The University of Nottingham
UNITED KINGDOM · CHINA · MALAYSIA

Meletiou, A (2018) Large-scale grid computing approaches towards mapping the conformational space of mycolic acids from Mycobacterium tuberculosis. PhD thesis, University of Nottingham.

Access from the University of Nottingham repository:
http://eprints.nottingham.ac.uk/50561/1/submit_to_internal_examiner.pdf

Copyright and reuse:

The Nottingham ePrints service makes this work by researchers of the University of Nottingham available open access under the following conditions.

This article is made available under the University of Nottingham End User licence and may be reused according to the conditions of the licence. For more details see: http://eprints.nottingham.ac.uk/end_user_agreement.pdf

For more information, please contact eprints@nottingham.ac.uk

Large-scale grid computing approaches towards mapping the conformational space of mycolic acids from *Mycobacterium tuberculosis*



Athina Meletiou

A thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

September 2017

Table of contents

Acknowledgments v
Abstract vii
List of abbreviations viii
Chapter 1: Introduction 1
Tuberculosis 1
Mycobacteria 6
MAs: properties and structure-function relationships 11
Experimental and computational conformational studies on MAs 18
Summary 32
Rationale 32
Chapter 2: Theory and Methods 35
Quantum Mechanics 35
Molecular Dynamics and Molecular Mechanics 39
Statistical analysis methods in MD simulations 42
Summary 45
Chapter 3: MD simulations and analysis 47
MD simulations: Introduction 47
Rationale 47
Computing on the UoN HPC facility (local approach) 48
Decision trees used for atom type assignment 51
Establishing effective conformational sampling 56
Simulation details 58
Computing on the WCG (grid approach) 61
Step-by-step build-up of the simulations 62
File handling on the WCG server and file naming convention 64
Simulation details 66
Analysis: Introduction 68
Rationale 68
Analytical approaches used 69
Clustering techniques 69
Intramolecular carbon atom distance matrices 71
Analytical workflow followed 72
Summary 73

Chapter 4: Results and Discussion 74
Results from local approach 74
Results from grid computing approach 97
Discussion 106
Chapter 5: Concluding Discussion 127
Summary of conclusions 127
Outlook and future work 129
Appendix A 135
Appendix B 165
References 171

Acknowledgments

I would like to thank the following people for their contributions towards the completion of this work:

To my supervisor, Dr Croft, thank you for the continued guidance, from the day I arrived in Nottingham as a new student until the submission of my thesis. I am grateful for your feedback and encouragement, especially towards the end. To my supervisor, Dr Jäger, for the tremendous help and patience, particularly with teaching me how to code. I appreciate all the hours you spent explaining basic programming and thank you for enduring all my questions.

To Dr Groenewald, thank you for the amazing communication, especially during my writing up. I am grateful for the discussions and the moral support.

I would like to thank Viktors Berstis, Jonathan Armstrong, Juan Hindo, Caitlin Larkin, Kevin Reed, Al Seippel, Sophia Tu, Erika Tuttle, Keith Uplinger, and the whole IBM team for their tremendous contribution to this work. It would not have been possible to complete this project without their support.

The HPC facility of the University of Nottingham has been instrumental to the completion of this work. Dr Colin Bannister is gratefully acknowledged, particularly for the helpful communication about software and storage requests.

I would like to thank David Burgess, particularly for the support in setting up and running our server.

To Dr Charles Laughton, for all the help with pyPCAzip and to Dr Gareth Shannon, for the help with developing the AWK script.

Funding has been provided by the EPSRC and the Faculty of Engineering.

To all the members of the BECT group, thank you for creating such a friendly and welcoming environment in the office and beyond. My three years in Engineering would not have been the same without you; I definitely miss our lunch break chats!

To my cat Arya, who kept me company for endless hours, sleeping next to me every day as I worked.

To my awesome family, thank you for cheering me on for all those years, and especially during the writing process. To my grandfather, a humble thank you for always believing in me. I know you would have loved to read my thesis, ask me questions, and study the history and origins of the words in the text.

And last but not least, a big thank you goes to my partner, Manos. I can say with absolute certainty that I could not have done it without your incredible encouragement and help this year. Thank you for everything.

Abstract

Tuberculosis (TB) has plagued humanity for centuries, and, despite being largely curable, it still claims thousands of lives daily. The causative agent, *Mycobacterium tuberculosis* (*M. tb*), shows resilience by evading immunological and antitubercular challenges. This is mainly due to its lipid-rich cell wall, whose main components are mycolic acids (MAs). MAs are long fatty acids with functional groups of precise stereochemistry. The MA chemical structure and composition balance in the cell wall modulates antigenicity, cell wall permeability, and virulence. MA chemical structure also steers MA folding. However, MA structure-function relationships are still not fully understood.

In this work, exhaustive MA conformational determination studies were performed in order to generate systematic detail into the correlation of MA structure and conformations, which may in turn hold the key to understanding TB at the atomistic level. Atomistic molecular dynamics (MD) simulations, in different solvents and temperatures, were performed on 166 MAs covering all three *M. tb* MA classes, both the *cis* and *trans* stereochemistries of the proximal cyclopropyl group, and a wide range of chain lengths. Analysis focused on dihedral angle and principal component analysis (PCA) clustering, as well as intra-molecular distance matrix calculations. The results presented in this work confirmed the MA functional groups as folding points and the spontaneity of the MA folding. Three major conformations were identified, namely four-chain, knot, and five-chain folds, with the four-chain fold being the dominant conformation. The knot fold demonstrated definable subtypes, whose further characterisation may offer potential for refinement of current serodiagnostic methods, since MAs have been used as antigens in TB serodiagnosis.

List of abbreviations

AMA	Alpha mycolic acid
AmB	Amphotericin B
BCG	*Mycobacterium bovis BCG*
CG	Coarse graining
CT	Computerised tomography
DFT	Density Functional Theory
ELISA	Enzyme-linked immunosorbent assay
EMB	Ethambutol
HF	Hartree-Fock
HIV	Human immunodeficiency virus
HPC	High performance computing
INH	Isoniazid
KMA	Keto mycolic acid
MP	Møller–Plesset
LCAO	Linear combination of atomic orbitals
MA	Mycolic acid
MAC	*Mycobacterium avium* complex
MC	Monte Carlo
MD	Molecular Dynamics
MDR-TB	Multidrug-resistant tuberculosis
M. kansasii	*Mycobacterium kansasii*
MM	Molecular Mechanics
MMA	Methoxy mycolic acid
MmpL	Mycobacterial membrane proteins, large
M. tb	*Mycobacterium tuberculosis*
NP	Nanoparticle
NMR	Nuclear magnetic resonance
OPLS	Optimised potentials for liquid simulations
QM	Quantum Mechanics
PCA	Principal Component Analysis

PES	Potential energy surface
PLGA	poly(lactic-*co*-glycolic acid)
PZA	Pyrazinamide
PPD	Purified protein derivative
RIF	Rifampicin
RMSD	Root-mean square deviation
RR-TB	Rifampicin-resistant tuberculosis
TB	Tuberculosis
TDM	Trehalose dimycolate (cord factor)
TMM	Trehalose monomycolate
UoN	University of Nottingham
WCG	World Community Grid
WHO	World Health Organization
XDR-TB	Extensively-drug resistant tuberculosis

Chapter 1: Introduction

Tuberculosis

TB is a disease that has existed for millennia and it persists as a major cause of morbidity and mortality worldwide until today. Despite being largely curable, this disease still claims thousands of lives daily. It is estimated that TB killed 1.4 million people in 2015 alone.[1] In the same year, an additional 0.4 million deaths resulted from TB co-infection among patients with human immunodeficiency virus (HIV).[1] Despite the number of TB deaths reducing by 22% between 2000 and 2015, the disease remained one of the top 10 causes of death worldwide in 2015.[1] TB is the most frequent opportunistic infection in HIV patients worldwide.[2]

TB poses a serious threat to countries of the developed and the developing world. According to the World Health Organization (WHO), the incidence of TB cases in 2015 per WHO region was 61% in the Asia Region (referring to South-East Asia and the Western Pacific), 26% in the African Region, 7% in the Eastern Mediterranean Region and 3% in the European Region and the Region of the Americas, respectively.[1]

The most characteristic symptoms of active TB include chronic cough, sometimes with blood-tinged sputum, chest pain, loss of appetite, fatigue, fever, night sweats and weight loss,[3] the latter giving rise to the term *consumption*, an old term for pulmonary TB. TB is primarily airborne and its transmission occurs through droplet aerosolisation by a patient with active pulmonary disease. TB most commonly presents as pulmonary, affecting the lungs, but its clinical manifestations also include extra-pulmonary TB. Common extra-pulmonary infection sites include the lymphatic system, central nervous system, pleura and the skeletal system.[3]

TB has plagued humankind throughout history. Its causative agent may have claimed more victims than any other microbial pathogen, attributing the

nickname "Captain among these men of death" to the disease.[4] TB reached epidemic proportions in Europe and North America during the 18th and 19th century.[4] Strikingly, 19th century societies responded to the epidemic by romanticising it as the disease of artists, a romantic disease that represented sensitivity and spiritual purity.[4]

Mycobacterium tuberculosis (*M. tb*) was identified as the causative agent of TB by Robert Koch in 1882.[5] Koch also extracted a substance from tubercle bacilli that he called *tuberculin*. The development of the first vaccine by Albert Calmette and Camille Guérin[6] from attenuated bovine-strain bacilli in 1919 was the first successful effort at immunisation against the disease. The vaccine was called BCG (*Bacille Calmette-Guérin*) and was first used on humans in 1921 in France, but received widespread acceptance after World War II.[4] Isoniazid, the first oral mycobactericidal drug, was developed in 1952 and rifamycins followed in 1957. The efforts of getting TB under control entered a new era: sanatoria closed, effective public health measures started being implemented, and treatment was expanding to include patients with latent TB infections.[4]

Today, TB diagnosis and treatment remain challenging with several tests usually needed to diagnose active TB. These include chest X-rays, sputum smear microscopy and culture, CT scans, and biopsies. However, some of the diagnostic tests have severe limitations. Smear microscopy has suboptimal sensitivity, successfully detecting only about 50% of TB cases.[7] TB can present with a variety of chest radiography findings[8-9] and its radiologic features can mimic other diseases including bacterial pneumonia, pleural effusion, and lymphadenopathy.[8] Often chest radiography is not enough to reach a definitive diagnosis, and cultures or biopsies are also necessary.[8]

Mantoux skin testing is used for diagnosing latent TB[10-11] and it replaced the Heaf test in the UK in 2005. Both involve intradermal injection of tuberculin-purified protein derivative (PPD) into the inner surface of the forearm to generate a skin reaction indicative of exposure to TB.[10] Mantoux testing

measures the degree of hypersensitivity to tuberculin rather TB immunity itself. The size of the reaction site is correlated with the risk of developing TB.[12]

Mantoux testing has low specificity and most positive skin reactions in low-risk patients are in fact false-positives.[13] Exposure to non-tuberculous mycobacteria or earlier BCG vaccination[14] may yield a false-positive result, even if the vaccine had been administered many years earlier. A negative result typically indicates that the patient has never been exposed to *M. tb*. However, very recent or very old TB infections, recent live-virus vaccinations, and bacterial or viral illnesses can result in a false-negative response.[15]

The standard treatment regimen for active drug-susceptible TB involves four drugs administered over a six month period.[1, 11] Isoniazid (INH), rifampicin (RIF), ethambutol (EMB), and pyrazinamide (PZA) are the first-line anti-TB drugs that form the basis of initial treatment plans.[16] However, the use of multiple drugs over long periods of time can cause serious adverse effects and leads to drug resistance.[17-18] Increased relapse risk, suboptimal response to treatment, and patient non-compliance are also well documented.[16]

Hepatotoxic events are the most serious common adverse effects of first line drugs. INH, RIF, and PZA have been associated with an increased risk of drug induced clinical hepatitis, especially in cases of underlying liver disease.[16] PZA has also been strongly associated with induced arthralgia and exanthema.[19] A substantial number of patients suffered from severe exanthema even when taking a reduced dose of PZA.[19]

Adverse reactions to first-line anti-TB drugs can cause life-threatening complications and may result in patient withdrawal from one of the basic drugs.[19] Complications can be managed if high risk patients are monitored in a hospital setting, but drug termination can lead to grave consequences. Although the terminated drug may be compensated by using an alternative, for example streptomycin, the patient will not be receiving the best available treatment against TB infection.[19] There is a crucial need to optimise current

treatment regimens and to ultimately develop novel, more efficient and targeted antibiotics.

Drug-resistant TB

Antimicrobial resistance threatens the effective diagnosis, prevention and treatment of TB. The emergence of drug-resistant TB strains is putting another obstacle in the efforts to control and ultimately eradicate the disease. In addition to multidrug-resistant TB (MDR-TB), which is resistant to front-line drugs isoniazid and rifampicin, extensively drug-resistant TB (XDR-TB) strains have been identified since 2006.[20] XDR-TB is resistant to isoniazid, rifampicin, any of the fluoroquinolones, and to at least one second-line drug, such as kanamycin.

Potential reasons for the emergence of drug resistant TB strains include patient non-adherence to the full course of the primary regimen,[17-18] failure to identify pre-existing resistance, use of an inadequate first-line treatment plan,[21] or limited access to sufficient and/or high quality drugs.[21] In 2015, there were an estimated 480,000 new cases of MDR-TB and 100,000 cases of rifampicin-resistant TB (RR-TB).[1] Although countries have been starting more patients on resistant-TB treatment over the last years, only an estimated 125,000 patients of those 580,000 resistant TB cases were reported to have been placed on second-line TB treatment plans.[1]

Resistant TB patients need treatment plans that are longer (9-12 months), more expensive, toxic, and less effective and accessible than first-line regimens.[22] There is a clear vicious circle in play: long treatment times and drug adverse effects are contributing to the emergence of MDR/RR-TB strains. At the same time, drug-resistant TB cases are tackled with inefficient drugs and even longer treatment plans. It has been shown that drug-resistance can be contained if appropriate prevention and control measures are implemented.[23] However, given the strong global presence of drug-resistant TB and the suboptimal treatment success rates, shorter treatment plans and more efficient drugs are

urgently needed to provide less toxic, less costly, and more effective treatment.[24]

HIV / TB coinfection

A 'syndemic', or synergistic epidemic, is defined as the combination of two or more diseases that act synergistically to augment the burden of disease in a population.[25] The association of TB with HIV has developed into a syndemic in recent years. Approximately 10% of patients living with HIV and with a latent TB infection, will develop active TB each year.[26] TB has been documented as the most common opportunistic infection in HIV patients worldwide. At the same time, HIV is the strongest risk factor for the development of TB.[2] In 2015, the African WHO Region had the highest HIV-associated TB burden, with 81% of TB patients having a documented positive HIV test result.[1]

HIV and TB interaction is governed by a bidirectional mechanism.[2] HIV critically impairs host immune responses to *M. tb*,[27] causing failure to restrict *M. tb* replication. Increasing immunodeficiency is correlated with increasing mycobacterial burden. At the same time, activation of mononuclear cells during the host immune response to TB leads to accelerated HIV replication.

Clinicopathological TB presentation is often unmodified in HIV patients with well-preserved immunity, but progressive immunodeficiency increases the likelihood of atypical TB presentation.[27] The clinical presentation of the disease tends to be more subtle and nonspecific and chest radiographs can appear normal sometimes despite extensive pulmonary infection.[2] Additionally, concurrent administration of antiretroviral and anti-TB drugs to patients with severe underlying disease often leads to adverse reactions.[28]

HIV infection is not only magnifying the TB problem but it has also been linked with MDR-TB incidents and with acquired rifamycin resistance.[29] HIV co-infection may be directly contributing to an increased MDR-TB occurrence among TB patients. A better understanding of the link between HIV infection

and resistant TB is crucial for the management of the resistant-TB epidemic and the future of HIV treatment.

Mycobacteria

The etymology of the word 'mycobacterium' echoes early observations that the growth characteristics and morphology of the bacterium resemble that of fungi.[30] Mycobacteria, a genus currently containing over one hundred and fifty recognised species, can be differentiated based on their phenotypic presentations. They can be classified into three principal groups: (i) strict pathogens, including the human pathogens *M. tb* and *M. leprae*, and the animal pathogen *M. bovis,* (ii) opportunistic pathogens, including *M. avium,* and (iii) rare pathogens, including saprophytes such as *M. smegmatis*.[31]

A common characteristic among mycobacterial species is their intrinsic antibiotic resistance. The chemical composition and structural organisation of the mycobacterial cell wall promote very low permeability and therefore contribute to drug resistance.[32] Additional exceptional properties observed in mycobacteria include resistance against chemical injury from various toxic compounds,[33] acid fastness,[34] dehydration resistance[35] and persistence within the host macrophage environment.[36-37] These can be directly attributed to the structure and composition of the *M. tb* cell wall and, more specifically, to the dominant presence of some unique fatty acids, called mycolic acids (MAs), in it.[38]

M. tb can persist within the host for decades despite a strong immune response. It has been shown that a mycobacterial gene cluster (mce4) enables the bacterium to degrade cholesterol from the host membranes and derive the carbon and energy necessary for its survival.[37] Cholesterol is not required for establishing infection in mice or for growth in macrophages, but it is crucial for persistence in the lungs of chronically infected animals and for growth in the macrophages.

The *M. tb* cell wall is exceptionally rich in lipids. MAs are the major components,[39] making up almost 30% of the bacterium dry weight[40] and over 40% of the cell envelope weight.[39] MAs occur in a distinct Gram-positive bacteria group classified in the suborder of *Corynebacterineae*, including the genera *Mycobacterium*, *Nocardia*, *Rhodococcus,* and *Corynebacterium*, mentioned in the order in which they contain large-and-complex to small-and-simple MAs.[41] In *Corynebacterium*, MAs provide a beneficial but not crucial permeability barrier,[42] whereas mycobacterial species depend on MAs for survival *in vitro* and growth *in vivo*.[43-44]

Mycobacterial MAs are among the most functionalised and long mycolates in nature[38] and they are composed of two aligned chains. The shorter one typically contains 24-26 carbon atoms and is a saturated, non-functionalised chain with a carboxylic acid motif. The longer one, the 'meromycolate' chain, typically contains 40-60 carbon atoms. It has a proximal and a distal position that can bear functional groups, indicated as [X] and [Y] respectively in Figure 1 below.

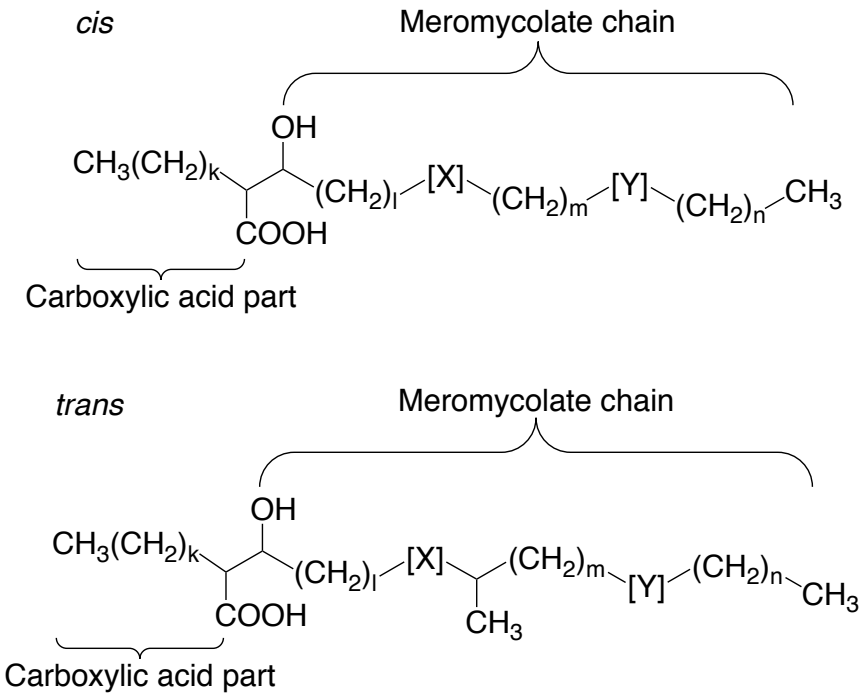


Figure 1. Schematic representation of the chemical structure of MAs, modified from Villeneuve *et al.*[45] [X] and [Y] represent the proximal and distal functional groups. The *k, l, m,* and *n* notations represent alkyl chain lengths. A complete list of the *k, l, m,* and *n* alkyl chain lengths of the different MAs modelled in this work can be found in Figure S1 in Appendix A.

MAs demonstrate variations in terms of functional group nature and stereochemistry, and their overall length. Depending on the distal functional group, MAs found in *M. tb* are classified in three main classes, as shown in Figure 2: (**1**) alpha- (AMA), where the distal group bears a *cis*-cyclopropane, (**2**) methoxy- (MMA), with the distal group substituted by a methoxy-methyl group, and (**3**) keto- (KMA) MAs, with the distal group substituted by a keto-methyl group. AMAs are the most abundant in many mycobacterial species.[39] All the pathogenic members within the *M. tb* complex (*Mycobacterium tuberculosis, africanum, bovis* and *microti*) synthesise a combination of the three MA classes, whereas the attenuated vaccine strain, *M. bovis* BCG Pasteur, only produces AMA and KMA.[46]

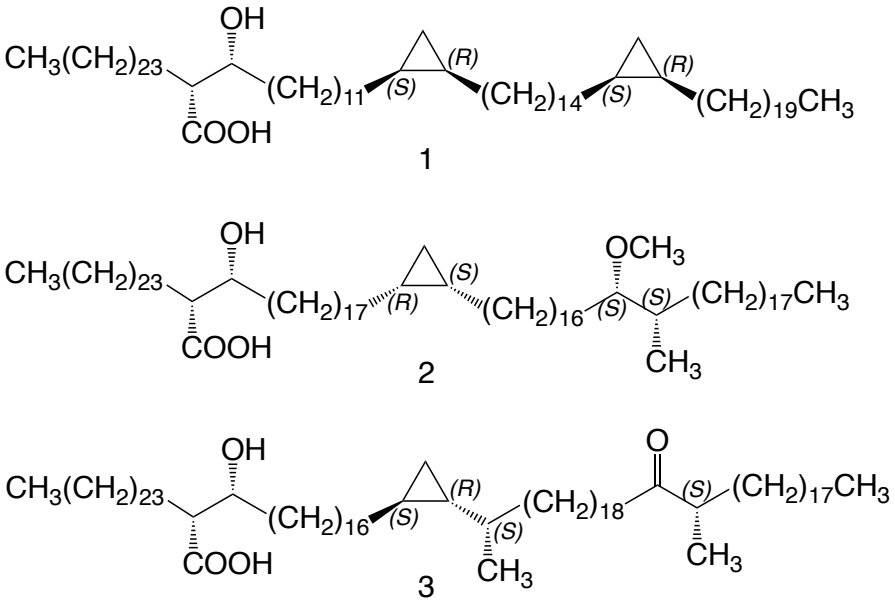


Figure 2. Chemical structures of the main components of MAs from *M. tb*. Adapted from Groenewald *et al*.[47] MA class is assigned according to the nature of the distal functional group, namely *cis*-cyclopropane (**1**, AMA), methoxy-methyl (**2**, MMA), and keto (**3**, KMA).

AMAs in *M. tb* have a *cis*-cyclopropyl group in the proximal position, whereas MMA and KMA can have either a *cis*-cyclopropane or an α-methyl *trans*-cyclopropyl group. The proximal (or distal, in AMA) position can be a double bond but the main presentation in most mycobacterial species has either *cis*- or *trans*- cyclopropane rings and lacks double bonds.[48] Mycobacteria have the

longest MAs, ranging from 60-90 carbon atoms, *Nocardia* have 46-60, *Rhodococus* 34-64 and finally *Corynebacterium* has the shortest MAs at 22-38 carbon atoms.[49]

In MMA and KMA, the oxygenated functional groups have an adjacent methyl group. MMA and KMA natural mixtures contain both *cis*- and *trans*-proximal cyclopropane rings, the latter also having an adjacent methyl group. Extensive NMR and mass spectroscopic analyses[48, 50] have been carried out to determine the nature and location of the groups in the meromycolate chains of MAs from representative mycobacteria. The major components of natural MAs from *M. tb* contain a *cis*-cyclopropane for MMA and an α-methyl *trans*-cyclopropane for KMA (Figure 2). The relevant stereochemistries have been identified through comparison with synthetic compounds.[51-54]

The mycobacterial cell wall has a complex structure that comprises three covalently linked building blocks: peptidoglycan, arabinogalactan and MAs. The peptidoglycan is linked to arabinogalactan *via* a phosphodiester bridge. Most MAs are covalently bound to peptidoglycan *via* arabinogalactan.[55] This covalent linkage creates a hydrophobic layer of exceptionally low fluidity,[56] which is referred to as the 'mycomembrane' (Figure 3). MAs can therefore be found in the outer layer of the lipid-rich mycobacterial cell wall and they form the inner layer of the mycomembrane. The low fluidity and permeability that MAs confer to the cell envelope, contributes to the antibiotic resistance of these bacteria.[38]

The outer part of the mycomembrane contains extractable lipids, most of which are mycobacteria-specific,[56] and porins.[57-58] The outermost cell-wall compartment, called the capsule, mainly contains polysaccharides (glucan and arabinomannan). MAs can also be found as trehalose mono- (TMM) or dimycolates (TDM)[41, 59] and as free hydroxyl acids. TMM consists of a trehalose sugar and one MA molecule, while TDM consists of a trehalose sugar and two of a set of MAs.

The current perception of the ultrastucture of the mycobacterial cell wall is shown in Figure 3. X-ray diffraction studies[60] have shown that MAs are packed side by side, parallel to each other, and perpendicular to the cell envelope plane. Thus, the complete structure is a bilayer, in which the MAs form the inner leaflet. Freeze-fracture studies[61] identified a second fracture plane in electron micrographs, confirming the existence of an outer hydrophobic bilayer structure, external to the plasma membrane. The mycobacterial cell wall is thus analogous to the Gram-negative bacterial cell wall, since it features an outer membrane covalently bound to peptidoglycan.[62] The outer layer of extractable lipids and porins which intercalates into the MA layer forms a "pseudoperiplasm".[63] Mycobacteria such as *M. tb*, *M. bovis*, *M. smegmatis*, and *M. avium* utilise drug efflux pumps that confer resistance from a variety of antibiotics and other toxic substances by pumping them into the pseudoperiplasm.[33] The mechanisms that the pumps employ to yield resistance are not fully understood. It is not yet known how such high levels of resistance are achieved simply by moving the antibiotics only into the pseudoperiplasm, inside the cell wall.[33]

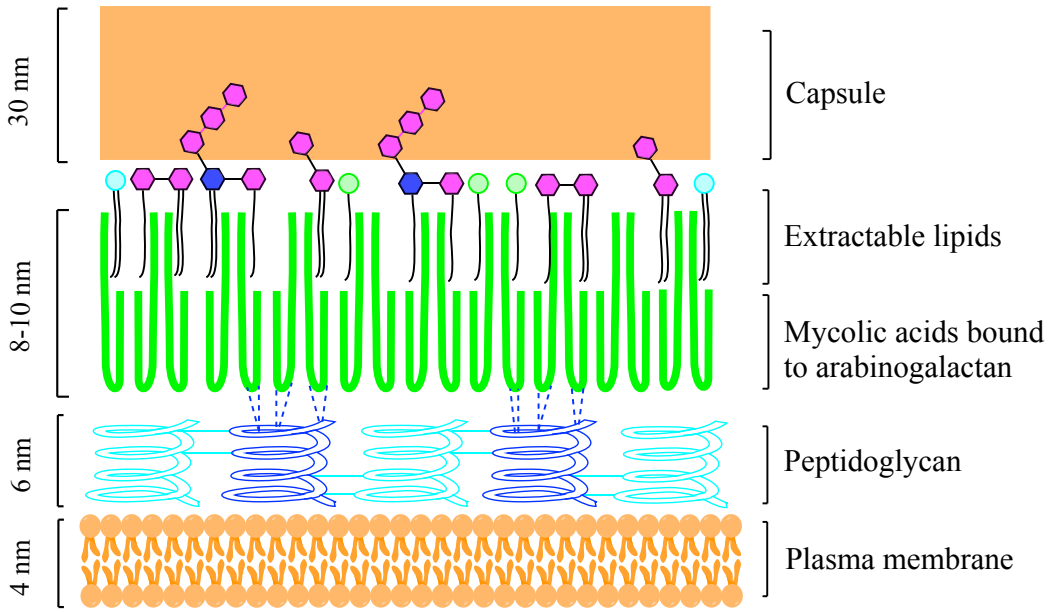


Figure 3. Schematic representation of the current view of the mycobacterial cell envelope. Adapted from Abdallah *et al.*[56] The measurements of the different cell envelope compartments[56, 64] highlight the need for tight packing of MAs inside the mycomembrane, which in turn results in a hydrophobic layer of extremely low fluidity.

TMM is a key precursor to TDM biosynthesis. At an initial step in the biosynthesis process, MAs are esterified to synthesise TMM.[38] TDM, also known as 'cord factor', is the most abundant extractable lipid from virulent *M. tb*, accounting for 90-95% of the extractable lipids from viable organisms.[65] Because of its extremely long alkyl chain length, ranging from 140-184 carbon atoms, TDM creates a significantly hydrophobic cell-surface environment and contributes to the overall low fluidity of the mycomembrane. TDM is highly immunogenic and has important biological activity in mycobacteria and other related genera.[66] Antibodies to TDM have attracted much attention as new surrogate markers that can aid in TB diagnosis.[67-68]

TDM was initially found as a virulence factor of the highly virulent cord-forming *M. tb* but it has been demonstrated that it is widely distributed among non-virulent mycobacteria and related microorganisms.[39, 69] The chemical and molecular properties of TDM are inextricably linked to the structure of the component MAs. It has been demonstrated that the chain length in mycoloyl glycolipids directly modulates microbial pathogenesis and toxicity,[70] thus indicating a strong relationship between chemical structure and biological function.

MAs: properties and structure-function relationships

MAs have some exceptional chemical characteristics that contribute to tighter packing, thus leading to low fluidity inside the *M. tb* cell wall inner leaflet.[71] Additional to their extraordinary chain lengths, MAs contain very few double bonds or cyclopropyl groups. Their shorter chain is always saturated and the long meromycolate chain has only two positions that can be substituted by a double bond, cyclopropyl group or oxygenated group. In the bilayer model of the mycobacterial outer cell wall, the MA-containing inner leaflet has very low fluidity, since lipids with saturated, longer carbon chains are packed more tightly.[71]

The outer leaflet, containing extractable lipids with shorter chains, is expected to demonstrate higher fluidity. Thus, there is a steep fluidity gradient across the thickness of the mycobacterial cell wall. This reveals another (apart from

the existence of an external bilayer, as discussed previously) analogy to the Gram-negative bacteria outer membrane, even though the fluidity gradients have opposite orientations.

MAs have been suggested as being important virulence factors of *M. tb.* TDM and glucose monomycolate from *Nocardia* and *Rhodococcus* demonstrated granuloma-forming activity in mice and elicited an immune response *in vitro.* Mycoloyl glycolipids possessing trehalose or glucose as a carbohydrate moiety exhibited immunomodulatory activity, and those with longer MAs demonstrated stronger activity in mice than those with shorter MAs.[72]

MAs have also shown potential to be considered for use in TB serodiagnosis, with anti-MA antibodies having been detected in human serum in an ELISA-based assay. Schleicher *et al.*[73] performed a study to explore the potential of utilising MAs as surrogate markers for TB, by detecting anti-MA antibodies in the sera of patients with active pulmonary TB, in a population with a high prevalence of either or both TB and HIV. Their work showed that TB/HIV co-infected patients maintained high MA antibody levels, demonstrating that natural mixtures of free MAs may offer potential to be used as antigens for TB serodiagnosis,[68, 73] even in countries with a high HIV burden such as Africa.[73-74] Disappointingly, the ELISA-based assay used by Schleicher *et al.*[73] demonstrated poor sensitivity and specificity and was deemed unsuitable to yield a reliable serodiagnostic result.

A possible explanation for the poor performance of the ELISA-based assay was the potential interaction of MAs with cholesterol. It was hypothesised that cholesterol may bind to MAs either *via* a non-specific van der Waals interaction, or *via* a more specific one, such as a hydrogen bond.[73] Benadie *et al.*[66] demonstrated that cholesterol and MAs were recognised equally well by Amphotericin B (AmB), which is a cholesterol-binding agent. Further, it was shown that cholesterol was attracted to immobilised MAs. It was thought that the attraction between MAs and cholesterol likely depended on MAs assuming

a rigid conformation,[66] similar to the folds that have been observed in MA monolayers.[45, 75]

Despite the low accuracy of the assay used in the work of Schleicher *et al.*,[73] the premise that the antibody response to MAs is preserved in HIV patients despite progressive immunodeficiency, makes the underlying principle of this assay a worthwhile avenue to explore in the efforts of TB diagnosis in HIV-burdened populations. Additionally, the association between cholesterol and MAs emphasises the need for better knowledge of the most commonly adopted MA conformations. This can aid the development of new serodiagnostic methods, since the basis of antigen-antibody recognition relies on structural affinity and specificity.

Mathebula *et al.*[74] demonstrated that MAs could be strongly integrated into a self-assembled monolayer (SAM) and used for detection of anti-MA antibodies in serum samples from HIV-positive patients. This study did not employ ELISA-based assays but used immunosensors, thus avoiding the limitations that the assays have shown in earlier works.[73]

Lemmer *et al.*[76] developed a novel nanoformulation for use as an anti-TB drug delivery system. Bone marrow-derived mouse macrophages, either uninfected or infected with different mycobacterial strains (*M. avium*, *M. bovis* BCG, or *M. tb*), were used. The infected cells were exposed to encapsulated isoniazid (INH)-poly(lactic-*co*-glycolic acid)polymer (PLGA) nanoparticles (NPs), with or without MAs as the targeting ligand. It was demonstrated that the encapsulation of the hydrophilic INH was greater in the NPs containing MAs. This directly implies that MA addition enhances drug loading and, thus, can increase the drug amount transferred into the infected macrophages. The inclusion of MAs into PLGA also increased the phagocytic uptake of NPs in infected macrophages and, thus, can result in more NPs transferring active drugs into the mycobacterium containing phagosomes. This work showed that innovative nanotechnologies may be able to overcome the limitations of

existing therapy regimens and contribute to the development of novel TB treatments.

In recent years, MmpL3 has received significant focus as a promising new target for novel anti-TB drug development. MmpL3 is part of the MmpL (mycobacterial membrane proteins, large) proteins family.[77] It is a vital inner membrane transporter of the resistance, nodulation, and division (RND) superfamily of transporters. The MmpL proteins play a crucial role in the building of the cell envelope of mycobacteria, being involved in the export of TMM to the periplasmic space, where the TMM can function as a MA donor in the building of the mycomembrane.[77-78]

Belardinelli *et al.*[77] demonstrated that impaired MmpL3 activity results in a decrease in MA transfer to arabinogalactan. Li *et al.*[78] investigated whether MmpL3 is essential for the replication and persistence of *M. tb in vitro* and *in vivo*. Their results indicated that silencing *mmpL3* resulted in a rapid bactericidal effect on replicating cells *in vitro*, and reduced the bacterial loads of acutely and chronically infected mice. Further, depletion of MmpL3 rendered *M. tb* hyper-susceptible to MmpL3 inhibitors, underlining the therapeutic potential of MmpL3.

Degiacomi *et al.*[79] confirmed the essentiality of *mmpL3* for *M. tb.* Downregulation of *mmpL3* stopped bacterial division and resulted in rapid cell death. MmpL3 was also shown to be essential for growth in human macrophages. MmpL3 depletion led to the repression of genes involved in MA and TDM biosynthesis. Further analysis indicated that MmpL3 depletion also led to modification of surface permeability. These studies confirm that MmpL3 is essential for *M. tb* to build a correct cell wall and therefore it is vital for cellular viability. Their results validate MmpL3 as a target of therapeutic interest with the potential to improve current TB chemotherapies.

Pathogenic mycobacteria exist inside eukaryotic cells in what can be described as a fairly nutrient-rich and privileged habitat. However, oxidative stress is a

major factor associated with this environment. Cyclopropanation of MAs occurs predominantly in pathogenic mycobacteria and has minor effects on the physical properties of lipids when compared with the corresponding olefin. Olefinic lipids should be more vulnerable to oxidative stress than cyclopropanated lipids, therefore this structural adaptation may be predominantly an adjustment to survive oxidative stress.[80]

The chemical structure of MAs and their composition balance in the cell wall have been shown to determine *M. tb* antigenicity, survival *in vitro*, growth *in vivo,* persistence in the host macrophage and establishment of a lethal infection. Data have shown that when *M. tb* strains lack MAs completely, they demonstrate increased cell wall permeability[81-82] and decreased virulence, thus underlining the importance of the cell wall structure for *M. tb* survival.[38] The discussion that follows addresses the effects of MA structure on mycobacterial infection and survival. However, at the same time it demonstrates that the structure-function relationships that determine these results remain complex and not fully understood.[41]

Yuan *et al.*[83] have studied the effects of complete KMA depletion in BCG and *M. tb* strains. Both mutant strains produced only MMA and AMA and demonstrated poor growth *in vitro* at reduced temperatures. Sensitivity to INH was unaffected but the cells were found to be hypersensitive to ampicillin and rifampicin. Finally, the KMA-depleted strains demonstrated severely reduced ability to grow within macrophages and glucose uptake was also significantly decreased. The above results confirm that mycolate composition plays a critical role in cell wall function and growth *in vitro* and *in vivo.* A distinct biological role for KMA was also revealed by Dkhar *et al.*[84] KMAs were shown of being capable of inducing foam cell biogenesis and contributing to granuloma formation. The foamy macrophages present around TB granulomas in chronic infection provide the bacterium with a rich nutrient source.

Dubnau *et al.*[46] used MAs from *M. tb* H37Rv to study the effect of MA oxygenation on cell wall permeability and virulence in mice, by synthesising a

mutant *M. tb* strain that lacked the *hma* gene. The *hma* gene is responsible for coding an enzyme involved in a key-step in all oxygenated MA biosyntheses. The mutant strain that was produced demonstrated significant alterations in its envelope permeability and virulence. Compared to the wild type, it was shown to be attenuated in mice and it grew slower in the lungs. This study not only illustrated that MA composition has a decisive effect on *M. tb* antigenicity and virulence, but also established *hma* as a good target for new antimycobacterial drug development.

Genomic studies have shown that *hadC*, a gene encoding a subunit of the HadBC dehydratase, is mutated in the attenuated *M. tb* H37Ra. Slama *et al.*[85] demonstrated that mutation or deletion of *hadC* substantially reduced the biosynthesis levels of MMA. Additionally, *hadC* mutation was shown to alter the fine structure of KMA and MMA, with mutant strains exhibiting a total loss of the *trans*-cyclopropane rings. AMA structures were not greatly affected. Further, *hadC* deficiency affected bacterial growth in biofilm communities, called pellicles, resulting in reduced thickness of the biofilm, and caused significant loss of virulence in mice. Finally, *hadC* deletion increased susceptibility to rifampicin.

Various studies have suggested that *M. tb* grows in biofilm communities rather than in dispersed cultures. It has been demonstrated that *in vitro* biofilms contain free MAs probably released from the enzymatic hydrolysis of TDM.[86] These biofilms have also been shown to harbour a drug-tolerant *M. tb* population that persists despite exposure to antibiotics.[87] Sambandan *et al.*[88] demonstrated that the ability of *M. tb* to synthesise KMA is essential for pellicle growth. In the same study, it was further shown that growth within this community confers drug resistance to the bacilli.

Vander Beken *et al.*[89] identified a strong relationship between the chemical structure and inflammatory function of MAs. By using single synthetic MA isomers of identical chain lengths to the most abundant *M. tb* MA homologues, they investigated the effect of oxygenation and cyclopropane stereochemistry

of MAs on host pulmonary responses. The oxygenated classes showed activity whereas AMAs were inert. KMAs with a *cis*-cyclopropane caused only mild inflammation whereas *trans*-cyclopropanated KMAs demonstrated an anti-inflammatory behaviour. In MMAs, *cis*-cyclopropanation resulted in a solid pro-inflammatory response, whereas *trans*-cyclopropanation resulted in partial loss of inflammatory activity.

Synthetic MAs have also been used to investigate recognition by patient serum antibodies. The study by Beukes *et al.*[90] determined that the fine structure of MAs is important for antibody recognition. Oxygenated classes were found to be more antigenic than AMAs. In both MMAs and KMAs, the α-methyl *trans*-cyclopropane in the proximal position of the meromycolate chain resulted in higher antigenicity than *cis*-cyclopropanation. Synthetic MMAs were the most antigenic and the ones bearing a *trans*-cyclopropane were even more antigenic than the natural mixture.

A mycobacterial gene, *pcaA*, has been shown to be required for cording and MA cyclopropanation in the cell wall of both BCG and *M. tb*.[91] Genetically modified mycobacteria from *M. tb* lacking *pcaA* replicated normally *in vivo* but failed to persist within the host and kill infected mice. These results defined cyclopropane synthetases as a novel class of persistence genes necessary for successful *M. tb* infection and outlined a role for cyclopropanated lipids in bacterial pathogenesis.

Different mycolates contribute differently to the cell function and cell wall permeability. In pathogenic mycobacteria, the stereochemistry and nature of the functional groups of MAs, as well as the varying chain lengths, play a significant role in determining the final conformation and packing of the chains.[47, 75] These features also determine virulence and establish a host-pathogen relationship that causes particular disease presentation.[92] The MA structure-function relationships are only beginning to be elucidated. MA modelling can provide insights into the correlation of MA conformation and

function and their complex structure-function relationships, ultimately aiming to contribute to a better understanding of the disease at the atomistic level.

Experimental and computational conformational studies on MAs

Little experimental data on the conformational patterns of MAs is available. This is because MAs are waxy and therefore they do not form crystal structures. Additionally, their hydrophobicity hinders solubility in most solvents and therefore leads to limited options in terms of suitable structural determination applications. MAs are, however, soluble in hexane and chloroform which makes them suitable for use as surfactants in monolayer experiments. MA monolayers on water[45, 75, 93-98] have been widely studied and they have generated extensive data about MA conformational preferences, underlining the importance of MA fine structure.

During a monolayer experiment, a solution of the surfactant in a suitable solvent is deposited onto a water layer of known surface area. Commonly, surfactants are dissolved in decane, or, less frequently, in chloroform. As the solvent evaporates, a barrier is used to apply lateral pressure. The monolayer forms at the air-water interface and a Langmuir film is created. As lateral pressure is applied, the average area per molecule decreases and the pressure in the monolayer increases. With increased lateral pressure, the molecules become packed increasingly closer together, their hydrophobic tails are straightened and stretched out in parallel, and the hydrophilic heads are coming in contact with the water surface. A thin plate, usually on the order of a few square centimeters in area and often made from filter paper or glass, called Wilhelmy plate, is used to measure the equilibrium surface or interfacial tension.

The various thermodynamic phases of the monolayer can be studied by plotting the surface pressure of the Langmuir monolayer versus the average area that the molecule is occupying, creating a graph depicting an isotherm, as seen in Figure 4. Surface pressure-average area isotherms can provide a measure of the average area per molecule and the compressibility of the

monolayer. They also contain information on phase transitions and on the collapse points of the monolayer. Monolayer experiments can therefore provide information on physical properties of MAs and can shed light into their conformational preferences, as shown in the discussion that follows. They, therefore, can provide a good approximation of the arrangement of MAs inside the cell wall, and the resulting data provide a substantial framework for theoretical models of MAs.

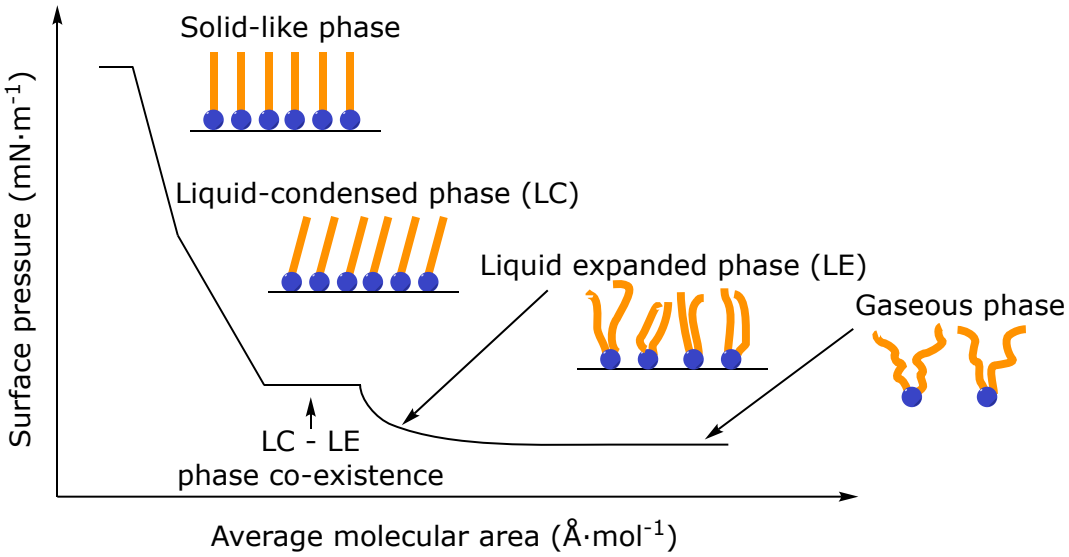


Figure 4. Idealised schematic representation of a general surface pressure-average area isotherm of a typical amphiphilic molecule, obtained by compressing an insoluble lipid monolayer formed at the air-water surface. The distinct regions of the isotherm follow the different level of ordering of the film during the experiment. Regions with a constant slope denote a homogeneous phase in the monolayer, whereas a change in the slope denotes phase transition at that pressure. In a typical amphiphilic molecule, the isotherm can be divided into a gaseous phase at low surface pressure, a liquid-expanded (LE) phase characterised by a decrease in intermolecular distance, a plateau region (LC-LE phase co-existence), a liquid-condensed (LC) phase characterised by maximum packing and minimum molecular area at the interface, and a solid-like phase, before the monolayer collapses.[99-100]

Langmuir monolayer experiments,[45, 75, 93-98] conducted at various different temperatures and targeted at MAs originating from a range of mycobacterial species, demonstrated that, in general, MAs tended to assume folded, energetically favoured conformations at low surface pressure and temperature conditions. Villeneuve and coworkers proposed[45, 75] the low-energy, tightly folded pattern to be a "W" conformation, where the MA functional groups

served as bending points and the four hydrocarbon chains were packed parallel to each other (Figure 5).

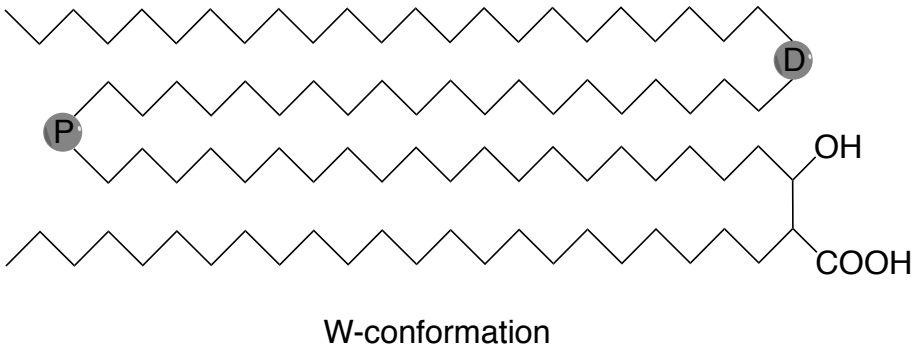


Figure 5. Schematic representation of a MA in the low-energy "W" conformation, as proposed by Villeneuve and coworkers.[45, 75] The proximal and distal functional groups are shown as P and D respectively. Adapted from Groenewald *et al*.[101]

The results of the monolayer experiments[45, 75, 93-98] were similar for all MAs despite the different origin, chain length, functional group nature and stereochemistry. The π (surface pressure) vs. A (average molecular area) isotherms indicated that the molecules assumed tightly folded conformations at low temperature and surface pressure conditions. Interestingly, MMAs and AMAs unfolded into stretched-out conformations as temperature and surface pressure increased.[45, 75, 93-97] KMAs were the ones that retained the tight folds even at high temperature and surface pressure,[45, 75, 94, 97] thus demonstrating a unique trend correlating with a potentially distinctive biological effect.

