. A scaling factor $\zeta$ is applied to this external potential, which is generally further weighted by the atom’s mass $m_i$. Several restraints may be applied to the structure to avoid over-fitting. These restraints may act to preserve secondary structure, the L-configuration of amino acid residues and proper handling of cis and trans isomers. It is key to check for errors before and after using MDFF. MDFF makes use of VMD\textsuperscript{86} plugins cispeptide and chirality to limit the propagation of errors during the fitting, while tools such as MolProbity\textsuperscript{87} may be used after the fitting.

### 2.3 Cluster analysis

Over the last decades, the amount of data in computational biology has increased significantly. Machine learning provides tools for its mining and sorting, paving the way for further advances. Tasks that require the grouping of unlabelled samples fall under the category of unsupervised classification methods. Among these, cluster analysis aims at identifying groups of samples that share common hidden features. Such groups are called clusters, which contain all samples that are more similar to themselves than to points belonging to other clusters. This definition leaves a lot to interpretation, as features are not always obvious and the chosen ones may turn out irrelevant to the initial goal, or better classification may come out from using a different set. This aspect however does not depend on the algorithm but on the user. However, spherical clusters, anisotropic distributions, background noise, density and population differences are all data properties that tend to divide clustering algorithms. Different properties may be relevant depending on which feature space the dataset is projected onto, making proper cluster analysis of biomolecular structures a complex endeavour.

Data preparation and feature selection thus cannot be dissociated from the algorithm as the success of the task depends on both. Hence, the more general an algorithm can be, the easier it will be to obtain results on new data and different feature sets, the alternative being trying as many approaches and parameter combinations as it takes. As a result of this complexity, a plethora of schemes exist, each targeting a set of specific data properties or aim at providing more general applicability to an existing algorithm. A number of key categories of clustering algorithms have risen over time and became staples of clustering, such as center- and centroid-based approaches, others based on density or those focusing on the connectivity
between points. A quick overview of each category will be provided here, before detailing the Density peaks algorithm of Rodriguez and Laio in preparation of chapter 3.

2.3.1 Center- and centroid-based clustering

The k-means algorithm is one of the oldest and yet most popular algorithm to date. This probably comes from its outstanding speed and simplicity, both to understand and implement, in contrast to the complex task that is clustering. While it has received countless iterations over the years, the base concept remains the same: the user selects a number of clusters that they think matches that of the data. Then, k-means randomly creates that many points in the feature space of the data. These points serve as putative cluster centroids, and data points are assigned to the centroid closest to them. Then, putative centroids are moved towards the mean of the points that were assigned to them. This is repeated until the position of the centroids shifts by a distance below a convergence threshold. The biggest drawbacks of this approach are three-fold: it is very sensitive to the random initial positions of the centroids, the reliance of the mean introduces a bias for spherical clusters (or requirement for well-separated ones) and selecting an initial number of clusters is often non-trivial. Common solutions include repeating the process to force different initializations, combining the process with heuristics, and sometimes selecting a higher number of clusters if those returned by the algorithm did not conform to expectations. Different metric than the mean, such as the median for the k-median variation, can be used instead to be more robust to outliers. The k-medoid alternatives choose actual data points as centers instead of centroids, which do not correspond to existing points in the dataset. A representation of the concept behind k-means is shown in Figure 2.5a.

2.3.2 Hierarchical clustering

Being unsupervised, clustering results are often not final. In practice, there can be many satisfactory clustering results, and even the ground truth may vary depending on expectations or goal. Hierarchical clustering offers not one clustering solution, but a dendrogram, where the user then chooses a specific hierarchy level. There are two main ways to proceed: a top-down, divisive approach, or a bottom-up, agglomerative approach. The former starts with all data points in a single cluster, and uses heuristics to divide it into clusters. Agglomerative clustering instead considers all data points as their own cluster, which are then merged according to
2.3. Cluster analysis

Figure 2.5 – Different categories of clustering. (a) An example of k-means. Two initial centroids are placed and represented by black crosses. Data points are assigned to their closest centroids. Then, centroids move towards the mean of the points assigned to them until convergence. (b) An example of agglomerative hierarchical clustering. The points are ordered by their euclidean distance, and subsequently merged with their nearest neighbor. Clusters at the selected hierarchy in the dendrogram (last panel) are shown in the middle panel. (c) An example of density-based clustering with DBSCAN. This algorithm defines an $\varepsilon$-neighborhood (red circle in the first panel), and a minimum number of points $\text{minPts}$ to be included within. The different possible outcomes during the process are shown in the middle panel. Data points colored in purple contain at least $\text{minPts}$ in their $\varepsilon$-neighborhood are core points. The data point in green is density reachable by a core point, as it is reachable by one but does not contain $\text{minPts}$ in its $\varepsilon$-neighborhood. Data points in orange are neither, and are hence categorized as noise. (d) An example of clustering by density peaks. A local density $\rho$ and minimum distance to higher density $\delta$ are computed for every point, and are plotted in a decision graph shown in the middle panel. Thresholds $\rho_{\text{thr}}$ and $\delta_{\text{thr}}$ are set to select cluster centers. Resulting clusters and outliers are shown in the last panel.
some metric. A linkage criteria needs also to be selected to merge clusters beyond the first level of the hierarchy. The agglomerative approach is more popular, with the Euclidean distance as metric and Ward’s method for merging clusters, which aims to minimize the within-cluster variance of all clusters within a level. A conceptual representation of agglomerative clustering is shown in Figure 2.5b.

2.3.3 Density-based clustering

Another class of clustering is those based on density. These algorithms focus on finding regions of higher point density separated by regions of lower density. This concept is at the core of DBSCAN, and one of its most well-known iteration, OPTICS. A conceptual representation of DBSCAN is shown in Figure 2.5c. DBSCAN takes two parameters: the radius of an $\varepsilon$-neighborhood and a minimum number points $minPts$ to be contained within such neighborhoods. Each data point is only visited once, making DBSCAN one of the fastest algorithm. A first point $p$ is selected and its $\varepsilon$-neighborhood is determined, i.e. its neighbours within that radius are identified. If there are at least $minPts$ in this $\varepsilon$-neighborhood, a cluster is initiated from $p$, which becomes a core point. All its identified neighbors are assigned to its cluster. The $\varepsilon$-neighborhood of each neighbor of $p$ is checked, which opens up two scenarios. One, the $\varepsilon$-neighborhood of $p$’s neighbor $n$ contains at least $minPts$ points and $n$ becomes another core of that cluster. Two, it contains less points than $minPts$ and $n$ is categorised as density-reachable instead. Such points are still assigned to $p$’s cluster, but will not sprout further search from their own neighbors. Once all neighbors originating from $p$ and related core points have been visited, the procedure starts again from the next unvisited point. Points that did not become core nor were density-reachable from clusters are categorised as noise. Hence, DBSCAN is able to find the proper amount of clusters without requiring the user to guess it a priori. The catch is that DBSCAN only works when a single density threshold can be found to distinguish clusters. If there are clusters of sufficiently different densities, DBSCAN will either classify lower density clusters as noise, or merge clusters if the density threshold is set too high or too low, respectively.

OPTICS aims at correcting this drawback by introducing two key concepts: a core distance and a reachability distance. The first is defined as the minimal radius of the $\varepsilon$-neighborhood of a core point $p$ so that it contains exactly $minPts$. The reachability distance is smallest of the
maximum between the core distance and the Euclidean distance to one of its neighbors that is also a core point. This distance is used to order the points within a cluster so that they are ranked close to their neighbors. The output of OPTICS is a plot that relates this ordering with the reachability distance, yielding a comprehensible unidimensional view on the dataset’s cluster structure. Clusters appear as valley, whose depth increases with cluster density. From there, it is possible to extract clusters either with a DBSCAN-like method by setting a threshold on the reachability distance, yielding a hierarchy of clusters, or with any algorithm detecting steep slopes to divide the plot into clusters as suggested by OPTICS’ authors.

### 2.3.4 Density peaks algorithm

The Density peaks\(^{56}\) (DP) algorithm by Rodriguez and Laio generated significant interest since its original publication in 2014. It is a surprisingly general algorithm, which is probably due to the fact that it does not fall in any of the categories outlined in previous section; rather, it borrows ideas from all of them. It is a center-based approach that provides some degree of hierarchy and relies on local density estimation. A conceptual representation is shown in Figure 2.5d. The novelty of DP lies in its definition of cluster centers: they are local density maxima located far from other points of higher density, which translates into two values computed for each data point. The local density \(\rho_i\) of data point \(i\) can be expressed as follows:

\[
\rho_i = \sum_j^N e^{-d_{ij}^2/d_c^2}
\]  \hspace{1cm} (2.16)

Where \(d_{ij}\) is the distance between data points \(i\) and \(j\), and \(d_c\) is a cut-off distance provided as only input to the algorithm. The minimum distance to a point of higher density \(\delta_i\) for data point \(i\) is in turn computed with:

\[
\delta_i = \min_{j: \rho_j > \rho_i} d_{ij}
\]  \hspace{1cm} (2.17)

The only exception to that is the point with highest density, which is attributed the largest distance between any two points in the dataset. The next step consists in plotting \(\rho\) and \(\delta\) for all datapoints in a decision graph (Figure 2.5d, middle). At this point, cluster centers should
become evident, and the user is tasked to select thresholds for both parameters. Data points above both thresholds are identified as cluster centers. Cluster assignment is done through a single loop over all points, ordered by decreasing $\rho$. This ordering allows each point to be assigned to the same cluster as that of their nearest neighbor of higher density, which is easily computed alongside $\delta$. The $\rho$ ordering ensures that the neighbor of higher density of a point being assigned either belongs to an existing cluster or is a cluster center. Moreover, by definition, the point of highest density in the dataset is itself a center. Outliers, which exhibit low $\rho$ but high $\delta$ values, are generally identifiable on the decision graph. To prevent them from being wrongly assigned to clusters, DP computes border densities between two neighboring clusters. For each such cluster pair, all the points that lie within a distance $d_c$ of a point belonging to the other cluster are recorded. The highest density within these points is referred to as border density $\rho_{\text{border}}$. Any point $i$ belonging to either cluster whose local density $\rho$ is lower than $\rho_{\text{border}}$ is unassigned and is considered an outlier. DP authors claim the algorithm to be effective regardless of cluster shape and dimensionality (for examples, see Figure 3.3 in the next chapter).

There are many limitations to the DP algorithm. The threshold selection step is rather unpractical, and cluster centers are often not obvious in practice. The decision plot also heavily depends on the cut-off distance $d_c$, practically yielding a 3-parameter algorithm requiring user input before and during clustering. Furthermore, the computation of both $\rho$ and $\delta$ depends on a full distance matrix, which makes the algorithm scale poorly with dataset size, both in terms of computation speed and memory use. The cut-off distance $d_c$ is also generally applied to the dataset, which leads to poor results when significant differences in density are present throughout a dataset. Furthermore, if two clusters of different densities are too close, the resulting border density $\rho_{\text{border}}$ may lead to considering most of the lower density cluster as outliers. A number of iterations have been published since to address some of these issues. The inclusion of k nearest neighbors (kNN)\textsuperscript{92,93,94} or heat diffusion\textsuperscript{95} allowed for a more robust estimation of $\rho$, leading to a better handling of datasets containing significantly different densities. However, manual intervention is still required to select thresholds for both $\rho$ and $\delta$. Conversely, Wang and Xu\textsuperscript{96} maximize an average silhouette index to select cluster centers automatically, although other input parameters are required instead. Liang and Cheng\textsuperscript{97} coupled principles from DBSCAN with a divide-and-conquer approach to recursively
2.4. Computer vision: local feature descriptors

2.4.1 Motivation and relevance for thesis

Computer vision focuses on providing computers with the ability to process and analyse image and video data to make informed decisions. Fields of application of computer vision include robot navigation, simultaneous localization and mapping, panorama and mosaic stitching, and most relevant to this thesis, the recognition of objects within a scene and the estimation of their pose within the latter. This can be done using local feature descriptors, which will be the main subject of chapter 4. Descriptors are generated in three main steps, the first being the detection of interest points. Such points, also referred to as key points, feature points or anchor points in this thesis, can be defined as a location in an image that is easily recognizable, and hence stable, under different views or transformations. In other words, an interest point has a well-defined position and its surroundings contain information usable to discriminate it from other interest points. The second step consists in providing each such point with an orientation so that it is invariant to rotation. This is done by processing the gradient of surrounding pixels. The third step consists in building a descriptor from the feature's rotation invariant surroundings, so that matching descriptors from different images may be identified and matched efficiently and exclusively. Successful descriptor matching in this sense leads to locating a first image into another or finding correspondences between them, depending on
Chapter 2. Methods

2.4.2 Feature detectors

Interest points are locations within an image that can be reliably recognized under viewpoint changes, rotation, or changes in scale, and contain significant texture in its surroundings. Blobs and corners are typical interest points (Figure 2.6). In this section, corner and blob detection principles and related methods will be detailed.

Laplacian of Gaussian, difference of Gaussians

Blob detectors aim at detecting regions that differ from their surroundings, such as in terms of

Figure 2.6 – Example of corner and blob features. (a) Source image of Helio the gecko. (b) Blobs detected by the Fast Hessian detector from OpenCV. The radius of the circles correspond to the features’ scale, and the line within denotes their orientation. Notice how the spot below the gecko’s eye is detected as single feature. (c) Corners detected by the Harris corner detector as implemented within OpenCV. Notice the finger junctions, eye corners and elbows being detected, but the spot has four different features attached. In both (b) and (c), specific thresholds were selected to better highlight the concepts behind the respective methods.

the target application.

In section 1.2 about integrative modelling, many hybrid modelling strategies were outlined. A core concept among them is the matching of structures or structural components obtained from different methods. In the framework of this thesis, this means locating high-resolution structural components into a low-resolution volumetric model of their assembly, which is directly translatable to typical application of local feature descriptors. In chapter 4, it will become clear how computer vision comes out as an attractive solution to current issues and limitations in integrative modeling.
2.4. Computer vision: local feature descriptors

**Figure 2.7 – Gaussian scale-space of an image.** An image of *Helio* the Gecko is consecutively blurred within two octaves. The upper octave has the image size halved, and its scales start from twice the initial sigma value of the previous octave. Difference of Gaussians (DoG) are obtained by subtracting two consecutive scales. Images with a red border are auxiliary images, so that three DoG images have two adjacent scales for scale-invariant feature detection, which is the optimal number of scales according to the original paper. The factor between scales here is $k = 2^{i/3}$, where $i$ is the index of the image in the octave ranging from 0 to 5.

Intensity. The scale-invariant detector from SIFT descriptors\(^{100}\) relies on an approximation of the Laplacian of Gaussian, which remains one of the most used nowadays, either as such or to train learned detectors\(^{101,102}\). Of note, the use of the Laplacian of Gaussian is also seen in hybrid modelling methods\(^{47,52,62}\) to better integrate density data by increasing shape contrast.

A key concept in feature detection and description is the scale. An image may contain features of different sizes, or an object may appear larger or smaller in the scene where it is to be recognised. Scale invariance is achieved through the building of a Gaussian pyramid of images\(^{103}\). This pyramid includes octaves, each of which halves the image size of the previous one, and a number of scales, which sequentially blurs the image further by increasing $\sigma$ so...
Figure 2.8 – Difference of Gaussian and Laplacian of Gaussian applied to a signal at different scales. The top panels show two Gaussian functions (left) used to generate the Difference of Gaussian (right), which is compared to the Laplacian of Gaussian. The bottom panels show a rectangular signal of different width being convolved with a Laplacian of Gaussian with different $\sigma$. For each signal, the maximum in the scale-space is shown by a red dot.

that its value doubles at each octave (Figure 2.7). A blurred image $I_\sigma$ at scale $\sigma$ is obtained as follows:

$$
I_\sigma(x, y) = G_\sigma(x, y) \ast I(x, y)
$$

(2.18)

where the Gaussian kernel the image is convolved with is defined as:

$$
G_\sigma(x, y) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)
$$

(2.19)

To extract interest points from this pyramid, SIFT then generates Difference-of-Gaussian (DoG) images by subtracting adjacent scales of the pyramid (Figure 2.7, in cyan). DoG is an approximation of the Laplacian of Gaussian (LoG) (Figure 2.8, top):
2.4. Computer vision: local feature descriptors

\[ DoG_{k\sigma} = (I_{k\sigma}(x, y) - I_{\sigma}(x, y)) \approx (k - 1)\sigma^2 \nabla^2 G \ast I(x, y) \quad (2.20) \]

where \( k \) is a factor chosen so that each octave starts with double the initial \( \sigma \) of the previous one, of, i.e. \( k = 2^{1/s} \) where \( s \) is the number of scales in an octave. SIFT found that 3 effective scales give the best results. In practice, this means that 3 auxiliary blurred images (and two DoG images) need to be additionally generated to have a complete octave (Figure 2.7, red frames). We notice that DoG do not require scale normalization as \( (k - 1) \) is constant over scales, whereas LoG requires each scale to be multiplied by \( \sigma^2 \) to be scale invariant as a consequence of the second derivatives.

A feature is extracted if its location \((x, y)\) is both an extremum in space and in scale. Effectively, \((x, y)\) is a feature if it is lower or higher than the 26 neighbours in the current scale and than the 27 neighbours in the adjacent ones, such as the red dots in Figure 2.8. From there, the location of the feature is refined to off-lattice (in scale and space) accuracy by taking the extremum of the Taylor expansion of the DoG:

\[ DoG(x) \approx DoG + \frac{\partial DoG(x)}{\partial x} x + \frac{1}{2} x^T H(x) x \quad (2.21) \]

\[ h = -\frac{\partial^2 DoG^{-1}}{\partial x^2} \frac{\partial DoG}{\partial x} = -H^{-1} \nabla DoG \quad (2.22) \]

where \( x \) and \( h \) are vector notations corresponding to \((x, y, \sigma)\) and \((\hat{x}, \hat{y}, \hat{\sigma})\), respectively, and the latter vector corresponds to the offset from the original on-lattice position. In matrix notation, this offset can be obtained by using the Hessian matrix and the gradient around the feature location using finite differences. If the offset in any direction is greater than 0.5, the feature is presumed to be closer to another pixel and the process is repeated starting from that pixel. If no satisfactory offset is found after few repetitions, the feature is discarded. Moreover, if the value of the sublocalized voxel is below a threshold, the feature is deemed unstable due to poor contrast. Finally, edge response is eliminated by looking at the principal curvature surrounding the feature. For this, the eigenvalues of the Hessian matrix from Eq. 2.22 are computed. If the ratio of between the largest and smallest eigenvalue is too high (a 10:1 ratio is
used in the paper), the feature is assumed to be located on an edge and is discarded. As will be further explained in a further paragraph (see also Figure 2.10), points located on an edge offer low stability due to the gradient being low in the edge’s direction, making any points along the edge poorly differentiable.

**Determinant of Hessian**

The feature detector of the Speeded Up Robust Features\textsuperscript{104} (SURF), the Fast-Hessian, is based on an approximation of the image’s Gaussian second order partial derivatives. On the premises that a Gaussian kernel is already discretized and that SIFT’s DoG approximation was at practically zero cost in accuracy, SURF authors further approximate the Gaussian kernel by using box filters instead (Figure 2.9). Such filters are surprisingly fast to apply to integral images, whose pixel are defined as the sum of all previous pixels in an image \( I \):

\[
U(x, y) = \sum_{i=0}^{x} \sum_{j=0}^{y} I(i, j) \quad (2.23)
\]

The most practical aspect of using a box filter on an integral image \( U \) is that it only requires a few addition and subtraction and no convolution operations. Furthermore, going over a Gaussian pyramid of images is also rendered useless, as only the size of the box filters need be increased to go over scales while higher strides will take care of higher octaves of the scale-space. The Fast-Hessian detector therefore only requires the base image to be integrated once for the whole scale-space. Interestingly, features extracted with the Fast-Hessian detector showed better performance on lower resolution images than DoG features when coupled with SIFT’s descriptor\textsuperscript{105}. The filtered map is made up of the determinant of the box filter-approximated Hessian matrix. Similarly to SIFT’s DoG detector, features are extrema located in scale and space in a 3x3x3 neighborhood which are then localized to subvoxel accuracy as before.

**Harris and Shi-Tomasi corner detector**

Corners are points of high interest for building descriptors: they correspond to a specific region in space, easily recognizable and stable under a wide range of viewpoints and transformations. Among the many corner detectors available\textsuperscript{106,107,108}, one that survived the test of time is that
2.4. Computer vision: local feature descriptors

Figure 2.9 – Gaussian partial derivatives, determinant of Hessian and box filters. The top panels show Gaussian first order and second order Gaussian derivatives, with the last panel showing the determinant of the Hessian matrix from second-order derivatives. The bottom panels show the equivalent for box filters as per SURF’s Fast Hessian detector.

of Harris and Stephens, developed in 1988\textsuperscript{106}. It is still used nowadays, although with a few key improvements, namely from Shi and Tomasi\textsuperscript{107} as well as a scale-invariant variation\textsuperscript{109}. The core concept behind this detector lies in measuring how the intensity within a window over the image varies by displacing it. For an edge, a change is only expected in one direction and for a corner, a change should be observed in both directions (Figure 2.10a). The sum of square differences $SSD$ between a window patch $W$ around point $(x, y)$ of an image $I$ and after a displacement $(u, v)$ can be written as:

$$SSD(u, v) = \sum_{W} [I(x + u, y + v) - I(x, y)]^2$$  \hspace{1cm} (2.24)

Then, a Taylor expansion can be applied to $I(x + u, y + v)$ and translating to matrix form, we obtain the following expression:

$$SSD(u, v) \approx [u, v] \left( \sum_{W} \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix} \right) [u, v] = [u, v] M [u, v]$$  \hspace{1cm} (2.25)

where $I_x$ and $I_y$ are the image’s partial derivatives in the $x$ and $y$ direction and $M$ is the structure tensor. The summation over the window $W$ may be done by convolving with a Gaussian function. Edges and corners can be identified by studying the eigenvalues of $M$. 
Harris and Stephens derived a corner response \( R \) to identify corner and edge regions:

\[
R = \text{det}(M) - k\text{Tr}(M)^2 = \lambda_1 \lambda_2 - k(\lambda_1 + \lambda_2)^2
\]  

(2.26)

where \( \lambda_1 \) and \( \lambda_2 \) are the eigenvalues of \( M \) and \( k \) is heuristically determined in the range of \([0.04, 0.15]\). At the time, this formula was derived in this manner to avoid the explicit and costly computation of the eigenvalues. If both eigenvalues are small, the window around the point \((x, y)\) covers a flat region. If one eigenvalue is large and the other one close or equal to zero, it is an edge. If both eigenvalues are large, then a corner is present in the window \( W \). The response \( R \) integrates this aspect in a single variable (Figure 2.10b). Shi and Tomasi further improved this formula by setting a threshold \( \lambda_{\text{thr}} \) and to consider the smallest eigenvalue as corner response instead:

\[
R = \begin{cases} 
\min(\lambda_1, \lambda_2) & \text{if } \min(\lambda_1, \lambda_2) > \lambda_{\text{thr}} \\
0 & \text{otherwise}
\end{cases}
\]  

(2.27)

In both cases, a response map is computed and non-maxima suppression is used to select
feature points. In essence, a point \((x, y)\) is a corner if it is a local maxima in a 3x3 window in the response map, which contains either the values of \(R\) or the value of the smallest eigenvalue in the case of the version of Shi-Tomasi. A drawback of these methods is the bias in favor of corners at a 90° angle, a direct consequence of the orthogonality of the eigenvectors of \(M\).

### 2.4.3 Feature descriptors

After detecting a set of interest points, the next step consists in describing their surrounding region in both a robust and discriminative fashion. Descriptors need to be robust to a variety of changes, such as viewpoint changes, occlusion, and illumination changes. Moreover, rotation invariance is a key aspect of most descriptors. In general, an object captured in two distinct images will not be aligned in the same way, meaning that the neighborhood of features needs to be reoriented accordingly. Descriptors need also to be discriminative so that they can be matched reliably between two different images or against a database of other feature descriptors with as few false positives as possible. There are a number of descriptor building schemes available, such as SIFT\(^{100}\) and variations\(^{110,111}\), SURF\(^{104}\), binary descriptors, which include (A)KAZE\(^{112}\), ORB\(^{113}\) or BinBoost\(^{114}\), while others take advantage of the recent potential uncovered by deep learning\(^{102,115}\). There are many benchmarks comparing these descriptors, and many report conflicting rankings, sometimes even on the same datasets\(^{116}\). This is not necessarily surprising, as the performance of descriptors typically depends on several hyper parameters and on their implementation. A recent study\(^{116}\) aimed at creating datasets for reproducible benchmarking of feature detectors and descriptors, and found that a SIFT variant\(^{110}\) still ranks among the best feature descriptors for image matching - even when compared to learned and deeply learned descriptors. Another study shows the same\(^{117}\).

Many descriptors are built with viewpoint, illumination and scale changes in mind. Considering the framework of this thesis, these aspects may not necessarily apply to biomolecular structures, which also include low-resolution grids from cryo-EM. A study made on infrared images aimed to measure performance of a variety of detectors and descriptors on lower resolution images from infrared cameras\(^{105}\). In that study, it was highlighted that SIFT descriptors are best combined with features extracted with the Fast-Hessian for this particular application, and ranked among the best descriptor building schemes. Still, there is no guarantee that these results are compatible to higher dimensions, or to a mix of low and high resolutions, as is the
case with cryo-EM density maps and atomic biological structures. Not all descriptors are easily translatable to 3D either. Most principles behind SIFT can be easily ported to 3D structures, and 3D versions exist for action recognition\textsuperscript{118} and medical images\textsuperscript{119,120,121} although they lack in critical aspects (see chapter 4). SIFT computes the gradient at every location around an interest point, creates and processes histograms based the gradient vectors’ orientations and magnitudes SURF is based on similar, but heavily approximated principles. Conversely, ORB, based on the FAST\textsuperscript{122} detector and BRIEF descriptor, may not be readily translatable to higher dimensions. FAST aims at identifying corner-like features as a series of contiguous point located on a circle around a potential interest point. Doing so in 3D would require porting the concept to a sphere, whose proper division is problematic (see section 4.3) and would certainly be costly. The orientation assigned to the corner is based on 2D measures with a single resulting angle, while rotation invariance in 3D requires at least a secondary angle, whose derivation may not be obvious using similar principles. The BRIEF\textsuperscript{123} descriptor may be amenable to 3D, as it simply computes intensity tests between pixel pairs, but does not provide rotation invariance, which is crucial for biomolecular structures. The rotated and learned variant of BRIEF used in ORB may be intractable, because the amount of possible binary tests in a 3D patch increases significantly on top of the secondary orientation issue previously mentioned. Learned and deeply learned approaches, such as TILDE\textsuperscript{101} and LIFT\textsuperscript{102} are powerful, but their training relied on SIFT features, thus requiring existing approaches to build training data. Of note, LIFT is trained on a patch under different viewpoints, which could be translated to different conformations of the same biological structure.

