The Making and Breaking of SAS-6: Structural Insights and Inhibitor Search for N-terminal Domain Dimerisation



Julia Maria Christiane Busch

Department of Biochemistry

University of Oxford

A thesis submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at the University of Oxford

Pembroke College

Trinity Term 2017

To free spirits and forgotten scientists.

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text. However, in consideration of the nature of scientific work being inherently collaborative and acknowledging that scientific work is not possible without the help, advice and feedback of other scientists, I decided to write this thesis using first-person plural "we" to refer to what otherwise would have been "I".

Acknowledgements

I want to thank my supervisor Dr. Ioannis Vakonakis for inviting me to work in his group, for answering endless questions, for giving me all opportunities to grow as a scientist and as a person. I want to thank my second supervisor Dr. Philip C. Biggin for including me in his group, for giving me all the chances to explore new techniques, for continued support and advice.

I want to thank my colleagues old and new, who all have been generous with help and advice, who have taught me so much and always made working with them feel effortless and the scientific community seem like a family. Colleagues from the Vakonakis group I want to specifically thank are: Dr. Michele Erat - for teaching me uncountable tricks and for inviting me along for great trips through the countryside, Leanne Slater - for putting up with me, Dr. Dirk Reiter - for emergency chocolate, Jodie Ford and Jemma Day - for being the best and most creative office-mates I could wish for, Dr. Anastasia Kantsadi and Dr. Qiuhong (Beatrice) Li - for their patience, generous help, for hugs and smiles. In the Biggin group I want to specifically thank: Dr. Maria Musgaard - for teaching me about computational biology and for her endless patience with me and Dr. Laura Domicevica - for help, friendship and exercise motivation. I also want to thank everyone working in the Rodney Porter building - for creating a welcoming working environment, for coffee break chats, cake, help, advice and movie nights.

I want to thank Dr. David Staunton, Dr. Ed Lowe, Mr Nick Soffe, Dr. Chris-

tina Redfield, Dr. Minos Matsoukas, Dr. David J. Scott and Dr. Jemma Harris - for help with experiments and techniques and for answering my many questions.

I also want to acknowledge Dr. Silke Hauf, who let me make my first steps in science in her laboratory, where I found my love for science and Dr. Ian Campbell, who inspired me and made me believe that I could actually be a scientist.

I am grateful to my funding bodies, the BBSRC Interdisciplinary Bioscience DTP, Pembroke college and the Biochemistry department. All these institutions have always made me feel welcome and I sincerely hope that I could contribute to their community and goals. The DTP has been a safe haven for me. I thank Dr. Gail Preston - for always taking the time to listen and to give advice. I thank Pembroke college - for giving me a place to belong. I thank everyone with open doors in the Biochemistry department - for invaluable advice and reassuring conversations.

I am deeply grateful for the friends I got to make during my time in Oxford. I have never met so many utterly nice people. I want to thank everyone in the 2013 DTP intake - for the amazing times we had and for being a source of continuous motivation, support and laughter. Thank you Lucy Taylor - for dragging me out of the lab sometimes and for being a rock in stormy situations. A big thank you to Andra Necula, Andrea Lei Leston and Samuel Cahill - for being my friends and housemates, for being there, for listening, sharing and caring, for every cup of tea and Red Star evenings, for being fun, loving and inspiring. Thank you so much guys.

I would not have been able to do this work without the support of my parents. I am very grateful, that they enabled me to have this experience - for their belief in me and for so much more, thank you. I thank my sister - for being the best little sister one can have and for not giving up on me.

Last, but decidedly not least, the honorable mention belongs to my husband Harald Busch, who has supported me through the tough days and has celebrated my

achievements, even when I could not, who has delivered late night food to the laboratory on his visits to Oxford and has send me countless letters to fill the void when he could not be there, who has not once in earnest complained about my absence, my inattention or my scientific madness. It is incredible to me, that I am so lucky to have you by my side. Thank you so much.

Abstract

"The Making and Breaking of SAS-6: Structural Insights and Inhibitor Search for N-terminal Domain Dimerisation" - Julia M. C. Busch - Pembroke College - Submission for Doctor of Philosophy in Trinity Term 2017

SAS-6 is the structural core of the forming centriole - a cylindrical protein complex, which is an essential component of the centrosome. Oligomerisation of SAS-6 is crucial for successful centriole duplication and is achieved through two dimerisation domains in the SAS-6 protein; a long C-terminal coiled-coil domain and a globular N-terminal dimerisation domain. As core components of the centrosome, centrioles help facilitate various cellular functions. They are involved in the anchoring of flagella and cilia to the membrane and in coordinating the spindle apparatus during chromosome segregation. A deeper insight into the molecular mechanisms at play in the centriole duplication process would have implications on our understanding of fundamental cell division processes and a number of related diseases.

Here the involvement of an unstudied loop region in the *C. elegans* SAS-6 N-terminal domain dimerisation is described. Combining structural biology, biophysical and computational techniques, the molecular interactions of this loop were explored, contributing to the oligomerisation of SAS-6 at the N-terminal dimer interface. Furthermore, the screening and testing of small molecule inhibitors of the SAS-6 N-terminal domain dimerisation is described, targeting a hydrophobic pocket in the domain. Two candidate compounds are presented as a result of the screens and next steps towards structure based compound design are suggested, based on computational analysis. The search for inhibitory compounds includes a set-up of an in-house virtual screening pipeline, as well as *in vitro* screening efforts and a new crystallographic structure of the *H. sapiens* SAS-6 N-terminal domain. By investigating the making and breaking of the SAS-6 N-terminal domain dimerisation, light is shed on so far neglected details of this essential protein-protein interaction and advancements towards a SAS-6 oligomerisation inhibitor described, which could ultimately be used for new approaches in cell cycle research and might open up new avenues for medical research by binding a disease relevant target.

Abbreviations

$x\mathbf{log}P$ calculated logarithmic partition-coefficient.

C. elegans *Caenorhabditis elegans.*

C. reinhardtii *Chlamydomonas reinhardtii.*

D. melanogaster *Drosophila melanogaster.*

D. rerio *Danio rerio.*

E. coli *Escherichia coli.*

H. sapiens *Homo sapiens.*

L. major *Leishmania major.*

X. laevis *Xenopus laevis.*

5-FAM 5-carboxyfluorescein.

ANA2 anastral spindle 2.

Asl asterless.

AUC analytical ultracentrifugation.

CBCA(CO)NH CβCα(CO)NH.

CBCANH CβCαNH.

CD circular dichroism.

CID cartwheel inner densities.

CPAP centrosomal P4.1-associated protein.

CV column volume.

DMSO dimethyl sulfoxide.

EDTA ethylenediaminetetraacetic acid.

EM electron microscopy.

ESI electrospray ionisation.

FID free induction decay.

FP fluorescence polarisation.

FPLC fast protein liquid chromatography.

FRAP fluorescence recovery after photobleaching.

HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid.

HMBC heteronuclear multiple-bond correlation.

HN(CO)CA HN(CO)Cα.

HNCA HNCα.

HSQC heteronuclear single quantum coherence.

IAEDANS 5-(2-[(iodoacetyl)amino]ethylamino)naphthalene-1-sulfonic acid.

IPTG isopropyl β-D-1-thiogalactopyranoside.

LB Luria Bertani medium.

log$P$ logarithmic partition-coefficient.

MALS multi-angle light scattering.

MCC multiciliated cells.

MD molecular dynamics.

MRE mean residue ellipticity.

MS mass spectrometry.

MTOC microtubule organising centre.

NEB New England Biolabs.

NMR nuclear magnetic resonance.

NOE nuclear Overhauser effect.

OD optical density.

PAGE polyacrylamide gel electrophoresis.

PBS phosphate buffered saline.

PCM pericentriolar material.

PCR polymerase chain reaction.

PCT pre-crystallisation test.

PES polyethersulfone.

PISA "present in SAS-6".

Plk4 polo like kinase 4.

PME particle mesh Ewald.

RMSD root-mean-square deviation.

SAR structure-activity relationship.

SAS spindle assembly abnormal protein.

SDS sodium dodecylsulfate.

SEC size exclusion chromatography.

SMD steered molecular dynamics.

SOC super optimal broth with catabolite repression.

SPB spindle pole bodies.

SPD spindle deficient protein.

STAN STIL/ANA2.

STIL SCL/TAL1 interrupting locus.

TCEP tris(2-carboxyethyl)phosphine.

UV ultraviolet.

ZYG-1 zygote 1.

Table of contents

Abstract **viii**

Table of contents **xv**

List of figures **xxi**

List of tables **xxvii**

1 Introduction **1**

1.1 Centrosomes, Centrioles and a Hint of History 1

1.1.1 Centrosome and Centriole in the Tree of Life 2

1.2 Centrosome and Centriole Function 4

1.2.1 Cell Division 5

1.2.2 Basal Bodies, Cilia and Flagella 6

1.3 Centrosomes and Centrioles in Health and Disease 7

1.4 Centriole Structure 9

1.4.1 The Cartwheel 14

1.5 Centrosome Duplication and Centriole Assembly 17

1.6 Spindle Assembly Abnormal Protein 6 (SAS-6) 23

1.6.1 SAS-6 Function and Regulation 24

1.6.2 Drug-target Potential of SAS-6 25

1.6.3 Architecture and Conservation 26

1.6.4 SAS-6 N-terminal Domain 29
1.7 Aim of This Thesis 34

2 Methods **35**

2.1 Cloning and Mutagenesis 35
2.1.1 Construct Design 36
2.2 Protein Production and Purification 39
2.2.1 Protein Expression 40
2.2.2 Cell Lysis 42
2.2.3 Affinity Chromatography 42
2.2.4 Proteolytic Tag Cleavage and Dialysis 43
2.2.5 Protein Quality Control 44
2.3 Biophysical Interaction Assays 46
2.3.1 Fluorescence Polarisation 46
2.3.2 Analytical Ultracentrifugation 49
2.4 Nuclear Magnetic Resonance Spectroscopy 52
2.4.1 Heteronuclear Single Quantum Correlation 53
2.4.2 Protein Backbone Assignments 55
2.4.3 Secondary Structure Information from NMR 60
2.4.4 Small Molecule NMR 61
2.5 X-ray Crystallography 62
2.5.1 Structure Determination of *H. sapiens* SAS-6 N-terminal Domain 64
2.5.2 Preliminary Structure Determination of *C. elegans* SAS-6 N-terminal Domain 65
2.5.3 Small Molecule X-ray Crystallography 66
2.6 Computational Techniques 66
2.6.1 Homology Modeling 67

2.6.2 Docking, Virtual Screening and Pipeline Set-up 67
2.6.3 Molecular Dynamics Simulations 75

3 Insights into the *C. elegans* SAS-6 α2-β5 Loop and Its Involvement in N-terminal Domain Dimerisation 85

3.1 Introduction 86
3.2 Biophysical Insight into the α2-β5 Loop 87
3.2.1 The α2-β5 Loop is Flexible and Disordered 87
3.2.2 The Presence of Loop α2-β5 Supports SAS-6 N-terminal Domain Dimerisation 90
3.3 Simulating Loop α2-β5 and Its Role in SAS-6 N-terminal Domain Dimerisation 93
3.3.1 Supporting Evidence for Loop 103-130 Involvement in SAS-6 Dimerisation from Steered MD 98
3.4 Conclusion 105

4 Virtual Screening for Inhibitors of SAS-6 Dimerisation 107

4.1 Introduction 108
4.1.1 Target Selection 108
4.1.2 Screening Method Selection 111
4.2 Results & Discussion 117
4.2.1 Preliminary Experiments 117
4.2.2 Virtual Screening for Compounds Binding *C. elegans* SAS-6 . . 122
4.3 Conclusion 138

5 *In Vitro* Search for SAS-6 Oligomersiation Inhibitory Compounds 141

5.1 NMR Based Screen for Small Molecules Binding SAS-6 N-terminal Domain 142

5.1.1 Compound Solubility 143
5.1.2 Screen for Compounds Binding *C. elegans* SAS-6 N-terminal Domain 146
5.1.3 Screen for Compounds Binding *C. reinhardtii* SAS-6 N-terminal Domain 153
5.1.4 Screen for Compounds Binding *H. sapiens* SAS-6 N-terminal Domain 156
5.2 Candidate Compound Effect on SAS-6 Oligomerisation 164
5.3 Conclusion 171

6 Structural Insights into Candidate Compound Binding to the SAS-6 N-terminal Domain 175
6.1 Introduction 175
6.2 Crystallographic Approach to Elucidating Binding Poses of Candidate Compounds 176
6.3 Investigating Candidate Compound Binding Positions with MD Simulations 180
6.4 Conclusion 189

7 Discussion and Future Directions 191
7.1 The Protein-Protein Interface of the SAS-6 N-terminal Domain Dimer . 191
7.2 Discussion on Next Steps in Structure Based Inhibitor Design Targeting SAS-6 Oligomerisation 198
7.3 Considerations on Future Use of SAS-6 N-terminal Dimerisation Inhibitors 212
7.3.1 N-terminal Domain Dimerisation Inhibitors as Research Tool . . 212
7.3.2 N-terminal Domain Dimerisation Inhibitors as Drug Candidates 214

8 Conclusion **217**

References **219**

Appendix A Chapter 3 **241**

A.1 Observations on the Role of S123 on Loop 103-130 Interactions from Molecular Dynamics Simulation 241

Appendix B Chapter 4 **249**

B.1 Virtual Screening Pipeline Script 250

B.1.1 split-mol2-3.0.py 250

B.1.2 pdbqt-converter.py 252

B.1.3 docking.py 254

B.1.4 docking-parallel.py 257

B.1.5 scoring.py 260

Appendix C Chapter 5 **265**

C.1 Complete NMR Screening Results for Compound Selection 265

C.2 Small Molecule Identity 269

C.2.1 NMR Assignments for A11 and A12 269

C.2.2 Mass Spectrometry Results for A11 and A12 274

C.2.3 Small Molecule X-ray Crystallography of A11 and A12 276

Appendix D Chapter 6 **279**

D.1 Crystal Morphology and Crystal Packing of hsSAS-6_N F131E 279

Appendix E Additional Work on Centriolar Protein STIL **281**

E.1 Investigation into the Phosphorylation Dependent Interaction of STIL and the SAS-6 Coiled-Coil Domain 281

E.1.1 Introduction 281
E.1.2 STIL STAN and SAS-6 Coiled-Coil Constructs and Purification 285
E.1.3 Localising STIL STAN Interaction Region on SAS-6 Coiled-Coil Domain 287
E.1.4 Development of Quantitative Interaction Assays for STIL STAN Phosphomimetic Constructs with SAS-6 289
E.2 Conclusion 293

List of figures

1.1 Phylogenetic Tree of Eukaryotes 4
1.2 Electron Microscopic Images of Centrioles 11
1.3 Architecture of Centrioles and Cartwheels 12
1.4 Cartwheel Stacks in Various Organisms 15
1.5 Centrosome Duplication in the Mammalian Cell Cycle 19
1.6 Centriole Duplication Pathways 20
1.7 SAS-6 Architecture 24
1.8 SAS-6 Oligomer Structure, Ring and Spiral 28
1.9 Structural and Sequential Conservation of SAS-6 N-terminal Domain . 30
1.10 SAS-6 N-dimer Interface Structures 31

2.1 Magnetisation Transfer in HSQC 53
2.2 Magnetisation Transfer in CBCA(CO)NH, CBCANH, HN(CO)CA and HNCA 57
2.3 Protein Backbone Assignment Strategy 59
2.4 High Throughput Docking Pipeline Set-up 71
2.5 Examples of SAS-6 Protein Structure in Simulation Box 83

3.1 Sequence Alignment - SAS-6 N-terminal Domain 87
3.2 Chemical Shift Index Analysis for Cα and Cβ Resonances of *C. elegans* SAS-6 N-terminal Domain 89

3.3 Heteronuclear $\{^1H\}$-^{15}N NOE Intensities for Residues of *C. elegans* SAS-6 N-terminal Domain 90
3.4 SEC and FP Experiments Following SAS-6 N-terminal Domain Dimerisation 91
3.5 ^{15}N-HSQC NMR Spectra of GLY Residues of SAS-6_N Constructs . . . 93
3.6 Model of SAS-6 N-terminal Domain Loop 103-130 in MD 95
3.7 Analysis of Interactions Formed During MD Between Loop 103-130 and α1-α2 97
3.8 SAS-6_N Δ 103-130 Constructs During SMD Simulations 100
3.9 Work Required for SAS-6_N Dimer Separation in SMD 101
3.10 SAS-6_N Constructs with Modelled Loop 103-130 During SMD Simulations 103
3.11 Work Required for SAS-6_N Dimer Separation in SMD 104

4.1 SAS-6 N-dimer Structure and Interface 110
4.2 Overlay of ^{15}N-HSQC NMR Spectra - *C. elegans* SAS-6 N-terminal Domain With Tryptophan, Cyclic Peptide and ZINC00392813 118
4.3 Chemical Shift Perturbation of Preliminary SAS-6 NMR Experiments . 119
4.4 High Throughput Docking Pipeline Set-up 124
4.5 Docking Parameter Test - Control Substances in Docking Position . . . 128
4.6 Vina Scores Reached for Top 10,000 Compounds 128
4.7 Structural Formulae for Top 20 Ranked Compounds 131
4.8 Docking Positions of Compounds in SAS-6 Structure 132
4.9 HSQC NMR Spectral Overlay for Example Compounds 133
4.10 Depiction of Pharmacophore Scaffold Design 135
4.11 Structural Formulae for Compounds Selected from Pharmacophore-like Screen 137

5.1 1D ^{1}H NMR Spectra - *C. elegans* SAS-6 N-terminal Domain With A11, C10 and A3 144
5.2 Overlay of ^{15}N-HSQC NMR Spectra - *C. elegans* SAS-6 N-terminal Domain With A11, A12 and B1 147
5.3 Chemical Shift Perturbation from ^{15}N-HSQC NMR Screen - *C. elegans* SAS-6 N-terminal Domain with A11, A12 and B1 149
5.4 Chemical Shift Perturbation from ^{15}N-HSQC NMR Screen - *C. elegans* SAS-6 N-terminal Domain with A11, A12 and B1 150
5.5 ^{15}N-HSQC NMR Titration - *C. elegans* SAS-6 N-terminal Domain with A11 152
5.6 Overlay of ^{15}N-HSQC NMR Spectra - *C. reinhardtii* SAS-6 N-terminal Domain With A11, A12 and B1 154
5.7 Chemical Shift Perturbation from NMR Screen - *C. reinhardtii* SAS-6 N-terminal Domain with A11, A12 and B1 155
5.8 Chemical Shift Perturbation from NMR Screen - *C. reinhardtii* SAS-6 N-terminal Domain with A11, A12 and B1 156
5.9 Overlay of ^{15}N-HSQC NMR Spectra - *H. sapiens* SAS-6 N-terminal Domain With A11, A12 and B1 158
5.10 Chemical Shift Perturbations from NMR Screen - *H. sapiens* SAS-6 N-terminal Domain with A11, A12 and B1 159
5.11 Chemical Shift Perturbations from NMR Screen - *H. sapiens* SAS-6 N-terminal Domain with A11, A12 and B1 160
5.12 ^{15}N-HSQC NMR Titration - *H. sapiens* SAS-6 N-terminal Domain with A11 162
5.13 ^{15}N-HSQC NMR Titration - *H. sapiens* SAS-6 N-terminal Domain with A12 163

5.14 Sedimentation Coefficient Distribution for *C. elegans* SAS-6 Oligomerisation with A11 . . . 166
5.15 Sedimentation Coefficient Distribution for *C. reinhardtii* SAS-6 Oligomerisation with A11 . . . 168
5.16 SEC MALS Results for hsSAS-6 1-199 and 1-205 . . . 169
5.17 Sedimentation Coefficient Distribution for *H. sapiens* SAS-6 Oligomerisation with A11 . . . 170

6.1 Crystal Structure of *Homo sapiens* (*H. sapiens*) SAS-6 N-terminal Domain 177
6.2 Sequence and Structural Conservation of SAS-6 . . . 178
6.3 Simulated A11 Binding Position on *C. elegans* SAS-6 N-terminal Domain 182
6.4 Simulated A11 Binding Position on *C. reinhardtii* SAS-6 N-terminal Domain . . . 183
6.5 Simulated A11 Binding Position on *H. sapiens* SAS-6 N-terminal Domain 184
6.6 Interaction Time of A11 with Pocket Residues in Simulation . . . 186
6.7 Simulated A12 Binding Position on *H. sapiens* SAS-6 N-terminal Domain 187
6.8 Interaction Time of A12 with Pocket Residues in Simulation . . . 188

7.1 Residues Involved in the SAS-6 N-terminal Dimerisation Interface . . . 193
7.2 Examples of Loop α2-β5 Interacting with Residues in the α1-α2 Region in the Dimerisation Interface . . . 194
7.3 SAS-6 N-dimer Interface Structures . . . 196
7.4 Predicted Binding Poses of A11 on SAS-6 N-terminal Domain . . . 200
7.5 Predicted Binding Poses of A12 on SAS-6 N-terminal Domain . . . 200
7.6 Identifying A11 Components Essential for SAS-6 Binding . . . 201
7.7 Sulfonamide Containing Compounds in Screening Selection . . . 202
7.8 Resonance Differences from NMR Screen with *C. elegans* SAS-6_N for Compounds A11, A5 and A10 . . . 204

7.9 Resonance Differences from NMR Screen with *C. reinhardtii* SAS-6_N for Compounds A11, A5 and A10 205
7.10 Resonance Differences from NMR Screen with *H. sapiens* SAS-6_N for Compounds A11, A5 and A10 206
7.11 Selection of Commercially Available Compounds for Next Generation Screening Based on A11 210
7.12 Selection of Commercially Available Compounds for Next Generation Screening Based on A11 and A12 211

A.1 Model of SAS-6 N-terminal Domain Loop 103-130 (S123A, S123E and S123D) in MD 243
A.2 Analysis of Interactions Formed During MD between Loop 103-130 and α1-α2 (S123A, S123E and S123D) 244
A.3 SAS-6_N Constructs with Modelled Loop 103-130 S123E During SMD Simulations 246
A.4 Work Required for SAS-6_N Dimer Separation in SMD 247

C.1 NMR Screening Results for Full Compound Selection - ceSAS-6 266
C.2 NMR Screening Results for Full Compound Selection - crSAS-6 267
C.3 NMR Screening Results for Full Compound Selection - hsSAS-6 268
C.4 Small Molecule NMR Analysis of A11 - 1D NMR ^{1}H NMR 270
C.5 Small Molecule NMR Analysis of A12 - 1D NMR ^{1}H NMR 271
C.6 Small Molecule NMR Analysis of A11 - 2D NMR ^{13}C-^{1}H HSQC and HMBC 272
C.7 Small Molecule NMR Analysis of A12 - 2D NMR ^{13}C-^{1}H HSQC and HMBC 273
C.8 Mass Spectrometry Result for A11 275
C.9 Mass Spectrometry Result for A12 276

C.10 Small Molecule Crystallography A11 and A12 278

D.1 Crystal Morphology and Crystal Packing of hsSAS-6_N F131E 279

E.1 STIL Architecture and Conservation 284

E.2 Possible Phosphorylation Sites in STIL STAN 285

E.3 ^{15}N ^{1}H HSQC NMR - STIL STAN and SAS-6 Coiled-Coil Interaction . 289

E.4 Interaction Assay for STIL STAN and SAS-6 Coiled-Coil 291

List of tables

1.1 Table of SAS-6 N-terminal Domain Structures 33

2.1 Protein Constructs 37
2.2 Primers 38
2.3 Polymerase Chain Reaction Protocol 38
2.4 Bacterial Strains 39
2.5 Molecular Weight of Protein Constructs 40
2.6 Proteases 43
2.7 Grid Configurations Used for Docking with *Autodock Vina* 68
2.8 Scripts for Virtual Screening Set-up 70
2.9 Test Runtimes for Developed Python Scripts 72
2.10 List of Compounds and Vendors 74

4.1 Control Substances for Docking Parameters 121
4.2 List of Top 20 Ranked Compounds 130
4.3 List of Compounds Selected by Pharmacophore-like Screen 136

5.1 Solubility of Screening Library Compounds 145

E.1 STIL STAN and SAS-6_{CC} Constructs 286

Chapter 1

Introduction

"It was the best of times, it was the worst of times..."

Charles Dickens

A Tale of Two Cities

This thesis describes research into the N-terminal domain dimerisation of spindle assembly abnormal protein (SAS)-6, a crucial component of the centriole duplication pathway. In this introductory section the biological context of SAS-6 and the current understanding of the structure and function of this protein are summarised.

1.1 Centrosomes, Centrioles and a Hint of History

Centrosomes are the primary microtubule organising centre (MTOC) in most eukaryotic cells. These organelles were first depicted in drawings by Van Benden, who studied mitosis in metazoan parasites in 1876. The term "centrosome" appeared when Theodor Boveri recognised this organelle as linked to cell division in 1887 (as reviewed in [1]). The term centrosome defines an organelle that 1. maintains itself in the cell centre due to microtubule nucleating/anchoring activity, 2. duplicates once per cell

cycle and 3. physically associates with the nucleus [2].

The term "centriole" was also coined by Theodor Boveri in his 1900 publication "Über die Natur der Centrosomen" [3]. In *Echinus microtuberculatus* egg slices he observed a dense granular structure within the previously discovered centrosome in the context of the cell cycle. The dense granules Boveri found within the centrosome are tightly linked to cell cycle progression in most eukaryotic cells and are connected to cancerous cell development [4–7]. Boveri correctly identified centrioles as separate entities within the centrosome. We now consider the centrosome of animal cells as the assembly of two centrioles surrounded by a less dense assembly of proteins called the pericentriolar material (PCM). We have now had more than 100 years to study these organelles, but our understanding of their role, regulation and life cycle is still far from complete.

1.1.1 Centrosome and Centriole in the Tree of Life

The centrosomes of animal cells consist of two centrioles surrounded by PCM. The centrosomes' role in microtubule organisation makes it a wide spread organelle throughout the evolutionary tree. However, it does not always consist of centrioles. In addition to animal cells, brown algae and some Amoebozoa also have internalised centrioles that organize the microtubule skeleton during certain life cycle stages (Figure 1.1) [8]. Other Amoebozoa and higher fungi have lost centrioles and cilia. Instead of centriole containing centrosomes, higher fungi have centrosome structures called spindle pole bodies (SPB) [9]. These are cylindrical organelles embedded in the nuclear envelope not consisting of centrioles. In some Amoebozoa such as *Dictystelium discoideum* the centrosome is a structure called the nuclear-associated body. This is a box-shaped three layered complex surrounded by amorphous material which can be compared to the PCM [10]. The centrosome found in diatoms, which are a form of microalgae, is the

so-called microtubule centre [11]. Higher plants have no centriole or centrosome equivalent and can form spindles in the absence of a distinguishable MTOC [12]. Despite the diversity of centrosomes between species, acentriolar centrosomes share structural commonalities with animal centrosomes, suggesting a common evolutionary origin [7]. For example an orthologue of the centriolar protein spindle deficient protein (SPD)-2 was found in the *Dictyostelium* genome and the protein locaslises to the *Dictyostelium* centrosome [7, 13].

In vertebrate cells the mother centriole can template the assembly of a primary or motile cilium [7, 14]. In this configuration the mother centriole is functionally equivalent to the basal body, and basal bodies and centrioles can interconvert. All species that have centrioles have cilia at some stage of their life cycle [14]. This might suggest that the ancestral function of the basal body and the centriole was the nucleation of cilia [15]. It is theorised that an *Amorphea* common ancestor of eukaryotes had flagella and actin based motion properties and might be the evolutionary point of origin of the centriole (Figure 1.1) [7]. Red algea, angiosperm, microsporidia and higher fungi comprise the only eukaryotes that seem to have lost centrioles [12], but they all appear to have descended from ancestors that once contained centrioles (Figure 1.1).

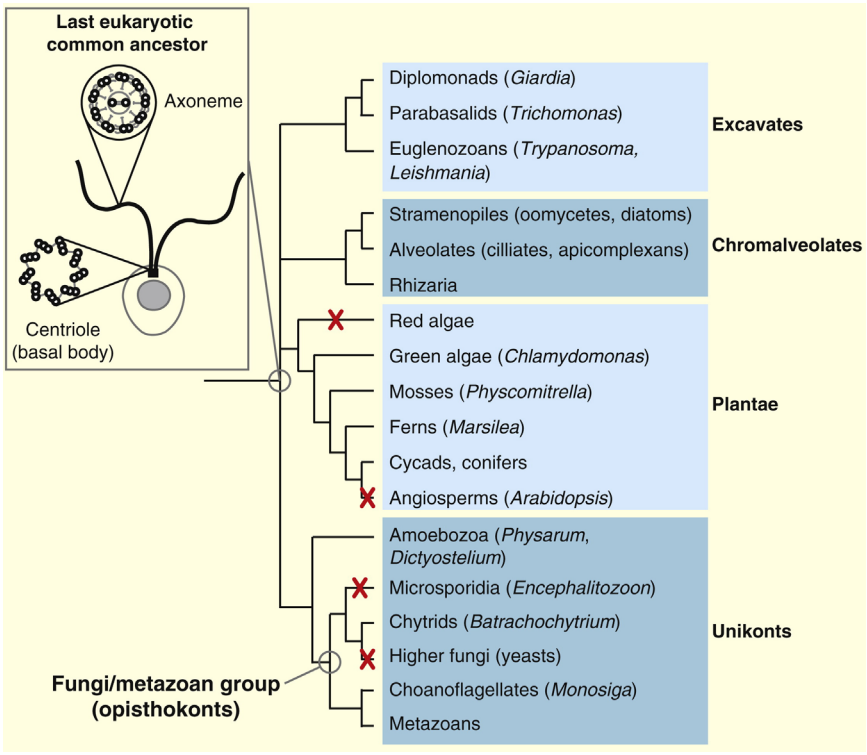


Fig. 1.1: Phylogenetic Tree of Eukaryotes: Depiction of Taxa in the phylogenetic tree of eukaryotes. Taxa in which the centriole was lost are indicated by a *red* cross. The branching off from the last eukaryotic common ancestor is indicated by a *grey* circle. Reprinted from Current Biology, 20, J. Azimzadeh and W. F. Marshall, *Building the centriole.*, R816-25, Copyright 2010, with permission from Elsevier. [16].

1.2 Centrosome and Centriole Function

Centrioles were first discovered in dividing cells as crucial components of the animal centrosome - an organelle found at the spindle poles and referred to as the primary MTOC [16, 17]. The MTOC also plays a role in signaling [18, 19], control of cell cycle progression [20], asymmetric cell division, establishment of cell polarity and DNA damage recognition [21–24].

1.2.1 Cell Division

Boveri described the centrosome as the 'true division organ of the cell' in 1887, identifying the apparent connection between the separation of chromosomes and the organisation of the mitotic spindle by the MTOC [1]. Nucleation of the microtubule network that forms the mitotic spindle is primarily linked to the PCM, whereas the centriole is believed to support correct PCM organisation [25]. During the centrosome maturation process in mitosis the PCM size increases. This leads to increased microtubule-nucleation and the formation of radially expanding astral and spindle microtubules (as reviewed in [26]). Whether it is the centrosomes that control spindle formation or the spindle formation that ensures the distribution of centrosomes into daughter cells is under debate [27]. A dependence on the presence of centrioles for successful cell division was shown for the division of *Xenopus leavis* eggs [28], fertilised embryos of *Caenorhabditis elegans* (*C. elegans*) [27, 29] and in the development of *Drosophila melanogaster* (*D. melanogaster*) [30]. Furthermore, the existence of centrioles at the spindle poles has also been shown to have a stabilizing effect on spindle formation [30, 31]. However, in general spindle formation is possible without the presence of centrosomes, but is error prone and less efficient [32]. In the absence of centrosomes, spindle formation can occur through control by a gradient of RanGTP in the vicinity of chromatin [33–35]. This gradient is part of a pathway that ultimately leads to the recruitment of the γ-tubulin complex and microtubule nucleation [36, 37]. Cell cycle progression without centrosomes is slowed down; an increased number of centrosomes on the other hand can lead to multipolar spindles and is linked to aneuploidy, a driver for tumorigenesis [4, 38].

1.2.2 Basal Bodies, Cilia and Flagella

After cell cycle exit, centrosomal centrioles may be converted into basal bodies [15]. As such they serve as the foundation of flagella and motile and primary cilia [12, 39–41]. Cilia and flagella are flexible membrane extensions which can be found on various cell types and occur on almost every cell in the human body [42]. Motile cilia and flagella enable the movement of unicellular organisms through fluids; their pulsed beating aids circulation of cerebrospinal fluid in the brain and spinal cord, helps to clear the respiratory tract and is linked to embryonic left-right asymmetry [43]. Structurally, these protein complexes are highly similar, however, whereas a single cell can feature hundreds of cilia, flagella are longer and are only found in small numbers on individual cells. Beating patterns also differ between cilia and flagella, reflecting their different roles [44]. Primary cilia, which are not involved in motility are essential for signal-receiving and propagating signaling cascades [45]. Thus centrioles are connected to a number of cellular functions in addition to cell division, such as cell motility and influential signaling pathways such as light-sensing and the hedgehog pathway which is crucial for cell differentiation, and mating pathways as in *Chlamydomonas reinhardtii* (*C. reinhardtii*) [20, 46–49].

The variety of processes in which centrioles are involved links errors in their structural appearance and duplication process to a number of diseases (as reviewed by [50, 51]). The impact the duplication process of these organelles has on cell function and disease is wide ranging, and is best illustrated by the variety of phenomena that can result from errors in centriole duplication.

1.3 Centrosomes and Centrioles in Health and Disease

The central role centrioles play in the cell cycle makes errors in their duplication a possible route for cancer development [38, 51–53]. Centrosome related abnormalities in mitosis can be found in a variety of cancer forms (as reviewed in [54]). The presence of centrosome aberrations in tumour cells has been established and their frequency has been linked to tumour progression [55, 56].

Centrosomal defects are typically classified into structural and numerical defects (as reviewed in [51] and [57]). The structural centrosome defects constitute either defects in centriole structure or in the formation of the PCM. Structural centriole defects in mammalian cancer cells are difficult to classify, as the small size of centrioles (0.2 - 0.5 μm length) makes accurate in-cell centriole length measurement difficult and the observation of increased PCM material could likewise be linked to an increased number of centrioles and thus be misinterpreted. Specific mutations of centriolar proteins have not yet been identified in cancer cells with structural centrosome defects, so their origin *in vivo* is still unclear [51, 58].

More often reported are the easier to detect numerical defects. Both overamplification and a lack of centrosomes can cause problems with cell division [4, 38]. Numerical defects could have a number of underlying causes, as the centriole duplication process is tightly regulated by various factors including protein expression levels, degradation and phosphorylation by the pathway controlling kinase zygote 1 (ZYG-1)/polo like kinase 4 (Plk4)/SAK [59–63]. The known tumour suppressor p53, which is indirectly linked to the control of Plk4 expression [64, 65], is often defective in cancer cells, which allows for proliferation even when abnormal centriole duplication activity occurs [61]. Centrosome overamplification causes chromosome segregation errors and asymmetric

cell division. Mammalian cells can form spindles and go through cell division even in the absence of centrosomes, however permanent centrosome loss leads to chromosome instability and aneuploidy [32].

Only few specific mutations of the centrosome related genes have been identified as causal factors in tumour cells [58]. Consensus studies of cancer genes across many tumour types did not find significant levels of alterations in centrosomal genes to be present [66–68]. One exception is the gene for CEP76, which encodes a protein that has been found to drive aberrant centriole amplification when depleted [69] and was identified in a focal deletion peak in acute myeloid leukemia and showed nonsense mutations in other tumour types [67]. The absence of striking alterations in genes encoding the key components of centriole and the centrosome could indicate that regulatory mechanisms often derailed in tumourgenesis, which have an influence on the progression of centriole duplication, cause the observed centrosomal abnormalities [58]. Another possible explanation would be that many different centrosomal genes are altered over the studied tumour cells, so that no one mutation is present to surface statistically. The culprits for centrosomal defects in cancer are still to be identified.

Primary recessive microcephaly is a genetic neurodevelopmental disorder that can be traced to single gene mutations many of which have been identified in proteins directly involved in centriole duplication (as reviewed in [50, 52, 53, 70]). Specifically, mutations in genes encoding for the centriole proteins SCL/TAL1 interrupting locus (STIL), centrosomal P4.1-associated protein (CPAP), Cep152, Cep135 or the mitotic spindle pole proteins WDR62 or ASPM are present in individuals with this syndrome. The mutations cause defects in the proteins in question, and either result in errors in protein interaction, or hinder their degradation, ultimately leading to a deregulation of centriole numbers. Cell division in cells with deregulated numbers of centrioles has

been shown to take longer due to a stress induced elongation of G1 phase mediated by p53 and p38 activity [50, 53, 71–73]. It is assumed that neuronal development is especially sensitive to the timing of cell division and that this leads to a reduced number of neuronal cells, which would account for decreased brain size, the main phenotype observed for this disease.

Ciliopathies are another group of genetic diseases linked to mutations in genes connected to centrioles (reviewed in [74]). This designation spans a variety of malfunctioning organs, mainly in the kidney, eyes, liver and brain, which are all linked to defects in cilia function or structure. It is assumed that the signaling function of cilia is most impacted in ciliopathies and likely to lead to the variety of possible symptoms. The genetic mutations that lead to cilia misformation are difficult to pin-point. Thirty genes have been identified with such mutations involved in ciliopathies; many of these genes are connected to centrosomal proteins (reviewed in [75, 76]).

Centriole duplication is a highly regulated process which occurs once per cell cycle. Errors in this process are linked to carcinogenesis, and mutations in proteins essential for the duplication process are implicated in brain diseases. Knowledge about the centriole duplication process is valuable for our understanding of fundamental biology and also influences our ability to address various conditions from cancer to neuronal diseases.

1.4 Centriole Structure

Centrioles are large cylindrical protein complexes that differ in size and morphology between organisms and cell types (Figure 1.2 and 1.3) [16, 77]. The majority of pro-

teins contained in the centriole complex is tubulin in the form of microtubules, which assemble in a cylinder by an arrangement of single, doublet or triplet microtubule blades with nine-fold radial symmetry (Figure 1.3 B). The microtubule arrangement comprising the centriole wall, as well as the length and diameter of the organelle, is dependent on the organism and cell type. Canonical centrioles such as occur in vertebrates have microtubule triplets (Figure 1.3 C). The triplet of microtubules is formed by a complete microtubule, the A-tubule, with two partially assembled microtubules (B and C-tubule) attached to it (Figure 1.3 B). Nine of these triplets assemble into a formation with a diameter of 100 to 250 nm in vertebrates and in vertebrates the full length of the organelle can range from 150 to 500 nm depending on the species and cell type in which it is found [16, 77].

The centriole cylinder features two distinct ends, a proximal and a distal end, which in the case of the vertebrate centriole is decorated with distal and subdistal appendages (Figure 1.2 A and 1.3 A and C) [78]. Vertebrate centrioles can be described as three parts along the longitudinal axis: a distal end, with microtubule doublets, a central part where microtubules are triplets, and a proximal end with microtubule triplets surrounding a central protein complex called the cartwheel (reviewed in [16, 77, 78]).

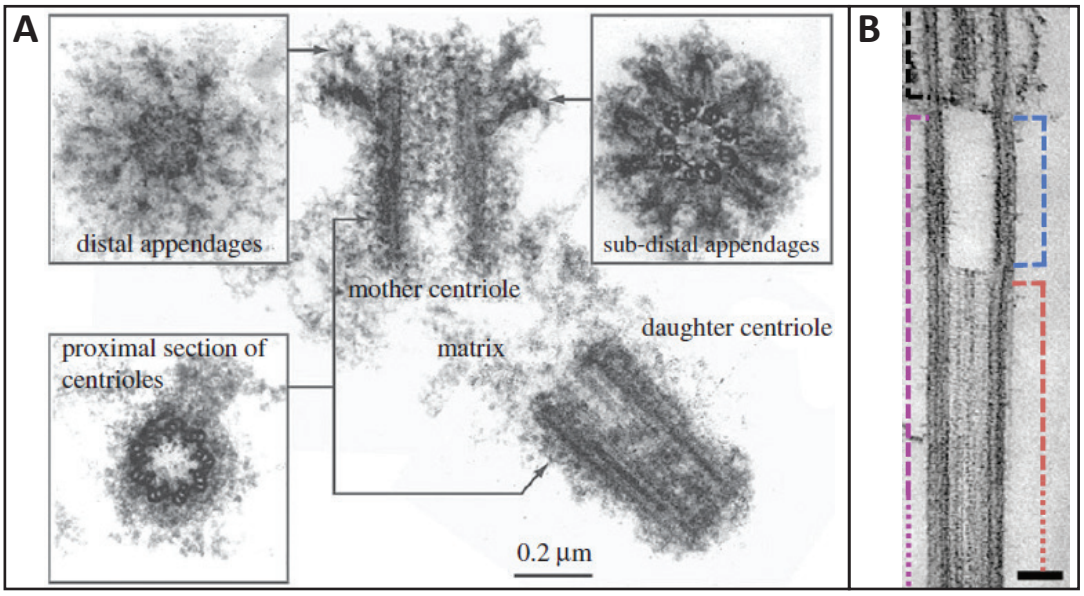


Fig. 1.2: Electron Microscopic Images of Centrioles: (A) Cross section and longitudinal section of entrosomal centrioles isolated from human cells. Mother centriole with distal and sub-distal appendages can be seen, as well as daughter centriole. Scale bar represents 200 nm. Figure adapted from [79] - Copyright Journal of cell science by Company of Biologists. Reproduced with permission of COMPANY OF BIOLOGISTS LTD. in the format Republish in a thesis/dissertation via Copyright Clearance Center. Original data from [80]. Reprinted from Journal of Structural Biology, 108, M. Paintrand, M. Moudjou, H. Delacroix, M. Bornens, *Centrosome organization and centriole architecture: Their sensitivity to divalent cations*, 107-128., Copyright 1992, with permission from Elsevier. (B) The distal end of a *Trichonympha* centriole (indicated with *purple* dashed line) as seen in EM, with part of the flagellum indicated by *black* dashed line). The distal cartwheelless region is indicated by a *blue* dashed line and a part of the cartwheel containing proximal region is indicated with a *red* dashed line. Scale bar represents 100 nm [81]. Reprinted from Current Biology, 23, P. Guichard, V. Hachet, N. Majubu, A. Neves, D. Demurtas, N. Olieric, I. Fluckiger, A. Yamada, K. Kihara, Y. Nishida, S. Moriya, M. O. Steinmetz, Y. Hongoh, P. Gönczy, *Native Architecture of the Centriole Proximal Region Reveals Features Underlying Its 9-Fold Radial Symmetry*, 1620 -1628, Copyright 2013, with permission from Elsevier.

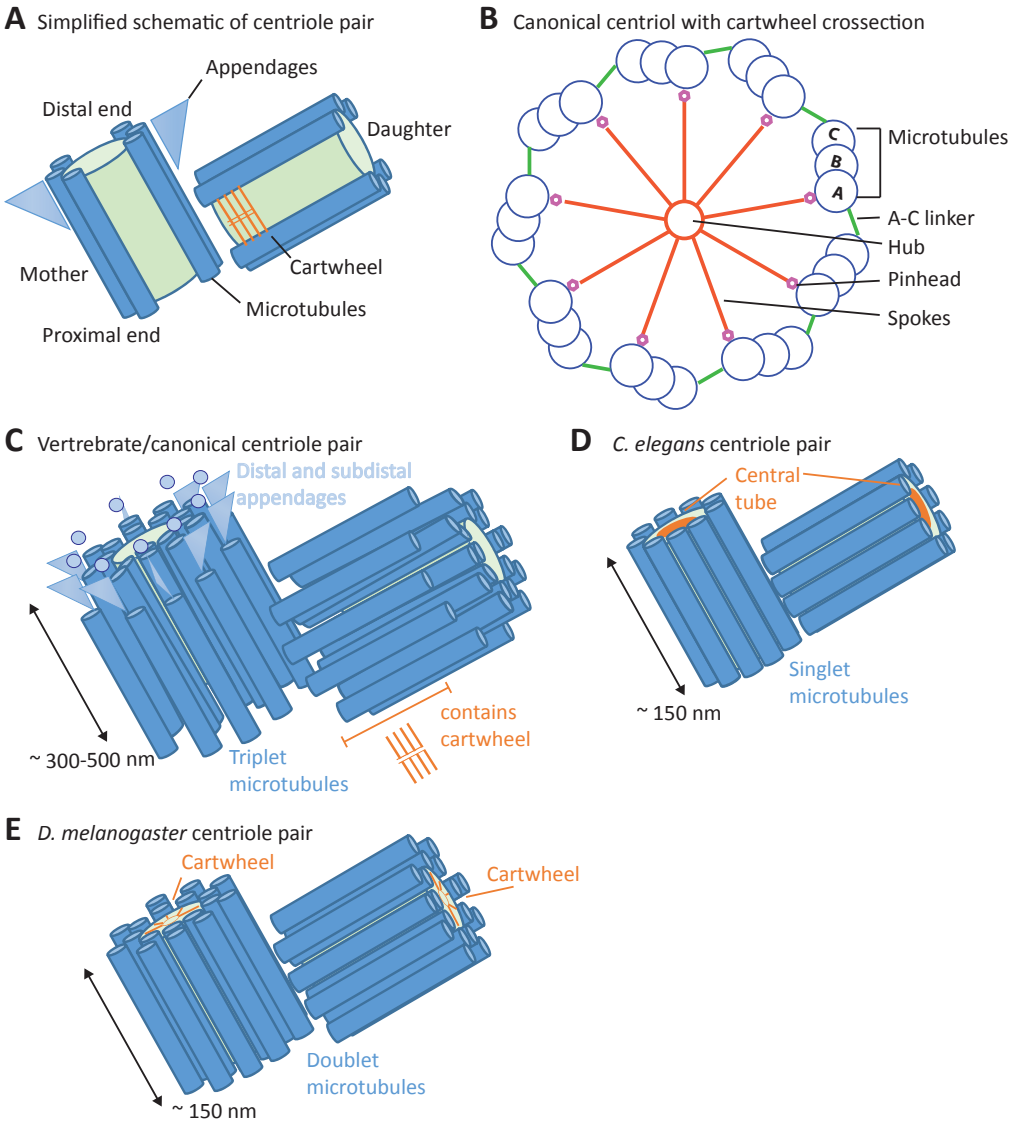


Fig. 1.3: Architecture of Centrioles and Cartwheels: (A) Simplified schematic depiction of a vertebrate centriole pair. Mother centriole with appendages and daughter centriole at proximal end perpendicular to the mother centriole are shown. Microtubules (*blue*) forming the centriole cylinder are not fully depicted. Cartwheel stacks are indicated in *red*. See (C) for full schematic depiction. (B) Schematic representation of a top down view on the cartwheel surrounded by nine-fold symmetric microtubule triplets as present in centrioles. Spokes and hub of cartwheel are depicted in *red*. The pinhead is depicted in *purple*. The A-C linker is depicted in *green*. (C) Schematic representation of a canonical vertebrate centriole pair with triplet microtubules, distal doublet microtubules and distal and subdistal appendages. The postition of the cartwheel of the daughter centriole is indicated. (D) Schematic depiction of *C. elegans* centriole pair, with singlet microtubules and the central tube indicated in *red*. (E) Schematic depiction of *D. melanogaster* centriole pair with doublet microtubules and cartwheel indicated in *red*.

In electron microscopy (EM) and cryo-electron tomographic images of centrioles, a density of unknown composition has been identified which seemingly links the microtubule triplets, and hence is called the A-C linker [81, 82] (Figure 1.3 B). The centriole cylinder is not just a hollow tube; sections of both the proximal and the distal end show densities within the lumen. Whereas little is known about what constitutes the diffuse densities found in the distal lumen, the proximal lumen contains a striking structure with a central hub and nine spokes radiating from the centre to the microtubule wall, fittingly termed the cartwheel (Figure 1.3 B). Side-slices of centrioles reveal that the cartwheel can have differing lengths and appears as a layered structure (Figure 1.2 B). The layered appearance observed in some species is assumed to suggest that the cartwheel consists of stacked rings, however this model has so far only been shown to be true for *Trichonympha agilis* centrioles [81, 83]. The length and appearance of the centriole is species specific. In mammalian cells cartwheel structures only occur in the daughter centriole; the cartwheel gets lost during maturation (Figure 1.2 A) [84].