It was assumed that the carbonyl group of KMAs was hydrated at the water surface, stabilising the molecule and thus accounting for KMAs assuming constrained W-folds.[45, 75] Inside the mycobacterial cell envelope, this interaction was hypothesised to occur either with arabinogalactan or with the MA hydroxy group, in an inter- or intra-molecular fashion, respectively. It was thought that MMAs and AMAs assumed more flexible conformations because of the lower degree of hydrophilicity of their functional groups. This also implied that the presence of oxygenated groups in MAs may be to promote conformational stabilisation through hydrophilic interactions. The α-methyl *trans* cyclopropyl

group was also revealed to contribute to the KMA compact folding by creating a hairpin-like kink.[94]

The thermodynamics study conducted by Villeneuve *et al.*[45] analysed AMA, MMA and KMA from *M. tb* and it revealed exceptionally temperature-dependent isotherm profiles. The relevant phase diagrams of the MA insoluble monolayers were quite different from each other. AMA and MMA films featured a wide liquid condensed phase region, whereas KMA films demonstrated a solid condensed and a liquid condensed phase. These results depict the relative flexibility of AMAs and MMAs, along with the rigidity of KMAs. The same team conducted a monolayer study[75] that focused on oxygenated MAs from BCG. This work also underlined the preference of KMAs to retain W-folds even at high temperature and surface pressure conditions. The results from complementary computer simulations performed in the same study[75] further confirmed the results of the thermodynamic analysis. KMAs retained their W-folds whereas MMAs gave either four-chain or stretched-out conformations, as seen in Figure 6.

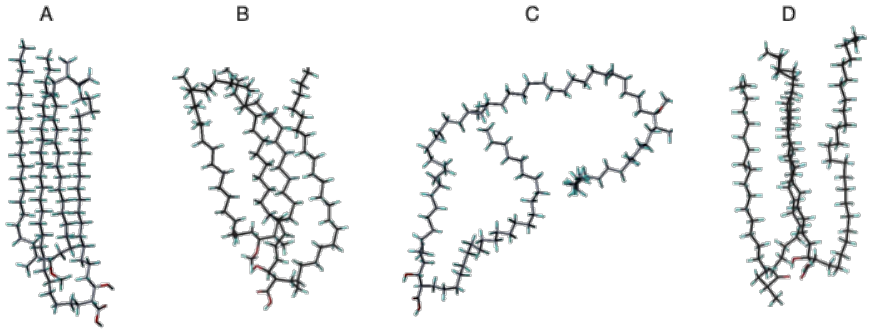


Figure 6. Resulting structures[75] of oxygenated MAs from BCG after having undergone Monte Carlo (MC) and MD simulations. (A) MMA, starting W-conformation, structure derived after MC calculation and minimisation, (B) MMA, retained W-fold after 20 ps MD, (C) MMA, stretched-out structure after further 20 ps MD, (D) KMA after 20 ps MD.

The conclusions regarding the molecular folding were reached by following a five-point (*a-e*) distance convention scheme as shown in Figure 7. According to this scheme, the five points represented: (a) the last carbon of the carboxylic acid part, (b) the carbon bearing the carboxyl group, (c) the distal carbon of the

proximal cyclopropane ring, (d) the carbon bearing the keto- (KMA) or the methoxy-group (MMA), or the distal carbon of the distal cyclopropane (AMA), and (e) the last carbon atom of the meromycolate chain. Resulting structures were classified into subgroups, namely W-folds, in-between structures or stretched-out structures, if they satisfied specific requirements for in-between *a-e*-point spacing. If the resulting structure did not fit into any of the subgroups it was described as a non-definable structure.

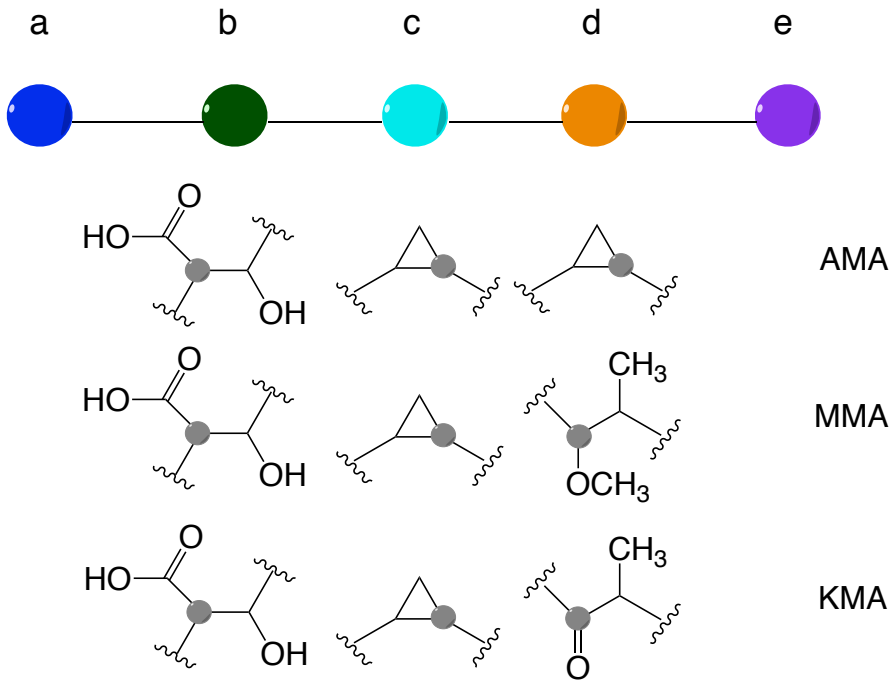


Figure 7. Schematic representation, modified by Groenewald *et al.*[47], of the a-e five-point distance convention used in the conformational analysis by Villeneuve *et al.*[75] Eight key distances (ab, bc, cd, de, ac, ae, ce, and bd) can be calculated from the five a-e reference points to describe each MA fold.

The potential intra-molecular hydrogen bonding that was discussed earlier may have contributed to an extent to the retaining of the W-fold in KMAs. However, one of the major factors that could explain the different behaviour between MMAs and KMAs were the different lengths of the individual methylene chain segments (*l-m-n* values) and specifically the relationship between the *l* and *m* chains. The major and minor components of BCG MMAs and KMAs, as established by Watanabe *et al.*,[50] correlated with their relevant

minimised energy levels after MD,[75] can be found in Table 1. As can be seen in Table 1, in MMAs *l* is larger than *m*, whereas in KMAs *l* is smaller than *m*. There appears to be a correlation between chain length and minimised energy value, which may contribute to the retention of the W-fold in KMAs. The value for *l* being smaller than *m* seems to correspond to significantly smaller energy levels (-117.20 to -159 as opposed to -29.30 to -50.21 kJ/mol) in the molecule.

Sample	**Cyclopropane stereochemistry and l-m-n values of major & minor MA components**	**Minimised energy (kJ/mol)**
MMA	cis 17-16-17 (major)	-54.39 to -79.50
	trans 17-16-18 (minor)	-29.30 to -50.21
KMA	cis 17-18-15 (major) or cis 17-18-17 (major)	-100.40 to -133.9
	trans 17-18-16 (minor)	-117.20 to -159

Table 1. Major/minor BCG MA components[50] and relevant minimised energies after MD.[75]

What the *l*>*m* relationship in MMAs practically means is that two bulky groups, the methoxy and its adjacent methyl group, are found at a position where they can obstruct the compact four-chain arrangement. In KMAs however, the relationship of the notations is *l*<*m*, which means that the bulky carbonyl and its adjacent methyl group either lie far down towards the end of the chain or stretch out of the four-chain structure. This may allow for a tighter arrangement and fewer vibrations that could lead to retention of the four-chain arrangement during the MD simulations.[75]

These complementary monolayer and computational studies on MAs have generated useful insights into the factors that affect the folding preferences. Fine structural details such as methylene segment chain lengths and nature of the distal functional group have been shown to affect the folding of MAs. KMAs have exhibited the tendency to retain a tight W-fold whereas AMAs and MMAs were shown to be somewhat more flexible.

Another monolayer and computational study by Villeneuve *et al.*[93] focused on the conformational behaviour of AMAs from different mycobacteria, namely *M. tb, M. kansasii, M. avium-intracellulare complex* (MAC), and *M. bovis* BCG (BCG), over a range of temperatures and covering a variety of different chain lengths. The results of the thermodynamic study were alike for all AMAs and agreed with the results from monolayer experiments in previous studies.[45, 75] The molecules assumed four-chain conformations at low temperature and surface pressure conditions but tended to unfold as temperature and surface pressure increased.

This work was the first published quantum chemical DFT study on a part of a MA structure. Villeneuve and coworkers[93] analysed the possible conformations of the mycolate moiety and the *trans* and *gauche* were found to be the most stable ones. MA models were then built assuming the *trans* and *gauche* conformations. The molecular models were prepared by MC calculations and were then subjected to MD simulations. The simulations confirmed the results of the thermodynamic study by attributing W-shaped structures to all AMAs. More specifically, after being subjected to the MD, applying no restrictions, all the *trans* and *gauche* models retained the starting W-shape structure.

After applying restrictions (on the carbon dihedral angle in $-\mathrm{CH_2}-\mathrm{CH(COOH)}-\mathrm{CH(OH)}-\mathrm{CH_2}-$ and to the range of the distance between the hydrogen of 3-hydroxyl group and the carbonyl oxygen of 2-carboxyl group), the *gauche* form molecules expanded to stretched-out forms and almost all the *trans* retained the W-conformation. AMA from *M. tb* and BCG tended to unfold more readily than those from *M. kansasii* and MAC. Since the molecules were of varied chain lengths, the analysis focused on six pre-determined intra-molecular distance points (Figure 8) in a slight variation of the *a-e* distance scheme seen earlier in Figure 7. In the *a-f* distance convention scheme, the six points taken into consideration are defined as follows: (a) the l^{th} carbon from *b*, (b) the carboxyl-bearing carbon, (c) the distal carbon of group Y, (d) the distal carbon of group X, (e) the m^{th} carbon from *d*, and (f) the last carbon of the meromycolate chain.

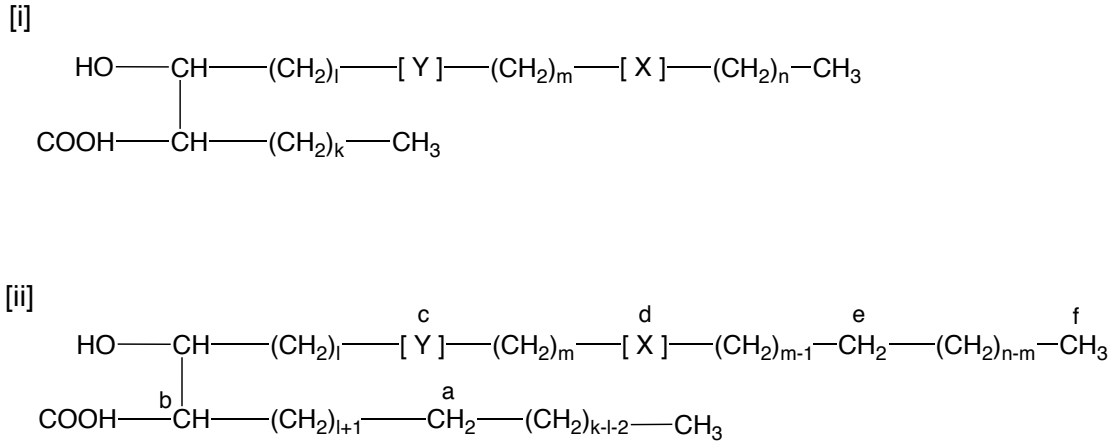


Figure 8. Schematic representation of the six-point *a-f* distance convention used by Villeneuve *et al.*,[93] as a slight variation of the analogous five-point *a-e* distance scheme proposed by the same group. (i) AMA general structure, where the various alkyl chain lengths are *k*=21 or 23, *l*=11–17 (odd numbers), *m*=14 or 16, and *n*=17 or 19. (ii) AMA general structure with markers *a–f* used to identify the six distance points used. In both (i and ii), [X] and [Y] denote the proximal and distal functional groups respectively.

A summary of some representative resulting conformations can be seen in Figure 9. These structures were defined, according to the six-point *a-f* variation of the *a-e* convention of Figure 7, as W-folds or extended shapes. Extended shapes consisted of extended (eS), folding extended (Se) or fully extended (Sf). The conformational patterns established that AMAs largely tended to assume W-folds but they stretched out quite readily to adopt different extended shapes in response to environmental temperature and pressure. Specifically, tuberculous AMA extended more readily than those from *M. kansasii* and MAC.[93] The overall flexibility of AMA may directly affect the physical properties of the lipid bilayer in the mycobacterial cell envelope. In the outer membrane lipid layer, AMAs may change their conformation from the W-shape to an extended shape to adapt to changes in the environment. The fact that AMAs are the most abundant MA class in many mycobacterial species[48] may imply that they are of significant importance to the structure-function relationships that govern the inner leaflet of the mycobacterial cell wall.[93]

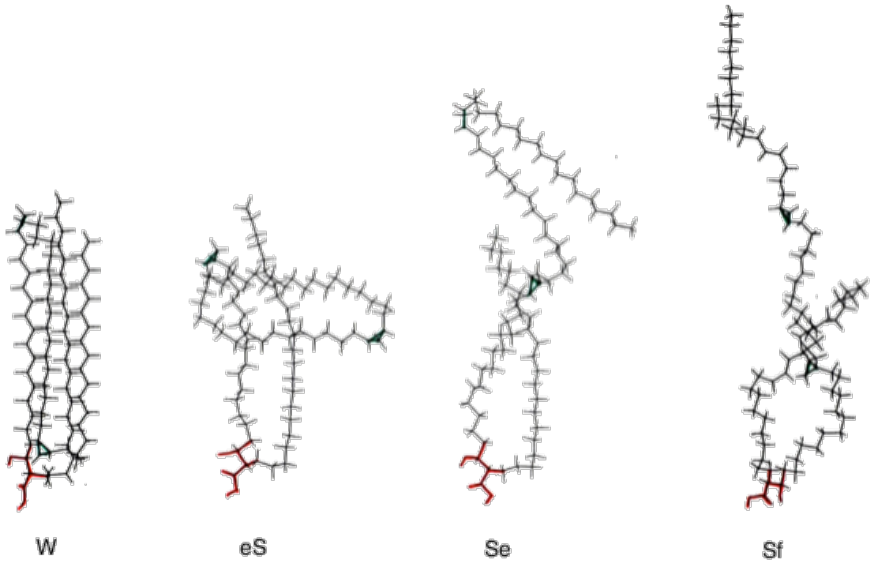


Figure 9. Representative conformations of AMAs from *M. kansasii* and *M. tb,* resulting from MC and MD simulations.[93] The conformational changes observed include the transformation from the starting W-shape to an extended shape (eS), then to a folding extended shape (Se), and in some cases to a fully extended (Sf) shape.

The length differences between the individual methylene chain segments was hypothesised to account for the resulting folding conformations.[93] Within the *M. tb* set of AMAs studied, which exhibited a tendency to unfold more readily than lipids from other species, those with shorter *l* segments unfolded quicker. Strikingly, in a previous study conducted by Villeneuve *et al.*[75] it was concluded that the $l > m$ relationship in MMAs allowed for potential obstruction of the compact four-chain arrangement, whereas in KMAs the $l < m$ led to retention of the four-chain fold. What becomes clear from this contradicting behaviour between different MA classes is that their inherent fine structure differences in combination with the different chain lengths contribute to varying trends in the conformational profiles of each molecule.

Another experimental monolayer study combined with MD simulations performed by Villeneuve *et al.*[94] focused on the effects of *cis*- as opposed to α-methyl *trans*-cyclopropanation on the proximal position of the meromycolate chain. The samples used in the monolayer experiment were MMA from the attenuated *M. tb* H37Ra strain, *M. canettii*, and *M. kansasii*, and KMA from *M. tb* H37Ra and MAC. The resulting data showed that α-methyl *trans*-

cyclopropane facilitated MA folding more than *cis*-cyclopropane did and resulted into more molecules assuming tighter U-shaped conformations. The presence of an adjacent methyl group to the methoxy and keto units in the case of the *trans* cyclopropanated mycolates might also have played a key role in allowing them to fold more efficiently.

Small model molecules were built to computationally assess the bend-forming capacities of the different stereochemistry cyclopropane.[94] Those model compounds had short alkyl chains of 3-5 carbon atoms and contained *cis*- versus α-methyl trans-cyclopropane. The results echoed the findings of the monolayer study by concluding that the *trans* moiety facilitated tighter U-shaped folds.

While the computational studies examined in the discussion above have provided an excellent framework in MA modelling and have generated intriguing results, they also had some disadvantages. These are mainly the very short simulation times that ranged up to a maximum of 35 ps,[93] the very little variation of starting conformations[75, 93] and the inconsistent number of replicates for different MAs.[75] Additionally none of these studies makes an explicit reference to the solvation scheme used, and therefore it is assumed that the simulations were performed either in vacuum or in implicit solvent.

Another computational effort in the study of MA conformational preferences was made by Groenewald *et al*.[47, 101] Groenewald investigated the conformational patterns and the structure-function relationships of MAs by employing a variety of computational methods over a range of levels of theory, namely electronic structure methods, atomistic MD simulations and coarse-grain simulations. The results obtained from the electronic structure methods[101] confirmed that the *cis* and the α-methyl *trans*-cyclopropyl groups provide distinct folding points in the MA molecules.

All three major *M. tb* MA classes were subjected to MD simulations both *in vacuo*[47] and in different explicit solvents, namely water and hexane.[47, 101] This

work therefore assessed the effects of intrinsic (MA chemical structure) and extrinsic (different solvent polarity) factors on the resulting MA conformations. Each 4 ns run started from an open structure.[47] The results showed that MAs folded spontaneously into reproducible conformational groupings. KMAs behaved differently to AMAs and MMAs, implying a potential distinct biological role.

For analysis,[47] five reference points were defined to identify the chain termini (**a**/**e**) and the specific atoms in the linking functional groups (b–d), as was proposed by Villeneuve *et al.*[75] and as can be seen in Figure 7. Eight key distances (ab, bc, cd, de, ac, ae, ce and bd) were used to describe the relevant MA folds. The idealised folding models seen in Figure 10 demonstrate the three general types the folds can be assigned to, namely 'W', 'U' and 'Z' overall shapes. Within the U-shape type, 'aU' and 'eU' subclasses have their extended chains terminated by **a** and **e** respectively, while 'sU' is symmetrical. Similarly, 'sZ' is symmetrical and 'aZ' and 'eZ' subclasses have their extended chains terminated by **a** and **e**, respectively.

Statistical analysis of the equilibrated datasets showed that AMAs had the highest frequency of 'WUZ'-folds (7.93% of structures), with lower frequencies occupied by MMAs and KMAs (1.35% and 3.01% respectively). Representative structures from the 'WUZ' folds of AMAs can be seen in Figure 11.

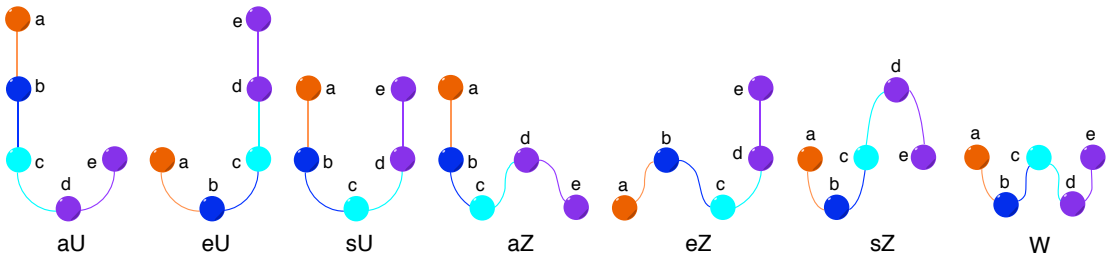


Figure 10. Idealised conformational arrangements of MAs with interacting parallel chains. The proposed folding types, 'W', 'U', and 'Z', can be collectively summarised as 'WUZ'. Modified from Groenewald *et al.*[47]

PCA analysis revealed that the unfolded starting structures progressed over the course of the simulation to folded ones, mostly because of interchain interactions. This suggests that, as expected, the weak van der Waals interactions *in vacuo* play a major role in modulating the folding preferences.[47] PCA analysis showed that KMAs folded more rapidly than AMAs and MMAs, implying that this MA class could have distinctive biological behaviour.

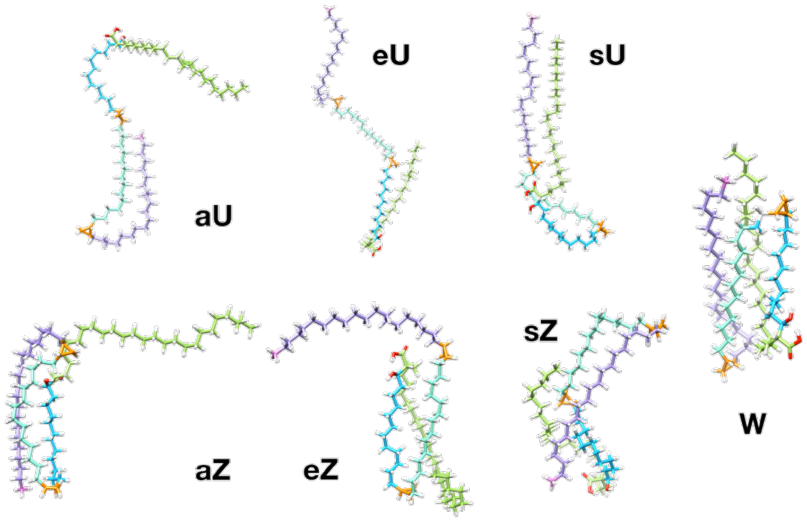


Figure 11. Representative 'WUZ' folds for AMA from *M. tb.* Adapted from Groenewald *et al.*[101] The 'W', 'U', and 'Z' folds were defined by implementing the model of the W-fold, as proposed by Villeneuve, while also taking into consideration structures where folding occurs at the MA functional groups.

AMAs exhibited the highest level of 'WUZ' folds,[47] it was the only MA class that folded into all of the seven possible 'WUZ' folds and it also spent more time in open conformers compared to MMAs and KMAs. AMAs have also been shown to be the least immune active.[89] Villeneuve *et al.*[75, 93] also concluded that AMAs are the most flexible compared to oxygenated MAs and may change their conformation from the W-shape to an extended shape to adapt to changes in the environment. MMAs showed folding *via* distinct folding pathways but also demonstrated the lowest percentage of 'WUZ' folds.[47] KMAs were shown to assume more compactly folded conformations.

Collating the results from the vacuum, water and hexane runs, Groenewald *et al.*[47, 101] concluded that all MAs preferred to assume W-folds at the highest

percentage, which aligns with the observation of Villeneuve *et al.*[75] that the four-chain conformation is of the lowest energy and thus constitutes the most stable conformer. A control group lacking meromycolate chains exhibited no 'WUZ' folds, therefore confirming that the functional groups provide clear bending points in the molecules.

The results from the explicit water solvation were quite similar to the ones observed *in vacuo*. Molecules mostly tended to assume W-conformations, even though there was naturally a higher level of flexibility observed in vacuum. Explicit hexane solvation resulted in mostly stretched-out conformations, with the exception of KMA which preferred assuming the 'sU' folding type. A summary of the observed results can be seen in Table 2.

MA	Solvent	W	aZ	eZ	sZ	eU	sU	aU	Total %
AMA	Vacuum	7.9	1.6	0.2	0.5	0.0	1.0	0.3	11.5
	Water	9.0	5.4	0.0	0.3	0.1	0.0	0.2	15.0
	Hexane	0.0	0.5	0.1	0.3	2.7	1.0	5.8	10.4
MMA	Vacuum	10.3	0.7	0.7	0.0	0.2	0.9	0.0	12.8
	Water	4.4	0.1	1.4	0.0	0.5	0.0	0.0	6.5
	Hexane	0.1	0.5	0.4	0.1	1.8	1.7	2.3	6.9
KMA	Vacuum	39.5	7.3	0.6	0.0	0.0	0.4	0.2	48.0
	Water	19.2	7.7	0.0	0.0	0.0	0.3	0.5	27.7
	Hexane	0.2	7.3	1.1	0.3	0.9	16.4	3.3	29.4

Table 2. Collated MD results of the percentage of 'WUZ' folds obtained for different classes of MA molecules in three solvent environments (water, vacuum, and hexane). Adapted from Groenewald *et al.*[101]

It is obvious from the table above that a large proportion of the resulting conformations could not be assigned to the 'WUZ'-folding models.[101] They were demonstrated to twist and bend at various points in the alkyl chains, thus consisting of various conformations and complex patterns of interacting chains that were difficult to define in two dimensions. These structures tended to be of higher energy compared to the 'WUZ'-folds. It was assumed that these newly

defined folds may be associated with mycobacterial biofilms where free MAs are present, since most single MA conformations are globular in nature.[101]

The results obtained from the atomistic MD simulations illustrated the distinct conformational differences among the three MA subclasses from *M. tb.* These differences can at least partially account for the relevant observed biological differences among MAs. Groenewald *et al.*[101] also presented a coarse-grained model for an AMA from *M. tb* using the MARTINI[102-103] force field. Parameterisation of bonded interactions was obtained from a single AMA in different environments, using atomistic MD simulations as reference. This model can simulate insoluble monolayers of different sizes and at varying temperatures, providing a comparison to existing experimental monolayer data for AMAs and a basis for future *M. tb* membrane simulations models.

The approximation that both cyclopropyl beads were treated as identical in the model may have led it to demonstrate less folding at the cyclopropyl groups. The folding analysis showed that that the MAs were mostly folded only at the head group and the two parallel chains came closer together and packed tighter as the surface area decreased. Despite the fact that the model could not account for explicit MA stereochemistry, it did replicate the basic findings of previous experimental monolayer studies.

The limitations of the work conducted by Groenewald *et al.*[47, 101] lie in the fact that simulations performed *in vacuo*[47] are likely to allow MAs for much more flexibility than if they were modelled in any solvent. Additionally, vacuum is not an entirely accurate representation of the MA natural environment in the membrane. The 'WUZ' analysis may also carry some limitations, as Groenewald showed that MAs largely assumed conformations outside the 'WUZ' scope. Finally, while the eight-point distances examined offer a certain level of detail, they lack the unbiased information that carbon atom distances from the backbone chain of the molecule can provide.

Summary

The discussion above has provided a comprehensive overview of the most current and relevant experimental and computational studies on *M. tb* and, more specifically, MA conformations. Further, this section has illustrated that the chemical structure of MAs and their composition balance in the *M. tb* call wall modulates and determines various aspects of infection and host-pathogen interaction. At the same time, it has been shown that MA chemical structure determines the range of MA conformational preferences, which in turn supports the premise that there is strong interplay between MA conformation and function.

Rationale

The work presented in this thesis focuses heavily on developing a systematic methodology for MA modelling. A total of 166 different MAs have been modelled, which include (i) a large variety of different chain lengths, (ii) all three major MA classes from *M. tb*, and (iii) both proximal cyclopropane stereochemistries. According to current literature, this is the first attempt to model such an extensive array of MAs. Previous computational studies have been performed on limited numbers of varying molecules and have not attempted to cover such a huge variety of different mycolates.

A disadvantage observed in previous computational works was the inefficient sampling of the starting conformations. A common concern in MD is structures getting trapped in local energy minima. This trapping can impair initial sampling, which is crucial for conformational determination studies. A methodical multi-step approach has been employed to ensure efficient sampling and to generate a complete representation of the conformational variability that MAs can exhibit. This approach consists of (i) an extensive variety of MA chain lengths being modelled, (ii) preliminary implicit solvation to allow molecules to explore the relevant conformational space freely, (iii) application of specialised software for effective starting conformation extraction, (iv) exploration of three solvent environments, (v) assessment of temperature effect on the folding.

Additionally, most of the previously discussed computational studies have only consisted of short MD simulations. As it has already been shown and is echoed by the findings of this work, MAs follow a complicated folding pathway towards their final conformational preference. This pathway involves many folding and unfolding events and it is therefore crucial to allow the molecules to explore the conformational space and generate a comprehensive representation of their folding variability. In order to avoid missing important events in the folding pathways, one of the two computational approaches undertaken in this work consisted of performing single 100 ns MD simulations.

Two different computational workflows have been employed. These were (i) single long MD simulations performed on the University of Nottingham high performance cluster, and (ii) multiple shorter simulations conducted in parallel on a distributed computing platform. This is the first attempt to perform MA conformational studies on a worldwide distributed computing platform and the outcomes will provide the groundwork for the computational setup of future MA simulations.

The analytical methods used in this work consisted of a multi-step and multi-approach process. A novel analysis protocol was implemented, focusing on the integrated application of dihedral angle clustering, PCA clustering, and the generation of time-resolved intramolecular carbon atom distances within the backbone chain. The integrated utilisation of the above tools aims to map the conformational space of MAs effectively, to identify major clusters in the folding profiles, and to provide detailed insights into the frequent folding and unfolding events. All previous computational studies on MAs, as they have been discussed earlier in this section, have employed distance-convention schemes focusing on pre-determined points between which the distances were calculated, namely five-point distances or eight-point distances. The current work therefore introduces a significant level of novel detail as the distances between all carbon atoms along the MD trajectory are calculated and assessed.

This work aims to utilise large-grid systems to employ atomistic MD simulations in order to map the conformational space of MAs efficiently. By exploring the conformational flexibility and folding preferences of MAs, as these are modulated by intrinsic (MA chemical nature) and extrinsic (temperature and solvent) factors, this work is in an excellent position to provide novel information and insights into MA structure and folding patterns. The chemistry and composition balance of MAs inside the bacterium has been shown[90-91, 104] to determine antigenicity, survival *in vitro*, and growth *in vivo*, as it has been discussed in detail earlier in this section. At the same time, the chemistry of MAs modulates their conformational preferences,[45, 47, 75, 101] while the MA structure-function relationships are only beginning to be unravelled. Therefore, the correlation of MA structure and conformation can be anticipated to have a distinct biological significance. It becomes clear that obtaining thorough and systematic detail into the correlation of MA structure and conformations may hold the key to a better understanding of TB at the atomistic level.

Chapter 2: Theory and Methods

Theoretical and Computational Chemistry

Theoretical chemistry can be defined as the mathematical description of chemistry.[105] It is a branch of chemistry in which mathematical methods are combined with fundamental laws of physics in order to study chemical phenomena.[106] Computational chemistry is a subfield of theoretical chemistry that mainly focuses on answering chemical questions by performing calculations. While there is strong interplay between theoretical and computational chemistry, the latter is not aiming at developing new theoretical methods but rather contains a strong practical element. Theoretical methods are translated into computer software that are employed in order to produce results.[106]

The main strength of computational chemistry is the ability to generate data from which the researcher can gain insight and therefore understand the behaviour of large classes of molecules. On the other hand, one of the main problems in computational chemistry is selecting a suitable theoretical method for a given problem and evaluating the quality of the resulting data.[106] Molecular modelling encompasses the theoretical chemistry methods and computational techniques that are utilised to study the structure, properties and behaviour of molecules.

Quantum Mechanics

Quantum Mechanics (QM) is a field of physics that quantifies the behaviour, interaction, and motion of atoms and subatomic particles. The mathematical models that QM employs describe molecules as a collection of nuclei and electrons.[106] The basis for nearly all computational chemistry methods is the fundamental time-independent Schrödinger equation. In its most simplified form it can be described as

$$\widehat{H}\Psi = E\Psi \qquad (1)$$

where $\hat{H}$ is the Hamiltonian operator, which operates on the wavefunction Ψ, and E is the energy of the system.[105] The Hamiltonian $\hat{H}$ is the sum of the kinetic and potential energy contributions of all the particles of the system and therefore is dependent only on the fixed properties (e.g., mass) of the electrons and nuclei. The wavefunction Ψ, describes the allowed spatial distribution of the electrons and nuclei of the system.[107] Therefore $\hat{H}$ is the "known" and Ψ and E are the "unknowns" of Schrödinger's equation. Solving this equation can, in principle, provide answers to essentially all chemical problems.[107]

Max Born's statistical interpretation of the wavefunction $\Psi(x,t)$ states that $|\Psi(x,t)|^2$ gives the probability of finding the particle at point x, at time t.[108] This introduces a notion of indeterminacy into QM, as the output of Born's law is essentially statistical information about the probability of the particle's position and not a certain calculation.[108] To this day the hydrogen atom, a one-electron system, is the only atomic or molecular species for which exact solutions of Schrödinger's equation have been derived,[107] but there have been numerous approximation methods developed in order to obtain the solution for multiple-electron systems.[105]

Ab initio methods

When the methods used to explicitly solve the Schrödinger equation do so without the use of any empirical parameters, meaning without any reference to experimental data, then they are called *ab initio*.[109] This term is derived from Latin and means 'from the beginning'. Of course this does not mean that the computations are exact and that no approximations are used, as the Schrödinger equation cannot be solved without approximations for more than a two-body system.[109]

The various different quantum chemical models available utilise different approximations. Ultimately, they demonstrate different reliability, capability, and cost in terms of computational power, time, and storage needed. No single model is suitable to provide correct insights to all chemical questions and the final choice lies on the accuracy and cost trade-off.[110] *Ab initio* methods are

computationally intensive, which hinders their ability to handle large sized molecules.[109]

The Schrödinger equation for multi-electron systems

Due to the large difference in mass, the motion of atomic nuclei is very slow relative to the motion of the electrons. Therefore, the first approximation that can be made to the Schrödinger equation is that the nuclei do not move and the electron motion is determined for a fixed position of the nuclei. This approximation, probably the most famous one in quantum chemistry and molecular physics, is called the Born-Oppenheimer approximation.[109] It is further assumed that electrons move independently of each other and their motions are separable.[109] If different arrangements of the nuclei are adopted and the calculation is repeated, then the Schrödinger equation can generate a set of solutions. These can then construct a molecular potential energy curve for a diatomic molecule, or a potential energy surface (PES) for a polyatomic molecule. The wavefunctions that result from this computation are called molecular orbitals.[109]

Another approximation that helps generate a solution to the Schrödinger equation for a multi-electron system is the Linear Combination of Atomic Orbitals (LCAO). It is a technique that constructs molecular orbitals as a linear combination of atomic orbitals.[111] Molecular orbitals can then be quantitatively calculated using the Hartree-Fock (HF) model. The HF model is an *ab initio* technique that attempts to solve the many-body problem by transforming it into a pseudo one-particle system by taking only the average electron-electron interaction into consideration.[106] In other words, this means that a distinct electron only "sees" the mean field of the other electrons.

One of the limitations of HF calculations is that they do not include the explicit electron-electron interaction.[105] The HF method provides a good understanding of the system and forms the basis for more refined computational methods[106] such as the Møller–Plesset (MP) method for correcting the HF wavefunction. The MP method corrects the HF calculations by adding the electron correlation

using the perturbation theory[109]. Energy corrections may be calculated to the second (MP2), third (MP3) or higher order.[107]

Semi-empirical methods

HF models also form the basis for the development of semi-empirical methods, which introduce additional approximations as well as experimental data to the calculation. One approximation that reduces the size of the problem is taking into account valence electrons only and ignoring core electrons. This simplifies QM calculations[110] and reduces the computational cost[106, 109] while having minimal adverse effects on the results.[110] A widely used semi-empirical method is Austin Model 1 or AM1.[112]

Semi-empirical methods can be considered a blend of *ab initio* methods and molecular mechanics because they encompass some *ab initio* elements but they replace some integrals with experimental data.[109] They are more capable in handling large sized molecules compared to *ab initio* models that are more computationally intensive.[109] Semi-empirical models can reliably calculate equilibrium geometries but they do not generate satisfactory results for thermochemical calculations.[110]

Density Functional Theory

The premise behind Density Functional Theory (DFT) is that the energy of a molecule can be determined from the electron density instead of a wave function.[105] In DFT, electrons interact with one another and with an external potential. In a molecule, this external potential is the attraction to the nucleus.[113] Hohenberg and Kohn provided the basis for the theory which was further developed by Kohn and Sham to produce Kohn–Sham DFT (K-S DFT).[105] In 1998, Kohn received the Nobel Prize in Chemistry for the Kohn–Sham equations and other work related to DFT. In K-S DFT, the expression of the ground-state electronic energy is divided into individually calculated components. Namely, the electronic energy occurs as the sum of the kinetic energy from the movement of electrons, the electron-nuclear attraction and nuclei-nuclei repulsion, the electron-electron repulsion (Coulomb self-

interaction), and the exchange-correlation energy.[110] Calculating the exchange-correlation energy is the main challenge in DFT and different methods to achieve it exist.

Molecular Dynamics and Molecular Mechanics

Molecular dynamics (MD) simulations study the time-dependent behaviour of a molecular system and they most often use molecular mechanics (MM) calculations to compute the energy of the system.[105] In principle, MD simulations could be performed for all the methods discussed above, but MM (or classical MD simulations) are most often used for larger systems. MD methods numerically solve Newton's second equation of motion for atoms on an energy surface by propagating a starting set of coordinates and velocities by a series of finite time steps. A trajectory of the atoms is then generated, that consists of a series of time-correlated points in phase space.[106] MD simulations can provide detailed information about all the particles in the system, making them suitable for calculating collective properties, and one of the most commonly used methods for conformational sampling.[106]

Molecular mechanics, or force field models, can be thought of as the alternative to quantum chemical models. Their starting point is a chemically reasonable depiction of molecular structure, rather than an exact theory such as the Schrödinger equation.[110] The underlying mathematical principle in MM describes molecules in terms of atoms and bonds whereas in QM molecules are regarded as a collection of nuclei and electrons.[106] MM methods use non-quantum (i.e. classical) mechanics to model molecular systems and calculate their geometries.[109] Processes that involve electronic rearrangements, such as chemical reactions, cannot be described at the MM level since electrons are ignored in MM force fields. These processes require a QM description of the electronic degrees of freedom.[114]

Quantum and classical mechanics are fundamentally different in the sense that classical mechanics is deterministic, while QM is also deterministic but the interpretation is probabilistic.[106] Deterministic means that Newton's equation

can predict where the particle is at a certain time. QM, on the other hand, can only calculate the probability of the particle being at a certain place at a certain time.[106, 108] In other words, a classical MD simulation with given coordinates and velocities for all particles must always result in the same result at time t.

Force fields

The fundamental assumption in MM models is based on the observation that molecules are made up from structurally and chemically similar units and therefore those properties are transferable. The basis of the transferability of properties is implemented into force fields[106, 110] in which atoms correspond to different atom types depending on their atomic number and bonding.[106] Force fields are the foundation that underlies MM models. They include parameters derived either from experimental (spectroscopic) data or *ab initio* calculations,[105] namely components that describe bond-stretching, angle-bending and torsional motions.[110] They also include non-bonded van der Waals and Coulomb interactions, as well as distortions from ideal bond distances and angles.[110]

The MM or force field energy (E_{FF}) expression does not use a wave function but consists of a simple algebraic equation:[105]

$$E_{FF} = E_{str} + E_{bend} + E_{tors} + E_{vdw} + E_{el} \qquad (2)$$

where E_{str} represents the energy for stretching a bond between two atoms, E_{bend} represents the energy required for bending an angle, E_{tors} is the torsional energy for rotation around a bond, E_{vdw} (van der Waals), and E_{el} (electrostatic) describe the non-bonded interactions.[106] In other words, the MM energy can be described as the sum of contributions resulting from distortions from an idealised form. This idealised form consists of "ideal" bond distances, bond angles and torsion angles, resulting in stretch, bend, and torsion contributions respectively, along with contributions due to non-bonded interactions (van der Waals and Coulomb).[110] The force field energy expression can differ between

force fields, namely in regards to the way in which the constants are derived and the number or the complexity of the energy terms.[105]

The main advantage of MM methods over QM methods is the speed at which calculations can be performed. This makes MM methods suitable for handling large systems[106] even in molecules comprising millions of atoms. MM can in fact model complex macromolecules, such as proteins and DNA segments which makes MM methods the primary tool for computational biochemists[105] and the pharmaceutical industry.[106] Recently, all-atom MD simulations of the complete HIV-1 capsid were reported.[115] The system modelled consisted of 64,423,983 atoms in explicit water, which was subjected to 1.2 μs MD. Notably, the size of a trajectory file was 315 Gigabytes for every 8 ns of simulation, and the performance yielded an impressive 8 ± 2 ns/day for the 64 million atom simulation.

Another main advantage of MM models over QM models is that force fields can be parameterised. This enables the development of force fields that are able to reproduce known experimental data accurately.[110] However obtaining good parameters is an ongoing and significant challenge. MM methods on the one hand perform well in predicting the properties for classes of molecules where empirical information is available, but on the other hand their scope is limited when the molecule of interest is even slightly unusual.[106]

'All-atom' force fields provide parameters for every single atom in the system. If simplification of the governing theoretical models is necessary, force fields offer the option to obtain a simplified representation where all the atomistic detail is not included. 'United-atom' force fields are designed for treating macromolecules and they introduce simplification by not considering hydrogens explicitly. A step towards further simplification is the generation of so called 'coarse-grained' (CG) force fields.[116] CG force fields do not provide an atomistic description of the system, but combine two or more atoms into groups called 'beads'. This approach significantly decreases the number of particles in the system and thus increases the potential system size to be

calculated using similar resources. The first step towards CG is implemented in most common atomistic force fields, such as AMBER, CHARMM and GROMOS,[106] by neglecting C - H bond vibrations.

Statistical analysis methods in MD simulations

From the first simulation of an enzyme[117] in 1977, MD simulations have come a very long way. Today, atomistic MD simulations of systems that contain millions of atoms[115, 118-120] are possible. The modelling timescales have also become significantly longer with the microsecond scale being practical.[121-122] It is clear that MD simulations at the atomistic level have routinely become more accessible, faster, and longer. Developments in software and hardware have given researchers the freedom to perform more demanding simulations, however this also translates into increasing volumes of raw data that need to be stored and analysed.

The resulting raw data from MD simulations can be enormous in size, particularly in the case of macromolecular systems and/or long timescale jobs. This is because the key raw data in MD are nearly exclusively trajectory files. In order for the trajectories to elucidate the dynamical evolution of the system, they include snapshots of the three-dimensional coordinates of each atom in the studied molecular system throughout the duration of the simulation.

The need for dimensionality reduction

Processing MD data in a way that generates accurate and meaningful analysis results with minimal loss of detail is an ongoing challenge. Analytical tools need to be fast and the resulting output files should ideally be of reasonable size. Chemical data include intrinsic information that is often difficult to grasp, and the inherent relationships among this data are particularly interesting. Data analysis methods should therefore be able to generate not only primary but also secondary information from datasets. The secondary information can, for example, be prediction models.[123] At the same time, the need to analyse and visualise large amounts of multivariate data also reveals a fundamental need for dimensionality reduction. The objective of dimensionality reduction is to

manage to map high-dimensional data into compact low-dimensional representations with as many coordinates as observed modes of variance.[124]

Machine learning methods are data analysis methods that can be used in order to elucidate the intrinsic relationships among the data.[123] Unsupervised learning is a type of machine learning algorithm that aims to build a representation of the data.[123] The objective in unsupervised learning is to describe the patterns and relationships among a set of data.[125] Approaches to unsupervised learning include clustering, principal component analysis, and outlier detection.[123]

Principal Component Analysis

Principal Component Analysis (PCA) is an unsupervised learning method that can reduce the dimensionality of the dataset while retaining its information content.[126] It is a method that can extract the systematic variance in a data matrix. What PCA does in mathematical terms is that it transforms a number of correlated variables into a smaller number of uncorrelated variables that are called principal components.[123] Those principal components describe the underlying structure of the dataset.[125]

The largest principal component is the direction that maximises the variance of the projected data. The second largest principal component is the direction where there is the second most variance and the smallest principal component minimises that variance.[125] The first few principal components carry the most relevant information and therefore the rest can be considered to be noise[123] as is illustrated in Figure 12. A dataset is given as 3-dimensional points (i) and then the three PCs are presented (ii), where it becomes clear that PC3 carries the least variance within the dataset. PC3 is therefore ignored in the data transformation where the dataset is projected onto the first two PCs (iii).

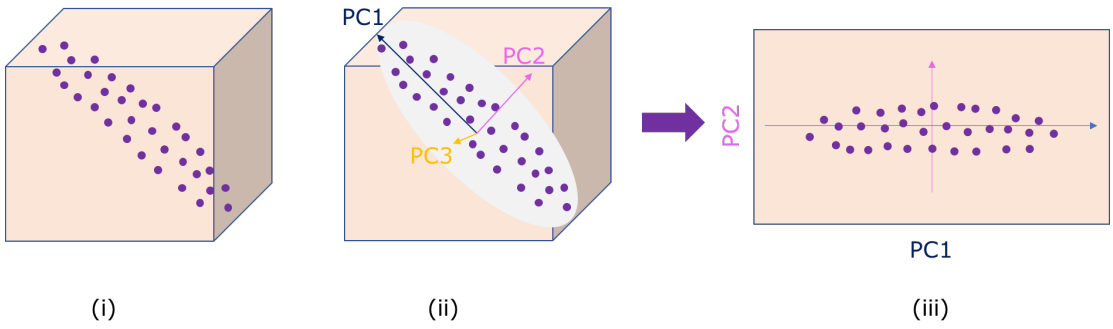


Figure 12. Illustration of of dimensionality reduction with PCA. In (ii) the PCs have been superimposed with the data set. In (iii) the data set is rotated so that the directions of maximum variance coincide with the canonical base. Simultaneously, PC3 is discarded as a direction contributing little or no data variability. Adapted from Kavraki.[127]

For MD data analysed with PCA, the data matrices represent atomic coordinates versus time. Hence, correlated and uncorrelated variables represent correlated and uncorrelated atomic movements which are in turn related to conformational changes. Therefore, in the PCA context the largest PC represents the strongest correlated movement.

Clustering

Cluster analysis is a key task in data mining. Clustering can be defined as segmenting and classifying a set of objects (or molecules) in distinctive groups called clusters. The way the segmentation takes place is by discovering similarities within the members of the set of objects and then dividing them in homogeneous clusters. It is a common and effective technique for statistical data analysis. Many clustering algorithms have been proposed and each differs in what it considers a cluster. A schematic representation of a general clustering application on a dataset can be seen in Figure 13. On the left, the dataset is in its original form, unclustered. The data is then divided in three clusters (right), demonstrated by different colours, in a way that promotes high intra-cluster similarity while inter-cluster similarity is low.

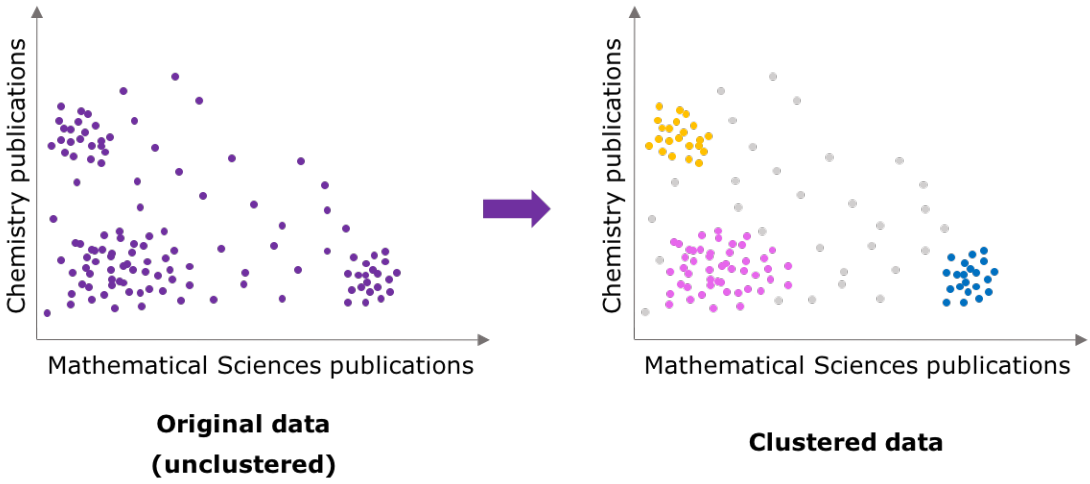


Figure 13. Schematic representation of a general clustering application on a given dataset. The dataset examines the publication activity of a research group and focuses on the number of contributions in Mathematical Sciences journals (x) and Chemistry journals (y) from each researcher. Within the clustered data, the upper-left corner (yellow) may denote a cluster of chemists, whereas the bottom-right corner (blue) may denote a cluster of mathematicians within the research group.