Due to the uncertainty about detector and descriptor performance on 3-dimensional, low-resolution data, the good overall performance of SIFT in a variety of applications (3D included) and its use to generate training data, the SIFT descriptor model has been chosen exclusively as basis to develop macromolecular descriptors for application to integrative modeling. The development of macromolecular descriptors will be fully presented in Chapter 4.

**Orientation assignment in SIFT**

The detectors previously outlined extracted the location in both scale and space of features. Orientation assignment provides rotation invariance to the final descriptors and is done by computing gradient magnitude and orientation in a patch around the feature location within
2.4. Computer vision: local feature descriptors

the blurred image at the feature’s detected scale, \( I_\sigma \) (Figure 2.11a). The size of the patch depends on the scale, reflecting the feature’s size. SIFT being patented, there is no related parameter provided but we can hypothesize that the size of the patch relates to that of a Gaussian kernel truncated at \( 3\sigma \) as it is commonly done. For each pixel in the patch, gradient orientation and magnitude are obtained with finite differences as follows:

\[
m(x, y) = \sqrt{(I_\sigma(x + 1, y) - I_\sigma(x - 1, y))^2 + (I_\sigma(x, y + 1) - I_\sigma(x, y - 1))^2}
\] (2.28)

\[
\theta(x, y) = \tan^{-1}\left[ \frac{I_\sigma(x, y + 1) - I_\sigma(x, y - 1)}{I_\sigma(x + 1, y) - I_\sigma(x - 1, y)} \right]
\] (2.29)

All values are sorted according to \( \theta \) in a histogram with bins every 10°. The contribution of each pixel is weighted by their magnitude \( m \) and by their distance to the feature through a Gaussian function, whose scale is 1.5 that of the feature point (Figure 2.11b). This provides robustness to changes that occur far from the feature point and reduces the effect of misalignment between object and scene, such as the corners of the patch. Quadratic interpolation around the maximum bin allows for a more accurate orientation, and other local maxima in the histogram with values above 80% of the global maxima are considered as secondary orientations, for which the feature is duplicated accordingly. If too many orientations are found, the feature is discarded. Else, the patch is rotated accordingly and subjected to description (Figure 2.11c).

Building descriptors

SIFT descriptors then describe oriented patches around features following a model of the receptive field within the primary visual cortex, which is more lenient on the spatial localization of features than their orientation. To reflect this, a trilinear interpolation step takes place (Figure 2.11d), which enables the division of the patch into a series of quadrants around the location of the feature. Each quadrant has their gradient magnitudes summed according to their orientation in a 8-bin histogram (Figure 2.11e), where a vector’s contribution to a given bin will be weighted according to its distance \( d \) to that bin’s center, yielding a factor of \( (1 - d) \) applied to its magnitude. This histogram, on top of the interpolation step, increases the descriptor’s robustness to structural changes around the feature point. This is of key impor-
Figure 2.11 – Orientation and description of a feature point. (a) A small example patch around a feature point, depicted by an orange pixel. The gradient orientation and magnitude are computed for each pixel in the patch and are represented by arrows. The gray background represents a Gaussian window function centered on the feature. (b) Each pixel is classified according to their orientation and weighed by their magnitude. The bin with the highest value is chosen as the feature's orientation. The gradient coordinates and orientations are then rotated accordingly. (c) The rotated patch, with the contributors to the dominant orientation colored in red. (d) A trilinear interpolation step then follows to create the grid holding the final descriptor. (e) The grid is subdivided in regions, where orientation histograms are computed, yielding the final descriptor. Here, a descriptor divided in 2x2 regions is shown, while in practice SIFT is built on descriptors divided in 4x4 regions.

Matching descriptors

Here, the application of interest for feature descriptors is the pose estimation of a given object within a scene. This is done by matching individually the descriptors of the object $D_o$ with those of the scene, $D_s$. To avoid computing all possible pairings, a nearest neighbor-based approach is typically followed. In essence, the two nearest neighbors of a descriptor $D_{o,i}$, $D_{s,j}$ and $D_{s,k}$, are used. The idea is that if $D_{o,i}$ matches $D_{s,j}$, then $D_{o,i}$ should be much closer to it than to $D_{s,k}$. If $D_{o,i}$ is close to both, then $D_{o,i}$ most likely doesn’t have a match in the scene. Pairs of descriptors that are truly matching can thus be identified by studying the NNDR, or Nearest-neighbor Distance Ratio between first and second neighbor. If repeated structures are to be expected in the images, then the ratio with a farther neighbor can be used instead, or a maximum distance between neighbors can be set and all descriptor pairs below the latter are kept. Because of the high dimensionality of their descriptors, SIFT resorts to an approximate neighbor search called Best-Bin First, as exact approaches offer no speed-up compared to exhaustive search in such cases. In 2009, Muja and Lowe published
2.4. Computer vision: local feature descriptors

RANSAC randomly selects pairs of points, and builds a model from that pair. In this case, a line is drawn between the selected points, and those that are within a distance $\varepsilon$ of that line are counted as inliers. The first two panels show inappropriate models with few inliers. The last panel show a good quality model build from the circled points. Points beyond the $\varepsilon$ distance are considered outliers.

Once potential matches have been identified, the homography between the object alone and the scene remains to be identified, which can be done using RANSAC, or random sample consensus\textsuperscript{127}. It is a probabilistic method that aims to find a set of inliers between two clouds of points, while simultaneously discarding outliers. RANSAC functions by randomly selecting samples, building an appropriate model from these samples, and then counting the number of inliers, or samples that agree with that model. A conceptual visualization of the process for a line model can be found in Figure 2.12. The models with the most inliers are retained, and a formula has been derived to determine the number of iterations $k$ to guarantee the finding of the best model with a given probability $p$:

$$k = \frac{\log(1 - p)}{\log(1 - w^n)}$$  \hspace{1cm} (2.30)

Where $p$ is the probability of finding a good estimation of the underlying model, $n$ the number of samples selected at each iteration and $w$ is an estimation of the number of inliers within the data. Progressive sample consensus (PROSAC\textsuperscript{128}) is an alternative to RANSAC, which is based on the same premise but pre-orders the samples according to some quality
Figure 2.13 – Example of SIFT. (a) The base images used in this example. On the left, the head of Helio the Gecko to be localized in the image on the right. (b) Interest points and descriptors as extracted by the SIFT implementation in OpenCV\textsuperscript{126} version 3.4.16. The radius of the circles represent the scale of each interest points, and the line inside depicts the descriptor’s orientation. (c) All interest points are shown in random colors. After identifying pairs of matching descriptors using NNDR, outliers were removed with RANSAC and the homography matrix was derived. The shape of the transformed image is shown at its estimated pose in the scene on the right. The final set of corresponding descriptors are shown in green. (d) The descriptor pairs rejected by RANSAC are shown in red. (e) The image in (a) has been placed where it has been posed by SIFT and RANSAC.

For descriptors, RANSAC would select four matching descriptor pairs at random and compute the corresponding homography between their respective feature points. The homography is applied to the remaining feature points, and the resulting model is evaluated by checking the number of selected feature points of the first image that are close enough to a feature point of the second image. The largest set of inliers found this way is kept, with the corresponding homography matrix. At this point, the pose of the object has been estimated within the scene of interest. A complete example of pose estimation using SIFT is shown in Figure 2.13. In that example, the spots on the gecko’s head may lead to many good features to be discarded due to their repetitions. As a result, the distance ratio with the fifth neighbor was used, which improved the number of inliers from 17 to 32.
3 CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

Disclaimer: this chapter is adapted from the following article published in Bioinformatics and available under the Creative Commons CC-BY-NC license, meaning that this work can be copied, redistributed and reused for any purpose except commercial ones.

Träger, S., Tamò, G., Aydin, D., Fonti, G., Audagnotto, M., Dal Peraro, M.  

**CLoNe: Automated clustering based on local density neighborhoods for application to biomolecular structural ensembles.**  
Bioinformatics, btaa742 (Accepted 21 August 2020)  
https://doi.org/10.1093/bioinformatics/btaa742

I am the main contributor for the method development as published. Dr. Giorgio Tamò aided in the initial discussion, project definition and early development stages. Dr. Deniz Aydin contributed to the analysis of the COQ9 system previously published\(^{75}\) and provided the corresponding molecular dynamics simulations. She contributed to the analysis of the TEM1 \(\beta\)-lactamase simulations provided by Vladimir Oleinikovas et al. as previously published\(^{82}\). Dr. Giulia Fonti contributed to the analysis of the KAP1 system, whose structural ensemble was co-developed by her and I as part of a previously published study\(^{129}\). Dr. Martina Audagnotto provided advice as well as the molecular dynamics simulations of the transmembrane \(\alpha\)-helix of the Amyloid Precursor Protein (APP) previously published\(^{73}\).
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

3.1 Background

The perception of molecular structures, especially proteins, is gradually shifting from the concept of one single and rigid structure to the idea that biomolecules natively exhibit a continuum of states. Protein folding, post-translational modifications, binding to other molecules or their involvement in catalytic events result in vast and complex conformational landscapes. Molecular Dynamics (MD), thanks to progress in both its technological and algorithmic aspects allows for the simulation of key biomolecular events (see section 2.2). Their observability, however, tend to be limited by currently accessible timescales. Researchers consistently come up with innovative protocols to push this limit further, granting us with the ability to capture protein folding as well as protein-protein, protein-membrane and protein-ligand interactions. State-of-the-art protocols for small-molecule docking, protein-protein docking and integrative modeling strategies in general have shifted towards the integration of dynamics in some form as well. All of the aforementioned aspects advocate dynamics as a cornerstone of modern structural biology and push the need for efficient tools to extract functional insight from structural ensembles in general.

However, these advances come at a price. The sheer size, the intrinsic complexity and redundancy of structural ensembles makes their successful analysis and computational integration non-trivial. Coarse-graining tools such as cluster analysis effectively reduce simulations of thousands of conformations to few key biological states and hence constitute a go-to approach with countless applications to date, reviewed in. Such states may serve as basis of Markov state models. To our knowledge however, an algorithm able to cluster data efficiently irrespective of their properties is still missing. Indeed, different cluster shapes, sizes and densities usually dictate which clustering approach is best suited for a given task.

Defining metastable states of proteins is non-trivial due to the large and often redundant number of internal degrees of freedom, yielding sampled conformational spaces with local minima often devoid of biological significance. We can make the assumption that, given enough sampling and a choice of relevant features, metastable states would lie in regions or clusters of high density, which would be separated by valleys of different density levels.
that would correspond to transitional states. Furthermore, no assumption can be made on the shape or relative densities of clusters, which would depend on both conformational sampling and target system. Unfortunately, structural biological data may exhibit regions at lower effective density but equal or even increased significance than others at higher densities. Moreover, the complexity of biological structures lead to numerous unique yet equally relevant choices of features, each with their own topology. This makes the analysis of such datasets by common clustering algorithm difficult, as small changes in the values of their input parameters may lead to drastically different and often unsatisfactory results, which implies that tedious fine-tuning steps have to be undertaken.

To remedy these issues and provide a clustering algorithm generally applicable to these cluster properties, we developed Clustering based on Local density Neighborhoods (CLoNe), recently published in Bioinformatics\textsuperscript{150}. It is an iteration of the Clustering by Fast Search and Find of Density Peaks by Rodriguez and Laio\textsuperscript{56}, which was detailed in section 2.3.4. In essence, the local densities of each point are computed using nearest neighbours and a Gaussian kernel and points associated with local density maxima are identified as putative cluster centers. To increase robustness to non-spherical cluster shapes, clusters are merged using the Bhattacharyya coefficient\textsuperscript{151} by comparing density distributions derived from putative cluster cores and boundaries. Finally, outliers from impromptu noise fluctuations are removed by means of a Bayes classifier.

CLoNe is able to find relevant clusters regardless of cluster shape, size, distribution and amount. We show its capabilities on many toy datasets with properties otherwise limiting the application of state-of-the-art approaches. Specifically, it improves on the original algorithm in key aspects, such as the reliance on a single, easy-to-tune and robust input parameter. First, we detail the different steps of clustering with CLoNe. Then, we apply CLoNe on a variety of datasets and compare it to other clustering algorithms\textsuperscript{56,88,89,90,91,99}. Then, we apply it to a variety of structural ensembles, including simulations of a membrane-binding protein\textsuperscript{75}, an allosteric pocket opening\textsuperscript{82}, oligomerization events within a membrane\textsuperscript{138} and a structural ensemble from a recent integrative modeling study from our lab\textsuperscript{129}.
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

3.2 Method development

3.2.1 Nearest-neighbor-based density estimator

A synthetic dataset was created to display the different steps behind CLoNe. It contains clusters of significantly different densities and various shapes. CLoNe starts by finding the k nearest neighbors of each point in a dataset \( X \) of \( N \) points using k-nearest neighbors (kNN), yielding a neighbor matrix \( M \) where each row \( i \) contains all the neighbors \( j \) of point \( i \) in increasing order of euclidean distance. To account for significant density differences between clusters, we initially assume that all points are cluster centers. In a first step, we estimate the local density \( \rho \) of each point \( i \) using a Gaussian kernel:

\[
\rho_i = \sum_{j \in kNN_i} e^{-\left(\frac{M_{ij}}{d_c}\right)^2} = \sum_{j \in kNN_i} \rho_{ij} \quad (3.1)
\]

Where \( kNN_i \) is the set of nearest neighbors of \( i \) in increasing order of distance and \( d_c \) is a cut-off distance defined as to be superior to a user-defined percentage \( pdc \) (the single input parameter of CLoNe) of all distances within \( M \), similar to the original DP algorithm (Rodriguez and Laio, 2014). We define the core of putative cluster \( i \) as the set of neighbors that contribute to \( \rho_i \) at least as much of the \( j-1 \) previous neighbors in average to a general cluster size \( p_{dc}N \):

\[
core_i = \left\{ j \in kNN_i \mid \rho_{ij} \geq \frac{1}{p_{dc}N} \sum_{k=0}^{j-1} \rho_{ik} \right\} \quad (3.2)
\]

We show in Figure 3.1a the cardinality (number of elements) of the core of each point in our synthetic dataset. As expected, this number is higher for points closer to real cluster centers and lower for points laying on the outskirt of clusters. The visualization of core cardinalities is an efficient way to observe the underlying topology of the dataset.

3.2.2 Automatic cluster center determination

In order to identify if \( i \) is a genuine candidate for cluster center, we identify its first neighbor \( j \) of higher density. If neighbor \( j \) belongs to \( core_i \), then neighbor \( j \) is a better candidate for cluster center in this region than \( i \). Conversely, \( i \) is a genuine candidate for cluster center if \( j \)
3.2. Method development

Figure 3.1 – Clustering based on local density neighborhoods. The dataset shown contains four croissant-shaped clusters of 500 elements and a central cross with 4000 elements. (a) The core cardinality of each point. (b) Results after the first stage of clustering. (c) The average Bhattacharyya coefficient between two clusters A and B. The upper plots show two clusters being merged. The lower plots show two clusters not being merged. Points belonging to the core of each cluster are shown in black, regular points in shades of grey and points belonging to the boundary in red. (d) Results after merging clusters. (e) The probability density functions of points belonging to any cluster cores (black), noise (blue), or to individual cluster cores (light blue). The range of local density of clusters classified as noise fluctuations are shown on the secondary y axis (dark blue). (f) The final results with outliers shown as black crosses.
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

is not in the core of \( i \). Cluster assignment is done in a single step by assigning a point to the same cluster as its nearest point of higher density, in order of decreasing local density, similar to the original DP approach. The results at this stage of CLoNe are shown in Figure 3.1a.

### 3.2.3 Merging subclusters

One of the drawbacks of the DP algorithm is its limited ability to deal with clusters with more than one peak or those with an elongated region of similar density, which may yield several subclusters instead of a single one (Figure 3.1c and some cases in Figure 3.3). Indeed, the Gaussian kernel in Eq. 3.1 is biased towards determining cores with spherical shapes and Eq. 3.2 is based on an approximative cluster size, which may not be accurate over the entire dataset. We can make the assumption that if two existing clusters \( A \) and \( B \) should be merged, then the density from one core to the other should be relatively constant. This can be estimated by looking at the core cardinality distribution of the points belonging to the core of both clusters as well as that of the points from the boundary between them (Figure 3.1c), which can be defined as:

\[
\text{boundary}_{AB} = \{ i \in A, j \in B \mid d(i, j) > d_c \}
\]  

(3.3)

Then, we define the following probability density function for the ensemble of points belonging to either cluster cores or the boundary from Eq. 3.3:

\[
P_S = \text{KDE}(\{\#\text{core}_i, i \in S\})
\]

(3.4)

Where \( S \) denotes one of the aforementioned ensembles and \( \#\text{core}_i \) the core cardinality of point \( i \) and KDE refers to the probability density function estimated by unimodal Gaussian kernel density estimation. Similarity between probability distributions can be measured using the Bhattacharyya coefficient (BC)\(^{151}\), which is bound between 0 and 1. Thus, the formula to compute the BC between the core of cluster \( A \) with the boundary from (3) becomes:

\[
\text{BC}_A = \sqrt{P_{\text{bound}}P_{\text{core}_A}}
\]

(3.5)
3.2. Method development

We take the decision of merging clusters A and B if the mean of their respective BC with boundary is above a threshold $\tau_{BC}$:

$$B(A, B) = \frac{1}{2} (BC_A + BC_B) > \tau_{BC} \quad (3.6)$$

Where $\tau_{BC}$ was chosen to be the 65th percentile to limit uncertainty and based on benchmarks. Taking the mean of both coefficients prevents the merging of a cluster whose probability density is similar to that of its boundary with a cluster of significantly higher density. This enabled us to identify clusters that can hardly be defined with a single density peak, such as uniform density over non-spherical shapes (Figure 3.1d and Figure 3.2ab). The point with highest $\rho$ is chosen as the new cluster center.

If $p_{dc}$ is chosen too small, clusters may be split into subclusters. Within our approach, these sub-clusters are likely to be merged into clusters matching the original topology, expanding the range of acceptable values for $p_{dc}$.

3.2.4 Outlier detection and removal

Another drawback of using the Gaussian kernel in Eq. 3.1 is that it may falsely identify impromptu local noise fluctuations as cluster centers. The original approach detects points within a distance $d_c$ of each other that belong to different clusters, and discards points below the highest density of the identified points (see section 2.3.4). This leads to clusters of lower density to be mostly discarded if a high-density cluster is nearby. To remedy this, we first recompute $\rho$ per cluster, so that influence from high-density clusters is no longer a concern. Then, we define two probability density functions. $M_{\text{cores}}$ is the probability density of the local density of all points belonging to any cluster core as per Eq. 3.4:

$$M_{\text{cores}} = \text{KDE}(\rho_x, \ x \in \text{core}_i \ \forall \ i \in C) \quad (3.7)$$

Where C is the set of cluster centers remaining after the previous merging step.
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

\[ M_{\text{outliers}} \] is the equivalent function for all points identified as outliers:

\[ M_{\text{outliers}} = \text{KDE}(\rho_{x}, \ x \in X \ if \ \rho_{x} < f_{\rho_{\text{center}}}) \] (3.8)

Where \(\rho_{\text{center}}\) is the local density of the center of the cluster \(x\) belongs to and \(f\) an arbitrary fraction chosen to be 0.1, after local densities have been recomputed for that cluster alone as mentioned above. This value does not need to be changed and was used for all examples presented in the article and this thesis. This is made under the assumption that points of lower density within an already identified cluster correspond to low probability points, which makes sense from a structural point of view but may hold true for many applications in general. This process enables us to estimate the frequency of noise within a dataset and to determine if an identified cluster is more likely to be derived from noise instead. To make this decision, we use a Bayesian classifier. For each cluster core \(c_{i}\), we derive the following posterior probabilities using Bayes’ theorem:

\[ p(Y|c_{i}) = \frac{p(Y)p(c_{i}|Y)}{p(c_{i})}; \ Y \in \{\text{outliers, cores}\} \] (3.9)

Where the prior probabilities \(p(Y)\) are defined as follows:

\[ p(Y) = \frac{|Y|}{N}; \ Y \in \{\text{outliers, cores}\} \] (3.10)

Where \(|Y|\) denotes the cardinality of the corresponding class. The likelihoods \(p(c_{i}|Y)\) can be computed by evaluating the previously defined probability distributions 3.7-3.8 at \(c_{i}\). Disregarding the evidence \(p(c_{i})\) common for both outliers and cores classes, we thus obtain the following Bayesian classifier:

\[ \hat{y} = \arg\max_{Y \in \{\text{outliers, cores}\}} p(Y) \prod_{x \in c_{i}} M_{y}(\rho_{x}) \] (3.11)

This classifier enabled us to remove all cluster centers arising from noise fluctuations. Combined with the previous merging step, the clustering is now complete (Figure 3.1ef).
3.3 Application to toy datasets

We applied CLoNe to a large set of common benchmark datasets, covering different key properties of clusters, such as non-spherical shapes, anisotropy, as well as significant size and density differences, all of which can be expected from real-world datasets from structural biology. In the previous section, we detailed how CLoNe automatically detects cluster centers, accurately merges clusters and removes outliers, succeeding in cases where previous iterations of DP did not (Figure 3.2a and Figure 3.3).

Similarly to the original DP algorithm, CLoNe requires a single input parameter $p_{dc}$, which relates to a cut-off distance used in the estimation of local densities (see previous section). In general, $p_{dc}$ takes a value in a small range and is intuitive to set. For instance, with $p_{dc}$ values of 1 or 2 local densities will be estimated considering neighborhoods small enough to identify individual spiral branches as clusters (Figure 3.2b). For higher values of $p_{dc}$, the scale of the Gaussian kernel in Eq. 3.1 will increase and merge individual branches into whole spirals, allowing the study of multiple hierarchies intuitively (Figure 3.2c). Other than clusters with non-spherical shapes, CLoNe identifies successfully the numerous Gaussian clusters.

**Figure 3.2 – General applicability of CLoNe on several datasets.** For each panel, the cardinality of the core of each point is shown on the left, highlighting cluster topology. The resulting clusters are shown on the right, with clusters in different colors and outliers as black crosses. (a) Noisy circles dataset. (b-c) Four instances of spiral dataset with different values of the input parameter yielding different yet valid clusters. (d) The a3 dataset containing 50 Gaussian clusters. (e) The s4 dataset with highly overlapping Gaussian clusters, some with anisotropic distributions. (f) A synthetic dataset with clusters of significantly different sizes and densities taken from Density Peaks Advanced.
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

Table 3.1 – CLoNe on high-dimensionality datasets. Each dataset contains 16 Gaussian clusters of 64 elements. In all cases, the 16 clusters are properly identified. The local density $\rho$ and the core cardinality of each cluster center are reported.

<table>
<thead>
<tr>
<th>Cluster#</th>
<th>64</th>
<th>128</th>
<th>256</th>
<th>512</th>
<th>1024</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>#core</td>
<td>#el</td>
<td>#core</td>
<td>#el</td>
<td>#core</td>
</tr>
<tr>
<td>1</td>
<td>29.05</td>
<td>27</td>
<td>64</td>
<td>29.05</td>
<td>32</td>
</tr>
<tr>
<td>2</td>
<td>29.02</td>
<td>31</td>
<td>64</td>
<td>29.02</td>
<td>31</td>
</tr>
<tr>
<td>3</td>
<td>29.03</td>
<td>31</td>
<td>64</td>
<td>29.03</td>
<td>31</td>
</tr>
<tr>
<td>4</td>
<td>29.04</td>
<td>32</td>
<td>64</td>
<td>29.04</td>
<td>32</td>
</tr>
<tr>
<td>5</td>
<td>29.05</td>
<td>32</td>
<td>64</td>
<td>29.05</td>
<td>32</td>
</tr>
<tr>
<td>6</td>
<td>29.06</td>
<td>33</td>
<td>64</td>
<td>29.06</td>
<td>33</td>
</tr>
<tr>
<td>7</td>
<td>29.07</td>
<td>35</td>
<td>64</td>
<td>29.07</td>
<td>35</td>
</tr>
<tr>
<td>8</td>
<td>29.08</td>
<td>35</td>
<td>64</td>
<td>29.08</td>
<td>35</td>
</tr>
<tr>
<td>9</td>
<td>29.09</td>
<td>35</td>
<td>64</td>
<td>29.09</td>
<td>35</td>
</tr>
<tr>
<td>10</td>
<td>29.10</td>
<td>35</td>
<td>64</td>
<td>29.10</td>
<td>35</td>
</tr>
<tr>
<td>11</td>
<td>29.11</td>
<td>36</td>
<td>64</td>
<td>29.11</td>
<td>36</td>
</tr>
<tr>
<td>12</td>
<td>29.12</td>
<td>36</td>
<td>64</td>
<td>29.12</td>
<td>36</td>
</tr>
<tr>
<td>13</td>
<td>29.13</td>
<td>36</td>
<td>64</td>
<td>29.13</td>
<td>36</td>
</tr>
<tr>
<td>14</td>
<td>29.14</td>
<td>36</td>
<td>64</td>
<td>29.14</td>
<td>36</td>
</tr>
<tr>
<td>15</td>
<td>29.15</td>
<td>36</td>
<td>64</td>
<td>29.15</td>
<td>36</td>
</tr>
<tr>
<td>16</td>
<td>29.16</td>
<td>36</td>
<td>64</td>
<td>29.16</td>
<td>36</td>
</tr>
</tbody>
</table>

of the A3 dataset (Figure 3.2d). Some degree of overlapping in real-world datasets is to be expected. The S4 dataset contains 15 highly overlapping Gaussian clusters of varying densities and shapes but equal size. As with the A3 dataset, CLoNe does not perform unnecessary merging with nearby clusters (Figure 3.2e) and is robust to large amounts of outliers on top of clusters with significantly different densities (Figure 3.2f). Similar to the original implementation of DP, CLoNe is applicable to high dimensionality datasets as well as evidenced by tests on 16 Gaussian clusters at 64, 128, 256, 512 and 1024 dimensions (Table 3.1).