As mentioned previously, the appearance of the centriole can differ between species and cell types. In addition to the canonical centriole found in vertebrate cells, different variations on the complex can be observed amongst model organisms. Centrioles found in *C. elegans* embryos have singlet microtubule blades, are only $\sim$ 150 nm in length, $\sim$ 100 nm in diameter, and do not contain a canonical cartwheel. Instead, these centrioles have been found to contain a so called "central tube" with a diameter of $\sim$ 60 nm [85] (Figure 1.3 D). *D. melanogaster* embryo centrioles feature doublet microtubules and are also usually shorter than the vertebrate centrioles with $\sim$ 150 nm length [78]. Both mother and daughter centrioles in *D. melanogaster* cells contain cartwheels, but have no appendages (Figure 1.3 E). The basal bodies of *C. reinhardtii* have triplet microtubules that turn into doublet-microtubules towards the distal end, as seen in vertebrate centrioles [12]. Mature *C. reinhardtii* basal bodies are $\sim$ 400

nm in length and probasal bodies are $\sim$ 86 nm in length. Both types have cartwheels and a sort of distal appendage can be observed [86]. The cartwheel appearance in *C. reinhardtii* changes during cell cycle progression [12, 86]. The cartwheel length increases in probasal bodies from 107 nm to 163 nm during telophase in mitosis.

Another remarkable centriole structure can be found in *Trichonympha agilis*. The basal bodies in this organism contain canonical triplet-microtubules and feature a distal region with appendages very similar in length to the canonical centriole (Figure 1.2 B) [81, 87, 88]. However, the cartwheel filled proximal region of these basal bodies is very long and can bring the length of the complex to $\sim$ 4 μm [40, 81, 88].

1.4.1 The Cartwheel

The cartwheel has been found to be instrumental for the complex duplication process centrioles undergo and the symmetry of the centriole complex [83, 89–92]. It has extensively been studied; however, our understanding of the structural components of this subassembly remains limited. It is assumed that the centre of the cartwheel, the so called hub, is mostly formed by SAS-6, a widely conserved protein that oligomerises to form rings or spirals with α-helical spokes arranged with nine fold symmetry. SAS-6 protein from several species has been observed to form these oligomers *in vitro* [93–96] and the expression of SAS-6 has been shown to be required for cartwheel formation *in vivo* [84, 90]. Cryo-electron tomographic images of *Trichonympha agilis* centrioles show densities in the lumen of the SAS-6 hub, termed cartwheel inner densities (CID), which suggest that there might be more components to the cartwheel structure [81, 83]. A notable density connecting the hub-spokes to the microtubules is called the pinhead. SAS-4/CPAP and Cep135 are believed to be located in the pinhead structure [81, 97, 98], but structural analysis of the subdomains of these proteins has not yet yielded a conclusive assignment, so that the exact composition of the pinhead

remains unclear.

Electron microscopy studies on the centrioles of several organisms revealed the cartwheel structure to be of different sizes and to consist of what appears to be layers [86, 99] (exemplified in Figure 1.4 A).

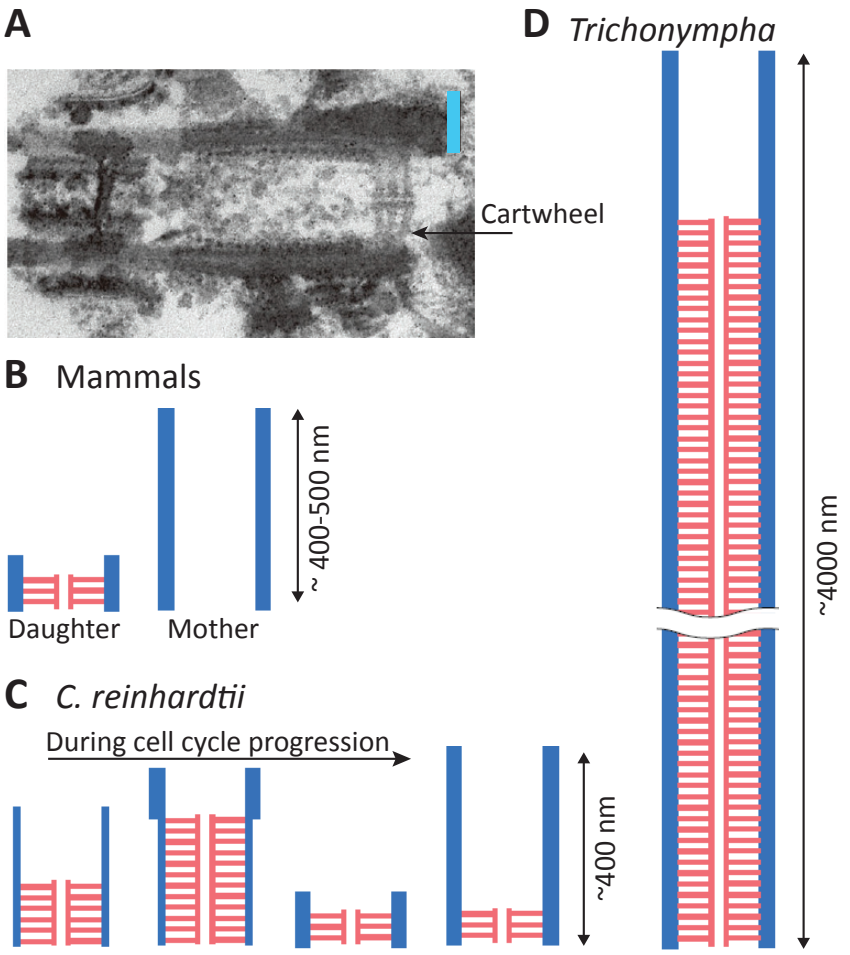


Fig. 1.4: Cartwheel Stacks in Various Organisms: (A) Electron microscopy image of resin embedded slice of *C. reinhardtii* centriole. The cartwheel is indicated with a *black* arrow. The scale bar corresponds to 100 nm and is shown in *light blue*. (B) Schematic depiction of cartwheel stacks as observed in mammalian centrioles. In the mature mother centriole the cartwheel is no longer present. (C) Schematic depiction of cartwheel stacks as observed in *C. reinhardtii*. The stack length in this organism changes during the cell cycle. (D) Schematic depiction of cartwheel stacks as observed in *Trichonympha*. The figure was adapted from [84] - Copyright 2014 M. Hirono. Published by the Royal Society under the terms of the Creative Commons Attribution License http://creativecommons.org/licenses/by/3.0/.

In vitro SAS-6 oligonmers have been observed to form stacks, though under unphysiological conditions [95, 96]. These stacks could explain the layered appearance of the cartwheel. The cartwheel can occupy $\sim$ 100 nm of a $\sim$ 400-500 nm long mammalian centriole [100, 101]. Electron microscopy images of *Trichonympha* centrioles have revealed that the $\sim$ 4 μm long centrioles contain a distal section of $\sim$ 420 nm, which corresponds to the length of the cartwheelless distal section in centrioles found in *C. reinhardtii* and a proximal extension occupied by the layered cartwheel that makes up for $\sim$ 90 % of the length of this extended centriole [81, 83]. Further investigation of the *Trichonympha* centriole using cryo-electron tomography revealed that the layers observed consist of a stacked cartwheel in this organism [81, 83]. Cartwheel stacks are assumed to be the canonical appearance of the cartwheel assembly. However, there are variations between species and it has yet to be shown that cartwheel stacks are the general appearance of the cartwheel. One such variation can be observed in *C. elegans* centrioles, where instead of the hub-and-spoke cartwheel, a "central tube" of $\sim$ 60 nm diameter was observed [85]. Here also the presence of SAS-6 was required for the appearance of the structure, so that it is assumed that this is a variation of a SAS-6 oligomer [85, 102]. Crystal structures of *C. elegans* SAS-6 have been used to reconstruct such a SAS-6 oligomer and it has been found that the angles present in the crystal structure would allow for a spiral of *C. elegans* SAS-6 proteins to form, the dimensions of which could fit into the central tube observed [103]. Furthermore, recombinant SAS-6 was seen to form long thin filaments in electron micrographs [103]. The SAS-6 spiral has not been observed *in vivo*, but it stands to reason that other organisms might have adapted alternatives to the stacked cartwheel structure.

The size of the cartwheel differs in cell cycle stages (as reviewed in [84], see Figure 1.4). Whereas the *Trichonympha* cartwheel stays intact after centriole maturation, in mammalian centrioles cartwheels are lost in the mother centriole (Figure 1.4 B).

In *C. reinhardtii* the cartwheel length is 7-10 layers in the developing centriole and reduces to 2-4 layers when mature (Figure 1.4 C) [12, 86]. The distance between the layers of the cartwheel stack also differs between species. Stacks of *Leishmania major* (*L. major*) have a narrow 4.3 nm periodicity in the crystallographic structure [95], whereas the *Trichonympha* cartwheel layers have an 8.5 nm periodicity [81]. However, the *L. major* SAS-6 stacks seen in the crystal structure form under unphysiological conditions and their spacing might be an artefact. Another noteworthy feature of the stacked cartwheel observed in the structural reconstruction from cryo-electron tomography of the *Trichonympha* cartwheel is that the cartwheel spokes seem to merge in between layers of the cartwheel stack at $\sim$ 20 nm from the centre of the hub [81, 83]. This creates an additional periodic spacing of $\sim$ 17 nm, which comes close to the height of two tubulin dimers ($\sim$ 16 nm). This might be another piece in the puzzle of the mechanism that enables SAS-6 to scaffold the nine-fold symmetric microtubule cylinder.

1.5 Centrosome Duplication and Centriole Assembly

Centrosomes need to duplicate once per cell cycle, as one of the two centrosomes is passed onto the daughter cell during cell division. Since centriole number needs to be tightly controlled to avoid errors in the cell division process, it is not surprising that centrosome duplication is a stringent pathway [104]. Conventionally, the duplication of centrosomes occurs in a semi-conservative manner, where a mother centriole must be present to facilitate the assembly of the daughter centriole (Figure 1.5). The daughter is primed at the proximal end of the mother centriole, where the daughter grows perpendicular to the mother. How the location of the duplication site and the creation

of only one daughter are controlled is not entirely understood, but it seems clear that control mechanisms are important for successful duplication.

The underlying principle of the duplication pathway of centrioles is conserved throughout most eukaryotes [105–108]. Dividing cells need to duplicate centrioles once per cell cycle. The process begins in G1-S phase transition with the formation of a procentriole or daughter centriole that grows perpendicular from the proximal end of the so called mother centriole. In S and G2 phase the procentriole elongates. During this stage and until the end of mitosis, the procentriole is physically linked to the mother centriole. Centriole disengagement is performed in telophase/G1 phase by a number of factors including separase, which also plays a key role in sister chromatid segregation [105, 109, 110].

Both centrioles in the centrosome serve as mother centrioles, resulting in a full set of four centrioles per cell prior to mitosis. The number of centrioles produced during duplication is strictly controlled, so that only one daughter centriole can develop per mother centriole. Exactly how this regulation is achieved is unknown.

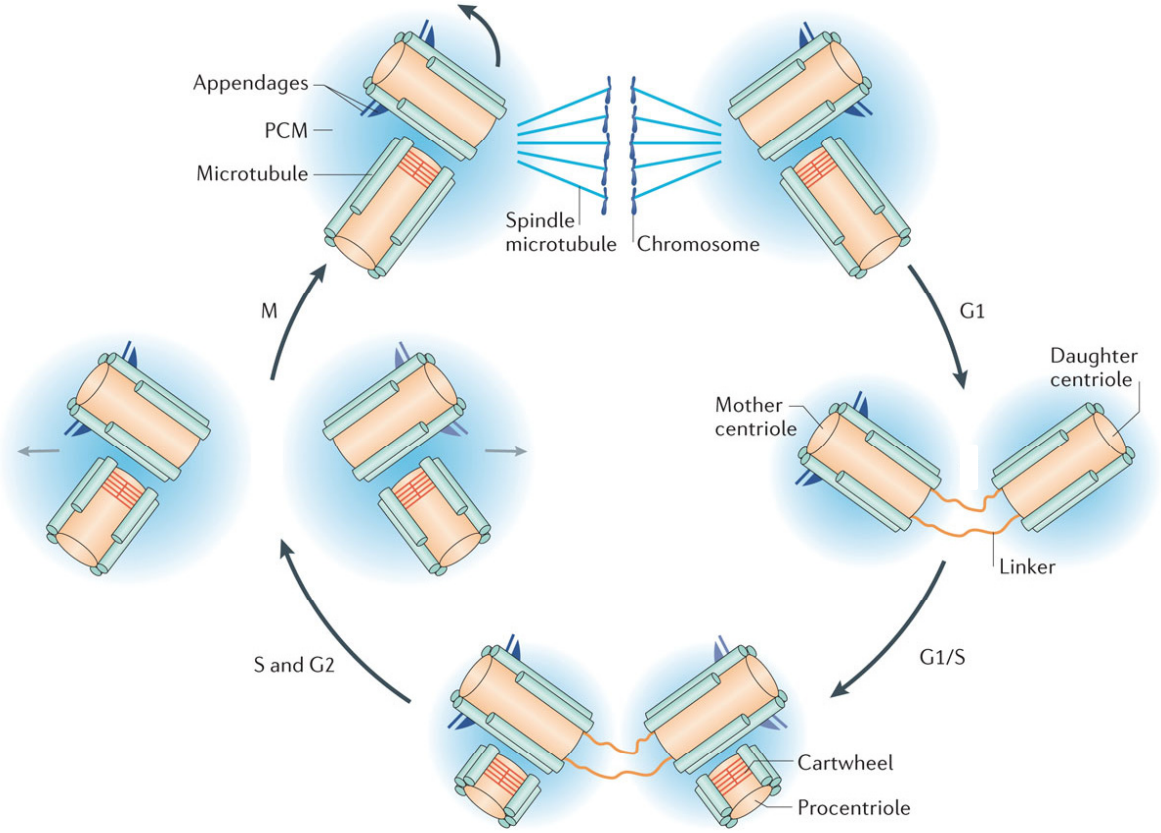


Fig. 1.5: Centrosome Duplication in the Mammalian Cell Cycle: The centrosomal centriole pair is depicted as a simplified schematic (see Figure 1.3 A for more detail). Centrosomes localise to the spindle poles during cell division. After successful division, only one centrosome remains. Both centrioles within the centrosome serve as sites of centriole assembly, which starts at the onset of G1/S phase. Procentriole elongation and PCM recruitment continues into S and G2 phase. Finally the connection of the two original centrioles is severed by separases, completing centrosome assembly. Reprinted by permission from Springer Nature: Springer Nature, Nature Reviews Cancer, *Centrosomes and cancer: revisiting a long-standing relationship*, P. Gönczy, copyright 2015 [58].

The key players of the centriole duplication pathway were first identified in *C. elegans* [59, 111, 112] and subsequently their orthologues and functional equivalents were identified in most eukaryotes. Although the central principles of the duplication are conserved, there are subtle differences in the proteins involved between species (Figure 1.6).

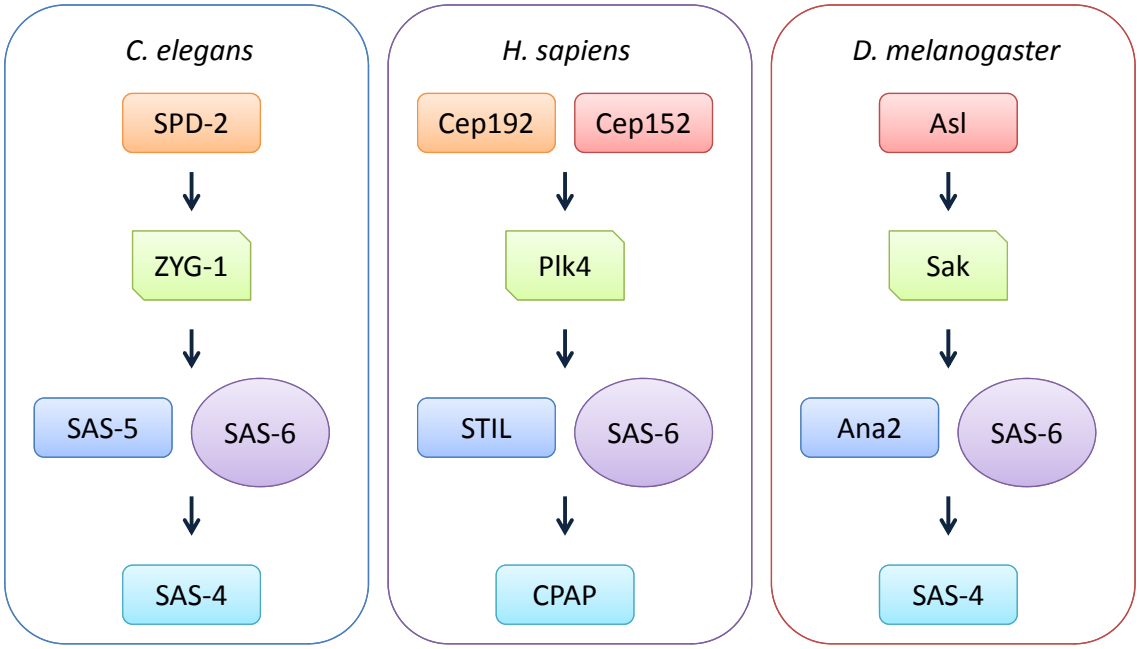


Fig. 1.6: Centriole duplication pathways: Schematic of the key proteins involved in centriole duplication. The basic components of the duplication pathway are conserved across species. Depicted are pathways found in *C. elegans*, *H. sapiens* and *D. melanogaster*.

Duplication begins with the priming of the duplication site on the mother centriole, where the first protein to localise is SPD-2 [113, 114] in *C.elegans* or Cep192 [115, 116]. In *D. melanogaster* the equivalent of SPD-2 has only been described to be involved in the assembly of the PCM [78, 117]; here, the first primer for centriole duplication is a protein called asterless (Asl) or Cep152 in mammalian cells. Mammalian cells require both Cep192 and Cep152 for successful priming of duplication [115]. Next, the pathway controlling kinase ZYG-1/Plk4/SAK (*C. elegans*/vertebrates/*D. melanogaster*) is recruited [59–63], which primes the duplication site further and seems to be involved in centriole number control. There are a number of proteins with which the kinase seems to interact, but all of its phosphorylation ligands may not have been identified yet. The localisation of ZYG-1/Plk4 to the duplication site leads to the recruitment of SAS-5 (*C. elegans*) [118]/anastral spindle 2 (ANA2) (*D. melanogas-*

ter) [118]/STIL (vertebrates) [119]. These are believed to be functional orthologues, with high sequence diversity. Only a shared 100 amino acid long region called the STIL/ANA2 (STAN) motif enabled the identification of these proteins as functional orthologues [120]. The proteins are able to homo-oligomerise and bind to SAS-6, SAS-4/CPAP and the pathway kinase ZYG-1/Plk4. The STAN motif in particular appears to be prone to modification by phosphorylation [121–124]. STIL has been observed to have a high turnover at the duplication origin, hinting at a potential shuttling function [119]. SAS-6 is recruited to the duplication site at a similar time as SAS-5/ANA2/STIL [125, 126]. The sequence similarity of SAS-6 between species is surprisingly low, however its structure and function are highly conserved [126]. The cartwheel is assumed to be formed by the oligomerisation of SAS-6 which gives rise to the characteristic nine-fold symmetry of centrioles and is essential for centriole duplication [126]. This structural component is strikingly visible in EM images [90, 127]. Lastly SAS-4 (*C. elegans*) [29, 128]/CPAP (mammalian cells) [129] and Cep135/Bld10 (vertebrates/*C. reinhardtii*) are recruited [98]. These proteins are likely to be structural components of centrioles, but are also implicated in the assembly of PCM. After maturation of the daughter centriole, the cartwheel is lost from the core of the centriole in some species, including *H. sapiens*. However, other species retain the cartwheel through the cell cycle, suggesting that the cartwheel is needed mostly for initial scaffolding of the microtubule wall.

Our understanding of the centriole duplication pathway is incomplete and ever evolving. Recently SAS-7 [130, 131] was identified as a protein possibly involved in the priming mechanism upstream of SPD-2 in *C. elegans*. More components of this pathway might be discovered in the coming years that will help close knowledge gaps. As it is, the interactions and affiliations of proteins so far identified are under intense investigation with new details surfacing rapidly. There is also evidence for *de novo*

formation of centrioles to be possible in S-phase in experimental settings. When centrioles are eliminated in experiments, they can be assembled *de novo* in subsequent cell cycles. However, the process has not been thoroughly elucidated: it seems to lack regulation and results in errors in centriole structure and function [132–135].

The doubling of the centrosome necessary for cell division, in which the canonical centriole duplication pathway is involved, is not the only occasion centrioles need to multiply. Cells can be equipped with several flagella and hundreds of cilia, all of which require basal bodies for membrane anchoring and ciliation. The mass multiplication of centrioles for basal body formation in these multiciliated cells (MCC) was assumed to be mediated by independent pathways: canonically formed procentrioles. These develop in a mother centriole dependent manner and acentriolar *de novo* formation of centrioles, which emerge from fibrous aggregates called deuterosomes [21]. The mother-centriole-dependent pathway resembles the typical centriole duplication and is specifically similar in phenotype to phenomena observed in cells overexpressing Plk4, SAS-6 or Cep152 [136–138]. Instead of a single daughter centriole, here multiple daughters nucleate from the mother centriole forming a rosette-like state [139]. The procentrioles are then released and dock to the membrane to aid ciliation as basal bodies. Deuterosomes have been studied as the origin of centriole formation in MCC for a while. Unfortunately, their composition and the mechanisms leading to deuterosome dependent centriole formation remain unclear [139]. The deuterosome is described as an electron-dense, ring-like structure [140] around which procentrioles form in a circular fashion. The deuterosome specific proteins Deup1 and Ccdc78 have been found to localise at this ring-like structure and be involved in deuterosome dependent centriole duplication [141, 142]. Interestingly, Deup1 is a paralouge of the protein Cep63, which is known to bind to Cep152 and can be found at the proximal end of mother

centrioles [143], suggesting a similarity between the canonical procentriole formation and the deuterosome dependent pathway [139]. Another indication of commonalities between the canonical duplication mechanism and the deuterosome-dependent centriole formation is the presence of SAS-6, Plk4 and Cep152 at the deuterosome [142]. A combined theory has emerged [144]. It concludes that the centrosomal daughter centriole could, in the presence of the regulating factor Deup1, form deuterosome and generate multiple rounds of procentriole seeding resulting in the typical procentrioles organised around the spherical deuterosome. This mechanism suggests that almost all basal bodies derive from the centrosomal daughter centriole.

1.6 Spindle Assembly Abnormal Protein 6 (SAS-6)

SAS-6 is the main component of the cartwheel structure that assembles at the start of centriole duplication. Its main purpose seems to be to create a scaffold that directs the nine-fold symmetry of the centriole [145]. It was among the first proteins identified in the centriole duplication pathway of *C. elegans*. A conserved sequence was then identified in the protein N-terminus, which was used to identify SAS-6 proteins in several other model organisms and termed "present in SAS-6" (PISA) [126] (Figure 1.7). Conserved features of SAS-6 include a globular N-terminal domain, a long coiled-coil domain and a C-terminal tail.

The function and structure of SAS-6 have been studied extensively. Nevertheless, questions remain about the mechanisms of the cartwheel formation, the regulation and control of SAS-6, and its interplay with other components of the centriole duplication pathway.

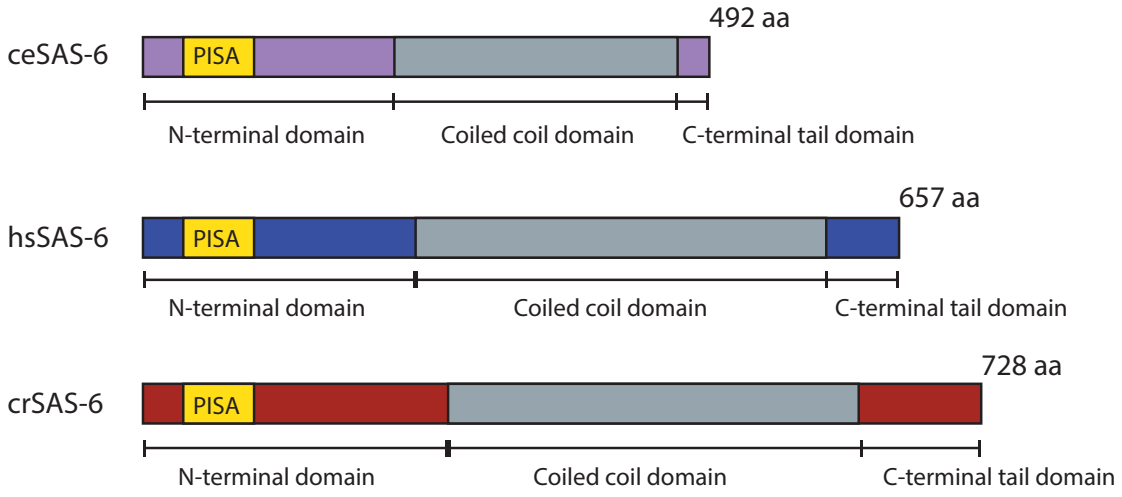


Fig. 1.7: SAS-6 Architecture: Schematic representation of SAS-6 architecture for the example of *C. elegans* SAS-6 (*purple*), *H. sapiens* SAS-6 (*blue*) and *C. reinhardtii* SAS-6 (*red*). The SAS-6 proteins consist of an N-terminal domain, a coiled-coil domain and a C-terminal tail. The locations of the conserved PISA domain are indicated in *yellow* and the predicted coiled-coil domains are indicated in *grey*.

1.6.1 SAS-6 Function and Regulation

The main function of SAS-6 appears to be in the formation of the cartwheel structure, which gives rise to the nine-fold symmetry of the centriole. SAS-6 siRNA experiments in *C. elegans* embryos showed that SAS-6 is essential for centriole duplication. Under SAS-6 siRNA, new centrioles cannot form [125]; this is also the case for human cells [126]. In *C. reinhardtii*, SAS-6 knockout was found to result in abnormal flagella formation, with non-circular basal bodies with incomplete microtubule walls being formed. These basal bodies lacked the central cartwheel in cross section images [90]. Loss of centriole duplication and errors in centriole elongation as well as symmetry were also observed in SAS-6 deletion mutants in *D. melanogaster* [127].

Localisation experiments show that SAS-6 localises to the mother centriole and remains stationary once recruited [145]. Required for SAS-6 localisation are ZYG-

1/Plk4 and the highly mobile SAS-5/ANA2/STIL, which both interact with the SAS-6 coiled-coil through a region in its mostly disordered C-terminus. This interaction is believed to be regulated through phosphorylation by ZYG-1/Plk4 [120, 121, 146–148]. Phosphorylated residues have been found at the C-terminus of SAS-5/ANA2/STIL. Conflicting evidence exists on phosphorylation of SAS-6. For example, SAS-6 has been shown to interact with ZYG-1, and *C. elegans* SAS-6 residue S123 has been suggested to be a potential site of modification through phosphorylation, however a mutational scan of possible phosphorylation sites in SAS-6 found no effect of modification of S123 [60, 149]. A major part of the regulation of SAS-6 is driven through cell stage dependent protein levels. SAS-6 levels are regulated by the KEN box present in its C-terminus, which mediates the anaphase-promoting complex related degradation [101, 150, 151].

1.6.2 Drug-target Potential of SAS-6

Centriole duplication plays a central role in cell cycle regulation. *In cellulo* experiments show how sensitive cell duplication control is to the presence and absence of centriolar proteins and to the regulation of their expression. Even though the necessity for regulated centriole duplication in human cell duplication can be overwritten with the deregulation of p53, commonly found in cancer cells [152], centriolar duplication control, specifically Plk4, has received a lot of attention as a potential cancer drug-target [153, 154]. Kinases are classical drug targets, making it likely that inhibitory compounds can be found with established screening methods. The tight connection between the regulatory effect of Plk4 on SAS-6 levels and localisation suggests that SAS-6 itself might be an interesting target. Kinase inhibitors need to be highly selective to avoid off-target effects, and Plk4 is involved in a variety of processes that are still being explored.

SAS-6 on the other hand has well defined functions during centriole duplication and in ensuring the structural symmetry of the centriole. However, the inhibition of SAS-6 assembly might not be sufficient to arrest cell division. The presence of SAS-6 is needed for cartwheel formation, yet the mechanism by which SAS-6 self-oligomerises was shown not to be influential enough to prevent a SAS-6 F131E mutant from facilitating centriole formation in human cell cultures during observations of *de novo* centriole formation [135]. Centrioles formed in this study where abnormal and defective but in p53 knockout background, cells could still duplicate.

Recent studies into *Plasmodium* parasites show that the presence and function of SAS-6 is crucial for flagella formation of male gamets [155]. In the absence of SAS-6, fertilisation and mosquito to mouse malaria transmission were greatly diminished. Further to this, a study in *Trypanosomes* showed that flagella formation was also disrupted in this organism and lack of SAS-6 resulted in cell death [156]. Assuming parasite SAS-6 self-assembly could be specifically blocked by a drug, centriole duplication might be deregulated enough to trigger cell-death pathways, or the centriole-like structures produced might eventually cause a slow down of cell duplication, due to accumulating abnormalities leading to functional defects. The rise in demand for unconventional drug targets, recent findings and future development might reveal hidden potential in inhibition of cartwheel assembly.

1.6.3 Architecture and Conservation

Like many centriole components, SAS-6 contains a colied-coil domain (Figure 1.7). A significant portion of the $\sim$ 500 amino acid long protein is reserved for the coiled-coil dimerisation (C-C dimer) region that forms the spokes emanating from the cartwheel hub. The remainder of the SAS-6 structure forms an N-terminal globular dimerisation

domain (N-dimer) and a disorderd C-terminal extension [93, 95, 103].

Structures of the SAS-6 N-terminal domain from several organisms have been solved by X-ray crystalography and are highly conserved. Several structures are depicted in Figures 1.9 and 4.1 to showcase their similarity. The structure consists of a seven stranded β-sandwich, capped by two α-helixes (α1 and α2) that form a helix-turn-helix motif. The N-terminal domain connects to the long α3-helix, which forms the coiled-coil domain.

Even though the structure of SAS-6 is highly conserved amongst species, the same is not true for its sequence [5]. A sequence alignment for the SAS-6 N-terminal domains of several organisms is shown in Figure 1.9 A. The essential cartwheel formation is facilitated by the presence of two dimerisation domains in SAS-6; a globular N-terminal domain and a long α-helix that forms coiled-coil dimers. The homo-oligomer this protein forms has a central hub, consisting of nine dimers of the globular head domain (Figure 1.8 A and C). The nine cartwheel spokes emanating from the hub are formed by nine coiled-coil dimers. These cartwheels appear in the centrioles of most model organisms and are thought to form stacks in the centriole core [90]. The cartwheel stacks are believed to be interconnected through further oligomerisation of SAS-6 coiled-coil strands between layers, as evidenced by cryo-electron tomographic reconstruction of *Trychonympha* cartwheel stacks [81].

Many SAS-6 variants identified have been seen to form cartwheel rings. These can be observed *in cellulo*, but also with recombinant protein in rotary metal shadowing electron microscopy, atomic force microscopy and X-ray crystal structures [93, 95]. However, there is evidence for SAS-6 adopting other oligomeric forms. *C. elegans* SAS-6 has been suggested to form a spiral that is believed to support the nine-fold microtubule scaffold by the formation of a double helix [103] (Figure 1.8 B).

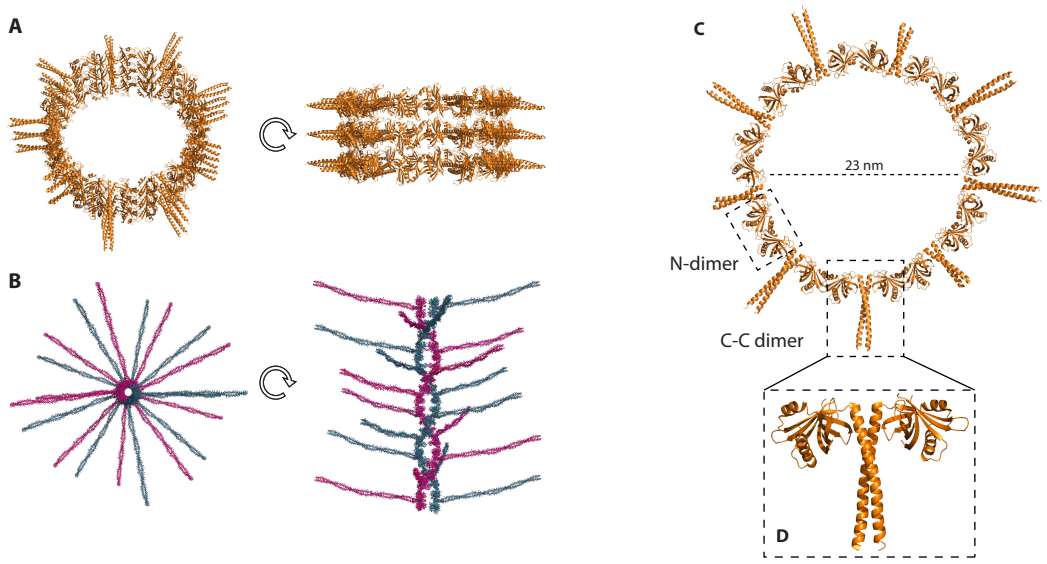


Fig. 1.8: SAS-6 Oligomer Structure, Ring and Spiral: (A) *L. major* SAS-6 ring stacks as seen in crystallographic structure PDB code 4CKP (3.45 Å resolution) [95]. (B) Model of the *C. elegans* SAS-6 spiral built using crystal structures PDB code 4GFC (2.85 Å resolution) and PDB code 4GFA (3.55 Å resolution) [103]. (C) *L. major* SAS-6 ring from structure PDB code 4CKP (3.45 Å resolution) [95]. The oligomer forms through nine C-C dimers coming together at the N-dimer interface and has a diameter of about 23 nm.

Recombinantly produced SAS-6 is able to form higher order oligomers on its own [94, 96, 157], which is also evident from EM and crystallographic structures. However, the individual dimerisation events are not sufficient to explain oligomer formation at the concentrations of SAS-6 likely to be present at the centriole. Computational simulations examining ring formation suggest that the ring-closure event is not likely to provide enough energetic advantage to explain spontaneous oligomerisation [158]. Further unknown scaffolding mechanisms or features so far overlooked in the SAS-6 protein itself might be involved. The cross-interactions of SAS-6 coiled-coils in cartwheel stacking might have an effect. Unexplained densities in the centre of the cartwheel visible in electron microscopy images taken from recombinant SAS-6 samples might also hint at further interactions not yet examined [81, 158]. Priming of the SAS-6

oligomer on the centriole duplication site or a microtubule/pinhead related scaffolding mechanism have been discussed in the past [84, 159, 160]. Evidence of cartwheel formation during *de novo* centriole formation with a SAS-6 mutant, which was shown not to be capable of forming higher order oligomers [135], *in vitro* seems to support the notion that there must be supporting mechanisms for SAS-6 oligomerisation we do not fully understand.

1.6.4 SAS-6 N-terminal Domain

The SAS-6 N-terminal domain crystal structure was first determined in 2011 [93] and since then structures of many variants of this domain have been solved (Figure 1.9). The wild type domain preferentially crystallises as a dimer (Figure 4.1). Structures from several organisms with various modifications have been solved. Available structures are listed in Table 1.1. It is immediately apparent from these that a key mediator of N-terminal dimerisation is the loop connecting β6-β7 (residues 151-158 in *C. elegans*), which positions a hydrophobic residue into a hydrophobic pocket in the dimer partner molecule. In *C. elegans* this hydrophobic key residue is I154, and in other species the equivalent residue is commonly phenylalanine (Figure 1.9 A and 4.1).

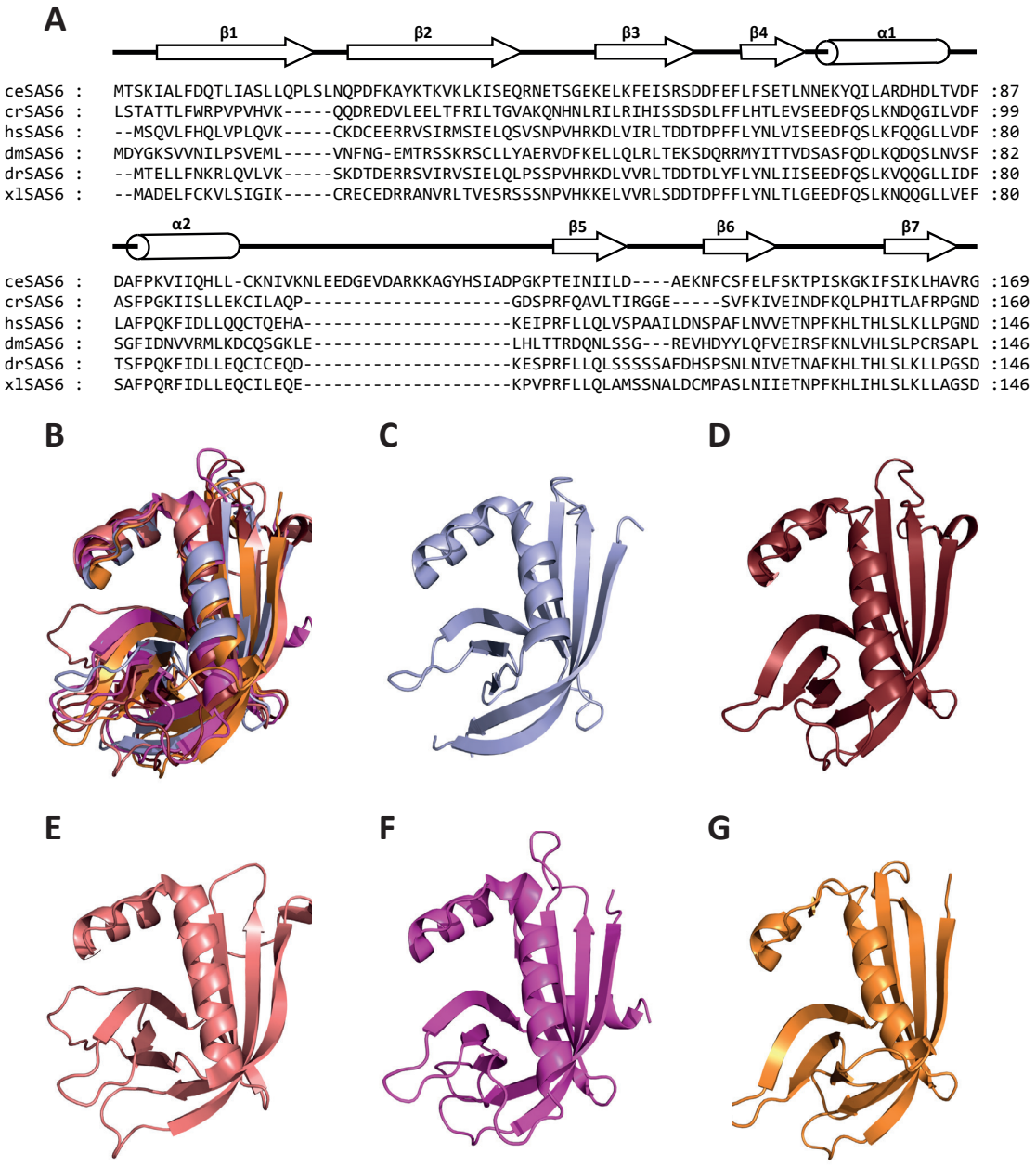


Fig. 1.9: Structural and Sequential Conservation of SAS-6 N-terminal Domain: (A) Sequence alignment of SAS-6 N-terminal domain sequences. (B) Aligned structures of SAS-6 N-terminal domains depicted individually in C-G. (C) *C. elegans*, PDB code 4G79 (1.8 Å resolution) [103]. (D) *C. reinhardtii*, PDB code 3Q0Y (2.1 Å resolution) [93]. (E) *D. melanogaster*, PDB code 5AL7 (2.92 Å resolution) [161]. (F) *Danio rerio* (*D. rerio*), PDB code 2Y3V (1.92 Å resolution) [94]. (G) *L. major*, PDB code 4CKM (2.15 Å resolution) [95].

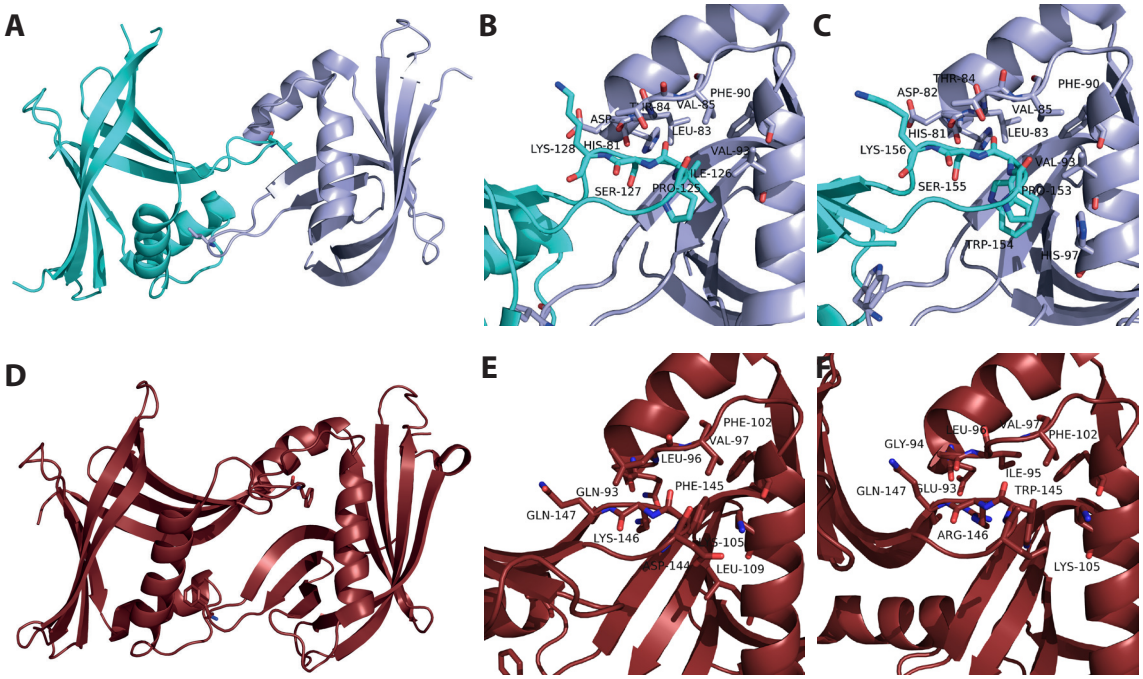


Fig. 1.10: SAS-6 N-dimer Interface Structures: (A) *C. elegans* N-dimer as present in structure PDB code 4G79 (1.8 Å resolution) [103]. (B) Interface residues in structure PDB code 4G79. (C) Interface residues in structure PDB code 4GEU with I154W mutation (2.65 Å resolution) [103]. (D) *C. reinhardtii* N-dimer as present in structure PDB code 3Q0Y (2.1 Å resolution) [93]. (E) Interface residues in structure PDB code 3Q0Y. (F) Interface residues in *C. reinhardtii* structure PDB code 4TO7 (1.25 Å resolution) [160]. Mutations present are Q93E, F145W and K146R; this combination of mutations is also called NN24.

The central role of the hydrophobic key residue was demonstrated in a number of *in vivo* experiments. Transgenic *C. elegans* embryos carrying SAS-6 I154E or I154G show errors in centriole duplication, and neither SAS-6 I154E nor I154G rescue centriole duplication in a SAS-6 depletion phenotype [93, 103]. In contrast, wild type SAS-6 or SAS-6 I154W show normal functioning centriole duplication and rescue SAS-6 depletion . The equivalent key residue in *D. melanogaster* mutated to an aspartate (F143D) exhibits similar errors in centriole duplication and again cannot rescue a SAS-6 depletion phenotype [161]. The equivalent SAS-6 aspartate mutant in *D. rerio* and human U2OS cells is not able to localise to centrioles [94]. In *C. reinhardtii*, SAS-6

F154D leads to non-functioning flagella [94].

In vitro experiments show that mutations of the key residue have a direct effect on the strength of the N-dimer formation. Comparing analytical ultracentrifugation (AUC) experiments from *C. elegans* SAS-6 wild type or SAS-6 I154E, oligomer formation is no longer detectable in the mutant [93]. The same is true for the *D. rerio* and human variant of SAS-6 with the equivalent aspartate mutation [94]. The *D. melanogaster* SAS-6 F143D protein shows reduced dimerisation in size exclusion chromatography (SEC) compared to wild type [161]. Fluorescence polarisation experiments with N-terminal domain fragments of *C. elegans* SAS-6 show that while the wild type fragment has a dimerisation affinity of 100 μM, an I154E mutation abolishes the interaction and the I154W variant dimerises with an increased affinity of 5 μM [103]. The key residue on loop β6-β7 and its hydrophobic interactions with the hydrophobic pocket are clearly central to the N-terminal dimerisation. However, other residues support the dimerisation.

.

Table 1.1: Table of SAS-6 N-terminal Domain Structures: Current published crystallographic structures of the N-terminal domain of SAS-6.

SAS-6 N-terminal Domain Structures				
PDB ID	Species	Construct	Resolution [Å]	Comment
3Q0Y	*C. reinhardtii*	1-159	2.1	N-dimer [93]
3Q0X	*C. reinhardtii*	1-226 F145E	3.02	N-domain with C-C dimer [93]
4TO7	*C. reinhardtii*	1-159 Q93E F145W K146R	1.25	N-dimer (NN24) [160]
4TQ7	*C. reinhardtii*	1-159 Q93E	2.64	N-dimer (NN10) [160]
4TPZ	*C. reinhardtii*	1-159 F145W	1.8	N-dimer (NN2) [160]
4TTZ	*C. reinhardtii*	1-159 G94E F145W Q147K	2.5	N-dimer (NN26) [160]
4TTY	*C. reinhardtii*	1-159 G94D F145W Q147R	1.9	N-dimer (NN25) [160]
4TTX	*C. reinhardtii*	1-159 G94C K146C	2.5	N-dimer (NN19) [160]
4TTW	*C. reinhardtii*	1-159 K105C F145C	1.2	N-dimer (NN18) [160]
4U2J	*C. reinhardtii*	1-159 Q93E F145W	2.2	N-dimer (NN27) [160]
4U2I	*C. reinhardtii*	1-159 Q93E G94D K146R Q147R	2.28	N-dimer (NN23) [160]
3PYI	*C. elegans*	1-168 S123E	2.1	N-dimer, loop 103-130 unresolved [93]
4GKW	*C. elegans*	256-342	3.3	C-C dimer fragment [147]
4GFC	*C. elegans*	1-187 I126G	2.85	N-domain with C-C dimer [103]
4GFA	*C. elegans*	1-187 I126G	3.55	N-domain with C-C dimer [103]
4GEX	*C. elegans*	1-168 S123E I154W	2.8	stabilised N-dimer [103]
4GEU	*C. elegans*	1-168 S123E I154W	2.65	stabilised N-dimer [103]
4G79	*C. elegans*	1-168 Δ 103-130	1.8	N-dimer without loop 103-130 [103]
2Y3V	*D. rerio*	1-156 S102N	1.92	N-dimer [94]
4CKP	*L. major*	97-320	3.45	N-domain with C-C dimer and ring formation [95]
4CKN	*L. major*	97-320 F257E	2.9	N-domain with C-C dimer [95]
4CKM	*L. major*	97-274	2.15	N-dimer [95]
5AL7	*D. melanogaster*	2-216 E36D F143D I207L	2.92	N-domain with C-C dimer [161]

The proteins involved in centriole duplication and especially SAS-6 have been a research focus for some time, as evidenced by the numerous crystallographic structures available (Table 1.1). Yet, open questions about the interactions of the proteins in this pathway and the molecular mechanisms of SAS-6 oligomerisation remain.

1.7 Aim of This Thesis

Cartwheel assembly is a central process in the centriole duplication pathway. Even though many advances in our understanding of the composition and structure of the cartwheel have been made, open questions remain.

This thesis focuses on the N-terminal dimerisation domain of SAS-6. It examines molecular mechanisms that govern the dimerisation in greater detail, and describes the path to developing inhibitory compounds for the dimerisation interaction, to be used ultimately for the study of the centriole duplication pathway.

Chapter 3 describes studies on a disordered loop region unique to nematode SAS-6 sequences and its effect on the dimerisation of this domain by biophysical techniques and molecular dynamics (MD) simulations.

Chapter 4 reports efforts to establish a high through-put fragment based virtual screening pipeline and compares results from this screen to results gained from a pharmacophore-like approach.

In Chapter 5, an account is given of the *in vitro* screening of small molecules by nuclear magnetic resonance (NMR) against *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6. Results of AUC experiments on the effect on oligomerisation with the compounds identified in the NMR *in vitro* screen are described.

Finally, Chapter 6 describes a novel crystallographic structure for the *H. sapiens* SAS-6 N-terminal domain. A combination of docking and molecular dynamics simulations are used to analyse the structure activity relationships of the identified inhibitory compounds, which are then used to propose a second generation compound library for future testing.

Chapter 2

Methods

"... it was the age of wisdom, it was the age of foolishness..."

Charles Dickens

A Tale of Two Cities

This chapter describes methods and protocols used for the cloning, purification and analysis of protein, as well as for experiments described in this thesis. If not otherwise stated all chemicals were purchased from Sigma-Aldrich.