In MD data analysis, clustering algorithms often group molecules together depending on atomic distances or conformational similarities. Clustering provides a refined view of how the molecule samples the conformational space. It can generate visualisations of the relationships between samples or parameters,[126] making clustering a useful tool in structural determination studies. Clustering algorithms commonly used in MD include kinetic clustering and geometric clustering. Kinetic clustering divides the conformational space in microstates which are then grouped together into metastable states according to kinetic proximity.[128] Kinetic proximity denotes the transition probability of one microstate to another within the same group of metastable states, which should be higher than the transition occurring to a microstate outside the group. Geometric clustering refers to grouping structures according to conformational similarity, such as protein conformations.

Summary

An overview of the theoretical background of this work has been presented in this section. The background of fundamental methodologies in computational

chemistry has also been discussed. Finally, a summary of commonly used statistical analysis methods in MD, specifically focusing in data dimensionality reduction, has been presented.

Chapter 3: MD simulations and analysis

MD simulations: Introduction

Methodology in this thesis was heavily focused on employing extensive atomistic MD simulations in different environments using Gromacs.[129] QM calculations using Gaussian 09[130] provided sufficiently optimised starting structures for the MD simulations. MD simulations were performed both on a local cluster and on a worldwide distributed computing grid provided by the World Community Grid (WCG). The local computing efforts were supported by the University of Nottingham (UoN) High Performance Computing (HPC) facility.

The computer power available for this work was therefore quite significant. Consequently, systematic automated approaches were necessary to support the computational workflows. The tasks at hand on a local level were to establish the MA systems, perform the initial set up of the simulations, and utilise a statistical approach to monitor energy convergence after equilibration. On the grid level, the challenge was to handle the step-by-step setup of the production runs and to interact with the established WCG server. Perl programs were written to undertake those tasks.

Rationale

MD simulations are a key tool in obtaining data that can shed light in the effects of intrinsic (MA chemical structure) and extrinsic (solvents and temperature) factors on MA conformations. They can also provide insightful information about the dynamic properties of MAs. Atomistic MD simulations can provide the atom-level information that is required for addressing the MA structure-to-function questions that remain unanswered. The cyclopropane units of MAs are unusual functional groups for which detailed modelling and an explicit consideration of their stereochemistry are necessary.[47] Finally, MD simulations are capable of handling large and flexible molecules, such as MAs.

Computing on the UoN HPC facility (local approach)

The availability of a local facility that offered large computer power and sufficient storage space was key for the groundwork phases of this study. The initial setup stages demanded a significant amount of interactive manual work that would not have been practical to complete on the grid. The established workflows reserved the use of the grid solely for data production, however a significant number of production runs were performed locally as well.

Selection of MA molecules

The MA molecules that were modelled cover all three MA classes that occur naturally in *M. tb*, both the *cis* and *trans* stereochemistries of the proximal cyclopropyl group, and a wide range of chain lengths. The combination of those variations allowed for many natural and synthetic MA molecules to be studied, and introduced a satisfactory level of sampling already from the first steps of the setup. The choice of chain lengths was made according to the variation scheme described in Table 3. The starting point of the selection was the chain lengths of MAs that are naturally occurring, predominantly in mycobacteria from the *M. tb* complex and in other pathogenic mycobacteria.[50]

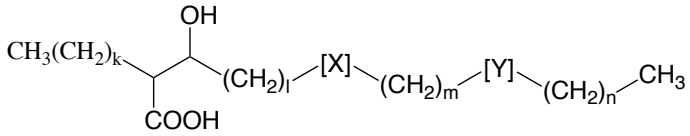
Chain lengths		
		
	l – **m**$^{+}$ – n	
	l – m – **n**$^{+}$	
l^{-} – m – n	**Natural MA**	**l**$^{+}$ – m – n
	l – m – **n**$^{-}$	
	l – **m**$^{-}$ – n	

Table 3. Variation scheme followed to generate the chain lengths of the studied molecules.

The *l*, *m*, and *n* chain length notations represent the lengths of different methylene segments along the MA molecules. According to the chain length variation scheme described in Table 3 above, the length of each MA molecule is chosen by keeping consistent two length notations (e.g. *m* and *n*) each time,

while varying the remaining length notation (*l*, in this example) by two methyl groups, either increasing (l^+) or decreasing (l^-) it. In this manner, a substantially wide variety of chain lengths was generated. The chain lengths that have been chosen for this work are summarised in Figure S1 in Appendix A. The naming convention used when referring to the various MAs modelled, stems from their chemical class and an index number attributed consecutively. For example, **a**lpha MAs with a ***c****is* proximal cyclopropane are abbreviated as AC. Further, molecule AC6 corresponds to the alpha MA with a chain length of 23-11-14-17, as can be seen in Figure S1 in Appendix A. This naming convention is used in an analogous manner for the remaining alpha, methoxy, and keto MAs modelled. The total number of individual MA molecules studied was 166.

Establishment of systems and initial computational setup

The workflow that was developed to establish the MA systems and initiate their computational setup involved a multi-step process and its overall schematic description is given in Figure 14. A combination of the SMILES[131-132] line notation and Discovery Studio[133] was used to build the MA structures. The geometry of the raw initial structures was then optimised at the semi-empirical AM1 level of theory using Gaussian 09.[130]

The backbone carbon chain was edited using Aten[134] in order to incorporate a logical ascending atom-numbering within the molecules. This corrected the random index numbers that were originally assigned to the carbon atoms. The carbon atoms were renumbered starting from the carboxylic acid end and proceeding through the chain towards the meromycolate part. This would allow for seamless and efficient molecule handling and analysis later.

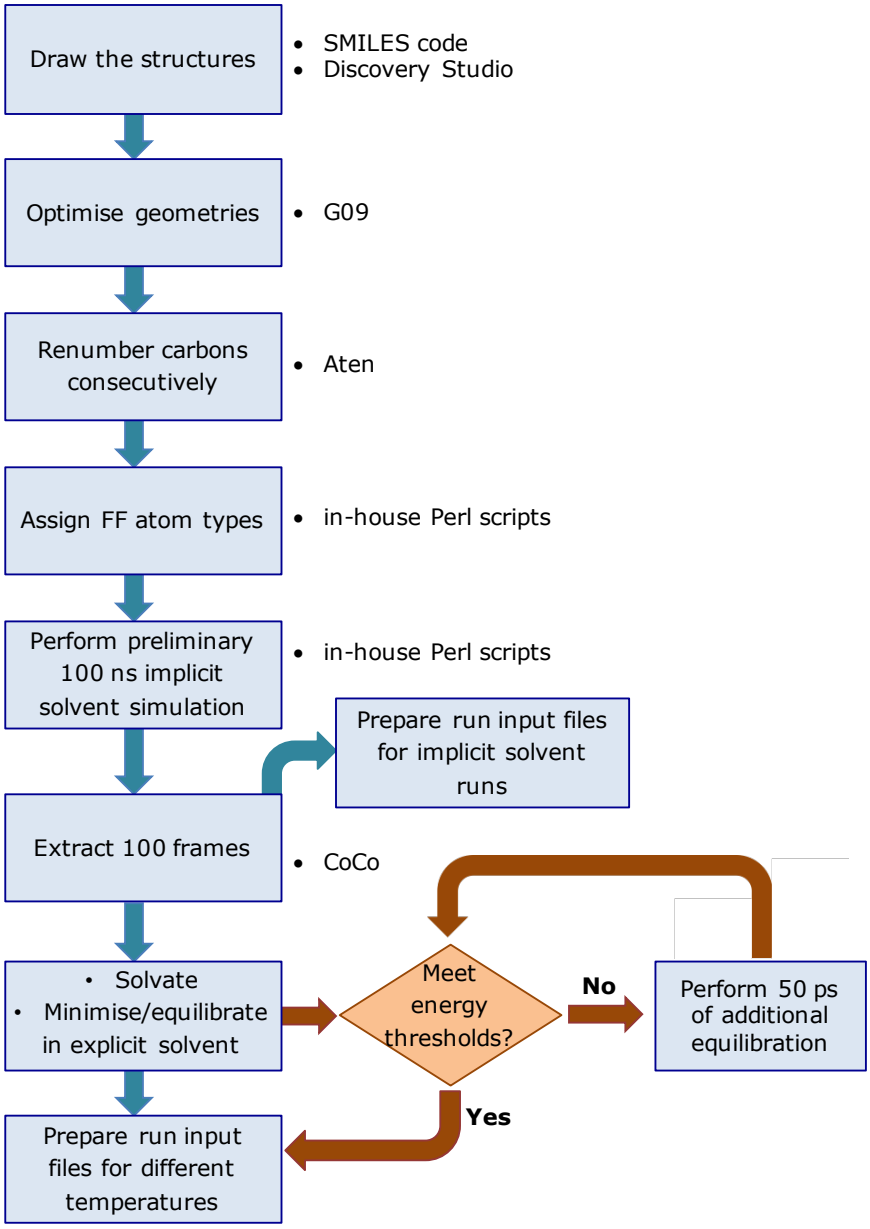


Figure 14. Schematic representation of the workflow followed for the initial setup of the simulations.

OPLS-AA force field

The force field chosen in this work was the **o**ptimised **p**otentials for **l**iquid **s**imulations **a**ll-**a**tom (OPLS-AA[135-136]) force field, which has been optimised for the simulation of liquids. Compared to the OPLS united-atom model, OPLS-AA is more extensively developed and tested and has broader applicability.[135-136] Parameters for bond stretching and angle bending have been implemented from the AMBER all-atom force field and torsional energy parameters have been generated by fitting to *ab initio* rotational energy profiles.[135-136] The non-

bonded parameters have been obtained alongside Monte Carlo statistical mechanics simulations.[135-136]

The OPLS-AA force field has required some re-parameterisation to be able to handle the lipid molecules accurately. Despite it being a commonly used force field for simulating biomolecules, the standard OPLS parameters for hydrocarbons could not accurately reproduce the properties of long alkanes in MD simulations. This was because only shorter alkanes were considered in the original parameterisation.[137] Longer alkanes were included in the re-parameterisation and the torsional parameters for hydrocarbons were refined by fitting to *ab initio* energy profiles. Additionally, the Lennard-Jones potential for methylene hydrogen atoms was adjusted, and partial charges were optimised.[137] Finally, the force field potential was refined. These adjustments gave remarkably improved lipid properties such as better description of the membrane area of a lipid bilayer,[138] more accurate distribution of *trans* and *gauche* states,[137] viscosity, and solvation free energies.[137]

In terms of its applicability for modelling MAs, the OPLS-AA force field lacked some cyclopropane parameters, namely one angle and several dihedral parameters. These were obtained from QM calculations.*

Decision trees used for atom type assignment

A crucial part of the initial computational setup was the assignment of the relevant atom types from the OPLS-AA[135-136] force field to the atoms of the studied molecules in an automated fashion in order to prevent unnecessary manual modification. This approach also offers the opportunity to quickly expand the selection of MA molecules for further studies. A list of the atom types that occur in the MA molecules is shown in Table 4 below. These atom types were used at the assigning step and were also incorporated in the decision-making processes described next.

* Dr Lee-Ping Wang is acknowledged for providing the cyclopropane parameters.

OPLS-AA atom type number	Atom	Molecule
135	C	alkane CH_3
136	C	alkane CH_2
137	C	alkane CH
140	H	alkane H
154	O	mono alcohol
155	H	mono alcohol
158	C	mono alcohol
180	O	dialkyl ether
181	C	C(H_3OR) methyl ether
183	C	C(HOR) ether
185	H	ether
267	C	CO in COOH
268	O	OH in COOH
269	O	CO in COOH
270	H	H in COOH
280	C	ketone
281	O	ketone
711	C	CH_2 in cyclopropane
712	C	CHR in cyclopropane

Table 4. The OPLS-AA atom types that occur in MA molecules.

An in-house Perl program read through the structure file of each molecule and identified the chemical nature of the atom in question. It then initiated a decision-making process to assign the appropriate atom type. This process is described schematically in the decision trees that follow (Figure 15, Figure 16, and Figure 18 for oxygen, hydrogen, and carbon atoms respectively). When assigning the appropriate atom type to an oxygen atom, the process was initially required to answer how many immediate neighbours this atom had. A single neighbouring atom meant that the oxygen in question had a double bond, therefore it had to be connected with a carbon atom (carbon 'alpha').

Thus, the oxygen belonged in a carbonyl group which could be found either in a ketone or a carboxylic acid molecule.

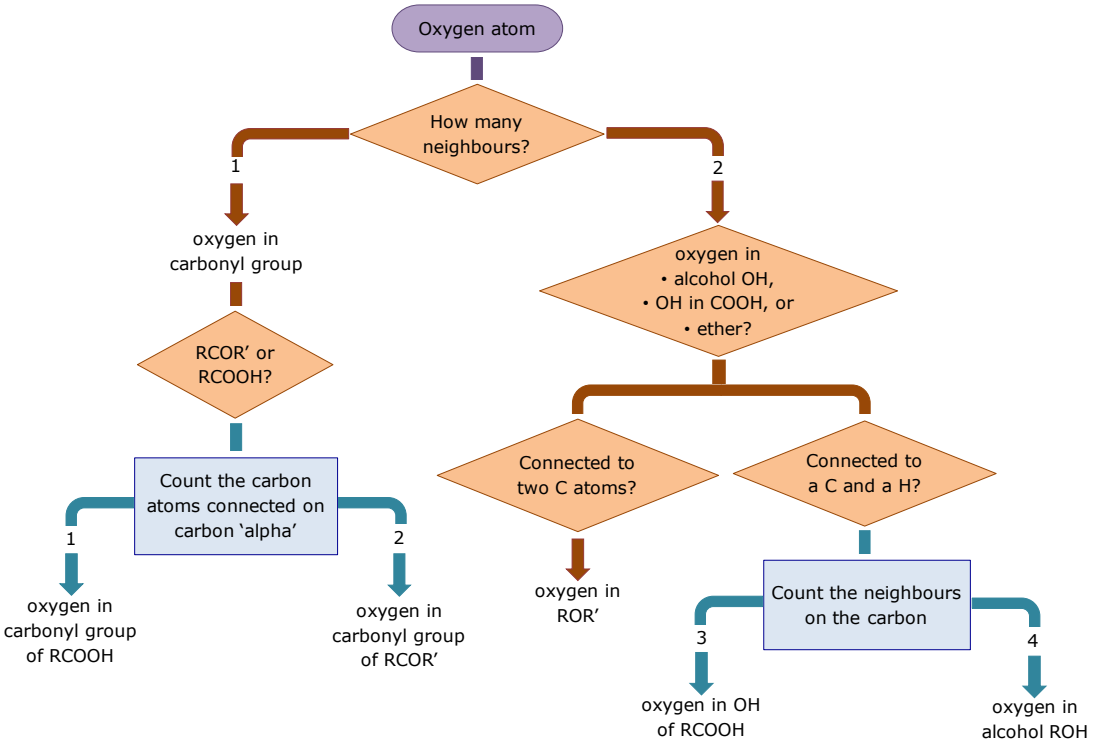


Figure 15. Flowchart of the decision tree followed to assign the relevant atom type to oxygen atoms in the MA molecules.

The program then differentiated between ketone and carboxylic acid by counting the number of carbons that were connected on carbon 'alpha'. If only one connected carbon was counted, then the oxygen in question was in fact found in the carbonyl group of a carboxylic acid. If two connected carbons were counted, then the oxygen belonged to a ketone molecule.

The script proceeded in a similar manner to examine the remaining outcomes of the decision tree. The same process was completed for all oxygen, hydrogen and carbon atoms of the chain and resulted in the creation of the relevant residue database files. Residue database files contain information on bonds, charges, charge groups, and improper dihedrals for the building blocks that comprise the molecules.[116] The residue database is used by Gromacs[129, 139] in order for its tools to be able to process the molecules.

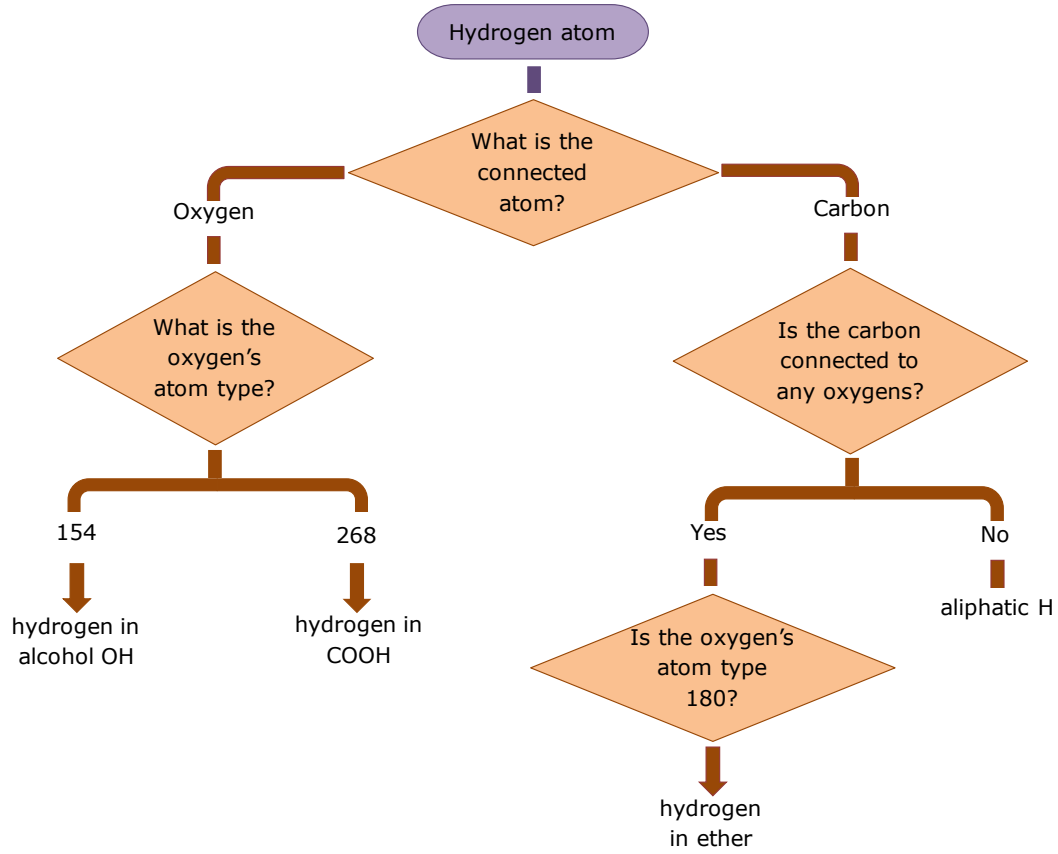


Figure 16. Flowchart of the decision tree followed to assign the relevant atom type to hydrogen atoms in the MA molecules.

While assigning the correct atom types to carbon atoms (Figure 18), the process reached a point where it had to determine if the carbon belonged to a cyclopropane or a regular alkane. This was resolved by measuring the angle between the carbon in question and its neighbouring carbons. The carbon in question was regarded as the central atom and its neighbours were thought to be at the end of two vectors that pointed away from the central atom, as shown in Figure 17. The angle between the two vectors was calculated according to the equation

$$\cos\theta = \frac{\vec{a} \cdot \vec{b}}{\|\vec{a}\| \cdot \|\vec{b}\|} \qquad (3)$$

where the numerator represents the dot product and the denominator represents the product of vector magnitude.

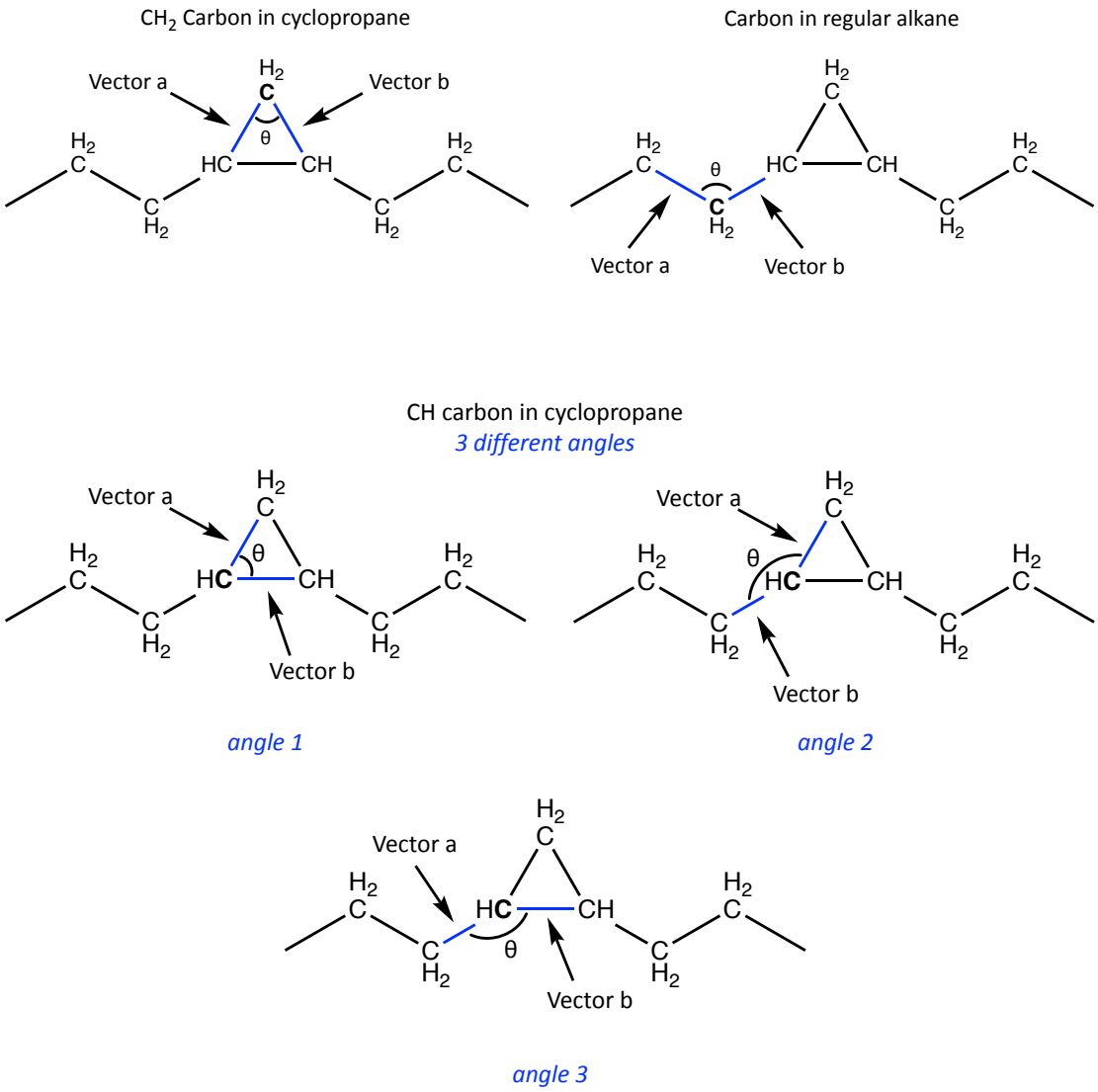


Figure 17. Schematic representation of the angle between the carbon in question and its neighbouring carbon atoms.

In the case of CH cyclopropane carbon, the process needed to examine three different angles (as can be seen in the second half of Figure 17). All three angles were measured and when one of them was found to be smaller than 70 degrees, it was concluded that the atom in question belonged in a cyclopropane. If no small angle was found, then the carbon atom was in fact a regular alkane carbon.

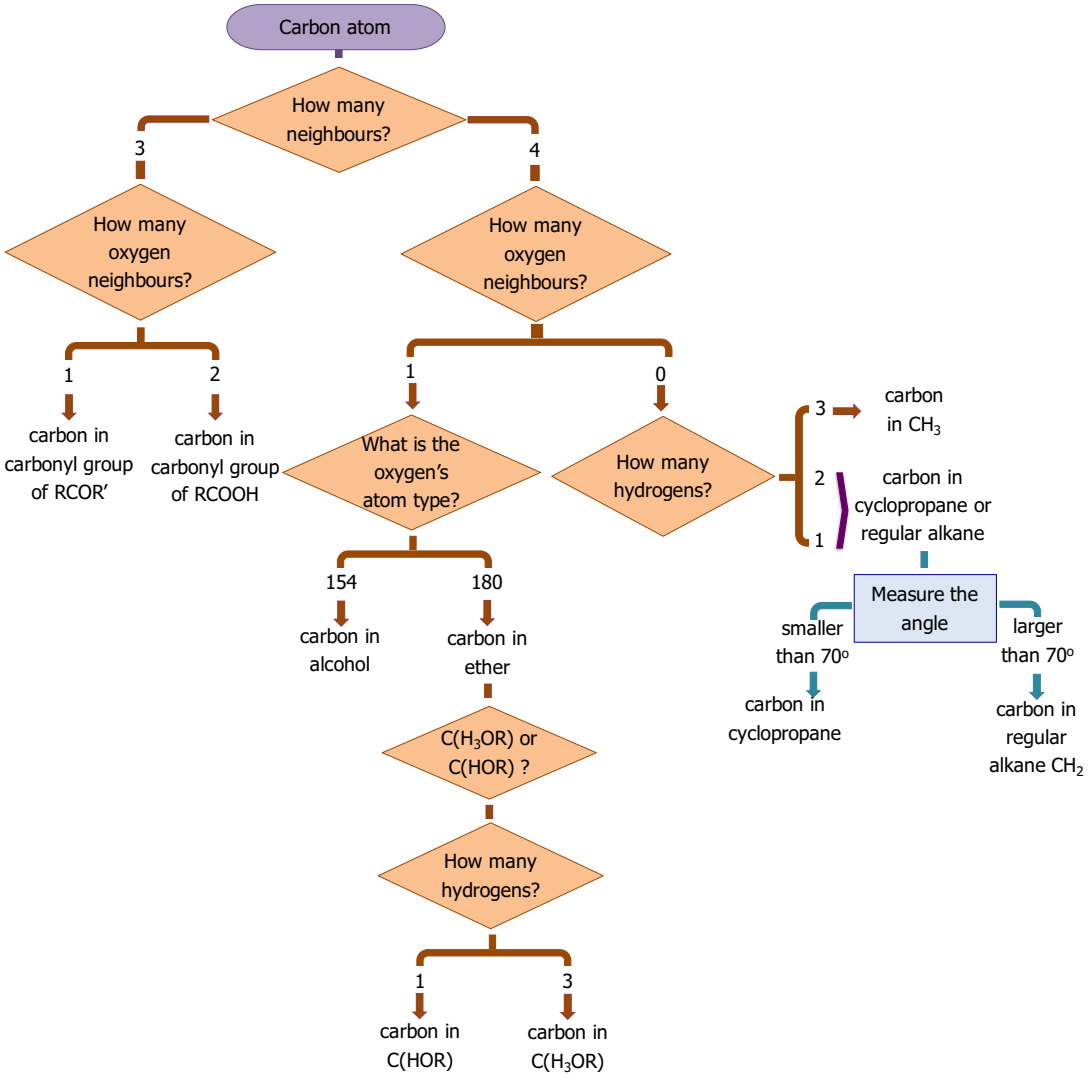


Figure 18. Flowchart of the decision tree followed to assign the relevant atom type to carbon atoms in the MA molecules.

Establishing effective conformational sampling

A common concern in molecular folding and conformational sampling studies utilising MD methods, is molecules getting trapped in local energy minima. This can limit the comprehensive sampling that is fundamental in structural and conformational determination studies. Efficient sampling was crucial in this work as the main objective was to conduct a comprehensive investigation of the spontaneous conformational preferences of single MAs. Comprehensive sampling would enable MAs to explore the conformational space freely and effectively, particularly because they tend to be quite flexible and to assume a

variety of folding patterns. Therefore, it was important to ensure that the starting conformations would be as diverse as possible.

While enhanced sampling methods, such as replica-exchange MD, can certainly contribute to efficient sampling, their application on a MA dataset running on a distributed computing platform was deemed impractical for this work. As the aim of the current study was to employ a multitude of shorter classical simulations doable on the WCG, it was crucial to incorporate the best practice to select diverse starting points early in the setup process.

CoCo,[140] which stands for "Complementary Coordinates", is a PCA-based method that was introduced in the workflow (Figure 14) in order to ensure as effective conformational sampling as possible. The way this method works is by analysing the distribution of an ensemble of structures in the conformational space and then generating a new ensemble that fills any potential gaps in the distribution. It is not guaranteed that the new structures are valid members of the ensemble, but they should rather be considered as potential solutions for refining the original data set and enhancing the conformational sampling.[140] CoCo[140] was used to extract one hundred frames from each resulting trajectory after the completion of a 100 ns simulation performed in implicit water.

Implicit solvation, or continuum solvation, is a mathematical model that uses a continuous medium, rather than explicit solvent molecules, to represent the solvent.[141] This technique essentially averages out the effect of a large number of explicit solvent molecules[141] and it provides an accurate representation of the electrostatic effects of the solvent.[116]

The main advantage of implicit compared to explicit solvation is the absence of viscosity, which allows the molecule to explore the conformational space more quickly.[116] This is particularly relevant to this work as it constitutes one more step towards effective conformational sampling. Other advantages include the significantly smaller computational cost in terms of simulation speed and storage space. On the other hand, the main disadvantage of implicit solvation

is the fact that the absence of explicit solvent molecules takes away the opportunity to derive information about how the molecule interacts with the solvent and how this interaction modulates the folding. The use of both implicit and explicit solvation will provide insights into the specific effects of the solvent on the folding preferences of the molecules. The three solvation systems that this work has focused on are explicit water, implicit water, and implicit hexane. The influence of implicit versus explicit solvation will be evaluated, as well as the effect of the different solvents themselves.

It has been demonstrated[45] in experimental work that the phase diagrams of the Langmuir monolayers of MAs exhibit various phases depending on the temperature and the surface pressure values. To assess the effect of temperature on the folding patterns of MAs elucidated by MD, explicit water simulations have been performed at four different temperatures (300, 325, 350 and 400 K). Simulations performed in implicit solvent (water and hexane) were conducted at 300 K.

The combined effect of the utilisation of these techniques has ensured a rigorous representation of the conformational variability that the MA molecules can demonstrate.

Simulation details

Preliminary implicit water run and frame extraction

Each system was energy minimised in vacuum using a steepest descent algorithm and a maximum of 300,000 steps. Even though equilibration is not typically needed when implicit rather than explicit solvent is involved, in this case a short preparatory run using a 0.5 fs timestep was required. This relaxed the system and prevented the simulations from crashing further on. Each system was then subjected to a preliminary implicit solvent run for 100 ns using a 1 fs timestep. CoCo[140] was used to extract one hundred different frames from the resulting trajectory.

An energy minimisation cycle in vacuum ensured that the extracted structures did not have inappropriate geometry. This was performed using a steepest descent algorithm and a maximum of 200,000 steps. The workflow at this stage reached a fork, as shown in Figure 19. From there, the Perl program proceeded to prepare the run input files for the implicit water and hexane production runs on the grid, and it also moved on to the solvation and equilibration phases that would eventually lead to the explicit solvent production runs.

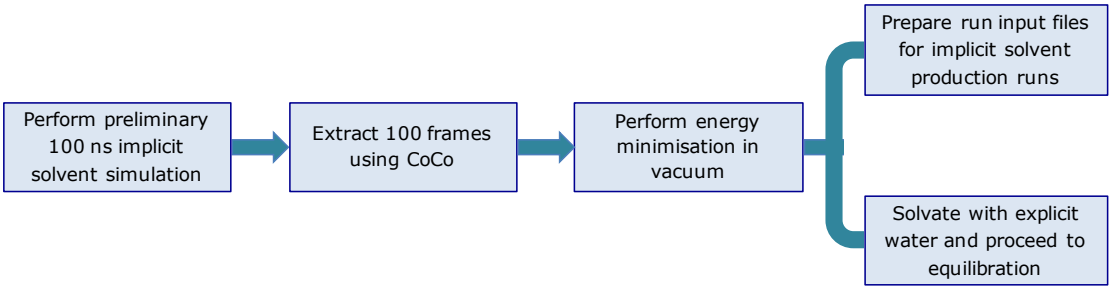


Figure 19. Workflow reaching a fork after the energy minimisation of the extracted frames.

Definition of box and solvation with explicit water

Each system was placed in the centre of a 10x10x10 nm cubic box and at least 1.0 nm from the box edge. This ensured that the minimum image convention was satisfied. The water model used was TIP4P water[142] and the number of water molecules used ranged from 32,923 to 32,957. TIP4P is a four-site model, with four interaction points: two hydrogen atoms, one oxygen atom and a dummy atom with a negative charge near of the oxygen. This model improves the electrostatic distribution around the water molecule. An energy minimisation cycle after the solvation relaxed the structure and ensured that the structure did not have inappropriate geometry. This minimisation was performed using a steepest descent algorithm and a maximum of 200,000 steps.

Equilibration stage and energy convergence checks

The equilibration was conducted in two phases. The first phase was 50 ps under an NVT ensemble (constant number of particles, volume and temperature) using the v-rescale thermostat.[143] Pressure coupling was not applied and the target temperature was 300 K. The second equilibration phase consisted of 50

ps under an NPT ensemble (constant number of particles, pressure and temperature) using the Berendsen barostat[144] for pressure coupling and continuing from the end of the NVT phase. For both equilibration phases, a timestep of 1 fs was used. A leap-frog algorithm was used to integrate Newton's equation of motion. Van der Waals interactions were modelled by using a shift function between 0.8 and 0.9 nm and electrostatic interactions were modelled by using the Particle mesh Ewald (PME)[145] method. Position restraints were applied on all carbon and oxygen atoms.

Before moving on to the production runs it was important to confirm that the total energy and the temperature of the system had reached satisfactory convergence values. If not, then the system was subjected to an additional round of 50 ps NPT equilibration and the energy levels were checked again. This would ensure that the system would not enter the production simulations without having converged first, which minimised the possibility of simulations crashing later on. A statistical approach was used, which involved an AWK program that calculated the standard deviation of the energy values from the last half of the NPT phase. The established thresholds were 0.5% for the total energy and one degree for the temperature. If the convergence thresholds were satisfied, then the system was deemed to have equilibrated sufficiently. The run input files for four different temperatures (300, 325, 350 and 400 K) were subsequently generated.

Production simulations in explicit water

For all explicit water simulations, a timestep of 1 fs was used. Electrostatics were treated as in the equilibration stage. The v-rescale thermostat was used with a time constant of 0.5 ps. The extended-ensemble Parrinello-Rahman barostat was used. The reason for using a different barostat at this stage is in order to yield the correct NPT ensemble, as the Berendsen algorithm yields a simulation with the correct average pressure but not the exact NPT ensemble.[116] No constraints were applied. Velocities were not generated again during the run, which allowed for an exact continuation from where the system was at the

end of the equilibration stage. The chosen scheme for the locally performed data production simulations was single long runs of 100 ns for each molecule.

All simulations were performed with Gromacs[129] 4.6 releases.

Computing on the WCG (grid approach)

The WCG is a worldwide volunteer-driven computer grid which was launched by IBM in November 2004. It consists of a virtual supercomputer that is constructed from the contributions of millions of personal desktop computers provided by volunteers and which is dedicated solely to humanitarian research. Collaborating with the WCG meant having access to enormous computer power, which in turn facilitated thousands of simulations to run in parallel. The total number of individual simulations run on the grid was 103,200.[†] This figure corresponds to the 166 MA molecules, from which one hundred frames were extracted, modelled in a total of six different environments (explicit water under 300, 325, 350 and 400 K temperature, and implicit water and hexane under 300 K temperature).

The discrepancy in the total number of simulations, as explained in the footnote, arose from initial sampling and setup approaches that involved extracting 10,000 frames from each molecule per simulation. Preliminary simulations on a test case proved that this was not the most efficient approach and the methodology of a combination of different temperatures and CoCo[140] was employed instead. Because the test simulations had already completed two equilibration cycles, it was deemed better to keep those 700 frames, despite posing a slight inconsistency, so as not to discard already generated data. The remaining of the dataset followed the 100 frames per molecule scheme.

[†]the discrepancy in the total number of simulations (99,600 is expected but the real total number of simulations is 103,200) comes from the fact that from one of the molecules 700 frames were extracted rather than 100. Therefore the correct number of total simulations is $(165 \cdot 100) \cdot 6 + (700 \cdot 6) = 103{,}200$. The scheme of 100 frames extracted from each of the 166 molecules is maintained for simplicity.

In CPU runtime, the simulations on the grid returned an average of 6 years per calendar day and 3073 years in total, since project launch in March 2016. Both runtimes are true for actual project runtime of sixteen calendar months.

Step-by-step build-up of the simulations

The nature of the WCG distributed computing dictated the need to perform multiple short simulations from individual previously sampled starting configurations. Chunks of short simulation periods ran on average desktop computers and led to a massive set of parallel simulations. In the local approach, the preferred scheme was single long simulations. The difference between the two schemes is addressed in the Discussion section where the efficiency and results of both approaches is assessed. The fact that the computer power was kindly donated by volunteers worldwide meant that simulations on the grid had to complete within reasonably short timeframes of 4 to 8 hours. It would not be practical to distribute work units that would take much longer to finish. The most efficient way to implement this was to develop an automated workflow as shown schematically in Figure 20.

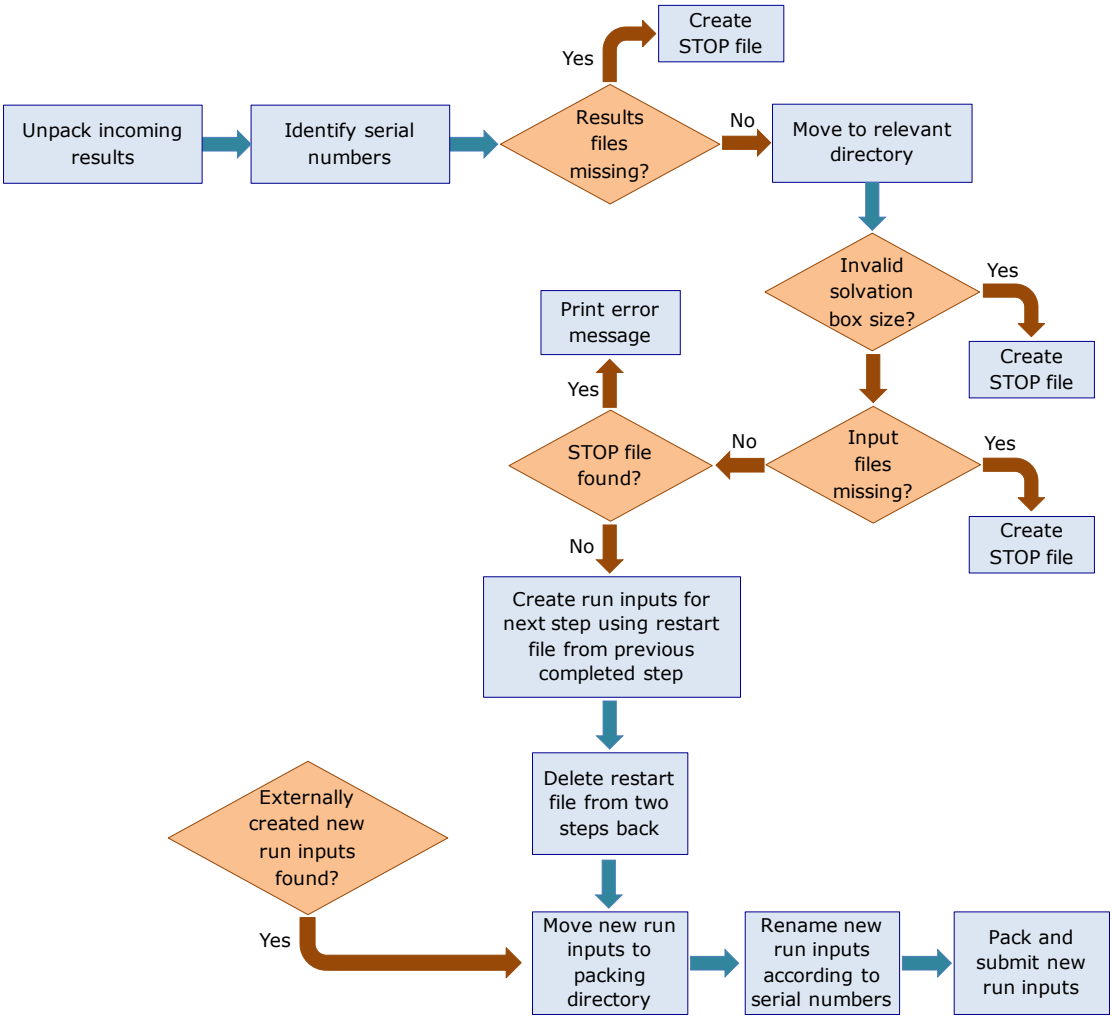


Figure 20. Schematic representation of the file handling protocol developed for the simulations performed on the WCG.

The key characteristic of this workflow was its step-by-step nature, as can be seen in Figure 21. This stepwise build-up meant that the progress of the simulations was in fact solely dependent on the rate that the results were being returned on the server. Each work unit completed a 100 ps (5 ns for implicit solvent) increment of MD before it was returned from the WCG server for handling. The script recognised and utilised the restart file from the last completed simulation step to produce the run input for the next step. To avoid creating a bottleneck and accumulating too many files, the program automatically identified and removed restart files two steps before the last incoming step of the simulation.

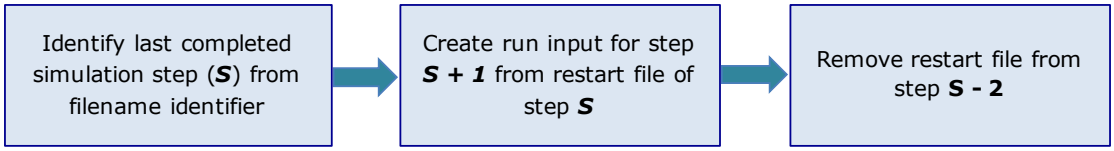


Figure 21. Step-by-step build-up of the simulations.

The main advantages of the stepwise build-up were its robustness and efficiency. The 100 ps increments finished within 5 hours and it was unlikely that a simulation would crash in such a relatively short timeframe. Despite the simulations being relatively short, the 100 ps increments facilitated a reasonably quick collation into longer trajectories with which meaningful analyses could be performed. Additionally, each simulation was run three times on the grid in order to ensure that valid results are retrieved. The obvious shortcoming of this method was its dependence on the incoming files rate.

File handling on the WCG server and file naming convention

The vast number of the simulations made the efficient organisation of the incoming and outgoing files a key aspect of the workflow. Effectively the total number of files handled on the server was much larger than 103,200 as each single simulation generated four result files per simulation step. A graphic representation of the way that the files were distributed on the WCG server is shown in the directory tree of Figure 22. For clarity, only the directory depth for MA1 is shown but the same applies identically for all 166 different MAs.

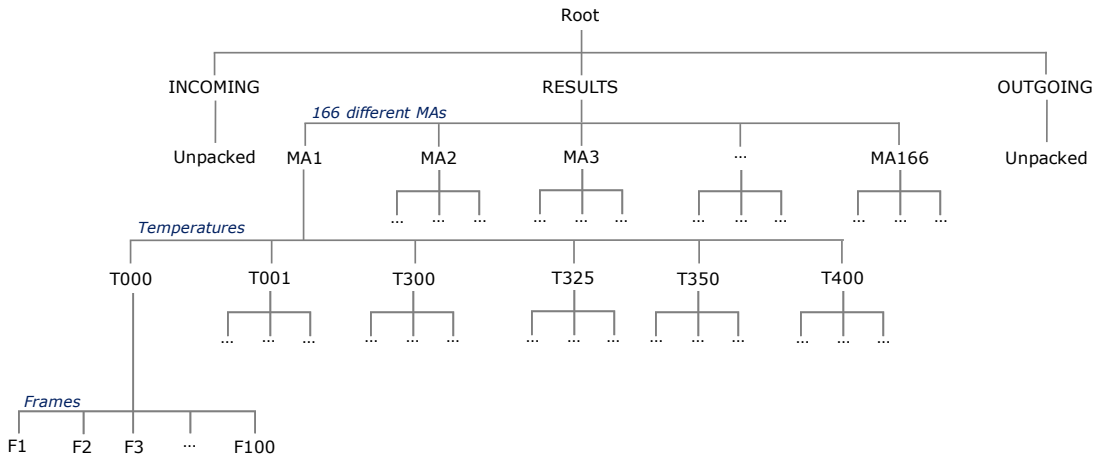


Figure 22. Directory tree explaining how simulation files were organised on the WCG server.

The rate at which results from the WCG were returned amounted to approximately 50 GB of cumulative data per day. Notably, this accumulation is after the trajectory size had been reduced by neglecting water molecules in trajectory files where explicit solvent is applied. Water molecules were only present in the restart files. The amount and frequency of incoming data and, most importantly, the step-by-step build-up nature of the simulations meant that it was crucial to check for new incoming results regularly. The script ran automatically every 1 hour, unpacked the incoming results and distributed them correctly to the corresponding subdirectory according to the MA molecule, temperature, and frame.

The correct destination for each incoming file was determined by interpreting the relevant indicators in their filenames. Each filename included indicators that suggested the identity of the specific MA and the temperature, frame and step of the simulation. The filename also included a project identifier used by the WCG and two more sets of numbers. Those represented the current outgoing batch of work units and the index number of each work unit respectively. The naming convention is shown schematically in Figure 23.

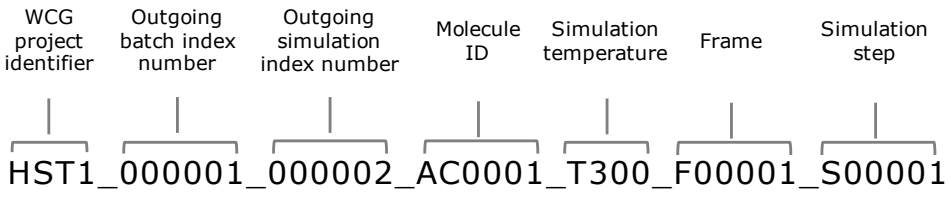


Figure 23. The naming convention used for incoming and outgoing files on the WCG server.

Checks implemented in the workflow

One of the most challenging aspects of the workflow was the comprehensive quality checks that needed to be implemented so that the jobs would run as seamlessly as possible. The script checked for missing result files and also confirmed that the size of the solvation box had no inappropriate increases, since this was an origin of failure detected during test runs on the WCG. Additionally, it monitored the directories for missing input files. If any of those checks came back as true, then a STOP file was created in the relevant directory.

A message informing about the nature of the problem was also printed. This would flag that specific simulation as failed (see also Figure 20).

Subsequently all the directories were scanned for STOP files in order to determine whether the process would continue to the next step or not. If such a file was found, then this clearly pointed to issues with that specific simulation. Those issues had to be corrected manually and the simulation had to be manually restarted. If no issues were found, then the program would proceed to create the run input for the next step of the simulation, using the restart file of the immediately previous completed step as an input. This allowed for an exact continuation of the trajectory.

After the new run input was created, it had to be renamed to reflect the ascending index number of the outgoing work units. The run inputs were then compressed and packaged in batches of 100 files each. This finalised the file handling process and marked the simulations as ready for submission. Finally, a process within the same program monitored the directories for manually created new run inputs that could have been generated externally. If any were found, they were also submitted for renaming and packaging.