We compared CLoNe to several commonly used algorithms in Figure 3.3, where CLoNe and OPTICS are the only algorithms that obtain optimal results on all datasets, although OPTICS tend to remove outliers inconsistently and may be slower. DP and DPA are clearly inferior to CLoNe on these datasets, where the latter has worse outlier removal and fails on the spiral dataset, where the original version succeeded. This general applicability of CLoNe coupled with a single, robust and easy to set input parameter is unique among the commonly used clustering algorithms found in the Scikit-learn package. The latest iteration from the original authors of Density Peaks, Density Peaks Advanced (DPA) also requires a single input parameter. We compared CLoNe with DPA in that regard on a dataset from their article (Figure 3.4). CLoNe’s $p_d c$ is indeed more robust than DPA’s $Z$, which exhibits a narrower range of valid input values.
3.3. Application to toy datasets

Figure 3.3 – Comparison between CLoNe and other clustering algorithms. The base code and parameters for each clustering algorithm is based on scikit-learn: https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.html, except for the last four rows and first three columns. Every column corresponds to a different clustering scheme. In order of left to right, there are CLoNe, Density Peaks, Density Peaks Advanced, minibatch K-means, affinity propagation, mean shift, spectral clustering, Ward hierarchical clustering, average linkage agglomerative clustering, DBSCAN, OPTICS, Birch and Gaussian mixture model. Rows from top to bottom correspond to the noisy circles, noisy moons, varied, anisotropic blobs, blobs, no structure, flame, spiral, aggregation and s4 datasets. The last four rows correspond to other relevant datasets, with optimal parameters obtained by scanning values. The runtime is shown on the lower left of each plot in seconds. Results for DP were obtained from a Python implementation of the original Matlab code and runtimes may differ. DPA results come from the original Fortran code.
Figure 3.4 – Large range of acceptable values for CLoNe’s unique input parameter. The upper dataset is one from Density Peaks Advanced\textsuperscript{99}, while the lower one is the same as in Figure 3.1. Clusters are shown in colors and outliers as smaller black crosses. In each case, the first row shows core cardinalities mapped on every data point as computed by CLoNe. Each column shows the clustering results for a different input value for CLoNe’s and DPA’s input parameter. In the case of DPA, the advertised range of value has been used. For CLoNe, the upper and lower rows showcase the clustering results before and after the merging step, respectively.
3.4 Application to structural ensembles

One of the principal aims of this work is to offer a clustering algorithm able to classify unlabelled biological structural ensembles into relevant states associated with their function and mechanism of action. We have applied CLoNe to real-world structural biology data reporting on the dynamic conformational space of a protein that associates with its specific biological membrane, cryptic allosteric pocket opening and dimerization of transmembrane proteins.

3.4.1 Membrane binding of Coenzyme Q9

COQ9 is a lipid-binding protein associated with the biosynthesis of coenzyme Q (CoQ), a redox-active lipid that is essential for cellular respiration\textsuperscript{157}. Recently, coarse-grained molecular dynamics (CG-MD) simulations and liposome co-flotation assays were used together to reveal that COQ9 accesses membranes in a multi-step fashion through a distinct, C-terminal amphipathic helix ($\alpha_{\text{10}}$)\textsuperscript{75}. In these simulations, COQ9 first diffused in the aqueous environment, then underwent various conformational changes upon membrane binding\textsuperscript{75}. We applied CLoNe to the CG-MD trajectory used in the latter study and sought to identify the main binding events pertaining to the protein itself. To this end, we extracted features characterizing both its movements in the aqueous environment through monitoring its distance to the membrane as well as key conformational changes based on the angle between its unique $\alpha_{\text{10}}$ and its globular domain (Figure 3.5a). Using these two features, CLoNe outputs three clusters, each of which seem to follow Gaussian distributions (Figure 3.5bc). One cluster regroups all conformations that correspond to diffusion movements in the aqueous environment, while the other two highlight the membrane association of $\alpha_{\text{10}}$ first followed by the globular domain as a converging step (Figure 3.5a), thus its higher density.

During the publishing process of CLoNe, reviewers asked whether similar results can be obtained hypothesis-free by using raw atomic spatial coordinates. To answer this, we extracted the coordinates of all 229 backbone atoms, yielding 687 features in total. After applying principal component analysis (PCA) and selecting the first two eigenvectors (eigenvalues of 0.65 and 0.15, respectively), clusters similar to the hypothesis-driven case were obtained (Figure 3.5d). Following CLoNe’s ability to handle high dimensionality datasets (Figure 3.5cd), we applied CLoNe directly on the 687 atom coordinates previously extracted. We projected the
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

Figure 3.5 – Cluster centers as key biological states of the COQ9 membrane protein. (a) COQ9 and its associated features, which include an internal angle $\theta$ and its distance to the membrane $d$. Conformations identified as cluster centers are shown in blue, red and orange. (b) The core cardinality of each frame of the simulation. (c) Clusters obtained, color-coded as in panel (a). (d) Clustering results of the first two principal components extracted from the coordinates of the 229 backbone atoms. (e) Clustering of the $x$, $y$ and $z$ coordinates of the backbone atoms without dimensionality reduction, projected on the principal space obtained in (d). (f) Same process, but projected on the feature space of panel (c). (g) Frame-by-frame label comparison between the different cases.

results onto both feature spaces to visually evaluate cluster labelling and obtain satisfactory results (Figure 3.5ef). A frame-by-frame label comparison is shown in Figure 3.5f. Through the visualization of cluster centers, CLoNe hence provides a quick overview of the main conformational changes occurring in the COQ9 simulation regardless of how the data is approached.

3.4.2 Allosteric pocket of TEM1 $\beta$-lactamase

In recent years, small-molecule docking software is no stranger to dynamics, taking into account ensembles of ligand conformations\textsuperscript{140} or receptor flexibility\textsuperscript{141,142,158}. A recent study highlighted a novel replica exchange-based molecular dynamics protocol combined with
3.4. Application to structural ensembles

benzene probes, where each replica harbors a different scaling of water-protein interactions. Using this method, the authors could observe the opening of cryptic allosteric pockets in several systems, including that of the TEM1 β-lactamase, which plays a critical role in antibiotic resistance. The simulations were started from the apo (i.e. without ligand) crystal structure with a closed allosteric pocket (Figure 3.6a). Out of the eight replicas of the simulation, we chose three with neutral (first), medium (fourth) and highest (last) scaling factors as a trade-off between maximizing the sampled conformational space and limiting redundancy of the over-represented closed conformations (Figure 3.6b). Along with key residue R244 on the opposite wall of the pocket, the opening of α-helices H11 and H12 and key residues L220 and N276 dictate pocket opening and allow two inhibitors to be accommodated, while the three mentioned residues form a triad when the pocket is closed. In addition to the opening of the two helices, visual inspection of the simulations indicated a deepening of the pocket. As a result, we chose features tracking the distance between the Cα of residues L220 and N276 as well as that of their sidechains to monitor pocket opening as well as the distance between the Cα of I263 and I279 as a measure of pocket depth (Figure 3.6a). The original study used fpocket to monitor pocket exposure in each replica (Figure 3.6b, top). The same was done on the clusters obtained by CLoNe (Figure 3.6b, bottom), showing different levels of pocket openness. Corresponding cluster centers highlight key structural differences between each state (Figure 3.6c), which are representative of the feature distribution per cluster (Figure 3.6d, top). Cluster assignation follows the observation of the original publication, where open states were more prevalent in the replica of medium scaling (Figure 3.6d, bottom).

3.4.3 Oligomerisation of Amyloid Precursor Protein

Another challenge in structural biology is the understanding of how biomolecules oligomerize to distinctive functional states. One of these cases, the transmembrane α-helix of the Amyloid Precursor Protein (termed APP hereafter), has recently been studied by our lab through the high-throughput MD protocol DAFT in order to identify which of two dimerization motifs is promoted depending on the lipid composition of the synaptic plasma membrane. The G700G704G708 motif is thought to direct the binding of APP to regulators promoting cholesterol biosynthesis, while the G709A713 motif would bind to cholesterol molecules (Figure 3.7a). Extracting features from molecular datasets is not always straightforward. Macromolecular
Figure 3.6 – Identification of different opening states of the allosteric cryptic pocket in TEM1 β-lactamase. (a) Holo and apo structures (left and right, respectively). Allosteric inhibitors are shown in grey and white. Features following helical opening include the distance between Cα atoms of N276 and L220 (medium blue) and the Cγ of their sidechain (light blue). Pocket depth is monitored by the distance between Cα-carbons of I263 and 279 (dark blue). (b) The pocket exposure calculated using the fpocket software for the original replicas (top) and for each clusters (bottom). The dotted line in both is the reference value of the holo crystal structure used in the original paper. (c) The center of each cluster in cartoon representation on top of a surface representation of the allosteric pocket, highlighting the different states of helical openness and pocket depth. The triad N276-L220-R244 governing pocket opening and closing are shown as grey sticks. (d) The distribution of each feature for each cluster (top) and the cluster assignment along the three chosen replicas (bottom).

movements possess an inherent redundancy due to the sheer number of internal degrees of freedom or prior knowledge may be lacking in order to select meaningful features, such as those highlighted in Figures 3.5 and 3.6. Dimensionality reduction methods, such as principal component analysis (PCA) have been used for the clustering of MD simulations and can help identifying coordinates of significance while discarding less useful dimensions.

The DAFT simulations of APP from a previous study are over 2 ms in total and contain countless states, many corresponding to unbound monomers. The first principal component based on the Cartesian coordinates of the coarse grain backbone covers 77% of the variability in the simulation, highlighting two clusters (Figure 3.7b). The blue cluster of lower amplitude corresponds to all states exhibiting unbound monomers (Figure 3.7e, left), while the second cluster regroups all the dimerized states regardless of motif. Focusing on that cluster, we calculated the pair-wise distances between the backbone atoms of each motifs in both helices.
3.4. Application to structural ensembles

Figure 3.7 – Classification and frequency estimation of APP dimerization motifs at the membrane. (a) The two known dimerization motifs of the transmembrane helix of APP. Membrane is depicted with grey spheres, the $G_{700}G_{704}G_{708}$ motif with teal spheres and the $G_{709}A_{713}$ motif with gold spheres. (b) The black line depicts the probability density distribution of the trajectory along the first principal component, which covers 77% of variability. Obtained clusters are shown in colors underneath, with their centers as black stars. Below the clusters is the distribution of core cardinalities, with the corresponding color bar below it. (c) The cluster of bound dimers of (b) was extracted and the pair-wise distances between the backbone atoms of each motifs in both helices was computed for every frame. The plot shows the first two principal components of the resulting dataset (with eigenvalues of 0.66 and 0.28, respectively), where each point is color-coded according to the cardinality of its core. Stars represent cluster centers. (d) The clusters obtained from the aforementioned dataset of bound conformations. (e) Renders of the unbound and dimerized APP cluster centers grouped by motif and their respective frequency after outlier removal. Dimerized APP centers shown correspond to the dark blue, dark brown and dark green clusters. (f) Renders of all cluster centers, color-coded according to (d). Sidechains have been hidden for better visualization.
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

(Figure 3.7a) and reduced these features to a two-dimensional principal space covering 94% of the variability before clustering (Figure 3.7c). We want to highlight CLoNe’s ability to analyze the neighborhood of low-density clusters without influence from high-density regions (Figure 3.7c). Clusters in blue in (Figure 3.7d) all depict states close to the \(G_{700}G_{704}G_{708}\) motifs and those in green the \(G_{709}A_{713}\) motif. In all cases, the darker-shaded clusters of each group correspond to the closest to the optimal motif arrangement (Figure 3.7e), while the others can be considered as closely related metastable states (Figure 3.7f). Similar to the original study\(^{138}\), CLoNe finds the preferred dimerization motif to be \(G_{700}G_{704}G_{708}\) as evidenced by the corresponding centers’ local densities, cluster population and core cardinality.

3.5 Case study: structural characterization of KAP1

A recent integrative modeling study from our group\(^ {129}\) involved the modelling of a macromolecular dimer known as KRAB-domain associated protein 1 (KAP1) or Tripartite Motif-containing protein 28 (TRIM-28). KAP1 plays a fundamental role in the regulation of gene expression by recruiting several transcription factors and altering chromatin organization\(^ {162,163}\). The N-terminal of KAP1 is comprised of a Really Interesting New Gene (RING) and two B-box domains followed by a coiled coil, collectively referred to as the RBCC domain (Figure 3.8a). A plant Homeodomain and bromodomain (PHD-Br) is present at the C-terminal after a long, disordered region containing a binding site for the heterochromatin protein 1 (HP1), which is a mediator of chromatin structure.

For over 10 years, KAP1 was wrongly thought to be trimeric\(^ {164}\). Indeed, size-exclusion chromatography can be used to estimate the molecular weight of globular proteins. However, KAP1 has an elongated structure, whose molecular weight tend to be overestimated by this type of chromatography experiments. This was taken into account in our study\(^ {129}\), where the molecular weight of KAP1 was estimated by size exclusion chromatography coupled with multi-angle light scattering (SEC-MALS) and analytical ultracentrifugation. These methods converged into KAP1 being a dimer in solution, although no information about its architecture could be drawn from these experiments. Given the extreme flexibility of the protein, high-resolution structural biology techniques such as X-ray crystallography and cryo-electron microscopy did not yield conclusive results. Instead, small-angle X-ray scattering (SAXS) was
used to structurally characterize KAP1. Raw SAXS data comes as a single curve, that can be used to retrieve the maximum length ($D_{\text{max}}$), thickness ($r_{\text{gyr}}$) and shape of the protein. A bead model can be created from this information, where each atom is approximated by a bead. Such models are often represented as a volumetric envelope (in contrast to cryoEM’s voxel grids) (Figure 3.8b). A theoretical SAXS curve can be drawn from a bead model and compared to the one obtained by the scattering of the sample solution. The final bead model corresponds to the one that maximizes the goodness of fit between the two curves according to the $X^2$ value. SAXS was performed on both the dimeric RBCC alone and on dimeric KAP1, showing the overall shape of the protein and an approximative internal architecture suggesting that the PHD-Br domains of KAP1 are asymmetrically arranged on the RBCC domains, as their presence only increased the thickness but not the maximum length of KAP1 (Figure 3.8b).

The SAXS curve really represents a weighted average of the different conformational states adopted in solution by the protein. As suggested by the SAXS bead models, KAP1’s flexibility likely stems from the PHD-Br domains and the 200 amino acid-long unstructured linker region before them. To explore the different possible configurations of the RBCC, linker and PHD-Br domains, we generated 1'000 random models of KAP1. Details of the molecular modeling process of the RBCC domain is not covered here as I did not contribute to it. Starting from that model of dimeric RBCC, whose validity was recently confirmed by X-ray crystallography\textsuperscript{165}, the two PHD-Br domains were randomly rotated and placed within a maximum distance of 140 Å perpendicular to coiled coils. The unstructured linker was modelled as random coils using the software MODELLER\textsuperscript{166}. An example model is shown in Figure 3.8c. The aim of this procedure is to obtain an ensemble of atomic models in agreement with the SAXS experimental data without bias of the initial domain configuration. As for bead models, theoretical SAXS curves can be derived from each of these random models, which can then be compared to experimental data with a $X^2$ value. The latter was optimized using the PEPSI-SAXS software\textsuperscript{167}, which relies on non-linear normal mode analysis\textsuperscript{168} to flex the models until convergence. This procedure results in 597 converged models with satisfactory $X^2$ values\textsuperscript{129}.

To characterize the architecture of KAP1 from these models, we measured the distances between the PHD-Br domain of the first KAP1 monomer $A$ and both RB modules (i.e. the RBCC domains minus the coiled coil), termed $d_{A1}$ and $d_{A2}$, and kept their minimum, termed $d_A$. The same was done for monomer B (Figure 3.8c). Prior to fitting, the resulting plot reveals
Figure 3.8 – Structural characterisation of KAP1. Panels b, d, h and i are adapted from their original publication \(^{129}\). (a) Sequence of KAP1 and its domains. (b) SAXS bead models as volumetric envelopes of the RBCC dimer and KAP1 full-length. (c) An example of an atomic model of KAP1, where the PHD-Br domains and linkers have been randomly placed and generated. \(d_A1, d_A2, d_B1\) and \(d_B2\) refer to the distances between the center of mass of the PHD-Br domains and that of the RB modules. (d) On the left, the distribution of \(d_A\), the minimum distance between \(d_A1\) and \(d_A2\), and \(d_B\), the minimum distance between \(d_B1\) and \(d_B2\) from the random initial models of KAP1. On the right, the same distribution after flexible optimization to maximize the goodness of fit between the random models and the experimental SAXS data. The white dots correspond to the models with the highest densities. (e) Core cardinalities as computed by CLoNe after sorting the distances \(d_A\) and \(d_B\) in ascending order, rendered possible by the evident symmetry of the previous panel. (f) The two resulting clusters identified by CLoNe. Centers are shown as black stars, and outliers as black crosses. (g) Surface representation of the two cluster centers following the same color code as in (f). Grey circles have been added to highlight the PHD-Br domains and RB modules, and different color shades denote individual monomers. (h) and (i) show a superimposition of all the final models of the red cluster of panel (f). A single RBCC dimer is shown in grey, while the PHD-Br domains closest to an RB module are shown in red and the others in blue. Panel (h) shows the models where one PHD-Br is closest to the RB module of monomer A while panel (i) shows the same for monomer B.
3.6. Comparison with other clustering algorithms

a homogeneous distribution with both PHD-Br domains located far from any RB modules as expected (Figure 3.8d, left). After the fitting, a clear and symmetric pattern arises, highlighting an asymmetric localization of the PHD-Br domains where one is close to an RB module and the other one far (Figure 3.8d, right).

Given the symmetry of the distribution, we sorted the distances $d_A$ and $d_B$ in ascending order, yielding $d_{Close}$ and $d_{Far}$. CLoNe was used on this new distribution and found two clusters Figure 3.8ef. The most populated cluster exhibits one PHD-Br domain close to an RB module at 60 Å while the other PHD-Br domain is located at about twice that distance to the nearest RB module. The other, lower density cluster is comprised of what can be interpreted as low probability transitional states, where both PHD-Br domains are located further away from the RB modules Figure 3.8g. As was the case for all the previous structural ensembles, the extracted clusters tend to follow Gaussian distributions. As such, cluster centers may be interpreted has higher probability states of the protein. We show in Figure 3.8hi a superimposition of both PHD-Br domains of the converged models of the main cluster. We can observe that they share the same asymmetric architecture of the cluster center. This specific feature came as a surprise and is currently unique to KAP1 within its family of proteins.

A table summarising cluster information of the structural datasets shown in this section is shown in Table 3.2.

3.6 Comparison with other clustering algorithms

We previously showed a comparison between CLoNe and other clustering algorithms (Figure 3.3). There, it was highlighted that only OPTICS had a comparable performance. However, these datasets are academic at best and do not reflect the complexity of biomolecular structural ensembles. Hence, we applied the main contenders of CLoNe to all the ensembles presented in this chapter. Results are shown in Figure 3.9. DBSCAN handles COQ9, KAP1 and APP well enough, though due to different density levels it misclassifies a few points in the case of KAP1 and is unable to detect more than one membrane-bound state of COQ9. A similar observation can be drawn for TEM1 and APP (bound), but in these cases the clustering results are not satisfactory. OPTICS, as advertised, is able to detect different density threshold and accurately clusters all cases (save an additional cluster in the case of COQ9), but tend to classify most
Chapter 3. CLoNe: Clustering based on Local density Neighborhoods for application to biomolecular structural ensembles

Table 3.2 – Statistics for clusters and centers from structural ensembles and related features obtained with CLoNe.

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>$\rho$</th>
<th># core elements</th>
<th>-outliers</th>
<th>$\theta$</th>
<th>$d$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusing states</td>
<td>69.1</td>
<td>69</td>
<td>708</td>
<td>767</td>
<td>128.54</td>
</tr>
<tr>
<td>Membrane bound #1</td>
<td>96.06</td>
<td>86</td>
<td>291</td>
<td>281</td>
<td>99.64</td>
</tr>
<tr>
<td>Membrane bound #2</td>
<td>78.81</td>
<td>98</td>
<td>398</td>
<td>392</td>
<td>150.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>$\rho$</th>
<th># core elements</th>
<th>-outliers</th>
<th>$d_{cc}(N263, L279)$</th>
<th>$d_{cc}(N276, L220)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep pocket</td>
<td>241.23</td>
<td>253</td>
<td>1039</td>
<td>986</td>
<td>12.75</td>
</tr>
<tr>
<td>Open pocket</td>
<td>395.3</td>
<td>449</td>
<td>2540</td>
<td>2208</td>
<td>9.89</td>
</tr>
<tr>
<td>Semi-open pocket</td>
<td>491.11</td>
<td>613</td>
<td>1823</td>
<td>1612</td>
<td>8.39</td>
</tr>
<tr>
<td>Closed pocket</td>
<td>758.34</td>
<td>811</td>
<td>9598</td>
<td>6839</td>
<td>8.83</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>$\rho$</th>
<th># core elements</th>
<th>-outliers</th>
<th>$d(G700)$</th>
<th>$d(G704)$</th>
<th>$d(G708)$</th>
<th>$d(G709)$</th>
<th>$d(A713)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GA #1 (light green)</td>
<td>124.83</td>
<td>192</td>
<td>203</td>
<td>135</td>
<td>12.21</td>
<td>1.94</td>
<td>9.06</td>
<td>8.97</td>
</tr>
<tr>
<td>GA #2 (green)</td>
<td>236.01</td>
<td>274</td>
<td>997</td>
<td>967</td>
<td>11.97</td>
<td>10.51</td>
<td>8.65</td>
<td>7.31</td>
</tr>
<tr>
<td>GA #3 (dark green)</td>
<td>190.24</td>
<td>232</td>
<td>430</td>
<td>404</td>
<td>11.77</td>
<td>10.46</td>
<td>9.87</td>
<td>5.1</td>
</tr>
<tr>
<td>Hybrid #1 (brown)</td>
<td>330.82</td>
<td>418</td>
<td>978</td>
<td>880</td>
<td>7.39</td>
<td>6.28</td>
<td>6.48</td>
<td>8.74</td>
</tr>
<tr>
<td>Hybrid #2 (dark brown)</td>
<td>246.34</td>
<td>348</td>
<td>734</td>
<td>671</td>
<td>8.94</td>
<td>7.69</td>
<td>9.55</td>
<td>11.82</td>
</tr>
<tr>
<td>GGG #1 (light blue)</td>
<td>292.95</td>
<td>347</td>
<td>1136</td>
<td>974</td>
<td>9.49</td>
<td>8.69</td>
<td>9.55</td>
<td>11.82</td>
</tr>
<tr>
<td>GGG #2 (blue)</td>
<td>381.7</td>
<td>428</td>
<td>2574</td>
<td>2219</td>
<td>7.9</td>
<td>7.36</td>
<td>7.68</td>
<td>9.75</td>
</tr>
<tr>
<td>GGG #3 (dark blue)</td>
<td>363.07</td>
<td>428</td>
<td>2168</td>
<td>1745</td>
<td>5.41</td>
<td>5.31</td>
<td>10.32</td>
<td>11.88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cluster Type</th>
<th>$\rho$</th>
<th># core elements</th>
<th>-outliers</th>
<th>$d_{Far}$</th>
<th>$d_{Far}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>KAP1 Far - Far</td>
<td>18.89</td>
<td>27</td>
<td>93</td>
<td>90</td>
<td>53.71</td>
</tr>
<tr>
<td>KAP1 Close - Far</td>
<td>36.81</td>
<td>42</td>
<td>504</td>
<td>420</td>
<td>59.66</td>
</tr>
</tbody>
</table>

points as outliers in these datasets (25% for COQ9, 84% for KAP1, 95% for TEM1, 51% for APP (all) and 60% APP (bound)). While DPA obtains comparable results to CLoNe for COQ9 and both APP datasets, it is unable to detect a second cluster in KAP1’s dataset and identifies two superfluous clusters for TEM1 (6 and 14 data points) in addition to classifying 50% of the frames as outliers. Of note, DPA was unable to process datapoints with identical values in the case of TEM1 and APP (bound), so the corresponding datasets were re-processed to contain only one data point with the same value. DBSCAN is the fastest algorithm of the lot, which is unsurprising due to its simplicity. It remains an algorithm of choice if different density levels are not expected. OPTICS is slower than CLoNe for smaller datasets, but scales better when size increases beyond 15’000 points, where CLoNe slows down considerably.

DPA is the slowest by a large margin. Of note, OPTICS is implemented in python within the scikit-learn package and DPA is coded in Fortran. This means that CLoNe and OPTICS’ speed performance can be compared on equal footing (although I do not pretend to be as efficient a programmer as them), while the low speed of DPA is definitely a surprise given the high performance of the Fortran language.