2.1 Cloning and Mutagenesis

The cloning of *H. sapiens* SAS-6 constructs (Table 2.1) and site directed mutagenesis of these constructs are described here. Constructs of *C. elegans* and *C. reinhardtii* SAS-6 used in this thesis were not created by the author. These constructs were created for previously published work and were provided by Dr. Ioannis Vakonakis [103, 160].

2.1.1 Construct Design

Protein constructs (Table 2.1) were designed using available structural information where possible. Where no previous information was available, disorder predictions on the basis of amino acid sequence were created with the *metaPrDOS* server [162] and initial protein structure prediction server *Phyre2* [163]. The periodicity of coiled coil domains was predicted using the *Coils* prediction tool [164]. Multiple sequence alignments were created using *MUSCLE* and *Clustal Omega* and where necessary were manually edited according to available structural information [165, 166]. To view and edit DNA and amino acid sequences, *ApE* software [167] and the *Benchling* webserver (benchling.com version 2017) were used. The *ProtParam* tool was used to predict protein properties from amino acid composition, such as calculated molar extinction coefficient, isoelectric point and molecular weight [168]. The *PISA* webserver was used to identify residues in dimerisation interfaces [169].

Table 2.1: Protein Constructs

Protein Constructs				
Construct	Organism of Origin	Plasmid	Tag and Cleavage Site	Reference
ceSAS-6_N Δ 103-130 I154E (ceSAS-6 1-168 Δ 103-130 I154E)	*C. elegans*	pHis	His,Thrombin	[103]
ceSAS-$6_{N\ CC}$ I154W (ceSAS-6 1-215 S123E I154W)	*C. elegans*	pHis	His, thrombin	[103]
crSAS-6_N F145E (crSAS-6 1-159 F145E)	*C. reinhardtii*	pSTCm1	His, thrombin	
crSAS-$6_{N\ CC}$ (crSAS-6 1-226)	*C. reinhardtii*	pSTCm1	His, thrombin	[160]
crSAS-$6_{N\ CC}$ NN24 (crSAS-6 1-226 Q93E F145W K146R)	*C. reinhardtii*	pSTCm1	His, thrombin	[160]
hsSAS-6_N F131E (hsSAS-6 1-152 F131E)	*H. sapiens*	pfoalt2	His,HRV 3C	NEW
hsSAS-6 1-199	*H. sapiens*	pfoalt2	His,HRV 3C	NEW
hsSAS-6 1-199 F131W	*H. sapiens*	pfoalt2	His,HRV 3C	NEW
hsSAS-$6_{N\ CC}$ (hsSAS-6 1-205)	*H. sapiens*	pfoalt2	His,HRV 3C	NEW
hsSAS-$6_{N\ CC}$ F131W (hsSAS-6 1-205 F131W)	*H. sapiens*	pfoalt2	His,HRV 3C	NEW

Primers were designed manually using the *Benchling* webserver (benchling.com version 2017) and the New England Biolabs (NEB) *Tm calculator* web tool (v1.9.6 and v1.9.7) (Table 2.2). For polymerase chain reaction (PCR), Phusion polymerase (NEB) with the supplied buffer was used (see protocol Table 2.3). PCR products were purified after agarose gel electrophoresis using Zymo Research Zymoclean Gel DNA Recovery kit and joined by Gibson assembly using the NEB Gibson assembly master-mix. The Gibson mixture was then subjected to DpnI (NEB) digest in NEB cutsmart

Table 2.2: Primers

Primers				
Construct Name	Primer Name	Sequence	Annealing Temperature	Comment
hsSAS-6 1-152, 1-199 and 1-205	hsSAS6 1 FWD	TCTGTTCCAGGGGCCCATGAGCCAAGTG CTGTTCCACCAACTA	56 °C	Gibbson assembly
hsSAS-6 1-152	hsSAS6 152 REV	TTAGCAGCCGGATCTCTTAAAATTTCTT TATCTCCACATCATTTCCAGGTAAAAG	56 °C	Gibbson assembly
hsSAS-6 1-199	hsSAS6 199 REV	TTAGCAGCCGGATCTCTTACCATTCATT CCGTAACTTATCTAATTCTTGTTT	56 °C	Gibbson assembly
hsSAS-6 1-205	hsSAS6 205 REV	TTAGCAGCCGGATCTCTTAGGCTGCTGT ATGTGACGCCCATTC	56 °C	Gibbson assembly
hsSAS-6 F131E	hsSAS6 FE QC FWD	GTGGTAGAGACAAATCCTGAAAAGCATC TTACACACCTC	60 °C	Quick change
hsSAS-6 F131E	hsSAS6 FE QC REV	GAGGTGTGTAAGATGCTTTTCAGGATTT GTCTCTACCAC	60 °C	Quick change
hsSAS-6 F131W	hsSAS6 FW QC FWD	GTGGTAGAGACAAATCCTTGGAAGCATC TTACACACCTC	60 °C	Quick change
hsSAS-6 F131W	hsSAS6 FW QC REV	GAGGTGTGTAAGATGCTTCCAAGGATTT GTCTCTACCAC	60 °C	Quick change

Table 2.3: Polymerase Chain Reaction Protocol

Polymerase	Protocol	Comment
Phusion	98°C 30 s	Initial denaturation
	27 cycles	
	98°C 10 s	Denaturing step
	(45 - 72)°C 20 s	Primer annealing temperature dependent on primer composition
	72°C 20 s/kb	Elongation
	end cycle	
	72°C 5 min	Final extension
	4°C hold	Storage

buffer for 4 h. A 1 μl sample of the mixture was then transformed with *Escherichia coli* (*E. coli*) competent cells (see description of heat-shock transformation in Section 2.1.1.1). For Quick change mutagenesis, PCR products were also digested using DpnI and purified using the Qiagen QIAquick PCR purification Kit.

2.1.1.1 Heat-shock Transformation

Plasmid constructs were transformed into competent *E. coli* cells (Table 2.4) using heat-shock treatment. For this 200 μl of competent cells were thawed on ice and 1 μl

to 2 μl of DNA added. The mixture was incubated for 20 min on ice, heat-shocked for 45 s in a 42 °C water bath and kept on ice for 2 min. 800 μl of warm super optimal broth with catabolite repression (SOC)-medium were added and incubated, shaking at 37 °C for 60 min. The suspension was then spun for 30 s at 8000 rpm and the supernatant discarded. The cells were resuspended in the remaining fluid and spread for selection on a Luria Bertani medium (LB)-agarose plate containing antibiotics. The plate was incubated at 37 °C overnight and checked for colonies the next day. If commercial ultra competent DH5α cells (NEB) were used, the transformation protocol was adjusted to include 30 min on-ice incubation, 30 s of heat-shock at 42 °C. Colonies were then picked and used to inoculate 3 ml of LB medium with added antibiotic over night at 37 °C. These cultures were then used for plasmid purification using the Qiagen QIAprep Spin Minprep kit. Plasmid concentration was estimated by absorbance at 260 nm on a NanoDrop 2000c (Thermo Scientific). A sample of the purified DNA was sent for commercial sequencing at Eurofins Genomics.

Table 2.4: Bacterial Strains

Strain	Genotype	Application
Lemo21 (DE3)	fhuA2[lon] ompT gal(λ DE3) [dcm] ΔhsdS7pLemo (Cam^R)	protein expression
DH5α	F$^-$ endA1 glnV44 thi-1 recA1 relA1 gyrA96 deoR nupG ϕ80dlacZΔM15 Δ(lacZYA-argF)U169, hsdR17(r$_K^-$ m$_K^+$), λ^-	cloning and plasmid amplification

2.2 Protein Production and Purification

The protein required for work described in this thesis were expressed and purified according to the protocols described in this section. The expression and purification conditions were optimised for the constructs listed in Table 2.5. The purification

protocol was based on the conditions published for the purification of ceSAS-6_N Δ 103-130 I154E [103] and altered according to the results of the optimisation process.

Table 2.5: Molecular Weight of Protein Constructs: Values calculated using *Protparam* [168].

Molecular Weight of Protein Constructs	
Construct	Mw [kDa]
ceSAS-6_N Δ 103-130 I154E	16.22
ceSAS-$6_{N\ CC}$ I154W	24.54
crSAS-6_N F145E	17.92
crSAS-$6_{N\ CC}$	25.66
crSAS-$6_{N\ CC}$ NN24	25.72
hsSAS-6_N F131E	17.57
hsSAS-6 1-199	17.6
hsSAS-6 1-199 F131W	17.63
hsSAS-$6_{N\ CC}$	23.6
hsSAS-$6_{N\ CC}$ F131W	23.63

2.2.1 Protein Expression

2.2.1.1 Test Expressions

Conditions for soluble protein expression were optimised with small scale test expressions. Temperature and length of incubation after induction of protein expression, as well as the strain of *E. coli* can have an effect on expression levels and solubility. For expression tests, 100 ml of LB medium with added antibiotics were prepared for each test sample and inoculated with a single colony at 37 °C until optical density $(OD)_{600}$ at 600 nm of ≥ 0.6 was reached. The batches were then incubated at 37 °C or 18 °C and

protein expression was induced with 0.25 mM isopropyl β-D-1-thiogalactopyranoside (IPTG). Cells were then incubated for 4 h at 37 °C or over night at 18 °C.

Samples of 1 ml of cell cultures were taken before induction and after the incubation period. The samples were spun down in a bench top centrifuge at 6,000 rpm and resuspended in 200 μl ultrapure water. Of the resulting suspension 5 μl were loaded on sodium dodecylsulfate (SDS) polyacrylamide gel electrophoresis (PAGE) to check for the presence and level of a protein of the expected molecular weight in the post-induction sample. Protein on SDS PAGE were visualised by staining with Quick coomassie stain (Generon). The remainder of the cell culture (98 ml) was centrifuged and lysed according to the cell lysis protocol described in Section 2.2.2 and the soluble lysis fraction loaded onto an affinity resin for affinity chromatography. The resulting fractions were analysed with SDS PAGE to check for levels of soluble protein.

2.2.1.2 Full Scale Expression

For full scale expression a single colony from a plate incubated over night was used to inoculate 100 ml of LB medium with the appropriate antibiotic. This suspension was left to incubate at 37 °C over night. Of this pre-culture 20 ml were used to inoculate 1 l of LB medium with antibiotics. The cultures were grown at 37 °C until $OD_{600} \geq 0.6$ was reached, then protein expression was induced with 0.25 mM IPTG and either incubated at 37 °C for 4 h or at 18 °C over night, according to optimal conditions found in test expressions. After overnight growth cells were harvested by centrifugation at 6000 rpm and 4 °C and resuspended in 20 - 30 ml of compatible lysis buffer (phosphate buffered saline (PBS) 150 mM NaCL pH 7.5). The suspension was stored until needed at -20 °C.

For the production of ^{15}N or ^{13}C ^{15}N labeled protein a similar protocol was followed. The pre-culture was prepared as described above. After over night incubation,

10 ml of pre-culture were added to 500 ml ^{15}N or ^{13}C ^{15}N M9 minimal medium.

2.2.2 Cell Lysis

For protein extraction cell pellets were thawed on ice. DNAse, 1 % Triton X-100, "cOmplete" mini ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor cocktail (Roche) and 100 mM $MgCl_2$ were added. The suspension was treated with sonication on ice in a glass beaker with short pulses of 15 s at a 2.5 W power output followed by 30 s of cooling down period for a total pulsation time of 5 min. The sample was then spun at 13,000 rpm for 30 min at 4 °C. Samples of the supernatant (1 μl and 9 μl sample buffer) and of the resuspended pellet (3 μl of 100 μl resuspension and 7 μl sample buffer) were loaded on SDS PAGE to check for soluble and insoluble fractions of protein of the expected molecular weight.

2.2.3 Affinity Chromatography

Prepacked HiTrap Talon columns (GE Healthcare) with 5 ml column volume (CV) were attached to an Äkta purification system (Äkta Purifier or Äkta fast protein liquid chromatography (FPLC)) at either room temperature or in a 4 °C cabinet depending on protein temperature stability. Maximum pressure was defined as 0.5 MPa. Equilibration of the column resin with lysis buffer took place at a constant flow rate of 2.5 ml/min. The column was loaded with the supernatant of the lysed cells at a flow rate of 1 ml/min. Subsequently, the column was washed with 20 CV of lysis buffer. A second washing step using lysis buffer and 2 % elution buffer - containing 250 mM imidazole - was performed. A gradient of 2 % elution buffer to 100 % elution buffer over 25 CV at a flow rate of 1.5 ml/min was used to elute the bound protein. The fraction

size for collection was set to 2 ml. The protein content of the fractions was traced by ultraviolet (UV) absorbance at 280 nm. Where possible additional wavelengths were used for absorbance measurement. Samples of fractions were evaluated by SDS PAGE.

2.2.4 Proteolytic Tag Cleavage and Dialysis

Protein-tags in constructs used in this project were designed to be separated from the main protein by a specific protease cleavage site which allowed the cleavage of the tag before conducting structural experiments (Table 2.6). Protease cleavage took place over night at 4 °C during dialysis. For dialysis, Spectra/por molecular membranes

Table 2.6: Proteases

Protease	Cleavage site	Comment
Thrombin produced in-house	Leu-Val-Pro-Arg-|-Gly-Ser	used to remove tags from proteins produced in pHis vector; His-tagged
HRV 3C Protease produced in-house	Leu-Glu-Val-Leu-Phe-Gln-|-Gly-Pro	used to remove tags from proteins produced in pGEX and pfloat vectors

were used. The membranes were washed with ultrapure water and then filled with the protein solution before incubation in 2 - 4 l buffer. The dialysis took place through stirring at 4 °C over a period of up to 12 h. The procedure was repeated two times.

2.2.4.1 Size Exclusion Chromatography

Protein containing elution fractions from affinity chromatography were further purified using SEC. For this, protein containing samples were purified with GE Sephadex 75 or Sephadex 200 columns of sizes 16/60 or 26/60. Chromatography was performed on an Äkta FPLC system at 4 °C. The buffer used to elute the protein from the column

was varied depending on the experimental method for which the protein was being purified. Maximum pressure was defined as 0.5 MPa and protein containing fractions were tracked using UV absorbance at 280 nm.

Protein containing elution fractions were analysed for their protein content and purity using SDS PAGE. The fractions containing pure protein were then combined and concentrated using spin concentrators, either Amicon Ultra (Millipore) with a membrane of regenerated cellulose or Sartorius Vivaspin with polyethersulfone (PES) membrane. For SAS-6 proteins purified for work in this thesis the use of PES membranes reduced aggregation during the concentration process. Protein concentration was measured using UV absorbance at 280 nm on NanoDrop 2000c (Thermo Scientific).

2.2.5 Protein Quality Control

For quality control, purified protein was analysed with circular dichroism (CD) and mass spectrometry (MS).

2.2.5.1 Circular Dichroism

When circular polarised light interacts with chiral particles, such as structured protein, the two types of circularly polarized light (right and left oriented) are absorbed to different extents. In a CD experiment, equal amounts of left and right circularly polarized light of a selected wavelength are incident on a protein solution. The wavelength-dependent difference of absorption is measured, yielding the CD spectrum of the sample.

The far-UV CD spectrum of proteins shows characteristic features depending on the secondary structure of the protein. This way α-helices (two spectral minima at 208 nm and 220 nm, a maximum near 192 nm) and β-sheets (spectral minimum at 217 nm and maximum near 198 nm) can be detected and an unfolded protein can be

distinguished from a structured protein (typically minimum at around 200 nm). Changes in conformation lead to a change of the spectrum making it possible to determine the conformational behavior of the protein in specific buffers or at set temperatures.

CD spectra were collected from samples in PBS buffer with 150 mM NaCl at pH 7.5. All spectra were collected on a Jasco (Eaton, MD) J-815 Spectropolarimeter in a quartz glass cuvette with a 1 mm path length (Hellman Analytics). The CD signal θ (in mdeg) was transformed into mean residue ellipticity (MRE) (in $\frac{deg \cdot cm^2}{dmol}$), which takes into account sample concentration c (in μM), the number of amino acids n_{aa} and the path length l (0.1 cm) as seen in equation 2.1.

$$MRE = \frac{\theta}{c \cdot n_{aa} \cdot l \cdot 10} \tag{2.1}$$

Information about the thermodynamic stability of a fold can be gained by measuring CD spectra over a temperature gradient. Protein construct stability was examined with series CD experiments conducted at temperatures between 10 - 80 °C.

2.2.5.2 Mass Spectroscopy

Intact mass spectroscopy measurements for protein and small molecule samples were performed with electrospray ionisation (ESI) mass spectroscopy with micromass liquid chromatography time of flight mass spectrometers. Measurements for protein samples were taken by Dr. David Staunton (Biochemistry Department, Oxford). Small molecule experiments were conducted by Dr. James Wickens (Chemistry Department, Oxford).

2.3 Biophysical Interaction Assays

2.3.1 Fluorescence Polarisation

A fluorescence polarisation (FP) based assay can be used to detect the change of motion behavior of large molecules in solution when - as a result of their interaction with another large particle - their volume changes. For this purpose the macromolecule is tagged with a fluorescence marker. Polarisation is an intrinsic property of fluorescence and can deliver directional information, which in combination with Brownian motion theory can be used to detect molecule volume changes, due to binding to another molecule or major structural changes.

When excited with planar (linear) polarised light, fluorescence molecules emit planar polarised light back. Since molecules in solution undergo Brownian motion, the molecular tumbling influences the measured fluorescence polarisation of the emitted light. The rotational correlation time (ϕ) is influenced by temperature (T), viscosity (η), the Boltzmann constant (k_B) and volume (V) of the molecule. The speed of the molecular motion is fast enough to affect polarisation within the nanosecond long fluorescence lifetime specific to the fluorophore.

$$\phi = \frac{\eta\, V}{k_B\, T} \tag{2.2}$$

In a fluorescence polarisation experiment the flurophore is excited with planar polarised light from a polarisation filter. The emitted light is detected through another polarisation filter. The intensity of the emitted light is measured at two settings relative to the excitation filter. $I_{\parallel}$ is the intensity measured when both filters are parallel; $I_{\perp}$ the intensity measured for a perpendicular setting. The difference in the

intensities $I_{\parallel}$ and $I_{\perp}$ is the measure of the extent of polarisation, P, observed.

$$P = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} \tag{2.3}$$

Another way to describe measured intensities is by anisotropy, A. This uses the total intensity ($I_{\parallel} + 2I_{\perp}$) by including the second perpendicular emission plane:

$$A = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}} \tag{2.4}$$

Both polarisation and anisotropy are unitless and describe the differences in the intensities measured. Since both P and A are composed of the same measurement values and the only difference is the doubling of value $I_{\perp}$ in the equation for A (see Equation 2.4 and 2.3), P and A can be converted into each other using the following relationship:

$$A = \frac{2}{3}\left(\frac{1}{P} - \frac{1}{3}\right)^{-1} \tag{2.5}$$

This can be shown to be true by substituting P (using Equation 2.3) in Equation 2.5 and simplifying the resulting equation as follows:

$$A = \frac{2}{3}\left(\frac{1}{\frac{1}{\frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}}} - \frac{1}{3}}\right) \tag{2.6}$$

$$A = \frac{2}{\frac{3(I_{\parallel} + I_{\perp})}{I_{\parallel} - I_{\perp}} - 1} \tag{2.7}$$

$$A = \frac{2(I_{\parallel} - I_{\perp})}{3I_{\parallel} + 3I_{\perp} - (I_{\parallel} - I_{\perp})} \tag{2.8}$$

$$A = \frac{2(I_{\parallel} - I_{\perp})}{3I_{\parallel} + 3I_{\perp} - I_{\parallel} + I_{\perp}} \tag{2.9}$$

$$A = \frac{2(I_{\parallel} - I_{\perp})}{2I_{\parallel} + 4I_{\perp}} \tag{2.10}$$

$$A = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + 2I_{\perp}} \tag{2.11}$$

Equation 2.11 resulting from rearrangement of Equation 2.5 with insertion of the Equation 2.3 for P (see Equations 2.6 to 2.11), is equal to the original Equation 2.4 for A. This shows that Equation 2.5 can be used to convert values of polarisation to anisotropy.

The polarisation and mobility of the macromolecule in solution are connected through the Perrin equation:

$$\left(\frac{1}{P} - \frac{1}{3}\right) = \left(\frac{1}{P_0} - \frac{1}{3}\right)\left(1 + \frac{\tau}{\phi}\right) \tag{2.12}$$

where P_0 is the intrinsic polarisation of the chosen flurophore and τ is the fluorescence lifetime specific to the flurophore.

For the anisotropy, Equation 2.12 can be rearranged using the relationship between P and A shown in Equation 2.5:

$$\frac{A_0}{A} = \frac{\left(\frac{1}{P} - \frac{1}{3}\right)}{\left(\frac{1}{P_0} - \frac{1}{3}\right)} \tag{2.13}$$

where A_0 is the intrinsic anisotropy of the chosen flurophore. Equation 2.13 applied to the Perrin equation leads to the following formula for anisotropy:

$$\frac{A_0}{A} = 1 + \frac{\tau}{\phi} \tag{2.14}$$

The rotational correlation time, ϕ (as defined in Equation 2.2), connects the anisotropy and the polarisation respectively, determined by the difference of the measured intensities, to the volume of the molecule, V:

$$\frac{A_0}{A} = 1 + \tau \frac{k_B T}{\eta V} \tag{2.15}$$

It follows that the anisotropy, A, increases when V, the volume of the flurophore particle, increases.

2.3.1.1 Labels

5-carboxyfluorescein (5-FAM) with peak excitation wavelength of 492 nm and peak emission wavelength of 518 nm with a fluorescence lifetime of $\sim$ 4 ns or 5-(2-[(iodoacetyl)-amino]ethylamino)naphthalene-1-sulfonic acid (IAEDANS) peak excitation wavelength of 336 nm and a peak emission wavelength of 490 nm with a fluorescence lifetime of 10 - 15 ns were used to label cysteine residues in purified protein following the standard labeling protocol described here. Polarisation was measured using a Clariostar monochromator microplate reader (BMG Labtech).

2.3.2 Analytical Ultracentrifugation

A method which uses sedimentary forces acting on proteins to determine their sedimentation coefficient is AUC. This way the oligomerisation state of a folded protein in solution, the overall homogeneity of a sample and other related factors can be determined. In simple terms a protein sample is centrifuged at a stable temperature at very high speed in a transparent sample cell. During that time the movement of protein molecules is monitored by a spectrometer and interference optics.

During ultracentrifugation a particle/protein is acted on by three forces: F_c, the centrifugal force, F_b, the bouyant force and F_f, the frictional force between the protein

and the solvent. At equilibrium these forces act together in the way described by equation 2.16. In the following definition of the forces, m is the mass of the protein, ω is the angular velocity, r is the distance from the centre of rotation, m_{dips} is the mass of displaced solvent, f is the frictional coefficient and v is the velocity of the particle.

$$F_c - F_b - F_f = 0 \tag{2.16}$$

$$F_c = m\omega^2 r \tag{2.17}$$

$$F_b = -m_{disp}\omega^2 r \tag{2.18}$$

$$F_f = -fv \tag{2.19}$$

When the density of the solvent and the partial specific volume of the particle are taken into account, this leads to equation 2.20, also known as the "Svedberg equation". In this equation s is the Svedberg coefficient, M is the molecular mass, $\bar{\nu}$ is the partial specific volume of the particle, ρ_f is the density of the solvent and N_{AV} is Avogadro's number.

$$s = \frac{v}{\omega^2 r} = \frac{M(1 - \bar{\nu}\rho_f)}{N_{AV} f} \tag{2.20}$$

This equation is used to describe the motion of the particle in terms of molecular weight and frictional coefficient. The Sverdberg coefficient is measured in Svedberg units, S, with 1 S = 10^{-13} s.

There are two kinds of experiments possible with AUC [170]. One is the equilibrium experiment in which a solution of protein is spun at comparatively low speed for a long period of time. The sedimentation coefficient is taken only after equilibrium has been reached.

The other type of experiment is the velocity experiment in which the actual process of sedimentation is of interest. This experiment was chosen and performed in this study. During the velocity experiment the rate at which the sedimentation boundary moves is a measure of the sedimentation coefficient, which in turn depends on the molecular weight and also on the molecular shape. The minimum width of the sedimentation boundary is related to the diffusion coefficient of the molecule; the presence of multiple species with similar sedimentation coefficients will cause the boundary to be broader than expected on the basis of diffusion alone.

Early experiments on AUC velocity were performed in the Biophysics facility in the Biochemistry Department, University of Oxford. Later experiments were conducted at the Rutherford Appleton Laboratory site on the Harwell Science and Innovation Campus (supervised by Dr. Jemma Harris and Dr. David Scott). The instrument used was a Beckman Coulter ProteomeLab XL-I Ultracentrifuge. Beckman Coulter An-50 Ti and An-60 Ti analytical rotors, cells and counterbalance were used with sapphire window inserts. The sedimentation velocity experiments were performed at 15 °C and 35,000 rpm with data collected simultaneously by absorbance measured at 280 nm wavelength and interference optics. The run was stopped once sedimentation was complete after about 200 - 250 scans. One scan is defined as a complete measurement of all cells per time point.

Samples were contained in 20 mM (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES) buffer and 150 mM NaCl pH 7.5 with 2 mM tris(2-carboxyethyl)phosphine (TCEP). Sample concentration for wild type SAS-6 constructs was 100 μM and 50 μM for mutant constructs. Where indicated 5 % v/v dimethyl sulfoxide (DMSO) was added to the sample. Small molecules were used at 2 mM concentration as indicated.

Data were analysed by calculating the sedimentation coefficient distribution ($C(s)$)

in *Sedfit*, and *Sednterp* (version 2) [171, 172]. The unit for the sedimentation coefficient distribution $C(s)$ is g ml^{-1} S^{-1} (where S is the Svedberg unit). The C(s) values are then normalised to the highest C(s) of the distribution, for ease of use; the unit becomes arbitrary.

2.4 Nuclear Magnetic Resonance Spectroscopy

NMR is a uniquely suitable technique to study the behavior of proteins in solution in great detail. It uses the magnetic moment of non-zero spin nuclei. These can be influenced by an applied magnetic field and a spin-state change can then be induced by radio-frequency pulses. The required frequency to induce the spin-state change is dependent on the magnetic field and the type of nuclei one wants to observe. In a static magnetic field, the spins align. After the pulse is applied a relaxation period follows, during which spins realign in the static magnetic field, this free induction decay (FID) is what is measured and after Fourier-transformation gives rise to the NMR spectrum. The FID contains information about the state of the selected type of nucleus and about the type of bond in which the atom is engaged. Since NMR is also highly sensitive to the chemical environment of each atom present in a molecule, spins connected through bonds influence each other via J-coupling, giving rise to multiplicity. A further influence on the resonances of individual atoms comes from nuclear shielding caused by the surrounding electron shells. By using pulses of varying frequencies and lengths, the spin perturbation can be transferred between other nuclei types before decay. This allows for the transfer of information on the environment of specifically targeted atoms and bonds within a molecule, and makes multidimensional experiments possible.

All spectra were collected using Bruker Avance II or III spectrometers and magnets

of field strengths of 11.7 T, 14.1 T or 17.6 T, fitted with cryogenic TCI probe heads. Experiments were controlled and recorded using *Topspin* software, data were processed using *NMRPipe* [173] and visualized for evaluation with *Sparky* [174].

2.4.1 Heteronuclear Single Quantum Correlation

2D ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) NMR experiments were used in this study to determine possible binding of small molecules and peptides to the SAS-6 N-terminal domain of *C. elegans*, *C. reinhardtii* and *H. sapiens*. The correlation of chemical shifts of ^{1}H and ^{15}N by HSQC experiments gives rise to one resonance peak per residue, resulting in a spectrum unique to the protein, which gives insight into its structural state (Figure 2.1).

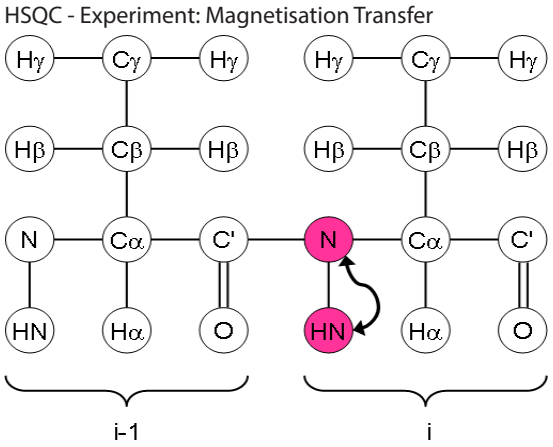


Fig. 2.1: Magnetisation Transfer in HSQC: Figure adapted from http://www.protein-nmr.org.uk with permission from Dr. Victoria Higman-Davies.

Protein samples were labeled with ^{15}N by expression in minimal medium enriched with ^{15}N (as described in Section 2.2.1.2). Protein concentration for NMR experiments was 100 μM contained in PBS (150 mM NaCl, 5 % DMSO and 7.5 % D_2O at pH 7). Spectra were collected at 25 °C. Small molecules were added to a concentration 2 mM.

2.4.1.1 Observing Interactions by NMR

Changes in HSQC spectra result from changes in the environment of the residues' N-H atoms. This can include changes in buffer composition, structural changes in the protein and ligand binding. Therefore it is possible to study interactions by studying the changes occurring upon addition of potential binding partners to a formerly measured ^{15}N labeled protein. The comparison of the resulting spectra can deliver information about the strength and location of the interaction.

The observed differences in the resonance frequencies arise from chemical exchange. Chemical exchange occurs on different time scales depending on the reaction or exchange rate k and can range from nanoseconds to seconds. How the chemical change manifests in NMR depends on the timescale of the exchange rate. If the exchange rate between two states A and B is slower than on the chemical shift timescale (μs - ms) both states give rise to individual resonance peaks. In the intermediate timescale regime, resonance peaks of both states can no longer be fully distinguished. If the resonance peaks result from overlapping resonances and appear as a broad peaks, we speak of line broadening. In the fast timescale one single sharp resonance is detected that reflects both states, with chemical shift reflecting the population averages of states A and B.

When studying interactions in HSQC experiments, a spectrum of the labeled protein is collected (spectrum 1), then the potential ligand (protein or small molecule) is added and another spectrum collected (spectrum 2). When interaction occurs, the chemical exchange will lead to perturbation of resonance position, which can be mapped to specific residues [175]. To analyse chemical shift perturbation $\Delta\delta$, resonance changes (δ_i) in H (δ_H) and N (δ_N) dimension are weighted with weighting factor α_i and combined. The weighting factors of $\alpha_N = 0.14$, applied in the N dimension, and

$\alpha_H = 1$, applied in the H dimension, are empirical values.

$$\Delta\delta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\alpha_i\delta_i)} \tag{2.21}$$

$$\Delta\delta = \sqrt{\frac{1}{2}(\delta_H^2 + 0.14 \cdot \delta_N^2)} \tag{2.22}$$

In this manner the Euclidean distance between the resonances originating from the same residue in two different HSQC spectra can be determined. Chemical shift perturbations in this study are considered notable if they exceed at least twice the intrinsic experimental error ($2\sigma_{exp}$), or twice the standard deviation (2σ) of chemical shift perturbation values over all residues in the experiment. The intrinsic experimental error depends on the resolution of the spectra collected, which derives from the values of sweep width and points collected. Digital enhancement of the FID after acquisition, applying for example window functions and zero-filling before the transformation of FID to the spectrum, increases the resolution. We estimate σ_{exp} to be ~ 0.01 ppm.

2.4.2 Protein Backbone Assignments

To assign NMR resonances to specific residues in a protein, 3D experiments were conducted which allow the transfer of magnetisation from ^{1}H to ^{15}N and ^{13}C. By controlling how the magnetisation is transferred, the connections of the resulting resonances can be probed. There are several experiments that can be combined to distinguish the residues by determining their position in the amino acid sequence through their connection.

Such sequence specific resonance assignments had already been created for the *C. elegans* SAS-6 N-terminal domain construct used in this study for NMR experiments and the equivalent domain of *C. reinhardtii*. For the assignment of resonances in the

H. sapiens SAS-6 N-terminal domain spectrum, ^{15}N and ^{13}C double labeled protein was produced in minimal medium containing ^{13}C glucose. The experiments were performed with protein at 500 μM protein concentration in PBS buffer with 150 mM NaCl, 5 % D_2O and pH 7.5 at 25 °C.

The two most fundamental experiments used for assignments are CβCα(CO)NH (CBCA(CO)NH) and CβCαNH (CBCANH) (Figure 2.2). The experiments are named after the nuclei involved. Here CA and CB refer to the Cα and Cβ atoms; CO refers to the carbonyl carbon. The order of nuclei in the experiment name is a reference to the sequence of magnetisation transfer during the experiment. The parentheses in CBCA(CO)NH reflect that the carbonyl carbon is involved in the magnetisation transfer, but a signal from it is not recorded.

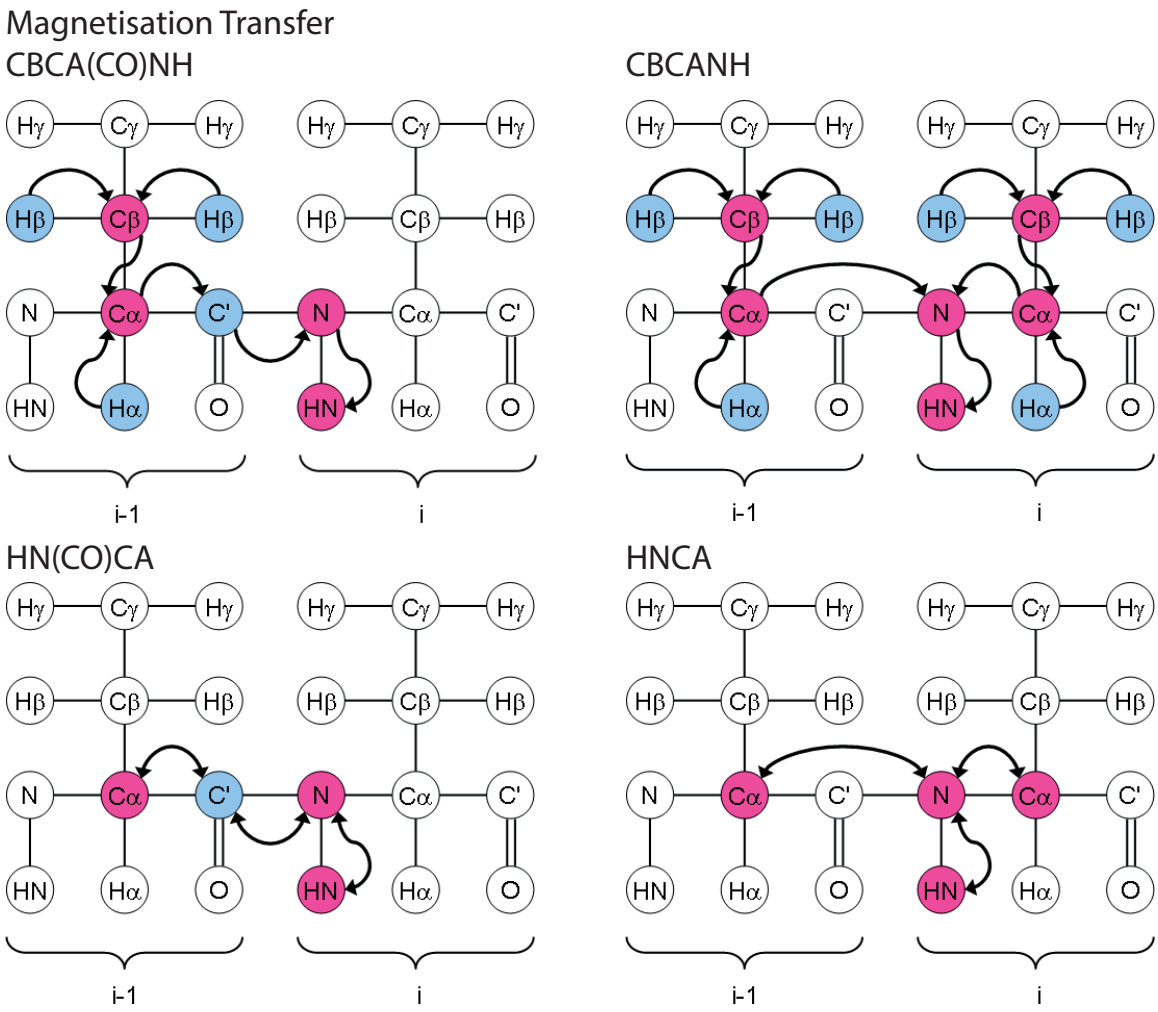


Fig. 2.2: Magnetisation Transfer in CBCA(CO)NH, CBCANH, HN(CO)CA and HNCA: Figure adapted from http://www.protein-nmr.org.uk with permission from Dr. Victoria Higman-Davies.

The CBCA(CO)NH experiment is designed to correlate the ^{1}H and ^{15}N amide resonances of one residue with both ^{13}Cα and ^{13}Cβ resonances of its preceding residue via the intervening ^{13}CO spin (Figure 2.2 and 2.3). The spectrum resulting from this shows resonances from Cα and Cβ of the amino acid in position i-1, correlated to N and H of position i. The CBCANH experiment correlates the ^{1}H and ^{15}N amide resonances with those of the intra- and inter-residue ^{13}Cα and ^{13}Cβ resonances. Resonances in the CBCANH spectrum correspond to Cα and Cβ of both amino acids in position i-1

and i. It is possible to annotate the peaks to the amino acid sequence. CBCA(CO)NH and CBCANH spectra can be taken together to trace the amino acid sequence, linking resonances to their corresponding residue (Figure 2.3).

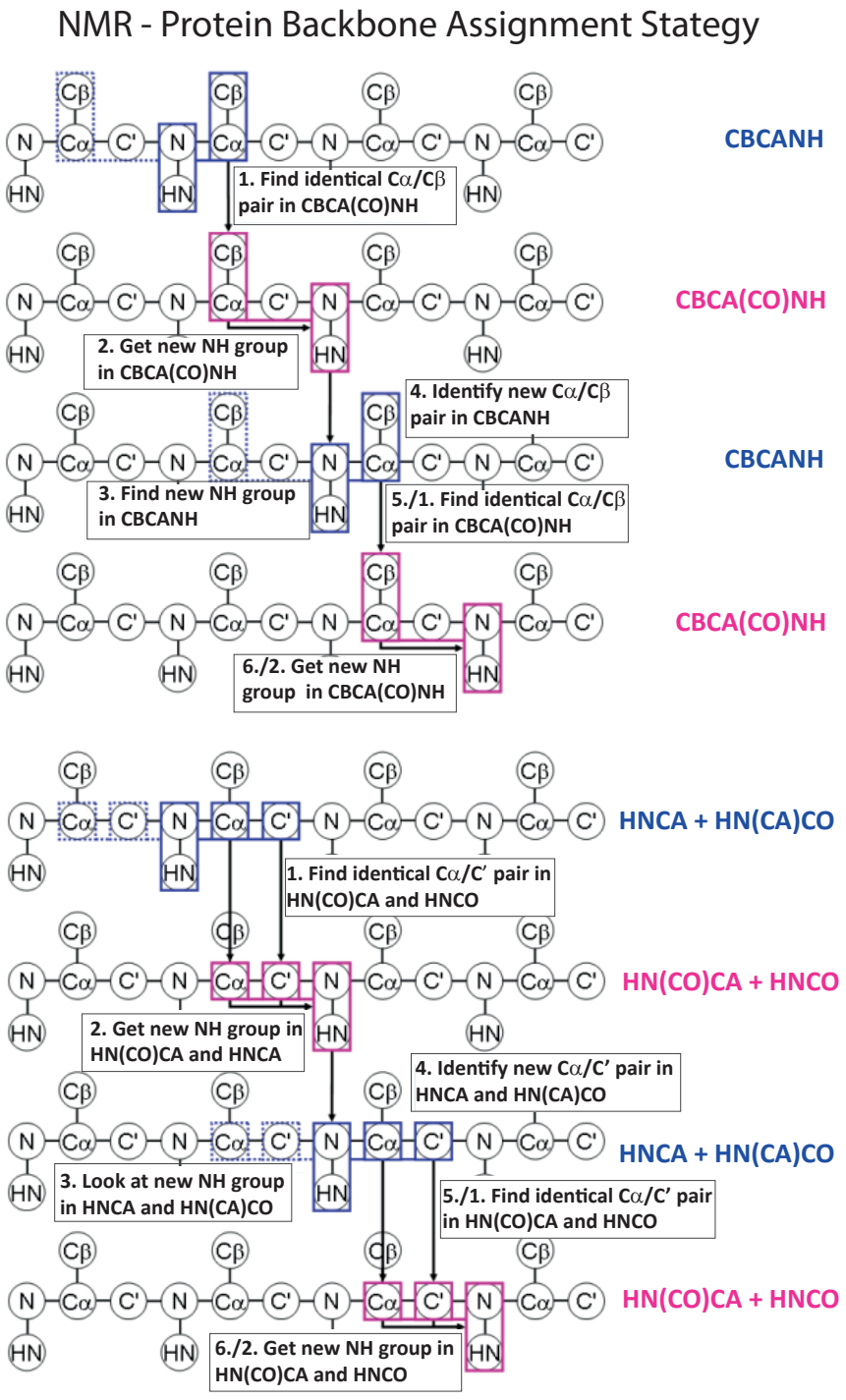


Fig. 2.3: Protein Backbone Assignment Strategy: Information from CBCA(CO)NH, CBCANH, HN(CO)CA and HNCA is sequentially linked to assign protein residues. Figure adapted from http://www.protein-nmr.org.uk with permission from Dr. Victoria Higman-Davies.

Additional information for the assignment of the *H. sapiens* SAS-6 N-terminal domain residues was collected from HNCα (HNCA) and HN(CO)Cα (HN(CO)CA) experiments. These experiments also provide information about i and i-1 connectivities as shown in Figure 2.3. Resonances in processed spectra were manually picked and preliminarily assigned using the *PINE* web server [176]. Automatic assignments were checked manually and needed to be manually completed.

2.4.3 Secondary Structure Information from NMR

Information from NMR spectra can also be used to estimate secondary structure properties of a protein. In this work two NMR techniques were used to confirm the unfolded state of loop 103-130 present in the *C. elegans* SAS-6 N-terminal domain. These experiments were not conducted by the author of this study. However, since the data were analysed by the author the principles of the techniques will be briefly described here.

2.4.3.1 Chemical Shift Indexing

Chemical shift indexing is a fast method for prediction of secondary structure from backbone ^{13}C chemical shifts [177, 178]. Cα resonances experience a downfield shift when located in α-helices and an upfield shift when located in β-strands. For Cβ resonances the opposite behavior can be observed. In addition to this information, experimental values for peptides without secondary structure were collected into a table of reference values for every amino acid. For secondary structure prediction using chemical shift indexing, these reference random coil values are subtracted from the Cα and Cβ values observed. Empirical threshold values are then applied to predict whether the corresponding residue shows behavior characteristic of residues in α-helices, β-strands or are likely to be contained in an unfolded region.

Here the backbone chemical shifts were collected for a *C. elegans* SAS-6 N-terminal domain construct spanning residues 1-168 and chemical shift index analysed with the reference table recorded in [178].

2.4.3.2 Using Heteronuclear Nuclear Overhauser Effect to Show Flexibility

The experiments described so far measure polarisation transferred via direct bonds. However, cross-relaxation makes it possible for spin polarisation to be transferred to non-bonded atoms that are within a certain range. This is called the nuclear Overhauser effect (NOE) which can be used to determine if residues are in close proximity to each other [179]. Here we use $\{^{1}H\}$-^{15}N heteronuclear NOE to reveal the difference in dynamic behavior of residues in folded secondary structure elements and highly flexible loop residues. What we are measuring is the effect of the correlation time on the cross-relaxation rate. The NH bond is subject to the overall tumbling motion of the protein, and a faster local motion, which has an influence on NOE signals. In the case of residues contained in a highly dynamic loop, the reorientation through the fast local motion is so rapid that the $\{^{1}H\}$-^{15}N NOE signal is lost or changes phase. This enables us to differentiate highly flexible residues in the *C. elegans* SAS-6 N-terminal domain from residues that are more rigid and most likely to be incorporated into secondary structure elements.

2.4.4 Small Molecule NMR

The identity of small molecules can be determined by a combination of NMR experiments and intact mass determination by MS. Due to the low number of atoms present in small molecules, 1D NMR experiments detecting ^{1}H or ^{13}C frequencies can

be used to distinguish the chemical groups contained. In the case of small molecules ^{13}C natural abundance is sufficient to detect ^{13}C resonances. More detailed information about how chemical groups are connected within the small molecule is gained by ^{1}H-^{13}C HSQC, showing resonances corresponding to every carbon atom attached to an H atom. This is further supplemented with ^{1}H-^{13}C heteronuclear multiple-bond correlation (HMBC) experiments which show long range correlations between carbons and protons that are separated by up to three to four bonds.

Small molecule NMR spectra were collected from samples at 10 mM concentration in deuterated DMSO. Spectra were analysed with *Topspin* (Bruker) and *NmrDraw* [173] and carbon atoms assigned manually with the help of resonance predictions made with *ChemNMR* in *ChemDraw* (Perkin Elmer).

2.5 X-ray Crystallography

To gain insight into the atomic details of protein-protein and protein-small molecule interactions, atomic resolution structural information is necessary. Only a few techniques are capable of delivering atomic resolution structures of proteins and protein complexes. The technique currently able of delivering the highest resolution structures is X-ray crystallography. This method is based on the elastic scattering of electromagnetic radiation with wavelength of around 0.1 nm. To use X-ray diffraction for the reconstruction of atomic structures, proteins need to be arranged in a highly symmetrical three dimensional array; a crystal. Radiation diffracted from an array causes both constructive and destructive interference effects. Due to the periodicity in the array, constructive interference causes measurable diffraction patterns from which - using Bragg‘s law - information about the lattice that caused the diffraction maxima can be extracted:

$$2d \sin\theta = n\lambda \tag{2.23}$$

where λ is the wavelength of the radiation, n is an integer, θ is the angle between the diffracted wave and the reflecting plane and d the spacing between layers of atoms.

The information contained in the diffraction pattern can be further examined under the consideration that the diffraction spots are the result of constructive interference and therefore a combination of wave functions. Every diffraction spot contains information about the position of every atom in the protein structure to be solved. This information is contained in the structure factor, F. Here h, k, l are the so called Miller indices that define the reciprocal crystal lattice, whereas x, y and z are real space coordinates of atoms. The atomic scattering factor is f_j and the number of atoms in the unit cell is N.

$$F_{hkl} = \sum_{j=1}^{N} f_j e^{2\pi i(hx_j + ky_j + lz_j)} \tag{2.24}$$

By back Fourier transform of this formula we can calculate the electron density, ρ using the structure factors, which are represented as amplitudes that can be determined from the intensities of the reflections:

$$\rho(x, y, z) = \frac{1}{V} \sum_{h} \sum_{k} \sum_{l} \mid F_{hkl} \mid e^{i\alpha_{hkl}} e^{-2\pi i(hx+ky+lz)} \tag{2.25}$$

Here V is the unit cell volume. The one component of the formula in addition to amplitudes necessary to determine the atom positions is the phase information contained in α, which is not recorded. The so called phase problem can be solved in several ways. One way is by experimental phase determination through the introduction of heavy atoms into certain positions of the protein crystals that can then be used as beacons. If the structure of a protein that is believed to be of similar fold is already known, this structure can also be used to calculate the missing phase information in a process called molecular replacement.

To grow protein crystals for the work presented here, protein concentrations suitable for crystallisation were identified using a pre-crystallisation test (PCT) (Hampton research). Broad crystallisation screens JCSG I, JCSG II, JCSG III, JCSG IV (Qiagen) and Morpheus and Morpheus II (Molecular Dimensions) at 18 °C and 4 °C were then used to find initial crystallisation conditions. To set up sitting-drop screening plates a Mosquito robot (TTP LabTech) was used. The drop volume was 200 nl and for each screening condition, two drops with screening buffer to protein ratio of 1:1 and 1:2 were set up. The screening plates were monitored in a Rigaku Minstrel HTUV crystallisation imager.