Simulation details

Production simulations in explicit water

The computational parameters applied are the same as the ones that were used for the production runs in explicit water in the local approach. Production runs in explicit solvent were performed in stepwise increments of 100 ps each. Velocities were taken from the previous restart file for exact continuation from where the system was at the end of the equilibration stage. Simulations were performed at four different temperatures (300, 325, 350 and 400 K).

Production simulations in implicit water and hexane

Production runs in implicit solvent were performed in stepwise increments of 5 ns each. A timestep of 1 fs was used for all implicit solvent simulation. A leap-

frog algorithm was utilised to integrate Newton's equation of motion. The GBSA model[146] was used to represent the implicit water. No constraints were applied. The dielectric constants used were 80 for implicit water and 1.89 for implicit hexane. Velocities were not generated again during the run which allowed for an exact continuation from where the system was at the end of the equilibration stage. Implicit solvent simulations in both water and hexane were performed at 300 K.

All simulations were performed with Gromacs[129] 4.6.1 release.

Analysis: Introduction

The number of simulations that were performed and the size of the trajectories dictated the need for effective analytical tools. Data-mining techniques, such as clustering and PCA, provide a robust way of understanding the information in a MD trajectory. The clustering algorithms used enabled the classification of the MAs most dominant folding patterns into different clusters. Investigating the relationships between the clusters and observing the frequency of folding events can elucidate the MA conformational behaviour further. The outcome of the analysis was a comprehensive map that provides information about the MAs flexibility, folding, and globularity.

Rationale

The analytical goal of this work was to unpick the folding preferences of different MAs and to associate them with the chemical nature of the MA molecules. This can in turn elucidate the innate structure-function relationships and the link between MA conformation and biological function. It is crucial to understand the effect that the oxygenated functional groups, cyclopropane stereochemistry and meromycolate chain lengths have on the folding of MAs and how these correlate with TB itself. The MD data generated from this work can provide a basis for future simulations of the complete bacterial cell wall.

The different environments in which the MAs were studied can shed light into the effects that the temperature and surrounding solvent can have on the spontaneous folding of the molecules. The two data production schemes employed in this work were taken into consideration in the analysis and are assessed in the Discussion. The comparison between them determined which sampling method returned more meaningful results. The conclusions set the groundwork for the computational setup of future MA simulations.

Analytical approaches used

Clustering techniques

Two clustering techniques were utilised to study the MA trajectories. Dihedral angle clustering using DASH[147] and principal component clustering using the pyPCAzip[148] suite.

Dihedral angle clustering with DASH algorithm

One of the analytical tools used in this work was DASH[147] (Dynamics Analysis by Salt and Hudson), which has been developed to utilise the sequential nature of MD simulation data. The DASH algorithm analyses MD trajectories by focusing on the torsion angles of rotatable bonds. It performs dihedral angle clustering by placing torsion angle ensembles into separate groups, thus distinguishing between individual representative conformations (DASH states).[147] The algorithm ignores high frequency oscillations and eliminates short-lived states. It extracts the major features of each torsion angle and combines them into DASH states that describe the trajectory of the molecule. The sequence of DASH states visited during the MD simulation, along with the length of time the molecule spent in each state, are combined to generate a compressed representation of the trajectory.[147] It is a useful tool for investigating structural characteristics, determining populations of conformational preferences and defining the dynamical relationships that arise from them.[147, 149-151] Additionally, this type of clustering retains the time information of the simulation. This means that the evolution of different conformations can be tracked. Haensele *et al*.[149] performed in-depth conformational dynamics of 11 μs long simulations of arginine-vasopressin (AVP) in aqueous solution. For the analysis, the DASH algorithm was utilised to focus on the different conformations the peptide assumed. Four DASH states were identified, each representing different ring conformations. The main conformations identified were all considered as candidates for biologically active conformations. Victor Paul Raj *et al.*[151] performed a semi-automated cluster analysis of individual Cα torsion angles of DNA polymerase I, in order to describe changes in the backbones of large systems on a residual length scale.

This analysis successfully identified the times at which major conformational changes occurred. Lanig *et al*.[150] performed a 4.1 μs MD simulation of the NR4A1 (hNur77) apo-protein, which is an "orphan" nuclear receptor. "Orphan" nuclear receptors exhibit no obvious ligand-binding pocket in their X-ray crystal structures, which in turn means that many of their biological functions are poorly understood. The DASH algorithm and classical RMSD-based clustering were utilised to observe conformational changes during the simulation. Notably, a novel, previously undetected, binding pocket was identified. Transitions between conformations identified by DASH, corresponded well to those identified by RMSD-clustering.

PCA clustering with pyPCAzip suite

In PCA, a variance–covariance matrix is constructed. The variability of each atom distance, as well as its co-variation with every other distance, is included. A vector is then identified, as a linear combination of the distances, in such a way as to capture the maximum amount of variance in the data. This is the first principal component. The second principal component is orthogonal to the first and maximises the remaining variance in the data.

The combination of PCA and clustering can reveal key conformational changes that take place in MD simulations. pyPCAzip[148] is a Python-based suite for PCA of MD simulation data. Two tools from this suite must be used sequentially to achieve PCA clustering. pyPCAzip itself performs the PCA on the input trajectory and outputs a compressed binary file. The compression can reduce the size of the trajectory down to a fraction of its original size with no significant loss of information. Then, pyPCzclust creates an N-dimensional histogram of the data, representing each snapshot in the trajectory by its position in the PC1, PC2, ..., PCN space. The local maxima in the histogram represent the different clusters and the user can control the number of PCs that are taken into consideration. The pyPCzclust output is a simple text file that consists of the PCs projections and the corresponding cluster ID.

Cluster transition maps

Mapping the cluster transitions in a bidirectional manner was the next step after completing the PCA clustering. The bidirectional monitoring of transitions between clusters offers detailed information in regard to cluster dependencies, provides insights into how short-lived clusters associate with dominant ones, and reveals the relationships between different clusters. It also gives a clear picture regarding the frequency of folding events.

Three PCs were taken into consideration when running the pyPCzclust analyses for this work. After experimenting with several different numbers of PCs it was shown that three PCs returned the most sensible balance of total number of clusters per molecule with limited noise in each case. The cluster transitions were also clearer and more cohesive when three PCs were used. With two PCs it was quite common that too few clusters were identified and the corresponding transitions were largely unilateral.

Intramolecular carbon atom distance matrices

Calculating the intramolecular carbon atom distances as a time series and storing them in matrices was an efficient tool in observing the frequency of folding events. Plotting the matrices provided an interesting visualisation of the MD data. The distance matrix plots can provide a visual depiction of how different the various folding patterns are, how long each molecule spends into a specific folding pattern and, most importantly, how often the folding events occur. They can also serve as a crude clustering technique or a way to confirm the classification that has occurred from the PCA clustering.

Root-mean-square deviation of carbon atom distance matrices

Measuring the root-mean-square (RMSD) deviation of the carbon distance matrices was an effective way to assess the MA flexibility. The correlation of the RMSD with the cluster itself and the cluster transition frequency offer a clear picture about the MA conformational behaviour. The RMSD between each cluster trajectory and the average structure from each cluster provides further information about flexibility. It is worth mentioning that classical

RMSD calculations are less favourable for the objectives of this work. Distance matrices are an internal measure and they can provide a substantial level of detail by neglecting global rotations. The same level of detail is hard to achieve with classic Cartesian RMSD.

Analytical workflow followed

Automation of the analytical workflow was critical, due to the high volume of the raw data. A Perl program was written to undertake this task and a schematic representation of the workflow can be found in Figure 24. The process started with the pyPCAzip/pyPCzclust computation that generated the clustering output. The rest of the workflow is solely focused on editing this output and using external tools to extract and generate useful information regarding the clusters that had been identified.

The program read the clustering output text file and that is when the process reached a fork. On the one hand the script focused on the column that contained the cluster ID index numbers and counted any occurring transitions in a bidirectional manner. Namely, not only the transition from cluster *a* to cluster *b* was counted, but also the one from cluster *b* to cluster *a*. Those transitions were subsequently stored in a matrix.

On the other hand, the clustering output was subdivided into individual files that each contained the PC projections and the cluster ID index for each cluster. Using the trjconv tool from the Gromacs[129] package, the program extracted the trajectory frames that corresponded to each cluster. Those were used to generate individual trajectory files for each cluster. Using the gdist tool from the Gromacs[129] package, the intramolecular carbon atom distances per cluster trajectory were calculated and stored in matrices. Then the RMSD of the matrices was measured.

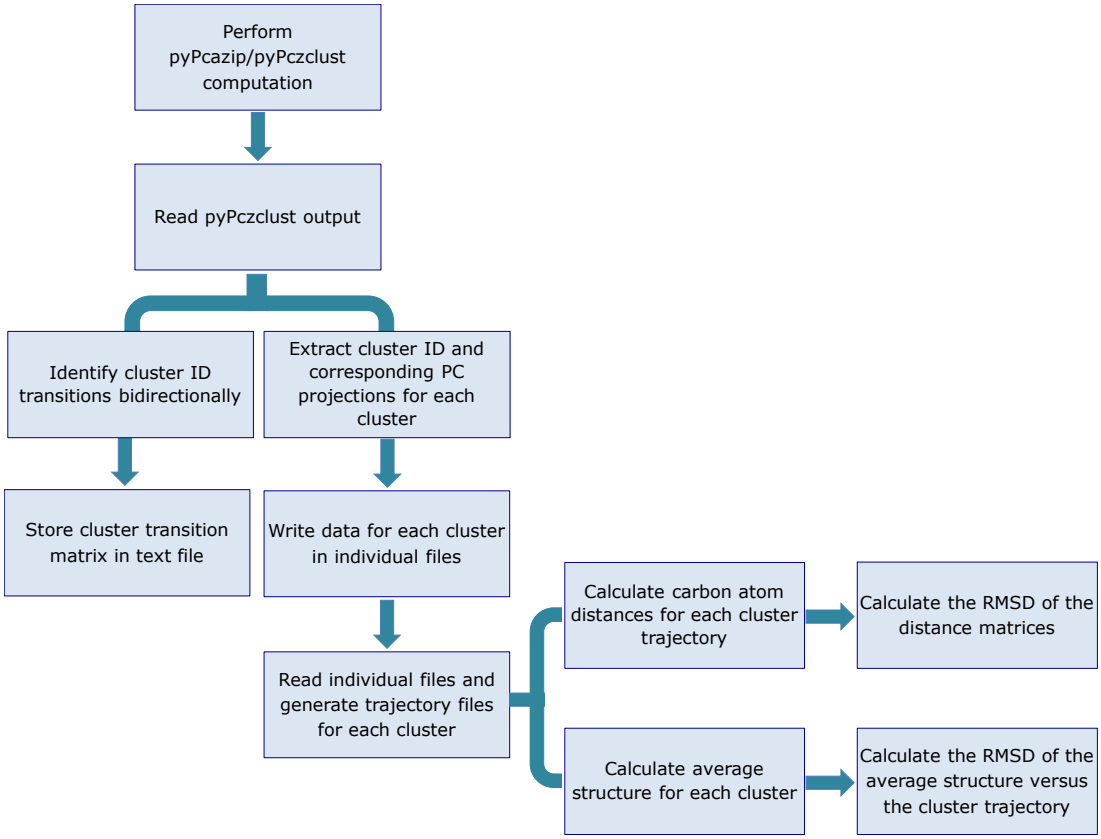


Figure 24. Schematic representation of the analytical workflow followed.

A representative average structure from each cluster trajectory was computed using the g_rmsf tool from the Gromacs[129] package. Finally, the RMSD of the average structure in relation to the cluster trajectory was computed.

Summary

The computing approaches and workflows employed in this work have been presented in detail in this section. The setup and sampling methods implemented have also been outlined, as well as the simulations details. Finally, an overview and the workflow of the analytical approaches used has been presented.

Chapter 4: Results and Discussion

Results from local approach

Single long simulations in explicit water at 300 K

Dihedral clustering

Dihedral clustering, using DASH, was performed on MD simulations of molecules AC6 and KC4. Each of these simulations originated from a total of five different starting conformations (five trajectories, 100 ns each). A collated summary of the DASH output from the respective molecules can be seen in Table 5 and Table S1 in Appendix A. The tables focus on the number of DASH states identified and the transitions between them over the course of the simulation, the snapshots that best represent each DASH state, and the length of time the molecule spent in each state.

	#1	#2	#3	#4	#5
DASH states / transitions	3 / 2	6 / 8	2 / 1	2 / 1	5 / 5
Representative snapshots	77,325 83,459 77,325	48,933 48,758 70,381 16,280 48,761 48,748	50,340 84,992	1 1	6,107 49,241 49,241 49,241 11,333
Timesteps spent in DASH State (DS)	DS 3: 44,676 DS 1: 27,392 DS 2: 27,933	DS 1: 1,132 DS 6: 4,229 DS 1: 875 DS 2: 3,841 DS 5: 19,846 DS 4: 30,121 DS 3: 16,261 DS 4: 6,729 DS 2: 16,967	DS 2: 33,036 DS 1: 66,965	DS 2: 63,125 DS 1: 36,876	DS 1: 766 DS 5: 2,288 DS 1: 2,919 DS 2: 41,947 DS 4: 30,389 DS 3: 21,692

Table 5. DASH analysis results for each of the five trajectories (#1-#5) of molecule AC6. The table offers detailed information on the DASH output for each trajectory. Namely, the first (#1A) trajectory has three DASH states (DS) with two transitions (from DS 3, to DS 1, to DS 2).

The representative snapshots for each DASH states, along with the time the trajectory spent in each state, can also be seen.

The number of DASH states identified for molecule AC6 ranged from two to six, and the number of transitions ranged from one to eight, as can be seen in Table 5. To get a clear understanding of what the dominant conformational patterns were, it was necessary to visualise the representative snapshots for each DASH state. The extracted structures, per trajectory, from molecule AC6 can be seen in Figure 25. The vast majority of the representative snapshots corresponded to highly similar four-chain folding conformations. Only one snapshot differed (trajectory #4) and it belonged to a two-chain folding conformation resembling a symmetrical gamma (γ) shape.

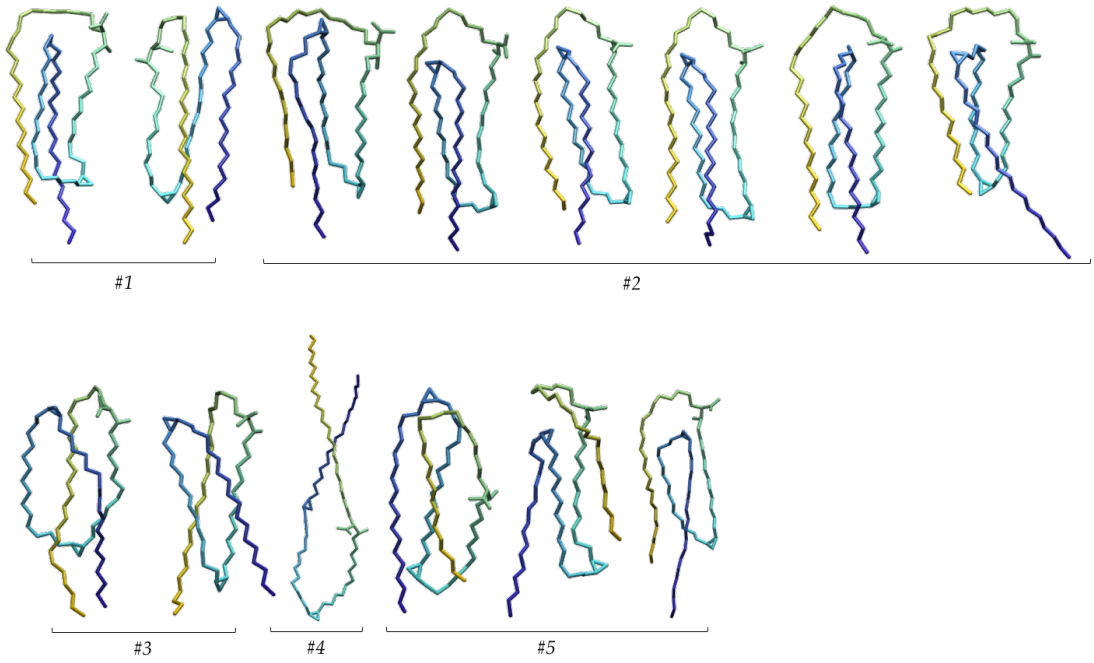


Figure 25. Representative snapshots from identified DASH states in each of the five trajectories (#1-#5) of molecule AC6. The colour coding of the molecular chains follows the consecutive numbering of the carbon atoms along the chain. More specifically, the numbering of the chain starts from the carboxylic acid part (yellow) and proceeds along through the end of the chain (blue). A schematic description explaining the 'start' and 'end' terms in correlation with the colour coding of the molecular chain is provided in Figure 28.

The cyclopropanes functioned as folding points in all the extracted snapshots. In the four-chain conformations, the carbon chain demonstrated clear kinks on

the proximal and distal cyclopropane. In the two-chain fold from trajectory #4, the folding occurred directly on the proximal cyclopropane.

The results for molecule KC4 were substantially similar. The number of DASH states ranged from two to seven and the number of transitions ranged from one to seven, as can be seen in Table S1 in Appendix A. The snapshots that represented each DASH state revealed a clear prevalence of the four-chain folding pattern, similarly to the AC6 molecule. Virtually all representative frames (Figure 26) consisted of nearly identical four-chain folds. Only two structures differed and these belonged to a two-chain (trajectory #1) and a three-chain (trajectory #5) folding pattern, respectively. The four-chain folds exhibited the proximal and distal cyclopropane as folding points. The proximal cyclopropane was a clear folding point in the two-chain conformation. On the other hand, the folding occurred at different positions in the three-chain conformation. There were two clear kinks, on the carbon atom adjacent to the carboxylic acid group and distal carbonyl group respectively.

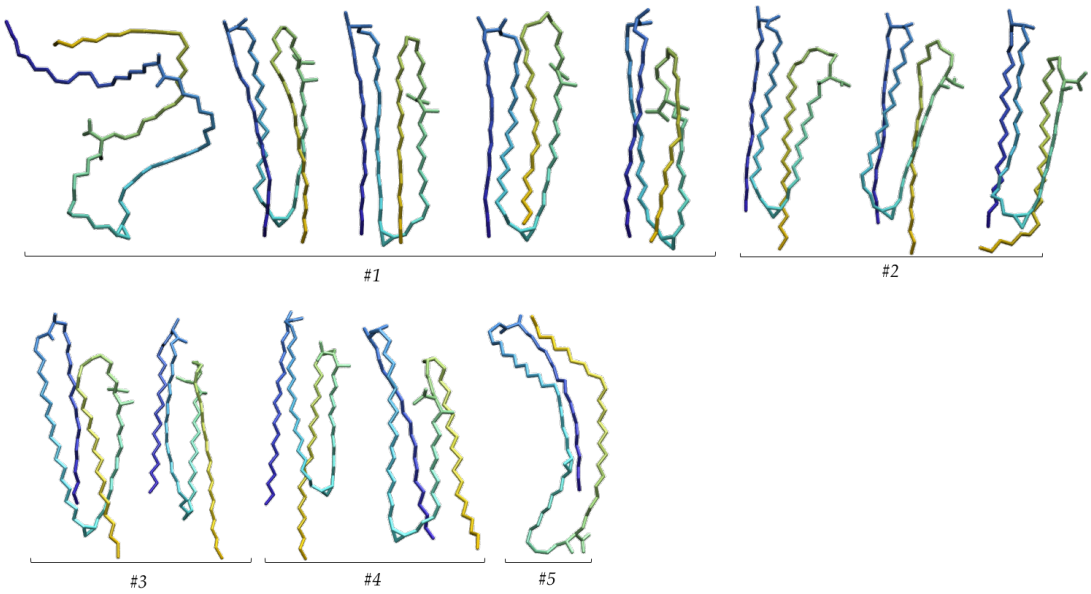


Figure 26. Representative snapshots from identified DASH states in each of the five trajectories (#1-#5) of molecule KC4. The colour coding of the molecular chains follows the consecutive numbering of the carbon atoms along the chain. More specifically, the numbering of the chain starts from the carboxylic acid part (yellow) and proceeds along through the end of the chain (blue). A schematic description explaining the 'start' and 'end' terms in correlation with the colour coding of the molecular chain is provided in Figure 28.

The DASH results from these two sets of simulations of different MAs were highly analogous. They both definitively identified the four-chain folding pattern as the dominant conformational preference of the molecules under the simulation conditions. They also identified the cyclopropanes, both proximal and distal, as the main folding points. However, these results offered negligible structural diversity. Therefore, DASH was deemed unsuitable to achieve the analysis that was needed for this work and it was decided that it would not be used on further datasets, as is elaborated in the Discussion section. At the same time, the PCA clustering approach employed reveals all the important different folding structures that the DASH analysis failed to identify.

PCA clustering and distance matrices

PCA clustering was performed on a comprehensive set of AMA and KMA molecules, simulated during the local computational approach. The same analysis was performed on a comprehensive set of AMA molecules and on test cases of KMA molecules simulated during the distributed computing approach, and these results are presented in the relevant section later in this chapter. The comparison of the outcomes between the two approaches will provide useful insights into the effect of sampling and computational setup on folding conformations.

The folding preference percentages, as they occur from the PCA clustering analysis, are shown in Table 6 and Table 7 for KMA and AMA molecules respectively. These results stem from analyses performed individually on each single trajectory from the five starting conformations (five trajectories, 100 ns each) within each molecule. Intramolecular distance matrices of all the carbon atoms in the molecular chain, and for every step of the trajectory, were also generated from the same datasets. The average matrix for each cluster trajectory was subsequently calculated. The RMSD matrix of each cluster trajectory was also generated to elucidate fluctuations in distances. For each different cluster, the resulting average distance matrix was substantially diverse. Therefore, one can discriminate between different clusters directly by looking at the average distance matrices, and it is not necessary to look up every

structure. A comprehensive collection of diverse average distance matrices is provided for reference in Figure 27.

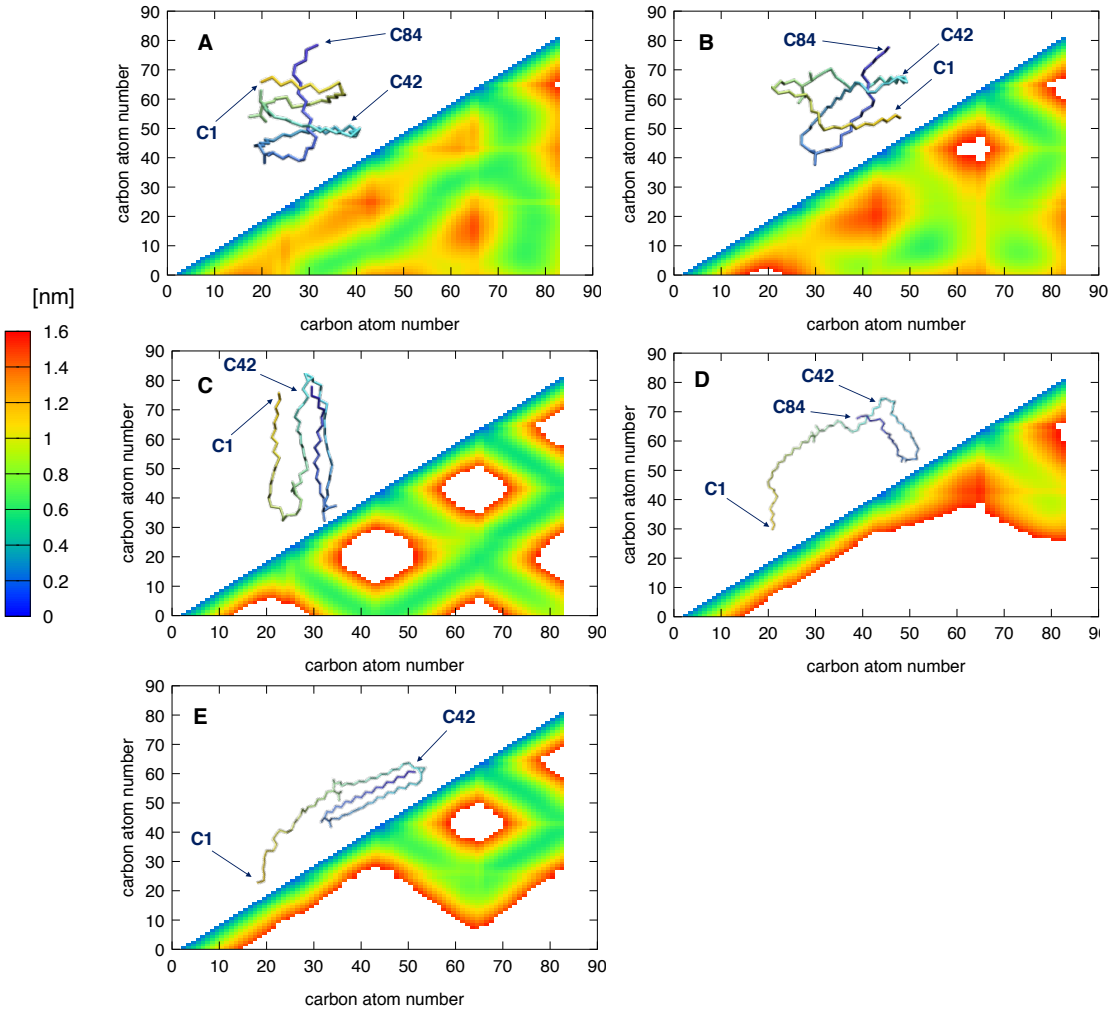


Figure 27. Representation of selected diverse average distance matrices of the five (A-E) structural clusters resulting from the trajectory of starting conformation #1, of molecule KC20, in explicit water (300 K). The numerical values on the axes represent the index number of each carbon atom in the MA chain. For clarity, selected carbon atoms (C1, C42, and C84) along the molecular chain are indicated with arrows. The intramolecular distances between all carbon atoms along the chain are given in nm and the colour scheme in the plots ranges from blue (0 nm) to red (1.6 nm), as can be seen in the legend on the left.

The PCA clustering results of single MA trajectories revealed that the overall most prevalent conformational preferences for both chemical classes of MAs corresponded to the four-chain, knot, and five-chain folding. The overall conformational presentation for the KMAs consisted of 57% four-chain folding, 28% knot folding, and 13% five-chain folding. The remaining 2% consisted of

three- or two-chain folding. With the exception of molecule KC20, in all other occurrences within the KMA class the percentage of four-chain conformations decreased as the *l* chain elongated (Table 6). Simultaneously, the percentage of knot folding and five-chain folding increased as the *l* chain elongated.

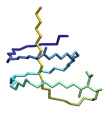
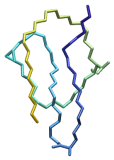
molecule	chain length (*k-l-m-n*)	% four-chain 	% knot 	% five-chain 
KC4	23-11-18-17	72	25	0
KC12	23-13-18-17	60	26	14
KC20	**23-15-18-17**	**65**	**16**	**13**
KC27	23-17-18-17	49	32	19
KC29	23-19-18-17	40	40	19

Table 6. Folding preference percentages, resulting from single MD trajectory analysis of each of the five trajectories per KC molecule. The columns, from left to right, represent the identifier of the KC molecule, the chain length, and the percentages of four-chain, knot, and five-chain folding. Representative structures of each cluster are also provided for clarity. The colour coding of the molecular chains follows the consecutive numbering of the carbon atoms along the chain, starting from the carboxylic acid part (yellow) and proceeding towards the end of the chain (blue). A schematic description of the colour coding is provided in Figure 28.

AMAs demonstrated a different conformational profile. The knot structures dominated the folding profile of single MD trajectories, by occupying 48% of the total pattern. Four-chain folding was at 45% and the remaining 6% corresponded to five-chain folding. As the *l* chain elongated (Table 7), the percentage of four and five-chain folding decreased and that of knot folding increased across the studied molecules. The only exception in this pattern was molecule AC33 which bore the longest *l* chain. It demonstrated an overall 4% prevalence of the four-chain folding over the knot structures while maintaining an 18% five-chain presence, which was the highest among the molecules examined in this chemical class.

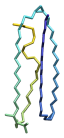
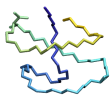
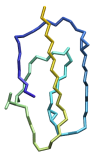
molecule	chain length (*k-l-m-n*)	% four-chain 	% knot 	% five-chain 
AC6	23-11-14-17	55	39	6
AC15	23-13-14-17	42	56	1
AC24	23-15-14-17	38	60	0
AC33	**23-17-14-17**	**43**	**39**	**18**

Table 7. Folding preference percentages, resulting from single MD trajectory analysis of each of the five trajectories per AC molecule. The columns, from left to right, represent the identifier of the AC molecule, the chain length, and the percentages of four-chain, knot, and five-chain folding. Representative structures of each cluster are also provided for clarity. The colour coding of the molecular chains follows the consecutive numbering of the carbon atoms along the chain, starting from the carboxylic acid part (yellow) and proceeding towards the end of the chain (blue). A schematic description of the colour coding is provided in Figure 28.

After completing the PC clustering analysis individually for each of the five trajectories of each molecule, further details were sought by performing the same type of analysis on the concatenated trajectories of each molecule. This provided a different point of view on the distribution of the principal components and highlighted the results of the individual analyses. The conformational profile that AMAs generated, as seen in Table 8, was different to what was shown previously in Table 7 from the individual analyses. Namely, the folding preferences do not seem to follow a pattern that corresponds to the individual methylene chain lengths. This may be due to either the change in the sampling method, or the efficiency of the clustering tool used.

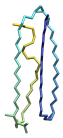
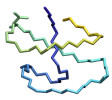
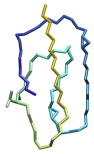
molecule	chain length (*k-l-m-n*)	% four-chain 	% knot 	% five-chain 	% three-chain 
AC6	23-11-14-17	58	42	0	0
AC15	23-13-14-17	43	0	57	0
AC24	23-15-14-17	91	5	0	4
AC33	23-17-14-17	39	61	0	0

Table 8. Folding preference percentages, resulting from concatenated trajectories analysis of AC molecules. The columns, from left to right, represent the identifier of the AC molecule, the chain length, and the percentages of four-chain, knot, five-chain, and three-chain folding. Representative structures of each cluster are also provided for clarity. The colour coding of the chains follows the consecutive numbering of the carbon atoms, starting from the carboxylic acid part (yellow) and proceeding towards the end of the chain (blue). A schematic description of the colour coding is provided in Figure 28.

Similarly, the results from KC molecules can be seen in Table 9. In this case, there is a trend that can be unpicked. As the *l* chain elongated, the knot percentage increased, something that changed at *l*=19, where the knot percentage decreased and the four-chain percentage increased.

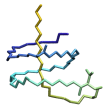
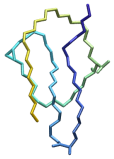
molecule	chain length (*k-l-m-n*)	% four-chain 	% knot 	% five-chain 
KC4	23-11-18-17	91	9	0
KC12	23-13-18-17	93	7	0
KC20	23-15-18-17	33	66	0
KC27	23-17-18-17	4	96	0
KC29	23-19-18-17	58	29	13

Table 9. Folding preference percentages, resulting from concatenated trajectories analysis of KC molecules. The columns, from left to right, represent the identifier of the KC molecule, the chain length, and the percentages of four-chain, knot, and five-chain folding. Representative structures of each cluster are also provided for clarity. The colour coding of the chains follows the consecutive numbering of the carbon atoms, starting from the carboxylic acid part

(yellow) and proceeding towards the end of the chain (blue). A schematic description of the colour coding is provided in Figure 28.

Visual inspection of the average distance matrix plots, from clusters across the five single trajectories within each molecule, and in comparison to the clusters resulting from the concatenated trajectories analysis, revealed clear similarities. The distributions were further explored by calculating the RMSD between the visually most similar average distance matrices. The resulting low RMSD values echoed the similarity between different clusters.

Comprehensive results from the principal component clustering are presented in tables for each molecule. Clusters are identified per-trajectory (#1-#5), for the individual analyses, and are named using a combination of the corresponding trajectory number and a letter notation (A-G). For example, for molecule AC6, cluster A resulting from trajectory 1 is abbreviated as cluster #1A. It is worth noting that the naming scheme used in the clusters resulting from the concatenated trajectories omits the initial index number that was used in the per-trajectory analyses. In the concatenated trajectories, clusters are identified as A, B, C, and so on, with the added clarification that they have been generated from the concatenated trajectories.

Each PCA clustering table contains comprehensive information regarding the identified clusters. The average structure of each cluster is included, and a more specific description of its conformation is also provided in order to add more detail regarding the identified subclass. As already discussed, the main identified conformations were four-chain folds, knots, and five-chain folds, however the description of four-chain folds and knots can carry further detail. The four-chain fold can be described in further detail by mentioning its variant, namely a tidy four-chain fold, where the four chains of the molecule fold in a highly structured way, while being stretched straight. The knot fold can be described in further detail by mentioning its variants, namely (i) a relaxed knot, (ii) a bow, where the start and the end of the carbon chain emerge through the loop, (iii) a knot - end, where the end of the carbon chain emerges through the

loop, and (iv) a knot - start, where the start of the chain emerges. For clarity, a schematic description explaining the 'start' and 'end' terms in correlation with the molecular chain is provided in Figure 28 below.

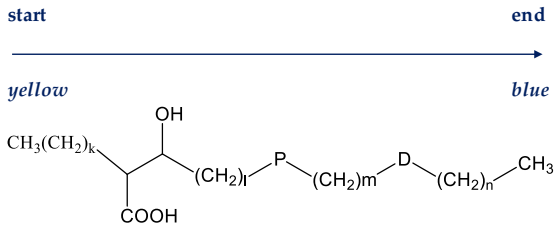


Figure 28. Schematic representation of the index numbering followed across the molecular chain. The numbering started from the carboxylic acid part and proceeded along the chain. The colour scheme used in the rendered structures was from yellow to blue.

The presence of each cluster in the overall trajectory is given as a percentage. Three columns with RMSD-related data per cluster are incorporated: (i) the RMSD value of the cluster average structure versus the cluster trajectory, (ii) the average RMSD value of the distance matrices, and (iii) the two lowest RMSD values calculated from the comparison between the visually most similar average distance matrices. This provides information about which were the most similar clusters to the examined one.

The results from the PCA clustering for each AMA simulated during the local computing approach are given in Table S2, Table S3, Table S4, Table S5 in Appendix A. Similarly, PCA clustering tables for each KMA simulated in the local approach are given in Table S6, Table S7, Table S8, Table S9, Table S10 in Appendix A. The results from PCA analyses performed on concatenated trajectories of AMAs are summarised in Table S11, Table S12, Table S13, and Table S14, as well as for the locally simulated KMAs in Table S15, Table S16, Table S17, Table S18, and Table S19 in Appendix A.

Alpha mycolic acids

In terms of similarities within the various clusters identified in molecule AC6, cluster #1A was shown (Figure 29) to be most similar with cluster B from the concatenated trajectories, and cluster #4A respectively. This underlined the

correlation of visually dissimilar structures, as it demonstrated the strong association between a knot (#1A) and a four-chain fold (#4A).

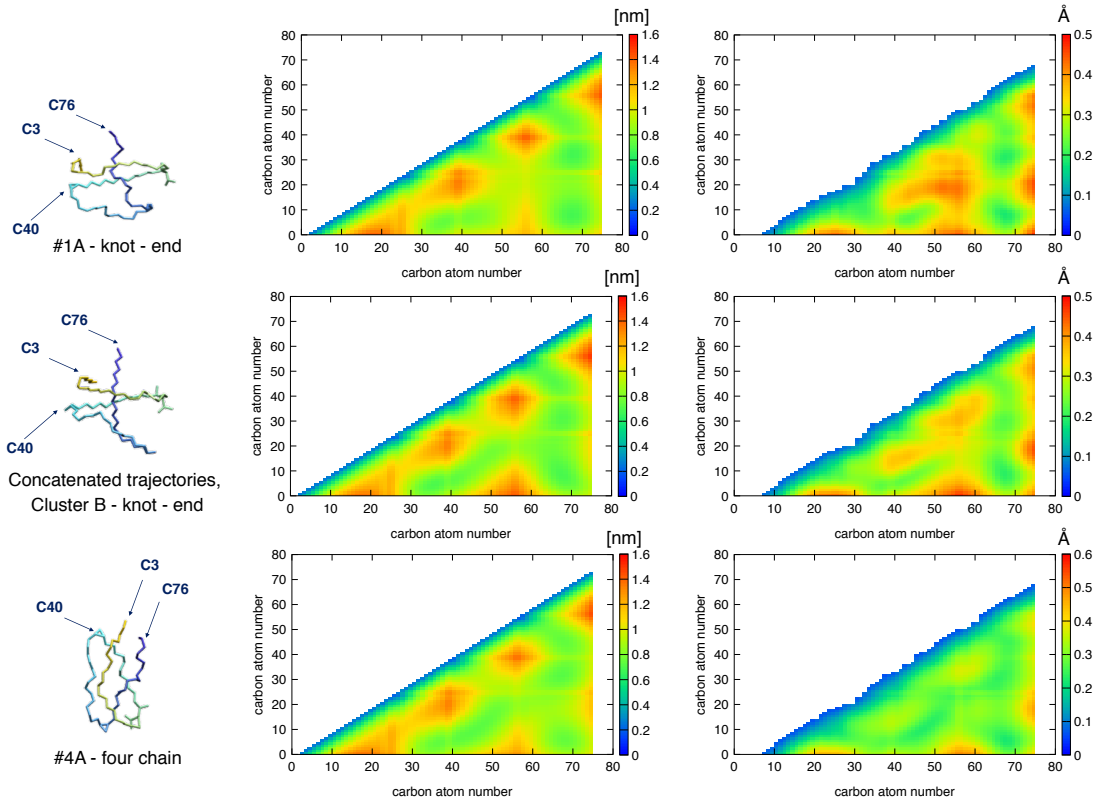


Figure 29. Average structure (left), average distance matrix (middle), and RMSD matrix (right) of clusters similar with #1A from molecule AC6 in explicit water (300 K). For clarity, selected carbon atoms along the molecular chain are indicated with arrows on the structures. The numerical values on the x and y axes (both types of plots) represent the index number of each carbon atom in the MA chain. The intramolecular distances between all carbon atoms along the chain are given in nm, and the colour scheme ranges from blue (0 nm) to red (1.6 nm) as can be seen in the legend on the right of each average distance plot. The fluctuations of the intramolecular distances are provided in Å, and the colour scheme ranges from blue (0 Å) to red (0.6 Å) as can be seen in the legend on the right of each RMSD plot.

The cluster transitions within trajectory #1, as shown in Table 10 below, echoed the findings shown in Figure 29 above, as they exhibited many transitions between knot (#1A) and four-chain folds (#1B, #1E). Namely, frequent bidirectional transitions were identified between clusters #1A and #1B (1566 and 1570 respectively) and between #1A and #1E (379 and 369 respectively).

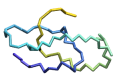
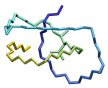
#1 AC6	A - knot 	B - four chain 	C - knot 	D - knot 	E - four chain 
A	0	1566	642	515	379
B	1570	40	46	104	521
C	643	40	0	129	14
D	520	106	122	0	0
E	369	529	16	0	0

Table 10. Number of cluster transitions occurring within trajectory #1 of molecule AC6. The first column represents the cluster identifier (A-E). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 1570 transitions from cluster #1B to cluster #1A, and 1566 transitions from #1A to cluster #1B.

Data from molecule AC15 reiterated the correlation between knot and four-chain folds. Specifically, cluster transitions in Table 11 and Table S20 indicated that the most frequent interactions, within trajectories #2 and #4 respectively, occurred between those two visually dissimilar structures. Namely, from Table 11, cluster #2A (four-chain) interacted most with clusters #2B and #2C, (knot folds). Many bidirectional transitions were identified (736 and 758 respectively between #2A and #2C, and 718 and 706 between #2A and #2B).

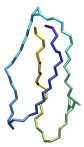
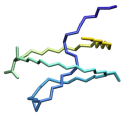
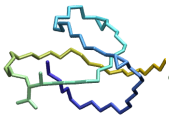
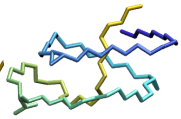
#2 AC15	A - four chain 	B - knot 	C - knot 	D - knot 	E - knot 
A	0	718	736	291	510
B	706	0	293	440	1
C	758	276	0	122	32
D	275	445	134	0	263
E	516	0	26	264	0

Table 11. Number of cluster transitions occurring within trajectory #2 of molecule AC15. The first column represents the cluster identifier (A-E). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 706 transitions from cluster #2B to cluster #2A, and 718 transitions from #2A to cluster #2B.

Finally, as can be seen in Table S20, there were 1281 and 1274 bidirectional transitions between clusters #4A (knot) and #4B (four-chain) respectively.

Molecule AC24 demonstrated a similar behaviour. Clusters #3A and #3B, that represent a four-chain and a knot fold respectively, interacted frequently. As can be seen in Table S21, there were 1864 and 1865 transitions between #3A and #3B respectively. Similarly, the distance matrix comparison (Figure S2) echoed the above.

Visual inspection of the average distance matrices, and subsequent RMSD calculation, for molecule AC15, revealed a correlation between knots and five-chain folds. As can be seen in Figure S3, cluster #4A (knot) was most similar with cluster A (five-chain) from the concatenated trajectories, and cluster #5B. A correlation between knot and five-chain folds was also exhibited in molecule AC33. The results in Figure 30 identified a link between cluster #2A (five-chain) and cluster A (knot) from the concatenated trajectories, and cluster #1B (knot), respectively.

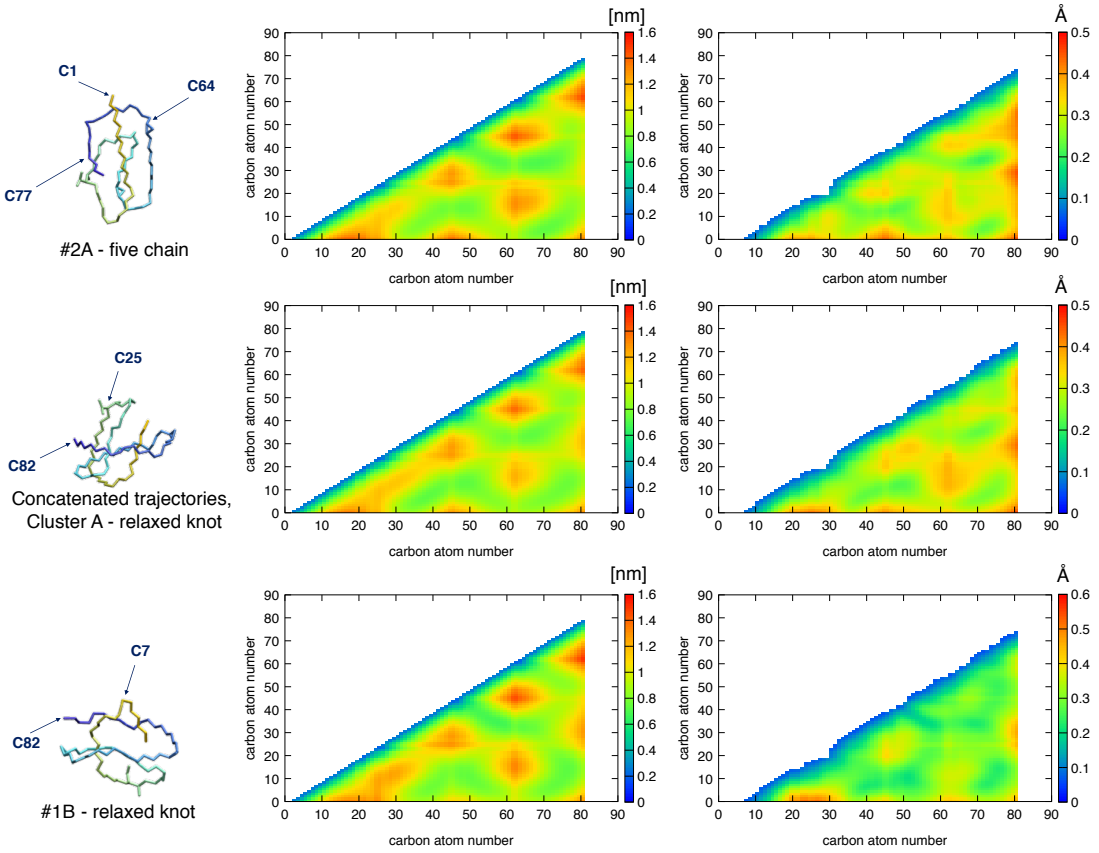


Figure 30. Average structure (left), average distance matrix (middle), and RMSD matrix (right) of clusters similar with #2A from molecule AC33 in explicit water (300 K). For clarity, selected carbon atoms along the molecular chain are indicated with arrows on the structures. The numerical values on the x and y axes (both types of plots) represent the index number of each carbon atom in the MA chain. The intramolecular distances between all carbon atoms along the chain are given in nm, and the colour scheme ranges from blue (0 nm) to red (1.6 nm) as can be seen in the legend on the right of each average distance plot. The fluctuations of the intramolecular distances are provided in Å, and the colour scheme ranges from blue (0 Å) to red (0.6 Å) as can be seen in the legend on the right of each RMSD plot.

Scarcely populated clusters, such as two-chain folds, were rare but nevertheless occurred in molecules AC15 and AC24. As can be seen in Table S12 in Appendix A, there is a single occurrence of a two-chain fold (cluster #5D) that occupied only a marginal percentage (0.005%) of the total folding pattern in molecule AC15. This structure was in fact very short-lived, limited only to the first few steps of the simulation when the molecule had not assumed its fully-folded state yet. The cluster transitions (Table 12) confirmed that this scarcely populated cluster had negligible interactions with the rest of the trajectory,

namely 10 and 11 bidirectional transitions between clusters #5B and #5D respectively.

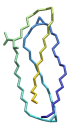
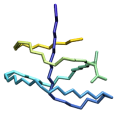
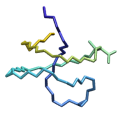
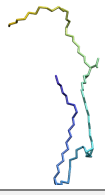
#5 AC15	A - four chain 	B - knot 	C - knot 	D - two chain 
A	0	1637	5	0
B	1636	0	1027	10
C	6	1026	0	0
D	0	11	0	0

Table 12. Number of cluster transitions occurring within trajectory #5 of molecule AC15. The first column represents the cluster identifier (A-D). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 1636 transitions from cluster #5B to cluster #5A, and 1637 transitions from #5A to cluster #5B.

A similar picture was observed in molecule AC24. As seen in Table 13, cluster #4F (two-chain), representing only a minor percentage (0.8%) of the total trajectory, had negligible interactions with all clusters within this trajectory.

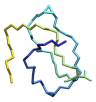
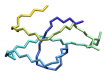
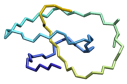
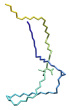
#4 AC24	A - knot 	B - knot 	C - four chain 	D - knot 	E - knot 	F - two chain 
A	0	950	136	304	743	13
B	948	0	602	557	1	46
C	143	591	0	145	163	17
D	301	559	143	0	0	2
E	742	6	159	0	0	0
F	11	48	19	0	0	0

Table 13. Number of cluster transitions occurring within trajectory #4 of molecule AC24. The first column represents the cluster identifier (A-F). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 948 transitions from cluster #4B to cluster #4A, and 950 transitions from #4A to cluster #4B.

Additionally, the above table showed that clusters #4D and #4E had no transitions between them. This was an interesting observation as both clusters represented the knot fold, albeit with a slight structural difference. This entailed that the opposite end (start/end) of the carbon chain extended through the loop each time. Another piece of information that can be extracted from the transitions above is that clusters #4B and #4E had, somewhat paradoxically, negligible bidirectional transitions, despite representing an identical knot fold. This could be explained by the rather fragmented clustering profile that trajectory #4 demonstrated, with numerous clusters being generated. 67% of those clusters corresponded to the knot fold and it is likely that this increased fragmentation into highly similar clusters has obscured the transition patterns.