This difference may be explained by the different approaches followed. DPA makes use
3.6. Comparison with other clustering algorithms

We selected the best performing algorithms highlighted in Figure 3.3 and compared their performance on our structural datasets to CLoNe. For each algorithm, we scanned candidate parameter values until clusters resembling those of CLoNe were achieved. Outliers are shown in black, and clusters are colored in the same manner as in the main text whenever possible. In the case of TEM1, due to the 3D projection, outliers have been hidden so that the clusters remain visible in the case of OPTICS and DPA. The runtime in each case is shown on the upper right of each panel.

of a recent density estimator developed by the same authors, named PAk\(^98\). In essence, it computes volumes around points, with radius based on the distance to nearest neighbors. This is done for every single point in a dataset and for as many nearest neighbors as needed until the density is no longer considered constant between successive shells, according to a likelihood ratio test. The cluster centers are identified following a similar definition as the one in the original approach, where a center should have high local density and high distance to a point of higher density, where the distance is based on the optimal neighborhood size highlighted with the likelihood ratio test. It is clear this is an expensive process, especially for large, dense clusters. The coarser approach followed in CLoNe is much faster as densities and
neighborhood radii can be computed by a simple cumulative sum of Gaussians, truncated where new neighbors contribute less than the average contribution of their predecessors. Despite the added information obtained in the process, DPA still requires additional steps to discard centers arising from local fluctuations and merging them accordingly, whose success seems to be quite sensitive on DPA's input parameter $Z$. This might be surprising as DPA is advertised as fully unsupervised and parameter-free, before $Z$ is introduced later in the article. However, DPA provides more than clusters and offers an in-depth topographical interpretation of the data, while CLoNe only does this partially through core cardinalities.

### 3.7 Implementation and functionalities for structural ensembles

The software has been written using Python 3.7. CLoNe itself behaves as other clustering algorithms of the scikit-learn library. When used on structural biological data, cluster centers are exported as separate PDB files, individual clusters as XTC trajectories and Tcl scripts are generated to load centers or clusters automatically within the visualization software VMD. These scripts are easily configurable to tune molecular representations following VMD’s standards and contain the necessary information to do so. Basic usage of CLoNe is show in Figure 3.10.

The loading of molecular dynamics trajectories as well as the saving of cluster centers and cluster sub-trajectories is done through the MDTraj package. We use the NumPy, SciPy, scikit-learn and Statsmodels packages for many operation and to compare CLoNe to other clustering algorithms.

Source code is available at https://github.com/LBM-EPFL/CLoNe/.

### 3.8 Conclusion

Many clustering methods rely on parameters that are often non-trivial to optimize or on random initial conditions that may drastically change the outcome. Commonly used algorithms are generally restricted to specific cluster properties, forcing the user through a process of trial and error involving choice of algorithm and input parameter values. Moreover, choosing relevant features from structural datasets is challenging and different features may generate
3.8. Conclusion

**Figure 3.10 – Basic usage of CLoNe.** On the upper left, the commands to run CLoNe on structural data are shown as well as a short explanation of the different inputs. The compatible formats for MD trajectory and topology files are those compatible with the MDTraj library, as it is used to load coordinates and to extract cluster centers and related trajectories. Similarly, the syntax used for atom selection is the one of MDTraj. The right-hand side of the figure shows some of the output generated by CLoNe. In addition to the plots similar to those shown in the main text, there is a separate file summarising the results and detailing statistical information on the clusters. The frames identified as cluster centers are shown below said summary. Not shown in this figure are the Tcl loading scripts to load either the cluster centers or the cluster trajectories into VMD for further visual inspection. On the lower left is shown the minimal code to run CLoNe as a general clustering algorithm as well as a separate script to run either a selection of benchmark datasets or one provided by the user along with a value for CLoNe’s input parameter. A sample input file is shown for the data to cluster. The same file format has to be used for structural features, as header names are used for some output files.

For structural datasets, CLoNe is able to extract clusters as separate trajectories and provides scripts for their automatic loading in the visualization software VMD. For larger different cluster topologies, sometimes irrelevant to the end goal. CLoNe was designed with these issues in mind and aims to provide a stream-lined analytic process to yield results rapidly along with helpful visualization scripts to analyze and confirm the relevance of the clusters in the target biological context. CLoNe’s only parameter regulates the size of the local neighborhood considered around each data point, which can be regarded as cluster sizing parameter. Its value need only be decreased if clusters seem too inclusive and vice-versa. Furthermore, we have shown on one example that different values may lead to different cluster hierarchies, and that the range of acceptable values is large, making CLoNe an intuitive algorithm to use in addition to its general applicability.

For structural datasets, CLoNe is able to extract clusters as separate trajectories and provides scripts for their automatic loading in the visualization software VMD.
macromolecules, the concept of a conformational state is blurry, hard to determine and often
depends on context. It is not always clear which features to use to obtain an accurate partition
of the structural ensemble. The results obtained on COQ9 can be obtained hypothesis-free
on raw spatial coordinates or using PCA to extract relevant features. The use of PCA was
successful in the case of APP, where it enabled the reduction a high-dimensional space to
a single, meaningful dimension and to remove redundancy within another, leading in both
cases to accurate results in dimensionalities that can be appreciated visually as well. However,
when other features than the targeted ones can be expected to exhibit motions of larger
amplitudes, PCA will favor the latter over the former. This is true for the TEM1 β-lactamase,
where internal structural motions will be more prevalent than the fluctuations of the selected
key pocket residues. In such cases, a feature-based approach is to be preferred. Alternatively,
some will advocate the use of time-lagged independent component analysis (TICA) (Naritomi
and Fuchigami, 2011) in stead. TICA was found to be the better alternative for building
Markov state models. However, in cases where large amplitude fluctuations are the
target or when there is redundancy in features, we believe that PCA remains a safe approach.
Alternatively, if one wishes to disentangle internal from overall motion, dihedral PCA has been
used successfully in peptide folding.

As the conformational ensembles presented in this study tend to exhibit Gaussian dis-
tributions, CLoNe may thus be used to extract cluster centers as higher probability states.
Such states offer an overview of the ensemble and may serve as starting models for building
Markov state models in general. Moreover, the precision of the classification achieved by
CLoNe enables the identification of dominant biological states from large datasets. Beyond
the case of APP, CLoNe identified different key pocket conformations in the case of TEM1
β-lactamase. Further clustering efforts on this system should target the different positions of
R244, which was not tracked in this study but was previously shown to play a dual role between
TEM1’s active site and allosteric pocket. CLoNe may then be used as a pre-processing tool
prior to small-molecule docking studies, where accounting for receptor flexibility is an active
development area.

We approached integrative modeling in section 1.2. Such hybrid strategies typically
combine low resolution data of whole complexes with high resolution structures of their
components so as to predict the quaternary structure of the former. This process is how-
ever severely hindered by structural dynamics differing between a complex and its isolated components. For this reason, many hybrid modeling strategies now incorporate some form of dynamics to bridge this gap\textsuperscript{12,39}. While we previously utilized classical molecular dynamics for the prediction of heptameric aerolysin pores\textsuperscript{37,38}, such an approach would not be feasible for heteromultimeric assemblies where multiple conformational ensembles are required simultaneously. Reducing them to their crucial components may enable the structural characterization of large macromolecular complexes, which may otherwise be intractable. CLoNe will be used in the upcoming chapters as well for this very purpose, as stated in section 1.3. Applied to the fields of small-molecule docking, integrative modeling and structural dynamics studies, CLoNe presents itself as a versatile and powerful tool for modern computational structural biology.
4 MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

4.1 Solving biomolecular puzzles with feature descriptors

A number of integrative strategies involving low-resolution cryo-EM data were introduced previously (see section 1.2). Many of these methods rely on feature extraction. The first to go down this path is SITUS, in which a strategy involves topology-representing neural networks to identify a series of codebook vectors or feature points to summarize structures. This reduced the problem of docking high-resolution structures into low-resolution map to a point-matching problem. Docking of high-resolution components into low-resolution maps comes at the cost of the full combinatorial exploration of aligning vectors from both high- and low-resolution data, which combined to parameter fine-tuning is an hour-long process with no guarantee of success. The same quantization was used to drive flexible fitting of pre-fitted components. Later, the combinatorial complexity was reduced by introducing a 3-anchor point matching procedure, where intermediate solutions may be discarded by observing the fitting of other features within a 15 Å radius. Zhang et al. extended this approach to multi-component docking, introducing an interesting solution to the combinatorial problem and a local, average density value assigned to feature points. The combinatorial problem was solved using Iterative Closest Point, with the possibility of including external restraints. Finally, in TEMPY, vector quantization has been used to define a unique position per component. These locations are used to pre-fit up to 8 components within the target density map, priming the
search space for a genetic algorithm\textsuperscript{49}, which is reminiscent of MultiFit\textsuperscript{43} or HADDOCK-EM\textsuperscript{53}.

All these methods were developed using simulated density maps at varying resolutions, and tested on few experimental systems in a purely rigid manner. Moreover, the best predictions are sometimes selected within many independent runs\textsuperscript{37,38,44,45,49}, which suggests that several runs may be needed to obtain a valid prediction. Further than being used to drive flexible refinement or reduce search space, feature points contain a plethora of useful and unique information in their neighborhood that may be used for more than average density estimations\textsuperscript{47}. In computer vision, a group of methods relying on feature or interest points and specific description of their surroundings are known as feature descriptors. They are used in many applications (see Section 2.4), generally able to find correspondences between two images, such as an object and a scene containing it, ultimately identifying the pose of the former within the latter. They further exhibit robustness to perspective change, partial occlusion and to noise. In 3D, we may hypothesize that such methods would be able to dock high-resolution structures in low-resolution cryo-EM maps of a complex they belong to. Advantages of feature descriptors in 2D may be interpreted as robustness towards conformational changes, background noise in cryo-EM maps or conformational noise from sample heterogeneity, and the wide resolution range in experimental data (see Section 2.1).

Here, we present Macromolecular Descriptors (MaD), which takes inspiration from the 2D scale-invariant feature transform (SIFT\textsuperscript{100}). A summary of the method is shown in Figure 4.1. Shown on a range of both simulated and experimental assemblies at different resolutions and containing a variety of components representative of real-life scenarios, MaD is able locate any number of components regardless of their conformational state at their expected positions in electronic density maps. In the case where a single component is present at multiple locations in the target assembly, a single copy is enough to identify all of them, making the method largely scalable. Further, docking is achieved regardless of symmetry and the results are obtained for individual components before being assembled, meaning that partially available components will not impair the modeling process. Used with conformational ensembles, a combination of local and global scoring enables to match locations in the presence of large conformational changes and to select the state that best fit the one captured by the target cryo-EM data. As pre-generated descriptors are stored in singular files, MaD facilitates further docking efforts at the cost of generating descriptors for new structures only. As interest
4.1. Solving biomolecular puzzles with feature descriptors

Figure 4.1 – MaD: Macromolecular Descriptors for density-guided quaternary structure prediction. A visual summary of MaD is shown. Starting from a low-resolution density map and an atomic structure of one of its component, anchor points are detected in both as local density maxima. Atomic components are first converted to simulated density maps of similar resolution and voxel spacing as the assembly map. These anchors inform on how to translate components into the assembly map. Regions surrounding anchor points are integrated to provide them with rotation invariance in a step called orientation assignment. This step provides the rotation matrices necessary to orient translated components within the density. To facilitate the identification of anchor pairs leading to successful docking, the regions surrounding anchors subdivided into sub-regions, adding robustness to conformational changes and local specificity. Converting these sub-regions into a unidimensional vector yields a descriptor. Their matching, if yielding a score above a threshold \( \alpha \), enables the identification of putative docking solutions. This acts as local scoring. To identify real solutions, translated and rotated cloud of anchors may be clustered. As all anchors are considered, this acts as global scoring. The largest clusters identified this way correspond to the translation and rotation information leading to correct component localizations within the assembly. While local scoring is robust to conformational changes, the global scoring may act as a conformational selector when a structural ensemble is provided. When multiple components are fed to MaD, a simple combinatorial procedure minimizing the overlap between identified solution or maximizing cross-correlation with the target map is enough to finalize assembly prediction.
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

is growing into *in situ* structural determination through cryo-electron tomography, now routinely outputting intermediate resolution structures, MaD presents itself as an optimal and transferable tool for the dynamic integrative modeling of large macromolecular complexes in their native context.

### 4.2 Anchor point detection in 3D biomolecular structures

Due to the intrinsic variability of low-resolution cryo-EM data, the choice of interest point detectors is crucial. Most feature detectors and descriptors were designed with 2D, high-quality images in mind. Lower quality images lead to significant differences in performance of either detectors and descriptors\(^{105}\), and the effect of adding another dimension may further affect performance. The Laplacian of Gaussian (LoG) (see Chapter 2, section 2.4.2), whose approximation is used in SIFT\(^{100}\), has been used within SITUS\(^{178,179}\) although the aim was to increase the global contour of components rather than localized information extraction, as is the goal here with feature descriptors. While the approach presented in this section utilizes the LoG, a Hessian-based method similar to the one used in SURF\(^{104}\) and a 3D corner detector based on the Shi-Tomasi variant\(^{107}\) methods were also implemented in 3D and tested on the same data presented here (see Appendix). Despite their effectiveness in computer vision, namely for lower resolution images\(^{105}\), they were found to be largely inferior when used with cryo-EM density maps.

Dealing with tridimensional structures comes with the advantage that there is no need to estimate any homography between object and scene. As will be amply demonstrated in the upcoming sections, a single interest point in 3D and its associated descriptor contains enough spatial and rotational information to pose-estimate an object within a scene - here, an atomic component within a low-resolution cryo-EM map. In that sense, feature points act as anchor points and will be referred to as such here onwards. Another crucial aspect for 2D feature descriptors is the scale. An object can appear larger or smaller in the scene where is needs to be recognized, and images usually contain features that are at different scales. For instance, an image of a flower field would probably contain flowers of different sizes, with some appearing bigger or smaller due to perspective. This does not affect biostructural data as it is encoded with respect to real-world measures at the molecular scale. Hence, in the sense of a LoG-based detector (see *Figure 2.8*), it might only be a question of finding an optimal,
4.2. Anchor point detection in 3D biomolecular structures

Figure 4.2 – Systems used to test anchor detection methods. The 6 systems used to test anchor detection methods are shown here, in order of increasing resolution and with subunits colored individually. (a) PDB: 6DBL, EMD: 7845 at 5 Å. (b) PDB: 5UP2, EMD: 8581 at 6 Å. (c) PDB: 5G4F, EMD: 3436 at 7 Å. (d) PDB: 3J4K, EMD: 5751 at 8 Å. (e) PDB: 2P4N, EMD: 1340 at 9 Å. (f) PDB: 3J3U, EMD: 5609 at 10 Å

scale-invariant $\sigma$ to find anchor points reliably. However, as mentioned in previous sections, experimental maps sometimes display a large resolution range which is conveniently reported as a single number in databases and most publications. A Gaussian pyramid is a stack of sequentially blurred images, and different resolutions may be interpreted as a convolution with a larger or smaller-scaled Gaussian. We can hypothesize that building a scale-space offers better handling of the local resolution of experimental maps, in contrast to typical 2D situations where this aspect was tailored for similar objects appearing at different sizes. Thus, both multi- and single-scale approaches will be evaluated in this section.

Anchor point detection using LoG has been tested on a range of systems covering the target resolution range, i.e. 5, 6, 7, 8, 9 and 10 Å (Figure 4.2). For each system, two situations
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

Figure 4.3 – Anchor detection with Laplacian of Gaussian. (a) On top, an atomic component present in an experimental map of an assembly, shown on the bottom. The component is converted to a density grid of similar resolution and voxel spacing as the experimental map. (b) Density maps are super-sampled by a factor 2, which yield two octaves counting the original maps. A scale space is built by convolution (denoted by *) with a Laplacian of Gaussian at increasing \( \sigma \) as explained in the main text. A single scale per octave is shown here. (c) Anchors are detected as local density maxima. (d) When the component’s location in the assembly map is known, component anchors may be transformed to that location for benchmarking purposes. Repeatable anchors between the component structure and assembly map are shown in their respective colors as spheres, while smaller dots correspond to non-repeatable anchors.

were explored. In both, the high resolution structure will be translated and rotated in space to force misalignment between experimental and simulated grids to reflect real cases. The first situation will compare the transformed structure with the low-resolution model as deposited in the EMDB\(^6\), while the second one will compare the structure to a density map simulated at the same resolution of the experimental one to act as positive control. In this particular step only, whole assemblies will be considered to match the entirety of the experimental data. The anchor detection process is summarized in Figure 4.3, although a single component is shown to be in line with the subsequent steps.

As density maps are represented as grids and atomic structures as a set of coordinates, all atomic components are converted to simulated density grids of resolution and voxel spacing equivalent to the target experimental data as detailed in Section 2.1.2 (Figure 4.3a). A Gaussian pyramid is built similarly to SIFT (see Figure 2.7) (Figure 4.3b). The first octave corresponds
4.2. Anchor point detection in 3D biomolecular structures

to the original map supersampled by a factor 2 to recover higher frequency features lost to
the initial blurring of the input map. A four-fold increase increase in repeatable features was
observed through this process in SIFT\textsuperscript{100}. The second octave is the original map itself, and
further octaves did not provide any significant improvements. Reasons include the voxel
spacing becoming increasingly large, impeding the accuracy of anchor points coordinates,
and the already low resolution of the data. Over both octaves, the first scale is LoG-filtered
with scale 0.5\(\sigma_{\text{init}}\) and \(\sigma_{\text{init}}\), respectively. Subsequent scales involve \(\sigma\) values increasing with
a factor \(2^{i/s}\), where \(i\) is the index of the scale starting from 0 and \(s\) is the number of scales,
which is set to 5 and includes two auxiliary scales similar to SIFT (Figure 2.7). Each map of
the pyramid is scale-normalized with a factor \(\sigma^2\) as explained in section 2.4.2 and Eq. 2.20.
The values of both simulated and experimental grids are normalized between 0 and 1 when
loaded (an initial contour level to half the recommended value is applied by default), and the
part of the LoG-filtered maps corresponding to the external shell is discarded. Indeed, it was
previously shown that its use was negligible compared to the interior densities\textsuperscript{179} and this was
confirmed by internal testing.

An anchor point is detected if a voxel is larger than its 26 neighbours in its scale and than
the other 54 neighbors from its adjacent scales (Figure 4.3c). A notable difference with the
2D case is the use of the Hessian matrix and its eigenvalues to detect saddle points. In 2D, if
the product of the eigenvalues is negative, then their signs are opposite and a saddle point
is present. This can be done by computing the determinant of the Hessian matrix. In 3D,
all three eigenvalues need to be computed explicitly as their product can be of either sign
regardless of the anchor’s saddleness. Subvoxel accuracy is done similarly to the 2D case (Eq.
2.22). A minimum intensity of \(5e^{-2}\) is set in order to filter extrema with poor contrast.

Performance will be judged on the base of repeatability, or the fraction of anchor points
that can be found at the same location in both map instances, i.e. the simulated density map
of the fitted assembly and the experimental density map. As the fitted solution is known, an
anchor from the latter can be traced back to the original position in the experimental map
and its correspondence is confirmed if the anchor point of the simulated density is within
a distance \(d_{\text{thr}}\) of another within the experimental density (Figure 4.3d). \(d_{\text{thr}}\) is defined as
the distance between two diagonally neighbor voxels. Of note, SIFT considers a point as
repeatable if it is within \(\sigma\) pixels of its putative match in the other image. This is too large for
density maps, where the voxel spacing can be as high as 2 Å, which would result in a shift of 4 Å of the structure, before accounting for orientation errors. Valid solutions in cryo-EM-guided integrative modeling strategies usually exhibit deviations below 5 Å of RMSD, which is why we limit repeatability to the neighboring voxels only. In the cases where voxel spacing is closer to 1 Å, the results presented in this section may thus be underestimated. Repeatability is generally obtained by dividing the number of correspondences by the number of detected anchors of the smallest structure, but the result can be biased if one structure contains significantly more features than the other. Here, the following formula will be used:

\[ R = \frac{2n_{corr}}{nP_{disp} + nP_{exp}} \] (4.1)

where \( n_{corr} \) is the number of correspondences or repeatable anchors found between the displaced structure and experimental density, and \( nP_{disp} \), \( nP_{exp} \) are the number of anchor points detected in each density map. While both experimental and simulated data theoretically contain the same structure, it may be that the experimental density contains significant noise or contain structural parts not included in the fitted structure and vice-versa. This can lead to different amount of detected features, which is accounted for with this formula while punishing detectors that react too strongly to noise in experimental maps. This metric will be used for the Laplacian of Gaussian detector detailed in the next section and for the 3D corner detector and the Hessian variant shown in the appendix.

We show in Figure 4.4 the repeatability and number of correspondences in the 6 mentioned systems after multi-scale anchor detection. Values for \( \sigma_{init} \) follow the values tested in the SIFT paper\(^{100} \), plus additional values to cover a wider range. The importance of upsampled structures and maps decreases with resolution, which is to be expected because higher levels of blurring will discard high-frequency information. However, the detected anchors seem to be unique to their respective octave, as the sum of correspondences from each individual octaves match the number of correspondences when both are considered simultaneously. With experimental maps, a \( \sigma_{init} \) of 1.0 yields the best results, although the upsampled octave does not display any detected anchors at this initial scaling. There is overall a striking difference in repeatability between simulated and experimental maps, where detected anchors in the latter are typically half those in the former. A global \( \sigma \) value generalisable to the cases...
4.2. Anchor point detection in 3D biomolecular structures

Figure 4.4 – Multi-scale detection with Laplacian of Gaussian. Each row corresponds to a system. The PDB and EMD access IDs are shown, starting with PDB 6dbl and EMD 7845 at a resolution 5 Å and ending with PDB 3j3u and EMD 5609 at 10 Å. The first column shows the repeatability as defined in Eq. 4.1, for the base, upsampled and both octaves simultaneously at different values of $\sigma_{\text{init}}$. Half of that value is considered for the upsampled octave. Both octaves have 5 scales, including two auxiliary ones so that 3 scales have two adjacent scales for comparison. Results obtained with the simulated map are shown in shades of red, and those obtained with the experimental map in shades of blue. In the second and third columns are shown the number of correspondences with simulated and experimental maps, respectively. The black bars are the sum of the correspondences obtained individually in the base and upsampled octaves.
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

Figure 4.5 – Single-scale detection with Laplacian of Gaussian. Each row corresponds to a system. The PDB and EMD access IDs are shown, starting with PDB 6dbl and EMD 7845 at a resolution 5Å and ending with PDB 3j3u and EMD 5609 at 10 Å. The first column shows the repeatability as defined in Eq. 4.1, for the base, upsampled and both octaves simultaneously at different values of \( \sigma_{\text{init}} \). Both octaves have a single scale, blurred with a Gaussian of scale \( \sigma_{\text{init}} \). Half of that value is considered for the upsampled octave. Results obtained with the simulated map are shown in shades of red, and those obtained with the experimental map in shades of blue. In the second and third columns are shown the number of correspondences with simulated and experimental maps, respectively. The black bars are the sum of the correspondences obtained individually in the base and upsampled octaves.
4.3 Orientation assignment

4.3.1 The sphere division problem

Once anchors have been extracted, the next step to descriptor building is to assign an orientation to each anchor. This is critical in cases where the object is not guaranteed to be oriented in a similar way in the scene where it must be pose-estimated. The same is generally true for biomolecular structures, as independent experiments will lead to different structure orientations.

In 2D, the orientation of a feature point can be obtained by examining the gradient of each pixel in a region around it. Their orientation and magnitude can be used to compute an orientation histogram, where the bin with the highest value corresponds to the orientation of the feature point. In the case of SIFT, this is done by dividing the unit circle into 36 bins...
covering 10° each. If the value of other bins are above 80% of the maximum bin value, these bins are counted as secondary orientations, unless they are too numerous, in which case the feature point is discarded as unstable.

In 3D, the problem becomes harder for a few reasons. First, at the very least, a second angle $\phi \in [0, \pi]$ is required on top of the azimuthal angle $\theta \in [0, 2\pi]$. This has been utilized in most 3D implementation of SIFT\textsuperscript{118,119,121}, which utilize angular histograms for each dimensions. However, the area covered by bins at regular angular intervals becomes smaller towards poles than on the equator, not to mention the different aspect ratios of the bins. This approach is therefore biased and to be avoided.

An interesting alternative exists. There are five exact solutions to dividing the sphere into a number of zones of equal area, or to have a uniform distribution of points on the sphere. These are known as the Platonic solids, which include the tetrahedron, hexahedron (cube), octahedron, dodecahedron and icosahedron with 4, 6, 8, 12 and 20 faces, respectively. All faces are made of the same regular polygon. Plato associated elements to each, based on their shape (Figure 4.6a-e). The cube is associated to earth, while fire is associated to the sharp edges of the tetrahedron. The smoothness of the faces of the octahedron reminds of air, and the many faces of the icosahedron relates the little balls of water, easily flowing from the hand. Plato said of the dodecahedron that God used it to arrange constellations, and hence represents the universe. Kepler then used these solids as basis for his Mysterium Cosmographicum, where a specific order of the solids was used to create 6 spheres, each representing one of the 6 known planets at the time (Figure 4.6f). This was the second attempt to physically prove heliocentrism, after Copernicus’ initial attempt, before being abandoned in favor of an elliptical model and works towards Kepler’s three laws of planetary motion.

A 3D SIFT method\textsuperscript{120} focuses on the icosahedron and dodecahedron to build descriptors, while the orientation is determined through the eigenvalues of the anchor point’s structure tensor. This method is applied to a range of medical data, including head magnetic resonance imaging (MRI) and abdominal computed tomography (CT) images. These images are less reliant on orientation, since patients are typically oriented similarly when images are acquired. Nevertheless, orientation assignation based on the structure tensor is appealing for its computational efficiency. However, the method provides a single orientation dictated
4.3. Orientation assignment

Figure 4.6 – Platonic Solids. (a) The cube or hexahedron, associated with the earth element by Plato, with Saturn as its the outer sphere and Jupiter as its inner sphere. (b) The tetrahedron, associated with fire, and with Mars as its inner sphere. (c) The dodecahedron, used to arrange constellations and with Earth as its inner sphere. (d) the icosahedron, associated with water and with Venus as its inner sphere. (e) The octahedron, associated with air and with Mercury as its inner sphere. (f) Kepler’s model of the solar system using Platonic solids.

by the eigen decomposition of the structure tensor, and discards features when eigenvalues are too similar instead of duplicating the feature to account for these differences and build robustness towards such effects. Given the low resolution of cryo-EM maps used in this thesis, this is not a sound alternative as key solutions will be discarded.