2.5.1 Structure Determination of *H. sapiens* SAS-6 N-terminal Domain

Construct hsSAS-6 1-152 F131E showed precipitation in PCT conditions from a sample of 98 mg/ml protein concentration in 20 mM Tris and 50 mM NaCl buffer, pH 7.5. This concentration was used for crystallisation screens. The structure was solved from plate shaped crystals of dimensions $\sim$ 0.2 - 0.25 mm $\times$ 0.1 mm $\times$ $\sim$ 0.02 mm, grown after 3 days at 18 °C in 0.1 M HEPES, 30 % PEG 6000, pH 7 (JCSG IV E12) in the presence of 2 mM A12. The crystals were cryoprotected by addition of 20 % v/v glycerol to the buffer solution and flash cooled in liquid nitrogen. Data were collected at Diamond Light Source (DLS, Harwell) on beamline I04-1 at 100 K to 1.46 Å resolution. The sample to detector distance was 264.56 mm and the beam wavelength was 0.97447 Å. Data were processed with XIA2 [180], and the phase problem solved by molecular replacement using a homology model based on the *D. rerio* SAS-6 structure PDB code 2Y3V (1.92 Å) [94] with *CCP4* [181], *Phaser* [182] and *Phenix autobuild* [183]. Iterative structure refinement was performed with *Phenix* [183]. Mo-

del building was done in *Coot* [184]. The space group was determined as P 2_1 2_1 2 with one protein chain per asymmetric unit and 4 asymmetric units per unit cell (unit cell dimensions [Å]: 59.59, 65.13, 38.06) (see Chapter 6, Figure 6.1 for a table of crystallographic data and refinement statistics and Appendix D for the crystal morphology and packing.). No density of compound A12 was detected. Visualisation of the structural model of hsSAS-6_N and alignment with homologous SAS-6_N structures as well as root-mean-square deviation (RMSD) calculations were performed using *Pymol* software (Schrödinger).

2.5.2 Preliminary Structure Determination of *C. elegans* SAS-6 N-terminal Domain

Construct ceSAS-6 1-168 Δ 103-130 I54E showed precipitation in PCT conditions from a sample of 6 mg/ml in 20 mM Tris and 50 mM NaCl buffer, pH 7.5. This concentration was used for crystallisation screens. Protein crystals of $\sim$ 0.18 mm $\times$ $\sim$ 0.18 mm $\times$ $\sim$ 0.03 mm formed after 5 days at 4 °C in 0.2 M Magnesium sulfate, 20 % (w/v) PEG 3350 (JCSG I F2) in presence of 2 mM A11. The crystals were cryoprotected by addition of 20% v/v glycerol to the buffer solution and flash cooled in liquid nitrogen. Data were collected at Diamond Light Source (DLS, Harwell) on beamline I04-1 at 100 K to 2.1 Å resolution. The sample to detector distance was 383.3 mm and the beam wavelength was 0.97447 Å. Data were processed with XIA2 [180], and the phase problem solved by molecular replacement using a homology model based on the ceSAS 1-168 Δ 103-130 crystallographic structure PDB code 4G79 (1.8 Å) [103] with *CCP4* [181], *Phaser* [182] and *Phenix autobuild* [183].

2.5.3 Small Molecule X-ray Crystallography

Crystals of small molecules A11 and A12 were grown in PCT conditions from 2 mM solution (20 mM Tris, pH 7.5, 5 % DMSO). The crystals were cryoprotected by addition of 20 % v/v glycerol and flash cooled in liquid nitrogen. Diffraction data were collected at the Soleil Synchrotron, Saint-Aubin, beamline Proxima-1 at 100 K. The sample to detector distance was 176 mm and the beam wavelength was 0.7999 Å. The data were phased using SHELX [185] and SUPERFLIP software [186] by Dr. Pierre Legrand (Soleil Synchrotron, Saint-Aubin).

2.6 Computational Techniques

Studying proteins, their structures and interactions in atomic detail is only possible using sophisticated biophysical techniques. These have stringent requirements in terms of equipment and sample quality. Biophysical techniques also have technical limitations. Inter- and intramolecular interactions can only be observed indirectly, and most techniques do not offer atomic resolution as well as dynamic information on a system.

Increased availability of computational power has led to the development of computational techniques that aim to predict the dynamic behaviour of proteins and their structures. Even though the calculations used for these predictions can only reflect a simplified version of the forces that govern atomic interactions, they offer valuable insights into possible mechanisms of interaction, where the limits of biophysical techniques collide with our requirements for information.

2.6.1 Homology Modeling

Homology modeling is used to predict the structure of proteins on the basis of structures solved of structurally related proteins and our understanding of secondary structure prediction and spatial restraints. For work presented in this thesis, the *MODELLER* software [187, 188] was used to create a homology model of the *H. sapiens* SAS-6 N-terminal domain structure on the basis of a crystallographic structure of the equivalent domain of the *D. rerio* SAS-6 homologue. This model was used for molecular replacement to solve the *H. sapiens* SAS-6 N-terminal domain structure form X-ray diffraction data (see Section 2.5.1 and Chapter 6). The template structure used was PDB code 2Y3V *D. rerio* 1-156 S102N which was solved to 1.92 Å resolution [94]. *MODELLER* was also used to build a model of loop α2-β5 (residues 103-130) into a crystallographic structure of the *C. elegans* SAS-6 N-terminal domain, PDB code 4G79 (1-168 Δ 103-130) [103], which was solved to 1.8 Å resolution [189].

2.6.2 Docking, Virtual Screening and Pipeline Set-up

Given the lack of a crystallographic structures of complexes of the SAS-6 N-terminal domain with potential dimerisation inhibitors discovered as part of work presented in this thesis (see Chapters 4, 5 and 6), possible binding poses of a small molecule in a target-pocket can be predicted using computational docking. There are many different docking algorithms, and a key difference between docking approaches lies in the way they score the interactions they predict.

Computational docking of small molecule structures to protein structures can also be used to screen for potentially binding compounds. Docking algorithms use simplified concepts of inter-molecular interactions and, although they do not reliably predict binding, the computational speed of this technique can allow for the screening of large

compound libraries. The reproduction of crystallographic binding position is more reliable with this technique.

In this study the docking software *Autodock Vina* [190] was used with docking parameters listed in Table 2.7. The *Vina* docking software uses a machine learning based algorithm to evaluate the positioning of a ligand structure on the target protein. The assigned score, termed "Vina score", is unit-less and used to compare *Vina* docking solutions. The protein structure file was prepared for docking using *autodock tools* [191], and the small molecule files were formatted using *OpenBabel* [192].

Table 2.7: Grid Configurations Used for Docking with *Autodock Vina*

Grid Configurations				
Species	Construct	PDB Code	Coordinates	
C. elegans	N-terminal Dimerisation Site	4G79 (monomer) 1.8 Å [103]	center_x center_y center_z size_x size_y size_z seed number of modes (max)	13.15 -33.8 8.6 23 23 23 2009 50
C. reinhardtii	N-terminal Dimerisation Site	3Q0Y (monomer) 2.1 Å [93]	center_x center_y center_z size_x size_y size_z seed number of modes (max)	13.15 -33.8 8.6 23 23 23 2009 50
H. sapiens	N-terminal Dimerisation Site	- (monomer) 1.49 Å Chapter 6	center_x center_y center_z size_x size_y size_z seed number of modes (max)	14.66 67.023 9.75 23 23 23 2009 50

2.6.2.1 High Throughput Fragment Based Virtual Screening Pipeline - Set-Up

To facilitate the docking of a large library of compounds, a number of Python [193] scripts were created for this work which set up a parallelised docking pipeline on a 64 core node of the "astrocyte" Linux cluster in the SBCB unit at the Department of Biochemistry, Oxford University. The pipeline was designed to create and prepare a ligand library on the node and dock all ligands contained in the set-up to a given target structure. The pipeline concept is shown in Figure 2.4. The scripts are listed in Table 2.8 and provided in Appendix B.

The "Frags Now" (version 25.10.2013) subset of small molecule structures from the ZINC database [194] was downloaded and prepared for docking by scripts developed during this project. The compounds in the chosen libraries have a molecular weight of under 250 Da and have a favourable calculated logarithmic partition-coefficient ($x\log P$) of under 3.5 [195]. The $x\log P$ values are provided for all compounds in the ZINC database and are calculated using the a fragment-based implementation of *Molinspiration* [194]. The compounds' solubility is characterised using $x\log P$ values instead of the experimentally determined logarithmic partition-coefficient ($\log P$) value, since no experimental solubility data for the selected compounds exists. The compounds listed in the ZINC database are purchasable, which enables follow-up *in vitro* experiments. The 527,585 structures are provided in a small number of .mol2 format files. These were converted to single-compound files and unpacked into the library (script Split_mol2.py). To avoid file-header overload, which leads to problems in most operating systems, the library was unpacked into a sub-folder structure. This creates a path utilising the eight digit ZINC ID allocated to each compound (an example path for compound ZINC12345678.mol2 would be /Working Directory/1/2/3/4/5/6/7/8/ZINC12345678.mol2). The compound files have charge

information provided at pH 7. During the library set-up, the .mol2 format is converted to .pdbqt file format using *OpenBabel* [192] to add charge distribution and torsion angles to the structural information (script pdbqt_converter.py).

The library was then docked to a manually prepared protein structure of a SAS-6 N-terminal domain (see Table 2.7), in a parallelised manner with *Autodock Vina* [190] utilising all available cores simultaneously (docking_parallel.py). The resulting docked conformations were sorted by their "Vina score" and the top 10,000 compounds extracted from the library (scoring.py).

Run times for the scripts used in this pipeline are provided in Table 2.9. The parallelised docking of 527,585 compound structures can be accomplished in little more than two weeks on 64 cores.

Table 2.8: Scripts for Virtual Screening Set-up

Virtual Screening Scripts			
	Name	Purpose	Requirements
Ligand Library	Split_mol2.py	Creates individual .mol2 files containing one structure per file	ZINC ID containing database.mol2 files
	pdbqt_converter.py	Converts .mol2 files to .pdbqt format, preparing the file for docking	Open Babel
Docking	docking.py	Docks all ligand molecules in ligand library to a single chain receptor.pdbqt within the grid information given in config.txt. Results are stored in ligand library folders.	Autodock Vina, config.txt, receptor.pdbqt, ligand library created with previous scripts
	docking_parallel.py	Parallel running version of docking.py (for more than 8 CPU threads)	Autodock Vina,config.txt, receptor.pdbqt, ligand library created with previous scripts, joblib python package
Analysis	scoring.py	Filters best ligand models for best Vina affinity score. Top results are stored in separate folder	docking.py result files, config.txt user input: number of top results wanted

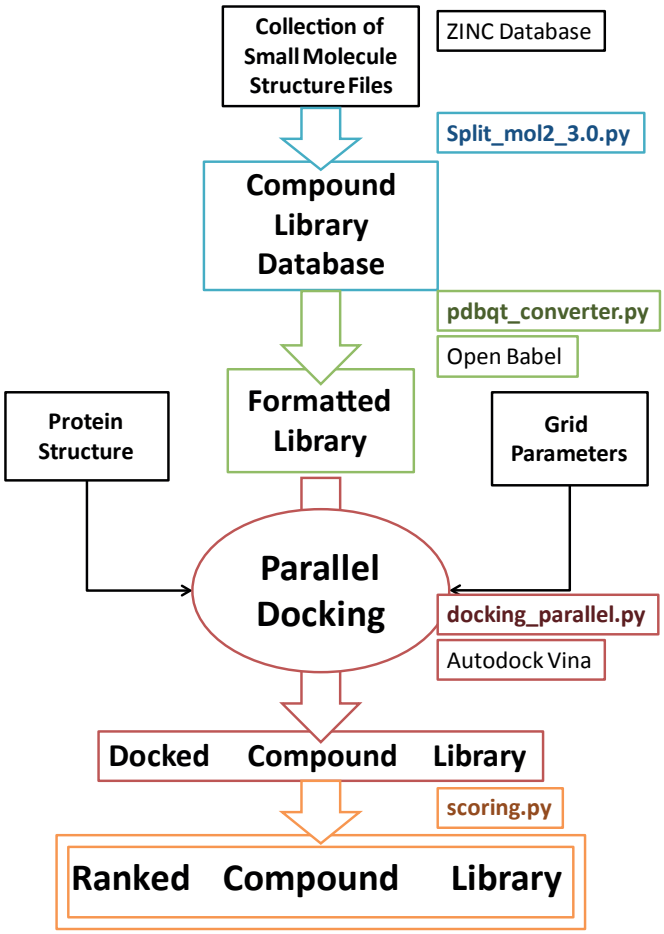


Fig. 2.4: High Throughput Docking Pipeline Set-up

Table 2.9: Test Runtimes for Developed Python Scripts

Timing of Pipeline Scripts			
98 Molecule Subset		Time	Comment
split_mol2_3.0.py	real	0 m 2.230 s	
	user	0 m 0.156 s	
	sys	0 m 0.252 s	
openbabel_converter.py	real	0 m 9.557 s	
	user	0 m 3.432 s	
	sys	0 m 1.084 s	
docking.py	real	10 m 5.268 s	8 CPU
	user	5 m 9.602 s	threads
	sys	0 m 4.320 s	
docking.py	real	9 m 3.106 s	12 CPU
	user	6 m 8.734 s	threads
	sys	0 m 2.160 s	
docking_parallel.py	real	6 m 4.216 s	12 CPU
	user	4 m 3.282 s	threads
	sys	0 m 1.963 s	
scoring.py	real	0 m 1.009 s	top 10
	user	0 m 0.448 s	
	sys	0 m 0.112 s	
scoring.py	real	0 m 1.466 s	top 20
	user	0 m 0.516 s	
	sys	0 m 0.180 s	
7555 Molecule Subset		Time	Comment
scoring.py	real	0 m 18.803 s	top 100
	user	0 m 16.157 s	
	sys	0 m 5.768 s	
scoring.py	real	0 m 55.651 s	top 300
	user	0 m 19.065 s	
	sys	0 m 5.928 s	
527,585 Molecule Library		Time	Comment
scoring.py	real	143 m 57.295 s	top 10,001
	user	116 m 15.262 s	
	sys	2 m 14.205 s	
docking_parallel.py		approx. 15 d	64 CPU
			threads
		⇒ 35,153/day	

2.6.2.2 Pharmacophore-like Virtual Screening

A further technique for computational screening used for this work is based on what is called a "pharmacophore". A pharmacophore is a scaffold of characteristics shared between known binding compounds that can then be used to find further candidate compounds matching the scaffold [196] (see Chapter 4). The pharmacophore scaffold was designed using *Discovery Studio* (Biovia) on the basis of the structure of loop β6-β7 in *C. elegans* SAS-6 N-terminal domain (3PYI 2.1 Å resolution [93]), the natural ligand of the target pocket. The ZINC "drug-like" database (version 25.5.2016) was then screened for matches against the pharmacophore structure, using *Discovery Studio* (Biovia). The screen and pharmacophore design were done by Dr. Minos Matsoukas (University of Barcelona).

2.6.2.3 Screening Compounds and Vendors

Compound selection purchased as a result of this screen were acquired through MolPort by vendors listed in Table 2.10.

Table 2.10: List of Compounds and Vendors

Compound Vendors		
Library ID	ZINC ID	Vendor
A1	ZINC71942314	ENAMINE Ltd.
A2	ZINC09970637	Specs
A3	ZINC72414547	ENAMINE Ltd.
A4	ZINC70040105	ENAMINE Ltd.
A5	ZINC58247642	ENAMINE Ltd.
A6	ZINC11487801	ENAMINE Ltd.
A7	ZINC55414762	ChemBridge Corporation
A8	ZINC54241093	Vitas-M Laboratory, Ltd.
A9	ZINC09584839	ENAMINE Ltd.
A10	ZINC13581301	Vitas-M Laboratory, Ltd.
A11	ZINC36638814	Specs
A12	ZINC09737303	UkrOrgSynthesis Ltd.
B1	ZINC06670228	Vitas-M Laboratory, Ltd.
B2	ZINC25772421	ENAMINE Ltd.
B3	ZINC13477106	Vitas-M Laboratory, Ltd.
B4	ZINC12534617	ENAMINE Ltd.
B5	ZINC15693151	ENAMINE Ltd.
B6	ZINC67921979	ChemBridge Corporation
B7	ZINC81336504	UkrOrgSynthesis Ltd.
B8	ZINC65591069	ENAMINE Ltd.
B9	ZINC81048793	UkrOrgSynthesis Ltd.
B10	ZINC69442707	ENAMINE Ltd.
B11	ZINC69870503	ENAMINE Ltd.
B12	ZINC76202639	UkrOrgSynthesis Ltd.
C1	ZINC28923015	ENAMINE Ltd.
C2	ZINC22976783	UkrOrgSynthesis Ltd.
C3	ZINC00067653	Vitas-M Laboratory, Ltd.
C4	ZINC79355337	UkrOrgSynthesis Ltd.
C5	ZINC04935336	Vitas-M Laboratory, Ltd.
C6	ZINC76041126	ENAMINE Ltd.
C7	ZINC63846942	ChemBridge Corporation
C8	ZINC64845906	ChemDiv, Inc.
C9	ZINC33306742	ChemDiv, Inc.
C10	ZINC02050114	Vitas-M Laboratory, Ltd.
C11	ZINC00678552	Specs
C12	ZINC13899084	Vitas-M Laboratory, Ltd.
D1	ZINC12920785	ENAMINE Ltd.
	ZINC00392813	ENAMINE Ltd.
	ZINC39122184	ENAMINE Ltd.
	ZINC01678901	ENAMINE Ltd.

2.6.3 Molecular Dynamics Simulations

To study the behavior of proteins in solution at atomic detail is the purpose of molecular dynamics simulations. The calculations that would allow the simulation of dynamics at a quantum level are still computationally too expensive to perform for systems that encompass all atoms in a protein. Molecular dynamics simulations rely on a number of simplified equations to make the simulation of proteins and solvent molecules feasible.

The rules that govern intra- and intermolecular interactions are contained in a so called force field [197, 198]. These can be expressed as energy potential $V(r)$, which is a summation of bonded and non-bonded interactions to be considered.

$$V(r) = \sum V_{bonds} + \sum V_{angles} + \sum V_{dihedral} + \sum V_{nonbonded} \tag{2.26}$$

Bonded Interactions Bonds (V_{bonds}) and angles (V_{angles}) are treated as elastic springs, between two (i, j) or three points (i, j, k). Hooke's law can be used to describe how the energy of the springs varies with the square of the changes from the reference bond length (l_0) and angle (θ_0). Force constants k_B and k_θ are introduced as penalties to account for unfavourable bond lengths and torsions.

$$V_{bonds} = \sum_{bonds} k_B(l_0 - l_{i,j})^2 \tag{2.27}$$

$$V_{angles} = \sum_{angles} k_\theta(\theta_0 - \theta_{i,j,k})^2 \tag{2.28}$$

Bond lengths and angles are naturally constrained to favourable conformations. The force constants are therefore set high for these parts of the potential. This also means that the influence of $V_{dihedral}$ and $V_{nonbonded}$ is more prominent. Dihedral angles $\phi_{i,j,k,l}$

describe the configuration of four points (i, j, k, l) and have a periodic nature.

$$V_{dihedral} = \sum_{dihedral} k_\phi(1 + cos(n\phi_{i,j,k,l} - y)) \tag{2.29}$$

The energy barrier for changes in dihedral angles is lower and more conformations are allowed. In addition to the force constant k_ϕ, integer n and a phase factor y, which determines a minimum, are needed to describe dihedral angles.

Non-bonded Interactions Non-bonded interactions are defined as Van der Waal's and Coulomb (electrostatic) interactions.

For Van der Waal's interactions, the Lennard-Jones potential is used to approximate the interaction between a pair of neutral atoms (i, j). The Lennard-Jones potential considers a strong short range repulsive term (r^{12}) (Pauli repulsion), which penalises overlapping orbitals, and a long range attractive term (r^6), which describes the attraction of spontaneous dipoles (London dispersion).

$$V_{LJ} = \sum_{LJ(i,j)} 4\varepsilon_{i,j} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] \tag{2.30}$$

The distance at which the interaction energy for atoms i and j is zero, is denoted as σ_{ij}, $\varepsilon_{i,j}$ describes the attraction and r_{ij} is the distance between i and j. The type of atom i and j represent has an influence on the parameters used here. The true Lennard-Jones potential describes a short range force and decays over a distance of r^{-6}. Here a discrete cut-off distance is used to simplify the calculations.

For electrostatic interactions charges and dipoles are treated as point charges.

Coulombs law:

$$V_C = \sum_{C(i,j)} \frac{k_e q_i q_j}{r_{ij}} \tag{2.31}$$

Here q_i and q_j represent partial charges of i and j, r_{ij} is the distance and k_e is the coulomb constant. V_C describes a long range force and decays with r^{-1}. Introducing a discrete cut-off would lead to discontinuity in the system and affect accuracy. Instead the particle mesh Ewald (PME) method is used to handle long range interaction forces [199]. This uses a fast Fourier transform to calculate the long-range component of V_C, whereas it calculates the short-range component in real space. This significantly reduces computation time.

Force Field Choice There are multiple force fields available that describe biomolecular systems [197, 198]. In this study the AMBER ff99SB-ILDN force field was chosen [200]. This force field was optimised to replicate data gathered with NMR experiments and has been shown to simulate protein systems well, compared to other force fields [201].

Water Models Similar to how the force field describes architecture and interaction of amino acids in proteins, the water model describes the water molecules that make up most particles in the simulation volume. The water model and protein force field need to be compatible. The TIP3P water model used in this thesis has been shown to perform in combination with AMBER ff99SB-ILDN [202]. This model describes water molecules with the use of three charged points creating two dipole moments.

2.6.3.1 Trajectory Calculation

To simulate the movement of particles, molecular dynamics simulations use Newton's equation of motion.

$$\overrightarrow{F_i} = m_i \overrightarrow{a_i}(t) \tag{2.32}$$

This describes the force $\overrightarrow{F_i}$ acting on atom i as the product of the atom's mass m_i and the acceleration the particle experiences $\overrightarrow{a_i}(t)$. Since acceleration and velocity are connected by: $a = \frac{dv}{dt}$ and velocity v can be described as follows: $v = \frac{dr}{dt}$, the equation describing force F_i can be rewritten to include a relation to the particle position r_i.

$$\frac{d^2 r_i}{dt^2} = \frac{F_{xi}}{m_i} \tag{2.33}$$

The force F_i can also be described by energy potential V. This allows the derivation of a relationship between force, position and time. In the biomolecule systems considered here, the motions of all particles are coupled. The resulting equation to describe the motion of all particles in the simulation cannot be solved analytically due to the complexity of the system. Instead a finite difference method is used for discrete time steps Δt. The Verlet algorithm is a common implementation for this solution . In the work described *GROMACS* (version 5.1) [203–205] is used for molecular dynamics simulations, which utilises the Verlet variant "leap-frog" [206].

The velocities v_i are calculated at time $t+\frac{1}{2}\Delta t$ based on velocities from the previous time step $t - \frac{1}{2}\Delta t$ and the acceleration at this time step $a_i \Delta t$.

$$\begin{aligned} v_i(t + \frac{1}{2}\Delta t) &= v_i(t - \frac{1}{2}\Delta t) + a_i \Delta t \\ v(t) &= \frac{1}{2}[v_i(t + \frac{1}{2}\Delta t) + v_i(t - \frac{1}{2}\Delta t)] \end{aligned} \tag{2.34}$$

Positions of particles can be calculated on the basis of their position at the previous time step and the evolved velocity.

$$r_i(t + \Delta t) = r_i(t) + v_i(t + \frac{1}{2}\Delta t) \tag{2.35}$$

The infinite difference approach truncates an infinite series and therefore creates an error. This truncation error is proportional to the size of the time step, however

small time steps make the simulation more computationally expensive. To balance these effects the time step used in the work presented here was set to be Δt =2 fs, which is a 5th of the time range of the fastest motions occurring in this simulations set up (C-H bond vibration motions occur within $\sim$ 10 fs) [206]. These fast vibrational motions are often restrained to reduce system complexity. This is achieved with the LINCS method [207].

The system is further simplified by limiting the volume simulated to a box with the protein to be studied surrounded by solvent. Periodic boundary conditions are introduced to create what are effectively copies of the simulated volume in all three dimensions. This circumvents problems that would otherwise arise from particles meeting the edges of the defined volume. With periodic boundary conditions in place a particle approaching the edge of the defined volume simply reenters the copy of the simulated volume. The box size chosen is adjusted to the size of the protein system to be simulated and must be large enough to ensure that the protein to be studied does not interact with the copies of itself.

2.6.3.2 System Equilibration

The molecular dynamics simulations used here study a system containing one to two copies of protein in a solvent filled box and either study dynamic bond formation in a protein dimer or the interaction of a small molecule ligand with the protein.

The simulations describe closed systems, with a fixed number of particles N, fixed volume V and fixed energy E. Parameters in this system which need to be controlled are temperature T and pressure P. In the work presented in this thesis, the systems were equilibrated to constant temperature (300 K) and constant pressure (1 bar) in two consecutive equilibration simulations of 200 ps and 1 ns length respectively.

The Berendsen thermostat and barostat were used [208, 209], which couple the system to a temperature/pressure bath via a coupling constant τ_T/τ_P.

$$\begin{aligned} \frac{dT(t)}{dt} &= \frac{1}{\tau_T}(T_{bath} - T(t)) \\ \frac{dP(t)}{dt} &= \frac{1}{\tau_P}(P_{bath} - P(t)) \end{aligned} \tag{2.36}$$

Whereas the pressure is controlled by the adjustment of the box volume, the temperature is scaled using the velocity, possible through the temperatures connection to kinetic energy K and Boltzman constant k_B.

$$K = \frac{1}{2}\sum m_i v_i^2 = \frac{3}{2}Nk_BT \tag{2.37}$$

After the system has been equilibrated for constant temperature and constant pressure, an energy minimisation simulation is performed. This serves to remove unfavourable interactions prior to the production simulation and is achieved by minimising the potential energy function. The steepest descend minimisation algorithm was used with a target energy of 100 kJmol^{-1}nm^{-1}(summarised in [210]).

2.6.3.3 Small Molecules in Molecular Dynamics Simulations

Parameters for commonly simulated particles like amino acids, water, lipids and some ions are contained in force fields. Non-standard molecules such as the small molecule ligands studied in this work needed to be newly parameterised. Molecule structure files were obtained from the ZINC database [194]. Hydrogens were added where missing using *autodock tools* [191]. *Gaussian* [211] was then used to optimise the geometry of the molecule before parameterisation. Parameters compatible with the AMBER force fields were generated using *antechamber* and *tLeap* [39].

The contacts of small molecules with protein residues over the course of the simu-

lation were analysed using *Lintools* [212].

2.6.3.4 Steered Molecular Dynamics Simulations

The computational cost of atomistic molecular dynamics simulations is high despite computational advances. Only short timescales can be feasibly explored in unbiased simulations. To compare protein-protein interaction strengths between mutants of SAS-6 N-terminal domain dimers in this thesis, steered molecular dynamics (SMD) is used. For this biased MD two centers of mass are defined which then are pulled with constant speed in opposite directions over a set distance. The Hamiltonian for this system changes contains a term which depends on time via a Harmonic potential centered on a point which moves linearly with time.

$$\begin{aligned} H_\lambda(X,t) &= H(X) + U_\lambda(X,t) \\ &= H(X) + \frac{k(t)}{2}(s(X) - s_0 - vt)^2 \end{aligned} \tag{2.38}$$

where $k(t)$ is the spring constant describing the force applied to pull the centre of mass at a certain time point and λ is the centre of the harmonic potential.

The pulling speed has an influence on the accuracy of the simulation and was chosen to be 1 nm/ns. The total distance pulled was 20 nm, which was chosen to be far enough to separate the SAS-6 dimer. The work required to pull the centre of mass with a constant speed over the simulation is defined as:

$$W = \int_0^{t_s} dt \frac{\delta H_\lambda(t)}{\delta t} \tag{2.39}$$

The value for the work required at every given time point is recorded over the simulation as a cumulative value and can be plotted against the distance pulled. This form of biased simulation does not yield reliable values, as the force applied makes the system

extremely unphysiological. However, in this study it is used as a means to compare how strong the interactions simulated in the dimerisation interface are estimated to be for different SAS-6 N-terminal domain mutants. *PLUMED* (version 2.2) [213] is used in combination with the*GROMACS* set up used for MD simulations in this work (described in Sections 2.6.3, 2.6.3.1, 2.6.3.2, 2.6.3.3 and 2.6.3.5) to facilitate the SMD.

2.6.3.5 MD Run Parameters, Software and Analysis

All simulations in this thesis were atomistic simulations performed using the *GROMACS* biomolecular simulations package version 5.1, with the AMBER ff99SB-ILDN force field and the TIP3P water model. The systems were simulated at 1 bar, 300 K and with time steps of 2 fs for 20 ns, 50 ns, 100 ns and 500 ns. All simulations were performed in triplicate where possible. SMD could only be completed in duplicate by the time of this thesis submission.

Cubic boxes were used filled with water and 150 mM NaCl. The box length for molecular dynamics simulations with SAS-6 N-terminal domain dimers was 10 nm (Figure 2.5 B), SAS-6 N-terminal domain monomers were simulated in boxes with 8 nm length (Figure 2.5 A), and steered simulations were performed in boxes with 11 nm dimension size.

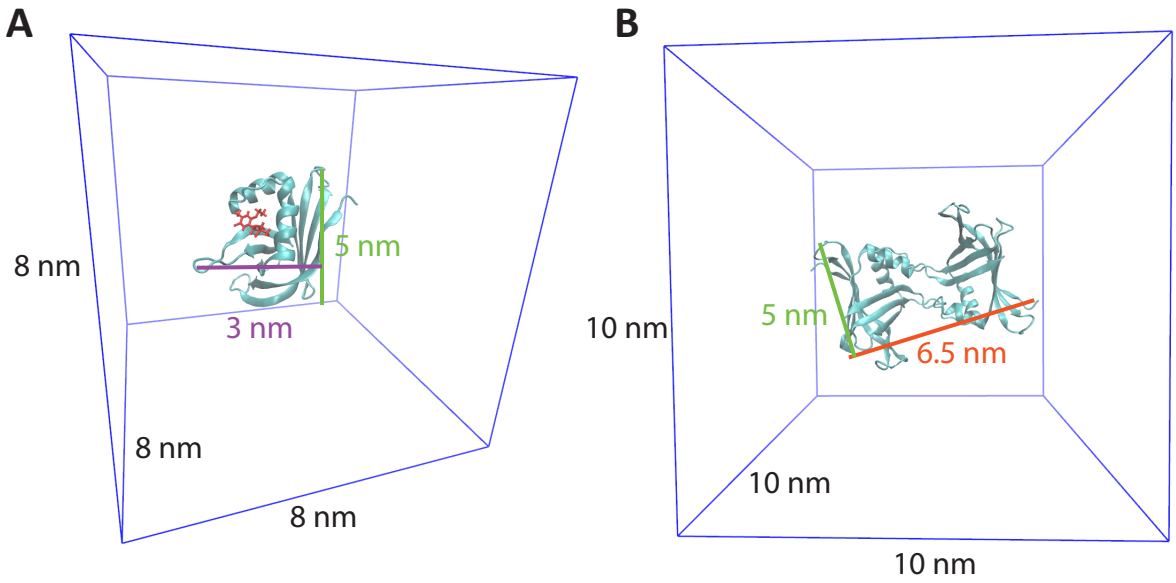


Fig. 2.5: Examples of SAS-6 Protein Structure in Simulation Box: (A) ceSAS-6_N monomer PBD code 4G79 (1.8 Å resolution [103]) with approximate dimensions in 8 nm x 8 nm x 8 nm simulation box. Small molecular compound A11 (ZINC36638814) is depicted in *red*. (B) ceSAS-6_N dimer PBD code 4G79 (1.8 Å resolution [103]) with approximate dimensions in 10 nm x 10 nm x 10 nm simulation box.

The *C. elegans* SAS-6 structure 4G79 [103], *C. reinhardtii* SAS-6 structure 3Q0Y [93] and the novel *H. sapiens* SAS-6 N-terminal domain structure described in this thesis were used in simulations. Modifications and point mutations were introduced using the *PyMOL* molecular graphics system (Schrödinger). Trajectory analysis and visualisation was performed using *VMD* [214] and *xmGrace*.

Chapter 3

Insights into the *C. elegans* SAS-6 α2-β5 Loop and Its Involvement in N-terminal Domain Dimerisation

---

"... it was the epoch of belief, it was the epoch of incredulity..."

Charles Dickens

A Tale of Two Cities

This chapter describes an investigation of the flexibility of loop α2-β5 - formed by residues 103-130 in *C. elegans* SAS-6 - and its influence on the N-terminal dimerisation of *C. elegans* SAS-6. Biophysical evidence is combined with computational simulations to form a hypothesis about the molecular interactions involved.

External contributions to work in this chapter: NMR data used for CSI analysis were collected and FP as well as HSQC NMR experiments were performed by Dr. Michele Erat. NOE experiments were performed by Ms Iris Blank. Analytical SEC experiments were conducted by Dr. Ioannis Vakonakis. The author was supported in tcl/tk scripting for data analysis by Dr. Maria Musgaard.

3.1 Introduction

The structure and function of SAS-6 have been studied extensively. The cartwheel structure visible in the centre of forming centrioles is believed to be built up from SAS-6 oligomers. It is this SAS-6 oligomer which is responsible for the nine-fold symmetry of the centriole. The SAS-6 oligomer forms by two dimerisation events; a parallel coiled-coil dimer and the N-terminal domain dimerisation. Affinities for these dimerisation events have been determined, defining the K_d of coiled-coil formation at $\sim$ 1 μM and the K_d of N-terminal dimerisation at $\sim$ 100 μM [93, 103]. Although, our knowledge of the individual dimerisation events should be sufficient to explain SAS-6 oligomerisation, open questions remain. The SAS-6 oligomerisation could not be recreated in computational studies using the known dimerisation affinities of the coiled-coil and N-terminal domains and intracellular protein concentration [158]. This suggests yet unknown mechanisms must aid oligomerisation. Unexplained density seen in the centre of the cartwheel in electron microscopic images indicate further mechanisms might be involved [81].

The evidence suggests that SAS-6 oligomerisation is more complex than expected and has not been studied in the necessary detail. Due to the central role of the SAS-6 oligomerisation in centrosome duplication, an understanding of all aspects of the interactions involved in SAS-6 oligomerisation is desirable. To close one of the gaps in our understanding of SAS-6 oligomerisation, we studied the loop region α2-β5 (residues 103-130) of the *C. elegans* SAS-6. This long loop has not been resolved in crystallographic structure or otherwise structurally explored [103]. The uncharacterised region appears to be unique to nematodes (Figure 3.1). Aside from inconclusive evidence on the possible phosphorylation of S123 located in this region [60, 149], the region comprising residues 103-130 has been neglected in studies so far.

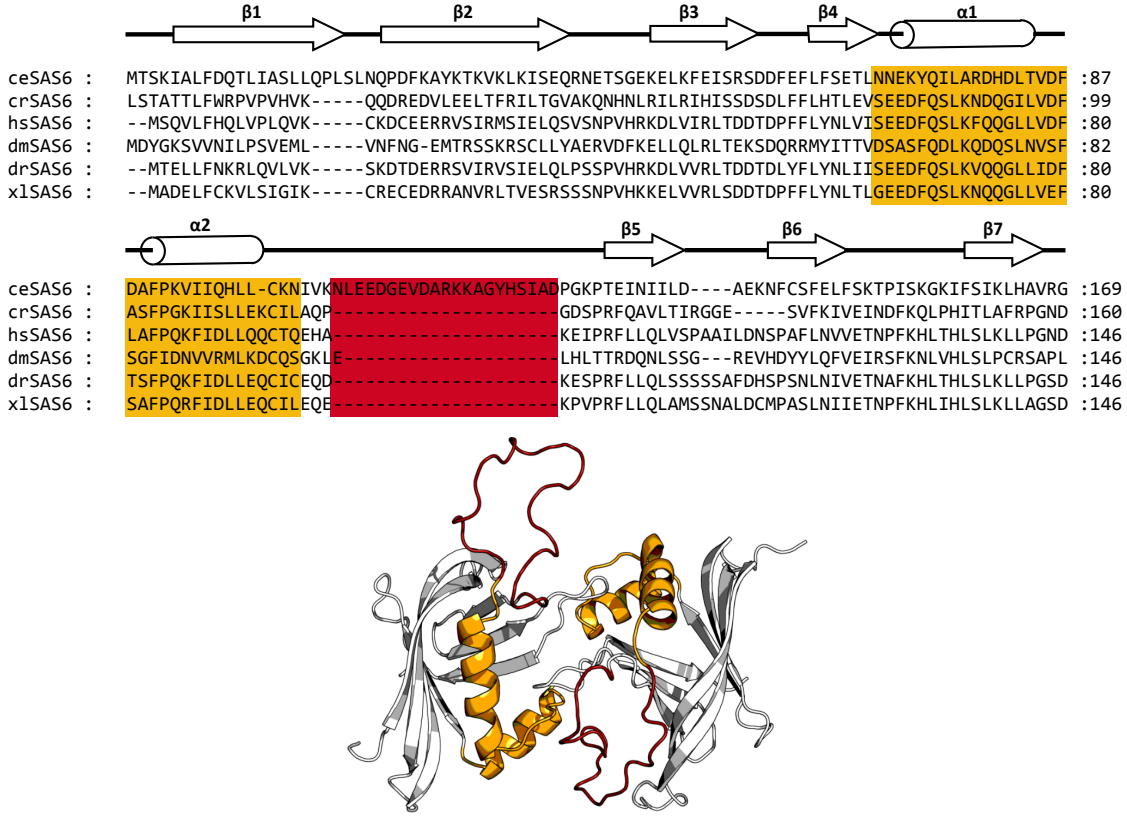


Fig. 3.1: Sequence Alignment - SAS-6 N-terminal Domain: Alignment of SAS-6 N-terminal domain protein sequences (*C. elegans* SAS-6 (ceSAS6), *C. reinhardtii* SAS-6 (crSAS6), *H. sapiens* SAS-6 (hsSAS6), *D. melanogaster* SAS-6 (dmSAS6), *D. rerio* SAS-6 (drSAS6) and *Xenopus laevis* (*X. laevis*) SAS-6 (xlSAS6)) and model of *C. elegans* SAS-6 N-terminal domain with loop region (based on structure 4G79 [149]). Region α1-α2 is highlighted in *yellow*, loop 103-130is highlighted in *red*.

3.2 Biophysical Insight into the α2-β5 Loop

3.2.1 The α2-β5 Loop is Flexible and Disordered

The structure of residues 103-130 could not be determined by X-ray crystallography [93], as no electron density was observed. This suggests that the region might be

a highly flexible loop. To examine this possibility with biophysical evidence, NMR experiments* were conducted.

Using ^{13}C CSI analysis, an insight into secondary structure can be obtained [177, 178]. This technique compares the ^{13}C chemical shifts of assigned residues to standard values for unfolded amino acids in random coil conformation. The analysis was performed on Cα and Cβ resonances of the *C. elegans* N-terminal domain (Figure 3.2). Positive Cα CSI values over a threshold of 0.7 are indicative of α helices, as are negative values under -0.7 for Cβ chemical shift differences [178]. The reverse applies for residues that are part of β-strands. Values close to zero or under the threshold indicate residues in disordered regions. The analysis correctly identifies major secondary structure features visible in crystallographic structures of this domain. For residues 103-130 Cα and Cβ chemical shift differences are close to zero, suggesting that this region is indeed dominated by random coil conformation.

*Data for ^{13}C CSI analysis were collected and backbone assignments were performed by Dr. Michele Erat. ^{13}C CSI analysis was performed by the author. NOE experiments were conducted by Ms Iris Blank

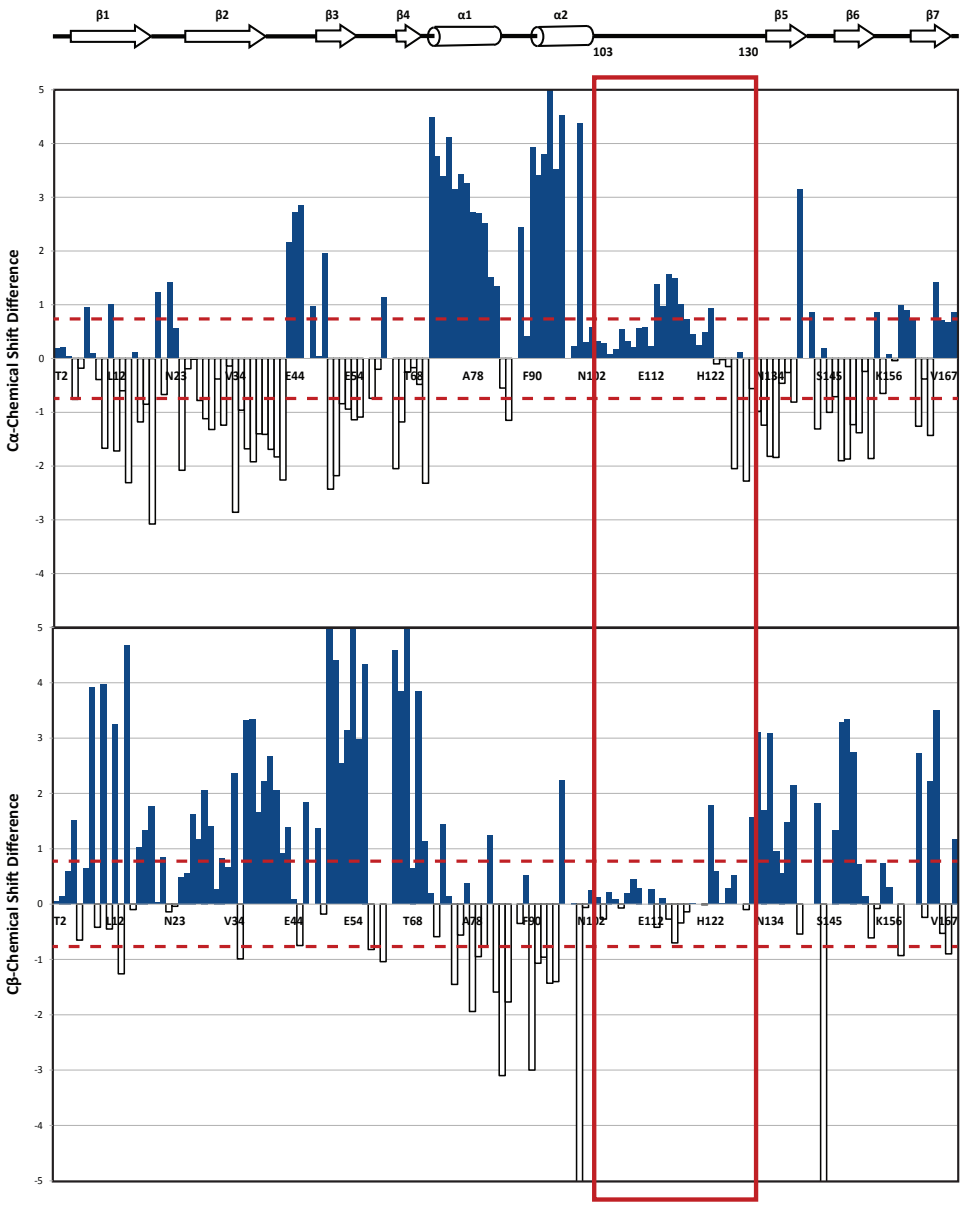


Fig. 3.2: Chemical Shift Index Analysis for Cα and Cβ Resonances of *C. elegans* SAS-6 N-terminal Domain: The threshold values of $\pm$ 0.7 are indicated by *red* line. The loop region 103-130 is indicated by a *red* box.

Further evidence that residues 103-130 form an unstructured flexible loop was gathered from heteronuclear NOE experiments (Figure 3.3). For residues in rigid secondary structure elements the NOE is strong, whereas flexible residues with mobility in the ps to ns timescale have lower NOE ratios. This can be observed for the hetero-

nuclear NOEs of residues 103-130, a clear indication of their flexibility. We therefore conclude that region 103-130 is a flexible unfolded loop.

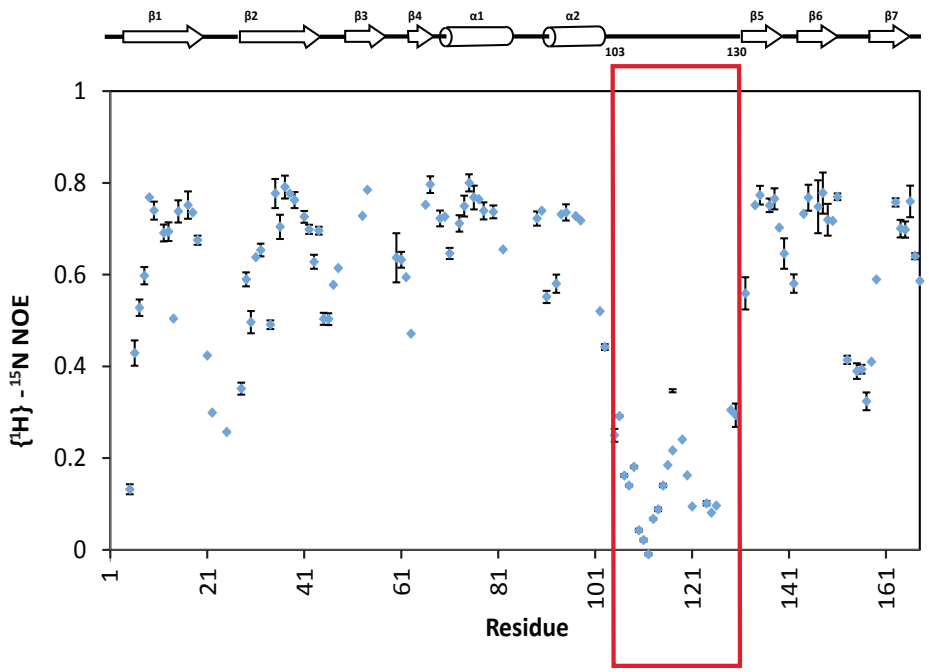


Fig. 3.3: Heteronuclear $\{^1H\}$-^{15}N NOE Intensities for Residues of *C. elegans* SAS-6 N-terminal Domain: Loop residues 103-130 are indicated by a *red* box.

3.2.2 The Presence of Loop α2-β5 Supports SAS-6 N-terminal Domain Dimerisation

SAS-6 is believed to be a purely structural component of the centriole with its main function being the formation of nine-fold symmetric oligomers. Given current evidence, it is reasonable to assume that all components of SAS-6 -including loop 103-130- are there to facilitate this function. To determine if the presence of loop residues 103-130 has influence on the dimerisation of ceSAS-6_N, analytical SEC was performed. The retention volume at which a particle elutes in SEC informs on the particle size. An increased presence of dimer particles was detected when SEC chromatograms of SAS-6_N

from samples with increasing concentrations were compared (Figure 3.4 A). For SAS-6_N Δ103-130 no change in retention volume could be seen (Figure 3.4 B), indicating that only the monomer species is present even at the highest protein concentrations (10 mg/ml).

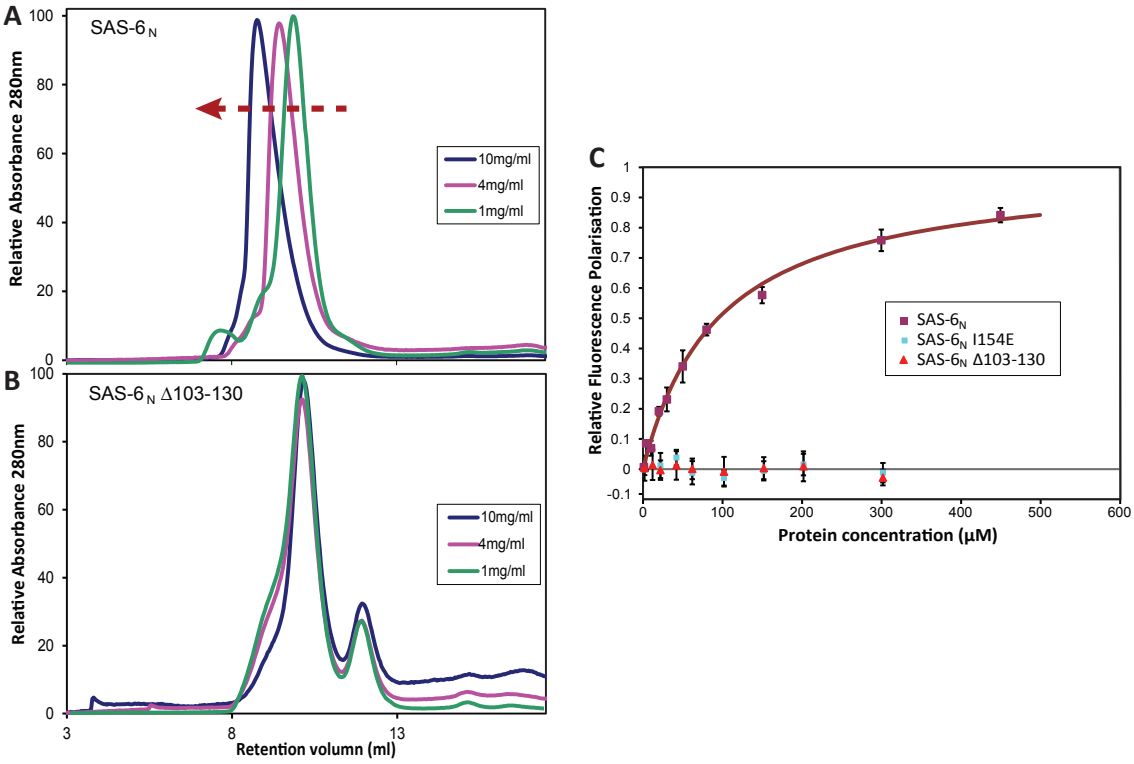


Fig. 3.4: SEC and FP Experiments Following SAS-6 N-terminal Domain Dimerisation: (A) Analytical SEC for SAS-6_N at 1 mg/ml (*green*), 4 mg/ml (*pink*) and 10 mg/ml (*green*). The arrow indicates the peak shift, suggesting increased dimer formation. (B) Analytical SEC for SAS-6_N Δ 103-130 at 1 mg/ml (*green*), 4 mg/ml (*pink*) and 10 mg/ml (*green*). (C) FP signal fit for SAS-6_N (*red*) and SAS-6_N I154E (*light blue*) [103]; and SAS-6_N Δ 103-130 (*orange*).