Visualising PCA results further, by plotting the first two principal components versus the timestep of the simulation in the trajectory, can reveal details about the folding pathway versus time. Two diagrams are given for molecule AC33, as they can be seen in Figure 31; the results from a single MD trajectory on the left, and from the relevant concatenated trajectories on the right. Each square point in the plot represents a successive timestep. The colour scheme used in the square points is derived from the idea of a heat map. The early timesteps in the trajectory are represented by dark blue, the middle is denoted with green, and the final timesteps are represented by red squares. The three segments of the simulation (start-middle-end) are readily distinguishable in the representative starting conformation plot (left). However, in the concatenated trajectories plot (right), there is a substantial level of diffusion and the pathway of the trajectory is not apparent.

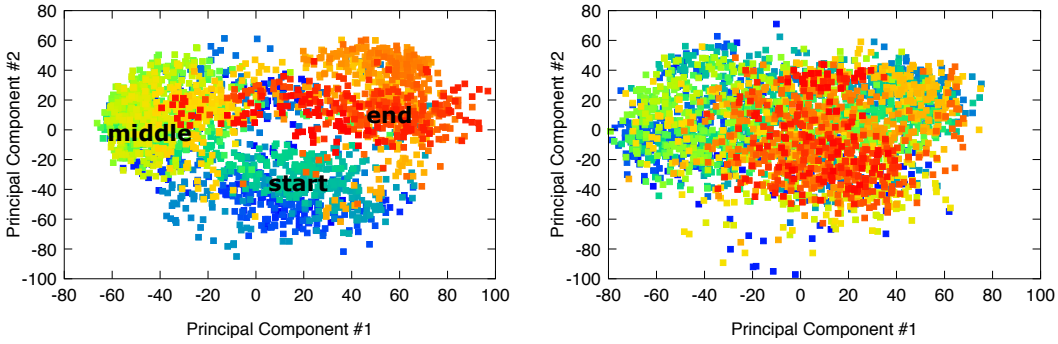


Figure 31. Projections of the trajectory of molecule AC33 onto the planes defined by the first two PCs. The plots are attempting to approach the potential folding pathway versus time. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). Each square is a different frame of the simulation, from the start (blue), to the middle (green), and finally to the end of the simulation (red). Results over the 100 ns trajectory (left) generated from the first starting conformation, and over the collective 500 ns (right) of the concatenated trajectories of the five starting conformations. The start-middle-end parts of the simulation are distinguishable in the left plot, but not in the concatenated trajectories plot (right).

Further PCA plots, as seen in Figure 32, can provide details into the contribution of each of the first two principal components towards each cluster. In the plot provided below, two representative frames from molecules AC24 (left, trajectory #4) and AC33 (right, trajectory #1) are analysed for comparison and in order to add further to the discussion above. The diagram on the left exhibits six different clusters and the high fragmentation that was mentioned in Table 13, in regard to trajectory #4 of molecule AC24, can be observed again. There is significant overlap between each cluster and most seem to be sharing identical areas on the plot. The diagram on the right reveals a different behaviour. The cluster representation is much more efficient, with only limited areas overlapping. However, some fragmentation is present in this plot too, as clusters #1C and #1D seem to be complementary to one another and can therefore be speculated to be in fact one cluster. Indeed, as can be seen in Table S5 in Appendix A, clusters #1C and #1D both represent four-chain folds.

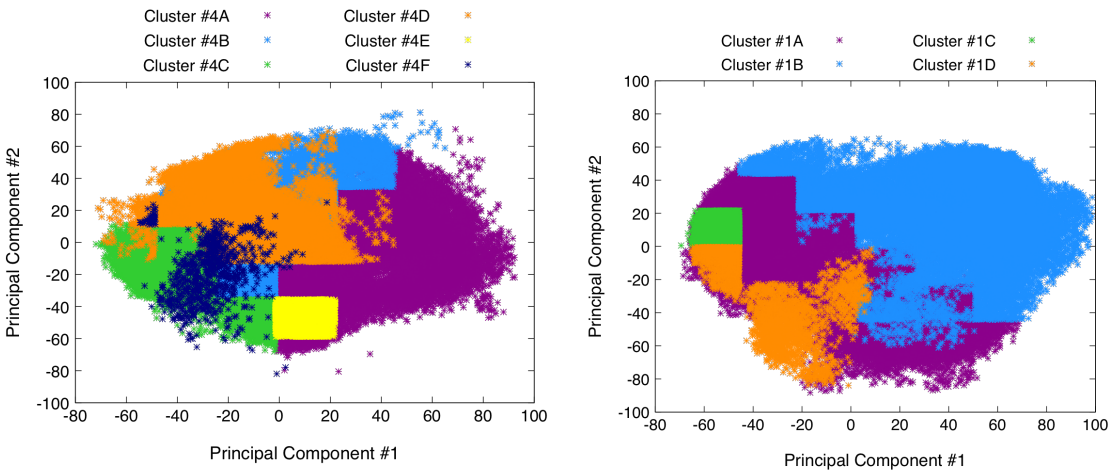


Figure 32. Projections of the trajectories of molecules AC24 (left) and AC33 (right) onto the planes defined by the first two PCs. The plots are exhibiting the contribution of the first two PCs towards each cluster, as can be seen by the different colours used for each cluster. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). Left: trajectory #4 of molecule AC24. Right: trajectory #1 of molecule AC33. Scarcely populated clusters like #4E and #4F (left) or #1C and #1D (right) can be seen occupying small areas of the plot, as opposed to major clusters like #4A and #4B (left) or #1A and #1B (right).

Another interesting correlation that may be unpicked, was a demonstration of negligible interactions between a very structured, tidy variant of the four-chain fold, and the knot fold. This can be seen in Table S21 for molecule AC24. Cluster #3C interacted multiple times with cluster #3A, both representing a four-chain fold, albeit with #3C fold being very structured and tidy. Namely, there were 1507 and 1509 bidirectional interactions between #3C and #3A respectively. However, there were only 20 interactions between the structured four-chain fold (#3C) and the knot-fold (#3B). A similar pattern was exhibited from molecule AC33, as seen in the cluster transitions table (Table S22 in Appendix A). Cluster #1C (tidy four-chain) interacted only with the similar clusters #1A and #1D, demonstrating no interactions with cluster #1B (knot).

Finally, all AMA molecules unsurprisingly demonstrated similarities among variants of the four-chain fold. This can be seen in Figure S4, Figure S5, Figure S6, and Figure S7 in Appendix A, for molecules AC6, AC15, AC24, and AC33 respectively.

Keto mycolic acids

Molecule KC12 demonstrated an interesting behaviour when examining visually similar average matrix distance plots. As shown in Figure 33, there was strong correlation between a tidy four-chain fold and a five-chain fold, represented by structures #1A and #5C respectively. This was newly identified, as all analysis so far had revealed an association among four-chain structures, and between the visually dissimilar four-chain folds and knots.

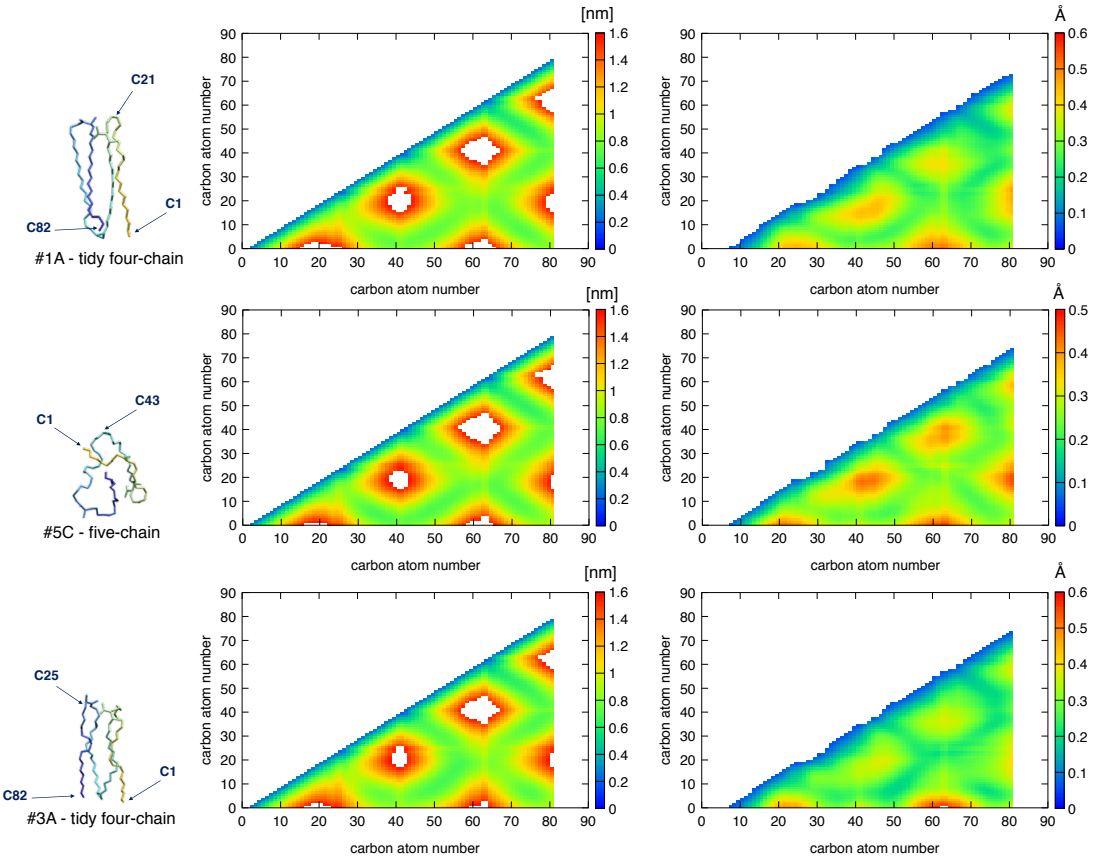


Figure 33. Average structure (left), average distance matrix (middle), and RMSD matrix (right) of clusters similar with #1A from molecule KC12 in explicit water (300 K). For clarity, selected carbon atoms along the molecular chain are indicated with arrows on the structures. The numerical values on the x and y axes (both types of plots) represent the index number of each carbon atom in the MA chain. The intramolecular distances between all carbon atoms along the chain are given in nm, and the colour scheme ranges from blue (0 nm) to red (1.6 nm) as can be seen in the legend on the right of each average distance plot. The fluctuations of the intramolecular distances are provided in Å, and the colour scheme ranges from blue (0 Å) to red (0.6 Å) as can be seen in the legend on the right of each RMSD plot.

Molecules KC12, KC27 and KC29 echoed the correlation between knot and five-chain folds seen in AMAs (AC15 and AC33). Regarding molecule KC12, the results shown in Figure S8 confirmed that knots (#1B) and five-chain folds (#5B) had similar average distance matrix plots. A similar picture was observed in molecule KC27, as shown in Figure S9, with cluster #1B (five-chain) being strongly correlated with clusters #5B (five-chain), and #4B (knot). Further, as seen in Table 14 for trajectory #1 in KC27, there were 1184 and 1174 bidirectional interactions respectively, between the knot (#1A) and the five-chain fold (#1B).

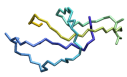
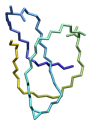
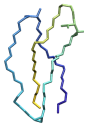
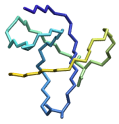
#1 KC27	A - knot 	B - five chain 	C - four chain 	D - knot 
A	0	1184	547	585
B	1174	0	635	6
C	556	626	0	184
D	585	5	185	0

Table 14. Number of cluster transitions occurring within trajectory #1 of molecule KC27. The first column represents the cluster identifier (A-D). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 1174 transitions from cluster #1B to cluster #1A, and 1184 transitions from #1A to cluster #1B.

The plot (left) in Figure 34 delivers further detail about the contribution of the two PCs towards each cluster. It demonstrates that clusters #1A and #1B are indeed tightly correlated, as they are shown to overlap substantially, and to occupy nearly an identical area. Further, the plot on the right offers a greater level of insight in regard to the timestep of the trajectory that the clusters in question occupy. Clusters #1A and #1B occupy a range of timesteps from the start up to the middle of the trajectory. Notably, the cluster distribution in this molecule was considerably balanced, with cluster #1A (relaxed knot) occupying 39% of the trajectory, cluster #1B (five-chain) at 31%, and #1C (four-chain) at 27%, as can be seen in Table S9. This could mean that molecule KC27 folded quickly, which resulted in an anomalous distribution of similar tightly

folded clusters. Further about the correlation between five-chain and knot folds, a similar behaviour for molecule KC29 can be seen in Figure S10.

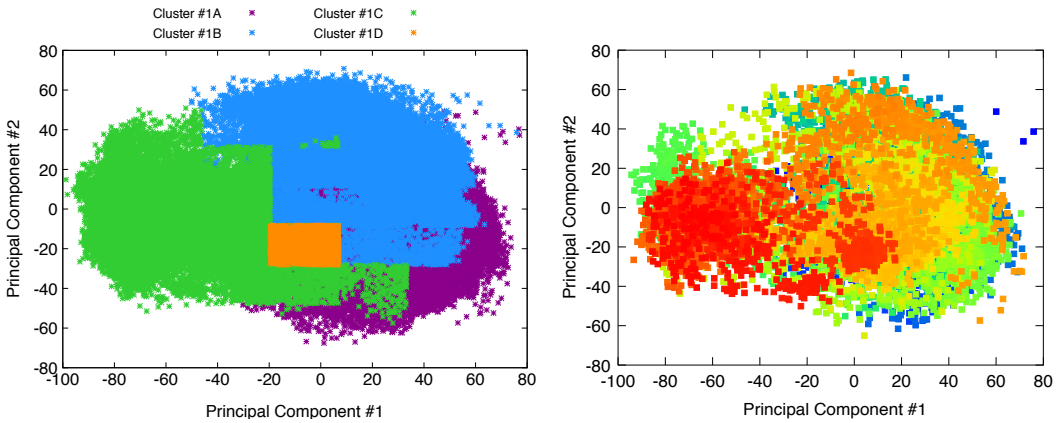


Figure 34. Projections of trajectory #1 of KC27 onto the planes defined by the first two PCs. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). Left plot: contributions of the first two PCs towards each structural cluster (#1A-#1D) of molecule KC27. The most scarcely populated cluster, #1D, can be seen occupying a small area of the plot, as opposed to the major clusters #1A, #1B, and #1C. Right plot: distribution of the 100 ns trajectory over the two first PCs. Each square is a different frame of the simulation, from the start (blue), to the middle (green), and finally to the end of the simulation (red). This plot is attempting to approach the potential folding pathway versus time, which in this case is diffused and the segments of the trajectory (start, middle, end) are not readily distinguishable.

Scarcely populated structures, such as two-chain and three-chain folds, were rare but occurred in molecules KC4 and KC20, similarly as in AMAs (AC15, AC24). In regard to molecule KC4, as can be seen in Table S23 in Appendix A, the two occurrences of two-chain folds (clusters #1D and #1E) had only negligible interactions with the other clusters in the clustering profile. These two-chain conformations were limited to the first few steps of the simulation, when the molecule had not folded fully yet.

Similarly, within the first trajectory of molecule KC20, clusters representing an open-three chain fold (#1D and #1E) had virtually no interactions (Table S24) with any other cluster within the trajectory. These structures represented only a very minor percentage (3.9%) of the total trajectory, limited only in its first few steps. The PCA plots below (Figure 35) offer further detail. The diagram

on the left demonstrates that clusters #1D and #1E are not occupying overlapping areas. At the same time, the diagram on the right confirms that those two clusters are indeed limited only within the first timesteps of the trajectory.

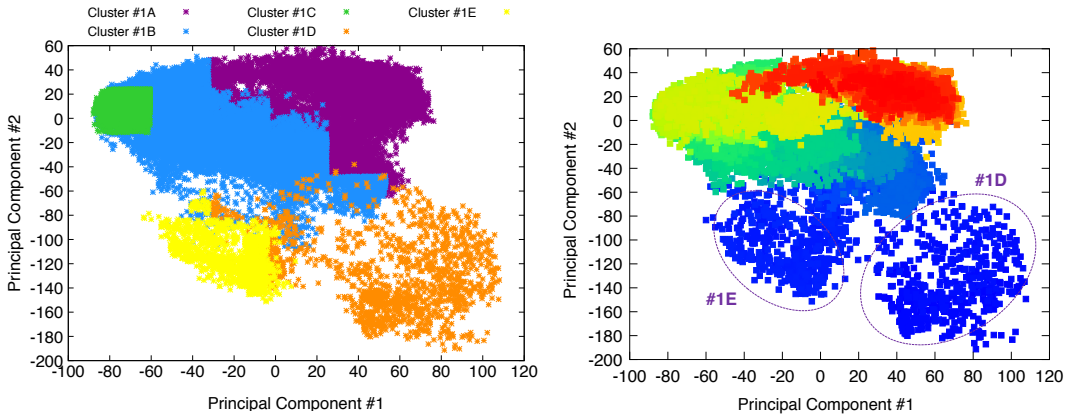


Figure 35. Projections of trajectory #1 of molecule KC20 onto the planes defined by the first two PCs. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). Left: contributions of the first two PCs towards each structural cluster (#1A-#1E) of molecule KC20. The most scarcely populated clusters, #1D and #1E, can be seen occupying small areas of the plot. Right: distribution of the 100 ns trajectory over the two first PCs. Each square is a different frame of the simulation, from the start (blue), to the middle (green), and finally to the end of the simulation (red). The circled areas indicate that the minor clusters #1D and #1E occupy early timesteps (blue squares) in the trajectory, before the molecule assumes its fully folded state.

Further in molecule KC20, a similar behaviour was exhibited within trajectory #4. The transitions, as they are shown in Table S25, demonstrated that the open two-chain folds represented by clusters #4D and #4E had virtually no interactions with any other cluster in the trajectory. These open two-chain conformations represented only a negligible percentage (2.4%) of the total trajectory.

An interesting implication was revealed in Table 15, detailing the cluster transitions within the second trajectory of molecule KC29. Cluster #2D (knot, start) interacted solely with cluster #2A, which also represented a knot structure with the start of the carbon chain emerging through the loop. There were 1043 bidirectional interactions between them. Cluster #2D had absolutely

no interactions with cluster #2C, which also represented a knot fold but in this case the end of the chain emerged through the loop, as opposed to the start. This could imply that the two knot fold variants are independent structures rather than arbitrary patterns of the overall knot fold.

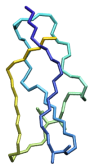
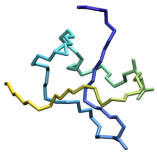
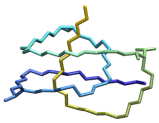
#2 KC29	A - knot 	B - four-chain 	C - knot 	D - knot 
A	0	401	243	1043
B	402	0	614	0
C	241	616	0	0
D	1043	0	0	0

Table 15. Number of cluster transitions occurring within trajectory #2 of molecule KC29. The first column represents the cluster identifier (A-D). The remaining columns represent the pairwise bidirectional transitions between each cluster, for example 402 transitions from cluster #2B to cluster #2A, and 401 transitions from #2A to cluster #2B.

A common theme within two of the examined KMAs was that the trajectories fragmented in many clusters representing nearly identical folds. This behaviour was particularly prevalent in KC4 and KC12. In KC4 (Table S23) this phenomenon manifested itself with each of the five clusters interacting only with cluster #1A and not at all among themselves. Further, within the third trajectory of the same molecule, clusters #3A and #3D both represented a very analogous tidy four-chain fold. In regard to molecule KC12, clusters #5B, #5C and #5D represented comparable five-chain folds, while clusters #5A and #5E represented four-chain packing.

The cluster transition analysis (Table S26) for trajectory #3 in molecule KC12, revealed very frequent bidirectional transitions between clusters #3A (tidy four-chain) and #3B (relaxed knot), namely 1597 and 1588 bidirectional transitions respectively. Further, in molecule KC20, as seen in Table S24, a high dependence between cluster #1A (knot, end) and #1B (distorted four-chain)

was revealed, with 907 and 905 bidirectional transitions between them respectively. This correlation between knot and four-chain folds, is perfectly aligned with what was exhibited from AMAs earlier in the chapter.

Finally, visual examination of average matrix plots followed by RMSD calculation showed that nearly all examined KMAs (KC4, KC20, KC27) demonstrated strong correlation within comparable knot structures, as can be seen in Figure S11 and Figure S12. Similarly, strong correlation was exhibited within analogous four-chain structures (KC4, KC20, KC29).

Results from grid computing approach

Multiple short simulations in various environments

For comparison, and to add further detail to the analysis already presented, selected simulations completed during the grid-computing approach were investigated by performing PCA clustering and internal distances analyses. Namely, the four AMAs (AC6, AC15, AC24, AC33) that have been looked at during the single long simulations analysis, were examined again under a different light. This new evaluation provides insights into how the different sampling, temperatures (300, 350, and 400 K), and solvation systems (explicit water, implicit water and hexane) affect the folding. The total length of each analysed grid-computed simulation can be found in Table S27 in Appendix B.

Additionally, a pair of KMAs is also presented. The chosen KMA molecules were KC25 and KC28 and their chain lengths were 23-17-16-17 and 23-19-16-17 respectively, provided as a short test case in comparison to analogous KMA molecules (KC27, KC29) that this chapter focused on earlier. A comprehensive collection of the cluster percentages is given in Table 16 below. Further, a collated table with comprehensive PCA clustering results is included in Table S28, in Appendix B.

Overall, an interesting insight is revealed from the implicit water simulations at 300 K. This environment left no room for structural variability, as all the

examined cases returned an absolute prevalence of four-chain folds. Interestingly, when a variety of user input options were assessed during calculations with the pyPCAzip package, namely when more principal components and a larger bin size were specified, the results did not deviate. In all cases, only two clusters were generated and they both represented four-chain folds.

Among AC6 and AC15 molecules, the prevalence of four-chain folds was maintained and the four-chain percentage increased through all different temperatures in explicit water, as well as in the two implicit solvent systems. Namely, molecule AC6 started at a four-chain prevalence of 58% at 300 K in explicit water, which increased to 78% at 300 K in implicit hexane, and 98% at 350 K in explicit water. Molecule AC15 started at a four-chain prevalence of 43% at 300 K in explicit water, which then increased to 60%, 90%, and 96%, at 300 K in implicit hexane, and 350 K and 400 K in explicit water, respectively.

Molecule AC24 demonstrated an interesting behaviour by generating identical results between two explicit water temperatures (300 and 350 K). More specifically, in both cases the results consisted of 91% four-chain, and 9% knot and open three-chain folding. At the same time, the prevalence of four-chain folds decreased across the different systems, with the simultaneous increase of the five-chain percentage across the 400 K explicit water (13%) and implicit hexane (50%). The highest four-chain prevalence was noted at 91%, which decreased to 87% at 400 K and 49% in implicit hexane.

Molecule AC33 maintained the knot fold prevalence and increased its knot fold percentage across the different systems. The lowest percentage was at 61% at 300 K explicit water, which increased to 71% and 90% at 350 K explicit water and implicit hexane, respectively. Simultaneously, five-chain folds were dominantly present at 400 K, occupying 98% of the AC33 cluster profile.

	local approach	stepwise production increments: 100 ps explicit / 5 ns implicit			
	explicit water			**implicit water**	**implicit hexane**
	300 K	350 K	400 K	300 K	300 K
AC6 23-11-14-17	58% four-chain 42% knot	98% four-chain 2% open chain	n/a	100% four-chain	78% four-chain 21% knot
AC15 23-13-14-17	57% five-chain 43% four-chain	90% four-chain 10% open chain	96% four-chain 4% open chain	100% four-chain	60% four-chain 39% knot
AC24 23-15-14-17	91% four-chain 5% knot 4% three-chain	91% four-chain 6% knot 3% open chain	87% four-chain 13% five	100% four-chain	50% five-chain 49% four-chain
AC33 23-17-14-17	61% knot 39% four-chain	90% knot 9% open chain	98% five 2% four-chain	100% four-chain	71% knot 29% four-chain
KC25 23-17-16-17	n/a	89% four-chain 11% open chain	71% knot 20% five-chain 8% four-chain	n/a	39% four-chain 35% five-chain 25% knot
KC27 23-17-18-17	96% knot 4% four-chain				
KC28 23-19-16-17	n/a	76% five-chain 23% four-chain	84% knot 16% five-chain	100% four-chain	72% knot 27% four-chain
KC29 23-19-18-17	58% four-chain 29% knot 13% five-chain				

Table 16. Folding preference percentages, resulting from concatenated trajectories analyses, of selected molecules in various systems.

The set of KMA molecules examined, KC25 and KC28, did not reveal a distinct pattern in their behaviour across different systems. Molecule KC25 demonstrated an alternation between dominant four-chain and knot folds. In explicit water, the prevalent folds were four-chain (89%) and knot (71%) in 350 and 400 K respectively. In implicit hexane, there was a balanced representation between four-chain and five-chain folds, at 39% and 35% respectively, with an additional substantial presence of knot folds (25%). Molecule KC28 also demonstrated a variety of dominant folds. In explicit water, five-chain folds (76%) and knot folds (84%) dominated the 350 and 400 K experiments, respectively. In implicit hexane, there was a 72% prevalence of knot folds, with an additional 27% of four-chain folds. In implicit water, there was an absolutely dominant presence of four-chain folds, as discussed earlier in this section.

The set of data for KMA molecules presented in this section is not as comprehensive as it should be in order to provide a complete description of their conformational behaviour across different systems. More analysis is needed on data from different solvation systems and temperatures, in order to build a complete and diverse conformational profile for KMAs. Because of the variance in the resulting conformational patterns and the lack of a distinct trend, it is not possible to pinpoint a clear correlation between the length of the individual methylene segments and the assumed folding.

Collated cluster percentage results from the modelled AMAs can be found in Table 17 below. The highest percentage (72%) of four-chain clusters can be found in the simulations performed at 350 K in explicit water. At the same time, the highest percentage of knot folds can be found at 300 K of explicit water, and that of five-chain folds at 400 K of explicit water.

	concatenated long trajectories from local approach	stepwise production, increments: 100 ps explicit / 5 ns implicit			
	explicit water			**implicit water**	**implicit hexane**
	300 K	350 K	400 K	300 K	300 K
four-chain	58%	70%	62%	100%	54%
open - chain	1%	6%	1%	0%	0%
five - chain	14%	0%	37%	0%	12%
knot	27%	24%	0%	0%	33%

Table 17. Collated cluster percentage results for AMAs. Data are collated across different temperatures (300, 350, and 400 K) and solvent environments (implicit and explicit water, implicit hexane), as well as through two different computational approaches (single 100 ns simulations and stepwise production in 100 ps increments for the explicit solvent environment / 5 ns increments for the implicit solvent environment).

The PCA clustering distribution, as can be seen in Table 18, verified that the pyPCAzip suite performed very consistently when used on data from the stepwise production simulations. Within AMA molecules, the number of resulting clusters was completely consistent through simulations under three different temperatures and three different solvent environments. However, there was a slight fluctuation in the number of clusters resulting from the local approach data, namely from the concatenated simulations.

Further, results stemming from analysis performed individually on the five different trajectories of each molecule exhibited wildly fluctuating numbers of clusters, even within the same molecule. For example, as can be seen in Table S3 for molecule AC15, the first trajectory generated seven different clusters, the second trajectory generated five and the third generated three. Additionally, as can be seen in Table S4 for molecule AC24, the first trajectory generated three clusters and the second generated five. This observation, along with the consistency showed in Table 18, implies that shorter simulations in stepwise increments, as opposed to single long ones, may provide more efficient sampling and may generate more consistent PCA clustering results.

	concatenated long trajectories from local approach	stepwise production, increments: 100 ps explicit / 5 ns implicit			
	explicit water			**implicit water**	**implicit hexane**
	300 K	350 K	400 K	300 K	300 K
AC6	**4**	3	n/a	2	4
AC15	**3**	3	3	2	4
AC24	**4**	3	3	2	4
AC33	**3**	3	3	2	4

Table 18. The number of clusters generated by the pyPCAzip suite across AMA molecules in different temperatures (300, 350, and 400 K) and through two different computational approaches (single 100 ns simulations and stepwise production in 100 ps increments for the explicit solvent environment / 5 ns increments for the implicit solvent environment).

PCA plots demonstrating the contributions of each PC towards each cluster can be seen in Figure 36. Plot (a) echoes the findings regarding folding behaviour in implicit water, as these were discussed earlier in this section and as can be seen in Table S28. It was noted that implicit water did not allow for any structural variability and generated a stubborn 100% dominance of two four-chain clusters every time, regardless of user input in the pyPCAzip suite. Plot (a) demonstrates that the two principal components were overlapping completely. Plots (b), (c), and (d), revealed clearer PC contributions towards each cluster, with much less overlapping. In these three diagrams, the main clusters complement each other and the most scarcely populated cluster overlaps in a small area of one of the major ones.

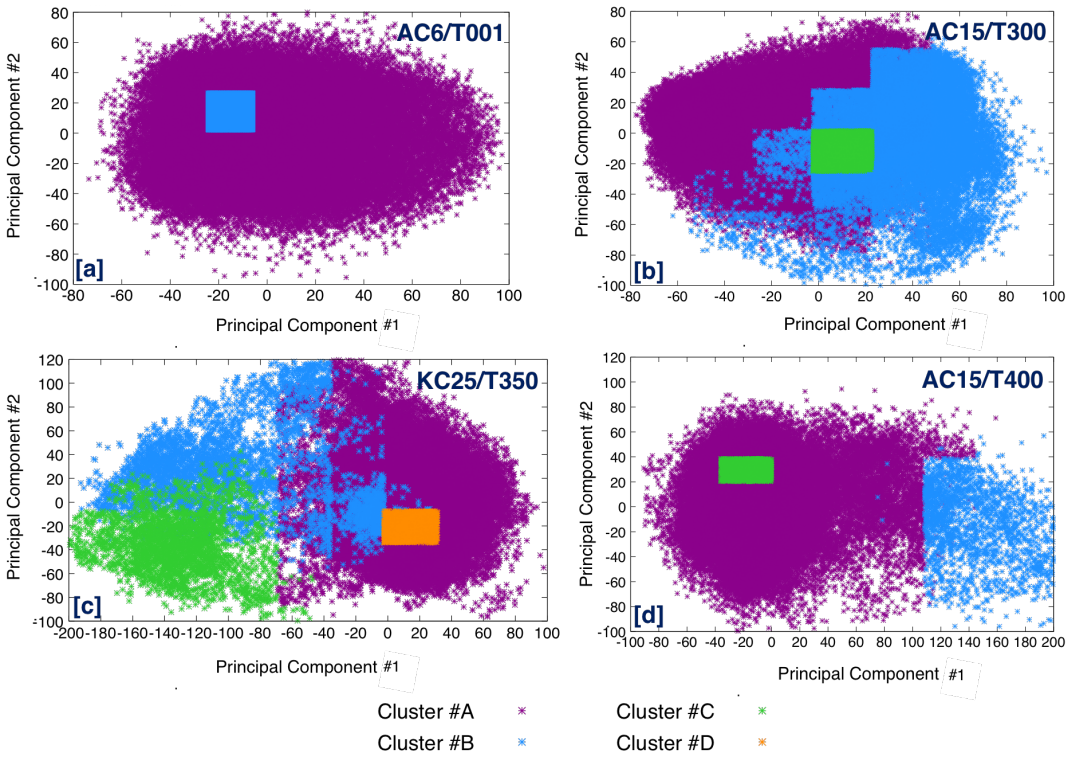


Figure 36.Projections of the trajectories of molecules AC6, AC15, and KC25 onto the planes defined by the first two PCs. The plots are exhibiting the contribution of the first two PCs towards each cluster, as can be seen by the different colours used for each cluster. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). The data in plots [a], [c], and [d] resulted from stepwise simulations performed on the WCG servers, whereas the data in plot [b] resulted from single 100 ns simulations performed on the local HPC facility. The trajectory of molecule AC6 in plot [a] was simulated in implicit water at 300 K; the trajectories of molecule AC15 was simulated in explicit water at 300 K (plot [b]) and 400 K (plot [d]); the trajectory of molecule KC25 in plot [c] was simulated in explicit water at 350 K. Scarcely populated clusters (cluster #B in plots [a] and [b] and clusters #D and #C in plots [c] and [d] respectively) can be seen occupying only small areas of the plot.

Further analysis included computing the intramolecular carbon atom distances within each molecule. Intramolecular distance matrices were created by calculating all the carbon atom distances along the molecular chain for every step of the trajectory. The average matrix for the complete trajectory of each cluster was subsequently calculated. In order to elucidate the effect of the different temperatures and solvents on the intramolecular distances, RMSD calculations were performed in a pairwise fashion between each entry of the average distance matrices from different environments.

The comparisons were performed between clusters within the following pairs: T000/T001, T000/T300, T000/T350, T000/T400, T001/T300, T001/T350, T001/T400, T300/T350, T300/T400, and T350/T400, therefore covering all possible combinations. It is noted that, for simplicity, 'T000' denotes simulations at 300 K in implicit hexane, and 'T001' denotes simulations at 300 K in implicit water. Comprehensive results from these comparisons, from all the molecules discussed in this section, can be seen in Table S29 in Appendix B. The lowest RMSD values, denoting the most similar clusters within different environments, can be found in Table S30, also in Appendix B.

As can be seen in Table S30, the most similar clusters were most often stemming from comparisons between clusters representing the same fold. More specifically, in 54% of the comparisons the highest similarity was between four-chain clusters, 17% between four-chain and knots, 13% between five-chain and knots, 9% between knots, and, finally 6% between four- and five-chain folds.

The lowest and second lowest RMSD values (0.028 Å and 0.033 Å) were noted in molecules AC33 and AC24 respectively, in comparisons between implicit hexane and explicit water (both at 300 K) clusters. In both cases, the clusters represented four-chain folds. The third lowest RMSD value (0.036 Å) was noted in molecule AC6, in a comparison between cluster C (implicit hexane) and cluster B (explicit water, 300 K), both representing knot folds.

One of the most similar pairs, with an RMSD value of 0.047 Å, of visually non-analogous clusters was cluster B (implicit hexane) and cluster A (explicit water, 300K) from molecule AC15. Cluster B represented a knot structure, whereas cluster A represented a five-chain fold. Their relevant average and RMSD distance matrices are given in Figure 37 below.

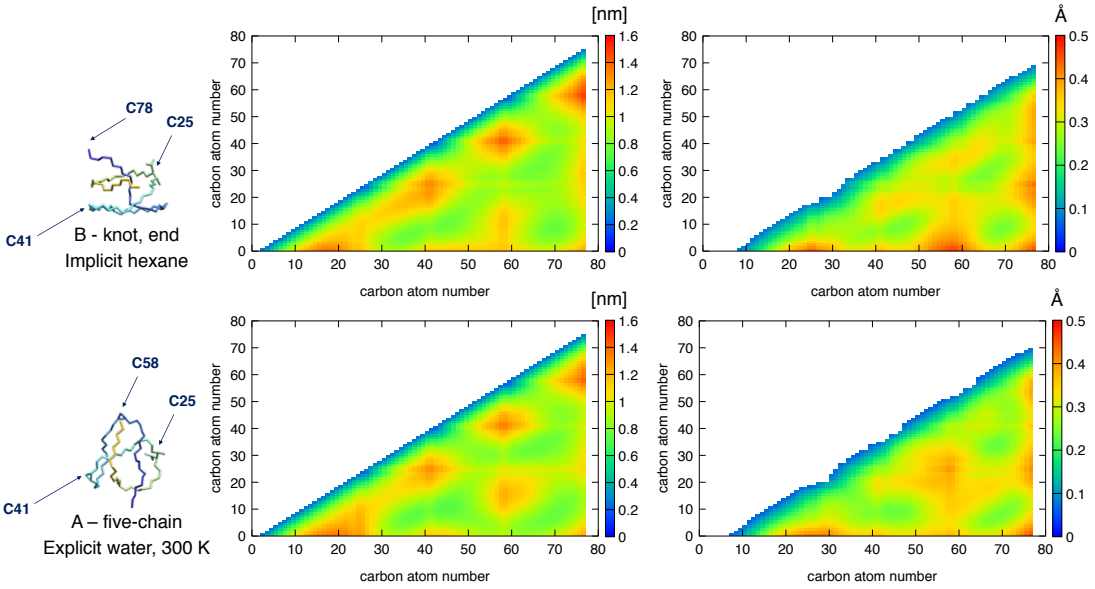


Figure 37. Average structure (left), average distance matrix (middle), and RMSD matrix (right) of cluster B (top, implicit hexane, 300 K) and cluster A (bottom, explicit water, 300 K) from molecule AC15. For clarity, selected carbon atoms along the molecular chain are indicated with arrows on the structures. The numerical values on the x and y axes (both types of plots) represent the index number of each carbon atom in the MA chain. The intramolecular distances between all carbon atoms along the chain are given in nm, and the colour scheme ranges from blue (0 nm) to red (1.6 nm) as can be seen in the legend on the right of each average distance plot. The fluctuations of the intramolecular distances are provided in Å, and the colour scheme ranges from blue (0 Å) to red (0.5 Å) as can be seen in the legend on the right of each RMSD plot.

Another pair of highly similar clusters, with an RMSD value of 0.036 Å, was cluster B (explicit water, 350 K) and cluster C (explicit water, 400 K), representing a knot and a four-chain fold, respectively. Their relevant plots and average structures can be seen in Figure 38 below.

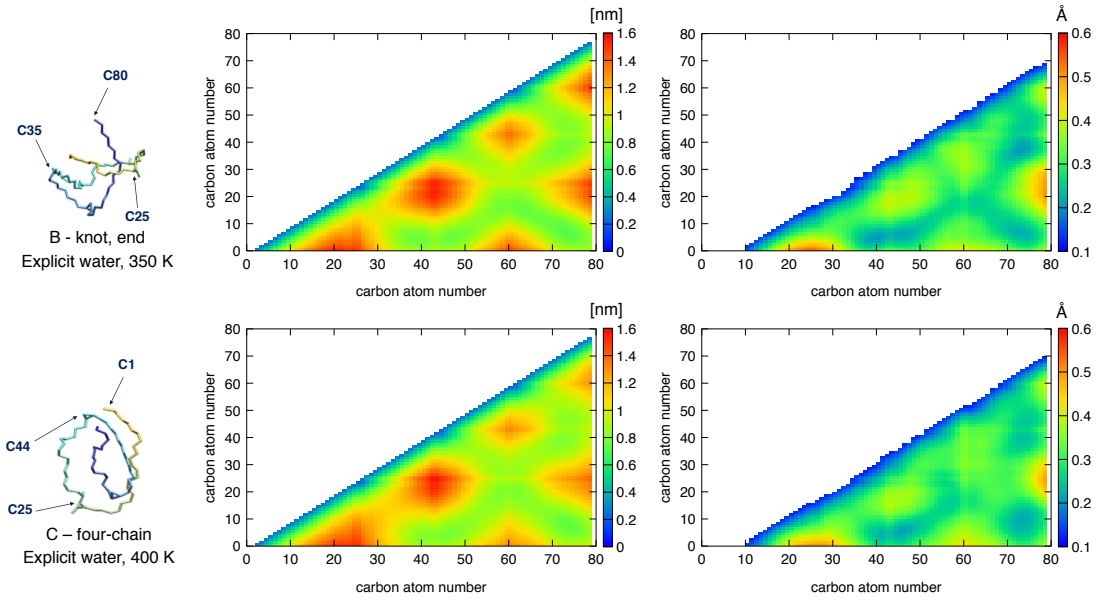


Figure 38. Average structure (left), average distance matrix (middle), and RMSD matrix (right) of cluster B (top, explicit water, 350 K) and cluster C (bottom, explicit water, 400 K) from molecule AC24. For clarity, selected carbon atoms along the molecular chain are indicated with arrows on the structures. The numerical values on the x and y axes (both types of plots) represent the index number of each carbon atom in the MA chain. The intramolecular distances between all carbon atoms along the chain are given in nm, and the colour scheme ranges from blue (0 nm) to red (1.6 nm) as can be seen in the legend on the right of each average distance plot. The fluctuations of the intramolecular distances are provided in Å, and the colour scheme ranges from blue (0 Å) to red (0.5 Å) as can be seen in the legend on the right of each RMSD plot.

Discussion

The chemical structure and composition balance of MAs in the inner leaflet of the mycobacterial cell wall determines *M. tb* survival *in vitro*, growth *in vivo*, and persistence in the host macrophage.[46, 78-79, 83] The physical properties and biological roles of free MAs and MA derivatives may be driven by their chemical make-up,[83] and various studies have highlighted the importance of the different MA functionalities. *Trans*-cyclopropanation of oxygenated MAs suppressed inflammation and decreased virulence in mice.[104] Oxygenated MAs have been proposed to play a major role in host lipid accumulation and foam cell formation at the site of infection, and are thought to facilitate long-term persistence in the host.[152] Mutant mycobacterial strains with altered MA composition exhibited increased cell wall permeability[79, 81-82] and decreased virulence.[85] The different chain lengths, and the stereochemistry and functional

group nature of the MAs, directly modulate MA conformational profiles in pathogenic mycobacteria.[47, 75] It is therefore clear that the relationship between the chemical structure and folding conformation of MAs is anticipated to be of biological significance in mycobacteria.

Various monolayer experiments[45, 75, 93-98] have focused on studying the conformations of MAs. The π vs. A isotherms indicated that the molecules assumed relatively constrained conformations at low temperature and surface pressure conditions. MMAs and AMAs were shown to unfold as temperature and surface pressure increased.[45, 75, 93-97] KMAs retained the tight fold even at high temperature and surface pressure,[45, 75, 94, 97] thus demonstrating a unique trend potentially correlating with a distinctive biological function.

Additionally, computational studies[47, 75, 93-94, 101] have focused on the conformational profiles of MAs. Groenewald *et al.*[47, 101] investigated the conformational patterns and the structure-function relationships of a group of naturally-occurring MAs. All three major *M. tb* MA classes were subjected to MD simulations *in vacuo*[47] and in explicit water and hexane.[47, 101] These studies demonstrated that MAs tended to fold spontaneously into manually-defined 'WUZ'-folds, with KMAs exhibiting the lowest frequency of 'WUZ' folds. Villeneuve *et al.*[75, 93-94] demonstrated that KMAs tended to assume tight conformations that did not stretch out, whereas MMAs and AMAs exhibited relative flexibility. A detailed presentation of the monolayer and computational studies discussed briefly in this section can be found in the Introduction.

The achievements of the current work and the improvements it introduces to the current state-of-the-art are the following: (i) modelling a large selection of both natural and non-natural MAs (all three *M. tb* classes, a wide variety of chain lengths, both proximal cyclopropane stereochemistries), (ii) using a multi-step initial sampling method to ensure sufficient sampling, (iii) performing long atomistic MD simulations, (iv) assessing the effects of intrinsic (chemical nature) and extrinsic (different temperatures and solvation systems) factors on MA folding , (v) utilising a novel analysis protocol which consists of

a variety of conformational determination methods and software (dihedral angle clustering, PCA clustering, time-resolved intra-molecular distance matrix calculations), and (vi) employing two different computational workflows (single long MD simulations on the University of Nottingham high performance cluster, and multiple shorter simulations conducted in parallel on a distributed computing platform). This is the first attempt to perform MA conformational studies on a worldwide distributed computing platform and on such an extensive array of MAs. This section elaborates on specific aspects of each analytical and computational approach, such as their shortcomings and strengths, and their wider applicability for structural determination efforts, particularly in biological studies. It also focuses on main findings as these were presented in the Results section.

Two computational approaches have been employed in this work, namely single long (performed on the local HPC facility) and multiple short simulations (performed on the IBM grid-computing platforms). Single long simulations had the clear advantage of a more straightforward file handling routine, with four output files (trajectory, structure, log, and energy file) being generated for each molecule. In comparison, running multiple short simulations in a stepwise manner resulted in four output files being generated per step and per molecule. That meant that file handling became more challenging, particularly in terms of the analysis, as multiple results files needed to be concatenated, and it also dictated a crucial need for a robust data management workflow so that all files were organised methodically.

Further, the interactive nature of the HPC facility enabled the equilibration and early setup stages to be completed seamlessly. As these preliminary and essential steps required frequent user input and assessment, they would not have been possible to complete on grid-computing platforms. In comparison, the nature of grid-computing platforms dictated the need for development and implementation of efficient data handling workflows.

Internal coordinates such as bond lengths, bond angles, and torsion angles (BAT), are natural coordinates for describing effectively a bonded molecular system.[153] Internal coordinate MD (ICMD) methods have been applied to approach a variety of structural biology problems, such as protein folding[154] or protein structure refinement.[155] Further, the selection of internal coordinates, such as torsion angles, for clustering, rather than opting for the 3N independent Cartesian coordinates of all atoms, is more effective for structural and conformational determination studies. Proteins possess thousands of degrees of freedom and this is one of the main reasons why ICMD approaches[153-155] have strong applicability in structural biology problems. MAs are long molecules, bearing chains of 60-90 carbons,[49] with many degrees of freedom. Utilising internal coordinates, such as dihedral angles, reduces the size of the data set, and is therefore much less computationally intensive than using the vastly larger data set of Cartesian coordinates. At the same time, because dihedral angles change more during folding events than bond lengths and angles do, they offer the potential to deliver insightful structural data.

The application of dihedral clustering using the DASH[147] algorithm on two MA molecules, namely AC6 and KC4, returned results with negligible structural diversity and therefore failed to address the frequent refolding events observed in the MA trajectories. The frequent refolding events and the large number of carbon dihedrals (namely, 75 dihedral angles for AC6 and 77 for KC4, respectively) deemed the extraction of meaningful data from the DASH output very challenging. The representative DASH states can be seen in Figure 25 and Figure 26 in the Results, for molecule AC6 and KC4 respectively. Both sets of results confirmed the dominant presence of the four-chain fold. This was echoed by the PCA clustering findings, as these can be seen in Table 6 and Table 7 in the Results, where the four-chain folds were the major clusters in both KC4 (72%) and AC6 (55%) molecules. It may be possible that the predominantly C-H structure of MAs, as opposed to a highly hydrogen bonding one, could be the molecular driver for the limited applicability of dihedral clustering in MAs. In contrast, dihedral clustering in proteins, where hydrogen bonding is the interaction that predominantly modulates the folding,[156] was shown to have

great applicability. In conclusion, despite the well-established applicability of dihedral angle clustering in a variety of conformational determination studies,[149-151] it was not a suitable approach for efficient structural analysis in this work, for the reasons outlined in the discussion above.

To assess the dynamic nature of MD simulations, a method that takes the temporal information that is inherent in trajectories into account is necessary. Recently, Mori *et al*.[157] developed a Dynamic Component Analysis (DCA) approach, that builds on PCA. PCA focuses on the static properties of the trajectory and does not make use of the temporal information.[157] In contrast, in DCA, the coordinates from the cross-time correlation function matrix (TCF) are derived and are used to construct a distance-map coordinate which is further converted to an N-dimensional vector. The DCA method was applied on the FiP35 WW protein domain, which has 35 residues. This meant that when the C^{α} atoms were used to construct the distance-map coordinate, a 595 $(= 35 \cdot 34/2)$-dimensional vector was constructed. The DCA method successfully described the overall folding/unfolding events in the trajectory, and further identified two misfolded states.

A simplified approach of the DCA method is used in the current work, with the time-resolved intramolecular distances of all the carbon atoms across the MA chain being calculated. Time-resolved distance matrices of backbone carbon atoms can assess the folding/unfolding frequency and are an excellent visual tool in directly observing the differences between the various MA folding patterns, without the need to look at each cluster structure separately. In contrast to the outcomes of the DASH[147] analysis, the distance matrices generated meaningful data about structural variability by managing to showcase the structural diversity that the analysis of MA trajectories can deliver. A comprehensive collection of diverse average distance matrices, along with their corresponding cluster, is provided for reference in Figure 27 in the Results.

The slight disadvantage of the intramolecular distance calculations is that distance matrices are approximately N-fold more computationally expensive to calculate and to store, N being the number of atoms in a molecule. Distance matrices possess $N \cdot (N-1)/2$ distance interactions as opposed to $(N-3)$ dihedral angles, so it becomes clear that the distance matrices rapidly become much more computationally expensive to calculate and to store, especially for larger molecules. For reference, a DASH output file for an entire 100 ns trajectory is around 60 Kilobytes large, and the calculation takes a few seconds to complete. In comparison, a distance matrix file for a specific timestep in a trajectory is also around 60 Kilobytes large, and the calculation takes several minutes to complete.