In 2D, SIFT utilizes a 36-bin histogram for orientation assignation. In 3D, a higher number of bins than those offered by Platonic solids might be needed to accurately cover the angular space with similar completeness. We need then to find spherical division schemes that allow for higher granularities. The geodesic discrete global grid system (ISEA3H GDDS) \cite{180,181} is most often used to map the Earth’s surface for meteorological and climatological studies. Rather than the impractical mix of hexagons and pentagons of the ISEA3H system, other approaches rely on pixelating the sphere similar to how igloos are built \cite{182}, or utilize quadrilaterals \cite{183}. Specifically, the zonal equal area sphere partitioning algorithm \cite{184} is compelling for orientation assignment and descriptor building. It provides an approximate division of the sphere into any amount of quadrilaterals, which are defined by their centers in spherical coordinates \((\theta_c, \phi_c)\) as well as longitudinal and latitudinal ranges \([\theta_i, \theta_f]\) and \([\phi_i, \phi_f]\) (Figure 4.7). There are two poles in the form of circular caps, which are well suited for orientation assignment as will be highlighted shortly. The rest of the zones are quadrilateral and arranged in a number layers so that all zones are approximately equally sized, proportioned and distributed over the sphere. This zonal configuration allows for easier bin assignation than hexagon and pentagons from geodesic grid systems. We will refer to a zonal equal area sphere partition as EQ sphere here onwards.
4.3.2 Using EQ spheres to orient anchors

Orientation assignment for macromolecular descriptors is made in several steps. First, a spherical region is extracted around the anchor point in the original map, at the octave it was detected in. We show this on an atomic structure for better visualisation in Figure 4.8ab, although in-software, the simulated map is considered. Considering a spherical region is a rather than a cubic patch around the anchor avoids bias from corners that are not likely to match between the component and the map due to differing orientations. Similar to other descriptors, a Gaussian window centered on the anchor point may be applied, so that the contributions of voxels further away are diminished. This makes sense, especially if different conformations are being compared or if the resolution or data quality fades with the distance from the anchor point.

Then, the density gradient is computed at every voxel within that patch, and assigned a zone on an EQ sphere based on its direction (Figure 4.8b). The azimuth $\theta$ and polar angle $\phi$ are computed as follows:

$$\theta = \tan^{-1}\left(\frac{g_y}{g_z}\right)$$  \hspace{1cm} (4.2)

$$\phi = \cos^{-1}(g_z)$$  \hspace{1cm} (4.3)
4.3. Orientation assignment

Figure 4.8 – Orientation assignment and orientation-free referential of structures. (a) A subunit of the PDB structure 5G4F, with the region considered for orientation assignment around an example anchor. For better visualization, the atomic structure is shown rather than its simulated, LoG-filtered density map. (b) The density gradient is computed at every voxel. Only a portion of the spherical region is shown for clarity. The gradient vectors are counted based on their direction following an EQ sphere partitioning model, here comprising 32 zones. (c) The gradient vector counts for each of the EQ zone is represented by an arrow proportional to the number of vectors with corresponding direction. The first sphere corresponds to the initial counts, with the dominant zone highlighted in red. The second sphere represents the counts after the first rotation, where the dominant zone is located on the upper pole of the EQ sphere. The next zone with highest vector count is shown in orange and corresponds to the secondary dominant zone, used to assign the second rotation of the anchor. After this step, the anchor is now rotation-invariant. (d) The structure from (a) in the orientation-free referential.

where \( g_a \) is the gradient in the \( a \) direction. Each zone of the EQ sphere is defined by its center in spherical coordinates and the angular ranges it covers (Figure 4.7), making assignation to any given zone a simple range check (Figure 4.8b). Each gradient vector is thus assigned to a zone this way. The EQ zone with the highest vector count is identified as the dominant orientation of the anchor, shown as a red arrow in Figure 4.8c. Any other zone whose count is within 80% of the maximum count is considered and the anchor is duplicated accordingly. The density patch is then rotated so that the dominant orientation lies on the upper pole of the EQ zone (Figure 4.8c, middle). To reach rotation invariance in 3D, another angle needs to be determined. After the previous rotation, voxels are again assigned to EQ zones corresponding to the orientation of their gradient. The zone with the maximum vector count (both poles excluded) is selected for the secondary dominant orientation, shown as an orange arrow in Figure 4.8c. Similar to the first step, any other zone with a vector count within 80% of the maximum is considered as well, leading to further duplication of the anchor.
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

if necessary. The anchor and its region is rotated so that the selected zone is the first of its layer on the EQ sphere (Figure 4.8c, right). Of note, this process does not depend on orthogonal coordinates as with an eigen decomposition of the structure tensor, allowing for more flexibility in assigning orientations in tridimensional space. For both dominant orientations, a maximum of 6 duplications is tolerated. After applying the same rotation steps on structure, it is now in an orientation-free referential (Figure 4.8d).

Other descriptors methods typically consider the magnitude of the gradient vectors to weight their contribution to the orientation histogram. Due to the densities not being regular in experimental maps and as such potentially inconsistent with their simulated counterparts, a simple vector count was chosen instead.

4.3.3 Localizing an atomic structure within an electronic density map

At this stage, the available information is sufficient for pose-estimation. The detector provides coordinates of anchor points in both atomic structure and low-resolution experimental map, thus the information on how to translate the former to the latter (Figure 4.9, top left). Orientation assignment provides the rotation matrices to finalize the subunit’s localization. In effect, the structure is rotated according to the anchor’s dominant orientations, setting it in the orientation-free referential. For a matching location and orientation found in the density map, the same referential is achieved. Thus, the angular position of the component may be retrieved with the inverse of the rotation matrices from the map’s anchor (Figure 4.9, bottom). Combined with the previous translational vector between anchors, the structure may thus be localized in the context of the cryo-EM data using a single pair of oriented anchors (Figure 4.9, top right).

4.3.4 Benchmark

Orientation assignment relies on several parameters whose value require benchmarking. We will consider the same systems as in the previous step (Figure 4.2). The nature of the problem typically involves individual components rather than full assemblies, so subunits of different shapes and sizes were selected (Figure 4.10a). The size of the spherical patch, initially a cube due to the structures being represented as grids, depends on the voxel spacing of the
4.3. Orientation assignment

Figure 4.9 – Estimating the position of a structure within a density map using a single pair of anchors. A subunit of the PDB structure 5G4F is shown as example. On the top left, the translation obtained between an anchor of the subunit and another of the target density map of the assembly is depicted. At this location, the subunit needs to be rotated in order to fit the density (top right). This process is shown on the bottom panel. The structure is rotated to the first dominant orientation of its anchor, then to the second one. From this orientation-free referential, the final position of the subunit may be obtained by applying the reverse operations from the orientation assignment of the target map’s anchor. Thus, anchor detection provides the translation vector and orientation assignment the rotation matrix.

Experimental map (Figure 4.10b). While this will remain a free parameter, it is important to test the robustness of the method to different voxel spacing values. In general, the voxel spacing of recent maps tend to be between 1 and 1.5 Å, while cases beyond 2 Å are rarer nowadays but still occur. Thus, the smallest patch size will be a cube of 13 voxels in length, which amounts to 18 Å (without counting the central voxel) for a voxel spacing of 1.5 Å, enough to span an entire α-helix and parts of other close-by structural elements. The largest size considered here will be a patch of 17 voxels in length, which amounts to 32 Å for assemblies with a voxel spacing of 2 Å, which should be enough to cover a patch of secondary structure elements without exceeding the size of smaller components. Choosing smaller sizes may not even span a single secondary structural element, which will lose in specificity and accuracy.
Figure 4.10 – Benchmark assemblies and relevant parameters for orientation assignment.

(a) The subunits and respective assemblies used in the benchmark, showing the PDB access code and the selected chain as well as the EMD number, resolution and voxel spacing of the corresponding experimental map. (b) Different sizes of the patch extracted around an anchor. Values shown refer to the diameter, including the planes crossing at the location of the anchor. A commonly found voxel spacing of 1.5 Å was used to render the patch. (c) The mask used to extract the spherical patch from the cubic region is shown. Each image is a slice from top to bottom, where the considered voxels are colored in red. Each contribution may be weighted by a Gaussian function centered on the anchor, so that contribution for points far from the center are diminished. The $\sigma$ of the Gaussian depends on the one used in the anchor detection step. (d) The different EQ sphere sizes considered.
Larger sizes may become problematic if the anchor is located close to an interface with other subunits in the complex, as orientation assignment will be biased by the absence of densities beyond the subunit’s surface, while that space would be occupied by other subunits in the experimental map of the complex. An obvious way to alleviate this bias and standardize the patch size would be to interpolate voxels at a fixed voxel spacing. However, this process lead to generally lower detector and orientation assignment performance, potentially due to interpolation artefacts and/or aliasing depending on the direction of interpolation.

The contribution of voxels within the spherical patch may be further weighted by a Gaussian window centered on the anchor. The $\sigma$ is dependent on the value used in anchor detection as in SIFT, which uses $1.5 \times \sigma$. This value was deemed too punitive, and factors of 2 and 3 will be tested instead along with the spherical mask alone (Figure 4.10c). The size of the EQ sphere used to count gradient vectors is probably the most crucial component. If set too small, the orientation may be too coarse and solutions may be too dependent on the initial orientation of the atomic structures. Conversely, a dominant orientation may be split across several zones if the EQ sphere contains too many, preventing satisfactory orientations from being derived. A set of sizes has been chosen so that different amount of layers between the two poles are considered (Figure 4.10d). To ensure that the process is not biased by the initial alignment between the atomic structure and the experimental map, each subunit was translated so that its center of mass is located at the origin and then rotated equally about the X, Y and Z axes with a step of 15° for a full revolution.

The different parameter sets were tested as follows. For each combination of subunit and experimental map shown in Figure 4.10a, the pairs of repeatable anchors are extracted using the LoG-based detection procedure previously described. Each of these anchors is then assigned an orientation, and the subunit is localized within the experimental map following the procedure summarized in Figure 4.9. Then, the localization is scored. The root mean square deviation (RMSD) is a common measure to estimate differences between structures. While it may be used effectively in general, it is most accurate on well-aligned globular structures. In our case, if the anchor is located on one side of an elongated structure and a small orientation shift exists between the localized subunit and the reference, the RMSD between the two will be dominated by the errors from atom pairs located far from the anchor, rendering this score less useful (Figure 4.11). Such errors are expected due to the coarseness of the EQ sphere.
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

Figure 4.11 – Bias from root-mean square deviation. The chain F of PDB structure 5UP2 is shown in orange. An example anchor is shown as a large red sphere. A copy of the structure is shown in blue after rotation about the anchor at a 15° angle. The root-mean square deviation (RMSD) is reported for two selections. The first includes all Cα atoms from the first domain and leads to an RMSD of 3.16 Å. The second includes all Cα atoms instead and lead to an RMSD of 8.35 Å due to the larger errors of the second domain, where the effect of the rotation is stronger.

Alternate scores have been developed to alleviate this bias\(^{185}\). Here, a different approach is adopted. As only repeatable anchors are considered, the translation component is valid by definition. Thus, only scoring the rotation component between the localized subunit and the reference is necessary. This can be obtained through the Kabsch algorithm\(^{186}\), which is also used in VMD\(^{86}\) to superimpose structures before computing the RMSD between them. After recentering the structures by subtracting the mean of their coordinates, we can obtain the optimal rotation matrix between the localized subunit and the fitted reference through singular value decomposition (SVD):

\[
C_{\text{localized}}^T C_{\text{fitted}} = U S V^T \tag{4.4}
\]

where \(C_{\text{localized}}\) and \(C_{\text{fitted}}\) are the centered coordinates of the localized subunits and its fitted reference in the complex, \(U\) and \(V\) rotation matrices and \(S\) a coordinate scaling matrix. The optimal rotation matrix \(R\) to bring the localized subunit on top of the fitted reference can
4.3. Orientation assignment

be obtained by multiplying the matrices $U$ and $V$:

$$R = V \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & d \end{bmatrix} U^T \quad d = \text{sign}(\det(UV^T))$$  \hspace{1cm} (4.5)

where $d$ ensures that the right-handed coordinate system is in use. The Euler’s rotation theorem states that any combination of rotations is equivalent to a single rotation against a single axis. That angle may be obtained with the following formula:

$$\theta = \arccos\left(\frac{\text{Trace}(R) - 1}{2}\right)$$  \hspace{1cm} (4.6)

Thus, the validity of subunit localization obtained from orienting a pair of repeatable anchors may be estimated by $\theta$. Here, we accept the localization if $-\frac{\pi}{3} < \theta < \frac{\pi}{3}$, which is a range where a fast and coarse rigid-body refinement procedure such as the one described in section 2.1.3 will converge easily if refinement is needed. We record, for each parameter set, the percentage of anchor pair that yield valid localizations as well as the minimum RMSD-Cα obtained over all angles.

The benchmark results are summarized in Figure 4.12. The best sets highlighted by the above procedure are shown in Figure 4.12a. Sub-3 Å localizations are easily found with many combination of parameters, with about 50% of repeatable anchor pairs leading to valid localizations. Larger patch sizes yield better results and 112 EQ zones rank best for both, while the smallest EQ sphere size of 32 zones consistently rank lowest. This is expected given that SIFT uses 36 bins in 2D. The lower scores obtained by a 128-zone EQ sphere suggest that the additional zones and layer may be detrimental to orientation assignment, probably because the relevant gradient information covers larger area and is thus split between different zones. While the application of a Gaussian window on the spherical patch does affect the minimum RMSD positively, the percentage of valid anchor pairs seems to suffer when it is applied. For both subunits and assemblies, an EQ sphere size of 112 zones combined with a patch radius of 7 voxels and no Gaussian window yield the best results, although the patch size of 15 voxels is statistically similar. These parameters will be used here onwards and by default. We further show the stability of the two scores over all angles considered (Figure 4.12bc). Both scores are
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

Figure 4.12 – Benchmark results for orientation assignment. (a) Rankings of parameter sets, including the patch diameter in voxels, the number of EQ zones and the relative $\sigma$ of the spherical mask, expressed as a multiple of the value used in the detection step, here $\sigma = 2$. The worst parameter sets are shown as white bars. (b-c) The minimum RMSD-C$\alpha$ and fraction of valid anchor pairs are shown for each angles between 0 and 360° with a step of 15°. The best-scoring set is shown in red, and the worst in black. (c) The average minimum RMSD-C$\alpha$ achieved per parameter value per subunit.

stable under any rotation, with the exception of 5UP2-F and 3J3U-A, which may be attributed to the smaller size and lower resolution, respectively, or to their elongated shape as mentioned previously in Figure 4.11. This is a crucial aspect, as there is no way to estimate the relative orientation between an cryo-EM map and the atomic structure of its components. Moreover, even for the worst parameters sets, relatively good and stable results were still obtained, which demonstrate the robustness of orientation assignment using EQ spheres.

We performed a similar benchmarking procedure on a set of 60 assemblies with pre-fitted PDB structures. They were chosen so that each resolution between 5 and 10 Å was equally represented (Figure 4.13a). The peak of the voxel spacing distribution is just short of 1.5 Å (Figure 4.13b). The same EQ sphere size and absence of Gaussian window highlighted for
4.4. Generating robust and discriminative descriptors

While we have shown in Figure 4.9 that the information derived at this stage is enough to localize subunits, this was only achieved because matching anchor pairs were known. In practice, for a subunit with 100 anchors and a density map with 1000 anchors, there are 100’000 possible pairs to compare before including the duplicates generated in orientation assignment. Generating descriptors from oriented anchors aims at facilitating the identification of valid pairs through the generation of descriptors. These take the form of unidimensional vectors summarizing the orientation-free density patch around anchors so that robust matching may occur despite noise or conformational differences while discarding as many false matches as possible.

Again, our procedure will follow the principles behind SIFT descriptors. This section will build on the example shown in the previous section (Figure 4.9). After orienting the structure, the patch around the feature is resampled (Figure 4.14a), this time without masking out the corners. When paired with a matching anchor, they are no longer a source of bias.
after orientation assignment and may enhance the discriminative power of descriptors. This resampling step is carried out using an interpolator generated during anchor detection to avoid rotating the whole structure or map for each anchor, saving considerable computing time. New voxel positions are interpolated so that no voxels are on any of the three 2D planes crossing the central voxel (Figure 4.14b). This enables the subdivision of the density patch into an array of descriptor regions (Figure 4.14c). This increases the discriminative power of descriptors by enforcing that information is not only present in the density patch, but at a specific location around the anchor. Moreover, this provides considerable robustness to conformational differences between the atomic structure and experimental map. As long as the information is present roughly in the same region, it will be assigned to the same EQ sphere and considered in the same way. A coarser EQ sphere is also used here to provide robustness to changes in gradient direction and avoid overfitting, as each sub-region contains significantly less voxels. Finally, the gradient counts for each zone of each EQ sphere may be concatenated into a single, unidimensional vector for efficient descriptor matching. The final descriptor is a unidimensional vector with $n \times r$ values, where $n$ is the number of sub-patches and $r$ the number of zones in each of the $n$ EQ spheres.
4.4. Generating robust and discriminative descriptors

Figure 4.15 – Benchmark for descriptor generation. (a) The density patch around an oriented anchor may be kept as is, or divided into 8 (2x2x2) or 64 (4x4x4) subregions after an interpolation step. (b) The different number of zones of the EQ sphere. Smaller sizes than in orientation assignment are tested due to the lower amount of voxels per subregion. (c) On the left, the average area under the receiver operating characteristics curve (AUROC) is shown for each subdivision scheme. On the right, the average AUROC computed in the same way but for the number of zones on each EQ sphere. *p*-values below 0.001 are indicated by ***, while non-statistically significant differences are denoted by *ns*. (d) The AUROC accomplished per subunit for a 4x4x4 subdivision scheme with 16 EQ zones compared to a descriptor without subdivision and an EQ sphere with 8 zones, corresponding to the worst performance achieved.

4.4.1 Benchmark

For each possible normalized descriptor pair, a cross-correlation score may be computed with a dot product, resulting into a scalar between 0 and 1 that can be interpreted as a probability of that descriptor pair being valid. Ideally, the best type of descriptors should fully prioritize positive descriptor pairings over negative ones. This can be done by computing the area under the receiver operating characteristics (AUROC) between the success pair labels from orientation assignment and the corresponding cross-correlation just computed. The advantage of this measure is that it does not depend on the confidence threshold nor is it affected by class imbalance, which in this case is strongly skewed towards negative descriptor pairs. An AUROC of 1 means the model discriminates perfectly positive and negatives samples, while an AUROC of 0.5 means the model does not make any such distinction and an AUROC of 0 means the model is prioritizing negative samples over positive ones.

Following the results of orientation assignment, a patch of 17x17x17 voxels interpolated to
one of 16x16x16 is used here exclusively. The number of sub-patches after interpolation and the number of EQ zones were benchmarked (Figure 4.15ab) for the same angular range and same structures as for the orientation assignment step. As expected, the descriptor without sub-regions performs worse while the 4x4x4 subdivision is the most distinctive (Figure 4.15c, left). However, results are less clear-cut for EQ zones, as differences beyond 8 zones are not statistically significant (Figure 4.15c, right). However, with a subdivision into 64 regions, each measuring 4 voxels in length, only 64 vectors are summarized in each EQ sphere. As such, 48 EQ zones may be prone to overfitting and not as robust to large conformational differences. Instead, 16 zones will be used. As for orientation assignment, the performance is stable over all angles tested (Figure 4.15d). The lower performance of 2P4N-A, which corresponds to a tubulin molecule as part of a larger filament, may be explained by the higher voxel spacing of the experimental map at 2 Å and its smaller size at 139 kDa. The patch size of 16 voxels leads to most descriptors bleeding over the density of nearby subunits in the assembly map (see Figure 4.2e), which is not present in the fitted PDB structure, impacting both orientation assignment and description.

### 4.5 Descriptor matching, scoring and post-processing

At this stage, we have a list of descriptor pairs that may lead to valid structure localizations within density maps by setting a confidence threshold on descriptor cross-correlation (Figure 4.16a). However, a single descriptor only covers a small portion of the structure, and thus is only a local score. Furthermore, it is expected that multiple descriptors lead to the same localization, which is amplified by feature duplication during orientation assignment. This multiplicity however may be put to good use, as valid localizations will be enriched whereas wrong pairs should lead to singular localizations instead, yielding an additional means of discarding false positives. As such, it is necessary to define a measure that scores descriptor matching on the scale of the whole structure rather than for smaller regions.

In 2D, it is common to use a joint outlier detection and clustering method, such as the Hough transform or random sample consensus (RANSAC$^{127}$, see section 2.4.3) to highlight the converging localization while estimating the homography between object and scene. However, RANSAC performs poorly when the number of inliers is below 25%. In section 4.2, we highlighted repeatability values ranging from 20% to 50% for experimental maps. In
4.5. Descriptor matching, scoring and post-processing

Figure 4.16 – Local and global descriptor matching. (a) For each anchor detected in a high-resolution component and a low-resolution density map containing the latter, descriptors are generated in the form of unidimensional vectors. Cross-correlating descriptor pairs yield a probability of a pair being a valid one. Pairs above a confidence threshold are selected for the next step. This constitutes local descriptor scoring. (b) Anchors from the descriptors pairs validated in (a) are localized in the assembly map according to the translation and rotation information from the detection and orientation assignment step. $R_c$ and $R_a$ refer to the rotation matrices from orientation assignment of the component and assembly anchors, respectively. The percentage of component anchors having a close neighbor in the target map is used to rank descriptor pairs. Example of matching pairs are circled in black, while lone anchors are circled in red. This constitutes global descriptor scoring. (c) Similar anchor localizations within the assembly map are clustered together. Anchors from the high-resolution component are color-coded by index for better visualization of their cloud’s orientation. (d) Final localization of structural components is obtained from the highest ranking clusters. Here, all 6 localizations of the single component in (a) are obtained simultaneously.
orientation assignment, a similar success rate was obtained, which yields a maximum of around 25% of inliers for the cases shown.

Thus, a different approach was designed based on progressive sample consensus\textsuperscript{128} (PROSAC), which orders samples using a quality function instead of relying on random sampling. RANSAC and PROSAC select a number of anchor pairs from both object and scene matching their dimensionality, estimate the homography and count inliers. Here instead, we found that using the translation and orientation information of each anchor provided better results than simply relying on a set of anchors. Essentially, the anchors leading to putatively matching descriptors are localized within the assembly map identically to Figure 4.9 and the fraction of inliers (repeatability) is computed (Figure 4.16b). This is done for each descriptor pair similarly to the benchmark of anchor detection and is a relatively fast process using a KD-tree.

By sorting each descriptor pair with this score, we can then proceed to a coarse, hierarchical clustering step, under the assumption that the best localizations will be ranked higher than others. The first localization is used as seed for the first cluster (Figure 4.16c). Then, the RMSD between the anchors from the second localization is compared to the anchors of the first localization. If it is below a threshold, the cluster is grown by 1, otherwise, a new cluster seed is created. This step is repeated until all selected descriptor pairs have been processed exactly once. This kind of clustering is sensitive to the initial ordering of the samples, which was here previously optimized both locally and globally to limit any bias. Moreover, this score will account for anchors that did not lead to matching descriptors but are repeatable between component and assembly map, making the most out of the data generated so far.

Scoring clusters is made by multiplying the repeatability of the seed with the cluster's population, here termed weight as it relates to the number of descriptors leading to that particular location in the assembly map:

\[
RW = \text{repeatability} \times \text{weight}
\]

The RW score ensures that the inliers also correspond to valid descriptor pairs and it was found more effective than either repeatability or weight alone. The cluster seeds with the best
score are localized within the assembly map as previously done (Figure 4.16d), yielding the final set of solutions. A major advantage of this method is that a single component will be localized at all its locations within the assembly map. Indeed, each component location within the assembly map has descriptors that will match similarly with those of the high-resolution structure, and this regardless of symmetry.

4.6 Docking and refining macromolecular components in EM maps

We show all macromolecular complexes previously mentioned after being fitted with MaD in Figure 4.17 and Figure 4.18, with results summarized in Table 4.1 at the end of this section. In all cases, individual components were extracted, randomly rotated and their geometric center was translated to the origin to limit bias from the initial positions of the components.

We first focus on the recombination-activating gene 1 and 2 (RAG1-RAG2) endonuclease complex (RAG) to illustrate the results at this stage. The RAG complex (Figure 4.17a) is involved in the adaptive immune response of vertebrates through the initiation of the V(D)J recombination and DNA cleavage. PDB 6DBL and the corresponding 5 Å map EMD-7845 contains two copies of each RAG1 and RAG2 along with recombination signal sequences (RSS) DNA molecules, which guide the RAG complex to the V, D and J gene segments of immature B and T cells. Docking a single copy of RAG1 successfully identified its two locations in the complex (Figure 4.17b). The RMSD-Cα of the second copy however is impacted by the Zinc finger and DNA-binding domains on the lower part of the structure, which is reverted with respect to the first monomer (Table 4.1).

However, while the localization is correct, the fit with density is not optimal. Displacements originating from the positions of the anchors that differ between RAG1 and the assembly map and from the coarseness of the EQ sphere can be observed (Figure 4.17b and e). On top of method bias, such displacements also originate from the resolution and quality differences between the component’s simulated map and the experimental assembly map. Thus, a quick refinement step is warranted, more so in cases where such displacements induce structural clashes with neighbor subunits. The procedure available in Chimera under the name Fit in map (see also Section 2.1.3)) has been reimplemented and integrated in MaD to correct this (Figure 4.17c, f and Table 4.1). After this fast procedure, it is now possible to compute a
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

Figure 4.17 – Local rigid refinement integrated in MaD. The RAG complex (PDB-6DBL and EMD-7845) is shown. (a) The fitted PDB and experimental map as found in their respective database. RAG1 protomers are shown in orange, RAG2 in red, and RSS DNA molecules in gray. (b) The RAG1 protomers docked by MaD prior to refinement. Close-up views of regions with imperfect fit are shown on the right. (c) Better fit to density and reference protomers after the refinement procedure mentioned in the main text. (d-f) The same panel organisation is shown for the RAG2 protomers.

cross-correlation between the final fit and the assembly map.