The K_d of ceSAS-6_N dimerisation has previously been determined by FP experiments to be $\sim$ 100 μM [103]. As can be seen from crystallographic structures of the domain, the N-terminal dimerisation is mediated by the short loop β6-β7, supporting contacts of residue I154 with a hydrophobic pocket in the interface - formed by and

between helices α1 and α2. The introduction of the I154E mutation disrupts the dimerisation interaction [103]. Interestingly, in FP experiments with SAS-6_N Δ 103-130 dimerisation was also not detectable (Figure 3.4 C).

In summary, these observations show that loop 103-130 is necessary for *C. elegans* SAS-6 N-terminal domain dimerisation. Furthermore, a set of ^{15}N-HSQC NMR experiments suggests the direct involvement of loop residues 103-130 in the dimerisation interaction (Figure 3.5). Resonances assigned to residues G120, and G157 are present in the SAS-6_N spectrum at low protein concentration (25 μM), but are no longer detectable at high protein concentration (500 μM). This disappearance of resonances is indicative of the involvement of residues in μs-ms interactions. Residue G157 is part of the short loop β6-β7, which is known to facilitate the dimerisation through the neighboring residue I154 and makes contact with a hydrophobic pocket. The involvement of G157 in the dimerisation interface, in which most of the protein engages at high concentration, is reflected in the disappearance of the resonance. The resonance assigned to loop 103-130 residue G120 is affected in the same way as the resonance for residue G157 at high protein concentration. This suggests that the environment of residue G120 drastically changes at high protein concentration, which suggests that this residue is involved in interactions not present at low protein concentrations. The changing factor at high protein concentration is the increased presence of SAS-6_N dimers. In contrast, resonances for residues G111 and G128 were not affected by higher protein concentration. Even though these residues are located in loop 103-130, they do not appear to be involved in the dimerisation interaction. This might be explained by their position at the edges of the loop.

The assumption that the disappearance of resonances for residues G157 and G128 are caused by their involvement in interactions in the dimerisation interface is suppor-

ted by the ^{15}N-HSQC NMR spectrum of SAS-6$_N$ I154E (Figure 3.5). In this spectrum resonances assigned to G120 and G157 are clearly visible at comparable intensity to resonances G128 and G111 even at 500 μM protein concentration. This suggests that the effect seen at 500 μM in the SAS-6$_N$ spectrum is not due to secondary effects deriving from high sample concentration, but caused by the dimerisation event.

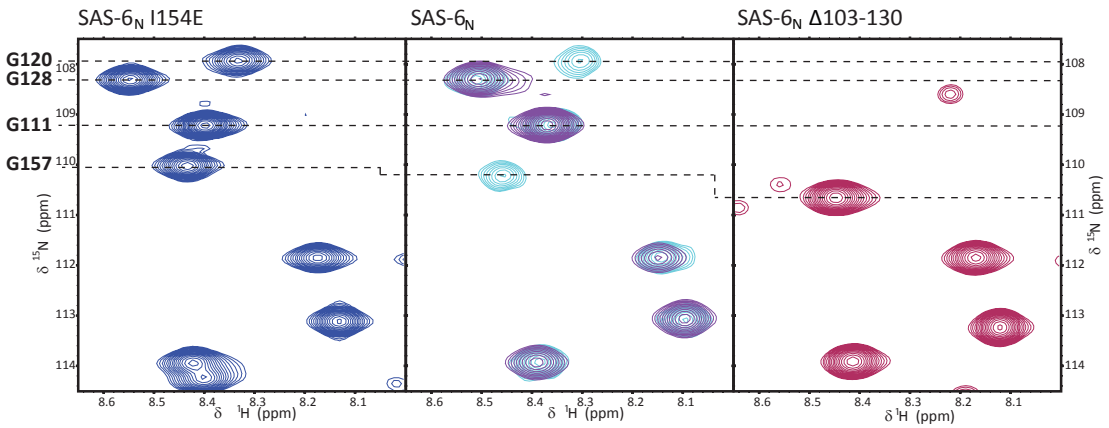


Fig. 3.5: ^{15}N-HSQC NMR Spectra of GLY Residues of SAS-6$_N$ Constructs: SAS-6$_N$ I154E at 500 μM (*dark blue*); SAS-6$_N$ at 25 μM (*light blue*) and at 500 μM(*purple*); SAS-6$_N$ Δ 103-130 at 500 μM (*red*). Resonances corresponding to residues G120, G128, G111 and G157 are indicated with a dotted line.

3.3 Simulating Loop α2-β5 and Its Role in SAS-6 N-terminal Domain Dimerisation

Further to the biophysical evidence gathered on the role of involvement of loop α2-β5 in ceSAS-6$_N$ dimerisation, we wanted to form a hypothesis about the interactions of the loop residues that aid in N-dimer formation. To gain atomic level insight into the loop dynamics, we used MD simulations. Since loop 103-130 is not resolved in existing crystallographic structures, we built a model of the loop into SAS-6$_N$ structure 4G79 (1.8 Å resolution [103]) - see Chapter 2 for details. There are fewer restrictions on the

geometry of loops than on β-sheets or α-helices. This complicates the prediction and assessment of the placement of loop residues. However, the starting position of the loop could have an effect on the energy landscape accessible during the simulation time. For this reason three models containing the dimer structure PDB code 4G79 (1.8 Å resolution [103]) and two individually modelled loops were used as starting positions for MD simulations (Figure 3.6). For each model, simulations were conducted in triplicate.

During the 50 ns simulations conducted, loops in all three chosen models rearranged into positions closer to the protein surface. After the loops adopted this proximal conformation, residues in the loop region formed interactions with helices α1-α2 of the opposing monomer. This "hugging" conformation may explain the increased dimerisation strengths observed in biophysical assays, as additional interactions could aid the dimerisation. Notably, the interactions formed between loop 103-130 and region α1-α2 were not static. Instead of settling into a stable conformation, the loop continued to move over the α1-α2 region, breaking and reforming interactions with several residues therein. Even when the simulation time was extended to 100 ns, the loop arrangement and dynamic interactions did not reach a more stable state. The highly dynamic behaviour of the loop in the simulation is consistent with the flexibility observed in biophysical experiments.

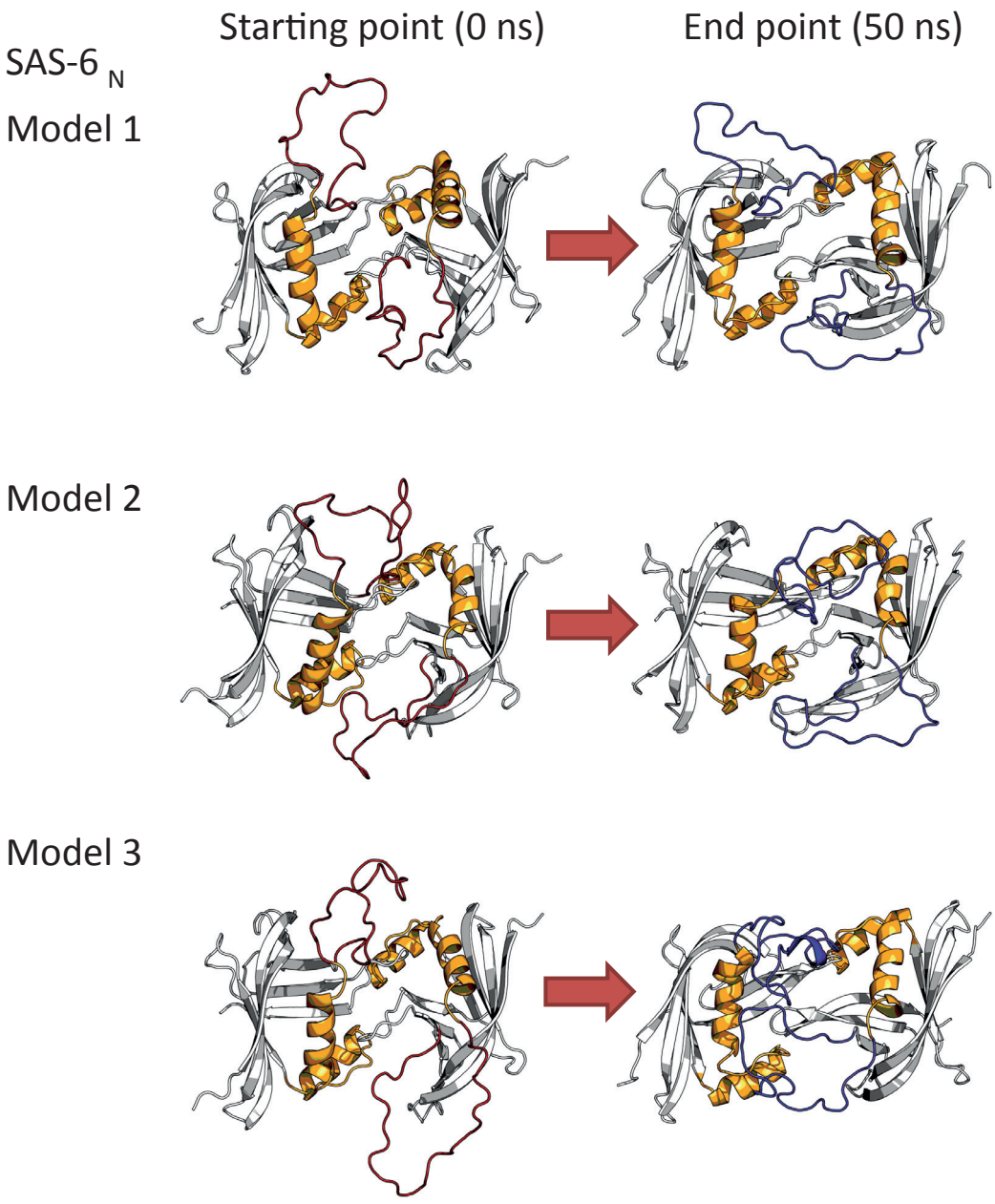


Fig. 3.6: Model of SAS-6 N-terminal Domain Loop 103-130 in MD: Structure of *C. elegans* SAS-6$_N$ dimer 4G79 incorporated in three models of loop 103-130 in SAS-6 dimer before and after 50 ns of MD. Loops in the dimer were modelled individually. Residues are highlighted in *red* for starting point and *blue* for end point of simulation. Residues of the helix region α1-α2 are highlighted in *orange*.

To quantitatively compare the loop behaviour between simulations, we analysed the time loop residues spent interacting with residues in region α1-α2 and *vice versa*

(Figure 3.7). In most simulations the majority of the residues contacted by the loop in the α1-α2 region were located in helix α1 (N70-H81) and the linker (D82-A89) connecting helices α1 and α2 (Figure 3.7 A). Residues on helix α2 (F90-N102) were also contacted in a minority of cases. The region of the loop involved in interactions with region α1-α2 includes nearly all loop residues, excluding residues at the loop ends. Residues 103-108 and residues G130 had little to no contact with the α1-α2 region. The residues spending most time in contact with the helix region are centered around residue G120 -the approximate middle of the loop (Figure 3.7 B). Notably residues G111 and G128 are amongst the residues rarely in contact with the helix region. Resonances for these two residues were seen in NMR experiments not to be affected by dimerisation, as discussed earlier, whereas G120 was.

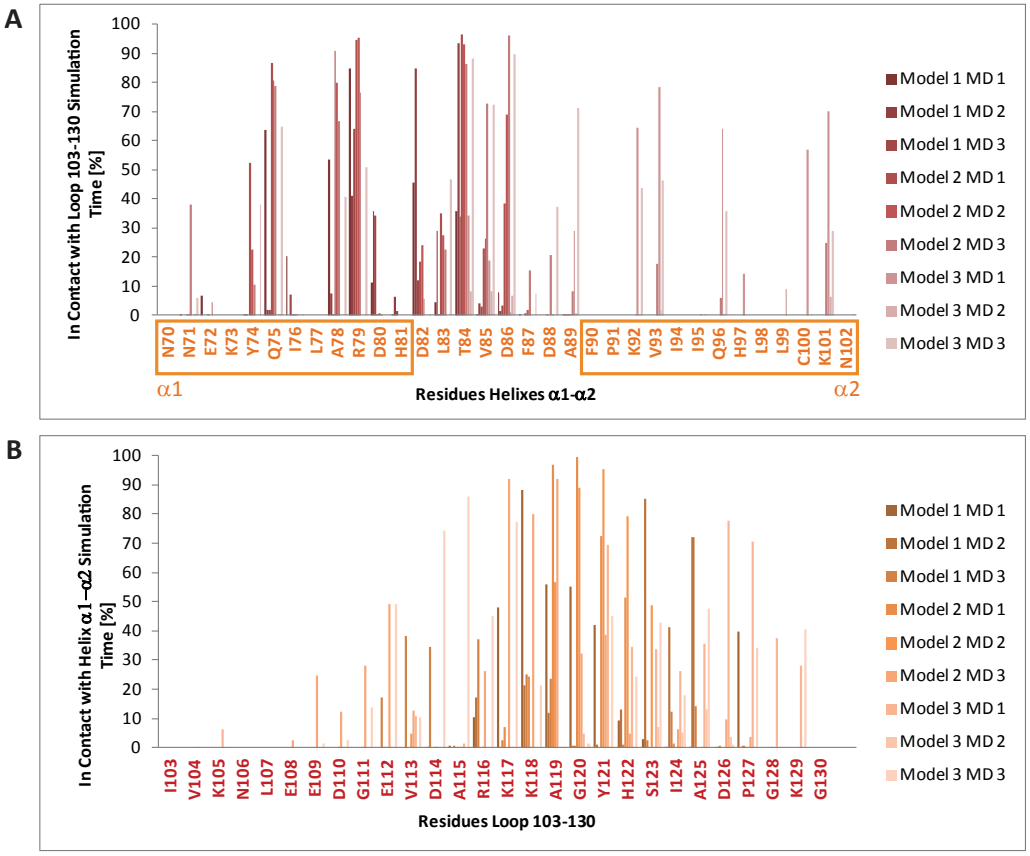


Fig. 3.7: Analysis of Interactions Formed During MD Between Loop 103-130 and α1-α2: Over 50 ns simulations conducted on three loop models with three repeats. (A) Time spent by residues in region α1-α2 in contact with residues of loop 103-130. (B) Time spent by residues in loop 103-130 in contact with residues of region α1-α2.

Loop 103-130 engages in interactions in the α1-α2 region in a similar manner in all simulations, regardless of the starting positions. The observed engagement in the dimerisation interface and the dynamic loop behaviour seem to fit with the evidence from biophysical experiments. In conclusion, we hypothesise that loop 103-130 interacts with α1-α2 region to support the N-terminal domain dimerisation. This hypothesis would need to be confirmed by further *in vitro* experiments. These could involve a mutational scan of the loop residues or the suggested helix-turn-helix motif. Part of helix α1 and the linker region between helices α1 and α2 are also involved in

the formation of the hydrophobic pocket, mediating most of the known dimerisation interface. Therefore any mutations induced in this region would have to be chosen very carefully. Due to the non-selective binding behaviour of the loop, selected mutations in the α1-α2 region might not have a visible effect on overall loop binding. Similarly, mutating selected residues in loop 103-130 might not have a detectable effect on dimerisation affinity, since most of the loop engages in interactions. One possible approach could be a simultaneous mutation of a range of loop residues around the tip of the loop with centre at residue G120. For example, constructs with alanine mutations of the region 116-127 could be used. Another way of experimentally confirming that the loop region interacts with the α1-α2 region could involve NOE based NMR or potentially FRET experiments. These techniques would not require mutations in the observed region and experiments could be performed without altering or disrupting the dimerisation interface.

3.3.1 Supporting Evidence for Loop 103-130 Involvement in SAS-6 Dimerisation from Steered MD

The classical MD simulations discussed earlier show that it is possible for the loop to play a role in the dimerisation interface by contacting residues in the α1-α2 region. The simulations also showed how the loop could stay dynamic while aiding dimerisation. Both the involvement of the loop in the dimerisation interface and the flexibility of the loop were detected *in vitro*. From FP and SEC experiments we also concluded that the presence of the loop has a strengthening effect on the SAS-6 N-terminal domain dimerisation and wanted to investigate if the simulation environment can provide support for this.

Since simulations enable us to look at molecular dynamics and interactions of sin-

gle particle systems in atomistic detail, we were interested in testing the difference in interaction strength of the N-terminal domain constructs *in silico*. Due to computational limitations the dimerisation event cannot be observed in classical MD. However, the artificial nature of the simulation allows us to directly interact with the system in ways not possible *in vitro*. Here we used SMD to pull apart the dimer during a MD simulation. Two centres of mass were defined, which incorporated all residues in a monomer, excluding the loop residues. Pulling force was applied onto these centres of mass in opposing directions. The pulling speed was set to 1 Å/ns over a 20 ns simulation. The work needed to pull gives an indication of how strong the interactions in the dimer interface were and allows us to compare the dimerisation of several constructs. Since the pulling in SMD simulations creates an artificial event, results from this simulation method have to be considered with caution. It is common practice to confirm observed results by repeating the simulation several times. In the time frame of the work for this thesis, the SMD experiments could only be performed twice. Thus, the results of the SMD simulations are discussed here as indicative evidence.

Besides the presence of the loop, mutations of the key residue I154 at the tip of a short loop (β6-β7) have been shown to have an influence on the K_d of dimerisation [93, 103]. The point mutant I154E shows no detectable dimerisation *in vitro* and the I154W mutation increases the dimerisation affinity by $\sim$ 20 fold. To test if the SMD could show the differences between the point mutations and compare their effect to the effect of the incorporation of the loop, we first performed SMD simulations with constructs lacking the loop (Figure 3.8).

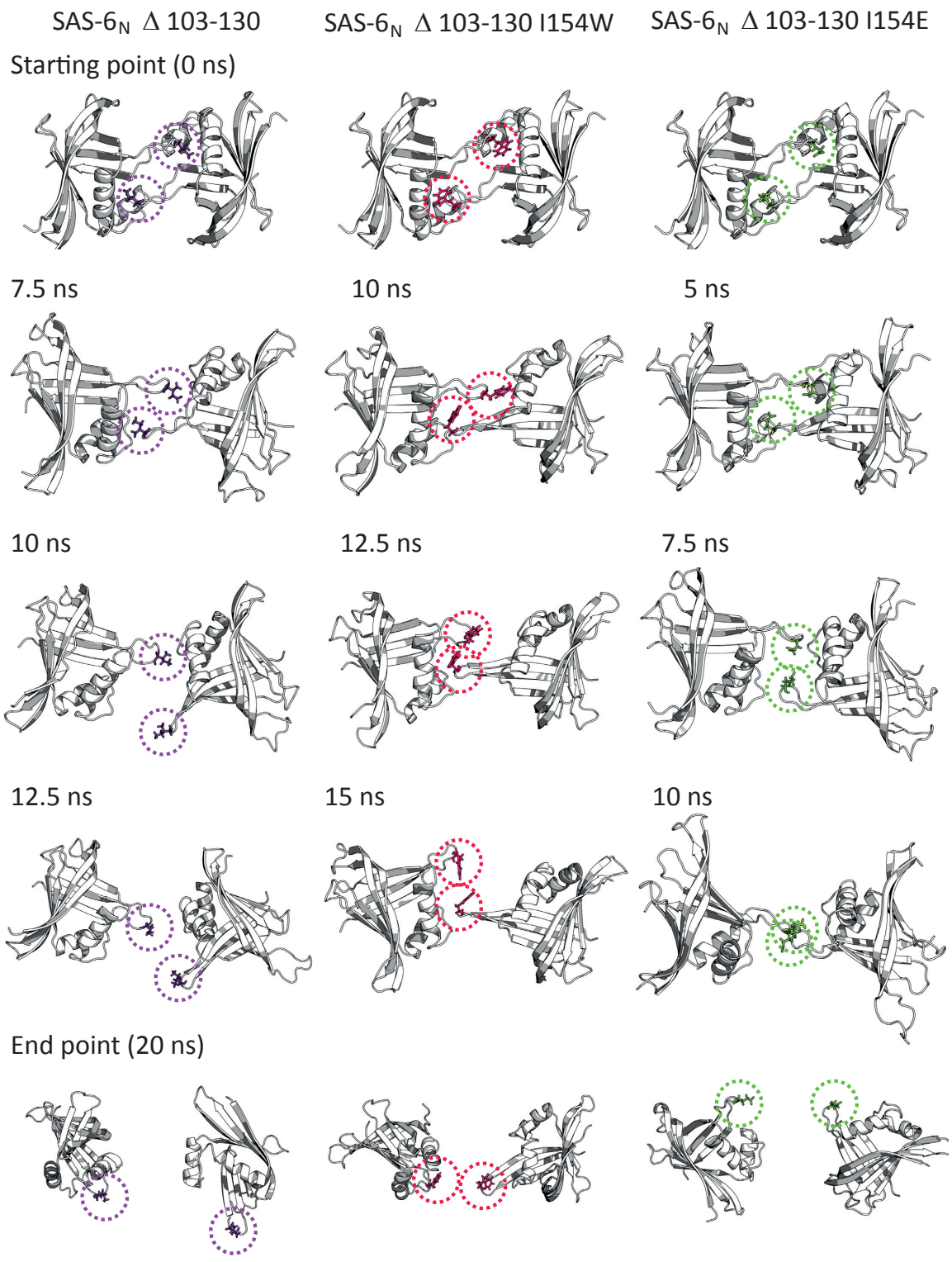


Fig. 3.8: SAS-6$_N$ Δ 103-130 Constructs During SMD Simulations: Stages of separation of SAS-6$_N$ Δ 103-130, SAS-6$_N$ Δ 103-130 I154W and SAS-6$_N$ Δ 103-130 I154E dimers during 20 ns SMD simulation. Residue I154 highlighted in *purple*, W154 highlighted in *pink* and E154 highlighted in *green*.

In the simulation of construct SAS-6_N Δ 103-130, the contact of one of the I154 residues in the dimerisation interface is broken after $\sim$ 10 ns. The dimer is completely separated after 12.5 ns. In comparison, the W154 residues of construct SAS-6_N Δ 103-130 I154W are still attached after 10 ns. The first detachment can be seen after 12.5 ns and complete separation after 15ns. For SAS-6_N Δ 103-130 I154E, the first E154 residue leaves the pocket it inhabits in the dimerisation interface after 7.5 ns, earlier than for both the I154 and W154 mutants.

The work required to pull these variants of the dimer apart is different, reflecting the mechanistic observation (see Figure 3.9). For the SAS-6_N variants with loop 103-130 deletion, the difference in work is small but discernible. In both SMD simulations performed, the I154W mutations raises the amount of work required for separation over the I154 variant ($\sim$ 25 kJ/mol). The cumulative work required to separate dimers with I154E mutation is closer to the value for I154, but still shows a trend to lower value.

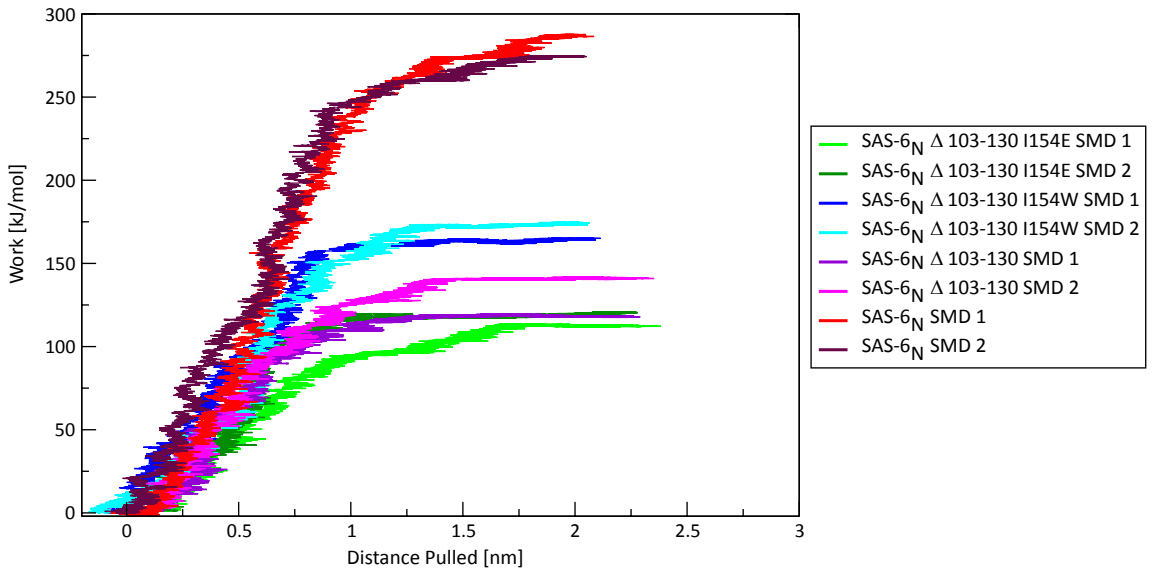


Fig. 3.9: Work Required for SAS-6_N Dimer Separation in SMD: Analysis of two repeats of simulations of SAS-6_N Δ 103-130 I154E (*green* and *dark green*), SAS-6_N Δ 103-130 I154W (*blue* and *dark blue*), SAS-6_N Δ 103-130 (*pink* and *purple*), SAS-6_N (*red* and *dark red*).

The addition of the loop causes a considerable increase of work required over all three loop deletion constructs with $\sim$ 150 kJ/mol difference between wild type and loop deletion (Figure 3.9 in *red*). This supports our assumption that the unstable loop interactions in the dimer interface are sufficient to strengthen the dimer interface. Mutants I154W and I154E were also simulated in constructs with the loop 103-130 model (Figure 3.10 and 3.11). For all loop containing constructs the SMD starting position chosen was the same as the end position of the classical MD simulations discussed earlier (Figure 3.10). This position was selected to ensure the loop was settled into an energetically favorable position. The loop residues interact with the dimer interface in these positions. It is the influence of these interactions on the discrimination strength we want to investigate.

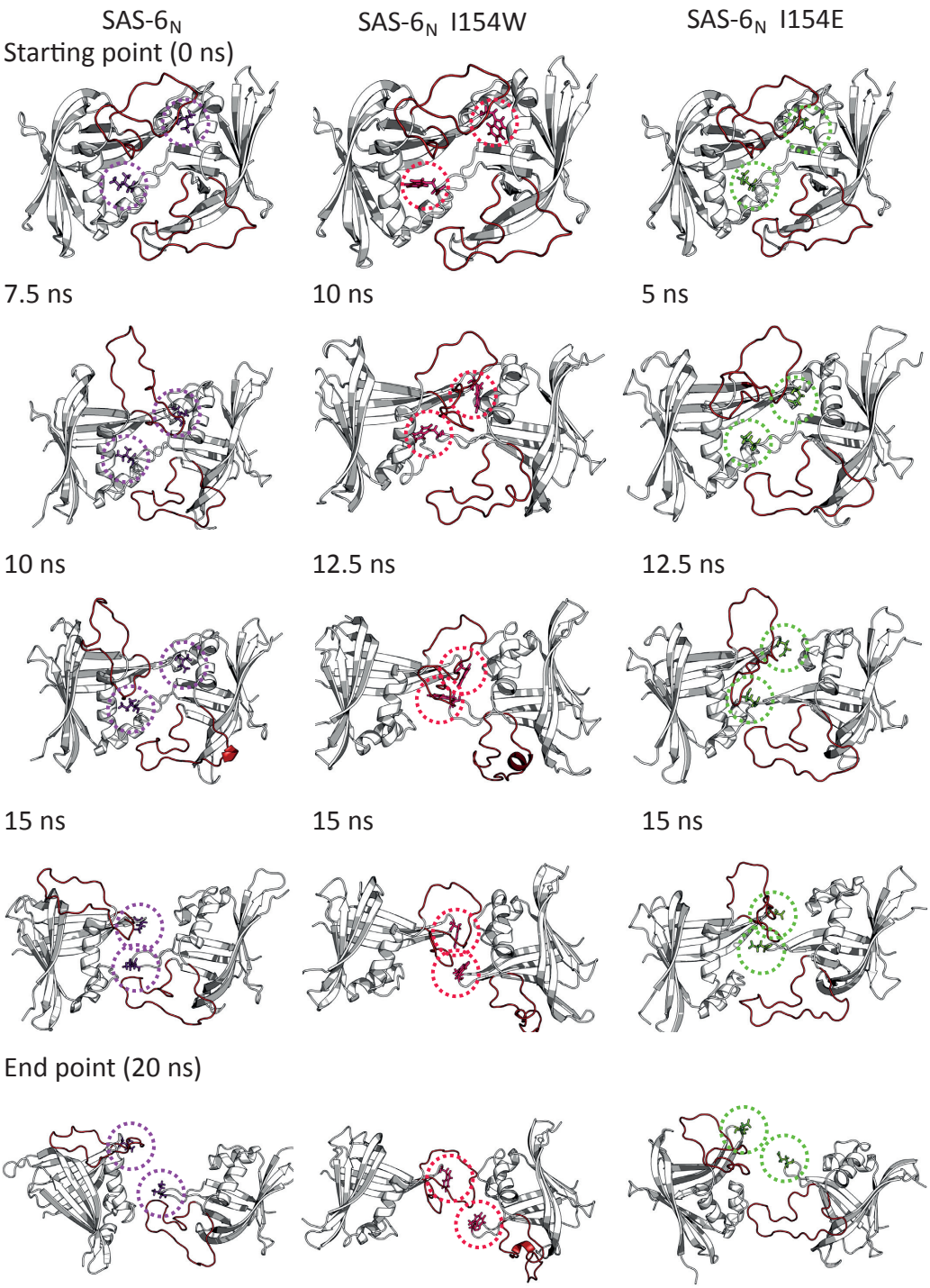


Fig. 3.10: SAS-6$_N$ Constructs with Modelled Loop 103-130 During SMD Simulations: Stages of separation of SAS-6$_N$, SAS-6$_N$ I154W and SAS-6$_N$ I154E dimers during 20 ns SMD simulation. Loop 103-130 highlighted in *red*, residue I154 in *purple*, W154 in *pink* and E154 in *green*.

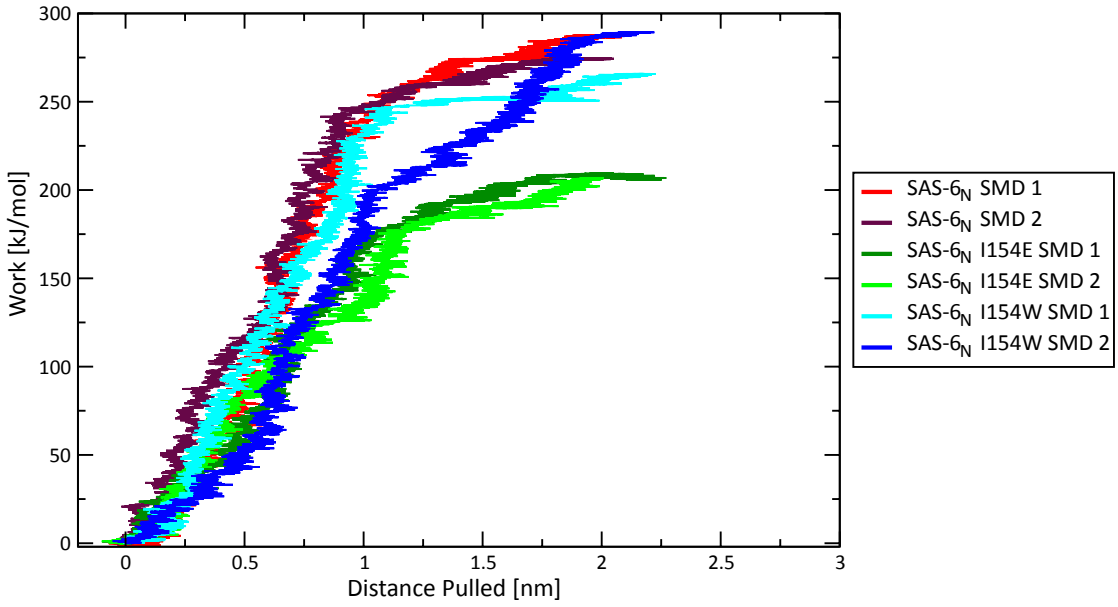


Fig. 3.11: Work Required for SAS-6$_N$ Dimer Separation in SMD: Analysis of two repeats of simulations of SAS-6$_N$ (*red* and *dark red*), SAS-6$_N$ I154E (*green* and *dark green*), SAS-6$_N$ I154W (*blue* and *dark blue*).

The wild type SAS-6$_N$ construct shows the first I154 leaving its dimerisation pocket after $\sim$ 15 ns. At this stage the loop of the monomer containing the separated I154 has already detached. The loop detachment happens after $\sim$ 10 ns. By the end of the simulation the second I154 residue has separated from its pocket, but the second loop is still interacting with parts of the α1-α2 region of the dimer interface. The W154 residues stay in interaction with the dimer interface until 15 ns into the simulation, when one of the two W154 residues detaches. This is later than the detachment of I154 and corresponds with our observation from the simulations with Δ103-130 constructs. Similarly to the SAS-6$_N$ I154 simulation, one of the loops detaches prior to residue W154 at $\sim$ 12.5 ns. After 20 ns, the second W154 residue is detached from the dimer interface, however the second loop is still interacting with region α1-α2. This incomplete separation is also seen in simulations with I154E mutation. The first E154 detaches after 12.5 ns, the second after 15 ns. The cumulative work required for the

separation of the I154 and W154 variant was similar in the simulations performed ($\sim$ 275 kJ/mol) (Figure 3.11). However, there is a clear difference to the E154 variant, where the work required for separation was notably lower, reaching $\sim$ 210 kJ/mol.

3.4 Conclusion

The analysis of NMR experiments using CSI and NOE shows that the loop region 103-130 of *C. elegans* SAS-6_N is disordered and flexible. This region is further shown to be necessary for *C. elegans* SAS-6 N-terminal domain dimerisation by analytical SEC and in FP assays. The comparison of ^{15}N-HSQC NMR spectra of SAS-6_N at low and high protein concentration, SAS-6_N I154E and SAS-6_N Δ 103-130, reveals a direct involvement of residues of loop 103-130 in the dimerisation interactions.

Using MD simulations on a dimeric structure of SAS-6_N with a model of loop 103-130, we hypothesise that the loop can interact with helix region α1-α2. The interactions formed appeared to be dynamic, which matches the high level of flexibility this loop showed *in vitro*. Furthermore, an effect of the presence of loop 103-130 on the dimerisation strength is indicated by preliminary results from SMD simulations.

Chapter 4

Virtual Screening for Inhibitors of SAS-6 Dimerisation

"... it was the season of Light, it was the season of Darkness,..."

Charles Dickens

A Tale of Two Cities

This chapter describes the first steps taken to find a small molecule inhibitor for the dimerisation of the *C. elegans* SAS-6 N-terminal domain. Two approaches to virtual screening for computational compound discovery are described and compared.

External contributions to work in this chapter: (1) NMR experiments on the effect of free tryptophan and cyclic peptide were not performed by the author of this thesis. (2) The pharmacophore design and screen was performed by Dr. Minos Matsuokas with consultative input by the author on prior data and experience.

4.1 Introduction

Control of expression levels of cell cycle related protein, such as the proteins linked to the centriole duplication pathway in cell biological experiments is crucial to reliably determine cause and effect in such experiments. To study elements of the cell cycle, often modified versions of the proteins studied are introduced into a system which can cause protein levels to deviate from the norm. Changes in protein levels of many of the factors involved in the centriole duplication pathway have been shown to influence centriole duplication [101, 215]. This complicates the study of the role of the individual components of the pathway. Using small molecules to modify the function of the proteins in this pathway might provide a way to study their role without interfering in the regulatory processes that control other protein levels and function.

4.1.1 Target Selection

This study aims to find compounds for the inhibition of SAS-6 oligomerisation, primarily for use in centrosome and cell cycle research. With this in mind, the SAS-6 variant chosen for initial screening is the *C. elegans* SAS-6. *C. elegans* is historically an important model organism for research into cellular processes (reviewed in [216, 217] and [218]). Even though elements of the centriole duplication pathway in this organism differ from other organisms, the general mechanism is similar and key player proteins are conserved to an extent that findings from this organism can be transferred to other systems. There are well established techniques to study cell division, centriole localisation and architecture in *C. elegans*. Another advantage of this model system is the existence of established research techniques and the simplicity of the system. Not being ciliated or flagellated, the only centrioles present are centrosomal, making defects easier to track through the cell cycle.

Even though the *C. elegans* SAS-6 structure is used, the conclusions we draw from this screen will be applicable to SAS-6 from other species, since SAS-6 variants have high structural similarity - as discussed in the introductory chapter of this thesis.

In targeting the oligomerisation of SAS-6 there are two dimerisation interfaces to consider. The coiled-coil dimer interface covers a large surface area, and furthermore the coiled-coil interaction is the stronger of the two dimerisation events, with a K_d of $\sim$ 1 μM compared to a K_d of $\sim$ 100 μM for the N-terminal dimerisation [93, 103]. SAS-6 is believed to predominantly exist as C-C dimer in the cytoplasm leaving the N-terminal domain interface exposed prior to oligomer assembly [145]. These factors make the N-terminal dimerisation domain the preferable choice for inhibition attempts.

The SAS-6 N-terminal domain features a hydrophobic pocket, the sides of which which are formed by helix α1 and α2. Crystal structures of the N-dimer reveal that the pocket forms a major component of the dimerisation interface (Figure 4.1 A and D). The dimerisation is achieved by the insertion of a key hydrophobic residue on loop β6-β7 into this pocket.

Comparing the loop-pocket interactions in the structure of one of these constructs with the wild type and *C. elegans* SAS-6 structures; both in wild type and the strengthened I154W mutant [93, 103], some supporting interactions can be identified (Figure 4.1 B and C).

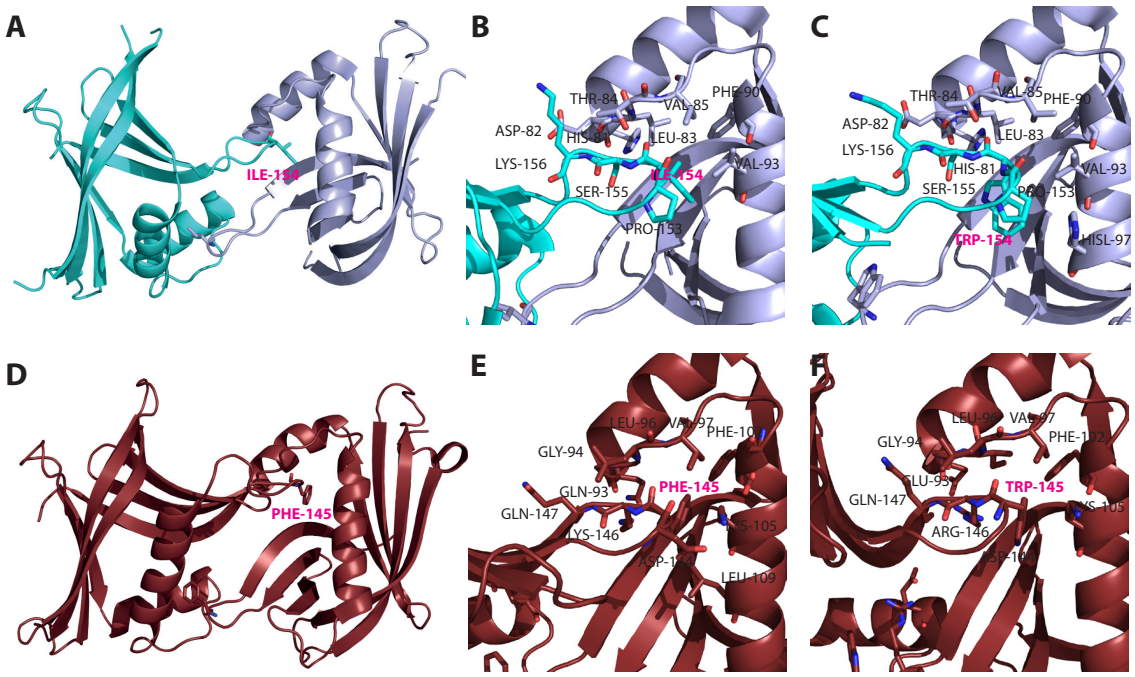


Fig. 4.1: SAS-6 N-dimer and Interface: (A) *C. elegans* N-dimer as present in crystallographic structure PDB code 4G79 (1.8 Å resolution [103]). (B) Interface residues in crystallographic structure PDB code 4G79 (1.8 Å resolution [103]). (C) Interface residues in crystallographic structure PDB code 4GEU with I154W mutation (2.65 Å resolution [103]). (D) *C. reinhardtii* N-dimer as present in crystallographic structure PDB code 3Q0Y (2.1 Å resolution [93]). (E) Interface residues in crystallographic structure PDB code 3Q0Y (2.1 Å resolution [93]). (F) Interface residues in *C. reinhardtii* crystallographic structure PDB code 4T07 (1.25 Å resolution [160]). Mutations present are Q93E F145W and K146R; this combination of mutations is also called NN24.

In the *C. elegans* structures, backbone interactions between the loop residues S155-K156 and residues D82 and T84 in the α1-α2 linker that forms the upper rim of the hydrophobic pocket can be seen. Additionally, P153 on the side of the key residue I154 is most likely to be stabilising the positioning of this residue. Compared to the I154W structure, both I154 and W154 seem to form general hydrophobic interactions with the pocket. Whereas the sidechain of I154 points into the hydrophobic pocket, the bigger W154 residue aligns lid-like against the pocket in the crystal structure.

A recent study reveals that mutations in *C. reinhardtii* SAS-6 influence centriole

symmetry [160]. Some of the constructs tested contain modifications in the dimerisation interface. Comparing the changes helps illustrate the interactions supporting the dimerisation mediated by the hydrophobic dimerisation pocket (Figure 4.1 E and F). Here the key residue contacting the hydrophobic pocket is F145. Both F145 and the mutant W145 residue align with the α2-helix and the side chain of K105 on this helix. The stabilising proline residue is not present next to the key residue, however here D144 supports the position of residue 145 by interacting with the end of the L105 side chain. The observed additional stabilisation over a simple F145W mutant construct NN24 is likely to stem from mutations Q93E and K146R that in the NN24 structure face each other to form an additional side chain interaction.

These structures highlight that a rigid positioning of the interactions of the key residue is a crucial feature of the dimerisation interface. The strengthened dimerisation affinity the tryptophan mutants exhibit over the isoleucine and phenylalanine wild type constructs, highlight the impact of hydrophobic interactions within the hydrophobic pocket. The factors that stabilise the positioning of the key residue differ between the two species discussed, which implies that it might not be essential to mimic particular interactions as long as a certain rigidity is achieved. At the same time, the small differences observed between these interfaces could be utilised to design an inhibiting compound.

4.1.2 Screening Method Selection

As the central structural component of the forming centriole, the SAS-6 oligomerisation is an interesting target for inhibition. However, protein-protein interactions are viewed as non-conventional targets for inhibition with small molecules and are often described as "undruggable".

The prediction of the druggability of a protein is mostly based on the presence of a pocket suitable for inhibitor design. Conventional drug targets and the druggable proteins have been reviewed by a variety of authors, see [219] and [220] as example. Typical druggable targets have deep accessible pockets, often with a well defined reaction centre that can be specifically inhibited by a small molecule with a binding area of $\sim$ 300 - 500 Å^2. Protein-protein interfaces on the other hand tend to span large areas of $\sim$ 1,000 - 2,000 Å^2 with many residues contributing to the interaction strength, making the design of specific inhibitory molecules a complex task. Even though the label "undruggable" suggests a certain finality, targets that fall in this category contain a large number of potential clinical targets and have received more attention in recent years. The decline in discovery of conventional drugs and drug targets has inspired research into "undruggable" targets (for comments on this see [221, 222]). Several recent reviews discuss difficulties in inhibitor design for these "undruggable" protein-protein interactions and their relevance for modern day drug discovery (for examples see [223–227]). Amongst the protein-protein interaction targets that have been the focus of drug discovery research are several bromodomain proteins and HIV integrase, drug candidates for which have been developed as far as clinical trials. This recent interest has also brought about a surge in the development of methods for small molecule discovery for protein-protein interaction targets.

The search for a novel small molecule inhibitor often begins with a broad screen of a large variety of compounds. These high-throughput screens are designed to cover a large area of chemical space. As a result of these screens, new starting points for drug development can be found and information about the functional groups that might be compatible with the target identified. Conventional high-throughput screening approaches are typically considered less promising for undruggable targets. Compound libraries used for these screens mostly contain compounds that were developed with

conventional drug targets in mind. The bias inherent to compound libraries means that in the search for small molecule inhibitors and modulators targeting protein-protein interactions, new methodology has to be applied to search chemical space. Moreover, whereas high-throughput screening assays for conventional drug targets often involve the detection of altered enzymatic activity, assays for identification of small molecules targeting undruggable targets measure the inhibition of protein-protein interactions. Biophysical techniques used for high-throughput screens to detect changes in protein-protein interactions include thermal shift assays, fluorescence polarisation assays and surface plasmon resonance. In a high-throughput manner the biophysical assays used here are more prone to false positive results caused by unwanted oligomerisation, protein aggregation or allosteric binding. Low-throughput methods that allow the study of protein-protein interactions in structural detail like NMR, X-ray crystallography, or isothermal calorimetry, then need to be used to distinguish true hits from false positive results.

The compound libraries used for screening also require consideration, as further limitations in the search for inhibitors of protein-protein interactions come with small molecule design. A selection of small molecules that can be considered drug-like underlies the "Lipinski rule of 5" [228]. Even though, protein-protein interfaces are large, the inhibitor size cannot exceed a certain molecular weight. The weight limit is set at around 500 Da for solubility properties of the compound, but also to make synthesis of the compound feasible. If the small molecule is to be used *in vivo* or considered as a drug candidate, the compounds have to be soluble in aqueous solutions, but also need to be able to cross the cell membrane. The log$P$ - where the partition coefficient is a ratio of the solubility of a compound in octanol and the solubility of the compound in water - of the compound to be considered, should be lower than 5. The small molecule

property recommendations are known as the "Lipinski rule of five" and also include guides for a favourable number of hydrogen bond donors and acceptors. These rules were set up on the basis of classical drugs and even for conventional targets should only be used as guidelines; there are non-Lipinski compounds that are approved drugs. The "Lipinski rule of five" can be summarised as follows:

- a molecular mass of $\leq$ 500 Da
- log$P \leq 5$
- $\leq$ 5 hydrogen bond donors
- $\leq$ 10 hydrogen bond acceptors

Using fragment based compound libraries has proven effective for high throughput screens of protein-protein interaction targets [229, 230]. Compounds that are defined as fragments are designed to comply with the "rule of three" [231] :

- a molecular mass of $\leq$ 300 Da
- log$P \leq 3$
- $\leq$ 3 hydrogen bond donors
- $\leq$ 3 hydrogen bond acceptors

Fragment compounds that bind even weakly to the target protein are considered a hit and are then used as a building block from which a final candidate compound is designed. In a fragment based screen an even larger portion of chemical space can be covered due to the combinatorial possibilities for which the fragments sized compounds allow. The main reason fragment screens are particularly suited for protein-protein interfaces are so called "hot-spots" that are often present in protein-protein interfaces (as reviewed for example by [226, 227, 232, 233]). These are small patches with

localised compound binding affinity. Blocking these patches with a well designed candidate compound can make it possible to block the protein-protein interaction with a Lipinski conform candidate compound. Hot-spots can be identified through fragment based screens; combining fragments identified to bind to several hot-spots then creates a candidate compound binding specifically to the protein-protein interface.

In vitro high throughput screening requires the existence of specialised equipment, compound libraries and considerable amounts of purified protein. Instead of spending these resources on initial screening against an undruggable target, computational screening techniques are often used to gather initial information. Virtual screening can be used to cover a wide range of chemical space in a time efficient manner. Several approaches to computational drug discovery are in common use; in this study fragment based and pharmacophore based virtual screens were used (reviews on computational drug discovery include [234, 235]). Since computational resource are ever becoming more powerful and available, the use of high-throughput virtual screening has also become a standard approach in industry and more recently in academic research. Considering the relatively recent advance of computational drug discovery methods, they have been used with some success (as reviewed in [236]).

During a virtual screen, a library of small molecule structures are docked to a protein structure of interest. The docking positions are scored and ranked to select a smaller subset of small molecule candidates. The use of a large compound library balances the high false positive and false negative rate that results from the low accuracy of the docking prediction. While hits from this method are likely to contain a high number of false positives, it can be used to create sub-sets of compounds for *in vitro* screening and reveal characteristics of compounds that will benefit the design of candidate compounds.