The inherent association between different structural clusters was investigated initially by visual inspection and subsequently with RMSD comparison between their relevant average distance matrices. This visual inspection aims to quickly detect associations between visually dissimilar folding conformations. However, this approach has limited applicability on large datasets, as visual inspection of distance matrices in that case would not be feasible. For extensive analyses of the complete array of data generated on the IBM grid-computing platforms during this doctoral work, an automated robust workflow is needed. An effective way to harness the strong structural detail contained in distance matrices would be to implement automated tools such as the one proposed by Kloczkowski *et al.*[158] Their mathematical approach delivers an efficient strategy for comparative molecular modelling, by enabling protein structure prediction and refinement, using interatomic distances.

After preliminary visual assessment of the similarity between two average distance matrices that corresponded to visually dissimilar clusters, the RMSD value between the two matrices was calculated to confirm the inherent structural similarities between the clusters. The pairs between which the lowest RMSD values were observed, were presented throughout the Results section, in figures analogous to the one seen in Figure 29. Among AMAs, strong association was observed between knots and four-chain folds, knots and five-

chain folds, and, unsurprisingly, among four-chain folds. The similarity between distance matrices of visually analogous clusters is expected. However, the strong association between visually dissimilar structures, such as four-chain folds and knots, reveals the strong interplay between different clusters, as this is demonstrated by their inherent characteristics (backbone carbon atom distances). In KMAs, strong association was also observed among knots and five-chain folds, four- and five-chain folds, and among knots.

The above correlations are important as they highlight the dependency between different clusters. Folding conformations that looked vastly different when visualised, were, in fact, clearly interconnected. Such relationships were further highlighted by calculating the bidirectional transitions between various clusters. It was demonstrated that there were many frequent interactions between visually dissimilar structures, thus echoing the results produced from the distance matrix comparisons as these were discussed earlier in this section.

Further, this analysis also showed that scarcely populated clusters had negligible interactions with the rest of the trajectory. This was due to the fact that these minor structures belonged to open two- or three-chain folds, were short-lived, and limited to the first steps of the trajectories. The insight that open, unfolded structures are only short-lived and they pertain to the start of the MD trajectories, led to an additional piece of information in regard to MA folding pathways, which was that MA molecules tended to assume folded conformations from early steps of the trajectory, and retain them until the completion of the simulation.

Finally, two variants of the knot fold, as discussed in detail in the Results section, demonstrated an interesting behaviour in regard to interactions with the other clusters in the trajectory. In two cases, namely within the clustering profiles of the second trajectory of molecule KC29 (#2 KC29) and the fourth trajectory of molecule AC24 (#4 AC24) respectively, no interactions were observed between two knot fold variants that involve the opposite end of the chain emerging through the loop (*knot start* versus *knot end*). The relevant data

can be seen in Table 9 (#4 AC24) and Table 11 (#2 KC29) in the Results. The lack of interactions between the two knot-fold variants could mean that they are more strongly associated with other clusters in the trajectory, and at the same time it implies that the knot variants are independent structures that do not emerge arbitrarily from the overall knot fold profile. They represent valid, independent conformational patterns, that may be worth exploring through further conformational determination studies. Knots and entanglements are topologies often observed at the molecular level.[159] In fact, recent surveys of the Protein Data Bank (PDB) identified hundreds of knotted protein structures, accounting for around 1% of total entries.[160] Knotted topologies can range from the simplest trefoil knot, up to the most complex Stevedore's knot.[159] The vast majority of knotted proteins belong in the simplest trefoil topology, which also resembles the knotted MA conformations (particularly the *bow* variant) identified in this work.

In nature, molecular knots can be found in DNA, RNA, and proteins.[159] Many knotted proteins exhibit well-characterised enzymatic functions,[161] such as RNA methyl transfer, and their folding pathways have been investigated experimentally and computationally. Recently, Mallam *et al.*[162] showed that two trefoil-knotted RNA methyl transferases, YibK from *Haemophilus influenzae* and YbeA from *Escherichia coli*, can spontaneously self-tie *in vitro* without molecular chaperones. The folding is enhanced by bacterial chaperonins *in vivo*. Experimental data confirmed that the knotting and folding pathways are robust, even when initiating from an unknotted linear state. Sułkowska *et al.*[163] performed a computational study on the folding mechanism of YibK and YbeA, using a standard C^{α} structure-based model. Analysis of the folding trajectories indicated that the knot folding of YibK and YbeA proceeds through a 'slipknot' intermediate configuration, as shown schematically in Figure 39 below.

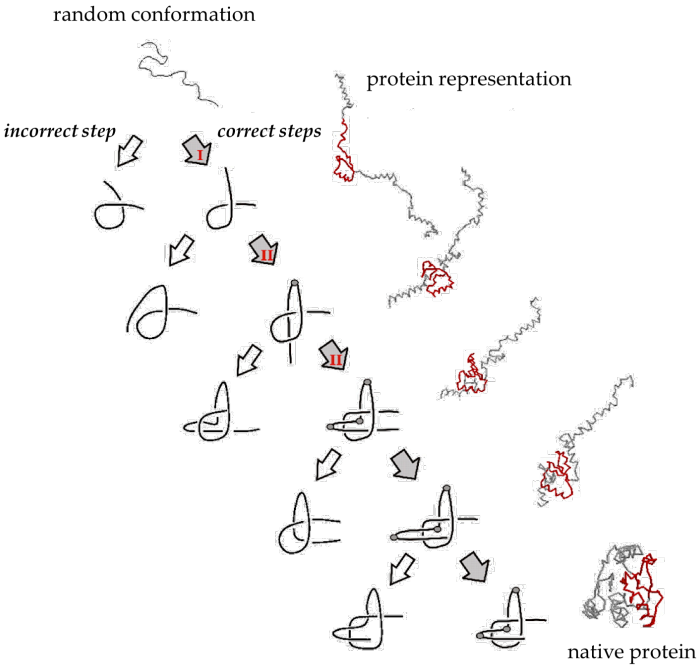


Figure 39. Schematic representation of the proposed folding mechanism for YibK and YbeA. The folding route leads to the native trefoil knot through an intermediate slipknot configuration. Adapted by Sułkowska *et al*.[163]

The progression of the proposed folding pathway from an unknotted shape, through to an intermediate slipknot configuration, and finally to the knot topology, could also pertain to the knot MA folding pathway. This is supported by the fact that the structures observed in the proposed mechanism, as seen in Figure 39 above, fit well with the observed MA knot variants (relaxed knot, knot - end, knot - start, and bow). Since most natural knots are thought to play a key role in the structural, dynamic and/or functional properties of the biological systems they are associated with,[159] and are progressively becoming targets of chemical synthesis,[164] it is clear that understanding how knot topologies form and how the properties of knotted structures differ from unknotted ones is vital.

Apart from the analytical methods already discussed, a PCA clustering approach was employed in order to reveal key conformational changes that take place in MD simulations. While PCA is a well-established analytical tool in protein studies, used to identify structural groupings,[165-168] its applicability to lipid structures, due to their chemical nature and smaller size, is less apparent.

However, as mycobacterial MAs are among the longest mycolates in nature, with a 60–90 carbon backbone,[49] PCA is suitable for use towards the conformational analysis of these molecules.[47] The PCA clustering package utilised in this work was pyPCAzip.[148] pyPCAzip has been successfully used for analysis of protein dynamics and has contributed to the identification of key structural characteristics within the studied molecules. Di Pietro *et al*.[167] employed a multistep computational approach towards the conformational analysis of BACE-1, a membrane-associated protease. This work consisted of 100 ns MD simulations on 200 replicates of the apo and ligand-bound structure of the enzyme. Selected frames were extracted from each trajectory and were concatenated. They were then subjected to PCA and clustering analysis using pyPCAzip.[148] The results were 13 clusters for the apo and 50 clusters for the ligand bound. The PCA analysis on the protein dynamics, together with studies of pocket druggability, detected a novel druggable pocket in the molecule.

To visualise the PCA clustering results, plots that focused on the contribution of the first two PCs towards each cluster were generated. Such plots can be seen in Figure 32 and Figure 34 (left) in the Results. In such plots, hard, straight lines can be observed in the distributions, as is highlighted with a white dashed line in Figure 40 below. The behaviour demonstrated in these plots is expected. The clustering tool in the pyPCAzip suite operates by putting data into multidimensional bins in the PC conformational space, and then using an algorithm to partition the bins into separate clusters. Every data point in a bin is assigned to the same cluster. The dimensions of those bins are defined by the user and are set by a 'grid' parameter. If a finer grid is chosen, the hard lines that can be observed in the plots below would be less obvious. However, this would in turn mean that there would be fewer data points per bin, leading to more noise in the distribution, and ultimately resulting into the generation of spurious low-occupancy clusters.

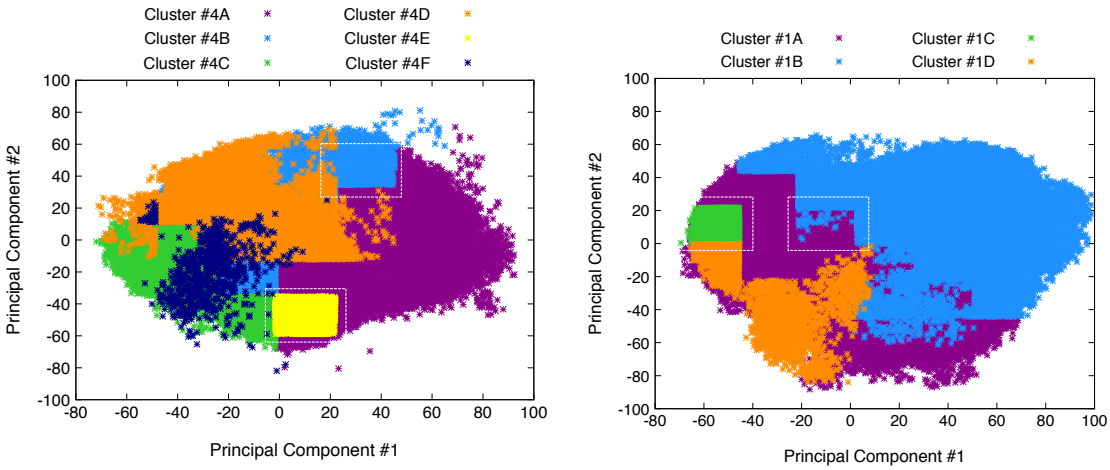


Figure 40. Projections of the trajectories of molecules AC24 (left) and AC33 (right) onto the planes defined by the first two PCs. The plots are exhibiting the contribution of the first two PCs towards each cluster, as can be seen by the different colours used for each cluster. The axes represent the first two principal components, which contain the maximum amount of variance of each atomic distance (Å). The hard, straight lines observed due to grid parameter are enclosed in white dashed lines in the plot. Left: trajectory #4 of molecule AC24. Right: trajectory #1 of molecule AC33. Scarcely populated clusters like #4E and #4F (left) or #1C and #1D (right) can be seen occupying small areas of the plot, as opposed to major clusters like #4A and #4B (left) or #1A and #1B (right).

In order to sample the conformational space in an efficient way and to obtain a comprehensive perspective on the data sets, pyPCAzip[148] was utilised to perform clustering analyses on data resulting from both computing approaches employed, namely the local HPC approach and the grid-computing approach, as follows: (i) separately, for each independent starting conformation of the studied molecules (data from local approach), (ii) collectively, on the concatenated trajectories of all different starting conformations (data from both approaches). Analyses were performed on a range of temperatures (300 K, 350 K, and 400 K) and in different environments (explicit water, and implicit water and hexane). Implicit solvation uses a continuous medium, rather than explicit solvent molecules, to represent the solvent,[141] while providing an accurate representation of the electrostatic effects of the solvent.[116] The advantage of implicit models versus explicit ones, is that the absence of viscosity in implicit solvation means that the studied molecules can explore the conformational space more quickly.[116] The disadvantage is that, by definition, implicit models

are unable to provide information about how direct solvent interaction modulates the folding.

PCA followed by structural clustering on concatenated MD trajectories is well-established in studies of the conformational space of proteins. Papaleo *et al.*[165] employed an integrated analytical protocol utilising PCA and subsequent structural clustering on trajectories of the native structure of holo-myoglobin. MD simulations were performed at 300 K, spanning 0.22 μs. Eleven independent 20 ns simulations (replicas) were carried out. PCA was performed individually on each replica and on the total eleven concatenated trajectories. RMSD matrices were computed from the concatenated trajectories by least-square fitting of the main chain atoms, and they were then subjected to clustering analysis. Even though PCA was performed on the individual replicas as well as the total concatenated trajectories, the structural clustering was only performed on the concatenated trajectories. This work highlighted that myoglobin does not undergo large conformational changes, but that its conformational ensemble is characterised by subtle conformational changes.

The major conformational preference of the concatenated trajectories of AMAs in all systems was consistently the four-chain fold. This was true across three temperatures (300, 350, and 400 K), across three solvation systems (explicit water, implicit water and hexane), and for both computing approaches (local simulations and grid-computing approach). This behaviour of AMAs is not fully aligned with the conclusions derived about this MA class from other computational[47, 93, 101] and monolayer[45] studies. The monolayer studies[45] had concluded that AMAs tended to assume a W-shape at low temperature and surface pressure, but these folds stretched-out when the temperature and the surface pressure increased. Further, computational studies[47] had concluded that AMAs simulated in vacuum spent more time in open conformations, compared to MMA and KMA. Additionally, AMAs had been shown[101] to assume the lowest percentage of 'W' folds, compared to MMA and KMA, in vacuum, explicit water and hexane.

Various reasons can explain why conclusions of the current study about the major AMA conformational preferences are not fully aligned with the conclusions of previous studies. Most notably, the length of the simulations in the computational study performed by Villeneuve *et al.*[93] was very short at only 35 ps, and there was very little variation of starting conformations. In the studies performed by Groenewald et *al.*,[47, 101] each run started from an open structure and was either 4 ns in duration, in vacuum,[47] or lasted 10 ns in vacuum, explicit water and hexane.[101] Further, in some cases there was no explicit mention of the solvent used[93] and therefore implicit solvent or vacuum is assumed. The multi-step approach employed in the current study to ensure efficient sampling, was thus more efficient and thorough than in previous computational MA studies.[47, 93, 101] It is therefore possible that the MA molecules investigated in those studies were not able to explore the conformational space fully, as the MA folding pathways involve many folding and unfolding events. It is crucial to allow the MA molecules to generate a comprehensive representation of their folding variability, in order to avoid missing important folding events.

Groenewald *et al.*[101] additionally observed that a large proportion of the resulting folding conformations could not be assigned to the 'WUZ' folds. Those were structures where the molecules twisted and bent at various points along the alkyl chains, thus resulting in various complex patterns of interacting chains that were difficult to define in two dimensions. These structures tended to be of higher energy compared to the 'WUZ' folds. More specifically, only 23% of the vacuum, 30% of the water, and 21% of the hexane simulation structures (as can be seen in Table 2 from the Introduction) were attributed to the 'WUZ' folds, and the rest were undefined.[101]

The 'WUZ' analysis carries limitations, as the conformational characterisation it achieves is based on specific requirements for two dimensional intramolecular distances, as was discussed in more detail in the Introduction. The two-dimensionality of the predefined cut-offs increases the possibility that structures that do not resemble the attributed 'WUZ' shape may nevertheless

be included in the 'WUZ'-defined folds, while at the same time structures that resemble 'WUZ' shapes are omitted.[101] The analysis tools used in the current study, most notably the pyPCAzip[148] suite, do not rely on predefined requirements for conformational characterisation, but rather utilise the intrinsic detail provided by principal components across the conformational space. Further, while the eight-point distances, as seen in Figure 7 in the Introduction, that are taken into consideration in the 'WUZ' scheme, offer a certain level of detail in the folding behaviour of the molecular chain, they lack the unbiased and comprehensive information that the intramolecular carbon atom distances offer.

In the current study, at 300 K and in explicit water, AMAs from individual analyses of single MD trajectories demonstrated a preference for the knot-fold, as can be seen in Table 7 from the Results. In this approach, 48% of the clusters were knot folds, 45% were four-chain folds, and 6% were five-chain folds. While the fact that the knot was identified as the main cluster in these analyses could be considered as an outlier, considering that the analysis from concatenated AMA trajectories indicated the four-chain fold as the major cluster, it was nevertheless a valuable piece of information to observe, as free MAs occurring in biofilms[87] are more likely to assume globular conformations.

Data from explicit water simulations demonstrated the highest degree of structural variability in terms of resulting clusters. This could be because the MAs interacted with the solvent in ways that modulated their folding. Hydrophilic interactions between the solvent and the MAs, particularly oxygenated ones like MMAs and KMAs, have been well documented in previous computational[75] and monolayer[45] studies. Additionally, the internal pressure applied from explicit water molecules to the mycolate could be promoting a slower reorganisation of the chains, thus allowing the clustering algorithm to identify more structures along the folding pathway. In contrast, the clustering analysis of implicit water simulations in this study resulted in only two four-chain clusters in all molecules examined. As MAs are molecules that spontaneously tend to assume folded conformations,[45, 47, 75, 101] it is

reasonable to assume that the lack of explicit interactions between the MA and the solvent in the implicit models, could have allowed MAs to fold quickly into four-chain folds.

The implicit hexane results observed in this study contrast with the conclusions derived from previous studies[101] in the same solvent. It had been previously shown that AMAs and MMAs tended to assume unfolded conformations in hexane, with many straight chains with various kinks, or 'aU' structures, according to the 'WUZ'-folds nomenclature. A schematic representation of the 'WUZ'-folds can be found in Figure 10 and Figure 11 in the Introduction. The current work has demonstrated that a high level of organisation is retained even in hexane for AMAs, with four-chain percentages in AC6 and AC15 at 78% and 60%, respectively.

Further, KMAs in hexane had been reported to assume tighter folds than the largely extended conformations of AMAs and MMAs,[101] with the KMA preferred fold being the 'sU'. Although the 'sU' conformation is certainly more organised than the unfolded shapes of AMAs and MMAs in hexane, the current study showed that KC25 had a 39% four-chain cluster, with KC28 following at 27% (with a 72% knot). Although more data are needed to be able to extrapolate a conclusion for a trend in KMAs, a possible explanation for the retention of organisation even in hexane could be intra-molecular hydrogen bonding. Further studies are needed into the behaviour of MAs of all classes in hexane, ideally in an explicit solvent model, as they offer the closest approximation to the environment that the MAs are finding themselves in within the mycobacterium cell wall leaflet, packed closely together and intercalating with extractable lipids.

In AMA concatenated trajectories, an interesting trend appeared across the different temperatures studied in explicit water, in both computing approaches. As can be seen in Table 17 in the Results, the percentage of the knot fold reduced from 27% (300 K), to 24% (350 K), to 0% (400 K). As the temperature increased, chains untangled and assumed simpler structures that

did not involve such extensive twisting and bending. Namely, and as can be seen in Table 17 in the Results, the four-chain percentage increased from 58% (300 K), to 70% (350 K), and then reduced slightly to 62% (400 K) where the five-chain presence increased to 37%. More specifically, molecule AC15 demonstrated an increased level of organisation with increasing temperature, as can be seen in Table 16 in the Results. Its four-chain percentage was at 43% at 300 K in explicit water, and then increased to 60%, 90%, and 96%, at 300 K in implicit hexane, and 350 K and 400 K in explicit water, respectively.

Molecule AC15, therefore, demonstrates a high degree of thermostability. In proteins, thermostability seems to be promoted by the compactness of the protein structure.[169] Thermophilic sequences exhibit a natural strategy for enhancing protein thermostability by lowering the entropy of unfolding through truncations of exposed loop regions. Thermostability in proteins has further been shown to be promoted by extra salt bridges and main-chain hydrogen bonds.[170-171] AMAs are the most abundant MA class in many mycobacterial species.[39] At the same time, it has been well established that MAs are crucial for cell wall integrity.[79, 81-82] The lipid nature of MAs and their tight packing inside the cell wall has been shown to promote low fluidity, which in turn contributes to antibiotics resistance.[38] If temperature increase could be promoting increased levels of organisation in MAs, such as the behaviour exhibited in molecule AC15, it is not a behaviour that should be overlooked as it may provide unique potential for understanding the mechanisms for cell wall impermeability and antibiotics resistance.

A similar trend of increased level of organisation, with increasing percentages of four-chain folds as the temperature increases, can be seen for molecule AC6 in Table 12 in the Results. The percentage of four-chain folds at 300 K in explicit water is at 58%, which increases to 98% at 350 K in explicit water. While regrettably there are no data available for the conformational behaviour of the molecule at 400 K, a 40% increase of the four-chain fold percentage from 300 K to 350 K can be observed. Additionally, the four-chain fold is the major cluster (78%) for AC6 in implicit hexane.

The chain lengths of the aforementioned AMA molecules, namely AC6 and AC15 are, respectively, 23-11-14-17 and 23-13-14-17 (*k-l-m-n* notation as can be seen in Figure 1 in the Introduction), in both of which cases the relationship between two of the individual methylene segments is *l*<*m*. Previously, it had been hypothesised[75] that a relationship of *l*<*m* in KMAs promoted the formation and retaining of low-energy 'W'-folds. Thus, it is possible that *l*<*m* promotes the four-chain folds in the case of molecules AC6 and AC15, as observed in this work. However, monolayer studies[45] have demonstrated that AMAs tended to unfold from the 'W' shape in high temperature and surface pressure conditions. Unfortunately, the available literature does not offer enough information to enable the interpretation of the stabilisation increase with temperature in molecules AC6 and AC15. Further experimental and computational studies are needed to identify if the above is a definite trend by investigating if it is observed in all AMAs with *l*<*m*. At the same time, further monolayer experiments may shed light into the behaviour of AMAs in high temperature conditions. The fact that AMAs are the most abundant MA class in many mycobacterial species[48] may imply that this trait could be of importance to the structure-function relationships in the cell wall.

A conformational trend associated with elongation of the *l* chain was unpicked in this study during the analyses of the single MD trajectories. In AMAs, as the *l* chain elongated, the percentage of four and five-chain folds decreased (Table 7 in Results). At the same time, the presence of knot folds increased across the studied molecules. The exception in this pattern was molecule AC33, which additionally bore the longest *l* chain. It demonstrated a 4% prevalence of the four-chain folds over knot folds, while maintaining an 18% five-chain presence, which was the highest in this class. The outlier in this pattern could be due to the fact that in AC33 the *l* and the *n* segments were of equal length, thus potentially prompting the molecule to increase its four- and five-chain presence, while reducing the percentage of knot folds. In a study performed by Villeneuve *et al.*,[75] the effect of the methylene chain lengths on MA folding was explicitly assessed, focusing on oxygenated MAs. The study concluded that depending on the relationship between the *l*, *m*, and *n* segments, the MA

molecule could either retain the 'W' fold or could deviate from it faster. The individual methylene chain lengths therefore have significant potential to direct MA conformations, which can be revealed by the wider scope initiated by the grid-based study.

In KMAs, the results stemming from the individual starting conformation analyses and those from the concatenated trajectory analyses were substantially aligned, as opposed to what was observed in AMAs. The individual conformations returned an overall 57% prevalence of the four-chain structure, a 28% for the knot folds, and a 13% for the five-chain folds. For comparison, the same PCA clustering analyses over the KMAs concatenated trajectories demonstrated a 56% prevalence of four-chain folds, a 41% of knot folds, and a 3% of five-chain folds. Therefore, the major cluster was consistently the four-chain fold across both analyses, with the knot fold and the five-chain being the minor clusters.

The separate analysis of the five individual starting conformations of KMAs, as can be seen in Table 6 in the Results, indicated that in all occurrences within the KMA class, with the exception of molecule KC20, the percentage of four-chain folds decreased as the *l* chain elongated. This echoes what was seen for the behaviour demonstrated by AMAs. At the same time, the percentage of knot and five-chain folds increased as the *l* chain elongated. Molecule KC20, with *l*= 15, behaved differently, with the four-chain folds increasing by 5%. The analysis from the concatenated KMA trajectories returned a similar result, with the percentages of the four-chain folds decreasing as the *l* chain elongated. The outlier in this pattern was molecule KC29, with *l*=19, where the four-chain folds increased. The consistency of the chain length effect on the KMAs folding, with four-chain folds decreasing and knot folds increasing as *l* elongates, albeit with the presence of outliers, sets the framework for future work to further investigate the effects of chain length on the conformational preferences of MAs.

At 300 K, in all examined solvents, the four-chain folds were the major clusters for the dominant majority of the KMA molecules studied. The only exceptions were observed in the concatenated trajectories of molecules KC20 and KC27, were the major clusters were the knot folds. At 350 K, molecule KC25 demonstrated a dominant four-chain fold. At 400 K, the four-chain percentage was only at 8%, while the knot fold was the major cluster at 71%. Similarly, molecule KC28 demonstrated a 76% dominance of five-chain folds at 350 K, which at 400 K changed to 84% knot presence, with only 16% five-chain. Therefore, a trend that is revealed within KMAs is that globularity increased with temperature. This contrasts to what was observed in AMAs where the gradient followed the opposite direction. A possible explanation for the different globularity gradients could be the different distal functional groups in each different MA class. In KMAs, the distal group is a carbonyl group, whereas in AMAs it is a cyclopropane. The KMA carbonyl is an excellent candidate to promote conformational stabilisation through hydrophilic interaction with the water solvent. In contrast, in AMAs the distal cyclopropane is not hydrophilic, which could in turn imply a higher degree of flexibility for these molecules.

Although the observed trend in the increase of KMA globularity is extrapolated only from two KMA structures, it is worth taking it into consideration for the following reasons. In contrast to what is observed in KMAs, the relevant conformations of AMAs exhibited the opposite gradient, which was towards a reduction in globularity, with increasing temperature. Additionally, KMAs have shown potential for a distinct biological role, as it has been elaborated in the Introduction. Yuan *et al.*[83] showed that KMA-depleted BCG and *M. tb* strains demonstrated extremely reduced ability to grow within macrophages, and Dkhar *et al.*[84] established that KMAs contribute to granuloma formation. Free MAs growing in mycobacterial biofilms[87] are more likely to assume globular conformations,[101] and biofilms have been shown[87] to harbour a drug-tolerant *M. tb* population that can persist despite exposure to antibiotics. Crucially, Sambandan *et al.*[88] demonstrated that the ability of *M. tb* to synthesise KMAs is essential for biofilm growth, and growth within this community was

shown to confer drug resistance to the bacilli. It is therefore important to confirm with further experiments in the future if the trend of the increased globularity carries across more chain lengths within this MA class. As the KMA globularity gradient may be relevant to their function and may be echoing a distinct biological role for KMAs, it is certainly worth pursuing with further experiments in the future.

pyPCAzip demonstrated different consistency levels across the two computational approaches employed. On the one hand, during the local approach (300 K, explicit water), pyPCAzip constantly returned a fluctuating number of clusters for each molecule. This was particularly true for the individual analyses performed on each of the trajectories from every molecule. When faced with single 100 ns trajectories, pyPCAzip returned vastly diverse numbers of clusters, even within the same molecule, as can be seen for example in Table S4 or Table S5 in Appendix A. However, it is worth noting that pyPCAzip[148] returned fairly consistent numbers of clusters for each molecule, when the individual starting conformation trajectories were concatenated into complete 500 ns trajectories (5 starting conformations x 100 ns each). The relevant clustering profiles can be seen in Table S11-Table S19 in Appendix A. In that case, five out of nine trajectories returned three clusters, three out of nine returned four clusters, and only one out of nine returned five clusters.

Further, pyPCAzip demonstrated a high level of consistency when the grid-computing results were analysed. The numbers of generated clusters remained completely unchanged across different temperatures and solvation systems, as can be seen in Table 18 in the Results. The concatenation of trajectories and the subsequent analysis of the concatenated data with pyPCAzip proved to be more efficient for the clustering algorithm employed, as the conformational space was sampled more effectively. More specifically, in the comparison between the concatenated stepwise-produced multiple simulations (increments: 100 ps for explicit, 5 ns for implicit - grid approach), versus concatenated 100 ns long MD trajectories (local approach), it seems that the incremental build-up of the simulations is a more efficient way to sample the

conformational space. The clustering tool identified a wide range of structural variability, showcasing a variety of structural clusters, while the pyPCAzip calculations returned consistent numbers of clusters for each case.

In conclusion, pyPCAzip proved an efficient clustering tool, identifying a wide range of structural diversity in the examined trajectories, and achieving generally high consistency in the generated results. One disadvantage was the difficulty in automatically discriminating between different clusters, but the carbon atom distance matrices provided insights and good visual representations of structural differences.

Chapter 5: Concluding Discussion

Summary of conclusions

A systematic and intensive computational study on the conformational behaviour of MAs has been conducted for this doctoral work, utlising atomistic MD simulations. The work that is presented in this thesis adds to the state-of-the-art computational research on MA conformations mainly by (i) focusing on a much wider MA dataset than previous studies have done, (ii) overcoming some of the critical limitations identified in previous works, (iii) employing a combination of computational workflow approaches to provide an extensible framework for exploring MA and related lipid structure, and (iv) using a novel analysis protocol. This is the first attempt to perform MA conformational studies on a crowd-sourced distributed computing platform. The outcomes and established workflows of this work will provide the foundation for performing further and more complicated MA and *M. tb* cell wall simulations in the future.

Overall, the analytical approaches used provided valuable insights, with the exception of the dihedral clustering tool used, DASH,[147] which was unfortunately not suitable to address the research questions this work focused on. In comparison, the PCA clustering process used, pyPCAzip,[148] returned meaningful results, and it was computationally cheap to run and store. Finally, the calculation of intramolecular distance matrices provided valuable insights into refolding events and an effective visual identification between different clusters, while however being expensive to run and store.

The MAs studied and analysed in this work have confirmed their tendency to assume a variety of conformations, as this has been shown in previous works, and have demonstrated that open structures are rare and limited to the early steps of the trajectories. More specifically, the concatenated trajectories of AMAs were shown to prefer tight four-chain folds in all of the different environments examined, which is in slight disagreement with conclusions from

previous MA computational[47, 93, 101] and monolayer[45] studies, where they were demonstrated to be more flexible. However, this disagreement likely stems from the fact that the current work has performed significantly longer simulations and has employed a more comprehensive and less biased analytical protocol. Further in AMAs, it was shown that as the *l* chain elongated, the percentage of four and five-chain folds decreased, while the presence of knot folds increased across the studied molecules. The exception in this pattern was molecule AC33, which bore the longest *l* chain, with *l*=*n*.

In KMAs, the results demonstrated that the major cluster was consistently the four-chain fold across all analyses, with the knot fold and the five-chain being the minor clusters. At the same time, a similar pattern to that observed in AMAs was also seen in KMAs. More specifically, the percentage of knot folds increased as the *l* chain elongated, while the four-chain folds decreased. Despite the presence of outliers, this observation sets the framework for future work to further investigate the chain length effect on MA conformations.

The effects of temperature on the folding were interesting for both chemical classes of MAs studied. For AMAs, as the temperature increased from 300 K to 400 K, the molecules assumed organised four-chain folds. In contrast, the trend observed in KMAs was that globularity increased with temperature, with the knot folds being the major cluster.

In terms of the effects of different solvents on the folding, implicit water demonstrated an impressive 100% dominance of the four-chain folds in all molecules studied. All implicit water results consisted of two four-chain clusters. Implicit hexane had a balanced representation of four-chain, five-chain, and knot-folds, with four-chain folds being the major cluster. In explicit water, the overall major cluster in all temperatures was the four-chain fold, with knot folds being the second most occupied cluster.

Outlook and future work

Distance matrix plots provide an efficient and direct visual representation of how different the various MA folding patterns are. By looking at the average distance matrix plots that have been included in this thesis (in the Results and Appendices) one can directly discern the structure of each different cluster the matrix is associated with, without needing to look up the structure. This is because each point on distance matrix plots corresponds to a measurable distance between two backbone carbon atoms. At the same time, PCA plots are not as capable to directly demonstrate these structural differences visually. Plots of projections of PCs offer information about the sampling and how the molecule maps the conformational space. PCA analyses on the distance matrices may therefore be useful to get the combined advantages of both analytical methods. This type of combined analysis may also be extremely useful for the terabytes of data generated on the grid-computing servers of the WCG. Robust automated analytical workflows should be employed that will be able to handle the size of the datasets and analyse them efficiently.

The effect of a wide array of temperatures (in 25 K increments from 300 K to 400 K) on MA folding has been investigated in this work. Recent temperature replica exchange simulations[172] on a small globular protein, protein G B1 domain (GB1), have investigated the conformational behaviour of the molecule in temperatures in 20 K increments from 240 K to 400 K. The protein started unfolding at around 320 K, while at 300 K it maintained its globular structure. It is therefore clear that even subtle temperature variations can lead to dramatically different conformations being assumed by the molecules in question. In the current work, atomistic MD data generated from 300 K, 350 K, and 400 K has been analysed and presented in earlier sections. Temperature has been found to play a major role in the conformational preferences of MAs, as has been shown in monolayer experiments,[45, 75, 93-94] with AMAs unfolding at higher temperature and surface pressure conditions and KMAs retaining their low-energy 'W' shapes. Temperature is therefore expected to elicit a conformational modulation on MAs that is anticipated to be different for each MA class. Ongoing simulations from this work are generating relevant data on

the IBM grid-computing platforms. For future work, the results from all studied temperatures should be scrutinised at a greater detail, in order to further elucidate the effect of temperature on the folding. The data from MMAs, which is an MA class for which data has not been presented in this thesis, and from different cyclopropane stereochemistries of AMAs and KMAs, would significantly enrich the current understanding of MA folding behaviour, and it is therefore important that the relevant generated data are investigated.

Chain lengths were shown to play a role in MA conformational preferences, with outliers in observed patterns being molecules AC33, KC20, and KC29. Particularly in regard to AMAs where a conformational trend (elongation of the l chain decreased the percentage of four and five-chain folding, and increased that of knot folding) was revealed, the relationship between the l and n segments was $l \leq n$. It has been previously highlighted that the individual methylene segment chain lengths can modulate the folding in different MAs.[75, 93] However, these studies did not manage to explicitly identify the limiting factor that steers this modulation in different MAs. It becomes clear that the inherent fine structural differences in MAs, in combination with the different chain lengths, contribute to varying trends in the conformational profiles of each molecule. This showcases that there is still a long way to go in order to demystify the chain length contribution to the MA folding preferences. The abundance of data generated during the current work on the IBM grid-computing servers will provide extremely fertile ground for further analysis that may ultimately be able to answer exactly how chain lengths modulate the folding.

Such detailed analyses can directly lead to very efficient mapping of the MA conformational space and to the definitive identification of their major classes and subclasses. Combined with the understanding of how inherent factors, such as the stereochemistry of the different MA functionalities, steer these conformations, such detailed studies would generate pivotal information about how to deal with the tuberculosis crisis. As is also highlighted further in this section, MAs have shown promising qualities as potential serodiagnostic tools

for TB,[73-74] while different MAs demonstrate different antigenicity in TB patients[90] and cyclopropanation loss impairs the ability of the bacterium to establish a lethal infection.[91] MAs have also been shown to enhance the drug load in INH-loaded nanoparticles.[76] It is therefore clear that the relationship between structure and MA conformation is expected to be of biological significance in the event of TB infection. Creating a detailed MA cloud-based "map" or database, similar to proteomics and genomics databases, could additionally provide a gateway for information access and exchange across areas of antitubercular research, thus further supporting the efforts in novel antitubercular drug design and TB diagnosis.

Apart from the different chain lengths, cyclopropanation and stereochemisty of MAs have both been shown to play a crucial role in the antigenicity and survival[104, 173] of the bacterium. Specific immune activities have been attributed to different MA classes, and oxygenation and *cis* versus *trans*-methyl cyclopropanation have been shown to steer such activities.[89-90] The major components of natural MAs from *M. tb* contain a *cis*-cyclopropane for MMA and an α-methyl *trans*-cyclopropane for KMA, as can be seen in Figure 41 below.

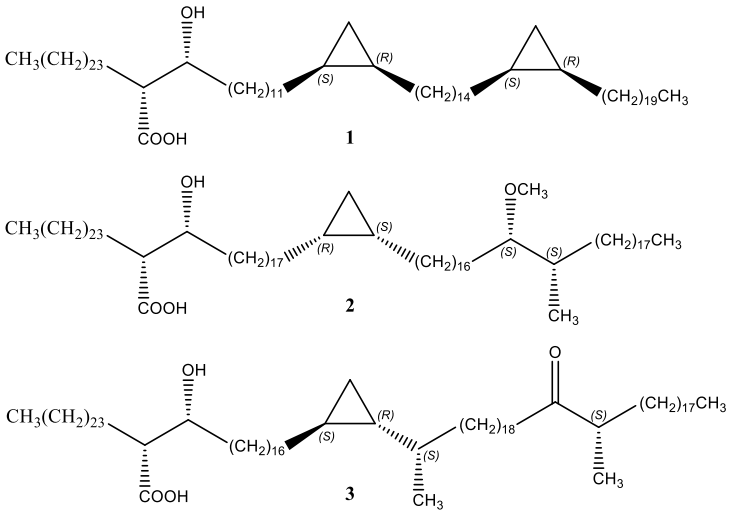


Figure 41. Chemical structures of the main components of MAs from *M. tb*. Adapted from Groenewald *et al.*[47] MA class is assigned according to the nature of the distal functional group, namely *cis*-cyclopropane (**1**, AMA), methoxy-methyl (**2**, MMA), and keto (**3**, KMA).

Model compounds of short alkyl chains, containing *cis*- and α-methyl trans-cyclopropane respectively, have been studied to assess the bend-forming capacities of the different stereochemistry.[94] The *trans* variant was shown to facilitate tighter folds. These results echoed findings derived from monolayer studies. Further, KMAs in atomistic MD studies in different solvents had the highest percentage of 'WUZ'-folds and mainly folded at the α-methyl *trans*-cyclopropyl group.[101] The current work has identified both the proximal and distal cyclopropanes in AC6, and the proximal cyclopropane in KC4, as distinct folding points through dihedral angle and PCA clustering analyses. It is therefore apparent that cyclopropanation and cyclopropane stereochemistry are excellent targets for further research, since they appear to play a crucial role in modulating the bacterial biology and the MA folding. The atomistic MD data that are being generated on the IBM grid-computing systems will provide an excellent framework for those molecular idiosyncrasies to be analysed and unpicked.

Two of the major conformers defined, the four-chain and the knot fold, exhibited readily recognisable and definable subtypes. The four-chain subtype was a tidy four-chain variant, with the four chains straightening out and packing closely together, as opposed to the non-tidy variant, where the chains were much more relaxed. For knots, there were four variants: (i) knot - end, where the end of the chain would emerge through the loop, (ii) knot - start, where the start of the chain would emerge through the loop, (iii) knot - bow, where both ends would intertwine and emerge through the formed loop, and (iv) relaxed knot. A substantially more detailed understanding of MA conformations can be derived by further defining these subtypes. MAs have the potential to be used as antigens for TB serodiagnosis,[68, 73] even in countries with a high HIV burden such as Africa.[73-74] The sensitivity and specificity of serodiagnosis depends on the detailed structure of the antigen and on the antibody used.[174] Since the mechanism for antigen-antibody recognition relies on structural affinity and specificity, the identification of MA subclasses and the recognition of the major structures within these subclasses may contribute to the development of more efficient serodiagnostic methods.

In terms of future outlook, simulations of interacting rather than single MAs would provide even more insightful information about the behaviour of MAs in the mycobacterial cell envelope. MAs inside the cell wall are packed side by side parallel to each other and perpendicular to the cell envelope plane,[60] while at the same time they interact with extractable lipids.[56] It has been shown through numerous studies that different MAs contribute differently to the cell function and cell wall permeability. Namely, oxygenated MAs are crucial for virulence in mice and for cell wall integrity.[46] MA cyclopropanation is essential for cell wall structure and resistance to antibiotics.[173] Finally, the recognition of MAs as promising serodiagnostic tools[73-74] emphasises the importance of detailed knowledge of the exact structure of major MA conformations. However, the MA structure-function relationships have still not been fully demystified.

While modelling single MAs delivers a wealth of information on the preferred conformations they may assume inside the cell wall envelope, it is not a realistic representation of their environment inside the bacterium. Utilising information from single MA simulations, such as the extensive data generated from this work, to build the necessary computational models, coarse-grained and atomistic, in order to study MA interaction, would be a very good approximation of the natural environment the MAs are finding themselves in within the mycobacterium. Such studies would provide crucial knowledge on the molecular interplay between the cell wall components, and may contribute towards better understanding of how the cell wall confers impermeability and antibiotics resistance.

Further, the development of novel computational cell wall models of the mycobacterium could deliver pivotal new information in antibiotics permeation inside the *M. tb* cell wall. Mycobacterial cell walls are crucial for the microorganism, as they promote very low permeability and therefore directly contribute to drug resistance.[32] MAs in *Corynebacterineae* were shown to play a major role in the functionality of the cell wall permeability barrier.[42] The mechanism of solute permeation through membranes is challenging to

study experimentally, because of the small scale and complexity of lipid bilayers. A novel dual-resolution MD method has been used to simulate the permeability of drugs and steroid hormones inside a lipid membrane.[175] The lipid and water molecules were represented by coarse-grain models, whereas the permeants were represented by atomistic models. Such an approach could be also applied towards the study of antibiotics permeation inside the mycobacterial cell wall.

Coarse-graining approaches, typically used for membrane work,[176] can represent the complexity of the modelled systems precisely and efficiently, with a smaller computational cost than atomistic MD simulations. The cumulative data of atomistic MD simulations of MAs in different temperatures and environments, that have been generated from this work, can be integrated to complete a meta-analysis in order to generate the parameters necessary to build much larger and computationally challenging coarse-grained cell wall models. Contrary to well-established cell membrane models, cell wall models are currently not accessible due to the lack of detailed data about the nature and complexity of the cell wall components, especially in regard to their explicit stereochemistry. A coarse-grained simulation of the complete cell wall could therefore provide insights into the mechanisms that govern antibiotics permeation into the cell wall, as well as cell wall integrity and structure. Such a development could lead to a breakthrough in antitubercular drug development.

A comprehensive conformational study on MAs has been presented. Atomistic MD simulations over a variety of environments and on a vast array of MAs have been employed. The data that have been generated from this work enrich the current understanding on MA conformations and can be utilised to deliver a complete model of the mycobacterial cell wall. This can put current TB research efforts in an excellent position to understand antibiotics permeation and resistance. The current work therefore offers tremendous potential not only to be further expanded, but most importantly to be employed in the forefront of antitubercular research.

Appendix A

Alpha MAs - ***c****is*	
23-07-14-19	AC1
23-09-14-19	AC2
23-11-10-19	AC3
23-11-12-19	AC4
23-11-14-15	AC5
23-11-14-17	AC6
23-11-14-19	AC7
23-11-14-21	AC8
23-11-14-23	AC9
23-11-16-19	AC10
23-11-18-19	AC11
23-13-10-19	AC12
23-13-12-19	AC13
23-13-14-15	AC14
23-13-14-17	AC15
23-13-14-19	AC16
23-13-14-21	AC17
23-13-14-23	AC18
23-13-16-19	AC19
23-13-18-19	AC20
23-15-10-19	AC21
23-15-12-19	AC22
23-15-14-15	AC23
23-15-14-17	AC24
23-15-14-19	AC25
23-15-14-21	AC26
23-15-14-23	AC27
23-15-16-19	AC28
23-15-18-19	AC29
23-17-10-19	AC30
23-17-12-19	AC31
23-17-14-15	AC32
23-17-14-17	AC33
23-17-14-19	AC34
23-17-14-21	AC35
23-17-14-23	AC36
23-17-16-19	AC37
23-17-18-19	AC38
23-19-14-19	AC39
23-21-14-19	AC40

Alpha MAs - ***t****rans*	
21-14-12-19	AT1
21-14-14-17	AT2
21-16-12-19	AT3
21-16-14-17	AT4
21-18-08-19	AT5
21-18-10-17	AT6
21-18-10-19	AT7
21-18-12-15	AT8
21-18-12-17	AT9
21-18-12-19	AT10
21-18-12-21	AT11
21-18-12-23	AT12
21-18-14-13	AT13
21-18-14-15	AT14
21-18-14-17	AT15
21-18-14-19	AT16
21-18-14-21	AT17
21-18-16-17	AT18
21-18-16-19	AT19
21-18-18-17	AT20
21-20-12-19	AT21
21-20-14-17	AT22
21-22-12-19	AT23
21-22-14-17	AT24

Methoxy MAs - ***c****is*	
23-15-16-17	MC1
23-17-12-17	MC2
23-17-14-17	MC3
23-17-16-13	MC4
23-17-16-15	MC5
23-17-16-17	MC6
23-17-16-19	MC7
23-17-16-21	MC8
23-17-18-17	MC9
23-17-20-17	MC10
23-19-12-17	MC11
23-19-14-17	MC12
23-19-16-13	MC13
23-19-16-15	MC14
23-19-16-17	MC15
23-19-16-19	MC16
23-19-16-21	MC17
23-19-18-17	MC18
23-19-20-17	MC19
23-21-12-17	MC20
23-21-14-17	MC21
23-21-16-13	MC22
23-21-16-15	MC23
23-21-16-17	MC24
23-21-16-19	MC25
23-21-16-21	MC26
23-21-18-17	MC27
23-21-20-17	MC28
23-23-16-17	MC29

table continues in next page

Methoxy MAs - ***t***rans	
23-14-16-17	MT1
23-16-16-17	MT2
23-18-12-17	MT3
23-18-14-17	MT4
23-18-16-13	MT5
23-18-16-15	MT6
23-18-16-17	MT7
23-18-16-19	MT8
23-18-16-21	MT9
23-18-18-17	MT10
23-18-20-17	MT11
23-20-12-17	MT12
23-20-14-17	MT13
23-20-16-13	MT14
23-20-16-15	MT15
23-20-16-17	MT16
23-20-16-19	MT17
23-20-16-21	MT18
23-20-18-17	MT19
23-20-20-17	MT20
23-22-16-17	MT21
23-24-16-17	MT22

Keto MAs - *cis*	
23-09-16-19	KC1
23-11-16-17	KC2
23-11-16-19	KC3
23-11-18-17	KC4
23-13-12-19	KC5
23-13-14-19	KC6
23-13-16-15	KC7
23-13-16-17	KC8
23-13-16-19	KC9
23-13-16-21	KC10
23-13-16-23	KC11
23-13-18-17	KC12
23-13-18-19	KC13
23-13-20-19	KC14
23-15-14-17	KC15
23-15-16-17	KC16
23-15-16-19	KC17
23-15-18-13	KC18
23-15-18-15	KC19
23-15-18-17	KC20
23-15-18-19	KC21
23-15-18-21	KC22
23-15-20-17	KC23
23-15-22-17	KC24
23-17-16-17	KC25
23-17-16-19	KC26
23-17-18-17	KC27
23-19-16-17	KC28
23-19-18-17	KC29

Keto MAs - ***t***rans	
23-10-18-17	KT1
23-12-18-17	KT2
23-14-14-15	KT3
23-14-16-15	KT4
23-14-18-13	KT5
23-14-18-15	KT6
23-14-18-17	KT7
23-14-18-19	KT8
23-14-18-21	KT9
23-14-20-17	KT10
23-14-22-17	KT11
23-16-14-17	KT12
23-16-16-17	KT13
23-16-18-13	KT14
23-16-18-15	KT15
23-16-18-17	KT16
23-16-18-19	KT17
23-16-18-21	KT18
23-16-20-17	KT19
23-16-22-17	KT20
23-18-18-17	KT21
23-20-18-17	KT22

Figure S1. Complete list of MA molecules simulated in this work. Tables include *k-l-m-n* chain length (left) and naming convention (right) used to describe each different MA.

molecule KC4	#1	#2	#3	#4	#5
DASH states / transitions	7 / 7	4 / 3	3 / 5	2 / 2	2 / 1
Representative frames	8,684 8,684 8,620 20,172 15,522 8,684 937	20,509 20,708 13,315 20,708	11,968 4,356 4,356	8,601 87,581	2,595 2,595
Timesteps spent in DASH state (DS)	DS 6: 744 DS 7: 246 DS 6: 8,005 DS 5: 11,061 DS 4: 997 DS 1: 31,741 DS 2: 19,043 DS 3: 28,164	DS 1: 56,877 DS 2: 13,083 DS 3: 11,504 DS 4: 18,537	DS 1: 38,248 DS 3: 5,436 DS 2: 3,050 DS 3: 37,752 DS 2: 4,315 DS 3: 11,200	DS 2: 6,759 DS 1: 2,817 DS 2: 90,425	DS 1: 62,541 DS 2: 37,460

Table S1. DASH analysis results for each of the five trajectories (#1-#5) of molecule KC4.