The repeatability previously derived is updated as well, as more component anchors may have a neighbor in the assembly map as a consequence of the refinement. False positives may then have reduced repeatability as well, increasing overall robustness. Furthermore, if two equivalent solutions were beyond the threshold for merging during the clustering step, they will be within range as well, thus preventing duplicate solutions. In such a scenario, their respective weight is summed and the best cluster seed is kept. The final solutions post-refinement are then ordered according to an updated RW score, which now includes the
updated repeatability, weight, and the newly computed cross-correlation:

\[
\text{RWC} = \text{repeatability}_{\text{updated}} \times \text{weight}_{\text{updated}} \times \text{CC} \tag{4.8}
\]

We now look at a triheteromeric NMDA receptor bound to FAB (PDB 5UP2, EMD-8581) solved at 6 Å\textsuperscript{188}. Different subunit combinations within NMDA receptors are the source of their diversity in function in post-synaptic responses and location in the brain. In the aforecited study, the NMDA receptor is comprised of two GluN1 subunits (with an RMSD-C\(\alpha\) of 2.56 Å between them), a GluN2A subunit, and GluN2B subunit attached to a highly specific Fab, used to aid in the reconstruction of the receptor from cryo-EM data\textsuperscript{188} (Figure 4.18a). The density of the Fab is resolved at lower resolution than the receptor itself, and thus constitutes an interesting test case for the method. All components were successfully docked within the assembly map with RMSD-C\(\alpha\) values between 0.36 and 1.68 Å (Figure 4.18b). Of note, both copies of GluN1 may be obtained by docking either, despite the conformational differences between them (Figure 4.18c). In all cases, the correct locations were ranked before any other solutions by the RWC score.

To further exemplify the ability of MaD to find all possible locations of a component within an assembly map, we show the Actin:Tropomyosin filament structure previously solved at 8 Å with 5 copies of F-actin\textsuperscript{189}. The five best solutions from MaD correspond to all possible locations of F-actin within the density map (Figure 4.18de). Similar to SIFT descriptors being robust to partial occlusion and perspective changes, MaD is robust if part of a component is not present in the density and if conformational differences are present, as shown in an 7 Å map of the VAT complex\textsuperscript{190} (Figure 4.18fgh). In this case, the chosen component exhibits RMSD-C\(\alpha\) ranging from 2.18 to 3.05 Å with the other protomers. The tubulin-\(\alpha\) and tubulin-\(\beta\) are docked with less than 1 Å RMSD-C\(\alpha\) with their reference and the kinesin within 1.7 Å in the corresponding 9 Å map\textsuperscript{191}. MaD also identified tubulin-\(\alpha\) within tubulin-\(\beta\)’s density and vice-versa, regardless of the presence of taxol, GTP and GDP molecules around tubulin-\(\beta\) (data not shown). Finally, the limit of the method is reached with the MecA-ClpC complex at 10 Å of resolution\textsuperscript{192}. While the ClpC protomers are all docked from a single copy (Figure 4.18k). RMSD-C\(\alpha\) values above 2.5 Å are obtained after refinement, certainly due to the lack of reliable density in the experimental map (Figure 4.18k). Moreover, the MecA subunits are
Figure 4.18 – Piece-by-piece docking with MaD. For each assemblies, components docked by MaD are shown in color while reference pre-fitted structures are shown in dark gray. (a) Triheteromeric NMDA receptor with Fab bound to GluN2B. Notice the absence of density around Fab at this contour level. (b) Comparison of MaD results with reference PDB. (c) Both locations of GluN1 protomers may be obtained from either of the two conformations present in the receptor. (d) F-actin:tropomyosin filament. (e) The F-actin protomers localized by MaD compared to the reference structure. (f) The VAT complex. Notice the absence of density for part of the red protomer. (g) Comparison with the reference structure. (h) Alternate view showing the lack of density and conformational differences between the docked protomer and the reference. (i) Microtubule structure with kinesin. (j) Comparison with the reference structure. (k) The MecA/ClpC complex. (l) Top: close-up view of the density squared in the previous panel, highlighting the lack of definition of secondary structure elements. Bottom: reference structure of a MecA monomer, which MaD was unable to localize within the density.
4.6. Docking and refining macromolecular components in EM maps

Figure 4.19 – Whole-structure docking with MaD. (a) Left: RMSD-Ca obtained for each of the 60 assemblies tested both pre-refinement (gray) and post-refinement (red). Right: similar distributions for the rank of each solution. (b-c) Map and pre-fitted structures of the two failed assemblies, EMD-1894 and PDB 2YGD as well as EMD-0876 and PDB 6LFG, respectively. A close up is shown in both cases to highlight the poor definition of secondary structure elements or absence of density, leaving significant parts of the pre-fitted structure unmatched.

too small to be located by MaD at such low resolution. Secondary structure elements may not be reliably distinguished, forming larger blobs instead (Figure 4.18l) that provide little to no support for macromolecular descriptors.

The same limits may be observed from docking the 60 assemblies previously mentioned in section 4.3.4. Pre-refinement, most assemblies are within 6 Å of the reference but within 1 Å after the rigid refinement step, and consistently rank on top (Figure 4.19a). When the expected solution is not ranked first, the assembly in question may be symmetrical, thus many equivalent results may be ranked before the specific component order from the reference structure (not shown). Out of these 60 assemblies, only two fail to be docked by MaD (Figure 4.19bc). Both exhibit densities that do not distinguish α-helices from β-sheets or even individual secondary structure elements, thus providing little to no support for building descriptors. Furthermore, many structural elements of the fitted atomic models do not have matching experimental densities, further impeding docking with MaD.
Table 4.1 – Summary of MaD applied to a set of experimental cryo-EM maps. The name of each system and component is shown alongside the EMD and PDB access IDs. The molecular weight, time for generating descriptors, matching and refinement are shown. When a component is repeated multiple times in the complex, only one protomer is reported as a single copy is used. The RMSD-\(C_{\alpha}\) is reported before and after refinement, and the total time to fit all component inside each cryo-EM map is shown in the last column. * The RAG1 protomers differ in the conformation of their Zinc finger domain, hence the higher RMSD-\(C_{\alpha}\).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RAG complex @ 5 Å</td>
<td>7945</td>
<td>6DBL</td>
<td>450</td>
<td>73.8</td>
<td>28.2</td>
<td>0.6</td>
<td>0.5</td>
<td>2.71</td>
<td>0.54</td>
</tr>
<tr>
<td>RAG1</td>
<td>131</td>
<td>119.7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>*11.09</td>
<td>*10.51</td>
</tr>
<tr>
<td>RAG2</td>
<td>59</td>
<td>15.9</td>
<td>0.4</td>
<td>0.3</td>
<td>1.74</td>
<td>0.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RAG1</td>
<td>1.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.56</td>
</tr>
<tr>
<td>NM1A receptor @ 6 Å</td>
<td>8581</td>
<td>5UP2</td>
<td>415</td>
<td>59.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>127.1</td>
</tr>
<tr>
<td>GluN1</td>
<td>94</td>
<td>13.6</td>
<td>1.2</td>
<td>4.1</td>
<td>2.53</td>
<td>0.84</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GluN2A</td>
<td>93</td>
<td>13.5</td>
<td>1</td>
<td>6.1</td>
<td>5.5</td>
<td>1.68</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GluN2B</td>
<td>94</td>
<td>13.2</td>
<td>1.1</td>
<td>3.7</td>
<td>4.01</td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fab</td>
<td>41</td>
<td>6.1</td>
<td>0.5</td>
<td>3.6</td>
<td>1.5</td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>VAT complex @ 7 Å</td>
<td>3436</td>
<td>5G4F</td>
<td>486</td>
<td>75.9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>99.9</td>
</tr>
<tr>
<td>VAT protomer</td>
<td>81</td>
<td>13.9</td>
<td>6.4</td>
<td>3.7</td>
<td>2.1</td>
<td>0.16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Actin/tropomyosin @ 8 Å</td>
<td>5751</td>
<td>3J4K</td>
<td>234</td>
<td>30.7</td>
<td>6.1</td>
<td>0.42</td>
<td>2.4</td>
<td>2.64</td>
<td>1.01</td>
</tr>
<tr>
<td>F-actin</td>
<td>42</td>
<td>6.1</td>
<td>0.42</td>
<td>2.4</td>
<td>3.01</td>
<td>1.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.28</td>
<td></td>
<td></td>
<td>1.29</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.44</td>
<td></td>
<td></td>
<td>1.55</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>4.3</td>
<td></td>
<td></td>
<td>1.19</td>
</tr>
<tr>
<td>Tubulin-kinesin @ 9 Å</td>
<td>1340</td>
<td>2P4N</td>
<td>139</td>
<td>13.4</td>
<td>3.3</td>
<td>6.2</td>
<td>3.2</td>
<td>2.17</td>
<td>0.48</td>
</tr>
<tr>
<td>Tubulin-alpha/beta</td>
<td>50</td>
<td>3.3</td>
<td>6.2</td>
<td>3.2</td>
<td>3.92</td>
<td>0.69</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Kinesin</td>
<td>36</td>
<td>2.9</td>
<td>0.1</td>
<td>0.2</td>
<td>3.52</td>
<td>1.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meca-ClpC complex @ 10 Å</td>
<td>5609</td>
<td>3J3U</td>
<td>696</td>
<td>48.5</td>
<td>9.5</td>
<td>5.7</td>
<td>6.117</td>
<td>5.00</td>
<td>2.64</td>
</tr>
<tr>
<td>Meca</td>
<td>26</td>
<td>1.6</td>
<td>0.8</td>
<td>1.5</td>
<td>n/a</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ClpC</td>
<td>90</td>
<td>9.5</td>
<td>5.7</td>
<td>6.117</td>
<td>5.68</td>
<td>2.67</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.72</td>
<td>2.90</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5.81</td>
<td>2.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6.43</td>
<td>3.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>8.41</td>
<td>3.48</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.7 Building assemblies from individually fitted components

While we advise using MaD for each subunit individually for a better control and validation of solutions, especially when there are large conformational differences, it is possible to build assemblies at practically no cost from the individual solutions. As they are already valid in themselves, there is no need to design a complex scoring function. Instead, we may simply perform a combinatorial exploration of the individual solutions, each in the appropriate amount making up the full assembly, while minimizing the overlap between them. We applied MaD to a set of 21 simulated assemblies used as benchmark set in another study at the same simulated resolution of 10 Å (Table 4.2). For all cases, we report the number of unique and total components separately, as only a single copy of each unique one needs to be processed by MaD. The results shown and the assembly building process however are based on the assumption that the number of copies is known a priori. If not, the building process is left to the user from the results of the each single docking steps as in the previous section. The RMSD-Cα is shown for the whole assembly, and compared to the best predicted and best scoring models of two previous methods. Of note, we report the results of γ-TEMPy after refinement with Flex-EM. MaD solves all cases within about 1 Å and clearly outperforms competitors in terms of either best predicted or best scored models. There is no need for a distinction between these two categories for MaD as it produced one single assembly in most cases and ranked the correct assembly first in others, with two exceptions (Table 4.2). Only 1Z5S and 1MDA required looking at other models than the best scoring one. In the former case, a subunit is wrongly rotated but placed correctly (Figure 4.20a) and in the latter case, a small component could not be located at all, despite the other copy being successfully placed (Figure 4.20b). We hypothesize that the reason behind this failure is that the component is small and surrounded by other subunits in the complex, which introduces bias in orientation assignment. Setting a smaller patch size may reduce the bias from the absence of density around the isolated subunit and from the neighbouring components in the assembly. Doing so rescued the failure of 1MDA (Figure 4.20c). All in all, we remind that MaD is not consistently applicable for resolutions above 10 Å (Figure 4.18l) due to its reliance on local, secondary structure information, in which situations the other methods thus remain superior.

For both experimental and simulated cases, total runtime is within 30 seconds to 2 minutes, depending on the molecular weight of the assembly. The bottleneck lies in descriptor
Table 4.2 – Summary of MaD applied to a set of simulated cryo-EM maps. The PDB, the number of unique chains and total chains is reported for 21 different systems. We report the number of successfully docked components and the RMSD-Cα on the whole assembly and the value obtained by the best prediction (BP) and highest scoring (HS) assembly obtained by competitors. We then report the rank of the best assembly obtained by MaD and the total number of models that were generated. The runtime is shown in the last columns for the building of descriptors for the assembly map, the components (which include descriptor matching and refinement), and assembly building. * run with a reduced patch size.

<table>
<thead>
<tr>
<th>PDB</th>
<th># Components</th>
<th>Unique</th>
<th>Total</th>
<th>Success</th>
<th>RMSD-Cα</th>
<th>Map Comp. Build Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1CS4</td>
<td>4</td>
<td>3/3</td>
<td>0.51</td>
<td>2.6</td>
<td>3.50</td>
<td>2.8 2.80 1/1 8.19 12.36 0.29 20.84</td>
</tr>
<tr>
<td>1GPQ</td>
<td>2</td>
<td>4/4</td>
<td>1.01</td>
<td>1.7</td>
<td>2.20</td>
<td>0.6 0.60 1/1 6.02 6.17 0.38 12.57</td>
</tr>
<tr>
<td>1SUV</td>
<td>3</td>
<td>6/6</td>
<td>0.34</td>
<td>5.2</td>
<td>5.30</td>
<td>n/a n/a 1/3 27.98 29.14 1.18 58.29</td>
</tr>
<tr>
<td>1TVQ</td>
<td>7</td>
<td>7/7</td>
<td>0.45</td>
<td>19.1</td>
<td>18.80</td>
<td>16.3 37.70 1/6 20.82 36.38 1.06 58.28</td>
</tr>
<tr>
<td>1VCB</td>
<td>3</td>
<td>3/3</td>
<td>0.65</td>
<td>1.8</td>
<td>2.20</td>
<td>4.8 25.50 1/1 3.98 6.50 0.32 10.80</td>
</tr>
<tr>
<td>2B8K</td>
<td>2</td>
<td>4/4</td>
<td>0.70</td>
<td>2.1</td>
<td>2.40</td>
<td>11.3 15.30 1/1 10.48 9.41 0.60 20.19</td>
</tr>
<tr>
<td>2BO9</td>
<td>2</td>
<td>4/4</td>
<td>0.49</td>
<td>1.3</td>
<td>1.60</td>
<td>4.0 10.30 1/1 11.66 10.36 0.27 22.30</td>
</tr>
<tr>
<td>2DQJ</td>
<td>3</td>
<td>3/3</td>
<td>0.90</td>
<td>2.0</td>
<td>2.50</td>
<td>0.7 1.70 1/1 3.93 7.92 0.48 12.33</td>
</tr>
<tr>
<td>2GC7</td>
<td>4</td>
<td>4/4</td>
<td>0.46</td>
<td>1.3</td>
<td>2.00</td>
<td>10.9 23.30 1/1 9.13 13.79 0.30 23.23</td>
</tr>
<tr>
<td>2HUX</td>
<td>2</td>
<td>4/4</td>
<td>0.35</td>
<td>1.1</td>
<td>1.50</td>
<td>n/a n/a 1/1 20.56 16.75 0.46 37.76</td>
</tr>
<tr>
<td>2NYY</td>
<td>1</td>
<td>3/3</td>
<td>0.30</td>
<td>0.9</td>
<td>1.30</td>
<td>n/a n/a 1/1 22.55 13.93 0.67 39.15</td>
</tr>
<tr>
<td>2YTH</td>
<td>3</td>
<td>3/3</td>
<td>0.26</td>
<td>1.7</td>
<td>2.20</td>
<td>n/a n/a 1/2 19.08 28.68 0.77 46.74</td>
</tr>
<tr>
<td>3LIJ</td>
<td>5</td>
<td>5/5</td>
<td>0.41</td>
<td>9.0</td>
<td>9.30</td>
<td>n/a n/a 1/4 43.10 86.83 1.32 131.26</td>
</tr>
<tr>
<td>2NVQ</td>
<td>2</td>
<td>2/2</td>
<td>0.33</td>
<td>0.9</td>
<td>1.00</td>
<td>n/a n/a 1/1 31.18 23.46 0.75 55.40</td>
</tr>
<tr>
<td>3PDU</td>
<td>1</td>
<td>4/4</td>
<td>0.63</td>
<td>1.3</td>
<td>1.60</td>
<td>n/a n/a 1/1 9.99 5.57 0.33 15.89</td>
</tr>
<tr>
<td>3PUL</td>
<td>4</td>
<td>5/5</td>
<td>0.39</td>
<td>22.7</td>
<td>23.40</td>
<td>n/a n/a 1/1 23.43 31.11 0.55 55.99</td>
</tr>
<tr>
<td>3R5D</td>
<td>1</td>
<td>3/3</td>
<td>0.53</td>
<td>1.4</td>
<td>1.90</td>
<td>n/a n/a 1/1 9.42 5.67 0.31 15.39</td>
</tr>
<tr>
<td>3SFD</td>
<td>4</td>
<td>4/4</td>
<td>0.44</td>
<td>1.4</td>
<td>1.50</td>
<td>n/a n/a 1/3 11.58 19.27 0.565 31.41</td>
</tr>
<tr>
<td>3SVD</td>
<td>2</td>
<td>2/2</td>
<td>0.24</td>
<td>1.6</td>
<td>2.10</td>
<td>n/a n/a 1/1 15.33 16.58 0.25 32.15</td>
</tr>
<tr>
<td>1CSS</td>
<td>4</td>
<td>4/4</td>
<td>0.48</td>
<td>8.7</td>
<td>9.00</td>
<td>n/a n/a 2/10 5.47 9.89 0.60 15.96</td>
</tr>
<tr>
<td>1MDA</td>
<td>3</td>
<td>6/6</td>
<td>0.58</td>
<td>7.8</td>
<td>8.10</td>
<td>12.0 14.10 2/2 14.23 11.80 0.48 26.51</td>
</tr>
<tr>
<td>1MDA*</td>
<td>3</td>
<td>6/6</td>
<td>0.58</td>
<td>7.8</td>
<td>8.10</td>
<td>12.0 14.10 3/3 10.65 10.76 0.51 21.91</td>
</tr>
</tbody>
</table>

4.8 Dealing with dynamics and large conformational changes

In the last sections, we showed how MaD could locate individual components or whole structures within experimental density maps reliably despite missing densities and conformational differences on a wide resolution range. The internal organisation of MaD allows it to go further with little effort. Once a database of descriptors has been generated, they can be matched to any other such database at the cost of loading the corresponding file, as the matching and refinement step are more than one order of magnitude faster than descriptor generation.
4.8. Dealing with dynamics and large conformational changes

Figure 4.20 – Borderline cases in the simulated benchmark. (a) The first two complexes obtained by MaD on PDB 1CS4. The first solution places correctly but misrotates a component, shown in red. The second solution however includes its correct orientation, shown in green. (b) With the default patch size of 16 voxels, MaD fails to place a small component, surrounded by others in the complex. An identical copy, shown in green, is however placed correctly. (c) Setting a smaller patch size at 12 voxels limits bias from surrounding subunits or absence thereof during descriptor building, yielding the full complex correctly predicted.

(Table 4.1). Thus, for any new conformation, only the component descriptors need to be generated, making the docking of structural ensembles an easy and fast extension of MaD.

To exemplify this, we will look at GroEL, a group I chaperonin comprised of two stacked homoheptameric rings (Figure 4.21a). GroEL, in tandem with the detachable lid GroES, encapsulates substrate proteins to provide optimal folding conditions. This process is ATP-dependent and involves a series of allosteric changes192. The single protomer of GroEL contains three distinct domains (Figure 4.21b). The equatorial domain remains symmetrical over both rings during the cycle. It transduces signals from ATP hydrolysis through the intermediate to the apical domain, the most flexible domain and asymmetrical between protomers to accommodate the substrate protein’s shape192.

An experimental cryo-EM map of GroEL at a global resolution of 6.1 Å and voxel spacing of 1.38 Å (EMD-5338) was selected from the EMDB (Figure 4.21c). This cryo-EM map does not have a pre-fitted structure, but PDB 1OEL (Figure 4.21a) matches the density relatively well after fitting an individual GroEL subunit with MaD (Figure 4.21d), where all 14 possible locations of the monomer within the map are identified without alternate, invalid results post-refinement.
Figure 4.21 – Assembling GroEL with MaD. (a) The two homoheptameric cis and trans rings of GroEL (PDB 1OEL) viewed from the top and front. (b) Close-up view of a GroEL subunit, showing the N-terminal in blue to the C-terminal in red, with each domain labelled. (c) Experimental cryo-EM map EMD-5338 at a global resolution of 6.1 Å with top and frontal views. (d) The GroEL component from (b) docked using MaD, successfully filling its 14 possible locations in the map.

Then, we ask whether MaD may identify which state from a structural ensemble best fits the assembly map. To this effect, a molecular dynamic simulation of a monomer from 1OEL (AMBER11 forcefield and PMEMD engine) was run for approximately 200 ns. More than 5000 conformational states strong, this ensemble is evidently too large to be used as is with MaD. Instead, the trajectory was projected on the first two principal components, which cover 68% of variability combined. CLoNe (see Chapter 3) was used to extract a set of unique conformations corresponding to cluster centers and to local energy minima in the projected conformational space sampled by the simulation (Figure 4.22a). The 5th cluster center is only 1.36 Å away from the reference structure 1OEL, the others are between 3.5 and 6.5 Å from the latter (Figure 4.22b-d), thus yielding a small but representative set of conformations. The 7 cluster centers were then docked individually within the assembly map using MaD. We show in Figure 4.22e the results from the state farthest from the initial structure, C1. Every location is found despite the conformational differences. However, the first 30 solutions need to be considered to identify them all, instead of the 14 expected ones. However, these additional, intermediate solutions carry meaning. The conformational differences between the expected structure and C1 may be explained by a hinge movement about the intermediate domain. Thus, depending on the cluster seed and location of the initial descriptor, it is possible that C1
4.8. Dealing with dynamics and large conformational changes

Figure 4.22 – Flexible modeling of GroEL with MaD. (a) The core cardinality of distribution obtained by using CLoNe on a 200 ns simulation of a GroEL monomer using AMBER. Red dots are the cluster centers extracted by CLoNe, and the white star corresponds to the starting conformation. (b) The RMSD-\(\alpha\) between the starting conformation and all cluster centers. (c-d) Comparison of the starting conformation with its closest and farthest state, respectively. (e) Reconstruction of MaD using the farthest state identified by CLoNe. The equatorial domain fits the density, but the intermediate and apical domain are outside the map. (f-g) In cases where conformational differences are present between the cryo-EM map and the component being docked, MaD may highlight different solutions for the same location, corresponding to different alignment to the density. In (f), the structure was aligned to the equatorial domain and in (g) to the apical domain. (h) MaD is able to highlight the conformational state that best fit the density from a structural ensemble, in this case the cluster center C5. (i) The average cross-correlation obtained by each cluster center on all solutions obtained by MaD. (j) C5 (blue) compared to C6 (orange), the second best conformational state from the ensemble.
is aligned to any of the equatorial, intermediate or apical domain (Figure 4.22fg). Therefore, these alternatives are not inherently wrong, but suboptimal.

Instead of checking each cluster center one by one, we may pool the results from the all cluster centers and sort them by their RWC score. Then, we may follow the same assembly building procedure as described in the previous section, with the difference that the component list to combine is now made up of different conformational states instead. It is thus possible to reconstruct GroEL with the state from the ensemble that best fit the cryo-EM density (Figure 4.22h). In this case, all components are equal due to the symmetry imposed during 3D reconstruction. However, this procedure is also applicable without modifications in the case where one or more location exhibit different conformational states or in the case of a heteromultimeric assembly. The cluster center identified this way is C5, which is indeed the closest to the reference structure (Figure 4.22b) and the one whose solution exhibit the highest average cross-correlation with the experimental map (Figure 4.22i). The second complex in line is one built from C6, whose main differences originate from the apical domain (Figure 4.22j).

4.9 Conclusions

We have developed MaD, a method based on feature descriptors in order to locate accurately any number of components within electronic density maps at the currently most problematic resolution range. Each step was thoroughly benchmarked and confirmed on whole-assemblies as well, confirming that the default parameters chosen are valid regardless of resolution, component size and voxel spacing. This makes MaD a certain method in the sense that initial conditions have little to no influence on the final results. We have further shown that MaD is robust to relatively large conformational changes, in which case multiple solutions are made available depending on the location of the main descriptor. While descriptors only integrate local information around an anchor point, incorporating the information from other descriptors was effectively turned into a conformational selector when a structural ensemble is provided. MaD is able to detect all positions of repeated structures within an assembly from a single copy, regardless of symmetry and with few to no false positives confusing results. Density cross-correlation is only used after local refinement, as it is sensitive to small displacements. Rather, scores based on descriptor cross-correlation and anchor cloud fit
proved to be a robust option until that step. While MaD processes subunits individually, the individual solutions are good enough that no complex scoring function needs to be developed when building assemblies. A fast combinatorial search is enough, and its simplicity further allows to consider multiple conformations for each component as well. However, we do find it easier to observe individual solutions manually instead, as it allows for better validation and easier identification of false positives, if any. We highlight the fact that MaD does not require any additional restraints, such as chemical cross-links, to obtain solutions. At the targetted intermediate range of resolution, such information is not needed, but may be added as post-processing to filter through existing solutions in case of uncertainty.