Docking is a computational technique that tries to model the position of a ligand on a protein structure (as reviewed in [237]). Structural and charge distribution information are used to maximize an affinity score function. It should be noted that the algorithms used for docking conventionally rely on simplified charge distributions and steric hindrance models to find binding poses, and scoring functions often favour computationally fast approaches over quantum mechanical calculations of higher complexity, which allow prediction of interaction affinities with higher accuracy. Although not able to reliably predict small molecule binding affinities, docking can be used to attain information about the general compatibility of ligands and potential binding sites when detailed experimental analysis of the binding event is not possible or feasible.

In a virtual screen, docking results are ranked based on the docking score. Since the docking algorithms utilised for high throughput screens are kept basic in favour of computational speed, the scoring is crude. Compound selections derived from virtual screening are used as enriched libraries for further *in vitro* screens to verify hits.

When targeting protein-protein interactions with virtual screening methods, hotspot identification can be pursued with fragment based libraries [238]. For docking approaches like this, the entirety of the protein-protein interaction surface needs to be covered by the docking search. One example of successful fragment based virtual screening approach in protein-protein inhibition is the identification of inhibitory compounds against the HIV target protein HIV-1 Nef [239, 240].

A further computational method used in this chapter to sift through chemical space is a more ligand based variety of virtual screening which utilises a so called pharmacophore to find matching compounds in a library (reviewed by [196]). The pharmacophore

is a three dimensional scaffold of chemical characteristics that is built on the basis of compounds with known affinity for the target and their binding position on the target. The pharmacophore scaffold is then used as a template to which the compounds in the screening library are matched. Results are typically scored for their compliance with their pharmacophore characteristics. The outcome of a pharmacophore screen is again a compound library enriched with potential hit compounds that then can be used for *in vitro* screening.

4.2 Results & Discussion

4.2.1 Preliminary Experiments

Since the I154W mutation increases N-terminal dimerisation in *C. elegans* SAS-6 [93, 103], tryptophan is a logical starting point in the search for a small molecule able to bind to the hydrophobic pocket.

An NMR titration series conducted* with a monomeric version of the *C. elegans* SAS-6 N-terminal domain (SAS-6 1-168 Δ103-130 I154E) does not show significant spectral changes, even at high concentrations of free tryptophan. The monomeric ceSAS-6_N is used in these experiments, as the hydrophobic target pocket is accessible for binding and no effects on the experiment from dimerisation have to be considered. Figure 4.2 A shows the resulting spectrum overlay.

*This experiment was conducted by Dr. Michele Erat, former postdoctoral researcher in the group of Dr. Ioannis Vakonakis.

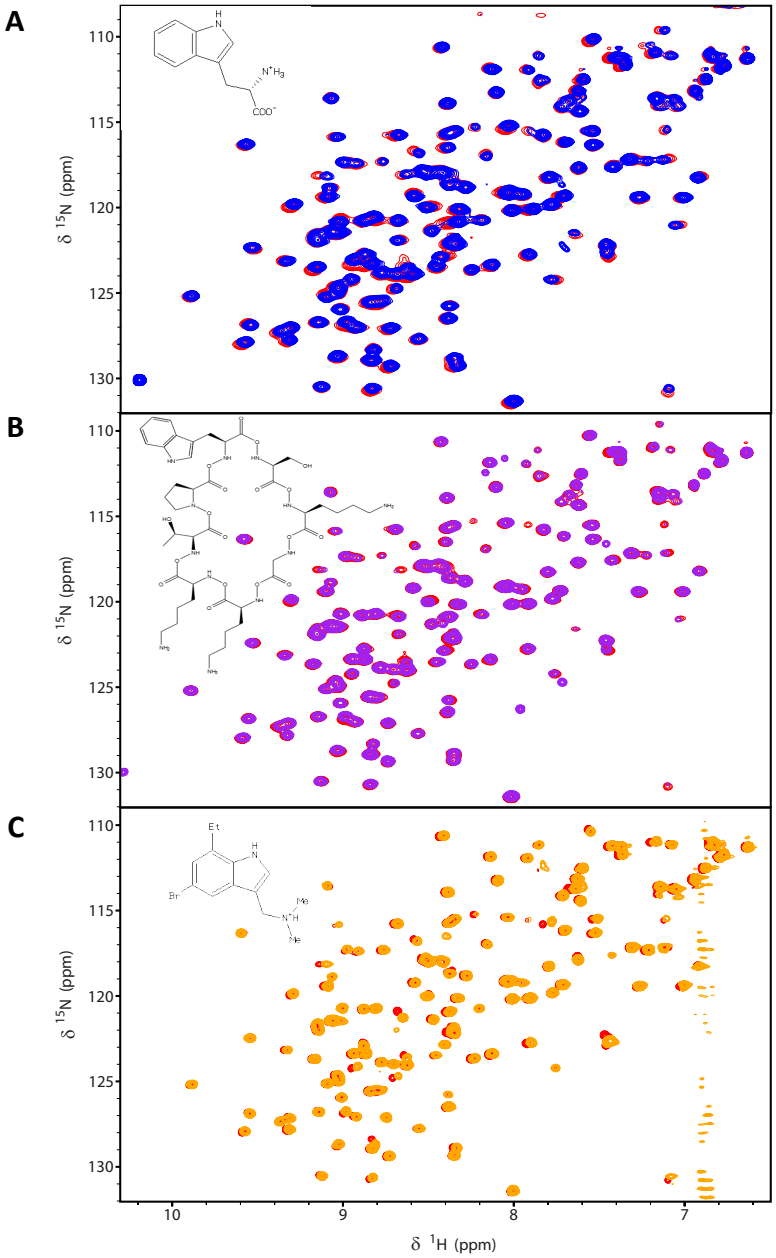


Fig. 4.2: Overlay of ^{15}N-HSQC NMR Spectra - *C. elegans* SAS-6 N-terminal Domain with Tryptophan, Cyclic Peptide and ZINC00392813: *Red*; spectra of ^{15}N SAS-6 1-168 Δ 103-130 I154E overlayed with: (A) *blue*; SAS-6 spectrum with 38.6 mM tryptophan, (B) *purple*; SAS-6 spectrum with 8 mM cyclic peptide KTPWSKGK. Additional resonances originate from added peptide due to high concentration and natural abundance of ^{15}N. (C) *orange*; SAS-6 spectrum with 2 mM ZINC00392813.

Chemical shift assignments for the *C. elegans* SAS-6 spectrum exist[†], so most residues can be associated with specific resonances and the effect of added tryptophan mapped onto the SAS-6 structure (Figure 4.3 A and B). In addition to showing only small resonance changes, residues affected do not localise to the hydrophobic pocket. From this we can conclude that free tryptophan does not bind to the hydrophobic pocket with specificity.

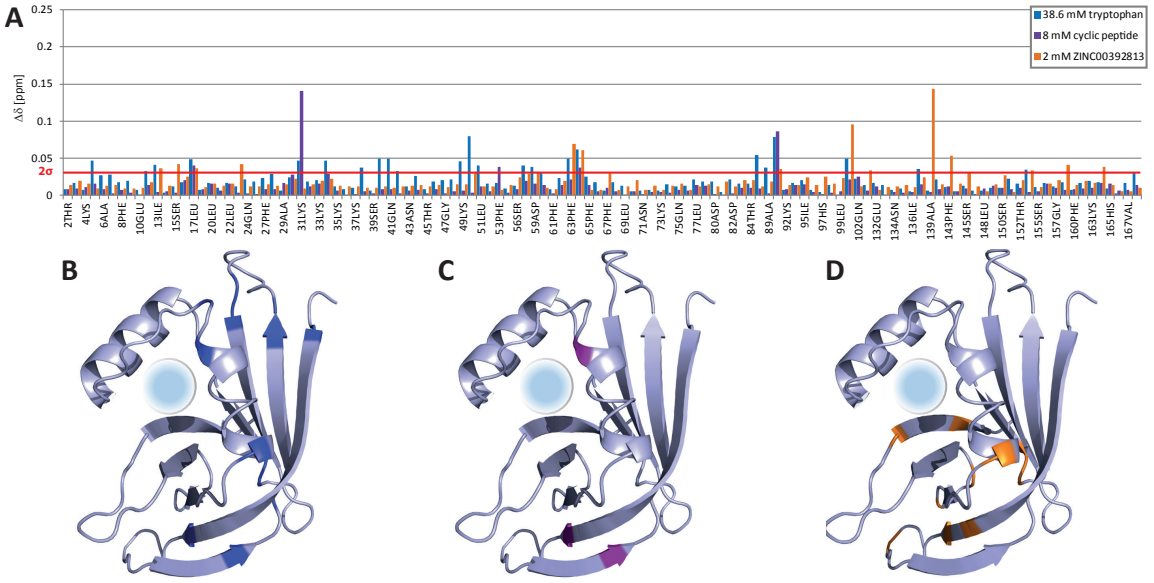


Fig. 4.3: Chemical Shift Perturbation of Preliminary SAS-6 NMR Experiments: (A) Bar chart showing chemical shift perturbations as a function of residues in ^{15}N SAS-6 1-168 Δ 103-130 I154E upon addition of: *blue* 38.6 mM tryptophan, *purple* 8 mM cyclic peptide KTPWSKGK and *orange* 2 mM ZINC00392813. 2σ indicated by *red* line is 2σ value indicator for all three experiments (2σ for tryptophan experiment = 0.031 ppm, 2σ for peptide experiment = 0.031 ppm, 2σ ZINC00392813 = 0.035 ppm). (B) Residues with notable (over 2σ) chemical shift perturbation upon addition of 38.6 mM tryptophan mapped onto SAS-6 structure. (C) Residues with notable chemical shift perturbation upon addition of 8 mM cyclic peptide KTPWSKGK mapped onto SAS-6 structure. (D) Residues with notable chemical shift perturbation upon addition of 2 mM ZINC00392813 mapped onto SAS-6 structure.
Chemical shift perturbations are small and do not occur consistently for residues in the target pocket region - pocket marked by blue sphere.

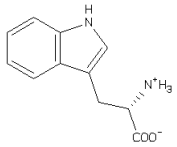
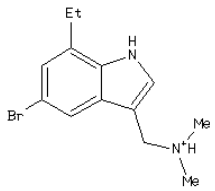
[†]Assigned by Dr. Michele Erat, former postdoctoral researcher in the group of Dr. Ioannis Vakonakis.

Another preferred approach to finding inhibitors for protein-protein interaction is the creation of a small peptide that mimics the interactions provided by the binding partner (reviewed by [241]). In this case a cyclic peptide was designed based on loop β6-β7 containing the key residue I154W (KTPWSKGK). The peptide was tested with NMR against the earlier mentioned SAS-6 construct[‡]. The spectrum overlay (Figure 4.2 B) also shows only small changes in resonance position upon addition of high concentrations of peptide. Additional resonances appearing can be explained as signal from naturally occurring ^{15}N in the cyclic peptide. When mapped onto the structure of SAS-6, the residues affected again do not seem to cluster in a recognisable manner at the target pocket (Figure 4.3 A and C). Again, the conclusion must be that this peptide does not bind to the hydrophobic pocket.

In 2014 a study identified a small molecule which showed signs of binding to the hydrophobic pocket of *L. major* SAS-6 in NMR experiments and was able to influence oligomerisation of *L. major* SAS-6 at millimolar concentrations [95]. Compound PK9119 (ZINC00392813) strongly resembles tryptophan (see Figure 4.2 or Table 4.1), through the indol ring that makes up most of its structure. It is further decorated with an ethyl and a bromide group, and differs from an amino acid mostly through the tertiary amine that replaces the amino group.

[‡]This experiment was conducted by Dr. Michele Erat, former postdoctoral researcher in the group of Dr. Ioannis Vakonakis.

Table 4.1: Control Substances for Docking Parameters

Control Substances for Docking Parameters	
Compound	Vina Score
ZINC00083315 (tryptophan) 	-6.4
ZINC00392813 	-6.4

The binding position of this compound in the *L. major* SAS-6 binding pocket is not clear from the published study, as no structure of the complex exists. The localisation of NMR resonances affected in the hydrophobic pocket rims suggests that the compound binds to the pocket. The presence of the compound mostly affects residues in the α1 helix at the rim of the pocket and the β7 strand which forms another pocket boundary. However, the exact position of the compound and molecular interactions cannot be derived from this.

We tested the effect of this compound on *C. elegans* SAS-6. However, at the same concentration as used in the *L. major* SAS-6 study, the compound did not show binding to *C. elegans* SAS-6. The resonance differences observed were small and mapping affected residues onto the *C. elegans* SAS-6 structure shows no binding to

the affected regions described in the *L. major* variant.

4.2.2 Virtual Screening for Compounds Binding *C. elegans* SAS-6

Based on the evidence gathered in preliminary experiments we aimed for a novel, non tryptophan-like compound able to bind to the *C. elegans* SAS-6 hydrophobic pocket. Since obvious starting points for structure based design, such as the key residue containing loop β6-β7, tryptophan and the related small molecule as candidate compounds did not show expected behaviour, new insight into the pharmacophoric characteristics of the target pocket was needed. We turned to computational approaches often used in structure based drug design. This enabled us to search through a large portion of chemical space to select promising compounds for *in vitro* experiments.

4.2.2.1 Fragment Based Virtual Screening

To facilitate parallelised high throughput virtual screening of a large number of small molecule structures on local computational clusters, a script based pipeline was developed as part of the work described here. Details of the pipeline are described in the methods section of this thesis and scripts are included in Appendix B.

Virtual screening has a high false positive and false negative rate. To combat this and to increase the chance to find a wide variety of compounds for *in vitro* screening, large compound libraries are docked in virtual screens. Even though the individual docking event requires only little computational time, considerable computational resources are required to facilitate high throughput docking in a reasonable time frame. Computational methods for drug discovery have become more wide spread with the

increase in availability of computational power. Docking pipelines that enable high throughput virtual screening - once the domain of the pharmaceutical industry - are now feasible to run on academic resources.

There are a number of docking programs available which differ in the underlying principle of the scoring algorithm used and their licensing. Commercial programs often come with an inbuilt "screening" function, but tend to have license limitations. This implies that individual docking events have to be calculated in series, whereas several docking events could conceivably also be calculated on several computational cores in parallel at the same time - given the availability of a large number of cores. Another disadvantage of commercial products is that their documentation is often not transparent. Open source docking programs do not usually have inbuilt "screening" options, meaning individual docking runs have to be performed manually for each compound.

Pipelines utilising open source docking programs have been developed for academic use, however the scripts and set-ups are often restricted to the use of the developing department, specific to a certain computational cluster or even the size of the compound library (examples include *DOVIS* [242], *APPLIED* [243] and *VinaMPI* [244], also reviewed in [234, 245]). Setting up a custom pipeline enabled us to choose the docking and ligand preparation software, adapt the pipeline to a potentially limitless library size, and parallelise the screening process to maximise the use of the available computational hardware (Figure 4.4 and see Chapter 2.6.2.1 for details and Appendix B for full scripts).

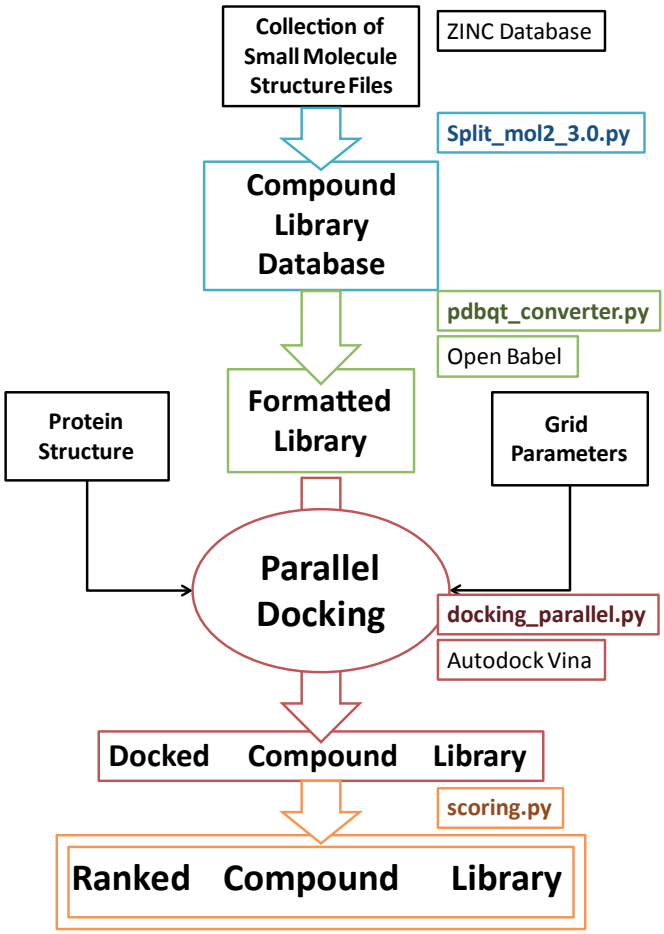


Fig. 4.4: High Throughput Docking Pipeline Set-up: Python scripts indicated by ending ".py" were developed for the work presented and can be found in Appendix B.

The pipeline described here uses *Autodock Vina* [190] as its docking program. *Vina* is an open source docking software which in comparative studies has been found to give results with a similar or even higher predictive quality than commercial docking suites [246, 247]. It uses a machine learning based scoring function to identify the best fit between ligand and receptor. The score calculated by the docking program is unit-less, but inspired by binding free energy. The more negative a Vina score is, the stronger the predicted interaction. Scores ranging from 0.0 to -10.0 can be expected. It should be noted that scores derived from docking programs are not predictions of

experimental vales, but are used purely as a way of comparing models.

The pipeline also facilitates the set-up of a compound library database, which circumvents a number of operating system related issues that can occur when handling a large number of files. A conversion step for compound structure files contained in the database to the format required by Vina is included. After the compound library has been docked, the pipeline includes steps to rank all docked compounds by Vina score and extract the top ranked compounds from the database.

The fragment based screening approach is thought of as particularly suited for the search for compounds which bind to protein-protein interfaces. Using fragment sized molecules in the docking library can reveal the hot-spots on the interaction surface which have increased affinity for compound binding. Several of these hot-spots can then be targeted at the same time by combining several fragments into one candidate compound. To gain insight into possible pharmacophoric elements and profile the druggability of the target pocket, a fragment based selection of compounds with low molecular starting weights of under 250 Da was chosen. This allows identification of low affinity binders interacting with different pocket features that can then be combined to result in a higher affinity ligand. Screening for small fragments reduces the complexity of the chemicals that need to be screened while also increasing the area of chemical space covered through the possibility of combining hit fragments into specific, strong binding substances. In addition, the chosen compound library is selected using criteria from the "rule of three". The 527,585 fragment containing compound library used here is the fragment subset of the ZINC database; a repository of compounds that are or have been available for purchase [194].

Another consideration for a fragment based virtual screen is the area of protein

that can be covered in the docking search. Docking time increases when the protein surface area included in the dock is increased, however in order to reveal interface hot-spots the whole interface needs to be covered. In addition, the inclusion of more protein surface in the docking search can be used to exclude compounds that might have an increased affinity to patches of the protein not located in the interface. To find a compromise between computational time and covered surface area, a so called docking "grid" is used, the centre and dimensions of which are defined prior to the screen (see Chapter 2.6 for details).

The monomer of the *C. elegans* SAS-6 N-terminal domain PDB code 4G79 (resolution 1.8 Å [103]) was used. This structure was derived from a SAS-6 construct lacking the α2-β5 loop. The construct ranges from residue 1 to 168 with a loop deletion of residues 103-130. This construct has been found to be stable during purification and the dimer contacts are present in its crystal structure.

The grid parameters required for docking were defined manually. The centre of the grid was set to the location of the I154 residue in the dimeric form of the SAS-6 structure, while the dimensions of the grid where maximized to include as much of the protein as possible and to sift out compounds with preference for other regions on the protein. The grid parameters were then tested for compound localistion by docking the structures of tryptophan (ZINC00083315) and ZINC00392813, the previously discussed *L. major* SAS-6 binding compound. Although these molecules did not measurably bind to the pocket *in vitro*, they have characteristics that make them reasonable candidates for structure based design. The docking algorithm is based on possible atomic intermolecular interactions. Since the pocket can be occupied by tryptophan - as shown in I154W and F145W mutants - if configured correctly we would expect these compounds to be positioned in the target pocket after docking, based

on the potential for intermolecular interactions between tryptophan and the pocket. However the score of the docking pose would be expected to be lower than for a true hit, since the intermolecular interactions possible between tryptophan and the pocket are evidently not enough for *in vitro* binding. Similarly, ZINC00392813 has the potential to fit into the pocket as it is related to the tryptophan structure, so the docking program should be able to place the compound into the pocket, even though the Vina score would be expected to reflect a difference between this compound and a compound that binds to the pocket *in vitro*. The chosen grid configuration and further parameters that were used to configure the docking process are listed in Section 2.6.2 of this thesis.

Docking results for the control substances can be seen in Figure 4.5. The Vina affinity scores are in the medium range on the possible 0 to -10 scale. For the purpose of configuring the docking parameters, the placement of the test substances in the pocket is the deciding criterion. The grid parameters have no influence on the Vina score attributed to a docking position. Figure 4.5 evidences that both ZINC00083315 and ZINC00392813 were docked successfully into the target pocket.

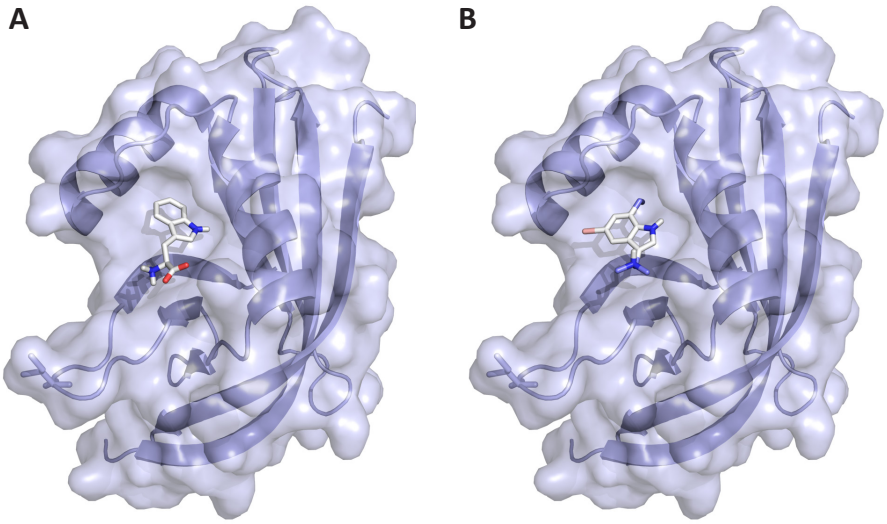


Fig. 4.5: Docking Parameter Test - Control Substances in Docking Position: (A) Tryptophan (ZINC00083315) and (B) ZINC00392813 localise in the target pocket.

To analyse the high-throughput virtual screen results, the docked compounds were sorted by their associated Vina affinity scores - with the most negative scores being ranked higher - and the top 10,000 scoring molecules extracted. Figure 4.6 shows the Vina scores of the best scoring 10,000 compounds.

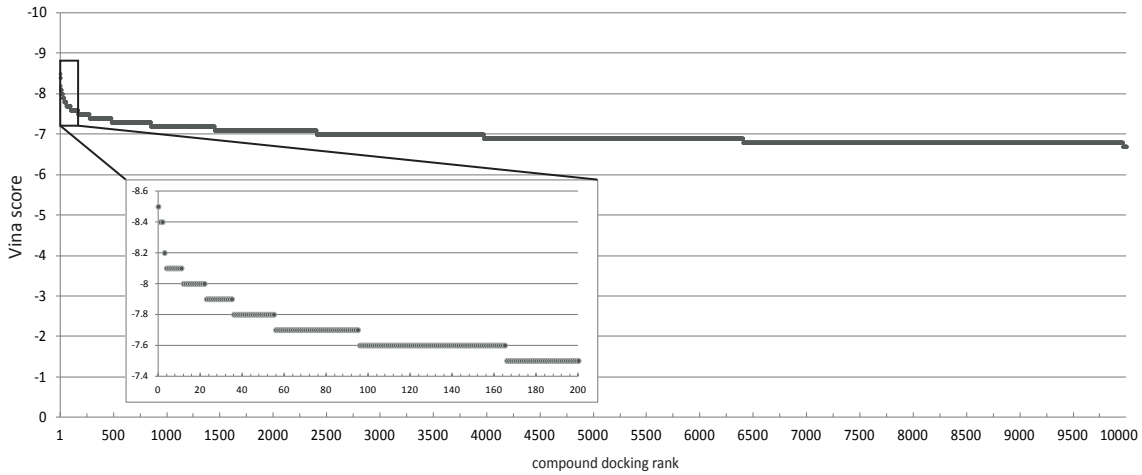


Fig. 4.6: *Vina Scores Reached for Top 10,000 Compounds.*

In this selection the 200 best scoring compounds have an exponential decline in score negativity. The top scoring compound reaches a Vina score of -8.5. Scores for the top 20 compounds are listed in Table 4.2. Also listed are molecular weights for these compounds and the $x\log P$ - the calculated partition coefficient. The $x\log P$ values are provided in the ZINC library instead of the experimentally obtained $\log P$, since many of the compounds used have not been studied before and no $\log P$ values exist. The calculations leading to the predicted $x\log P$ values are based on experimental data of the chemical building blocks of which the compound in question consists, however these values can only ever be an estimate. The Vina score is calculated based on atoms "interacting" with atoms in the target, so since this score also takes van der Waals forces into consideration, the score is biased towards compounds with more atoms i.e. higher molecular weight. This explains why the top 20 compounds do not show examples from across the range of molecular weights represented in the library.

Table 4.2: List of Top 20 Ranked Compounds.

Fragment Screening Results				
Rank	ZINC ID	Vina Score	MW[g/mol]	$x\log P$
1	ZINC39122184	-8.5	248.28	3.28
2	ZINC34435359	-8.4	248.28	3.23
3	ZINC34435365	-8.4	248.28	3.23
4	ZINC03847874	-8.2	247.3	3.41
5	ZINC19200219	-8.1	249.32	2.92
6	ZINC01728537	-8.1	249.27	2.51
7	ZINC01678901	-8.1	240.26	2.12
8	ZINC04951858	-8.1	246.22	1.13
9	ZINC03847855	-8.1	247.3	3.45
10	ZINC37237807	-8.1	248.32	1.33
11	ZINC34432656	-8.1	234.25	2.90
12	ZINC34435368	-8.1	248.28	3.23
13	ZINC01568410	-8.0	238.29	1.66
14	ZINC01600706	-8.0	236.27	3.42
15	ZINC05190800	-8.0	246.31	1.37
16	ZINC05332262	-8.0	246.27	3.40
17	ZINC06598438	-8.0	249.31	3.25
18	ZINC03643364	-8.0	235.22	0.87
19	ZINC03847852	-8.0	249.31	2.47
20	ZINC34484113	-8.0	248.28	3.23

The structural diversity of the top 100 compounds is unexpectedly low. To exemplify this the top 20 compounds are shown in Figure 4.7. The selection is dominated by compounds containing large hydrophobic ring systems. The docking algorithm recognised the hydrophobicity of the target pocket correctly. However, this obvious feature of the pockets was known and does not further inform our candidate search. Figure 4.8 shows the docked positions of the top 100 compounds in the SAS-6 structure. They are all positioned within the pocket, but are not spread from the centre enough to consider fusing multiple compounds into a combined candidate. Interestingly, none of these compounds reach into the pocket, like I154 does, but were docked with the ring systems aligned lid-like onto the hollow of the pocket (Figure 4.8 **D**).

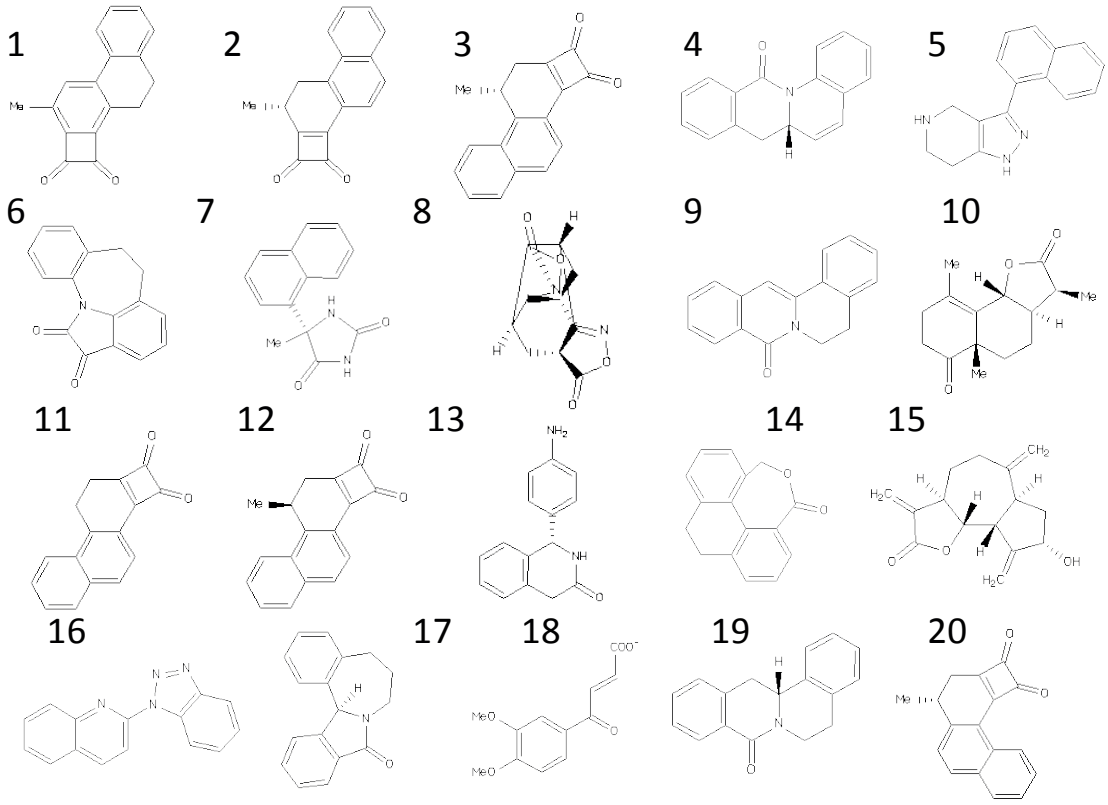


Fig. 4.7: Structural Formulae for Top 20 Ranked Compounds.

Details of the docking positions are exemplified here by the top most scoring compound and the compound ranked seventh. These were chosen as examples to represent structural diversity and two sets of commonly seen docking positions (Figure 4.8 **E** and **F**). For both examples the aromatic ring systems are directed towards the hydrophobic pocket. The dione decoration of compound ZINC39122184 points towards H97 on the α2 helix, however the hydantoin ring of compound ZINC01678901 is oriented towards the α1-α2 loop on the other side of the pocket. This provides an indication of the possibility of designing compounds to interact with the pocket rims.

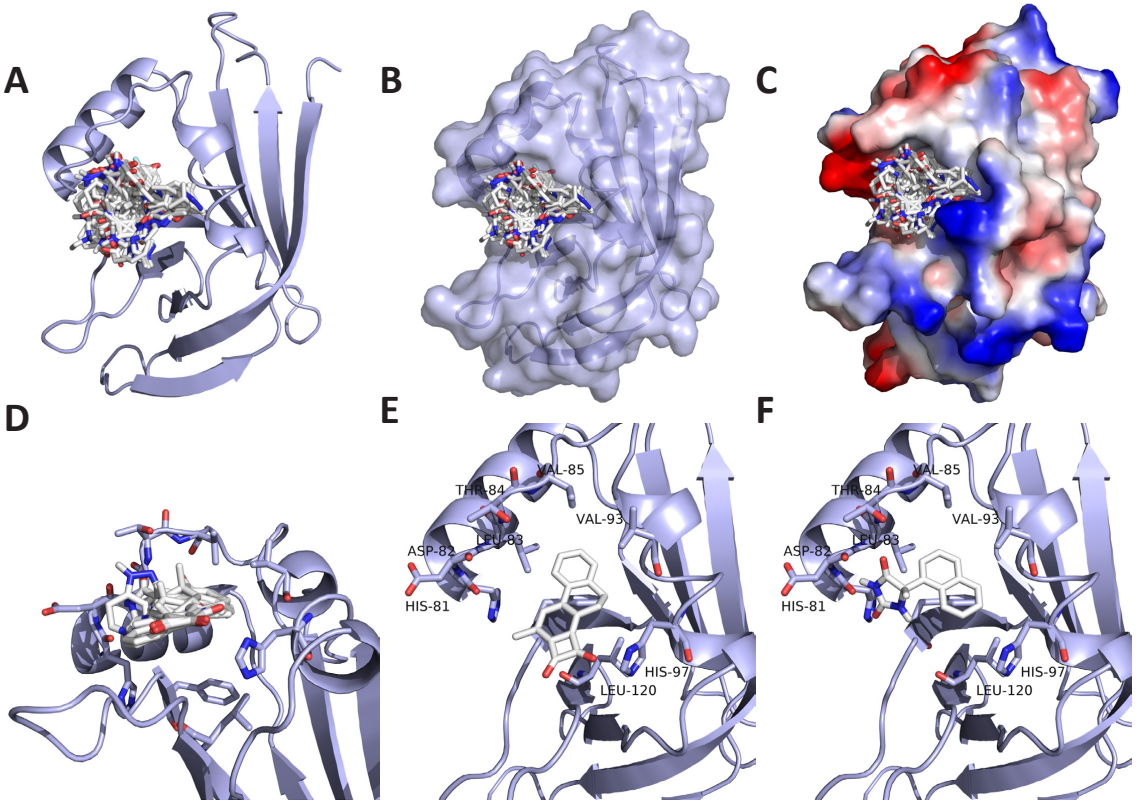


Fig. 4.8: Docking Positions of Compounds in SAS-6 Structure: (A) Overlay of docking positions of top 100 ranked compounds. (B) Overlay of docking positions of top 100 ranked compounds in a surface representation of the protein structure. (C) Overlay of docking positions of top 100 ranked compounds in a surface representation of the protein structure with added charge information. (D) Zoomed and tilted view of *A*. (E) Example of single compound docked in the hydrophobic pocket - rank 1 ZINC39122184. (F) Example of single compound docked in the hydrophobic pocket - rank 7 ZINC01678901.

To test the virtual screen prediction, compound ZINC39122184 (rank 1) and ZINC01678901 (rank 7) were tested for their interaction with *C. elegans* SAS-6 using NMR. The acquired spectral overlays can be seen in Figure 4.9. No chemical shift perturbations can be observed at a concentration of 2 mM.

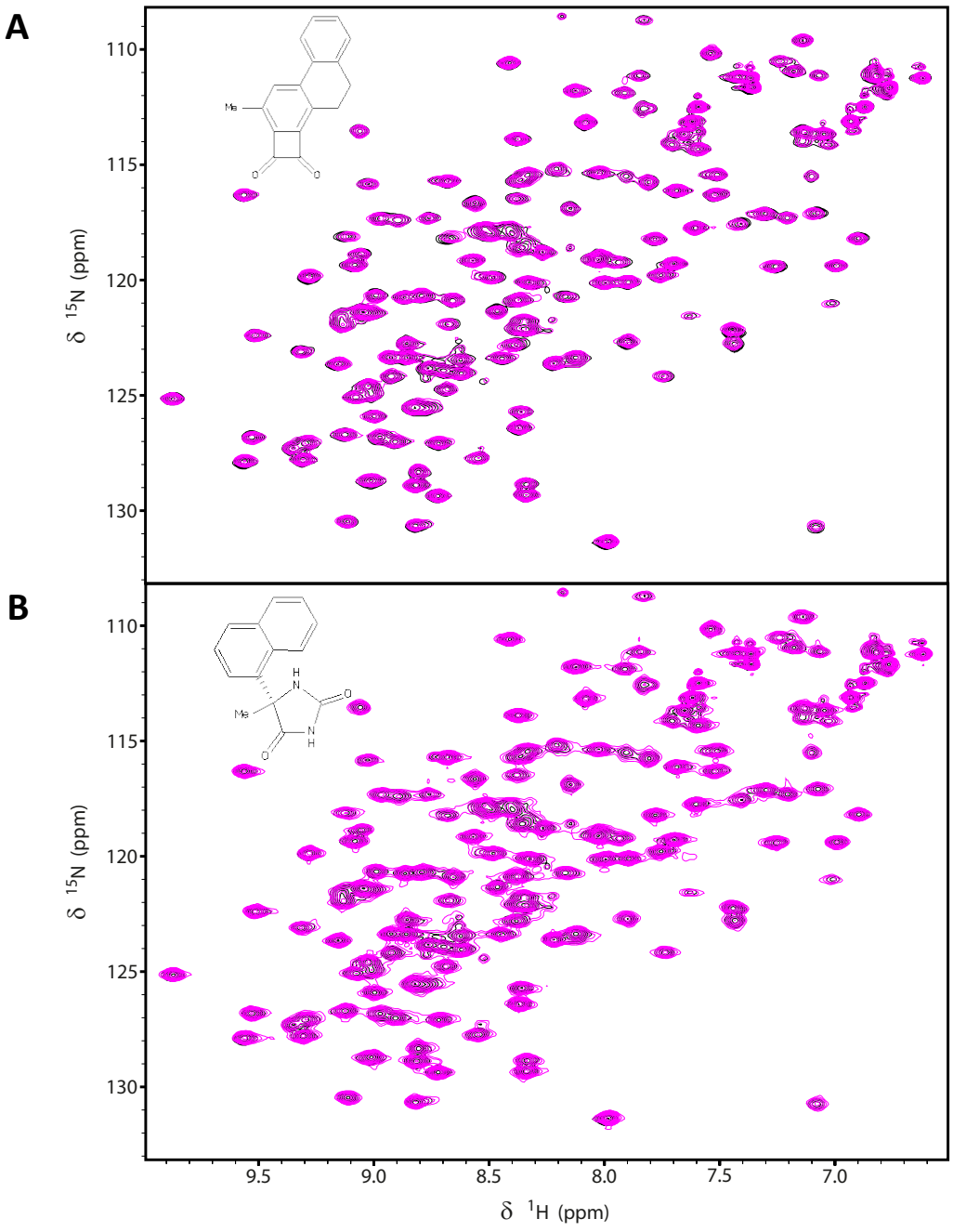


Fig. 4.9: HSQC NMR Spectral Overlay for Example Compounds: *Black* spectrum of ^{15}N SAS-6 1-168 Δ 103-130 I154E. (A) Overlay with spectrum collected with 2 mM ZINC39122184 - rank 1 in *pink*. (B) Overlay with spectrum collected with 2 mM ZINC01678901 - rank 7 in *pink*.

This result does not rule out the possibility that other compounds within the highest ranking selection have *in vitro* binding affinity for the target pocket. However, the selection of compounds in the top ranking results does not look promising for further compound development due to their aromatic bulk and incompatibility with chemical joining.

On the basis of the docking positions observed in the top ranking compounds, we conclude that both the α1-α2 loop as well as residue H97 are potential targets for a successful compound, but lack the tools for *de novo* compound design or synthesis. Instead of pursuing this technique, we turn to a different screening approach to find existing compounds that have favourable attributes.

4.2.2.2 Pharmacophore-like Based Virtual Screen

A pharmacophore approach requires a selection of preliminary candidate compounds with known binding capacity. In this case we lack an existing variety of small molecule ligands. Instead we base the pharmacophore model on loop β6-β7 of the SAS-6 N-terminal domain, which is the natural ligand for the target pocket. The position of the loop is highly similar in all solved SAS-6 structures. For this particular purpose the model for the *C. elegans* SAS-6 N-terminal structure 3PYI 1-168 S123E with 2.1 Å resolution was used.[§]

[§]The pharmacophore design and the associated virtual screen were performed by Dr. Minos Matsuokas (University of Barcelona). The process was informed by experience gained through the fragment based virtual screen described earlier and results discussed in the process.

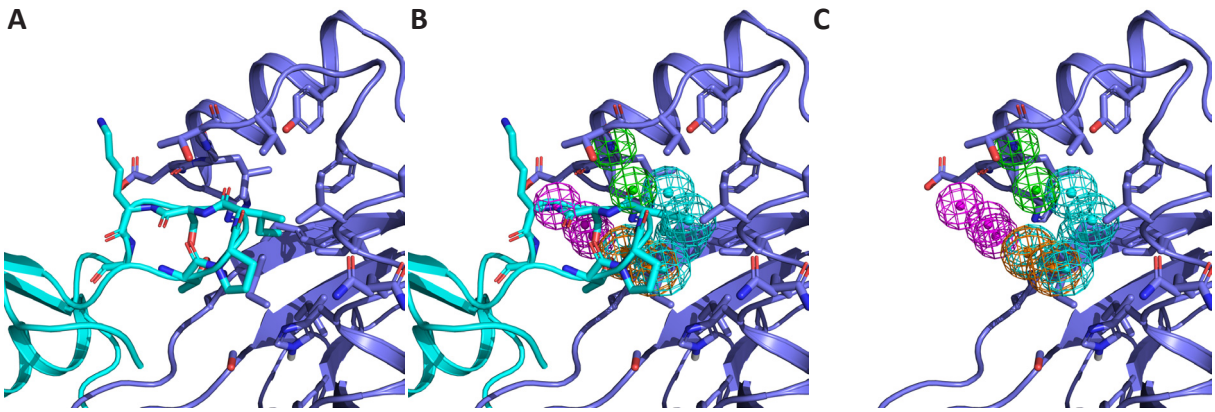


Fig. 4.10: Depiction of pharmacophore scaffold design: (A) Zoom of loop β6-β7 in hydrophobic pocket in SAS-6 N-terminus. (B) Pharmacophore scaffold - represented by spheres - as built onto loop β6-β7 where sphere colours represent characteristics assigned to scaffold positions (*pink/green* hydrogen bond donor/acceptor, *light blue* hydrophobic, *orange* aromatic). (C) Pharmacophore scaffold in hydrophobic pocket.

The pharmacophore model can be seen in Figure 4.10. Characteristics of the loop, such as hydrophobic interactions and H-bonds formed, that are involved in the binding of the pocket residues are built into a property scaffold to which the compound library is matched. The scaffold focuses on utilising backbone interactions between S155 and K156 and the α1-α2 loop, depicted in Figure 4.10 by pink and green spheres. Furthermore, it defines an area of hydrophobicity for the desired compounds (light blue spheres) and an element of aromatic bulk in the I154 position to introduce rigidity and potentially mimic the effect the I154W mutation (orange spheres).

The resulting property scaffold was then used to perform a virtual screen of all drug-like compounds contained in the ZINC database. This screen identified 2500 compounds matching the scaffold characteristics. These were then manually curated and 50 compounds selected based on chemical diversity. Finally, 37 of the compounds could be purchased and represent the final compound selection on which the *in vitro* experiments described in Chapter 5 were based (Table 4.3 and Figure 4.11).

Table 4.3: List of Compounds Selected by Pharmacophore-like Screen.

Pharmacophore Curated Compound Selection				
Library ID	ZINC ID	Score	MW[g/mol]	$x\log P$
A1	ZINC71942314	-6.5	321.42	2.63
A2	ZINC09970637	-6.0	321.4	-1.26
A3	ZINC72414547	-6.2	327.49	4.41
A4	ZINC70040105	-5.5	305.47	1.95
A5	ZINC58247642	-6.7	308.34	1.6
A6	ZINC11487801	-6.2	317.46	3.83
A7	ZINC55414762	-5.7	303.43	2.5
A8	ZINC54241093	-5.6	325.39	1.49
A9	ZINC09584839	-6.2	320.48	3.01
A10	ZINC13581301	-5.6	300.38	1.18
A11	ZINC36638814	-5.8	307.8	2.15
A12	ZINC09737303	-6.4	326.42	1.69
B1	ZINC06670228	-6.1	326.42	2.27
B2	ZINC25772421	-6.6	314.39	1.62
B3	ZINC13477106	-6.0	310.42	2.79
B4	ZINC12534617	-6.5	325.44	3.17
B5	ZINC15693151	-7.4	314.39	2.55
B6	ZINC67921979	-6.4	303.39	2.29
B7	ZINC81336504	-7.4	348.79	2.38
B8	ZINC65591069	-6.3	344.42	2.02
B9	ZINC81048793	-6.1	337.17	1.56
B10	ZINC69442707	-6.5	330.39	0.62
B11	ZINC69870503	-6.2	332.37	2.44
B12	ZINC76202639	-6.5	334.42	0.78
C1	ZINC28923015	-7.7	349.83	4.25
C2	ZINC22976783	-6.4	338.45	2.32
C3	ZINC00067653	-7.1	341.42	2.46
C4	ZINC79355337	-6.7	330.39	0.96
C5	ZINC04935336	-6.6	376.41	2.05
C6	ZINC76041126	-6.2	356.47	3.35
C7	ZINC63846942	-7.7	378.48	2.31
C8	ZINC64845906	-7.1	376.50	3.37
C9	ZINC33306742	-7.4	403.57	4.22
C10	ZINC02050114	-6.5	445.43	4.45
C11	ZINC00678552	-6.7	438.94	2.92
C12	ZINC13899084	-6.7	415.38	6.54
D1	ZINC12920785	-6.9	375.5	3.09

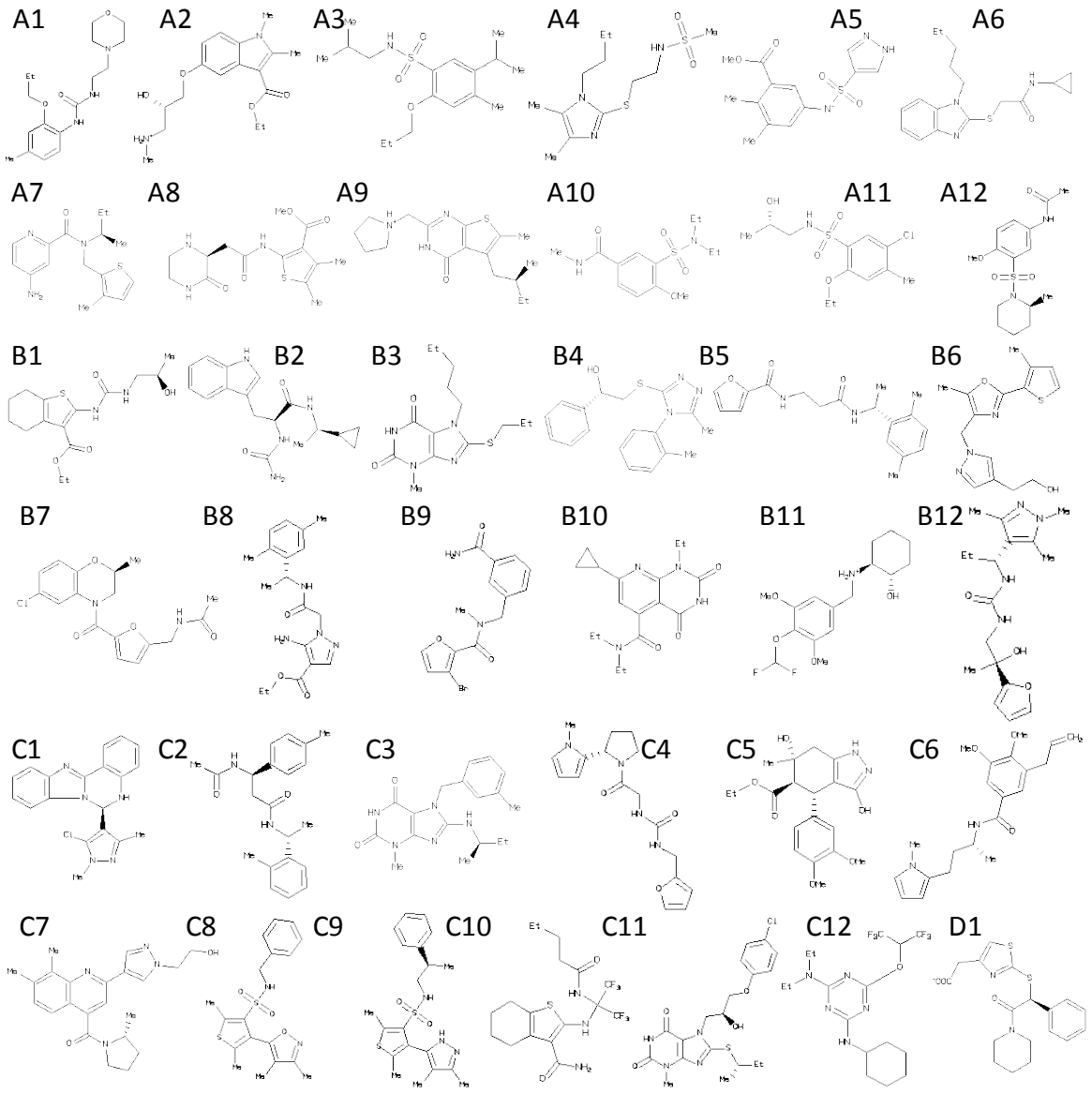


Fig. 4.11: Structural Formulae for Compounds Selected from Pharmacophore-like Screen.