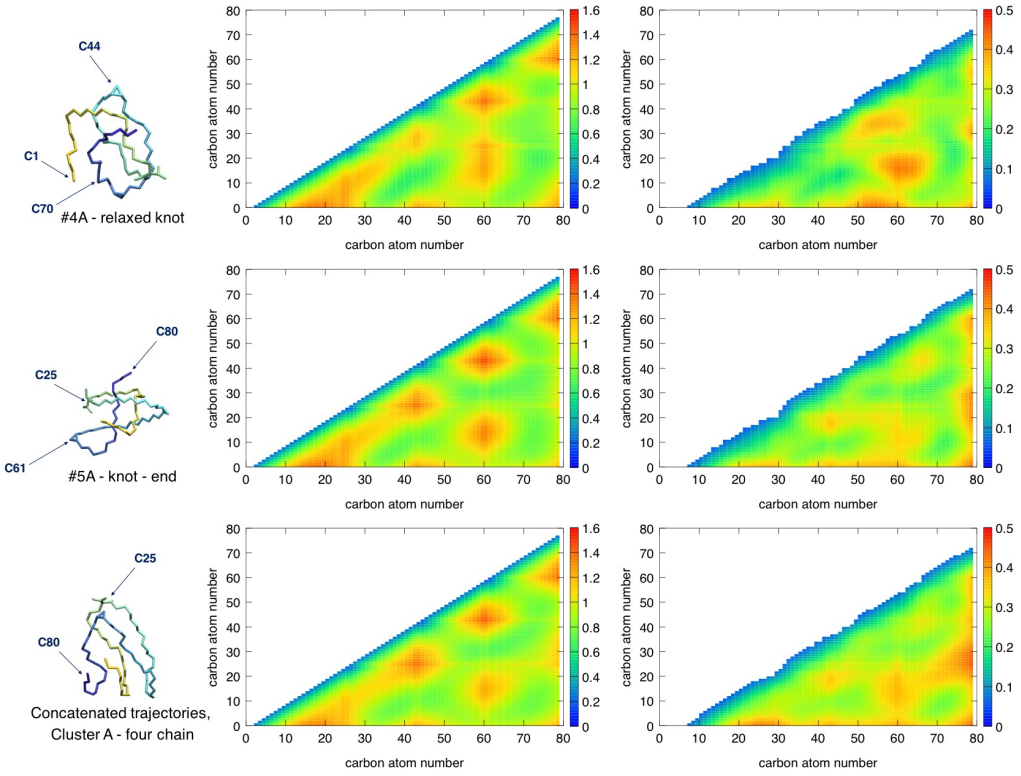


Figure S2. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #4A from molecule AC24.

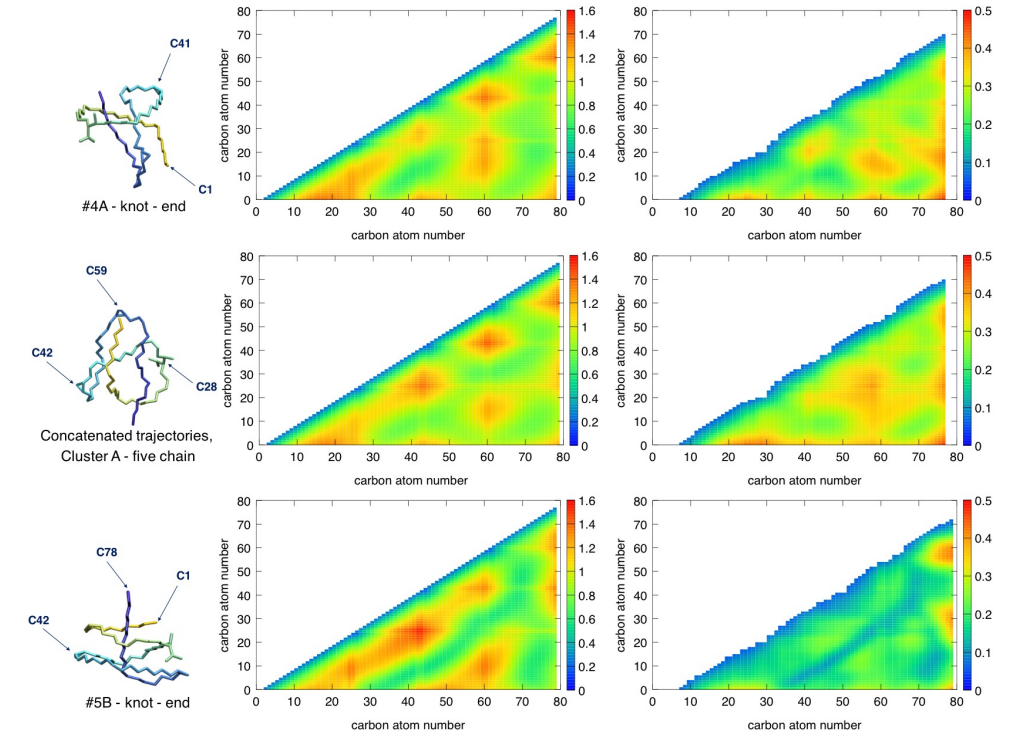


Figure S3. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #4A from molecule AC15.

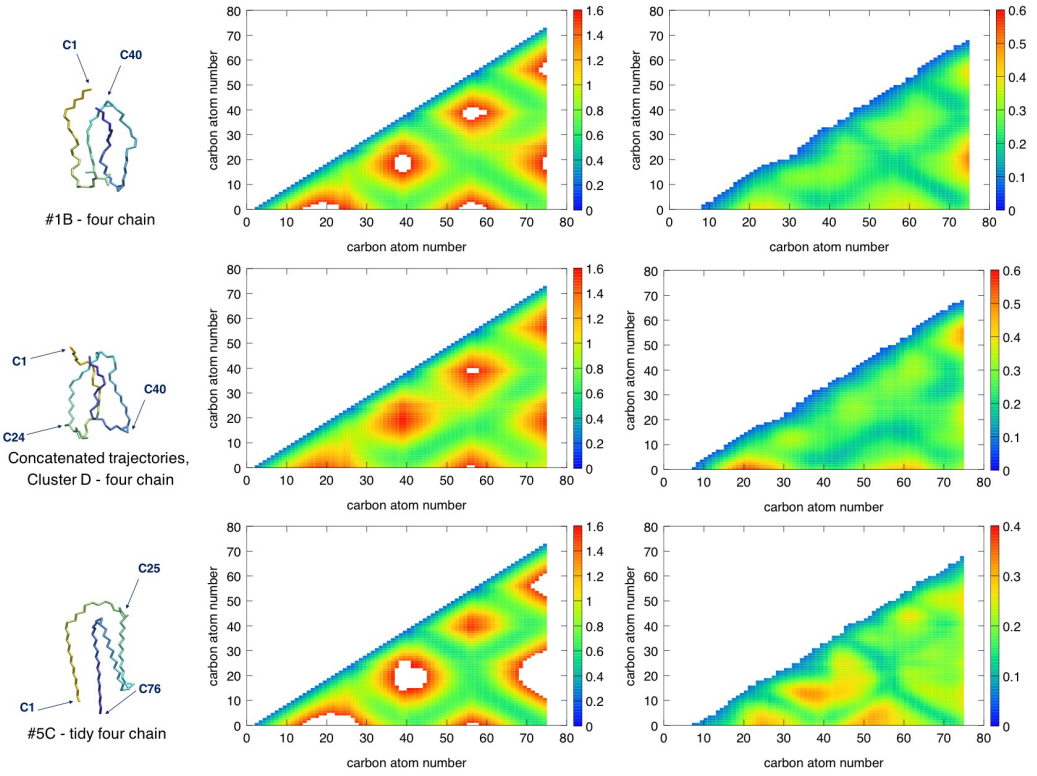


Figure S4. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1B from molecule AC6.

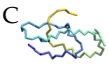
AC6		average structure	shape	cluster percentage	RMSD of average structure	average RMSD of distance matrices	most similar with cluster (RMSD value)
Snapshot #1	A		knot - end	35.9	0.4541	0.1153	Concatenated trajectories Cluster B (0.0087) #4A (0.0092)
	B		four chain	29.4	0.3793	0.1020	Concatenated trajectories Cluster D (0.0165) #5C (0.0313)
	C		knot - bow	17.4	0.2649	0.0761	
	D		knot - bow	12.3	0.2750	0.0806	
	E		tidy four chain	5.0	0.2158	0.0662	
Snapshot #2	A		four chain	60.0	0.4345	0.1154	
	B		relaxed knot	29.6	0.3654	0.0959	
	C		knot - end	10.4	0.2143	0.0711	
Snapshot #3	A		knot - start	79.5	0.3936	0.1086	
	B		four chain	16.0	0.3763	0.1064	
	C		knot - start	4.5	0.2195	0.0672	
Snapshot #4	A		four chain	93.1	0.4676	0.1152	
	B		knot - end	6.8	0.2803	0.0862	
Snapshot #5	A		four chain	65.2	0.4590	0.1179	
	B		five chain	30.1	0.4384	0.1225	
	C		tidy four chain	4.6	0.2983	0.0803	

Table S2. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule AC6.

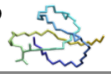
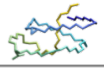
AC15		average structure	shape	cluster % percentage	RMSD of average structure	average RMSD of distance matrices	most similar with clus (RMSD value)
Snapshot #1	A		knot - start	30.3	0.4378	0.1131	
	B		four chain	26.7	0.3972	0.1171	
	C		relaxed knot	13.0	0.4616	0.1228	
	D		tidy four chain	9.5	0.1950	0.0678	
	E		relaxed knot	8.4	0.2665	0.0819	
	F		five chain	6.3	0.3563	0.1034	
	G		three chain	5.8	0.3391	0.1048	
Snapshot #2	A		four chain	30.5	0.4247	0.1126	#3A (0.0129) Concatenated trajectories Cluster B (0.0133)
	B		knot - end	27.9	0.4342	0.1135	
	C		knot - end	22.1	0.3698	0.1060	
	D		relaxed knot	15.5	0.3355	0.0911	
	E		knot - start	3.9	0.2501	0.0723	

table continues in next page

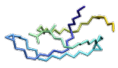
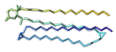
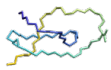
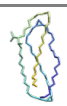
	**four chain**	**71.7**	**0.4806**	**0.1261**	
	knot - end	23.8	0.4327	0.1115	
	tidy four chain	4.5	0.2055	0.0691	
	knot - end	63.6	0.4465	0.1087	Concatenated trajectories Cluster A (0.0086) #5B (0.0152)
	four chain	19.2	0.4155	0.1144	
	relaxed knot	9.0	0.2730	0.0789	
	knot - start	8.1	0.3962	0.1046	
	tidy four chain	47.1	0.4540	0.1163	
	knot - end	45.1	0.4607	0.1228	
	knot - end	7.7	0.2187	0.0701	
	two chain	0.05	0.3440	0.0923	

Table S3. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule AC15.

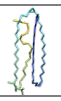
AC24		average structure	shape	cluster % percentage	RMSD of average structure	average RMSD of distance matrices	most similar with clus (RMSD value)
Snapshot #1	A		four chain	57.3	0.4908	0.1238	
	B		knot - end	38.2	0.5298	0.1057	
	C		knot - start	4.5	0.2267	0.0733	
Snapshot #2	A		four chain	39.9	0.4168	0.1207	
	B		knot - start	31.7	0.4505	0.1121	
	C		knot - end	18.0	0.4244	0.1137	
	D		four chain	5.4	0.3177	0.0943	
	E		three chain	4.9	0.5638	0.1870	
Snapshot #3	A		four chain	62.2	0.4706	0.1224	#1A (0.0081) Concatenated trajectories Cluster B (0.0089)
	B		knot - end	31.2	0.4304	0.1067	
	C		tidy four chain	6.6	0.3400	0.0997	
Snapshot #4	A		**relaxed knot**	**30.9**	**0.4701**	**0.1077**	**#5A (0.0126) Concatenated trajectories Cluster A (0.0156)**
	B		knot - end	24.4	0.4357	0.1143	
	C		tidy four chain	20.8	0.3731	0.1122	
	D		knot - start	18.3	0.3046	0.0848	
	E		knot - end	4.8	0.2426	0.0709	
	F		two chain	0.8	0.4153	0.1421	
Snapshot #5	A		knot - end	91.2	0.4317	0.1097	
	B		knot - end	8.7	0.2350	0.0786	

Table S4. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule AC24.

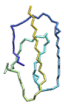
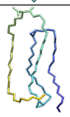
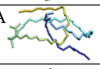
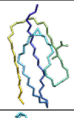
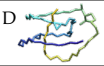
AC33		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with clus (RMSD value)
Snapshot #1	A		four chain	55.2	0.4528	0.1220	#2B (0.0091) #3A (0.0095) Concatenated trajectories Cluster B (0.0095)
	B		relaxed knot	34.9	0.4449	0.1167	
	C		tidy four chain	6.2	0.2180	0.0645	
	D		four chain	3.7	0.4225	0.1303	
Snapshot #2	A		five chain	45.1	0.4725	0.1180	Concatenated trajectories Cluster A (0.0103) #1B (0.0128)
	B		tidy four chain	34.2	0.4509	0.1186	
	C		four chain	16.8	0.4729	0.1210	
	D		relaxed knot	3.9	0.3326	0.0928	
Snapshot #3	A		four chain	96.2	0.4965	0.1247	
	B		four chain	3.8	0.3288	0.0944	
Snapshot #4	A		**knot**	**77.3**	**0.5423**	**0.1222**	
	B		five chain	19.4	0.4079	0.1098	
	C		five chain	3.3	0.3066	0.0929	
Snapshot #5	A		knot - bow	34.0	0.3858	0.1041	
	B		five chain	22.4	0.4398	0.1111	
	C		knot - bow	21.9	0.3486	0.0969	
	D		knot - start	17.1	0.3558	0.0928	
	E		knot - start	4.6	0.1662	0.0535	

Table S5. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule AC33.

KC4		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with clu (RMSD value)
Snapshot #1	A		four chain	76.7	0.4498	0.1099	Concatenated trajecto Cluster D (0.0117) #3A (0.0140)
	B		five chain	13.7	0.4706	0.1087	
	C		tidy four chain	8.5	0.1958	0.0695	
	D		two chain	0.7	0.4216	0.1529	
	E		two chain	0.3	0.3764	0.1297	
Snapshot #2	A		four chain	48.4	0.4316	0.1206	
	B		four chain	40.2	0.5261	0.1222	
	C		tidy four chain	11.2	0.2180	0.0769	
	D		two chain	0.2	0.2143	0.0606	
Snapshot #3	A		tidy four chain	61.0	0.4162	0.1102	
	B		knot - end	23.5	0.3264	0.0906	#5B (0.0124) Concatenated trajecto Cluster C (0.0126)
	C		four chain	8.8	0.5023	0.1175	
	D		four chain	6.7	0.2459	0.0805	
Snapshot #4	A		four chain	43.1	0.4961	0.1282	
	B		four chain	39.8	0.4937	0.1105	
	C		four chain	13.3	0.3974	0.1138	
	D		tidy four chain	3.8	0.2764	0.0872	
Snapshot #5	A		knot - end	84.1	0.4374	0.1128	
	B		knot - end	15.8	0.2195	0.0707	

Table S6. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule KC4.

KC12		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with clu (RMSD value)
Snapshot #1	A		tidy four chain	69.2	0.4249	0.1192	Snapshot #5 Cluster (0.0078) Snapshot #3 cluster (0.0079)
	B		relaxed knot	19.6	0.4456	0.1105	Snapshot #3 Cluster (0.0181) Snapshot #5 cluster (0.0229)
	C		tidy four chain	11.2	0.2609	0.0879	
Snapshot #2	A		tidy four chain	80.8	0.4030	0.1171	
	B		five chain	13.3	0.4315	0.1223	
	C		tidy four chain	5.9	0.2174	0.0771	
Snapshot #3	A		tidy four chain	48.4	0.4024	0.1116	
	B		relaxed knot	34.6	0.4759	0.1108	
	C		relaxed knot	13.8	0.4155	0.1075	
	D		five chain	3.3	0.2270	0.0737	
Snapshot #4	A		knot	50.9	0.3885	0.1098	
	B		four chain	40.0	0.4644	0.1368	
	C		knot - start	9.0	0.2174	0.0690	
Snapshot #5	A		four chain	35.1	0.3850	0.1092	
	B		five chain	21.1	0.4892	0.1141	
	C		five chain	21.0	0.4242	0.1117	
	D		five chain	12.6	0.3939	0.1196	
	E		tidy four chain	10.2	0.2015	0.0707	

Table S7. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule KC12.

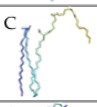
KC20		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with clu (RMSD value)
Snapshot #1	A		knot - end	45.1	0.3472	0.0949	#5B (0.0330) #2B (0.0334)
	B		four chain	38.9	0.4781	0.1222	
	C		tidy four chain	12.1	0.2819	0.0864	#3A (0.0136) #4C (0.0208)
	D		open three chain	2.6	0.5464	0.1972	
	E		three chain	1.3	0.3738	0.1133	
Snapshot #2	A		tidy four chain	52.0	0.4093	0.1219	
	B		knot - start	27.8	0.3147	0.0967	
	C		five chain	12.4	0.4354	0.1231	
	D		four chain	7.8	0.1942	0.0689	
Snapshot #3	A		tidy four chain	80.2	0.3941	0.1050	
	B		four chain	16.0	0.2681	0.0923	
	C		open three chain	3.8	0.4713	0.1562	
Snapshot #4	A		five chain	52.3	0.4976	0.1260	
	B		tidy four chain	32.6	0.4070	0.1310	
	C		tidy four chain	12.6	0.3257	0.0958	
	D		open two chain	2.0	0.4301	0.1237	
	E		open two chain	0.4	0.4359	0.1340	
Snapshot #5	A		tidy four chain	85.0	0.4048	0.1180	
	B		relaxed knot	8.9	0.4018	0.1162	
	C		tidy four chain	6.0	0.1959	0.0701	

Table S8. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule KC20

KC27		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with clu (RMSD value)
Snapshot #1	A		relaxed knot	39.1	0.4527	0.1142	
	B		five chain	30.9	0.4097	0.1089	#5B (0.0091) #4C (0.0227)
	C		four chain	26.8	0.4372	0.1177	
	D		knot - end	3.1	0.4383	0.0993	
Snapshot #2	A		four chain	71.9	0.4144	0.1208	
	B		relaxed knot	12.9	0.4001	0.1146	
	C		four chain	9.6	0.2067	0.0719	
	D		relaxed knot	5.6	0.4758	0.1209	
Snapshot #3	A		four chain	48.9	0.5031	0.1241	
	B		tidy four chain	31.4	0.3965	0.1184	
	C		five chain	14.4	0.4294	0.1251	
	D		four chain	5.3	0.1955	0.0659	
Snapshot #4	A		four chain	44.4	0.4600	0.1236	
	B		knot	26.1	0.4349	0.1162	
	C		knot	23.9	0.4808	0.1216	
	D		tidy four chain	5.5	0.2815	0.0866	
Snapshot #5	A		knot - start	47.7	0.4548	0.1165	
	B		five chain	47.7	0.3914	0.1093	
	C		five chain	4.6	0.2989	0.0876	

Table S9. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule KC27.

KC29		average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices	most similar with cluster (RMSD value)
Snapshot #1	A		four chain	57.5	0.3870	0.1104	#4B (0.0119) #2B (0.0170)
	B		knot - start	22.9	0.3137	0.0877	
	C		five chain	13.6	0.4232	0.1197	
	D		tidy four chain	4.3	0.2537	0.0799	
	E		five chain	1.7	0.2860	0.0800	
Snapshot #2	A		knot - start	57.2	0.4207	0.1112	Concatenated trajectories Cluster C (0.0085) #4A (0.0091)
	B		four chain	21.2	0.3868	0.1143	
	C		knot - end	16.3	0.4104	0.1201	
	D		knot - start	5.2	0.2146	0.0660	
Snapshot #3	A		relaxed knot	76.9	0.4306	0.1093	
	B		knot - end	16.7	0.3620	0.0956	
	C		knot - start	6.4	0.1854	0.0600	
Snapshot #4	A		five chain	58.3	0.3911	0.1110	
	B		tidy four chain	32.0	0.3776	0.1146	
	C		tidy four chain	5.5	0.2091	0.0758	
	D		five chain	4.2	0.2824	0.0873	
Snapshot #5	A		tidy four chain	56.8	0.3352	0.0931	
	B		five chain	17.7	0.3954	0.1126	
	C		tidy four chain	15.0	0.4192	0.1116	
	D		tidy four chain	10.4	0.2409	0.0813	

Table S10. Collated PCA clustering results for each of the five trajectories (#1-#5) of molecule KC29.

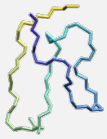
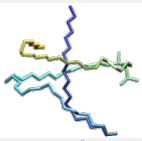
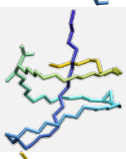
molecule AC6	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		four chain	55.9	0.4979	0.1175
B		knot - end	24.8	0.4144	0.1141
C		knot - end	16.8	0.3695	0.1037
D		four chain	2.49	0.3823	0.1042

Table S11. Collated PCA clustering results from the five concatenated trajectories from molecule AC6.

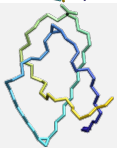
molecule AC15	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		five chain	57.4	0.4852	0.1169
B		tidy four chain	39.3	0.4436	0.1177
C		four chain	3.3	0.4044	0.1059

Table S12. Collated PCA clustering results from the five concatenated trajectories from molecule AC15.

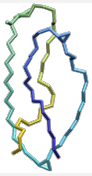
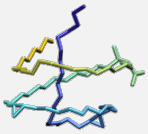
molecule AC24	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		four chain	50.2	0.5260	0.1165
B		four chain	40.6	0.4469	0.1212
C		knot - end	4.8	0.2480	0.0807
D		three chain	4.4	0.4920	0.1404

Table S13. Collated PCA clustering results from the five concatenated trajectories from molecule AC24.

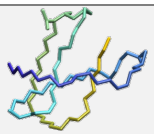
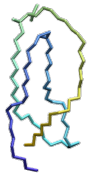
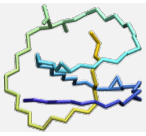
molecule AC33	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		relaxed knot	57.4	0.4828	0.1169
B		four chain	38.7	0.4449	0.1211
C		knot - start	3.9	0.3886	0.1017

Table S14. Collated PCA clustering results from the five concatenated trajectories from molecule AC33.

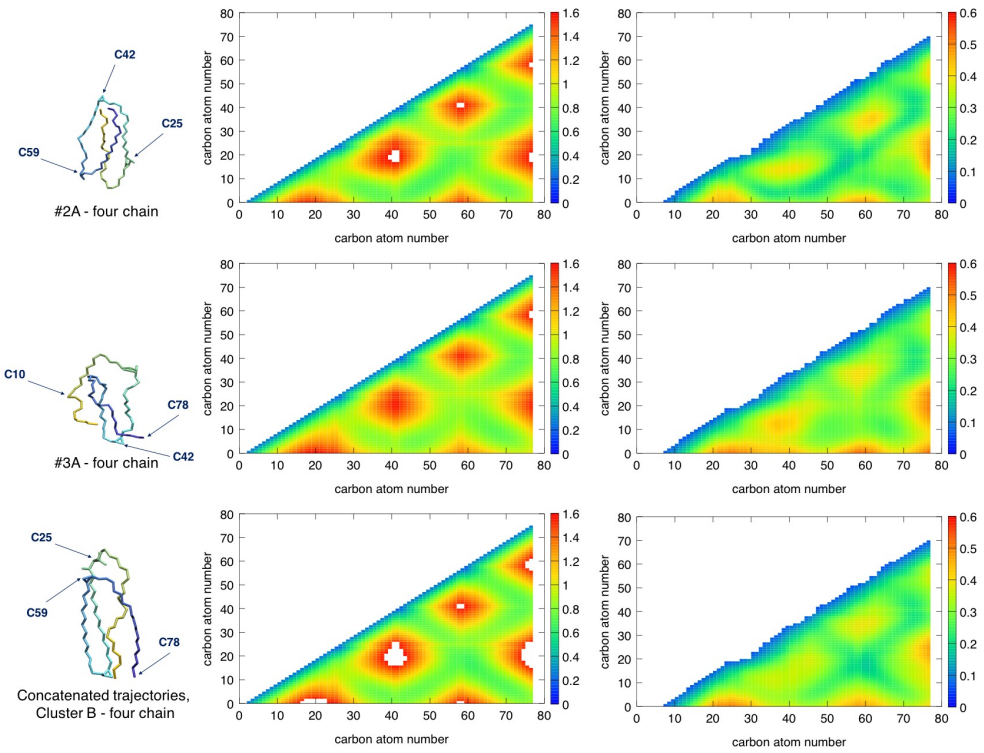


Figure S5. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #2A from molecule AC15.

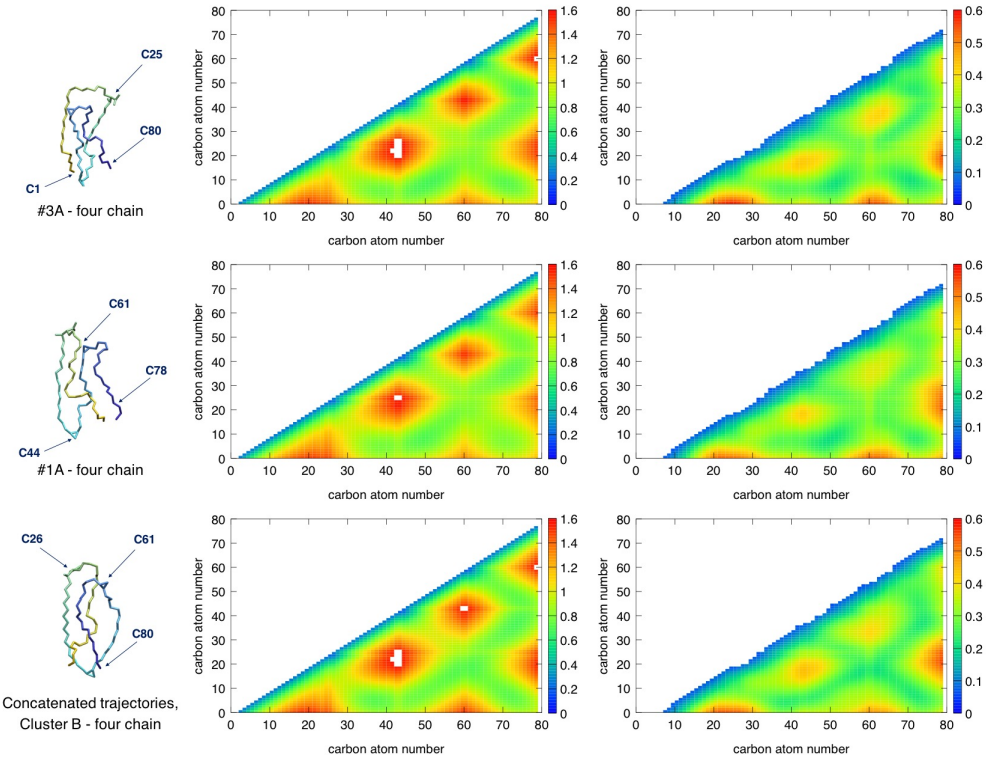


Figure S6. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #3A from molecule AC24.

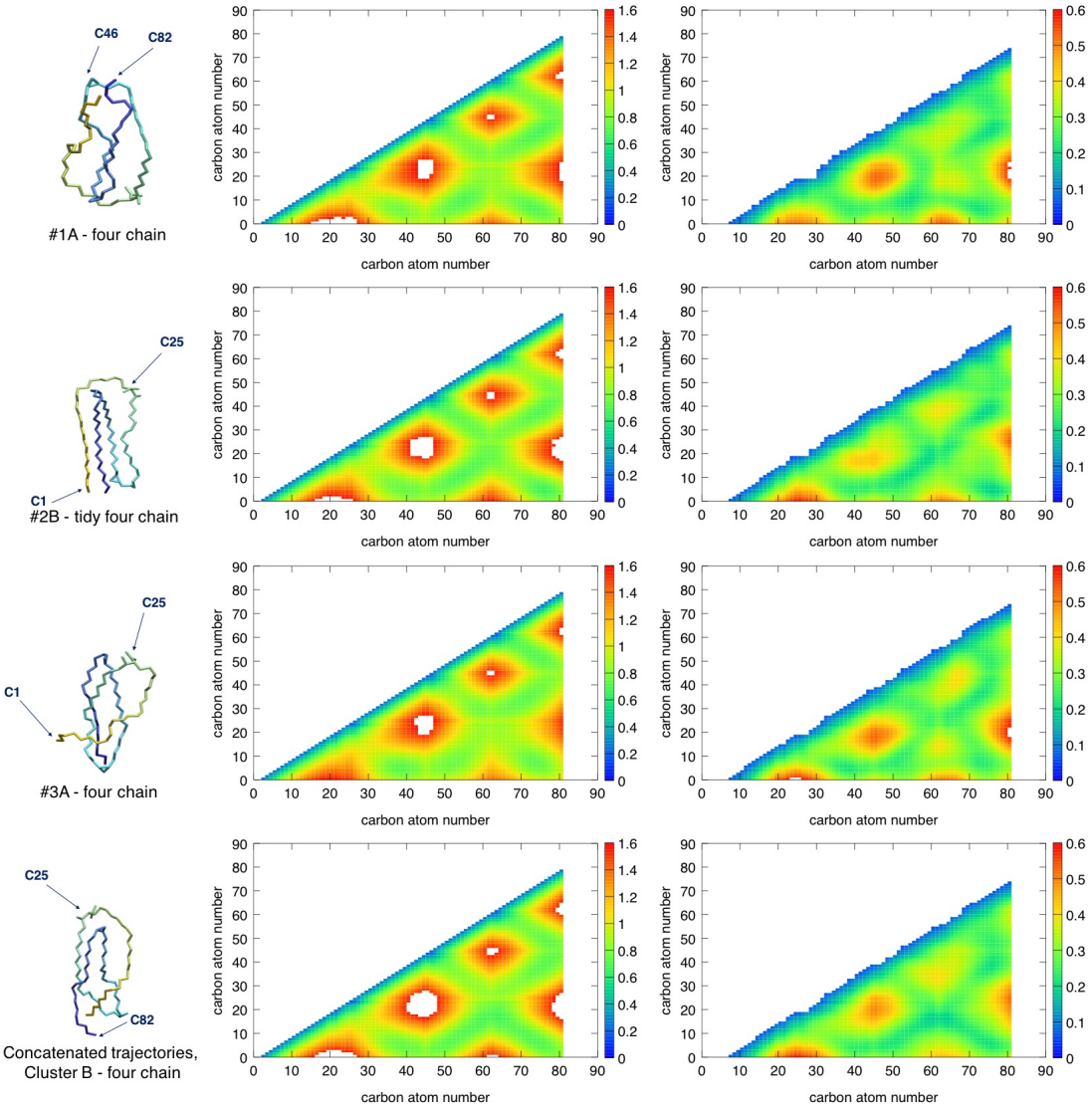


Figure S7. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1A from molecule AC33.

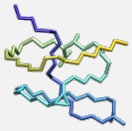
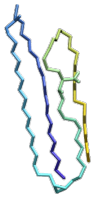
molecule KC4	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		four chain	53.5	0.5292	0.1239
B		four chain	28.9	0.4655	0.1287
C		knot - end	9.1	0.3617	0.1030
D		tidy four chain	8.4	0.3741	0.1002

Table S15. Collated PCA clustering results from the five concatenated trajectories from molecule KC4.

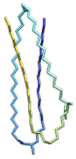
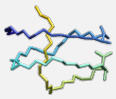
molecule KC12	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		four chain	84.8	0.4864	0.1294
B		tidy four chain	7.8	0.3013	0.0889
C		knot - start	7.4	0.3464	0.1004

Table S16. Collated PCA clustering results from the five concatenated trajectories from molecule KC12.

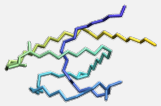
molecule KC20	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		knot - end	66.5	0.5158	0.1320
B		tidy four chain	17.1	0.4728	0.1731
C		tidy four chain	16.4	0.3216	0.0945

Table S17. Collated PCA clustering results from the five concatenated trajectories from molecule KC20.

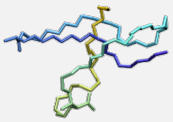
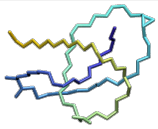
molecule KC27	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		knot - start	63.3	0.4999	0.1313
B		knot - bow	32.4	0.4789	0.1207
C		tidy four chain	4.3	0.2676	0.0821

Table S18. Collated PCA clustering results from the five concatenated trajectories from molecule KC27.

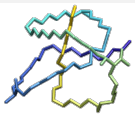
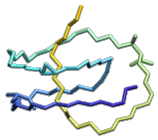
molecule KC29	average structure	shape	cluster % percentage	RMSD of average cluster structure	average RMSD of distance matrices
A		four chain	53.9	0.4935	0.1345
B		knot - start	24.2	0.4361	0.1212
C		five chain	13.3	0.4347	0.1135
D		knot - start	4.8	0.3117	0.0928
E		tidy four chain	3.7	0.2117	0.0756

Table S19. Collated PCA clustering results from the five concatenated trajectories from molecule KC29.

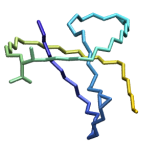
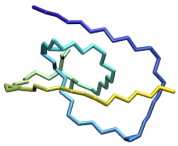
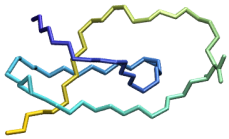
#4 AC15	A - knot 	B - four chain 	C - knot 	D - knot 
A	0	1281	1021	346
B	1274	0	0	661
C	1015	0	0	35
D	359	654	29	0

Table S20. Number of cluster transitions from trajectory #4 of molecule AC15.

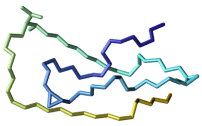
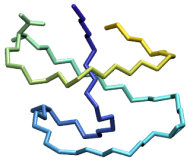
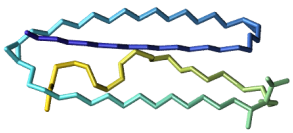
#3 AC24	A - four chain 	B - knot 	C - four chain 
A	0	1864	1509
B	1865	0	19
C	1507	20	0

Table S21. Number of cluster transitions from trajectory #3 of molecule AC24.

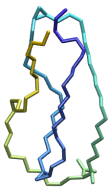
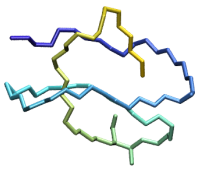
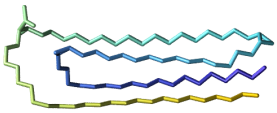
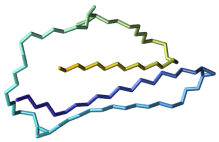
#1 AC33	A - four chain 	B - knot 	C - four chain 	D - four chain 
A	0	1206	1157	620
B	1209	0	0	0
C	1162	0	0	10
D	613	2	15	0

Table S22. Number of cluster transitions from trajectory #1 of molecule AC33.

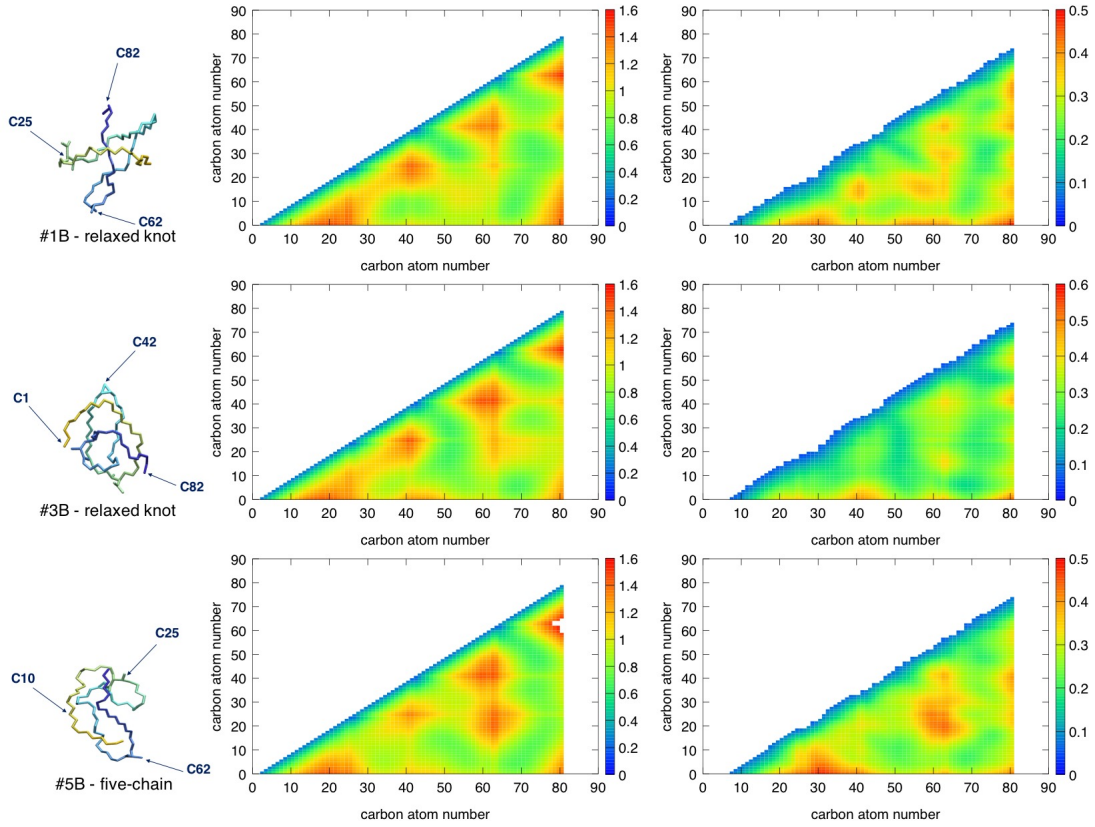


Figure S8. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1B from molecule KC12.

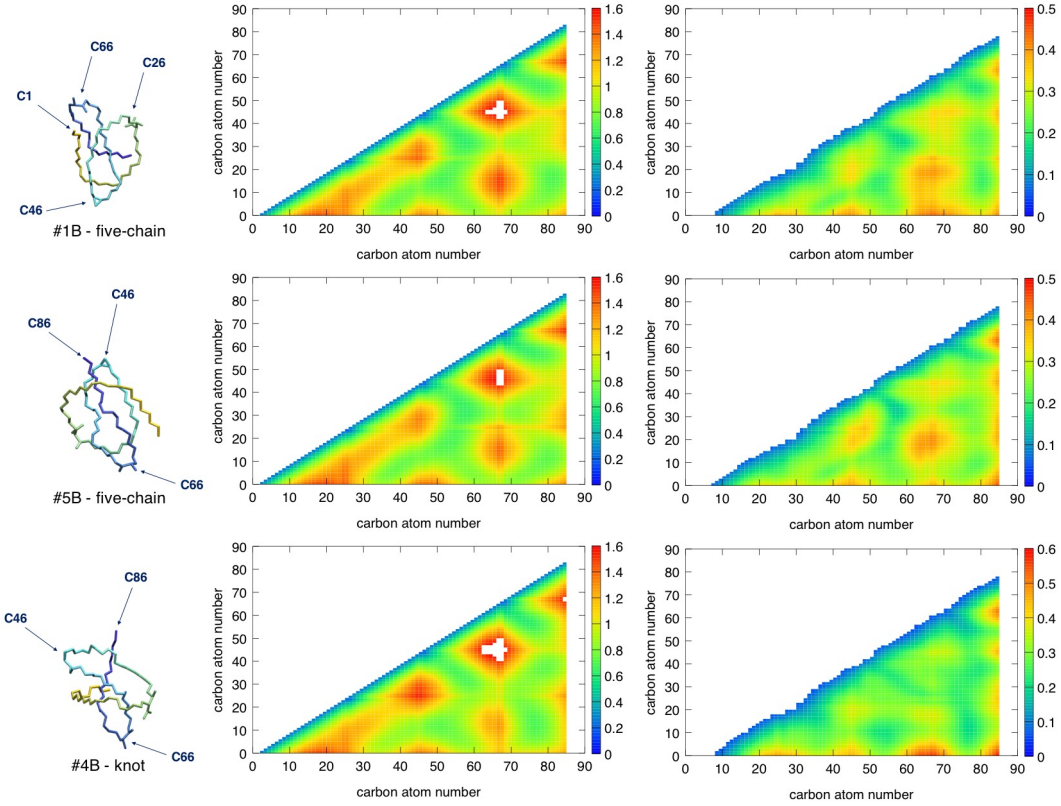


Figure S9. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1B from molecule KC27.

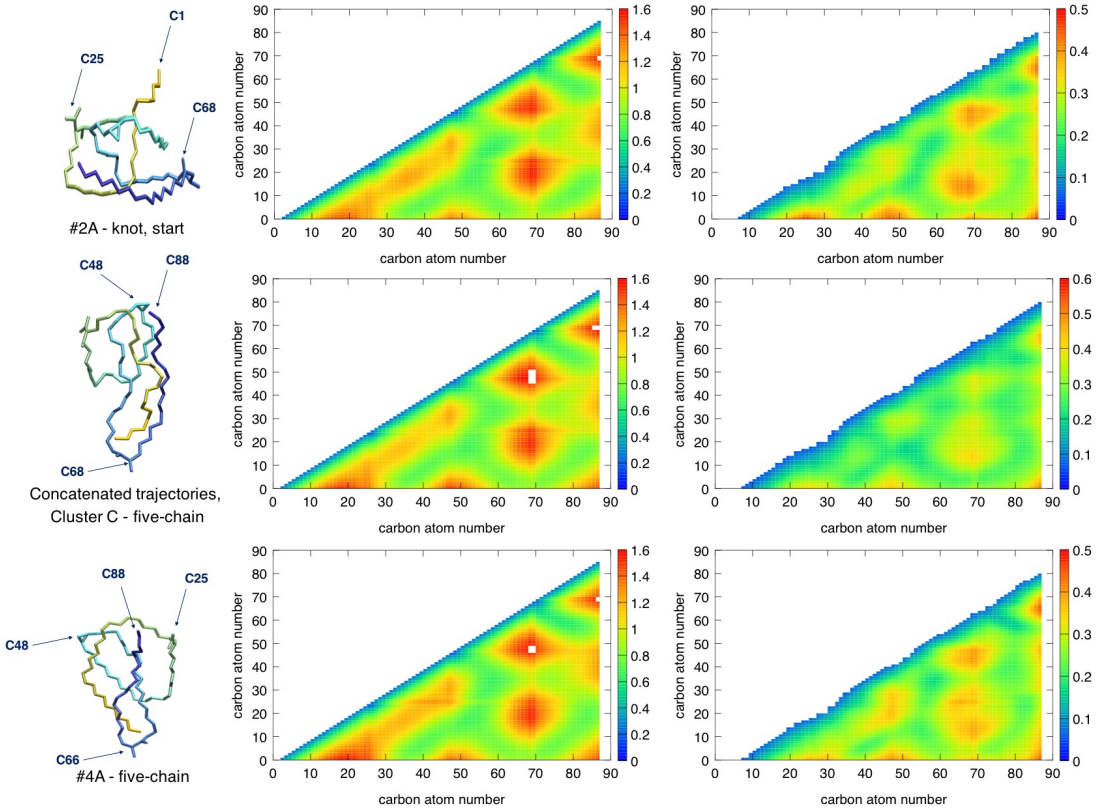


Figure S10. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #2A from molecule KC29.

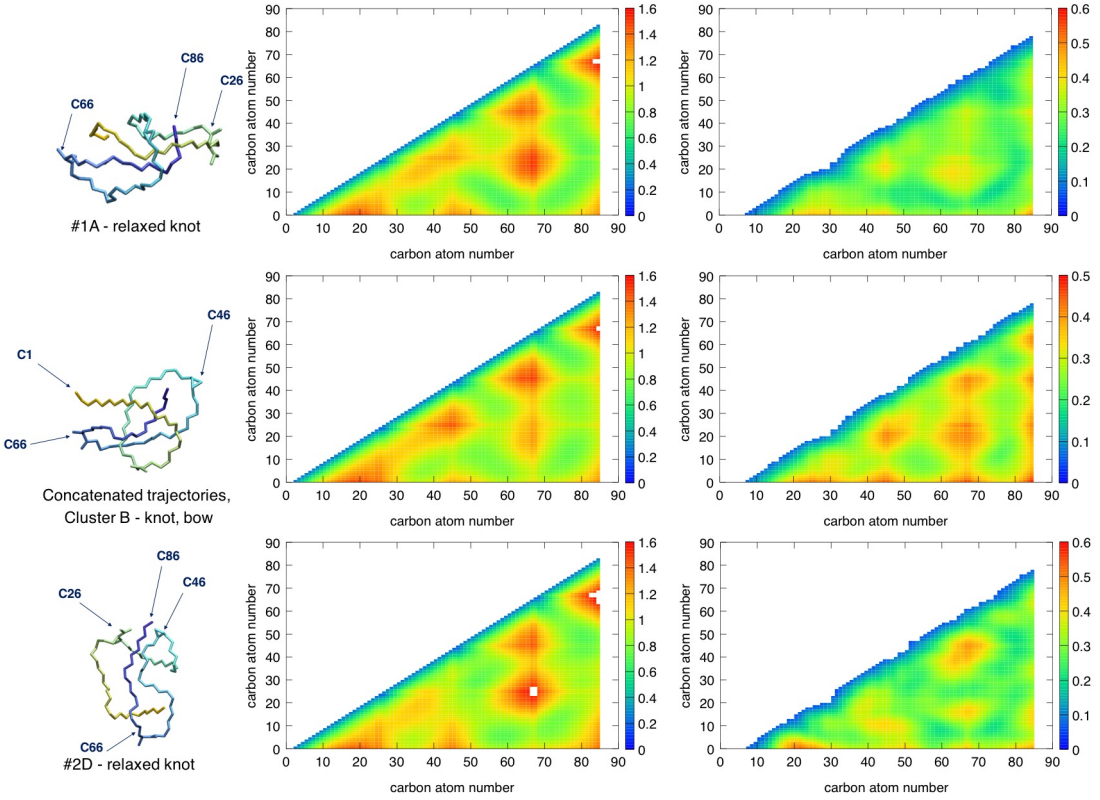


Figure S11. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1A from molecule KC27.