The final output structures of MaD are models more than structures. Components have been obtained from other, independent experiments such as X-ray crystallography, NMR, computational modeling\textsuperscript{12}. While MaD was shown to be robust to conformational differences, inter-subunit contacts are not optimized and further efforts need to be carried out. Molecular dynamics flexible fitting\textsuperscript{31} is our method of choice to bridge this conformational gap due to its speed and ease of use within VMD\textsuperscript{86} and NAMD\textsuperscript{193}. Specifically, the GroEL assembly predicted by MaD from the farthest cluster (Figure 4.22e) was successfully refined into the density using MDFF (data not shown). Alternatively, a refinement protocol similar to that of HADDOCK-EM could be applied, which include a semi-flexible simulated annealing step and another in explicit water\textsuperscript{53}.

Due to the nature of cryo-EM maps, the resolution and the voxel spacing are the only parameters that we have no control of. The latter may be avoided by interpolating experimental maps at a standardized voxel size, but attempts in doing so lead to generally lower performance due to interpolation artefacts in the form of aliasing or further loss of resolution. Nevertheless, we have shown that a patch size of about 25 Å in length is generally valid. For larger structures, larger patches may be considered for a slight performance increase. For smaller components, smaller sizes may be warranted instead, especially when the component is buried in the complex. The patch size remains a free parameter for this reason. The $\sigma$ of the LoG used at the anchor detection step may also be increased when the components are larger and resolution approximately below 8 Å. With higher blurring, less anchors are extracted, reducing the runtime of descriptor generation, which is the main bottleneck of the method. Due to the variability of experimental data, it is difficult to offer more precise guidelines and
Chapter 4. MaD: Macromolecular Descriptors for integrative dynamic modelling of large macromolecular assemblies

presented all data using the same default parameters for simplicity.

EQ spheres proved to be a convenient vector for orientation assignment. Multiple schemes have been tested to limit boundary effects, including the introduction of a local optimization to find the optimal alignment between the EQ sphere and the density or simply weighing the contribution of each gradient vector with respect to its distance to nearby EQ zone centers. However, while slightly more performant, these approaches were at least an order of magnitude slower. Thus, simple bin assignment has been chosen over alternative methods. We were initially weighing the contribution of vectors according to their magnitudes. However, due to the inconsistency of voxel values within experimental cryo-EM maps, we found that it was no better than simple vector counts, which is simpler and can be kept as integers. This is important for descriptor generation, where the optimal number of sub-regions for descriptor generation was determined to be 64. With each sub-regions being summarized by an EQ sphere with 16 zones by default, this leads to vectors of 1024 values. Storing thousands of such descriptors is more efficient as integers than floating numbers.

In section 2.1, we have mentioned the isovalue, or contour level used for visualization of electronic density maps. This value is critical for MaD. Without proper thresholding, noise particles may be present around the structure and will generate spurious anchors, which may severely slow down the docking procedure. The author-recommended contour level from EMDB is not consistent and should be adjusted manually. Regardless, suboptimal choices for contour levels are unlikely to affect the quality of the results beyond extending runtime. Half the recommended value was used for all cases presented in this chapter without issues. Maps downloaded from EMDB often contain the structure of interest surrounded by large amounts of empty space. MaD includes functions to eliminate voxels below a specific isovalue and truncating the empty space, making the loading of experimental maps easier and lighter.

Many other IM methods were tested on resolutions down to 20 Å at the time. Nowadays, such resolutions are much less common, with an average of 6.3 Å for single particle analysis in 2020. In this chapter, MaD relied on secondary structure elements within local density patches, which makes it most applicable up to resolutions of 10 Å, limit above which secondary structure is less distinguishable and density becomes blob-like. Early development of MaD included success on cases of resolutions down to 20 Å on simulated assemblies. The cleanliness
of simulated data is however too different from experimental data, making MaD intractable at this particular scale in real-world scenarios with the size of the regions considered in this chapter. In general however, the applicability of MaD is a matter of choosing a scale. Tube-like structures within 12 Å of diameter may correspond to $\alpha$-helices, and similar structures within 50 or 60 Å in diameter may correspond to tertiary structural motifs of larger components, making it applicable to lower resolutions as well if the pieces are large enough and patch size chosen accordingly. Even further in scale, MaD could be applied actin filaments, microtubules or even cilia or flagella in specific biological contexts. The underlying data is still in the same, grid-like format and would require little adjustments. From that, we can hypothesize that MaD would be applicable to predict the localizations of atomic or pseudo-atomic models of larger assemblies, subcellular components within whole tomograms, painting high resolution pictures of specific biological contexts. On the other end of the molecular scale, MaD may be applied to small-molecule docking. An early version of the anchor detector presented here was used to extract features from energetic grids derived from pockets (see EPFL Theses 8630\textsuperscript{194} and 9641\textsuperscript{195}). In that particular application, these energetic features were utilized by a genetic algorithm\textsuperscript{196} to dock small molecules in their target pockets. All in all, this general potential of MaD coupled to its robustness to conformational differences, incomplete structures and significant levels of noise makes it a powerful tool for multiscale integrative modeling in general.
Modeling the group II chaperonin Mm-Cpn with MaD-CLoNe

5.1 Background

Proteins generally require folding into 3D structures after translation of their messenger RNA in order to be functionally active. Without specific conditions, aberrant folding may occur, in turn leading to dysfunctions within the cell and ultimately to a variety of conditions, including neurodegenerative diseases or cancer. Among molecular chaperones, chaperonins are a class of proteins that provide ideal conditions to ensure proper protein folding and thus contribute to proteostasis. Chaperonin complexes are formed by two stacked, back-to-back rings, which have diverged into two groups. Group I chaperonins are present in prokaryotes as well as mitochondria and chloroplasts. The most common member is GroEL (see chapter 5, section 4.8), which requires a detachable lid GroES to encapsulate protein substrates within its nanocage in response to ATP binding to GroEL\textsuperscript{197}. Group II chaperonins instead are present in eukaryotes and archaea and their monomers are diverse, sometimes heterogeneous within the same complex as in the case of TRiC\textsuperscript{198}. In lieu of a GroES-like detachable lid, each subunit includes a built-in lid extending the apical domain. Group II chaperonin rings are aligned in phase in contrast to the staggered rings of their group I counterparts (see Figure 4.21). Subsequently, their regulation differs as well and the resulting complexity makes chaperonins common targets for hybrid modeling strategies\textsuperscript{199,200}.

Mm-Cpn is a group II chaperonin from \textit{Methanococcus maripaludis}. The apical domain serves for substrate recognition through a built-in lid extension and receives information from the ATP-binding equatorial domain through the hinge-like intermediate domain (Figure 5.1a). In absence of ATP, the subunits remain in open, dynamic states\textsuperscript{199} as evidenced by cryo-EM.
Figure 5.1 – Architecture and known structures of the Cpn complex from *Methanococcus maripaludis*. (a) Structure of a monomer in the closed state. The different domains are colored. The lid is shown in blue although it is part of the apical domain. A close-up view of the ATP-binding site in the equatorial domain is shown with the catalytic residues D386 and K161 of the intermediate domain mentioned in the main text. (b) A cryo-EM structure at 8 Å of the open state, currently unreleased, provided by Prof. Wah Chiu and Dr. Yanyan Zhao of Stanford University. Two contour levels are shown to highlight both the high resolution around the equatorial domains and the low resolution of the other domains. (c) The structure the D386A complex in open, locked ATP-bound form (PDB: 3IZH, EMD-5244). (d) The atomic structure obtained from a 3.5 Å cryo-EM map currently unreleased, provided by the same group in Stanford. Top and front views of cartoon and sphere representations are shown.

models with low densities near the apical domains (Figure 5.1b). The binding of ATP leads to a 45° rotation of the apical domain, leading to a more compact open state (Figure 5.1c). ATP hydrolysis by D386 in the intermediate domain further closes the lid and releases the substrate into the folding cage. ATP hydrolysis is sensed by K161 in a Nucleotide Sensing Loop (NSL), which is shifted, transducing this change allosterically to the apical domain through α-helices within the intermediate domain. A loop in the apical domain is then interacting with those of nearby subunits, disrupting binding of the substrate and subsequently releasing it in the folding chamber. In that closed state, the apical domains from each ring form an iris as evidenced by a recent, unreleased structure determined from a 3.5 Å cryo-EM map form the lab of Prof. Wah Chiu in Stanford (Figure 5.1d). In slight contradiction with these observations, Mm-Cpn was crystallised in complex with AMP-PNP (unhydrolysable ATP) in a
closed conformational state\textsuperscript{202}. Regardless, ATP hydrolysis is required to release the substrate into the chamber and thus for folding to occur\textsuperscript{200}. In a structural homolog\textsuperscript{204}, it was found that the substrate is released with ADP and phosphate at the end of the cycle\textsuperscript{204}. At least in the case of a $^{35}$S-rhodanese substrate and Mm-Cpn, multiple rounds of ATP binding and release, thus of opening and closing, occur in the time required to fold the substrate polypeptide, which binds the complex prior to ATP binding\textsuperscript{200}. Hydrophobic interactions are involved in substrate binding in both group I and II chaperonins, and a substrate-binding region has been approximatively mapped to a hydrophobic helical patch on the apical domain\textsuperscript{205}. The latter also exhibits a number of both positively and negatively-charged residues pointing inwards the folding chamber, essential for peptide folding\textsuperscript{205}.

Thus, there currently exist many yet mostly equivalent atomistic models of Mm-Cpn. In the closed state, PDB 3IYE (later updated to 3LOS) was obtained from a 4.5 Å map, 3IZM from a 7.2 Å map, 3KFB from x-ray crystallography at 3.3 Å\textsuperscript{205} and 3RUV in complex with AMP-PNP. Other closed models are available from a study on the dual role of ATP in Mm-Cpn\textsuperscript{200}, although not in a mutated form lacking lid residues. Alternatively, a DNA-bound, semi-open state\textsuperscript{200} (Figure 5.1c) has been modelled from an 15 Å map in the same study after mutating catalytic residue D386 to alanine to prevent ATP hydrolysis (PDB 3IZH). In a fully open state, PDB 3IYF was modelled from 3IYE into a density map at 8 Å using DireX\textsuperscript{34}, albeit in lidless form due to the poor resolution of the apical regions. Another model, also lidless, was obtained from low-resolution X-ray crystallography at 6 Å\textsuperscript{205} (PDB 3KFK). As part of a collaborative effort with Prof. Wah Chiu and Dr. Yanyan Zhao from Stanford University, recent yet unpublished structures were shared with us. They include a closed state structure determined from a 3.5 Å cryo-EM map\textsuperscript{203} obtained from Mm-Cpn particles in high concentration of ATP, and an 8 Å map of the open state in presence of AMP-PNP, an unhydrolysable form of ATP. In the latter, the resolution at the equatorial domain may be estimated to 5 Å as side-chains are not visible, and decreases to beyond 10 Å towards the apical domain (Figure 5.1b). After focused classification at the subunit level from a sample of Mm-Cpn in presence of ATP (1 Mm-Cpn particle per 8 ATP molecules), 20 cryo-EM maps at varying resolutions and at different intermediate states between open and closed conformations were further shared with us (Figure 5.2).

While previous studies mostly imply a concerted behaviour between subunits, collaborators uncovered that ATP binding is in reality statistically random between subunits. Moreover,
Figure 5.2 – Mm-Cpn subunit in 20 different states determined by cryo-EM. Maps provided by collaborators in Stanford University after focused classification on full complex particles with a subunit mask. An estimation of the resolution for each map is reported. The maps are ordered from an open to closed state following a procedure explained in a later section.
they report that one subunit binding ATP does not induce allosteric effects onto neighbour subunits, within either the same ring or in the other. This corroborates the flexibility of open states seen in cryo-EM maps (Figure 5.1b) and collaborators hypothesize that further studies of Mm-Cpn could thus be simplified to studying a single component.

Our goal was to further investigate the dynamics of chamber closing from the perspective of both individual subunits and full particles. The results presented in this chapter constitute a work in progress and were obtained through a hybrid modeling effort involving the aforementioned structures and cryo-EM maps, molecular dynamics, MaD, CLoNe as well as other, existing software such as PHENIX\textsuperscript{206} and ChimeraX\textsuperscript{30}.

### 5.2 Molecular dynamics and docking with MaD-CLoNe

Our collaborators previously reported higher structural heterogeneity in presence of ATP\textsuperscript{203}. An initial, ATP-bound monomeric structure extracted from the atomic structure of the closed-state complex was thus used to setup molecular dynamics simulations (Figure 5.3a). Performed with CHARMM-GUI\textsuperscript{70} and GROMACS\textsuperscript{207}, the 1 \(\mu\)s simulation exhibits large move-
ments of the intermediate and apical domains. Under the fair assumption that Mm-Cpn subunits remain anchored on their equatorial domains, the trajectory was aligned on the latter prior to a principal component analysis. The first component covers 68% of the variability covered in the simulation and progresses from a closed state to an extremely open one (Figure 5.3b, bottom). CLoNe was then used to extract a set of 124 intermediate states from that unidimensional dataset (i.e. cluster centers) (Figure 5.3c), down from almost 6000 frames, which results in an efficient subset of the conformational landscape sampled during the simulation.

Each state was then docked using MaD within each of the 20 maps using the protocol described in the previous chapter (Figure 5.4). For each map, the conformational state that best fit the cryo-EM data is thus found automatically, only by feeding the list of maps and structural ensemble to MaD (Figure 5.4). Due to the high degree of flexibility of the structure, specifically near the apical region, the fit to the density was further refined using a short MDFF run in all of the 20 cases (Figure 5.5).

5.3 Modeling the mechanism of chamber closing of Mm-Cpn

At this stage, the combination of molecular dynamics and MaD-CLoNe enabled high-resolution atomistic models of a range of intermediate states of Mm-Cpn subunits. Of note, the apical domains lack in accuracy due to their flexibility and poor resolution in the cryo-EM maps. Regardless, the enhanced resolution of the subunit models now enables their accurate ordering, mostly based on the equatorial and intermediate domains, by projecting them on the first principal component after the same alignment as for the ATP-bound trajectory (Figure 5.6a). Of note, the maps shown in Figures 5.2, 5.4 and 5.5 were already shown in that order. The ATP-bound state (PDB: 3IZH) is closer to the open states while the newest closed structure is closest to the closed states while both conveniently cross midway between the estimated order (Figure 5.6b).

While the resolution of most subunit maps is not high enough to accurately detect the presence of ATP, even less distinguishing between ATP and ADP, reasonable estimations could be made (Figure 5.6c). The more closed states contain relevant densities that could be attributed to either ATP or ADP. Specifically, the density is not covering the \( \gamma \)-phosphate in the last state (#20), which may suggest that it was hydrolysed or that it was present only
5.3. Modeling the mechanism of chamber closing of Mm-Cpn

Figure 5.4 – Docking of Mm-Cpn subunit ensemble with MaD-CLoNe. The maps are ordered from an open to closed stated following a procedure explained in a later section.
Figure 5.5 – Flexible refinement of docked Mm-Cpn subunits with MDFF. The maps are ordered from an open to closed stated following a procedure explained in a later section.
5.3. Modeling the mechanism of chamber closing of Mm-Cpn

Figure 5.6 – Modeling the closing mechanism from subunit models. (a) The 20 fitted structures were ordered according to the first principal component, which describes the closing of the Mm-Cpn subunit. The most open and closed states are shown in blue and orange, respectively. (b) The RMSD-Cα was computed between a monomer of the ATP-bound structure (PDB: 3IZH) and all MaD-fitted models in order from open to closed states (blue). The same was made with a monomer from the provided closed structure (red). Vertical bars identify the states where nucleotide density could be observed. (c) Close-up views of the cryo-EM density that may be attributed to ATP or ADP. (d) Comparison between Mad-MDFF fitted structures and the closest states sampled by molecular dynamics from both apoform and ATP-bound Mm-Cpn subunits. The structural ensemble extracted from the latter by CLoNe is also shown.
in some of the particles used in the 3D reconstruction of the density map. The resolution of the corresponding map was estimated at 4.9 Å by our collaborators, which is not enough to distinguish between hydrolysed ATP or already released ADP. Alternatively, similar densities may be observed in the second and tenth states, whose original maps were estimated at 5.0 Å and 4.6 Å, respectively. In these cases, it is not possible to determine whether the density originates from ATP binding of a new cycle, corresponds to transitional, opening states post-ATP hydrolysis or is at all relevant.

We then computed the RMSD-\(\alpha\) between the subunit models and every frame of the ATP-bound simulations, as well with each state from the structural ensemble from CLoNe (Figure 5.6d). To avoid bias from the poorly resolved lid residues in the cryo-EM map, they were not considered. In general, the states from CLoNe are not far from the optimal frames available in the simulation, with RMSD-\(\alpha\) between 1.33 and 3.88 Å, and thus provide a significant upgrades from the previously available structures of the closed and ATP-bound open states (Figure 5.6b). Finally, a simulation of an apoform Mm-Cpn subunit was made. Consistent with the report of our collaborators\(^{203}\), the ATP-bound simulation is more heterogeneous and sampled states closer to those captured by cryo-EM than the apoform trajectory. All in all, these results further confirm the applicability of CLoNe to create relevant and efficient structural ensembles for hybrid modeling tasks.

We then aimed to integrate the subunit models to predict the dynamics of closing at the level of the full particle. While the trajectory was previously aligned on the equatorial domain, the internal dynamics of Mm-Cpn cannot be summarized by the movements of the intermediate and apical domains alone. A rocking mechanism was previously observed on the equatorial domain upon closing of the folding chamber\(^{199}\). The newly shared map of the open state exhibit high resolution of the equatorial domain (Figure 5.1b), enabling a more accurate modeling than with the previous map (Figure 5.7a), which involved an elastic deformation of the closed structure\(^{199}\). Comparing that domain with its counterpart of the closed state, a rotation is observed (Figure 5.7b), which may be accurately estimated given an RMSD-\(\alpha\) of 1.6 Å between the two domains (0.9 Å excluding flexible loops), thus preventing bias from different conformational states. A similar movement was previously highlighted\(^{199,205}\). Of note, the resolution of the ATP-bound, open state (PDB: 3IZH and EMD: 5244) do not allow for a precise estimation of the equatorial conformation and could not be accounted for (Figure 128).
5.3. Modeling the mechanism of chamber closing of Mm-Cpn

Figure 5.7 – Modeling the closing of the folding chamber of Mm-Cpn. (a) Docking of an equatorial domain within the open state cryo-EM map. (b) Comparison of the structure from (a) with the equatorial domain of the atomic structure of the closed state in front and top view. The Euler angles corresponding to the rotation are reported. (c) The equatorial domain of the intermediate, ATP-bound open state (PDB: 3IZH and EMD: 5244). The density does not allow for accurate estimation of the equatorial position relative to those in (b). (d) Reconstruction of the full Mm-Cpn complex from the first, most open subunit model (left) and last, most closed model (right). (e) The height of the complex model corresponding to the 17th subunit (in order towards chamber closing) is compared to the last two complexes, with the closed structure from our collaborators in white outline. A compaction from approximatively 170 Å to 145 Å is observed. (f) Halves of last two reconstructions closest to the closed state are shown in the order of the chamber closing (grey and green, respectively, with the respective state number reported). A rotation of the apical domains (depicted by colored circles on the right) is observed.
5.7c Nevertheless, our collaborators reported no allosteric change propagated to neighboring subunits in either ring upon binding of ATP. This suggests that no large change occurs on the equatorial domain upon ATP binding.

Under the assumption that all 20 subunit models are spaced evenly in the closing cycle of the complex and by decomposing the rotation into smaller steps, each subunit model was thus rotated in order to better reflect the movement of the equatorial domain during the closing of the chamber. Taking advantage of the D8 symmetry of Mm-Cpn, full complexes have then been modelled at each intermediate state (first and last of which are shown in Figure 5.7d) using the apply ncs tool in PHENIX\textsuperscript{206}. Interestingly, the last complex models display a compaction from 170 Å across the rings to 145 Å (Figure 5.7e), 25 Å more compact than the closed structure from our collaborators. Furthermore, an en bloc rotation of the apical domains may be observed between the last two complex models (Figure 5.7f). Both of the compaction and apical movements are previously unreported and may correspond to new aspects of the Mm-Cpn folding pathway.

5.4 Further insight from large-scale molecular dynamics

We finally sought to elucidate the transition dynamics and energy barriers between each intermediate states modelled in the previous section and previously published structures, in a similar fashion to a previous study\textsuperscript{208}. To this end, we performed 1\(\mu\)s-long simulations (CHARMM forcefield, simulated with GROMACS) of each subunit model from Figure 5.5. We plot the first two principal components from all simulations after aligning each frame with respect to backbone atoms excluding the lid residues and flexible regions. Two main basins are sampled through the simulations (Figure 5.8a). The first basin is the most dense and closest to the states from cryo-EM, whose progression until lid closing is reported along with the closed structure from our collaborators and the ATP-bound open state (PDB 3IZH) (Figure 5.8b, left). Consistent with our collaborators including ATP during sample preparation (ratio of 1 Mm-Cpn hexadecamer to 8 ATP molecules), the subunit models derived from cryo-EM seem to transition from an ATP-bound, open state (rather than a fully open state) to a fully closed state. Although the available data do not allow to distinguish between pre- and post-hydrolysis conformations, the models derived from the cryo-EM maps seem to progress further than the closed struture. This is consistent with the observations made in the previous section, where
5.4. Further insight from large-scale molecular dynamics

Figure 5.8 – Combined molecular dynamics of individual subunit models. (a) Projection of the 20 simulations of 1 µs onto their first two principal components. Purple depicts high density and orange low density. (b) The path to chamber closing derived from the initial 20 subunit models is shown in black, with the open and closed state referred to with O and C, respectively (left). The white cross corresponds to the ATP-bound, open state (PDB: 3IZH) and the white diamond corresponds to the closed structure from the 3.5 Å cryo-EM map. We show the cluster centers identified by CLoNe as black stars (right). The arrows depict our hypothesis for mapping the ATP binding and hydrolysis events relating the basins. (c) Structures of the cluster centers color-coded as per their density level from shown on top of a surface representation of the complex built from the most open state (left, O in panel (b)) or the most closed state (right, C in panel (b)). (d) Cartoon representation of the cluster centers. The arrows report the rotation of the apical domain followed by the closing of the chamber, in tandem with the rotation of the equatorial domain. SBR refers to the substrate-binding region. On the right, a view shown from above highlighting the rotation of the apical domain. (ef) Relevant panels from a previous study of our collaborators199, showing the rotation of the apical domain upon ATP-binding and closing of the chamber upon ATP hydrolysis.
further compaction and rotation of the apical domains were observed.

To understand the nature of the secondary basin, which is unexplored by the cryo-EM data at our disposal, CLoNe was used to extract representative structures (cluster centers) (Figure 5.8b, right). We hypothesize that the secondary basin corresponds to an ATP-free and substrate-accepting state, as it is leads to an extremely open Mm-Cpn particle (Figure 5.8c). Comparison of this putative ATP-free state with the center closest to the ATP-bound state highlighted a rotation of the apical domain (Figure 5.8d), which conveniently exposes the substrate-binding region to the particle's surroundings (Figure 5.8d, blue highlight). This rotation of the apical domain has been previously reported, albeit at a smaller angle (relevant figure panels from their original article201 are shown in Figure 5.8ef). This rotation may occur to promote the individual subunits binding nearby substrate polypeptides prior to ATP-binding.

Finally, we report the space sampled by each individual simulations in Figure 5.9. Many states remain close to the main basin along the sampled trajectory (models 1, 7, 8, 12, 13, 14 and 16). Interestingly, several others transition from the main to the secondary substrate-accepting basin (models 2, 4, 5, 9, 10, 15). While it cannot be expected that a subunit in isolation behaves similarly to its oligomeric form, two distinct transition paths between the main basins may be observed. One path travels upwards (models 2, 5, 9, 10 and 15), while others pass by the lower part of the principal space (models 17 and 19). Since the simulations only include ATP-bound subunits, a more accurate assignation of pathways is not possible with this data alone. Nonetheless, combined with the hypothesis derived from Figure 5.8b and the cryo-EM models diverging from the closed structure, we can further suggest that Mm-Cpn undertakes a different path post ATP hydrolysis than upon substrate and ATP binding. We may thus hypothesize that Mm-Cpn first binds its substrate first and ATP next, which induces a rotation of the apical domains. This brings the substrate peptide towards the inside of the chamber. ATP is then hydrolysed, the particle closes and the substrate is released within the folding chamber. A tandem movement of compaction and rotation of the apical domains then induces the re-opening of the chamber, enabling the release of both substrate, ADP molecules and cleaved phosphates.
Figure 5.9 – Individual conformational sampling of subunit trajectories. In each panel, the conformational space sampled by all 20 simulations are shown in grey. Then, for each of the atomic models derived from the 20 cryo-EM maps in order towards chamber closing, we show their respective sampling. Blue dots refers to frames at the start of the simulation and red dots to frames near the end of their respective trajectories.
5.5 Conclusion

We have obtained atomistic models of 20 states of the group II chaperonin Mm-Cpn by integrating molecular dynamics and cryo-EM with both CLoNe and MaD. We resorted to classical molecular dynamics because of its general applicability and higher biophysical accuracy but strategies based on biased molecular dynamics schemes or normal mode approaches could have been used towards enhancing the resulting conformational sampling. Nonetheless, the classical approach proved successful in previous integrative studies by our group. Here, few yet relevant states were extracted from molecular dynamics simulations using CLoNe. This structural ensemble was then used in MaD, which was able to process, filter and select the conformational states that best fitted those captured by cryo-EM data. Of note, the frames selected by CLoNe were among the closest ones to the final cryo-EM models obtained after refinement. Finally, the extensive benchmarks from the previous chapter enabled a fluid docking workflow, with all results presented obtained with just the default parameters and best ranking solutions, exhibiting the ability of the MaD-CLoNe combination to tackle complex and flexible macromolecular assemblies.