The molecular weight distribution of the results is again not representative of the spread in the library used. The compound library was not chosen for fragment based design, so the size of the candidate compounds was not a consideration. As can be seen for the depiction of the 37 selected compound structures in Figure 4.11, the compounds share common features. This is due to the scaffold used to identify these compounds. Still, a variety of heterocyclic compounds with various decorations are

represented in the manually curated selection.

To compare the results of the two screening approaches, the 37 compounds selected from the pharmacophore screen were docked with the docking methodology used for the fragment based screen. The Vina scores of these 37 compounds are listed in Table 4.3. Even though the 37 compounds have a higher molecular weight than the compounds identified in the fragment based screen, overall they do not reach the same scores as the top ranking compounds in the fragment based screen. Indeed the scores here would mostly rank lower than the top 500 compounds identified in the other screen. By value they are closer to the scores found for the control compounds tryptophan and ZINC00392813. This shows that the molecular weight bias is not as severe as the first screen suggested, but also indicates that the two methods highlight entirely different sets of putative hits.

4.3 Conclusion

Neither high concentrations of free tryptophan, a cyclic peptide mimicking loop β6-β7 I154W nor a compound previously identified to bind *L. major* SAS-6 show binding to the hydrophobic target pocket in the *C. elegans* SAS-6 N-terminal domain. In our search for inhibitors for the N-dimer formation in *C. elegans* SAS-6 we turned to computational methods to search the chemical space for suitable starting points for compound design.

We set up a high throughput virtual screening pipeline for a fragment based search for compounds with characteristics favourable for binding to the hydrophobic pocket in the *C. elegans* SAS-6 N-terminal domain. This pipeline can be adapted to use

different libraries to diversify this method for not only fragment-based screening. The fragment based screen resulted in a large number of high scoring compounds. These top scoring compounds however are not suitable for compound generation through fragment combination due to their size and overlapping position in the target pocket. However, the compounds have the expected properties, for example a big hydrophobic core and orientation to residues in both of the rims of the hydrophobic pocket. Selected compounds from this screen did not show signs of binding to the protein in NMR experiments.

A pharmacophore-like screen with a scaffold based on properties of loop β6-β7 was performed, informed by experience gained from the fragment based virtual screen and a curated 37 compound diversity set selected. This compound selection contained small molecules with higher diversity in functional groups. The selection was purchased and used for *in vitro* screens and oligomerisation assays described in Chapter 5.

Chapter 5

In Vitro Search for SAS-6 Oligomersiation Inhibitory Compounds

"...it was the spring of hope, it was the winter of despair,..."

Charles Dickens

A Tale of Two Cities

This chapter describes the *in vitro* screening of a small molecule selection assembled for the purpose of finding an inhibitor for the N-terminal dimerisation of SAS-6. The compound selection was tested for binding to the hydrophobic pocket of the SAS-6 N-terminal domain, which mediates the dimerisation.

The *in vitro* screen was conducted with the use of ^{15}N HSQC NMR experiments on SAS-6 homologues from three model organisms and hits were tested on their ability to inhibit SAS-6 oligomerisation using AUC.

External contributions to work in this chapter: AUC experiments were conducted with help from Dr. David Staunton at the Biophysics facility of the Biochemistry Department, University of Oxford and Prof. David Scott and Dr. Jemma Harris at Rutherford Appleton Laboratory, Harwell Science Campus.

5.1 NMR Based Screen for Small Molecules Binding SAS-6 N-terminal Domain

A compound selection containing 37 small molecules was compiled with the help of virtual screens investigating possible small molecule ligands for the hydrophobic pocket essential for N-terminal dimerisation in *C. elegans*; as described in Chapter 4.

The compounds were selected on the basis of a chemical-functionality scaffold defined using the interactions of loop β6-β7, the natural ligand of the pocket, forms with the pocket residues. To test if the selected compounds bind to the target pocket as predicted, we conducted an ^{15}N HSQC NMR-based *in vitro* screen. This method was chosen because it allows for the identification of binding compounds, and also can identify the residues affected by the binding event if resonances are assigned [175].

For the screen, first an HSQC spectrum was collected of a monomeric construct of the SAS-6 N-terminal domain. After addition of a small molecule compound, another spectrum is collected, which can then be compared to the original protein spectrum. Differences between the two spectra can be interpreted as changes in the chemical environment of particular residues indicative of binding or structural rearrangement. To capture compound binding to the hydrophobic pocket of SAS-6, the NMR screen was performed with monomeric SAS-6$_N$ constructs. In the monomeric state, the hydrophobic dimerisation pocket is exposed. The compound can access the pocket without having to outcompete the dimerisation. Even though we aim to identify compounds that can ultimately inhibit SAS-6 oligomerisation, at this stage we want to study only the compounds' ability to bind to the target pocket.

5.1.1 Compound Solubility

Compounds in the screening collection were selected for their compliance with Lipinski rules [228], which includes a consideration of solubility in the form of a limit on the log$P$. Ideally the log$P$ should be between -0.4 and 5.6 to ensure reasonable solubility of the small molecule in both hydrophilic and hydrophobic conditions [195]. However, log$P$ values are experimentally determined and are not available for most compounds. Instead $x$log$P$ values are used, which are calculated on the basis of experimentally determined values for the compounds' building blocks. Since $x$log$P$ predictions are error prone, some compounds in the library might still show unfavourable solubility.

Compounds were dissolved in DMSO to a 60 mM stock and then added to the aqueous protein solution to a final concentration of 5 % DMSO and 2 mM compound. Upon exposure to the aqueous buffer, some compounds showed slight precipitation. To test if the compound remained soluble despite the visible precipitation, 1D ^{1}HNMR spectra were collected for each sample. Of the 37 compounds tested, eight compounds were insoluble in aqueous buffer and five additional compounds showed decreased solubility in the aqueous solution, but were detectable as additional peaks appearing in the ^{1}H-NMR spectra (Figure 5.1 and Table 5.1, where $x$log$P$ values are listed in comparison to the solubility observed in the aqueous buffer used in NMR experiments described in this thesis).

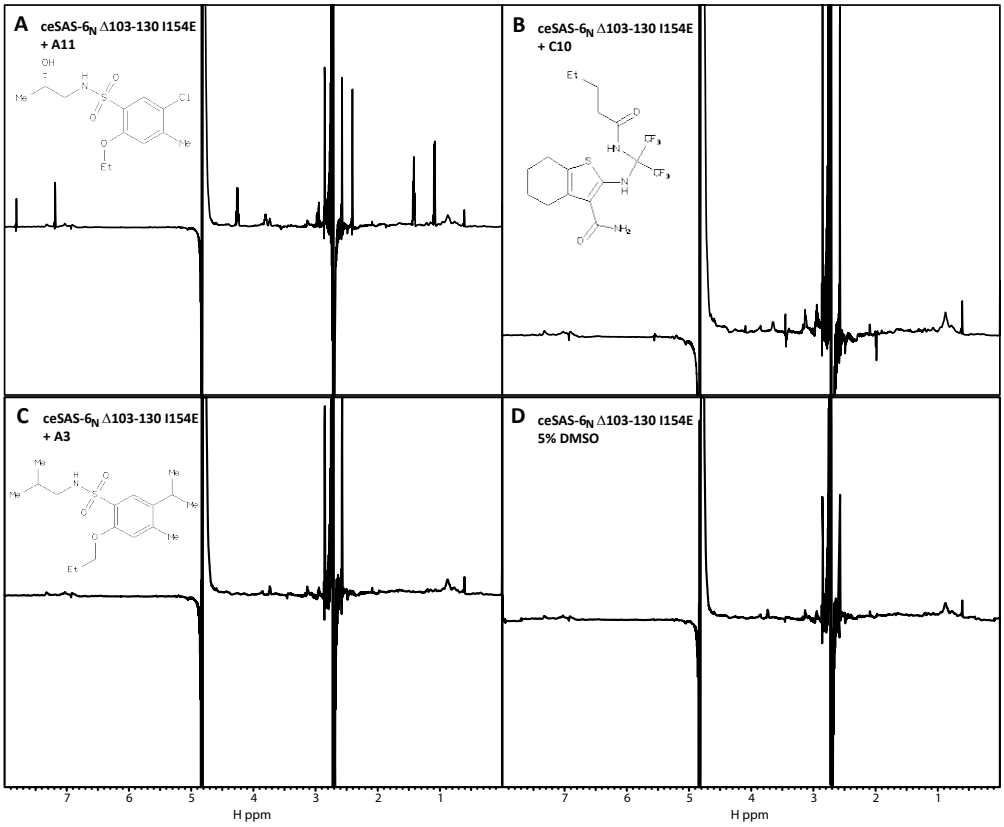


Fig. 5.1: 1D ^{1}H NMR Spectra - *C. elegans* SAS-6 N-terminal Domain With A11, C10 and A3: Spectra of ^{1}H SAS-6$_N$ Δ 103-130 I154E with: (A) 2 mM A11 - as anexample of a soluble compound, (B) with 2 mM C10 - as an example of a partially soluble compound, (C) spectrum with 2 mM A3 - as an example of an insoluble compound and (D) 5 % of DMSO added - as control sample.

Table 5.1: Solubility of Screening Compounds: The $x\log P$ values listed here are calculated $\log P$ values sourced from the ZINC library.

Solubility of Screening Compounds			
Selection ID	ZINC ID	$x\log P$	Solubility in Aqueous NMR Buffer Determined by 1D NMR
A1	ZINC71942314	2.63	
A2	ZINC09970637	-1.26	
A3	ZINC72414547	4.41	insoluble
A4	ZINC70040105	1.95	
A5	ZINC58247642	1.6	
A6	ZINC11487801	3.83	
A7	ZINC55414762	2.5	
A8	ZINC54241093	1.49	
A9	ZINC09584839	3.01	insoluble
A10	ZINC13581301	1.18	
A11	ZINC36638814	2.15	
A12	ZINC09737303	1.69	
B1	ZINC06670228	2.27	
B2	ZINC25772421	1.62	
B3	ZINC13477106	2.79	insoluble
B4	ZINC12534617	3.17	
B5	ZINC15693151	2.55	
B6	ZINC67921979	2.29	
B7	ZINC81336504	2.38	
B8	ZINC65591069	2.02	
B9	ZINC81048793	1.56	
B10	ZINC69442707	0.62	
B11	ZINC69870503	2.44	
B12	ZINC76202639	0.78	
C1	ZINC28923015	4.25	insoluble
C2	ZINC22976783	2.32	partially soluble
C3	ZINC00067653	2.46	insoluble
C4	ZINC79355337	0.96	
C5	ZINC04935336	2.05	
C6	ZINC76041126	3.35	partially soluble
C7	ZINC63846942	2.31	
C8	ZINC64845906	3.37	partially soluble
C9	ZINC33306742	4.22	insoluble
C10	ZINC02050114	4.45	partially soluble
C11	ZINC00678552	2.92	insoluble
C12	ZINC13899084	6.54	insoluble
D1	ZINC12920785	3.09	partially soluble

5.1.2 Screen for Compounds Binding *C. elegans* SAS-6 N-terminal Domain

For the purpose of finding a small molecule inhibitor for the SAS-6 N-terminal domain dimerisation we wanted to identify compounds that bind to the hydrophobic pocket, which facilitates the dimerisation. A monomeric construct of the *C. elegans* N-terminal domain was used (ceSAS-6_N (1-168) Δ103-130 I154E) [103]. For *C. elegans* SAS-6, a I154E mutation has been shown to abolish dimerisation *in vitro* [103]. The additional deletion of loop α2-β5 (residues 103-130) further decreases the affinity of dimerisation (see Chapter 3). These alterations leave the hydrophobic pocket unaltered and solvent exposed.

A comparison of changes in the NMR spectra allows a qualitative differentiation of compounds with affinity for the ceSAS-6_N construct (Figure 5.2). Here, the effects of compounds A11, A12 and B1 are discussed to exemplify the extent of changes observed across the screen. Addition of compounds A11 and B1 to ceSAS-6_N (Figure 5.2 A and C) causes notable chemical shift perturbations, indicating potential binding. In contrast A12 does not have a visible effect on ceSAS-6_N.

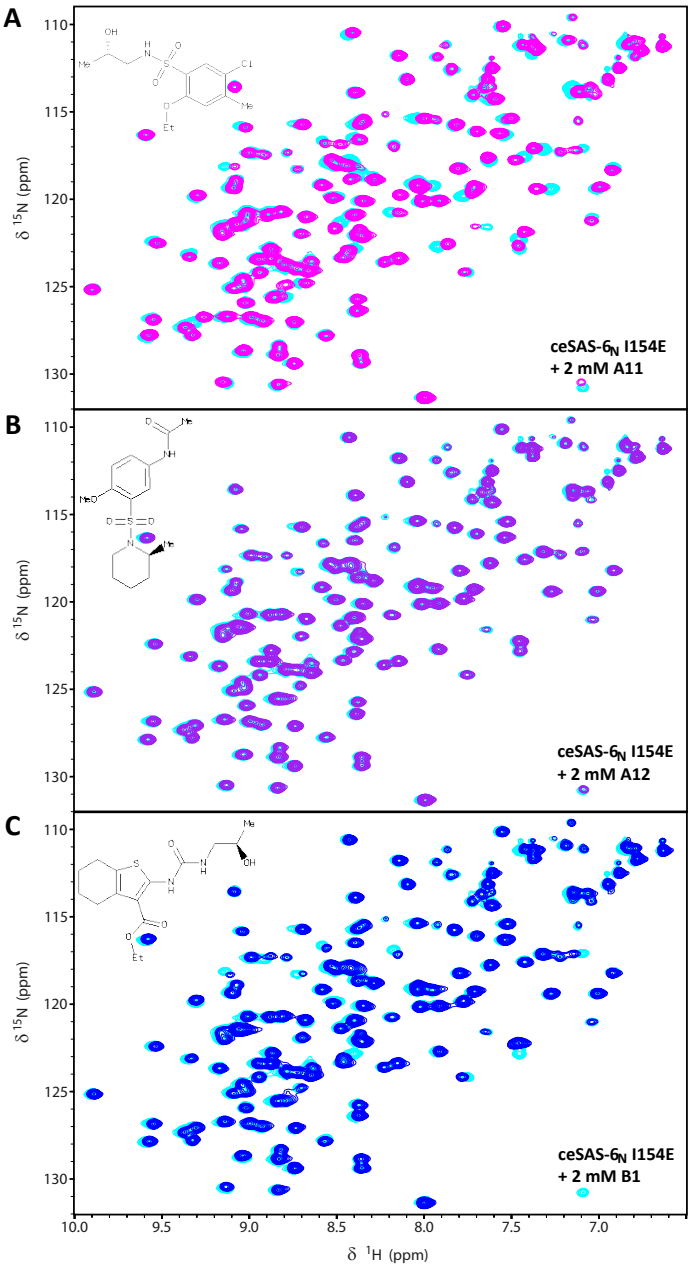


Fig. 5.2: Overlay of ^{15}N-HSQC NMR Spectra - *C. elegans* SAS-6 N-terminal Domain With A11, A12 and B1: Spectra of ^{15}N SAS-6$_N$ Δ 103-130 I154E *turquoise* overlayed with: (A) *pink* spectrum with 2 mM A11, (B) *purple* spectrum with 2 mM A12 and (C) *blue* spectrum with 2 mM B1.

This qualitative observation can be further analysed by measuring the chemical shift perturbation as a weighted difference in resonance position. Assignments of the

resonances for this construct necessary for this analysis were available*. The resulting perturbation values (Figure 5.3 A) are used to identify residues with significant perturbations. The position of the affected residues in the ceSAS-6_N structure is indicated in Figure 5.3 and 5.4 B, C and D for compounds A11, A12 and B1 respectively.

*NMR assignments done by Dr. Michele Erat

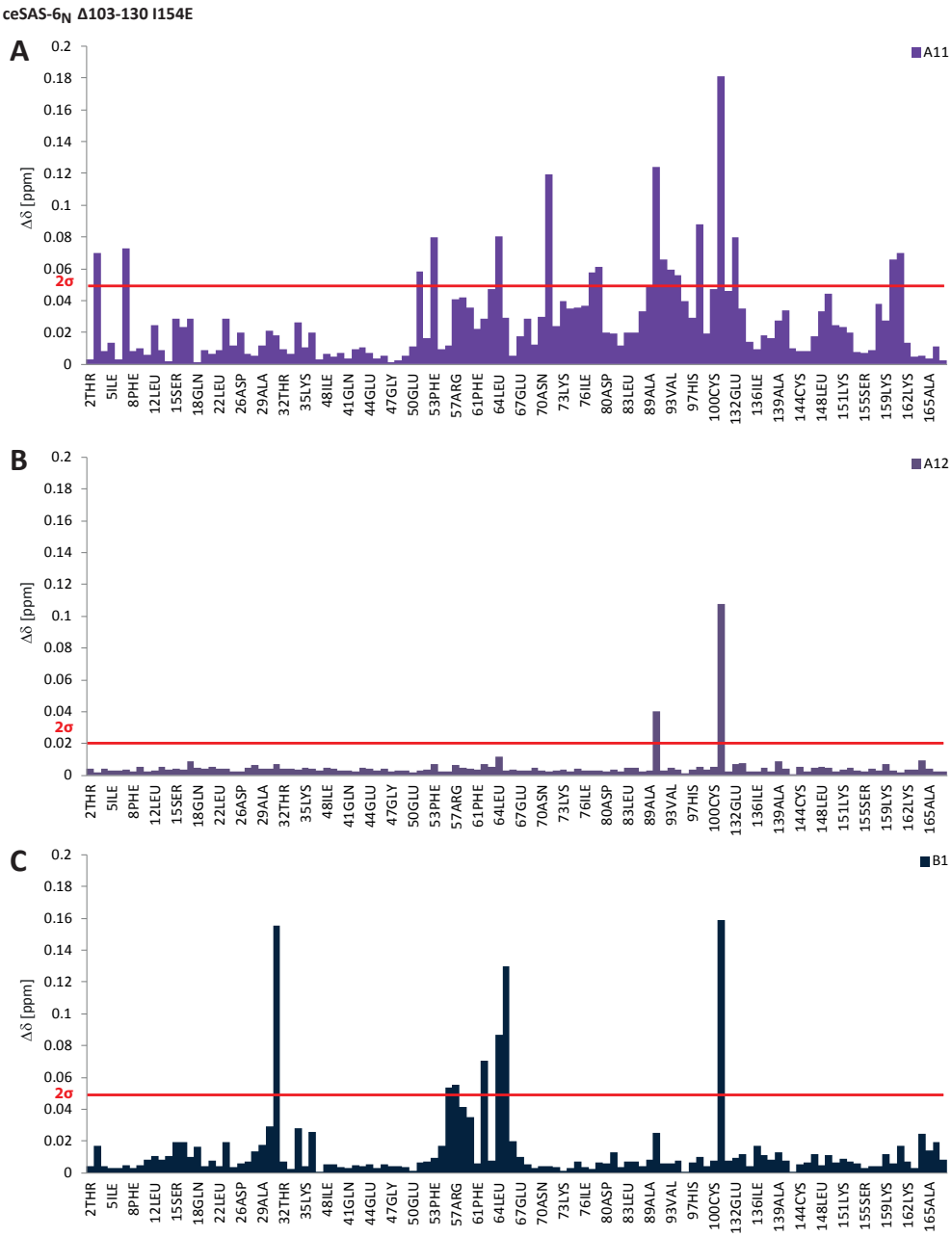


Fig. 5.3: Chemical Shift Perturbation from NMR Screen - *C. elegans* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Table of weighted chemical shift perturbations for assigned residues for *C. elegans* SAS-6_N Δ 103-130 I154E spectra with: *purple* 2 mM A11; $2\sigma = 0.056$ ppm. (B) *dark purple* 2 mM A12; $2\sigma = 0.02$. (C) *dark blue* 2 mM B1; $2\sigma = 0.05$.

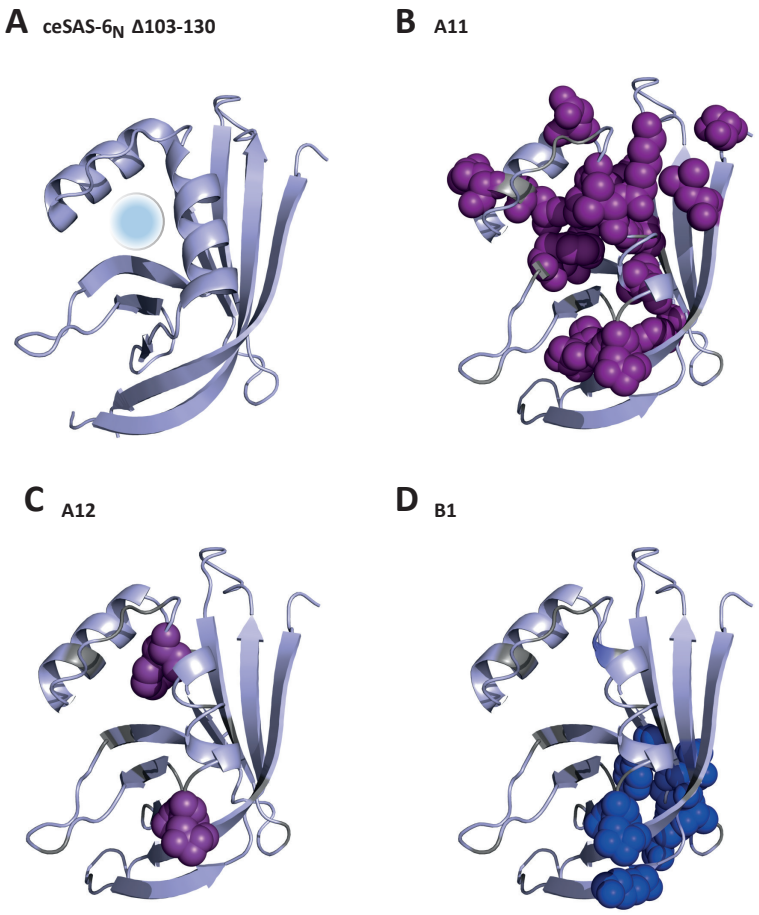


Fig. 5.4: Chemical Shift Perturbation from NMR Screen - *C. elegans* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Structure of SAS-6_N Δ 103-130 I154 (PDB code 4G79, 1.8 Å resolution) [103]; the *blue hazed* sphere marks the hydrophobic target pocket. (B) Structure depicted in (A) with residues experiencing significant chemical shift perturbation upon addition of 2 mM A11 shown as *purple* spheres. (C) Structure shown in (A); *purple* spheres highlight residues with significant chemical shift perturbation upon addition of 2 mM A12. (D) Structure depicted in (A); *blue* spheres show residues experiencing significant chemical shift perturbation from addition of 2 mM B1.

The addition of A11 has a strong effect on residues located in the target pocket. A12 only showed chemical shift perturbation for two residues. A minimal effect like this is not considered enough evidence for binding. Various factors, such as buffer composition and temperature fluctuations could have caused these unspecific resonance changes. Even though B1 addition shows significant chemical shift perturbations for

several residues, the affected region is not in the target pocket. Amongst the 29 soluble compounds there were several non-binders such as A12 and several examples of compounds showing significant chemical shift perturbation, which did not affect residues in the target pocket similar to the B1 effect. Tables of chemical shift perturbations for every compound screened can be found in Appendix C.

A11 was the only compound emerging from this screen with a chemical shift perturbation pattern indicating binding to the target pocket. A11 is the first compound shown to bind to the hydrophobic pocket of the *C. elegans* SAS-6 N-terminal domain.

To estimate the binding affinity of A11 to the SAS-6 N-terminal domain, NMR spectra of samples with different A11 concentrations were collected (Figure 5.5). The titration effect is visualised on chemical shift perturbations on residues G157, V93 and I95. Residues V93 and I95 are located directly in the target pocket, whereas residue G157 is located at the bottom rim of the pocket. The titration did not reach saturation, but indicates that the binding affinity must be in the millimolar range. This would be considered weak binding for a small molecule and would need to be improved for use in biological systems.

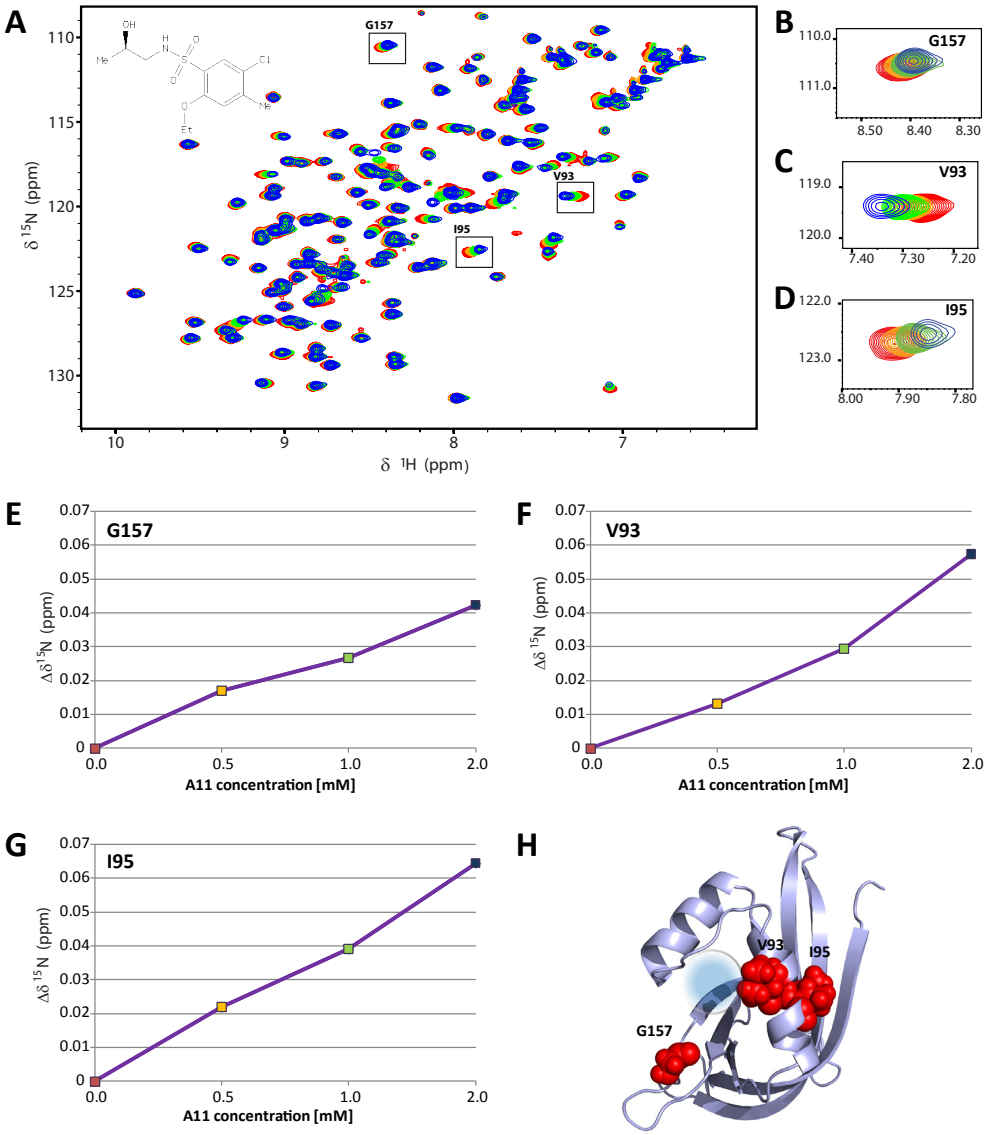


Fig. 5.5: ^{15}N-HSQC NMR Titration - *C. elegans* SAS-6 N-terminal Domain with A11: (A) Overlay of spectra of ^{15}N SAS-6$_N$ Δ 103-130 I154E with: no added A11(*red*), 0.5 mM A11 (*yellow*), 1 mM A11 (*green*) and 2 mM A11 (*blue*). (B) Zoom in of (A) on resonances assigned to pocket residue G157. (C) Zoom in of (A) on resonances assigned to pocket residue V93. (D) Zoom in of (A) on resonances assigned to pocket residue I95. (E) Chemical shift perturbation of residue G157 plotted against concentrations of A11 (0 mM, 0.5 mM, 1 mM and 2 mM). (F) Chemical shift perturbation of residue V93 plotted against concentrations of A11 (0 mM, 0.5 mM, 1 mM and 2 mM). (G) Chemical shift perturbation of residue I95 plotted against concentrations of A11 (0 mM, 0.5 mM, 1 mM and 2 mM). (H) Structural representation of ceSAS-6$_N$ Δ103-130 (PDB code 4G79, 1.8 Å resolution) [103]; residues I157, V93 and I95 depicted as *red* spheres; the *blue hazed* sphere marks the binding pocket.

5.1.3 Screen for Compounds Binding *C. reinhardtii* SAS-6 N-terminal Domain

Although there are substantial sequence differences between SAS-6 variants amongst species, the similarity of structures is very high. Essential features like the β6-β7 loop and the hydrophobic target pocket are conserved in their function. We were therefore intrigued to test the compound selection derived from *in silico* screening of the *C. elegans* SAS-6 variant on other homologues of the protein.

C. reinhardtii is a model organism of great interest to centriolar research. This single celled organism has two flagella used for signalling and movement. Defects in basal bodies can be studied in this system for their impact on various levels of flagella development and function [12, 39]. This makes it a relevant target for the design of SAS-6 oligomerisation inhibitors.

The same compound library was tested in an ^{15}N HSQC NMR screen with a monomeric protein construct of *C. reinhardtii* SAS-6 N-terminal domain (crSAS-6$_N$ (1-159) F145E) and analysed in the manner described for the *C. elegans* screen. Assignments for this construct were available[†]. Results for A11, A12 and B1 are discussed as examples of the overall observed screening results (Figures 5.6, 5.7 and 5.8). Chemical perturbation tables for all compounds screened can be found in Appendix C. Similarly to the ceSAS-6 screen, we observe significant chemical shift perturbations for A11 and B1, but none for A12. Residues affected by the presence of A11 are localised to hydrophobic pocket residues. B1 shows a smaller effect on *C. reinhardtii* SAS-6. The residues affected do not connect to an off-target binding region, but appear unspecific. Across all compounds screened, A11 was the only compound identified to bind in a specific manner to the hydrophobic pocket of crSAS-6.

[†] Assigned by Dr. Anna Fusery

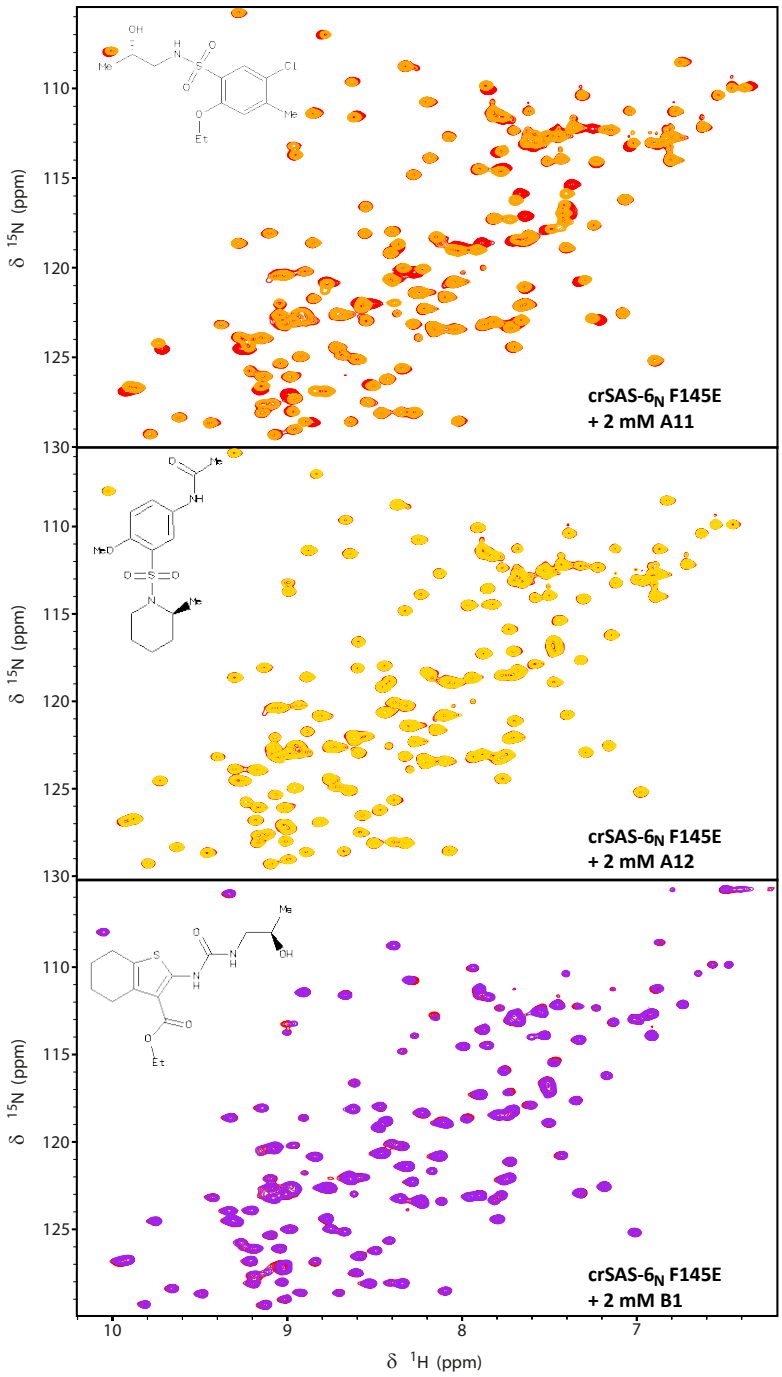


Fig. 5.6: Overlay of ^{15}N-HSQC NMR Spectra - *C. reinhardtii* SAS-6 N-terminal Domain With A11, A12 and B1: Spectra of ^{15}N SAS-6$_N$ F145E *red* overlayed with: (A) *orange* spectrum with 2 mM A11, (B) *yellow* spectrum with 2 mM A12 and (C) *purple* spectrum with 2 mM B1.

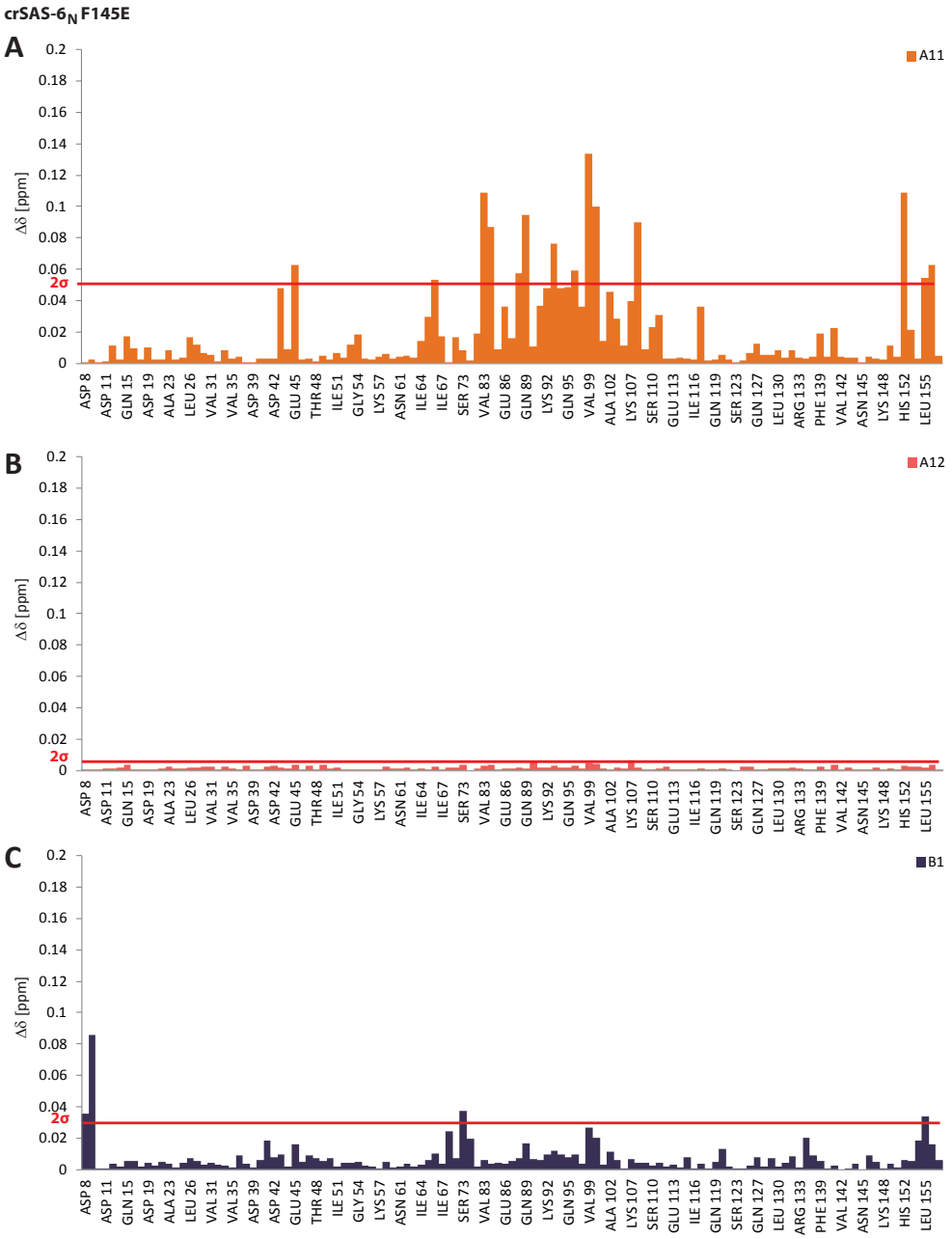


Fig. 5.7: Chemical Shift Perturbation from NMR Screen - *C. reinhardtii* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Table of weighted chemical shift perturbations for assigned residues for crSAS-6$_N$ F145E spectra with *orange* 2 mM A11; $2\sigma = 0.05$ ppm. (B) *salmon* 2 mM A12; $2\sigma = 0.002$ ppm and (C) *dark blue* 2 mM B1; $2\sigma = 0.03$ ppm.

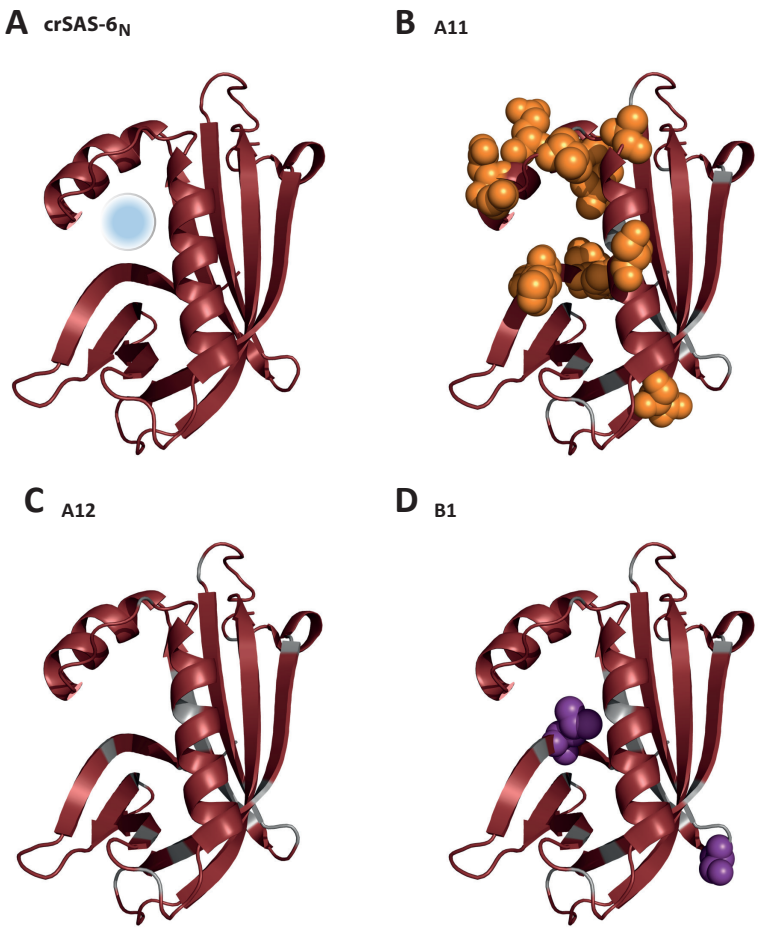


Fig. 5.8: Chemical Shift Perturbation from NMR Screen - *C. reinhardtii* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Structure of *C. reinhardtii* SAS-6 1-159 (PDB code 3Q0Y, 2.1 Å resolution) [93]. (B) Structure shown in (A) with residues experiencing significant chemical shift perturbation upon addition of 2 mM A11 shown as *orange* spheres. Residues with missing assignments are coloured *grey*. The *blue hazed* sphere marks the target pocket. (C) Structure shown in (A), no significant chemical shift perturbations upon addition of 2 mM A12 were observed. (D) *Purple* spheres show residues experiencing significant chemical shift perturbations from addition of 2 mM B1.

5.1.4 Screen for Compounds Binding *H. sapiens* SAS-6 N-terminal Domain

No matter how similar the model organisms we study, one aim of research is to understand biology in the human context. The study of centriole duplication in human

cells is particularly relevant for their involvement in the cell cycle and therefore their connection to cancer development and diseases like ciliopathies and microcephalies [70].

Constructs for the study of the *H. sapiens* SAS-6 variant had not previously been studied and were created for this work. For the purpose of the NMR screen, the N-terminal domain construct was designed with a F131E mutation (hsSAS-6 1-152 F131E), which - based on our experience with SAS-6 constructs of other species - should create a greatly reduced dimerisation affinity. For the purpose of this study, preliminary assignments for the majority of this constructs' residues were derived. Data collected for assignment purposes will be used to complete assignment after further analysis.

Screening results are exemplified using A11, A12 and B1 (Figure 5.9, 5.10 and 5.11). Interestingly, we observe significant chemical shift perturbation for residues localising to the hydrophobic pocket upon addition of A11 and A12. Residues affected by the addition of B1 were less specifically localised. The NMR assignments for the construct used are incomplete. However, the analysis of the preliminarily assigned residues in conjunction with the observable differences in the spectra collected let us conclude that A11 and A12 are potential candidate compounds. Both compounds have a visible effect on the chemical shift perturbations and show signs of binding to the hydrophobic pocket of the *H. sapiens* SAS-6 N-terminal domain.

A12 is the first compound to be identified to bind selectively to hsSAS-6 - compared to *C. elegans* and *C. reinhardtii* SAS-6. A11 on the other hand is of interest to study as a potential "broadband" compound, binding to all three tested SAS-6 variants.

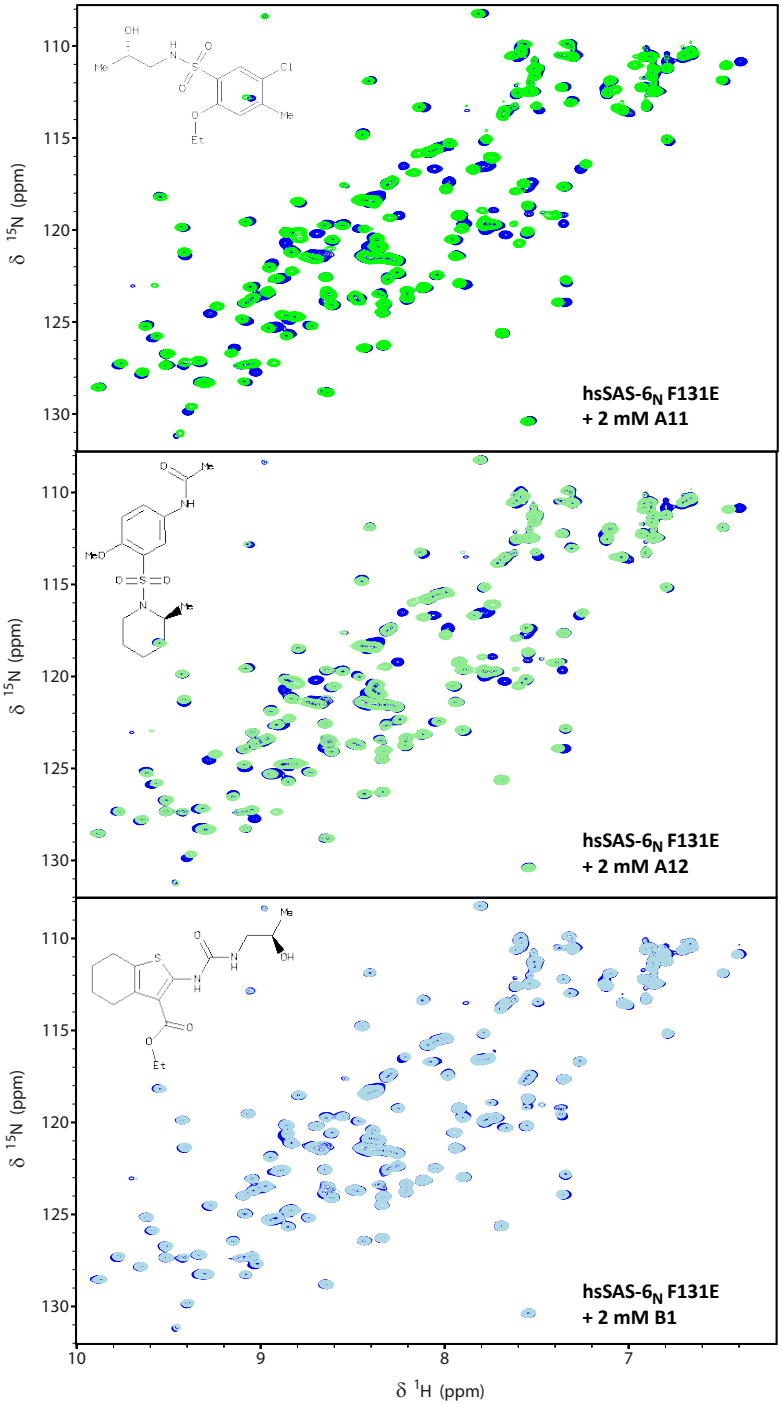


Fig. 5.9: Overlay of ^{15}N-HSQC NMR Spectra - *H. sapiens* SAS-6 N-terminal Domain With A11, A12 and B1: Spectra of ^{15}N SAS-6$_N$ F131E *blue* overlayed with: (A) *green* spectrum with 2 mM A11, (B) *light green* spectrum with 2 mM A12 and (C) *light blue* spectrum with 2 mM B1.

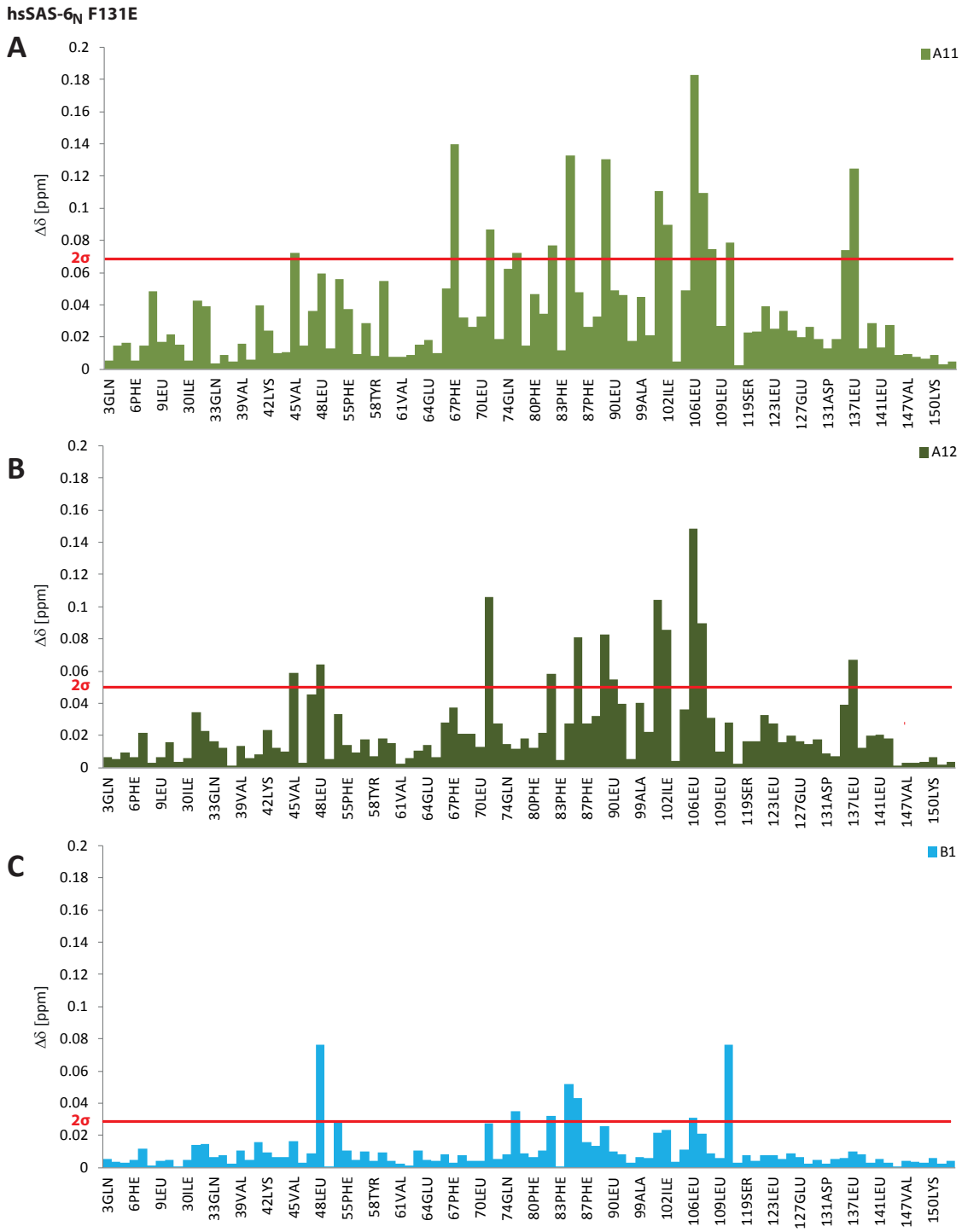


Fig. 5.10: Chemical Shift Perturbation from NMR Screen - *H. sapiens* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Table of weighted chemical shift perturbations for assigned residues for *H. sapiens* hsSAS-6_N F131E spectra with: *light green* 2 mM A11; $2\sigma = 0.07$ ppm. (B) *dark green* 2 mM A12; $2\sigma = 0.05$ ppm and (C) *light blue* 2 mM B1; $2\sigma = 0.03$ ppm.