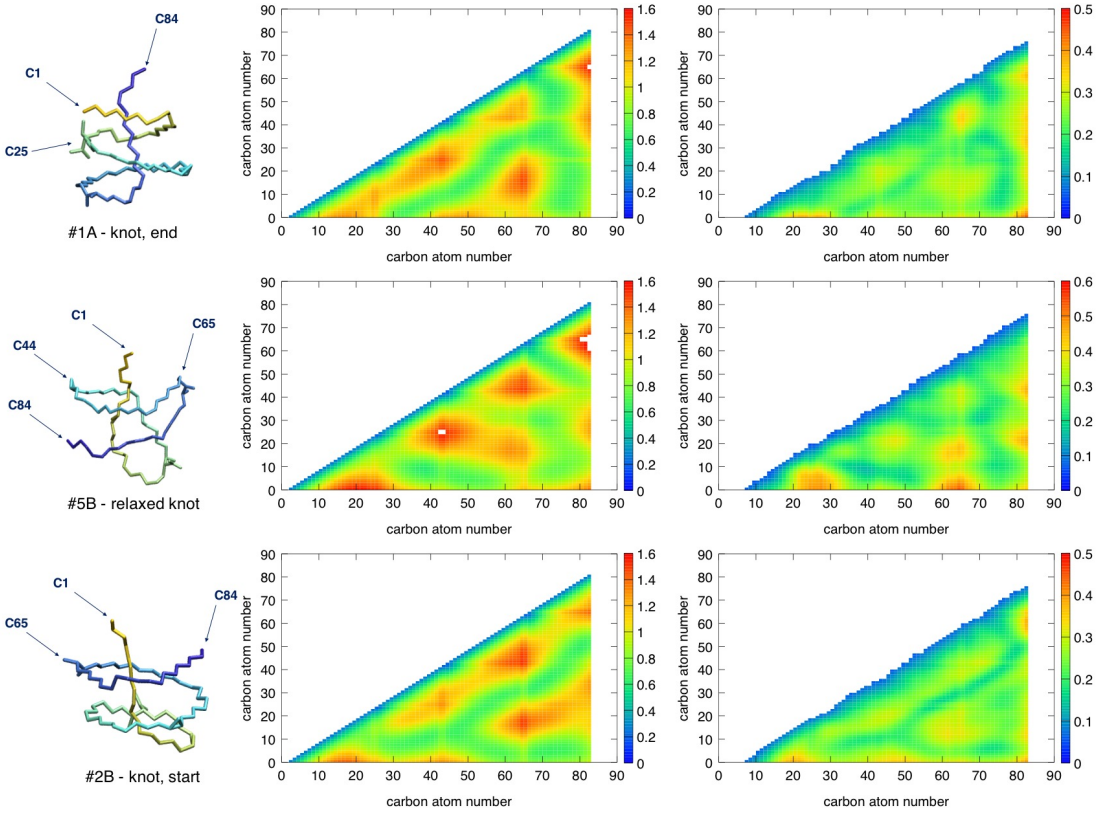


Figure S12. Average distance matrix (left) and RMSD matrix (right) of clusters similar with #1A from molecule KC20.

#1 KC4	A - four chain	B - five chain	C - four chain	D - two chain	E - two chain
A	0	547	1215	17	2
B	547	0	0	0	0
C	1214	0	0	0	0
D	18	0	0	0	6
E	2	0	0	6	0

Table S23. Number of cluster transitions from trajectory #1 of molecule KC4.

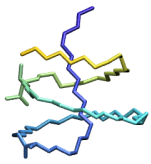
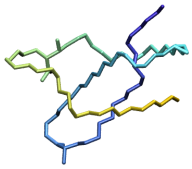
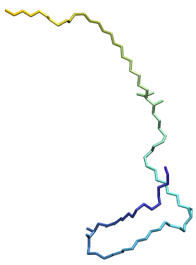
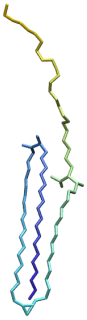
#1 KC20	A - knot 	B - four chain 	C - four chain 	D - three chain 	E - three chain 
A	0	907	2	0	0
B	905	0	1107	4	10
C	5	1104	0	0	0
D	0	3	0	0	65
E	0	12	0	63	0

Table S24. Number of cluster transitions from trajectory #1 of molecule KC20.

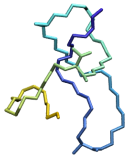
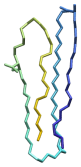
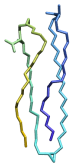
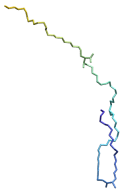
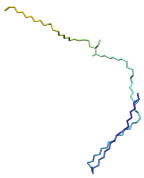
#4 KC20	A - five chain 	B - four chain 	C - four chain 	D - two chain 	E - two chain 
A	0	1932	707	0	0
B	1947	0	1085	5	30
C	692	1100	0	0	0
D	0	6	0	0	33
E	0	29	0	34	0

Table S25. Number of cluster transitions from trajectory #4 of molecule KC20.

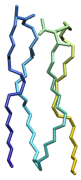
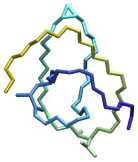
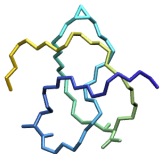
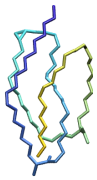
#3 KC12	A - four chain 	B - knot 	C - knot 	D - five chain 
A	0	1597	181	172
B	1588	0	1065	380
C	190	1062	0	246
D	173	374	251	0

Table S26. Number of cluster transitions from trajectory #3 of molecule KC12.

Appendix B

	implicit hexane 300 K	implicit water 300 K	explicit water 350 K	explicit water 400 K
AC6	3.5 μs	5.3 μs	150 ns	N/A
AC15	7.5 μs	6.8 μs	136 ns	59 ns
AC24	5.6 μs	11.5 μs	105 ns	59 ns
AC33	3.3 μs	8.8 μs	70 ns	97 ns
KC25	5.8 μs	N/A	69 ns	94 ns
KC28	5.3 μs	1.43 μs	143 ns	128 ns

Table S27. Total lengths of the grid-computed simulations.

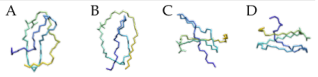
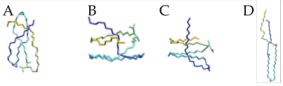
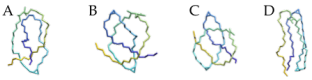
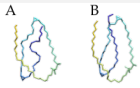
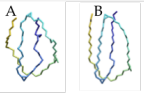
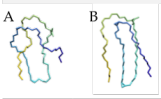
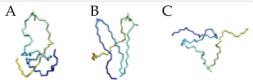
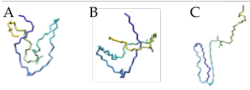
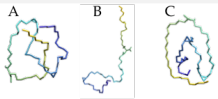
	AC6				AC15				AC24			
	implicit hexane				**implicit hexane**				**implicit hexane**			
average structure												
shape	four chain	four chain	knot - end	knot - end	four chain	knot - end	knot - end	open chain	four chain	five chain	five chain	four chain
cluster %	42	36	19	2	60	37	2	0.001	48	36	14	1
	implicit water				**implicit water**				**implicit water**			
average structure												
shape	four chain	four chain			four chain	four chain			four chain	four chain		
cluster %	97	3			97	3			97	3		
	T350				**T350**				**T350**			
average structure												
shape	four chain	four chain	open chain		four chain	open chain	four chain		four chain	knot - end	open chain	
cluster %	94	4	2		86	10	4		91	6	3	
					T400				**T400**			
												
					four chain	open chain	four chain		four chain	five chain	four chain	
					93	4	3		82	13	5	

table continues in next page

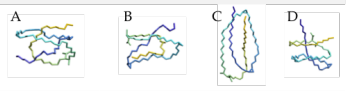
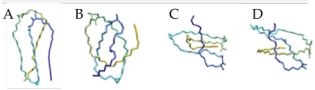
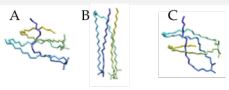
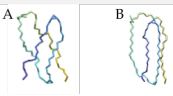
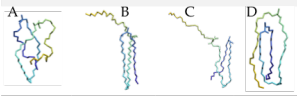
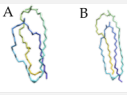
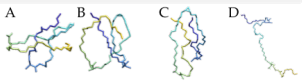
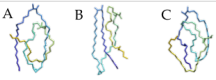
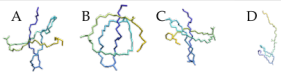
	AC33				KC25				KC28			
	implicit hexane				**implicit hexane**				**implicit hexane**			
average structure												
shape	knot - start	knot - end	four chain	knot - end	four chain	five chain	knot - end	knot - end	knot - end	four chain	knot - end	
cluster %	40	29	29	2	39	35	23	2	70	27	2	
	implicit water				**T350**				**implicit water**			
average structure												
shape	four chain	four chain			four chain	open chain	open chain	four chain	four chain	four chain		
cluster %	97	3			86	7	4	3	97	3		
	T350				**T400**				**T350**			
average structure												
shape	knot - start	open chain	knot - start		knot - end	five chain	four chain	open chain	five chain	four chain	five chain	
cluster %	86	9	4		71	20	8	0.2	74	23	2	
	T400								**T400**			
average structure												
shape	five chain	five chain	four chain						knot - end	five chain	knot - end	open chain
cluster %	70	28	2						81	15	3	0.1

Table S28. Collated PCA clustering results for selected AMA and KMA molecules from the grid-computing approach.

	AC6			
T000 / T001	0.075	0.107		
	0.053	0.053		
	0.070	0.099		
	0.128	0.160		
T000 / T300	0.056	0.051	0.058	0.094
	0.052	0.074	0.100	0.051
	0.060	0.036	0.054	0.081
	0.123	0.063	0.059	0.133
T000 / T350	0.066	0.084	0.534	
	0.068	0.064	0.519	
	0.068	0.084	0.532	
	0.119	0.142	0.541	
T001 / T300	0.077	0.086	0.114	0.092
	0.070	0.118	0.138	0.066
T001 / T350	0.044	0.094	0.495	
	0.102	0.062	0.522	
T300 / T350	0.080	0.039	0.541	
	0.083	0.101	0.540	
	0.102	0.113	0.553	
	0.108	0.078	0.543	

	AC15		
T000 / T001	0.047	0.069	
	0.085	0.126	
	0.144	0.181	
	0.371	0.408	
T000 / T300	0.088	0.074	0.064
	0.047	0.129	0.074
	0.102	0.180	0.129
	0.417	0.407	0.426
T000 / T350	0.048	0.410	0.061
	0.054	0.421	0.080
	0.120	0.442	0.134
	0.371	0.301	0.387
T000 / T400	0.093	0.824	0.069
	0.100	0.834	0.073
	0.149	0.846	0.125
	0.329	0.659	0.373
T001 / T300	0.107	0.091	0.097
	0.149	0.042	0.098
T001 / T350	0.051	0.384	0.080
	0.103	0.411	0.099
T001 / T400	0.065	0.794	0.070
	0.127	0.815	0.114
T300 / T350	0.074	0.438	0.088
	0.106	0.416	0.086
	0.078	0.445	0.069
T300 / T400	0.119	0.847	0.077
	0.135	0.826	0.111
	0.128	0.857	0.082
T350 / T400	0.063	0.811	0.047
	0.344	0.504	0.400
	0.100	0.827	0.045

table continues in next page

AC24

T000 / T001	0.044	0.082		
	0.074	0.137		
	0.133	0.202		
	0.084	0.054		
T000 / T300	0.104	0.033	0.153	0.175
	0.051	0.081	0.107	0.156
	0.052	0.149	0.084	0.148
	0.155	0.050	0.194	0.222
T000 / T350	0.090	0.044	0.536	
	0.078	0.074	0.548	
	0.109	0.138	0.561	
	0.131	0.069	0.540	
T000 / T400	0.089	0.138	0.063	
	0.091	0.118	0.071	
	0.132	0.122	0.131	
	0.126	0.177	0.093	
T001 / T300	0.113	0.070	0.165	0.157
	0.171	0.078	0.218	0.236
T001 / T350	0.068	0.061	0.513	
	0.143	0.087	0.525	
T001 / T400	0.060	0.114	0.067	
	0.133	0.182	0.107	
T300 / T350	0.100	0.104	0.565	
	0.117	0.050	0.550	
	0.151	0.152	0.584	
	0.154	0.186	0.550	
T300 / T400	0.118	0.125	0.094	
	0.114	0.165	0.074	
	0.169	0.172	0.144	
	0.155	0.155	0.180	
T350 / T400	0.040	0.062	0.082	
	0.093	0.136	0.036	
	0.473	0.494	0.536	

AC33

T000 / T001	0.101	0.180	
	0.090	0.167	
	0.077	0.039	
	0.163	0.235	
T000 / T300	0.038	0.144	0.082
	0.048	0.131	0.081
	0.172	0.028	0.161
	0.094	0.198	0.127
T000 / T350	0.074	0.397	0.065
	0.073	0.393	0.066
	0.136	0.387	0.154
	0.144	0.430	0.127
T000 / T400	0.071	0.118	0.079
	0.061	0.116	0.068
	0.110	0.164	0.106
	0.141	0.171	0.137
T001 / T300	0.123	0.079	0.135
	0.199	0.053	0.186
T001 / T350	0.068	0.354	0.103
	0.160	0.386	0.181
T001 / T400	0.049	0.104	0.069
	0.135	0.184	0.131
T300 / T350	0.091	0.409	0.070
	0.133	0.397	0.145
	0.122	0.424	0.102
T300 / T400	0.088	0.129	0.088
	0.104	0.163	0.094
	0.107	0.153	0.093
T350 / T400	0.040	0.073	0.079
	0.365	0.290	0.373
	0.066	0.128	0.089

KC25

T000 / T350	0.115	0.362	0.624	0.080
	0.074	0.407	0.657	0.120
	0.077	0.411	0.660	0.131
	0.081	0.414	0.668	0.118
T000 / T400	0.119	0.163	0.101	1.122
	0.092	0.080	0.100	1.154
	0.096	0.075	0.105	1.157
	0.100	0.099	0.099	1.163
T350 / T400	0.047	0.064	0.050	1.119
	0.341	0.390	0.383	0.884
	0.597	0.623	0.637	0.682
	0.104	0.120	0.075	1.137

KC28

T000 / T001	0.121	0.230		
	0.079	0.061		
	0.125	0.222		
T000 / T350	0.045	0.136	0.118	
	0.174	0.083	0.267	
	0.073	0.139	0.144	
T000 / T400	0.063	0.085	0.048	0.686
	0.150	0.206	0.171	0.668
	0.084	0.112	0.071	0.698
T001 / T350	0.108	0.063	0.210	
	0.218	0.130	0.309	
T001 / T400	0.082	0.141	0.105	0.652
	0.195	0.247	0.215	0.676
T350 / T400	0.037	0.066	0.041	0.668
	0.098	0.161	0.128	0.641
	0.151	0.102	0.136	0.697

Table S29. RMSD values (Å) from average distance matrix comparisons. Different temperatures can be found on the left-hand side (bold).

molecule	cluster / temperature	shape	RMSD value (Å)
AC6	B / T000 and A / T001	four-chain	0.053
	B / T000 and B / T001		0.053
	C / T000 and B / T300	knot (end)	**0.036**
	B / T000 and B/ T350	four-chain	0.064
	B / T001 and D / T300		0.066
	A / T001 and A / T350		0.044
	A / T300 and B / T350		0.039
AC15	A / T000 and A / T001	four-chain	0.047
	B / T000 and A / T300	knot (end) and five-chain	0.047
	A / T000 and A / T350	four-chain	0.048
	A / T000 and C / T400		0.069
	A / T001 and C / T300		0.097
	A / T001 and A / T350		0.051
	A / T001 and A / T400		0.065
	C / T300 and C / T350		0.069
	A / T300 and C / T400	five- and four-chain	0.077
	C / T350 and C / T400	four-chain	**0.045**
AC24	A / T000 and A / T001	four-chain	0.044
	A / T000 and B / T300	four-chain	**0.033**
	A / T000 and B / T350	four-chain and knot (end)	0.044
	A / T000 and C / T400	four-chain	0.063
	A / T001 and B / T300		0.070
	A / T001 and B / T350	four-chain and knot (end)	0.061
	A / T001 and A / T400	four-chain	0.060
	B / T300 and B / T350	four-chain and knot (end)	0.050
	B / T300 and C / T400	four-chain	0.074
	B / T350 and C / T400	knot (end) and four-chain	0.036
AC33	C / T000 and B / T001	four-chain	0.039
	C / T000 and B / T300		**0.028**
	A / T000 and C / T350	knot (end and start)	0.065
	B / T000 and A / T400	knot (end) and five-chain	0.061
	B / T001 and B / T300	four-chain	0.053
	A / T001 and A / T350	four-chain and knot (start)	0.068
	A / T001 and A / T400	four- and five-chain	0.049
	A / T300 and C / T350	knot (relaxed and start)	0.070
	A / T300 and A / T400	knot (relaxed) and five-chain	0.088
	A / T300 and C / T400	knot (relaxed) and four-chain	0.088
	A / T350 and A / T400	knot (start) and five-chain	0.040

KC25	B / T000 and A / T350	five- and four-chain	0.074
	C / T000 and B / T400	knot (end) and five-chain	0.075
	A / T350 and A / T400	four-chain and knot (end)	**0.047**
KC28	B / T000 and B / T001	four-chain	0.061
	A / T000 and A / T350	knot (end) and five-chain	0.045
	A / T000 and C / T400	knot (end)	0.048
	A / T001 and B / T350	four-chain	0.063
	A / T001 and A / T400	four-chain and knot (end)	0.082
	A / T350 and A / T400	five-chain and knot (end)	**0.037**

Table S30. Lowest RMSD values (Å) from average distance matrix comparisons, as described in Table S29. Values are provided per molecule and per temperature. For simplicity, T000 represents implicit hexane and T001 represents implicit water, both in 300 K. The overall lowest RMSD value per molecule is given in bold.

References

1. World Health Organization, *Global TB report 2016*. World Health Organization: Geneva, 2016.

2. Lawn, S. D., *Medicine* **2009,** *37*, 654-656.

3. Sia, I. G.; Wieland, M. L., *Mayo Clin. Proc.* **2011,** *86*, 348.

4. Daniel, T. M., *Respir. Med.* **2006,** *100*, 1862-70.

5. Koch, R., *Berl. Klin. Wochenschr.* **1882,** *19*, 221-230.

6. Calmette, A., *Br. J. Tuberc.* **1928,** *22*, 161.

7. Cattamanchi, A.; Dowdy, D. W.; Davis, J. L.; Worodria, W.; Yoo, S.; Joloba, M.; Matovu, J.; Hopewell, P. C.; Huang, L., *BMC Infect. Dis.* **2009,** *9*, 53.

8. Burrill, J.; Williams, C. J.; Bain, G.; Conder, G.; Hine, A. L.; Misra, R. R., *RadioGraphics* **2007,** *27*, 1255.

9. Padyana, M.; Bhat, R. V.; Dinesha, M.; Nawaz, A., *N. Am. J. Med. Sci.* **2012,** *4*, 221.

10. Centers for Disease Control and Prevention (CDC), *Mantoux Tuberculosis Skin Test Facilitator Guide, Part One: Administering the Mantoux Tuberculin Skin Test*. CDC: Atlanta, 2013.

11. National Institute for Health and Clinical Excellence (NICE), *Clinical diagnosis and management of tuberculosis, and measures for its prevention and control*. NICE: London, 2011.

12. Nayak, S.; Acharjya, B., *Indian Dermatol. Online J.* **2012,** *3*, 2.

13. Starke, J. R., *Pediatrics* **1996,** *98*, 123.

14. Chaturvedi, N.; Cockcroft, A., *Occup. Med.* **1992,** *42*, 179.

15. American Thoracic Society, *Am. J. Respir. Crit. Care Med.* **2000,** *161*, 1376.

16. American Thoracic Society; Centers for Disease Control and Prevention; Infectious Diseases Society of America, *Am. J. Respir. Crit. Care Med.* **2003,** *167*, 603.

17. Moonan, P. K.; Quitugua, T. N.; Pogoda, J. M.; Woo, G.; Drewyer, G.; Sahbazian, B.; Dunbar, D.; Jost, K. C., Jr.; Wallace, C.; Weis, S. E., *BMC Public Health* **2011,** *11*, 19.

18. Hirpa, S.; Medhin, G.; Girma, B.; Melese, M.; Mekonen, A.; Suarez, P.; Ameni, G., *BMC Public Health* **2013,** *13*, 782.

19. Schaberg, T.; Rebhan, K.; Lode, H., *Eur. J. Immunol.* **1996,** *9*, 2026.

20. Gandhi, N. R.; Nunn, P.; Dheda, K.; Schaaf, H. S.; Zignol, M.; van Soolingen, D.; Jensen, P.; Bayona, J., *Lancet* **2010,** *375*, 1830.

21. Sharma, S. K.; Mohan, A., *Chest* **2006,** *130*, 261.

22. Falzon, D.; Schünemann, H. J.; Harausz, E.; González-Angulo, L.; Lienhardt, C.; Jaramillo, E.; Weyer, K., *Eur. Respir. J.* **2017,** *49*, 1602308.

23. World Health Organization, *Global TB report 2013*. World Health Organization: Geneva, 2013.

24. Kurbatova, E. V.; Dalton, T.; Ershova, J.; Tupasi, T.; Caoili, J. C.; Van Der Walt, M.; Kvasnovsky, C.; Yagui, M.; Bayona, J.; Contreras, C.; Leimane, V.; Via, L. E.; Kim, H.; Akksilp, S.; Kazennyy, B. Y.; Volchenkov, G. V.; Jou, R.; Kliiman, K.; Demikhova, O. V.; Cegielski, J. P., *Emerging Infect. Dis.* **2015,** *21*, 977.

25. Kwan, C. K.; Ernst, J. D., *Clin. Microbiol. Rev.* **2011,** *24*, 351.

26. Montales, M. T.; Chaudhury, A.; Beebe, A.; Patil, S.; Patil, N., *Front. Public Health* **2015,** *3*, 281.

27. Lawn, S. D.; Butera, S. T.; Shinnick, T. M., *Microbes Infect.* **2002,** *4*, 635-46.

28. McIlleron, H.; Meintjes, G.; Burman, W. J.; Maartens, G., *J. Infect. Dis.* **2007,** *196*, S63.

29. Wells, C. D.; Cegielski, J. P.; Nelson, L. J.; Laserson, K. F.; Holtz, T. H.; Finlay, A.; Castro, K. G.; Weyer, K., *J. Infect. Dis.* **2007,** *196 Suppl 1*, S86.

30. Willcocks, S.; Wren, B. W., *Future Microbiol.* **2014,** *9*, 657.

31. Rastogi, N.; Legrand, E.; Sola, C., *Rev. Sci. Tech.* **2001,** *20*, 21-54.

32. Jarlier, V.; Nikaido, H., *FEMS Microbiol. Lett.* **1994,** *123*, 11.

33. Li, X.-Z.; Nikaido, H., *Drugs* **2009,** *69*, 1555.

34. Barksdale, L.; Kim, K. S., *Bacteriol. Rev.* **1977,** *41*, 217.

35. McMurray, D. N., Mycobacteria and Nocardia. In *Medical Microbiology*, 4th ed.; University of Texas Medical Branch: Texas, 1996.

36. Armstrong, J. A.; Hart, P. D., *J. Exp. Med.* **1971,** *134*, 713.

37. Pandey, A. K.; Sassetti, C. M., *Proc. Natl. Acad. Sci. U. S. A.* **2008,** *105*, 4376.

38. Barry, C. E., 3rd; Lee, R. E.; Mdluli, K.; Sampson, A. E.; Schroeder, B. G.; Slayden, R. A.; Yuan, Y., *Prog. Lipid Res.* **1998,** *37*, 143.

39. Minnikin, D. E., Lipids: complex lipids, their chemistry, biosynthesis and role. In *The Biology of Mycobacteria*, Ratledge, C., Stanford, J., Ed. Academic Press: London, 1982.

40. Rouhi, A. M., *Chem. Eng. News* **1999,** *77*, 52.

41. Verschoor, J. A.; Baird, M. S.; Grooten, J., *Prog. Lipid Res.* **2012,** *51*, 325.

42. Gebhardt, H.; Meniche, X.; Tropis, M.; Kramer, R.; Daffe, M.; Morbach, S., *Microbiology* **2007,** *153*, 1424.

43. Portevin, D.; De Sousa-D'Auria, C.; Houssin, C.; Grimaldi, C.; Chami, M.; Daffe, M.; Guilhot, C., *Proc. Natl. Acad. Sci. U. S. A.* **2004,** *101*, 314.

44. Vilchèze, C.; Morbidoni, H. R.; Weisbrod, T. R.; Iwamoto, H.; Kuo, M.; Sacchettini, J. C.; Jacobs, W. R., *J. Bacteriol.* **2000,** *182*, 4059.

45. Villeneuve, M.; Kawai, M.; Kanashima, H.; Watanabe, M.; Minnikin, D. E.; Nakahara, H., *Biochim. Biophys. Acta* **2005,** *1715*, 71.

46. Dubnau, E.; Chan, J.; Raynaud, C.; Mohan, V. P.; Laneelle, M. A.; Yu, K.; Quemard, A.; Smith, I.; Daffe, M., *Mol. Microbiol.* **2000,** *36*, 630.

47. Groenewald, W.; Baird, M. S.; Verschoor, J. A.; Minnikin, D. E.; Croft, A. K., *Chem. Phys. Lipids* **2014,** *180*, 15.

48. Watanabe, M.; Aoyagi, Y.; Ridell, M.; Minnikin, D. E., *Microbiology* **2001,** *147*, 1825.

49. Minnikin, D. E.; Goodfellow, M., *Soc. Appl. Bacteriol. Symp. Ser.* **1980,** *8*, 189-256.

50. Watanabe, M.; Aoyagi, Y.; Mitome, H.; Fujita, T.; Naoki, H.; Ridell, M.; Minnikin, D. E., *Microbiology* **2002,** *148*, 1881-1902.

51. Al Dulayymi, J. a. R.; Baird, M. S.; Roberts, E.; Deysel, M.; Verschoor, J., *Tetrahedron* **2007,** *63*, 2571.

52. Al Dulayymi, J. a. R.; Baird, M. S.; Roberts, E., *Tetrahedron* **2005,** *61*, 11939.

53. Al Dulayymi, J. a. R.; Baird, M. S.; Roberts, E.; Minnikin, D. E., *Tetrahedron* **2006,** *62*, 11867-11880.

54. Al Dulayymi, J. a. R.; Baird, M. S.; Mohammed, H.; Roberts, E.; Clegg, W., *Tetrahedron* **2006,** *62*, 4851.

55. Daffe, M.; Brennan, P. J.; McNeil, M., *J. Biol. Chem.* **1990,** *265*, 6734.

56. Abdallah, A. M.; Gey van Pittius , N. C.; DiGiuseppe, P. A.; Cox , J.; Luirink , J., *Nat. Rev. Microbiol.* **2007,** *5*, 883.

57. Niederweis, M., *Mol. Microbiol.* **2003,** *49*, 1167-77.

58. Daffé, M.; Draper, P., *Adv. Microb. Physiol.* **1998,** *39*, 131-203.

59. Minnikin, D. E.; Kremer, L.; Dover, L. G.; Besra, G. S., *Chem. Biol.* **2002,** *9*, 545.

60. Nikaido, H.; Kim, S. H.; Rosenberg, E. Y., *Mol. Microbiol.* **1993,** *8*, 1025-1030.

61. Puech, V.; Chami, M.; Lemassu, A.; Laneelle, M. A.; Schiffler, B.; Gounon, P.; Bayan, N.; Benz, R.; Daffe, M., *Microbiology* **2001,** *147*, 1365-82.

62. Bhamidi, S.; Shi, L.; Chatterjee, D.; Belisle, J. T.; Crick, D. C.; McNeil, M. R., *Anal. Biochem.* **2012,** *421*, 240.

63. Hett, E. C.; Rubin, E. J., *Microbiol. Mol. Biol. Rev.* **2008,** *72*, 126-156.

64. Sani, M.; Houben, E. N. G.; Geurtsen, J.; Pierson, J.; de Punder, K.; van Zon, M.; Wever, B.; Piersma, S. R.; Jiménez, C. R.; Daffé, M.; Appelmelk, B. J.; Bitter, W.; van der Wel, N.; Peters, P. J., *PLoS Pathog.* **2010,** *6*, e1000794.

65. Bloch, H.; Sorkin, E.; Erlenmeyer, H., *Am. Rev. Tuberc.* **1953,** *67*, 629.

66. Benadie, Y.; Deysel, M.; Siko, D. G.; Roberts, V. V.; Van Wyngaardt, S.; Thanyani, S. T.; Sekanka, G.; Ten Bokum, A. M.; Collett, L. A.; Grooten, J.; Baird, M. S.; Verschoor, J. A., *Chem. Phys. Lipids* **2008,** *152*, 95.

67. Fujiwara, N.; Pan, J.; Enomoto, K.; Terano, Y.; Honda, T.; Yano, I., *FEMS Immunol. Med. Microbiol.* **1999,** *24*, 141-149.

68. Pan, J.; Fujiwara, N.; Oka, S.; Maekura, R.; Ogura, T.; Yano, I., *Microbiol. Immunol.* **1999,** *43*, 863-9.

69. Fujita, Y.; Naka, T.; McNeil, M. R.; Yano, I., *Microbiology* **2005,** *151*, 3403.

70. Ueda, S.; Fujiwara, N.; Naka, T.; Sakaguchi, I.; Ozeki, Y.; Yano, I.; Kasama, T.; Kobayashi, K., *Microb. Pathog.* **2001,** *30*, 91-9.

71. Lewis, R. N. A. H.; McElhaney, R. N., The mesomorphic phase behaviour of lipid bilayers. In *The Structure of Biological Membranes*, CRC Press: 1991.

72. Han, Y.; Kawada, N.; Yano, I., *Osaka City Med. J.* **1998,** *44*, 201-17.

73. Schleicher, G. K.; Feldman, C.; Vermaak, Y.; Verschoor, J. A., *Clin. Chem. Lab. Med.* **2002,** *40*, 882-7.

74. Mathebula, N. S.; Pillay, J.; Toschi, G.; Verschoor, J. A.; Ozoemena, K. I., *Chem. Commun. (Cambridge, U. K.)* **2009,** *0*, 3345-7.

75. Villeneuve, M.; Kawai, M.; Watanabe, M.; Aoyagi, Y.; Hitotsuyanagi, Y.; Takeya, K.; Gouda, H.; Hirono, S.; Minnikin, D. E.; Nakahara, H., *Biochim. Biophys. Acta* **2007,** *1768*, 1717.

76. Lemmer, Y.; Kalombo, L.; Pietersen, R. D.; Jones, A. T.; Semete-Makokotlela, B.; Van Wyngaardt, S.; Ramalapa, B.; Stoltz, A. C.; Baker, B.; Verschoor, J. A.; Swai, H. S.; de Chastellier, C., *J. Control. Release* **2015,** *211*, 94.

77. Belardinelli, J. M.; Yazidi, A.; Yang, L.; Fabre, L.; Li, W.; Jacques, B.; Angala, S. k.; Rouiller, I.; Zgurskaya, H. I.; Sygusch, J.; Jackson, M., *ACS Infect. Dis.* **2016,** *2*, 702-713.

78. Li, W.; Obregon-Henao, A.; Wallach, J. B.; North, E. J.; Lee, R. E.; Gonzalez-Juarrero, M.; Schnappinger, D.; Jackson, M., *Antimicrob. Agents Chemother.* **2016,** *60*, 5198-207.

79. Degiacomi, G.; Benjak, A.; Madacki, J.; Boldrin, F.; Provvedi, R.; Palù, G.; Kordulakova, J.; Cole, S. T.; Manganelli, R., *Sci. Rep.* **2017,** *7*, 43495.

80. Barry, C. E.; Mdluli, K., *Trends Microbiol.* **1996,** *4*, 275.

81. Wang, L.; Slayden, R. A.; Barry, C. E., 3rd; Liu, J., *J. Biol. Chem.* **2000,** *275*, 7224-9.

82. Etienne, G.; Villeneuve, C.; Billman-Jacobe, H.; Astarie-Dequeker, C.; Dupont, M. A.; Daffe, M., *Microbiology* **2002,** *148*, 3089.

83. Yuan, Y.; Zhu, Y.; Crane, D. D.; Barry, C. E., 3rd, *Mol. Microbiol.* **1998,** *29*, 1449-58.

84. Dkhar, H. K.; Nanduri, R.; Mahajan, S.; Dave, S.; Saini, A.; Somavarapu, A. K.; Arora, A.; Parkesh, R.; Thakur, K. G.; Mayilraj, S.; Gupta, P., *J. Immunol.* **2014,** *193*, 295-305.

85. Slama, N.; Jamet, S.; Frigui, W.; Pawlik, A.; Bottai, D.; Laval, F.; Constant, P.; Lemassu, A.; Cam, K.; Daffe, M.; Brosch, R.; Eynard, N.; Quemard, A., *Mol. Microbiol.* **2016,** *99*, 794-807.

86. Ojha, A. K.; Trivelli, X.; Guerardel, Y.; Kremer, L.; Hatfull, G. F., *J. Biol. Chem.* **2010,** *285*, 17380-9.

87. Ojha, A. K.; Baughn, A. D.; Sambandan, D.; Hsu, T.; Trivelli, X.; Guerardel, Y.; Alahari, A.; Kremer, L.; Jacobs, W. R., Jr.; Hatfull, G. F., *Mol. Microbiol.* **2008,** *69*, 164-74.

88. Sambandan, D.; Dao, D. N.; Weinrick, B. C.; Vilcheze, C.; Gurcha, S. S.; Ojha, A.; Kremer, L.; Besra, G. S.; Hatfull, G. F.; Jacobs, W. R., Jr., *mBio* **2013,** *4*, e00222-13.

89. Vander Beken, S.; Al Dulayymi, J. a. R.; Naessens, T.; Koza, G.; Maza-Iglesias, M.; Rowles, R.; Theunissen, C.; De Medts, J.; Lanckacker, E.; Baird, M. S.; Grooten, J., *Eur. J. Immunol.* **2011,** *41*, 450-60.

90. Beukes, M.; Lemmer, Y.; Deysel, M.; Al Dulayymi, J. a. R.; Baird, M. S.; Koza, G.; Iglesias, M. M.; Rowles, R. R.; Theunissen, C.; Grooten, J.; Toschi, G.; Roberts, V. V.; Pilcher, L.; Van Wyngaardt, S.; Mathebula, N.; Balogun, M.; Stoltz, A. C.; Verschoor, J. A., *Chem. Phys. Lipids* **2010,** *163*, 800.

91. Glickman, M. S.; Cox, J. S.; Jacobs, W. R., *Mol. Cell.* **2000,** *5*, 717.

92. Dao, D. N.; Sweeney, K.; Hsu, T.; Gurcha, S. S.; Nascimento, I. P.; Roshevsky, D.; Besra, G. S.; Chan, J.; Porcelli, S. A.; Jacobs, W. R., *PLoS Pathog.* **2008,** *4*, e1000081.

93. Villeneuve, M.; Kawai, M.; Watanabe, M.; Aoyagi, Y.; Hitotsuyanagi, Y.; Takeya, K.; Gouda, H.; Hirono, S.; Minnikin, D. E.; Nakahara, H., *Chem. Phys. Lipids* **2010,** *163*, 569.

94. Villeneuve, M.; Kawai, M.; Horiuchi, K.; Watanabe, M.; Aoyagi, Y.; Hitotsuyanagi, Y.; Takeya, K.; Gouda, H.; Hirono, S.; Minnikin, D. E., *Microbiology* **2013,** *159*, 2405.

95. Hasegawa, T.; Nishijo, J.; Watanabe, M.; Funayama, K.; Imae, T., *Langmuir* **2000,** *16*, 7325.

96. Hasegawa, T.; Leblanc, R. M., *Biochim. Biophys. Acta* **2003,** *1617*, 89.

97. Hasegawa, T.; Amino, S.; Kitamura, S.; Matsumoto, L.; Katada, S.; Nishijo, J., *Langmuir* **2003,** *19*, 105.

98. Hasegawa, T., *Biopolymers* **2004,** *73*, 457.

99. Schone, A. C.; Roch, T.; Schulz, B.; Lendlein, A., *J. R. Soc. Interface* **2017,** *14*.

100. Borrell, J. H.; Domènech, Ò.; Keough, K. M. W., *Membrane Protein – Lipid Interactions: Physics and Chemistry in the Bilayer*. Springer International Publishing: 2016.

101. Groenewald, W. Diverse approaches to modelling conformations of mycolic acids from *Mycobacterium tuberculosis*. PhD Thesis. Bangor University, Bangor, Wales, 2013.

102. Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H., *J. Phys. Chem. B* **2007,** *111*, 7812.

103. Marrink, S. J.; Tieleman, D. P., *Chem. Soc. Rev.* **2013,** *42*, 6801.

104. Rao, V.; Gao, F.; Chen, B.; Jacobs, W. R., Jr.; Glickman, M. S., *J. Clin. Invest.* **2006,** *116*, 1660-7.

105. Young, D., *Computational Chemistry: A Practical Guide for Applying Techniques to Real World Problems*. Wiley: New York, 2004.

106. Jensen, F., *Introduction to Computational Chemistry*. Wiley: New York, 2007.

107. Weinhold, F.; Landis, C. R., *Valency and Bonding: A Natural Bond Orbital Donor-Acceptor Perspective*. Cambridge University Press: Cambridge, 2005.

108. Griffiths, D. J., *Introduction to Quantum Mechanics*. Pearson Prentice Hall: New Jersey, 2005.

109. Mueller, M. P., *Fundamentals of Quantum Chemistry: Molecular Spectroscopy and Modern Electronic Structure Computations*. Springer: New York, 2007.

110. Hehre, W. J., *A Guide to Molecular Mechanics and Quantum Chemical Calculations*. Wavefunction: Irvine, CA, 2003.

111. Haken, H.; Wolf, H. C., *The Physics of Atoms and Quanta: Introduction to Experiments and Theory*. Springer: New York, 2000.

112. Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P., *J. Am. Chem. Soc.* **1985,** *107*, 3902.

113. Cramer, C. J., *Essentials of Computational Chemistry: Theories and Models*. Wiley: New York, 2005.

114. Groenhof, G., *Methods Mol. Biol.* **2013,** *924*, 43.

115. Perilla, J. R.; Schulten, K., *Nat. Commun.* **2017,** *8*, 15959.

116. van der Spoel, D.; Lindahl, E.; Hess, B.; and the GROMACS development team GROMACS User Manual 4.6.7. www.gromacs.org (2014).

117. McCammon, J. A.; Gelin, B. R.; Karplus, M., *Nature* **1977,** *267*, 585.

118. Lindner, B.; Petridis, L.; Schulz, R.; Smith, J. C., *Biomacromolecules* **2013,** *14*, 3390.

119. Wang, X.; Xu, F.; Liu, J.; Gao, B.; Liu, Y.; Zhai, Y.; Ma, J.; Zhang, K.; Baker, T. S.; Schulten, K.; Zheng, D.; Pang, H.; Sun, F., *PLoS Pathog.* **2013,** *9*, e1003132.

120. Tarasova, E.; Farafonov, V.; Khayat, R.; Okimoto, N.; Komatsu, T. S.; Taiji, M.; Nerukh, D., *J. Phys. Chem. Lett.* **2017,** *8*, 779.

121. Roy, J.; Laughton, C. A., *Biophys. J.* **2010,** *99*, 218.

122. Klepeis, J. L.; Lindorff-Larsen, K.; Dror, R. O.; Shaw, D. E., *Curr. Opin. Struct. Biol.* **2009,** *2*, 120.

123. Gasteiger, J.; Engel, T., *Chemoinformatics: A Textbook*. Wiley: New York, 2003.

124. Roweis, S. T.; Saul, L. K., *Science* **2000,** *290*, 2323.

125. Hastie, T.; Tibshirani, R.; Friedman, J., *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer: New York, 2009.

126. Livingstone, D. J., *A Practical Guide to Scientific Data Analysis*. Wiley: New York, 2009.

127. Kavraki, L. E. Dimensionality Reduction Methods for Molecular Motion. http://cnx.org/contents/02ff5dd2-fe30-4bf5-8e2a-83b5c3dc0333@10 (accessed July 15, 2016).

128. Keller, B.; Daura, X.; van Gunsteren, W. F., *J. Chem. Phys.* **2010,** *132*, 074110.

129. van der Spoel, D.; Lindahl, E.; Hess, B.; Groenhof, G.; Mark, A. E.; Berendsen, H. J., *J. Comput. Chem.* **2005,** *26*, 1701-18.

130. Gaussian 09, Revision D.01, Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Mennucci, B.; Petersson, G. A.; Nakatsuji, H.; Caricato, M.; Li, X.; Hratchian, H. P.; Izmaylov, A. F.; Bloino, J.; Zheng, G.; Sonnenberg, J. L.; Hada, M.; Ehara, M.; Toyota, K.; Fukuda, R.; Hasegawa, J.; Ishida, M.; Nakajima, T.; Honda, Y.; Kitao, O.; Nakai, H.; Vreven, T.; Montgomery, J. A., Jr.; Peralta, J. E.; Ogliaro, F.; Bearpark, M.; Heyd, J. J.; Brothers, E.; Kudin, K. N.; Staroverov, V. N.; Kobayashi, R.; Normand, J.; Raghavachari, K.; Rendell, A.; Burant, J. C.; Iyengar, S. S.; Tomasi, J.; Cossi, M.; Rega, N.; Millam, N. J.; Klene, M.; Knox, J. E.; Cross, J. B.; Bakken, V.; Adamo, C.; Jaramillo, J.; Gomperts, R.; Stratmann, R. E.; Yazyev, O.; Austin, A. J.; Cammi, R.; Pomelli, C.; Ochterski, J. W.; Martin, R. L.; Morokuma, K.; Zakrzewski, V. G.; Voth, G. A.; Salvador, P.; Dannenberg, J. J.; Dapprich, S.; Daniels, A. D.; Farkas, Ö.; Foresman, J. B.; Ortiz, J. V.; Cioslowski, J.; Fox, D. J., Gaussian, Inc., Wallingford CT. 2009.

131. Weininger, D., *J. Chem. Inf. Model.* **1988,** *28*, 31-36.

132. Weininger, D.; Weininger, A.; Weininger, J. L., *J. Chem. Inf. Model.* **1989,** *29*, 97-101.

133. Dassault Systèmes BIOVIA, Discovery Studio Modeling Environment, Release 4.5, San Diego: Dassault Systèmes. 2015.

134. Youngs, T. G. A., *J. Comput. Chem.* **2010,** *31*, 639-648.

135. Jorgensen, W. L.; Maxwell, D. S.; Tirado-Rives, J., *J. Am. Chem. Soc.* **1996,** *118*, 11225.

136. Kaminski, G.; Duffy, E. M.; Matsui, T.; Jorgensen, W. L., *J. Phys. Chem.* **1994,** *98*, 13077-13082.

137. Siu, S. W. I.; Pluhackova, K.; Böckmann, R. A., *J. Chem. Theory Comput.* **2012,** *8*, 1459-1470.

138. Shinoda, W.; Okazaki, S., *J. Mol. Liq.* **2001,** *90*, 95-103.

139. Hess, B.; Kutzner, C.; van der Spoel, D.; Lindahl, E., *J. Chem. Theory Comput.* **2008,** *4*, 435-447.

140. Laughton, C. A.; Orozco, M.; Vranken, W., *Proteins: Struct., Funct., Bioinf.* **2009,** *75*, 206-216.

141. Lewars, E. G., *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics*. Springer Netherlands: 2010.

142. Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L., *J. Chem. Phys.* **1983,** *79*, 926-935.

143. Bussi, G.; Donadio, D.; Parrinello, M., *J. Chem. Phys.* **2007,** *126*, 014101.

144. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R., *J. Chem. Phys.* **1984,** *81*, 3684-3690.

145. Darden, T.; York, D.; Pedersen, L., *J. Chem. Phys.* **1993,** *98*, 10089-10092.

146. Still, W. C.; Tempczyk, A.; Hawley, R. C.; Hendrickson, T., *J. Am. Chem. Soc.* **1990,** *112*, 6127-6129.

147. Salt, D. W.; Hudson, B. D.; Banting, L.; Ellis, M. J.; Ford, M. G., *J. Med. Chem.* **2005,** *48*, 3214-20.

148. Shkurti, A.; Goni, R.; Andrio, P.; Breitmoser, E.; Bethune, I.; Orozco, M.; Laughton, C. A., *SoftwareX* **2016,** *5*, 44-50.

149. Haensele, E.; Banting, L.; Whitley, D. C.; Clark, T., *J. Mol. Model.* **2014,** *20*, 2485.

150. Lanig, H.; Reisen, F.; Whitley, D.; Schneider, G.; Banting, L.; Clark, T., *PLoS One* **2015,** *10*, e0135246.

151. Victor Paul Raj, F. R. D.; Exner, T. E., *J. Mol. Model.* **2014,** *20*, 2196.

152. Peyron, P.; Vaubourgeix, J.; Poquet, Y.; Levillain, F.; Botanch, C.; Bardou, F.; Daffé, M.; Emile, J.-F.; Marchou, B.; Cardona, P.-J.; de Chastellier, C.; Altare, F., *PLoS Pathog.* **2008,** *4*, e1000204.

153. Vaidehi, N.; Jain, A., *J. Phys. Chem. B* **2015,** *119*, 1233-1242.

154. Balaraman, G. S.; Park, I.-H.; Jain, A.; Vaidehi, N., *J. Phys. Chem. B* **2011,** *115*, 7588-7596.

155. Larsen, A. B.; Wagner, J. R.; Jain, A.; Vaidehi, N., *J. Chem. Inf. Model.* **2014,** *54*, 508-517.

156. Hubbard, R. E.; Kamran Haider, M., Hydrogen Bonds in Proteins: Role and Strength. In *Encyclopedia of life sciences*, Wiley: New York, 2001.

157. Mori, T.; Saito, S., *J. Chem. Phys.* **2015,** *142*, 135101.

158. Kloczkowski, A.; Jernigan, R. L.; Wu, Z.; Song, G.; Yang, L.; Kolinski, A.; Pokarowski, P., *J. Struct. Funct. Genomics* **2009,** *10*, 67-81.

159. Lim, N. C. H.; Jackson, S. E., *J. Phys.: Condens. Matter* **2015,** *27*, 354101.

160. Jamroz, M.; Niemyska, W.; Rawdon, E. J.; Stasiak, A.; Millett, K. C.; Sułkowski, P.; Sulkowska, J. I., *Nucleic Acids Res.* **2015,** *43*, D306-D314.

161. Hsu, S.-T. D., *Sci. Rep.* **2016,** *6*, 39357.

162. Mallam, A. L.; Jackson, S. E., *Nat. Chem. Biol.* **2011,** *8*, 147-53.

163. Sułkowska, J. I.; Sułkowski, P.; Onuchic, J., *Proc. Natl. Acad. Sci. U. S. A.* **2009,** *106*, 3119-3124.

164. Barran, P. E.; Cole, H. L.; Goldup, S. M.; Leigh, D. A.; McGonigal, P. R.; Symes, M. D.; Wu, J.; Zengerle, M., *Angew. Chem., Int. Ed.* **2011,** *50*, 12280-12284.

165. Papaleo, E.; Mereghetti, P.; Fantucci, P.; Grandori, R.; De Gioia, L., *J. Mol. Graphics Modell.* **2009,** *27*, 889-899.

166. Tama, F.; Miyashita, O.; Kitao, A.; Go, N., *Eur. Biophys. J.* **2000,** *29*, 472-80.

167. Di Pietro, O.; Juárez-Jiménez, J.; Muñoz-Torrero, D.; Laughton, C. A.; Luque, F. J., *PLoS One* **2017,** *12*, e0177683.

168. Maisuradze, G. G.; Liwo, A.; Scheraga, H. A., *J. Mol. Biol.* **2009,** *385*, 312-329.

169. Thompson, M. J.; Eisenberg, D., *J. Mol. Biol.* **1999,** *290*, 595-604.

170. Kumar, S.; Tsai, C.-J.; Nussinov, R., *Protein Eng., Des. Sel.* **2000,** *13*, 179-191.

171. Pack, S. P.; Kang, T. J.; Yoo, Y. J., *Appl. Biochem. Biotechnol.* **2013,** *171*, 1212-1226.

172. Wei, S.; Ahlstrom, L. S.; Brooks, C. L., *Small* **2017,** *13*, 1603748.

173. Barkan, D.; Liu, Z.; Sacchettini, J. C.; Glickman, M. S., *Chem. Biol.* **2009,** *16*, 499-509.

174. Jones, A.; Pitts, M.; Al Dulayymi, J. a. R.; Gibbons, J.; Ramsay, A.; Goletti, D.; Gwenin, C. D.; Baird, M. S., *PLoS One* **2017,** *12*, e0181414.

175. Orsi, M.; Essex, J. W., *Soft Matter* **2010,** *6*, 3797-3808.

176. Voth, G. A., *Coarse-Graining of Condensed Phase and Biomolecular Systems*. CRC Press: Boca Raton, FL, 2009.