Our models represent the pathway to the closing of the folding chamber in unprecedented details, from fully open, to ATP-bound and towards chamber closing upon ATP hydrolysis in unprecedented details. Two new, previously unreported conformational changes highlight movements that may occur post-ATP hydrolysis. We observed significant flexibility within a secondary basin that we attributed to an ATP-free, substrate-accepting Mm-Cpn with an exposed substrate-binding region. This is compatible with previous studies reporting binding of substrate peptides prior to the binding of ATP. Thus, the crossing between the two basins may be hypothesized to be driven by the binding of the substrate peptide. The modeling of individual subunit dynamics was focused on an alignment on the equatorial domain, putting emphasis on the hinge-like opening of the monomer and mimicking the anchoring of the components within each ring of the complex. The location of the subunit models on the landscape sampled by our simulations support this approach and is compatible with previous studies stating that, after ATP-binding, only a closing of the chamber occurs, internal allosteric changes notwithstanding.

While the subunit cryo-EM maps were technically obtained from purified Mm-Cpn in
5.5. Conclusion

presence of ATP, it is safe to assume that not all Mm-Cpn components in the micrographs would be ATP-bound. However, the inherent flexibility of the structure probably prevented reliable reconstructions of more open states to be obtained. This is corroborated by the extremely low density of the intermediate and apical regions in the newest cryo-EM map of Mm-Cpn's open state shared by our collaborators. Furthermore, even if some states had estimated resolutions below 5 Å, neither the presence of ATP and its distinction from ADP may be determined with certainty. Nucleotide densities could still be observed in the more closed states, which may belong to Mm-Cpn particles prior or after ATP hydrolysis. The nucleotide density in the last state exhibits partially uncovered phosphate, which supports the observation that our subunit models diverge from the closed structure on the conformational landscape obtained from our large-scale simulations.

The newer cryo-EM structures of the open and closed states allowed for a precise determination of the rotation of the almost rigid equatorial domain during the cycle, in line with previous studies despite their reliance on lower resolution data. This movement enabled to transpose subunit models to full particles. As the modeling originates from singular components that were refined individually in density that do not display high enough resolution to highlight side-chains, any inter- and intra-ring contact that may be observed in our models (beyond the most closed state) may not be reliable. Nevertheless, our analysis highlighted previously unreported movements in the compaction of the closed state and further rotation of the apical domain. Consistent with the poorly defined ATP density of the most closed subunit model and the conformational landscape from the large-scale simulations, this movement may be hypothesized to occur post-ATP hydrolysis as the Mm-Cpn particle reopens. Of note, the direction of that rotation cannot be elucidated from our data alone as it is probable that some intermediate states in the presented order may have been permuted due to the PCA mostly focusing on the closing of the chamber.

Despite our hypotheses, the exact conformational changes occurring upon ATP hydrolysis and chamber opening post-hydrolysis are still unknown. The presence of a smaller basin (approximative PC coordinate (0, -5) in Figure 5.8a) may correspond to such an intermediate state, which would differ from the path towards chamber closing probably due to the release of ADP and its cleaved phosphate. Similar alternative paths may be observed in few individual simulations. The presence of ATP in our simulations most probably prevents the accurate
exploration of post-hydrolysis states in general. Additionally, it is common for subunits to behave differently in isolation than in their oligomeric form\textsuperscript{37,38}. The latter basin and any other unexplained sampling may be due to such differences. To investigate these aspects, we plan on running further simulations of each subunit models, this time in absence of ATP or presence of ADP (and/or phosphate). Finally, to better estimate the conformational landscape accessible by the subunit in the context of the full particle, triplets of components with similar nucleotide conditions may be simulated as well. If the secondary, substrate-accepting basin is not only relevant for Mm-Cpn monomers but for the full particle as well, then the central protomer should be able to sample that basin despite the hindrance from neighbour protomers. Finally and to our knowledge, all structures previously published and those shared by our collaborators are substrate-free. Thus, we suggest that further experimental efforts should be focused on the structural determination of the complex in presence of known substrates, which should be a tractable feat post-resolution revolution.
This work started as an extension to our optimization platform for enhancing resolution, pow\textsuperscript{er} \textsuperscript{37,38,196}. This platform was limited to symmetrical homomultimeric assemblies, and included molecular dynamics data to sample component flexibility. An exception to this is a modeling effort of the Huntingtin protein\textsuperscript{210}, which used a different optimizer tailored for that specific case\textsuperscript{194}. Genetic algorithms have been applied to heteromultimeric assemblies\textsuperscript{49}, but authors advertised a limit of 8 rigid components, certainly due to the growing complexity of the task and search space dimensionality. The CLoNe project started with the single aim to reduce the size of conformational ensembles and relieve the burden of the optimizer so that the feat of aerolysin\textsuperscript{38} and HIV capsomer\textsuperscript{37} could be upgraded to heteromultimeric assemblies. This quickly proved intractable, with a key issue being the reliance of our optimizers on a continuous search space, making pow\textsuperscript{er} incompatible with CLoNe.

Thus, I chose to go down an alternative path, revisiting the less popular and less successful feature-driven methods\textsuperscript{47,48} with a new perspective from computer vision. Nowadays, many vision tasks are dominated by deep learning since the advent of convolutional neural networks\textsuperscript{211}, which enabled the integration of extremely large amounts of image data with incredibly few parameters compared to traditional neural nets. Even with this progress, SIFT\textsuperscript{100} and similar descriptor schemes\textsuperscript{110,111} are still competitive\textsuperscript{116,117} compared to learned techniques, which mostly relied on SIFT features for training\textsuperscript{101,102,115}. For macromolecular applications, no such methods existed prior to MaD, and the EMDB\textsuperscript{6} was significantly smaller, with fewer cryoEM structures having a pre-fitted PDB. This lack of properly labelled data made it difficult to design partially learned or end-to-end learning frameworks from scratch.
Chapter 6. Conclusion and perspectives

Nowadays, due to the growth of the EMDB, such avenues are now approachable. At the time of writing, attempts using Siamese convolutional networks (similar to LIFT\textsuperscript{102}) are being tested on a set of 200 assemblies with pre-fitted structures carefully selected from EMDB\textsuperscript{6}, which result in a few millions descriptors. The biggest bottlenecks to proper training arise from the different voxel sizes and from the poor quality of many cryoEM maps. The former may be circumvented using inception modules, which let the model learn which kernel size to use and the latter may bring significant confusion during training for regions with poorly defined densities. This is reminiscent of SURF’s approach with box filters and integral images\textsuperscript{104}. However, current models perform worse than hand-crafted macromolecular descriptors, although this is a work in progress.

The resolution revolution of cryoEM certainly affected the range of application of MaD. While most structures from single particle analysis are nowadays within reach of de novo building methods, many assemblies are still stuck at intermediate resolutions. Many such cases are part of larger groups of structures submitted simultaneously, and often one map is resolved at near-atomic resolutions while others correspond to alternative biological states. In such cases, MaD may not be the most appropriate method. However, such situations are not yet true in general, a prime example being the group II chaperonin Mm-Cpn presented in Chapter 5. In such cases, dynamics-including hybrid approaches are in high demand and the combination of MaD and CLoNe are ideal for such tasks.

Cryoelectron tomography is converging to the same intermediate resolutions when combined with single particle analysis, while subtomogram averaging currently achieves resolutions of 15 Å in average. As outlined in the conclusions of chapter 4, MaD may be easily adapted to this resolution range when sufficiently large components are being docked. It may be pushed further to dock whole macromolecular complexes into tomograms, painting high resolution pictures of key biological environments, or even an atlas of the cell depending on further developments of cryo-electron tomography\textsuperscript{212}, proteomics and super-resolution techniques. On the other end of the molecular scale, we have previously integrated an anchor detection scheme for small-molecule docking involving energy grids\textsuperscript{194,195}. Undoubtedly, descriptors could be applied in that area as well, although not without several adjustments. In general, the robustness to conformational differences and to noise of local feature descriptors is enticing, and makes MaD a promising avenue for multiscale integrative modeling. Of note,
a clever use of descriptors may be found within MaSIF\textsuperscript{213} (molecular surface interaction fingerprinting), which may be applied to a variety of relevant prediction tasks including interface prediction and pocket classification.

At the scale investigated in this thesis, cryoEM maps do not contain enough information to optimize inter-component contacts and interfaces. As high-resolution component structures are obtained from independent experiments, structural and energetic clashes are to be expected. While MDFF is a powerful tool to circumvent these issues up to a degree, it may not be sufficient and sometimes introduce unnatural structural changes. These may be detected through the use of tools such as MolProbity\textsuperscript{214} and when predictions are sufficiently close to their optimal conformation, short molecular dynamics simulations may be performed instead, similar to HADDOCK-EM\textsuperscript{53}. This only covers the cases where the conformational states of the individual components are already conformationally close to the low resolution data. In general, discriminating good from bad solutions is a recurrent problem in integrative modeling\textsuperscript{196,215}, which is further made difficult when larger conformational changes are present. In such situations, electrostatics or structural clashes may prevent the energetic component of scoring functions to be reliable.

Alternatively, experimental inputs are often introduced in integrative studies\textsuperscript{39,196,216}. They include chemical crosslinks and FRET, which provide distance information, H/D exchange and mutagenesis experiments which provide interface and binding sites mapping, or SAXS, which provides low resolution shape information. However, their proper integration requires specifically designed scoring functions on top of the previously mentioned energetic components. Due to the different nature of these inputs, their relative contribution needs to be weighed accordingly and, in general, whenever new information is added\textsuperscript{196}. This is not only tedious but best-ranking assembly models may also not be properly highlighted\textsuperscript{215}, forcing the users to dive into countless possibilities. This is made further difficult when flexibility needs to be accounted for. As a result, many docking approaches instead rely on such scoring to sort through the generated models\textsuperscript{215}. We recently published an alternative approach to solve this issue within our optimization platform, pow\textsuperscript{er}, which relies on a memetic viability approach (pow\textsuperscript{er}-mVie). By splitting the scoring function into objective and constraints, treated as inequalities, it is able to bypass individual weighing of energetic and experimental constraints\textsuperscript{196}. By extension, such information independence also facilitates the inclusion of
new information and different yet valid models satisfying different sets of constraints could be identified in some cases. In practice, pow$^{er}$ could only be reliably applied to symmetrical assemblies or to two-component complexes. Higher order assemblies required a number of dimensions too large for pow$^{er}$-mViE to navigate and converge reliably. Similar to other feature-driven approaches$^{46,49}$, MaD was originally designed to reduce the search space for pow$^{er}$-mViE by pre-docking components into low-resolution data. Then, pow$^{er}$-mViE could select assembly models that best fitted a set of experimental restraints while further optimizing the fit to low-resolution cryo-EM maps. Even in rigid cases with no conformational differences, it was however not able to reliably predict assemblies of more than 3 components. Thus, a more specialized development of macromolecular descriptors was initiated. In its final form, MaD bypasses the need for a traditional scoring function altogether, at the cost of relying on existing cryo-EM data as a strict requirement. Instead, MaD relies purely on internal descriptor scoring, which conveniently doubles up as conformational selector. This stems from the nature of descriptors, which integrate local information first and foremost. In general, the more matching descriptors that may be identified between two structures, the better the match between the component and the target low-resolution data. Large conformational changes often involve the movement of domains with respect to one another rather than internal atomic rearrangements, from the perspective of low-resolution cryo-EM data at the very least. This was shown with GroEL, where 14 components deviating by more than 6.5 Å of RMSD could not only be located correctly but also from the perspective of multiple domains as well. Relevant solutions ranked before any false positives or only included few of them. In general however, validating such solutions is difficult and the prior building of a structural ensemble, for example through normal modes$^{168}$ or molecular dynamics, is worthwhile. If confusion between solutions still remains, experimental inputs should be added as a post-processing step or through our optimization platform pow$^{er}$-mViE$^{196}$.

In this thesis and previously in our lab, we have utilized molecular dynamics to sample the conformational landscapes of atomic components. The goal was to sample states close to those captured by cryoEM$^{38}$ or fitting a set of experimental constraints$^{38}$. One may wonder about the potential of molecular dynamics to bridge the gap between conformational selection and induced fit, or its effectiveness in sampling the conformational landscape of proteins. This was approached in a recent article$^{217}$, which focused on establishing the best flexible
approaches for protein-protein docking. Molecular dynamics was found to be effective in bridging that gap, but like all other methods presented, it was not able to complete the task even when pushed in the proper direction along normal modes or principal components. The simulation time was however extremely short, at only 60 ns. In collaboration with Dr. Luciano Abriata, Dr. Giorgio Tamò and now Lucien Krapp from our group, a large-scale simulation effort was initiated with the Swiss National Supercomputing Center (CSCS). About 25 systems were selected from the Protein-Protein benchmark 5.0\textsuperscript{14}, covering protein-protein interfaces of many kinds. Several simulations of over 1 \( \mu \)s were made for each selected systems, involving the protein complex and either receptor and ligand isolated in their bound and unbound forms. While this project is undergoing, interesting conclusions may already be drawn. In several cases, significant overlap between the conformational spaces sampled by the either the complex or its components was observed. Combined with the previous use of molecular dynamics by our group and within this thesis, the potential of classical molecular dynamics to sample relevant regions of conformational landscapes is clear. Combined with tools such as CLoNe and MaD, molecular dynamics may be used \textit{a priori} to sample key functional states, aiming to approach those captured by other methods, such as cryoEM. Then, it may be used \textit{a posteriori} to refine inter-component interfaces, yielding atomic models of large and flexible assemblies at unprecedented levels of details. Molecular dynamics in conjunction with hybrid approaches are thus on the front line of modern structural biology.
Appendix

A 3D Hessian detector

The Fast-Hessian of SURF\textsuperscript{104} is based on computing an approximated determinant of Hessian through Gaussian box filters. The concept of integral images is easily translated to 3D using Numpy’s cumsum function. Box filters were fully implemented and tested in 3D, but were too slow to be tractable on density maps. This is because of Python’s infamously slow for loops, in this case the loop over voxels. It took 2.5 sec to process a 67x59x80 map (EMD: 1340, previously shown in Figure 2.1) and 86 sec to process the same map super-sampled by a factor 2. This for loop unfortunately could not be replaced by any of NumPy’s efficient array functions, which in general yield speed-up of a couple of orders of magnitudes. It was found that it is more effective to convolve a Gaussian pyramid with regular Gaussian filters using well-optimized libraries, in this case SciPy\textsuperscript{171}. We thus need to make some considerations about the use of the determinant of the Hessian matrix in 3D and its normalization over scales. The 3D Hessian matrix can be expressed as follows:

$$H = \begin{bmatrix}
\frac{\partial^2 M}{\partial x^2} & \frac{\partial^2 M}{\partial x \partial y} & \frac{\partial^2 M}{\partial x \partial z} \\
\frac{\partial^2 M}{\partial y \partial x} & \frac{\partial^2 M}{\partial y^2} & \frac{\partial^2 M}{\partial y \partial z} \\
\frac{\partial^2 M}{\partial z \partial x} & \frac{\partial^2 M}{\partial z \partial y} & \frac{\partial^2 M}{\partial z^2}
\end{bmatrix}$$ \hspace{1cm} (1)

Where M is a density map. As second-order Gaussian partial derivatives are used, a normalisation factor per element in $H$ should be multiplied by $\sigma^2$ similar to the case of the Laplacian of Gaussian (see Eq. 2.20). The determinant of a 3D matrix is computed from the
matrix’ minors:

\[
|A| = \begin{vmatrix}
    a & b & c \\
    d & e & f \\
    g & h & i
\end{vmatrix} = a(ei - fh) - b(di - fg) + c(dh - eg) \tag{2}
\]

As each minor include the product of three second-order partial derivatives, each normalised by $\sigma^2$, it follows that each scale needs to be normalised by a factor equal to $\sigma^6$. The generation of the scale space is the same as for the Laplacian of Gaussian. A minimum intensity value of $10^{-4}$ is set. Lower values usually lead to extreme amounts of anchors being detected. Results for multi-scale detection with the determinant of Hessian will not be shown. The 3D determinant of Hessian detector performs significantly worse than the LoG detector, with the maximum repeatability between the structure and the simulated map reaching 75% at best. Only with $\sigma_{\text{init}} = 3$ could repeatability values rise up to 25% for experimental maps. Moreover, only the upsampled octave provided repeatable features except for 3j3u at 10 Å, where the reverse is true, surprisingly, suggesting a poor stability of this detector applied to biological structures at varying resolutions.

We show the results of single-scale detection from the determinant of Hessian in Figure A.1. The trend is similar but worse than the LoG operator, although detected correspondences are significantly more numerous with the determinant of Hessian, as evidenced by the similar number of correspondences but lower repeatability. This may come from the combined effect of the minimum intensity threshold, scale normalization by $\sigma^6$ (compared to $\sigma^2$ in the LoG case) and any inconsistent contour level between experimental and simulated densities. Indeed, the voxel values are normalized between 0 and 1 upon loading, and that range is affected by the chosen contour level of the experimental map (see section 2.1.2). Any difference in scaling between simulated and experimental map could be amplified by the high normalization factor, and the minimum intensity threshold may wrongly discard voxels in the reverse case. As previously mentioned, lowering that threshold only produced extreme amount of features without improving repeatability and correspondences.
A. 3D Hessian detector

Figure A.1 – Single-scale detection with the determinant of the Hessian matrix in 3D. Each row corresponds to a system. The PDB and EMD access IDs are shown, starting with PDB 6dbl and EMD 7845 at a resolution 5Å and ending with PDB 3j3u and EMD 5609 at 10 Å. The first column shows the repeatability as defined in Eq. 4.1, for the base, upsampled and both octaves simultaneously at different values of $\sigma_{\text{init}}$. Both octaves have a single scale, blurred with a Gaussian of scale $\sigma_{\text{init}}$. Half of that value is considered for the upsampled octave. Results obtained with the simulated map are shown in shades of red, and those obtained with the experimental map in shades of blue. In the second and third columns are shown the number of correspondences with simulated and experimental maps, respectively. The black bars are the sum of the correspondences obtained individually in the base and upsampled octaves and can be compared to the correspondences obtained when both octaves are used simultaneously.
B 3D Harris corner detector

Harris and Stephens’ corner and edge detector\textsuperscript{106} along with the variant from Shi and Tomasi\textsuperscript{107} have been covered in section 2.4.2. The structure tensor at the heart this detector is well-defined in 3D and can be expressed as follows:

\[
M = \begin{bmatrix}
I_x^2 & I_x I_y & I_x I_z \\
I_x I_y & I_y^2 & I_y I_z \\
I_x I_z & I_y I_z & I_z^2
\end{bmatrix}
\] (3)

Where \( I_x \) denotes the first-order Gaussian derivative of the image in the \( x \) direction. The eigenvalues of \( M \) are extracted at every non-zero voxel, and a response map is made from the smallest eigenvalue \( \lambda_{\text{low}} \) in the sense of the Shi-Tomasi variant\textsuperscript{107}. We show in Figure B.1 the repeatability achieved at different thresholds of \( \lambda_{\text{min}} \), meaning that candidate anchors must display a \( \lambda_{\text{low}} \) higher than \( \lambda_{\text{min}} \). For the sake of simplicity, we show only the results corresponding to the \( \sigma \) value of the Gaussian first-order derivatives that lead to the highest overall repeatability on experimental data (\( \sigma=3 \)). Other values of \( \sigma \) follow the same trend. Increasing the threshold on \( \lambda_{\text{min}} \) reduces repeatability and number of correspondences drastically, which suggests that the high repeatability for lower thresholds is due to the extreme density of detected anchors and that their location is not stable. We show the detected anchors of the simulated density of the structure pair PDB 6qd3 and EMD 4518 in Figure B.2a and immediately observe that features are detected mostly on the external surface of the map and at different locations despite the grids being simulated from the same atomic structure, save a translation and rotation prior to detection to avoid bias from a perfect lattice alignment. Unstable anchors are typically located on edges and planar surfaces, which typically have two and one large eigenvalues, respectively. We show the distribution of the eigenvalues of the detected anchors in Figure B.2b. The latter is observed, and we conclude that this detection method is not suitable for application to low resolution voxel grids.
B. 3D Harris corner detector

Figure B.1 – Single-scale detection with Shi-Tomasi corner detector in 3D. Each row corresponds to a system. The PDB and EMD access IDs are shown, starting with PDB 6dbl and EMD 7845 at a resolution 5Å and ending with PDB 3j3u and EMD 5609 at 10 Å. The first column shows the repeatability as defined in Eq. 4.1, for the base, upsampled and both octaves simultaneously at different values of $\sigma_{init}$. Both octaves have a single scale, blurred with a Gaussian of scale $\sigma_{init}$. Half of that value is considered for the upsampled octave. Results obtained with the simulated map are shown in shades of red, and those obtained with the experimental map in shades of blue. In the second and third columns are shown the number of correspondences with simulated and experimental maps, respectively. The black bars are the sum of the correspondences obtained individually in the base and upsampled octaves and can be compared to the correspondences obtained when both octaves are used simultaneously.
Figure B.2 – Anchors detected by the Shi-Tomasi detector. (a) A render of the system EMD-4518 and its fitted PDB 6qd3. Detected anchors are shown in green and orange for the simulated density and the transformed atomic structure, respectively. (b) The eigenvalues of the anchors detected on the simulated map. Top graph refers to the lowest eigenvalue, the middle graph to the middle eigenvalue, and bottom graph to the highest eigenvalue.


[38] Matteo T. Degiacomi, Ioan Iacovache, Lucile Pernot, Mohamed Chami, Misha Kundryashev, Henning Stahlberg, F. Gisou van der Goot, and Matteo Dal Peraro. Molecular assembly of the aerolysin pore reveals a swirling membrane-insertion mecha…


Bibliography


Bibliography


Bibliography


[152] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau,


Bibliography


Bibliography


CV - SYLVAIN TRAEGE

Nationality  Swiss
Date of birth  1st March 1990
Contact  sylvain.traeger@gmail.com | +41 79 480 65 10
Address  Rue du Sablon 16, 1110 Morges, CH

Strengths
Programming - Machine learning - Structure modeling - Molecular dynamics
Data processing and visualization - Project design and management
Adaptability - Intercultural communication - Teamwork - Technical writing

Education

Ecole Polytechnique Fédérale de Lausanne (EPFL, Lausanne, CH)
Ph.D in Computational and Quantitative Biology  2016 - 2021
   New Methods for the Integrative Dynamic Modeling of Biomolecular Structures
Master in Life Sciences and Technology - Molecular Medicine  2013 - 2016
   Bioinformatics, Computational Biology, Cancer biology, Infectiology, Pharmacology
      (GPA: 5.27/6)
Bachelor in Life Sciences and Technology  2010 - 2013
   Computer programming, Mathematics, Physics, Molecular Biology, Biochemistry
      (GPA: 5.01/6)

Ecole de culture générale et de commerce (Nyon, CH)
   Commercial studies diploma (CFC de commerce)  2007 - 2009

Core experience

Interfaculty Institute of Bioengineering (IBI - EPFL)
   Laboratory for Biomolecular Modeling (Prof. Dal Peraro)  2015 - present
   My objective was to design and manage the development of new
   methods for the analysis of biological structural data from various
   sources. Namely, I developed and published a novel cluster analysis
   method for large database of molecular structures. In collaboration
   with internal and external laboratories, I developed a new integrative
   modeling strategy based on computer vision concepts, genetic
   algorithms and machine learning.
   Notable collaborations:
   Prof. Wah Chiu (Stanford University)
   Prof. Maya Topf (Birkbeck, University of London)

Global Health Institute (GHI - EPFL)
   Laboratory of Molecular Microbiology (Prof. Blokesch)  2014
   2-month internship on the the Type IV pilus and natural competence
   in V. cholerae.
   BSL-2 environment, standard molecular biology techniques
CV - SYLVAIN TRAEGER

Debiopharm Int. SA (Innovation Park, EPFL) 2014
Summer internship. Early PK studies of active compounds
(UV-spectrophotometry and chromatography, MS)

Swiss Institute for Experimental Cancer Research (ISREC - EPFL)
Mol. Mech. of Lung Cancer Development (Prof. Meylan) 2014
Semester project. The antiviral innate immune response in non-small cell lung cancer
(Mice, flow cytometry, cellular and molecular biology)

Mech. of Centrosome Duplication, Cell Division Timing and Asymmetric Cell Division (Prof. Gönczy) 2013
Bachelor thesis, various internships. Spindle positioning and cortical forces in the C. elegans embryo
(DIC microscopy, immunofluorescence, RNAi, worm culture)

Technical skills
Scientific Python (NumPy, SciPy, scikit-learn, pandas, matplotlib, Keras, PyTorch)
Matlab, R, C++
Slurm Workload Manager
Linux (CentOS, Ubuntu), Windows
Molecular dynamics (AMBER, CHARMM, GROMACS)
I-TASSER, SwissModel, ChimeraX, PyMol, VMD
LaTeX, Adobe Illustrator, Microsoft Office

Translational skills
EPFL Open Days (Portes Ouvertes) 2019
Popular science, design and display of educational activities about structural biology through virtual and augmented reality

Student mentoring for Bachelor and Master thesis
Project design, development, coaching, daily supervision, mentoring and support in thesis writing and corrections

Teaching assistant
(EPFL) Biomolecular structure and mechanics (SV Faculty)
(EPFL) Informatique I (SV, MA/PH Faculties)
Supervision of exercise sessions, semester projects, grading and surveillance of examination

Publications
MaD: Macromolecular Descriptors for the modelisation of large and flexible macromolecular complexes
(in preparation)
CV - SYLVAIN TRAEGE

CLoNe: Automated clustering based on local density neighborhoods for application to biomolecular structural ensembles. 
*Bioinformatics*, accepted 21 August 2020

KAP1 is an antiparallel dimer with a functional asymmetry.  
*Life Science Alliance*. 2019

Modelling structures in cryo-EM maps  
*Current Opinion in Structural Biology*. 2019

Disentangling constraints using viability evolution principles in integrative modeling of macromolecular assemblies  
*Scientific Reports*. 2017

Clathrin regulates centrosome positioning by promoting acto-myosin cortical tension in *C. elegans* embryos.  
*Development*. 2014

**Additional certifications**

Machine Learning A-Z (Udemy)  
Deep Learning Specialization (Coursera)

**Languages**

French  (native proficiency)  
English  (full professional proficiency)  
Italian  (limited working proficiency)

**Extracurricular activities**

Music (bass guitar, electric guitar)  
Herpetology  
Study of the Japanese language