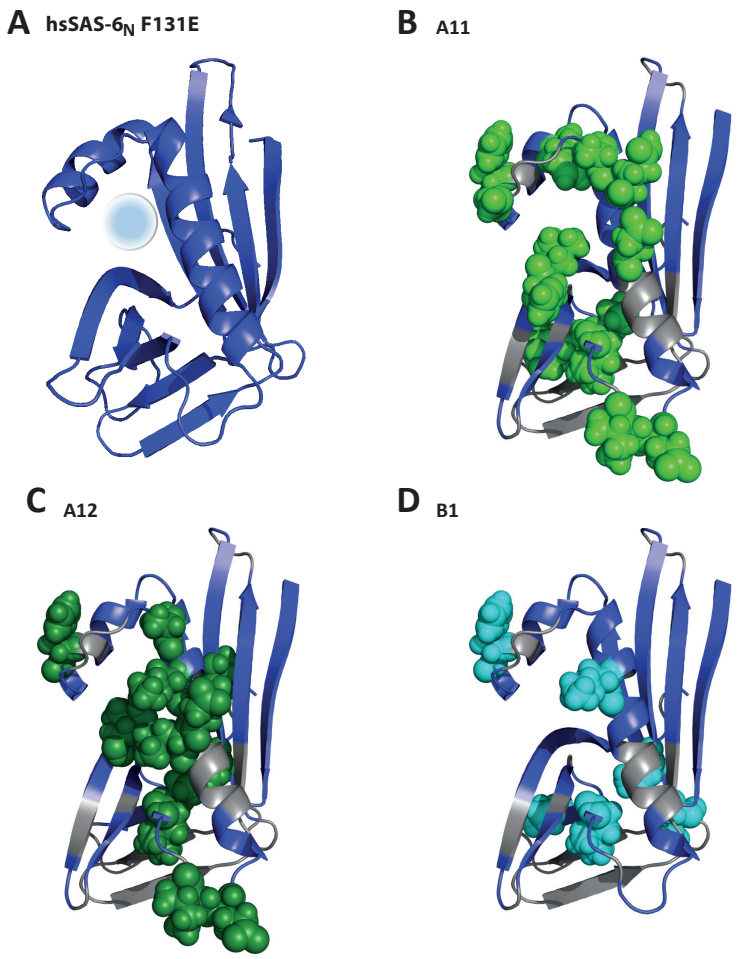


Fig. 5.11: Chemical Shift Perturbation from NMR Screen - *H. sapiens* SAS-6 N-terminal Domain with A11, A12 and B1: (A) Structure of hsSAS-6_N F131E (see Chapter 6 for a description of this structure, 1.49 Å resolution). The *blue hazed* sphere marks the target pocket. (B) Structure depicted in (A) with residues experiencing significant chemical shift perturbations upon addition of 2 mM A11 shown as *green* spheres. Residues with missing assignments are coloured *grey*. (C) Structure shown in (A), *dark green* spheres highlight residues with significant chemical shift perturbations upon addition of 2 mM A12. (D) *Light blue* spheres show residues experiencing significant chemical shift perturbation from addition of 2 mM B1.

NMR was used for titration experiments of A11 and A12 against the *H. sapiens* SAS-6 N-terminal domain (Figure 5.12 and 5.13). Residues F87, L70, H136 and I46 located in the hydrophobic binding pocket are used to follow the chemical shift perturbation over the course of the titration. In both cases the titration did not reach

saturation. For A11, this indicates that the binding affinity is again weak and in the millimolar range. The same conclusion was reached for experiments with A12 binding. In addition, this experiment revealed a solubility limit for this compound. The sample with the highest concentration of A12 added showed precipitation, which is also reflected in the titration series as a reduction of chemical shift perturbation for the last data point.

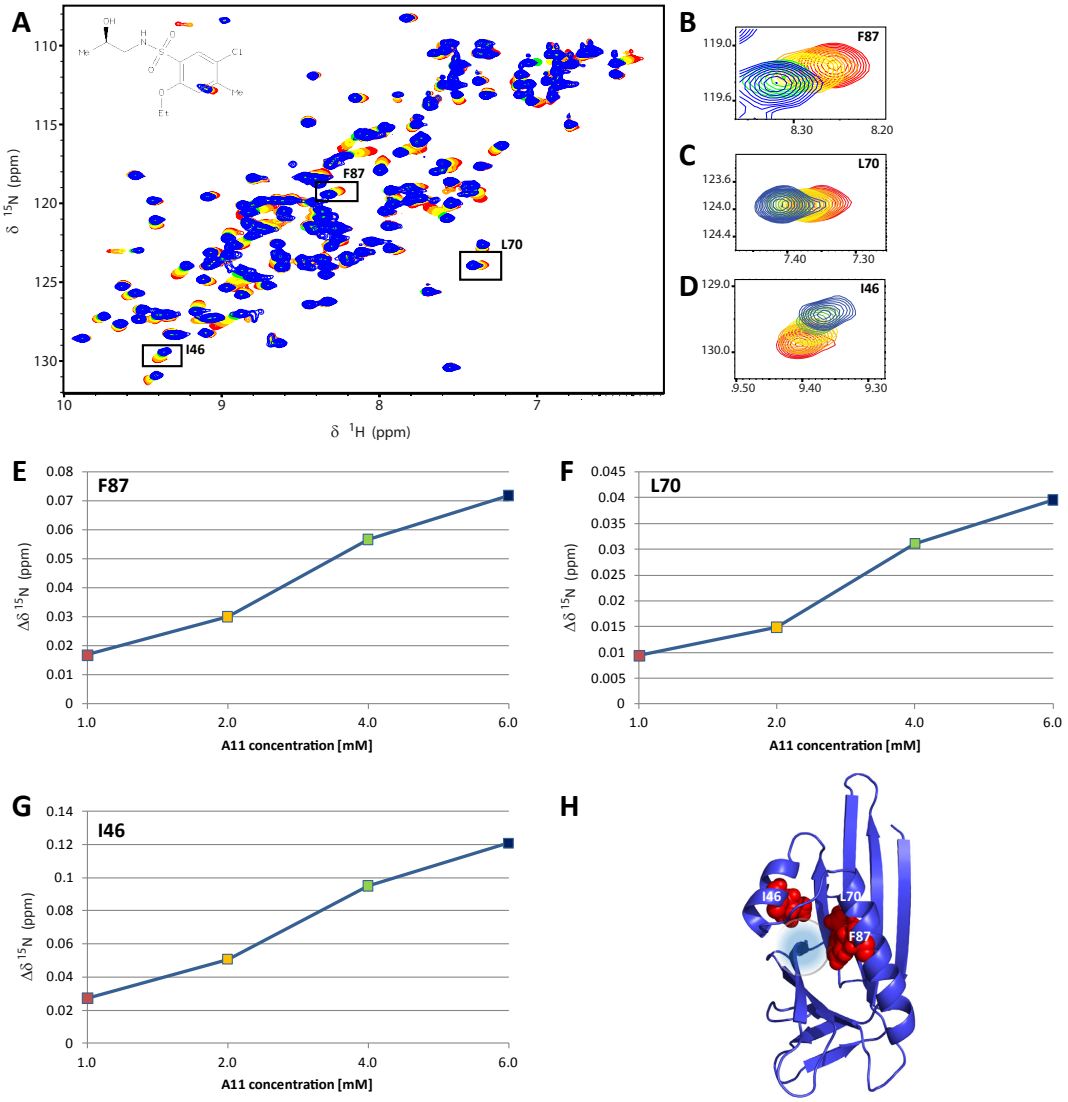


Fig. 5.12: ^{15}N-HSQC NMR Titration - *H. sapiens* SAS-6 N-terminal Domain with A11: (A) Overlay of spectra of ^{15}N hsSAS-6$_N$ F131E with no added A11 (*red*), 1 mM added A11(*yellow*), 2 mM A11 (*orange*), 4 mM A11 (*green*) and 6 mM A11 (*blue*). (B) Zoom in of (A) on resonances assigned to pocket residue F87. (C) Zoom in of (A) on resonances assigned to pocket residue L70. (D) Zoom in of (A) on resonances assigned to pocket residue I46. (E) Chemical shift perturbation of residue F87 plotted against concentrations of A11 (1 mM , 2 mM, 4 mM and 6 mM). (F) Chemical shift perturbation of residue L70 plotted against concentrations of A11 (1 mM , 2 mM, 4 mM and 6 mM). (G) Chemical shift perturbation of residue I46 plotted against concentrations of A11 (1 mM , 2 mM, 4 mM and 6 mM). (H) Structural representation of hsSAS-6$_N$ (structure described in Chapter 6, 1.49 Å resolution); residues F87, L70 and I46 are depicted as *red* spheres; the *blue hazed* sphere marks the binding pocket.

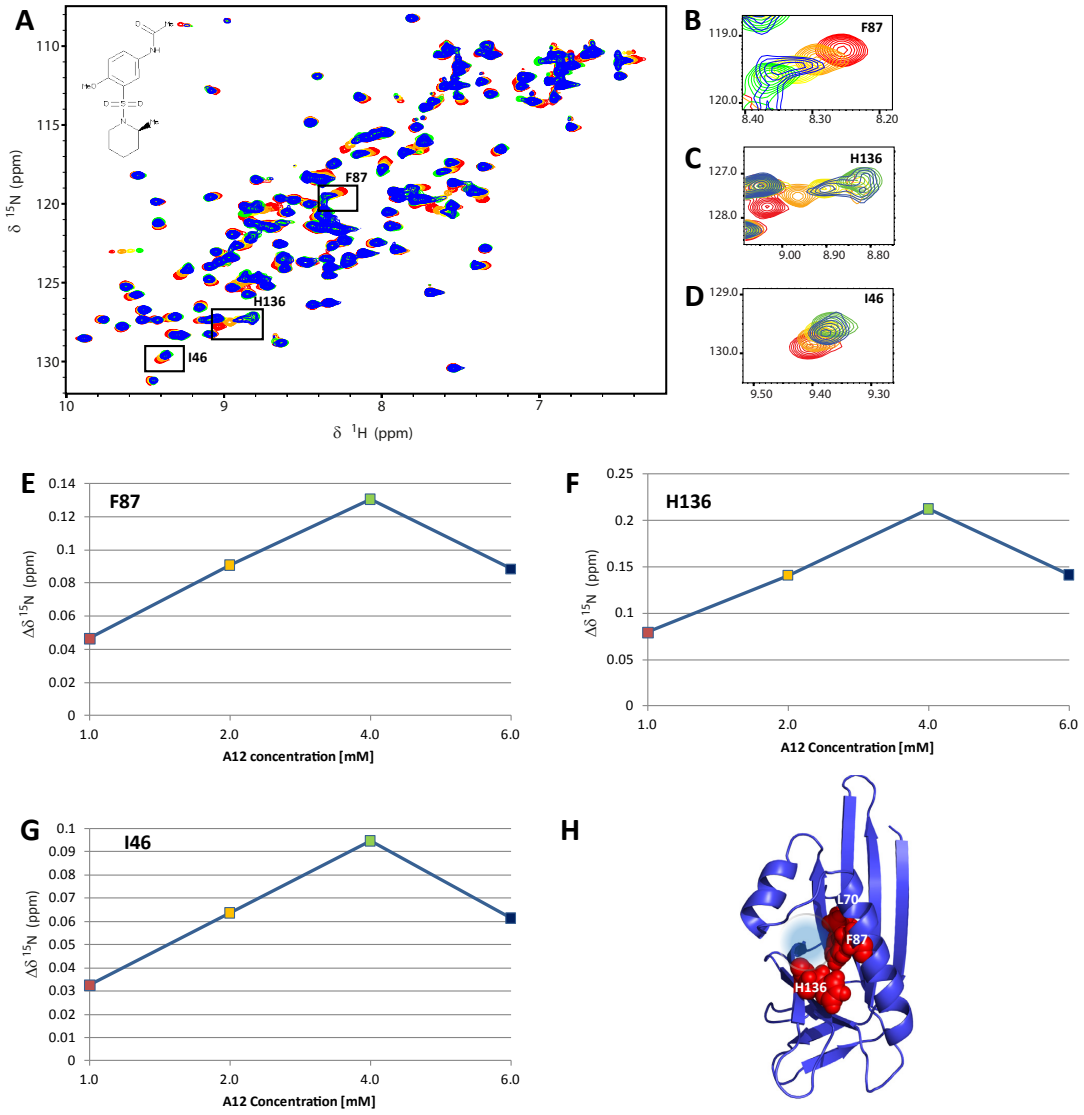


Fig. 5.13: ^{15}N-HSQC NMR Titration - *H. sapiens* SAS-6 N-terminal Domain with A12: (A) Overlay of spectra of ^{15}N hsSAS-6$_N$ F131E with no added A12 (*red*), 1 mM added A12(*yellow*), 2 mM A12 (*orange*), 4 mM A12 (*green*) and 6 mM A12 (*blue*). (B) Zoom in of (A) on resonances assigned to pocket residue F87. (C) Zoom in of (A) on resonances assigned to pocket residue H136. (D) Zoom in of (A) on resonances assigned to pocket residue I46. (E) Chemical shift perturbation of residue F87 plotted against concentrations of A12 (1 mM , 2 mM, 4 mM and 6 mM). (F) Chemical shift perturbation of residue H136 plotted against concentrations of A12 (1 mM , 2 mM, 4 mM and 6 mM). (G) Chemical shift perturbation of residue I46 plotted against concentrations of A12 (1 mM , 2 mM, 4 mM and 6 mM). (H) Structural representation of hsSAS-6$_N$ (structure described in Chapter 6, 1.49 Å resolution); residues F87, H136 and I46 depicted are as *red* spheres; the *blue hazed* sphere marks the binding pocket.

5.2 Candidate Compound Effect on SAS-6 Oligomerisation

The NMR screen gave us an indication of compounds showing signs of binding to the hydrophobic pocket we aimed to target. Next, we wanted to investigate if the candidate compounds are able to inhibit N-terminal domain dimer formation. Due to the UV absorbance of the compounds, biophysical techniques that use UV based detection methods can not be used to study samples containing these compounds. We chose to test the effect of the compounds by sedimentation velocity AUC experiments that allow us to monitor the existence of SAS-6 oligomers with interference optics instead.

For these experiments SAS-6 constructs capable of forming higher order oligomers were used to monitor the candidate compounds' effect on SAS-6 oligomerisation. Oligomerisation for *C. elegans* and *C. reinhardtii* SAS-6 has been studied with the help of AUC previously [103, 160] and several oligomerisation states could be observed with constructs containing the N-terminal domain and a section of the coiled-coil domain (SAS-$6_{N\ CC}$). The *C. elegans* construct SAS-$6_{N\ CC}$ (1-215 S123E) I154W was chosen for AUC experiments, as it has been shown to form more distinguishable oligomers than wild type in a previous study [103]. All AUC experiments included samples with and without addition of 2 mM A11. Since A11 is dissolved in DMSO, as a result samples containing A11 also contained 5 % DMSO. To control for possible effects of the addition of DMSO on the dimerisation, in addition to the buffer control sample, a sample with 5 % DMSO was also measured to observe possible changes on the oligomerisation states.

In our AUC experiments ceSAS-$6_{N\ CC}$ I154W exists as a dimer species - representing the C-C dimer - with a sedimentation coefficient of $\sim$ 3 S (Figure 5.14 A) and

shows a second peak at a sedimentation coefficient of $\sim$ 5 - 6 S representing a SAS-6 tetramer-hexamer equilibrium which shows the additional formation of N-dimer C-C dimer species. This observation is in accordance with the published AUC data for this construct [103]. For the sample containing 5 % DMSO, the multimer peak is slightly shifted to lower sedimentation coefficients. This might be indicative of the slight reduction of N-dimer affinity in the presence of DMSO. The addition of 2 mM A11 has a notable effect on the sedimentation coefficient distribution as the multimer peak disappears, showing a clear inhibition of the N-terminal dimerisation. The experiment was repeated (Figure 5.14 B) with similar results. Further experiments with A12, which does not bind to ceSAS-6 showed no effect on the N-terminal dimerisation, as the sedimentation coefficient distribution for this sample remains unaltered (Figure 5.14 C).

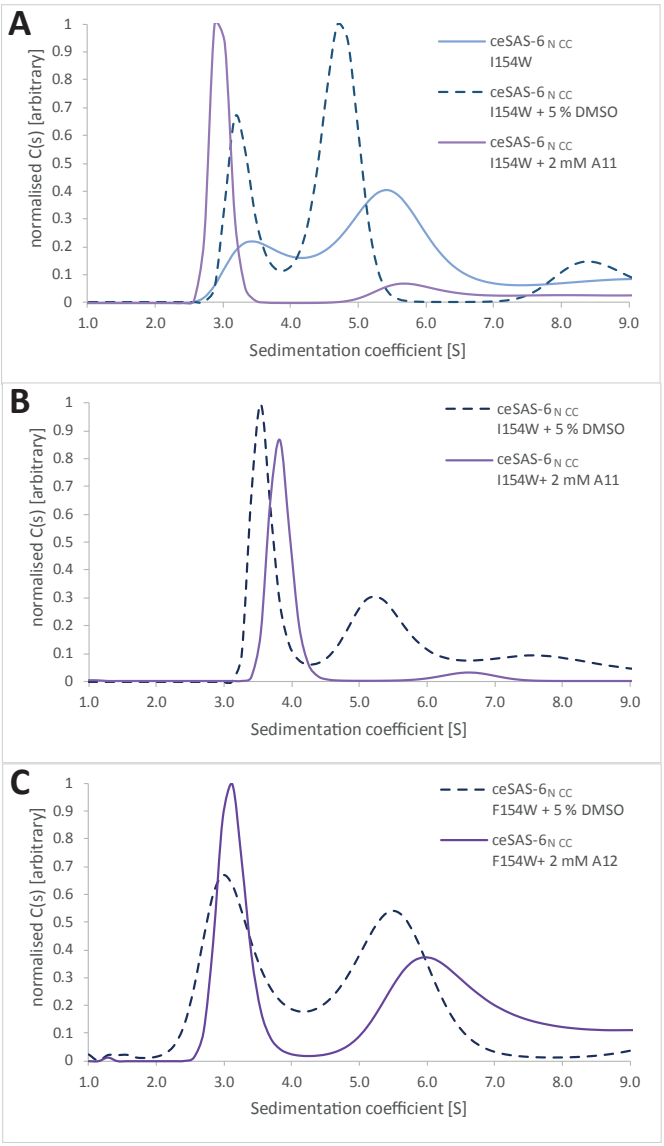


Fig. 5.14: Sedimentation Coefficient Distribution for *C. elegans* SAS-6 Oligomerisation with A11: (A) AUC sedimentation velocity experiment conducted with 50 μM *C. elegans* SAS-6$_{N\ CC}$ I154W. The buffer control sample is depicted in *blue*, the DMSO control sample in *dark blue* and the A11 containing sample in *purple*. (B) Repeat of experiment (A), without buffer control. (C) AUC sedimentation velocity experiment conducted with 50 μM *C. elegans* SAS-6$_{N\ CC}$ I154W. The buffer control sample is depicted in *blue*, the DMSO control sample in *dark blue* and A12 containing sample in *purple*. 1 S = 10^{-13} s

For AUC experiments with the *C. reinhardtii* SAS-6 homologue, crSAS-$6_{N\ CC}$ (1-226) and crSAS-$6_{N\ CC}$ NN24 (Q93E F145W K146R) were both tested [160]. In our AUC experiment, the crSAS-$6_{N\ CC}$ constructs showed several peaks in the sedimentation coefficient distribution indicating oligomerisation (Figure 5.15 A). However, the oligomer fraction was reduced on the addition of 5 % DMSO, so that an effect of A11 could not be seen. For construct crSAS-$6_{N\ CC}$ NN24 which was shown to have a stronger N-terminal dimerisation [160], the addition of 5 % DMSO had no detectable effect on oligomerisation (Figure 5.15 B). The addition of 2 mM A11 had minimal effect on the oligomerisation of this construct.

Since the oligomerisation of the crSAS-6 wildtype construct was inhibited by the presence of DMSO in the control sample, we cannot determine if A11 alone would have had an effect on the wildtype oligomerisation of crSAS-6. The construct NN24 was shown to have increased N-dimerisation affinity with a K_d of $\sim$ 1 μM, compared to the wildtype *C. reinhardtii* SAS-6 N-dimer which dimerises with a $\sim$ 100 fold higher K_d [160]. At the concentration used, A11 was not able to inhibit the oligomerisation of this strengthened construct. Experiments with a construct with stable enough N-terminal dimerisation to withstand the disruptive effect of DMSO and a K_d closer to the wildtype construct would be needed to determine if A11 has an effect on the oligomerisation of *C. reinhardtii*. Alternatively, a further developed compound based on A11 could be tested against crSAS-$6_{N\ CC}$ NN24. Since several other constructs with different mutation of this SAS-6 variant were shown to have strengthened N-dimerisation, a suitable construct might be identified for AUC experiments [160].

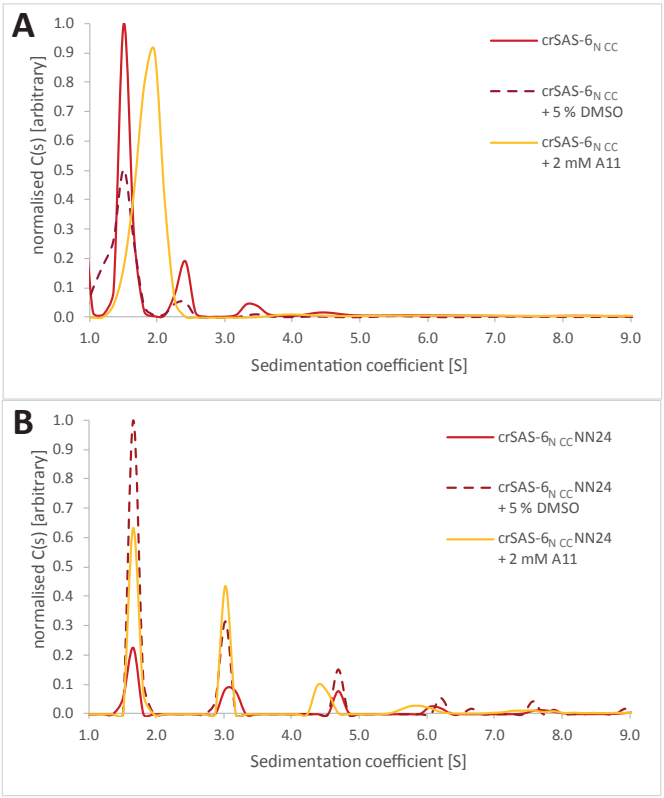


Fig. 5.15: Sedimentation Coefficient Distribution for *C. reinhardtii* SAS-6 Oligomerisation with A11: (A) AUC sedimentation velocity experiment conducted with 100 μM *C. reinhardtii* SAS-6$_{N\ CC}$. The buffer control sample is depicted in *red*, the DMSO control sample in *dark red* and the A11 containing sample in *orange*. (B) AUC sedimentation velocity experiment conducted with 50 μM *C. reinhardtii* SAS-6$_{N\ CC}$ NN24. The buffer control sample is depicted in *red*, the DMSO control sample in *dark red* and the A11 containing sample in *orange*. 1 S = 10^{-13} s = 100 fs

New protein constructs were created for AUC experiments with the *H. sapiens* SAS-6 variant. To create a partial coiled-coil domain which would dimerise reliably, heptad repeat predictions were used to identify construct boundaries. Guided by the construct design of ceSAS-6$_{N\ CC}$ and crSAS-6$_{N\ CC}$ which showed oligomer formation, two constructs of *H. sapiens* SAS-6 were initially created: hsSAS-6 1-199 containing 5 heptad repeats of the predicted coiled-coil domain and hsSAS-6 1-205 with 6 heptad repeats. The coiled-coil dimerisation was tested with SEC multi-angle light scattering

(MALS) (Figure 5.16). For hsSAS-6 1-199, no C-C dimerisation could be detected at 100 μM, whereas for hsSAS6 1-205 an additional peak with higher predicted molecular mass was detected, indicating a monomer-dimer equilibrium.

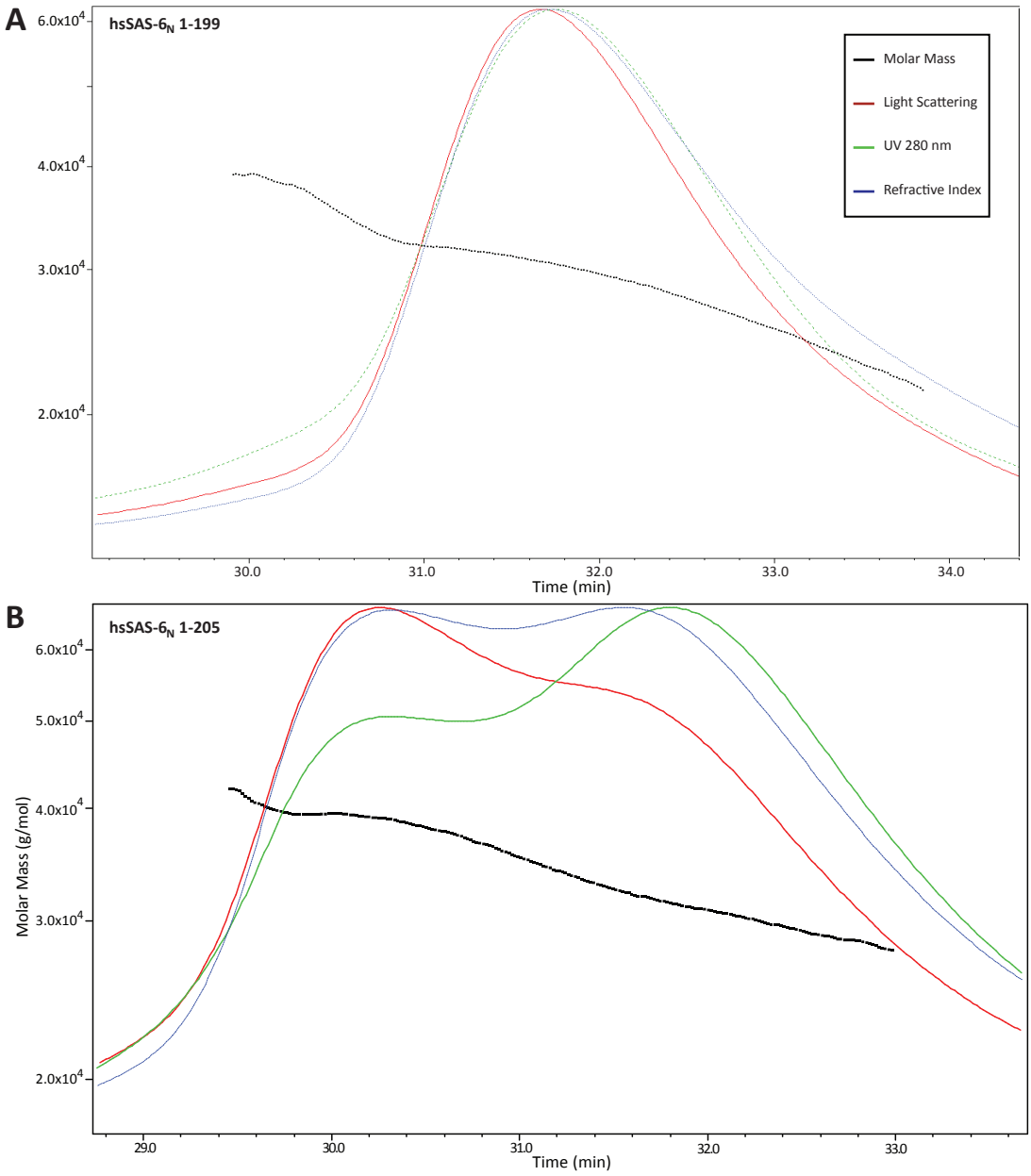


Fig. 5.16: SEC MALS Results for hsSAS-6 1-199 and 1-205: Light scattering shown in *red*, UV Absorbance spectra for 280 nm shown in *green*, refractive index shown in *blue* and molar mass prediction shown in *black* for SEC MALS with: (A) 100 μM hsSAS-6 1-199 and (B) 100 μM hsSAS-6 1-205.

Constructs hsSAS-$6_{N\ CC}$ (1-215) and hsSAS-$6_{N\ CC}$ F131W were used for AUC experiments with A11, and in the case of hsSAS-6 1-215 F131W, also with A12 (see Figure 5.17).

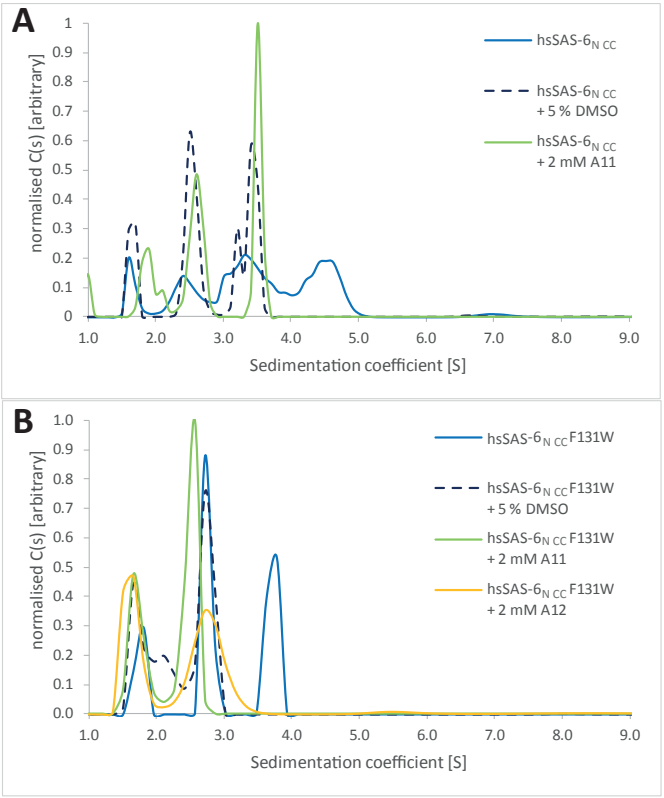


Fig. 5.17: Sedimentation Coefficient Distribution for *H. sapiens* SAS-6 Oligomerisation with A11: (A) AUC sedimentation velocity experiment conducted with 100 μM *H. sapiens* SAS-$6_{N\ CC}$. The buffer control sample is depicted in *blue*, the DMSO control sample in *dark blue* and the A11 containing sample in *green*. (B) AUC sedimentation velocity experiment conducted with 50 μM *H. sapiens* SAS-$6_{N\ CC}$ F131W. The buffer control sample is depicted in *blue*, the DMSO control sample in *dark blue*, the A11 containing sample in *green* and the A12 containing sample in *yellow*. 1 S = 10^{-13} s

The wild type construct shows several peaks in the sedimentation coefficient distribution analysis indicating the presence of several oligomerisation states (Figure 5.17 A). The higher-order oligomer peak however is not present for the sample containing DMSO. No further differences were seen in the distribution for the A11 containing

sample, so that no effect above the DMSO signal could be detected.

Whereas tryptophan mutations of the key residue in loop β6-β7 have been shown to strengthen N-terminal dimerisation for *C. elegans* and *C. reinhardtii* [103, 160] SAS-6, this mutation in *H. sapiens* SAS-6 had not been studied previously. The sedimentation coefficient distribution for this sample also shows several peaks indicating oligomerisation (Figure 5.17 B). For the DMSO control, the peak with the highest sedimentation coefficient is not present showing that this constructs oligomerisation capability is also hindered by the presence of DMSO. The distributions of samples containing A11 and A12 show peaks comparable to the DMSO control. Hence no effect in addition to the disruption caused by DMSO could be detected.

More experiments are needed to establish whether A11 or A12 have an effect on hsSAS-6 oligomerisation. The *H. sapiens* SAS-6 homologue has not been characterised as extensively as the *C. elegans* and *C. reinhardtii* SAS-6 and no information about the effect of mutations is available. To find a construct suitable for experiments in the presence of DMSO, the F131W and further mutations which could improve the N-terminal domain dimerisation need to be characterised. As another conclusion from these experiments and specifically on the influence of the presence of DMSO in samples, the solubility in aqueous buffers of A11 and A12 need to be improved in future compound design steps.

5.3 Conclusion

We tested a selection of 37 compounds - 29 of which were soluble - chosen for their computationally predicted compatibility with the hydrophobic target pocket in the *C. elegans* SAS-6 N-terminal domain with an NMR based screen against the *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6 N-terminal domain constructs.

We identified compound A11 as binding to the target pockets of *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6 in this screen. In a previous publication, compound (PK9119 (ZINC00392813)) was identified as binding to the *L. major* SAS-6 variant, however our NMR experiments showed no binding to the target pocket of *C. elegans* SAS-6 as discussed in Chapter 4. A11 is the first compound identified to bind to *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6, and is the first compound shown to bind to the SAS-6 of several organisms.

The screen described here further identified compound A12, which showed binding in NMR only to the *H. sapiens* SAS-6 N-terminal domain.To facilitate the analysis of the *H. sapiens* SAS-6 screen, experiments for the NMR assignments of this construct were conducted and so far $\sim$ 80 % of residues have been assigned. Further data analysis will lead to improved assignment coverage.

To show the effect of the compounds on SAS-6 oligomerisation, AUC experiments were conducted. The addition of A11 had a notable effect on the oligomer distribution of *C. elegans* SAS-6. An effect on the oligomerisation of *C. reinhardtii* and *H. sapiens* SAS-6 oligomerisation could not be shown in our experiments. This is mostly due to the disruptive effect of the presence of DMSO on the oligomerisation. It is possible that A11 and A12 do affect SAS-6 oligomerisation for *C. reinhardtii* and *H. sapiens* SAS-6, so further experiments are necessary to establish whether these compounds affect N-dimer formation. To achieve this, N-terminal domain constructs for these SAS-6 variants with modulated dimerisation affinity would need to be created. For the strong N-dimer forming *C. reinhardtii* SAS-6, the AUC experiment could be repeated with increased A11 concentration.

Additionally, other biophysical assays would be needed to support our observations. The high UV absorbance of the candidate compounds might influence results

obtained from many biophysical methods based on UV measurements. Biophysical assays which allow the monitoring of dimerisation or oligomer formation and do not rely on UV based detection methods could, for example be performed using surface plasmon resonance or microscale thermophoresis experiments.

Since NMR based titration experiments with A11 and A12 on *C. elegans* and *H. sapiens* SAS-6 indicate a weak binding affinity of the compounds in the millimolar range, the affinity of these compounds needs to be optimised. In addition the solubility of future candidate compounds based on A11 and A12 would need to be improved to remove the necessity for the addition of DMSO. We aim to optimise these candidate compounds with structure based compound design methods.

Chapter 6

Structural Insights into Candidate Compound Binding to the SAS-6 N-terminal Domain

"...we had everything before us, we had nothing before us,..."

Charles Dickens

A Tale of Two Cities

This chapter contains a description of efforts to co-crystallise candidate compounds for the inhibition of SAS-6 N-terminal domain dimerisation with the SAS-6 N-terminal domain and insights into the possible binding pose of candidate compounds in the *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6 N-terminal domains from MD simulations.

6.1 Introduction

Small molecule screens identified compound A11 (ZINC36638814) binding to the dimerisation interface of the SAS-6 N-terminal domain of the *C. elegans*, *C. reinhardtii* and *H. sapiens* (see Chapters 4 and 5). Further to this, compound A12 (ZINC09737303)

was shown to bind to the dimerisation interface of the *H. sapiens* SAS-6 N-terminal domain. NMR titration experiments suggested that the affinity of these compounds for the hydrophobic pocket they target is low (see Chapter 5). Nonetheless, compound A11 was shown to disrupt *C. elegans* SAS-6 oligomerisation in AUC experiments (see Chapter 5).

To be of use as SAS-6 oligomerisation inhibitors in a cellular context, the affinity and solubility of the candidate compounds A11 and A12 need to be improved. Hit compound improvement is commonly approached by analysing interactions of the compound with the target using structural information to identify the key components connected to the activity and affinity of the compound [248–250].

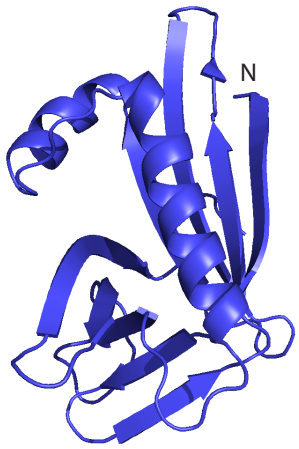
6.2 Crystallographic Approach to Elucidating Binding Poses of Candidate Compounds

For structure based ligand optimisation, atomic resolution information about binding poses is required. Crystal structures of ligand protein complexes can provide such insight. Co-crystallisation and subsequent soaking of *H. sapiens* SAS-6 1-152 F131E with 2 mM A11 or 2 mM A12 were attempted. The broad crystallisation screen resulted in diffracting crystals. Diffraction data collected from these crystals to 1.49 Å resolution was used to solve the first crystallographic structure of *H. sapiens* SAS-6 (Figure 6.1). However, no small molecule density could be identified.

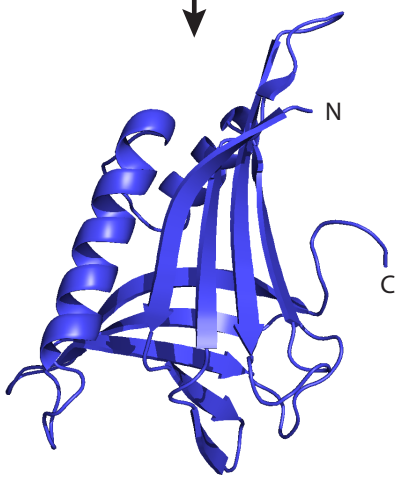
The affinity of protein and ligand is a key factor that influences co-crystallisation [249, 251]. Since NMR titration experiments suggest the binding affinity of A11 and A12 to be in the millimolar range, co-crystallisation might not be achievable (see Chapter 4). The observed tendency of the compounds to form small molecule crystals

at concentrations above 1 mM further hinder co-crystallisation and soaking efforts.

A



B



C

hsSAS-6 1-152 F131E	
Data collection statistics	
Space group	P 2_1 2_1 2
Unit Cell (Å)	a= 59.59 b= 65.13 c= 38.06
Beam line	DLS/i03
Wavelength (Å)	0.97947
Resolution range (Å)	21.98 - 1.46
High resolution shell (Å)	1.512 - 1.46
R_{merge}	0.055(1.009)
R_{pim}	0.02 (0.423)
Completeness (%)	99.9 (99.5)
Multiplicity	6.1 (5.4)
Mean I/σ(I)	15.2 (1.7)
CC-half	0.999 (0.598)
Refinement statistics	
R_{work} (reflections)	18.63 % (26339)
R_{free} (reflections)	23.6 % (1247)
Number of atoms	
Protein (excluding H)	1229
Water	163
Ligands	0
Average B factors (Å^2)	
Protein	31.71
Water	52.78
RMSD from ideal values	
Bonds (Å)	0.011
Angles (°)	1.05
MolProbity statistics[a]	
Ramachandran favoured (%)	98.65
Ramachandran disallowed (%)	0.00
Favoured rotamers (%)	97.22
Poor rotamers (%)	0.00
Clash score (percentile)	0.81 (99th)
MolProbity score (percentile)	0.75 (100th)

Fig. 6.1: Crystal Structure of *H. sapiens* SAS-6 N-terminal Domain: (A) and (B) model of hsSAS-6_N F131E protein structure. (C) Table of crystallographic data and refinement statistics. Values in parentheses correspond to those for the highest resolution shell.

Amongst SAS-6 homologues for which structures exist, *H. sapiens* and *D. rerio* SAS-6_N have the highest sequential conservation (Figure 6.2 A). For this reason, the drSAS-6_N structure (PDB code 2Y3V, 1.92 Å resolution [94]) was used as a template for molecular replacement on hsSAS-6 diffraction data. The novel structure of *H. sa-*

piens SAS-6_N shows high similarity with the published structure of *D. rerio* SAS-6_N. As previously observed the structural semblance between SAS-6 homologoues is high (Figure 6.2 B, C, D).

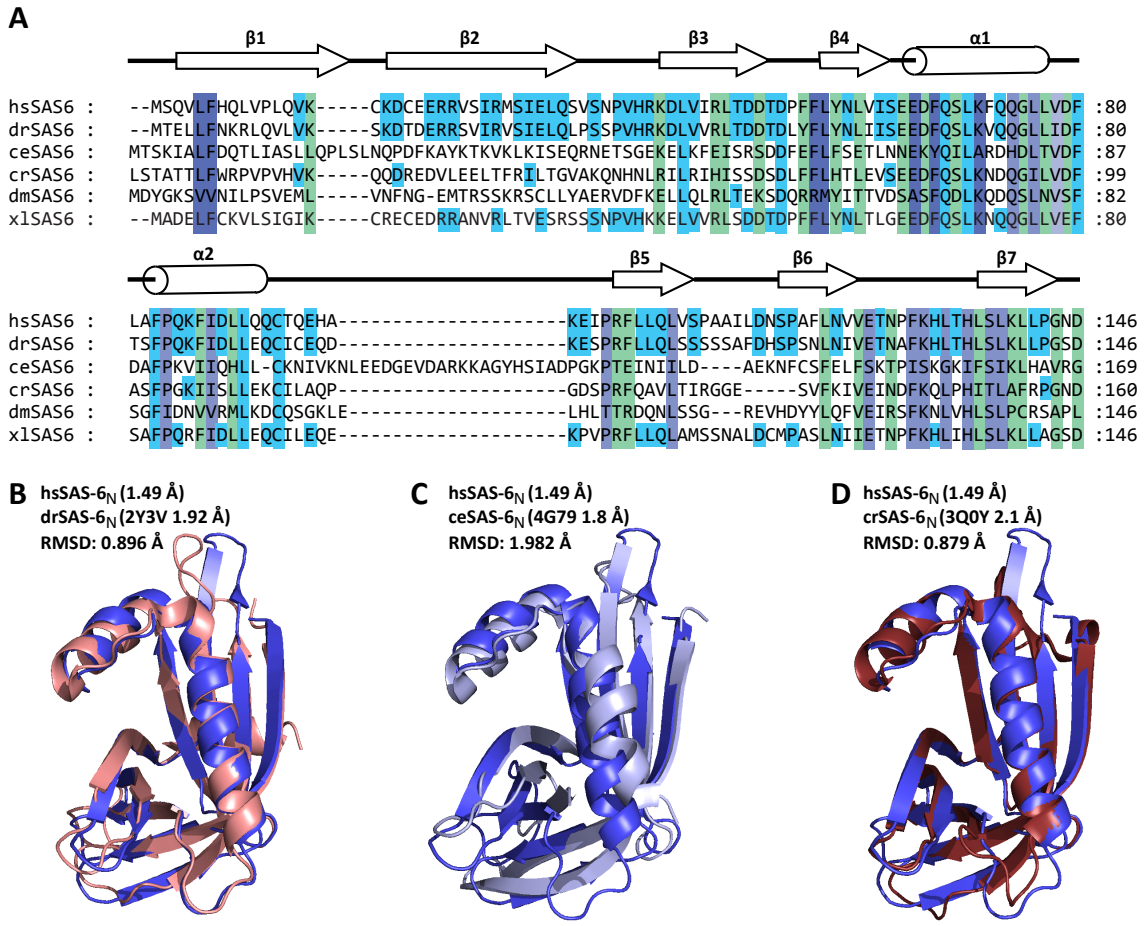


Fig. 6.2: Sequence and Structural Conservation of SAS-6: (A) Alignment of SAS-6 N-terminal domain protein sequence with similar sequence sections indicated by colour.
Structural alignments of new hsSAS-6_N structure (1.49 Å resolution, 150 residues) with: (B) drSAS-6_N structure PDB code 2Y3V, 1.92 Å resolution, 158 residues [94] - RMSD of 0.896 Å (134 Cα atoms used for alignment); (C) ceSAS-6_N structure PDB code 4G79, 1.8 Å resolution, 134 residues [103] - RMSD of 1.982 Å (119 Cα atoms used for alignment); (D) crSAS-6_N structure PDB code 3Q0Y, 2.1 Å resolution, 146 residues [93] - RMSD of 0.879 Å (136 Cα atoms used for alignment).

The overall fold of the N-terminal domain is conserved in the novel structure. This

is also reflected in low RMSD values (< 2 Å) calculated for the structural deviation of the Cα backbones of the structural alignments of hsSAS-6$_N$ with drSAS-6$_N$ (PDB code 2Y3V, 1.92 Å resolution, 158 residues present in structure [94]) - RMSD of 0.896 Å, with 134 Cα atoms used for the alignment (Figure 6.2 B) - , hsSAS-6$_N$ with ceSAS-6$_N$ (PDB code 4G79, 1.8 Å resolution, 134 residues present in structure [103]) - RMSD of 1.982 Å, with 119 Cα atoms used for the alignment (Figure 6.2 C) - and hsSAS-6$_N$ with crSAS-6$_N$ (PDB code 3Q0Y, 2.1 Å resolution, 146 residues present in structure [93]) - RMSD of 0.879 Å, with 136 Cα atoms used for the alignment (Figure 6.2 D). One notable difference is an elongation of β strands four and five present in the structural model of hsSAS-6$_N$ in comparison to the structures of the other homologues. Another difference immediately obvious from structural overlays is a difference in the length of helix α2 between the *C. elegans* homologue (residues 88-101) and other previously solved structures, including the new human structure (residues 80-97) (Figure 6.2 C). The longer α2 helix as well as differences in the fold of the β-sheet seem to account for the higher RMSD between hsSAS-6$_N$ with ceSAS-6$_N$. The extended loop α2-β5 (residue 103-130) in the ceSAS-6$_N$ sequence is not present in the recombinant protein used to solve the ceSAS-6$_N$ structure (PDB code 4G79) and is therefore not included in the RMSD calculation, making the structural difference between these two homologues appear smaller. The RMSD values for hsSAS-6$_N$ with drSAS-6$_N$ and hsSAS-6$_N$ with crSAS-6$_N$ are comparably low despite the higher sequential similarity of hsSAS-6$_N$ with drSAS-6$_N$.

Several constructs of the SAS-6 N-terminal domain of *C. elegans* have been successfully crystallised [93, 103, 147]. Attempted co-crystallisation of A11 with *C. elegans* SAS-6 1-168 Δ 103-130 I154E in broad crystallisation screens resulted in both small molecule and protein crystals (2.1 Å, P2$_1$2$_1$2$_1$ space group). As mentioned in previous

chapters, this construct is monomeric and so is the most likely construct to crystallise with an exposed binding interface, making it accessible for A11 binding. Preliminary analysis of diffraction data collected of these protein containing crystals revealed no small molecule electron density in the expected binding site. Many crystal structures of the *C. elegans* SAS-6 are available, thus, in the interest of other priorities pursued in this study, full structure refinement was not undertaken after an initial model was obtained (see Chapter 2 for methods). No additional small molecule density could be detected in data resulting from soaking of these crystals with A11 solution prior to data collection. Tight crystal packing and the low affinity of the compound to the protein could explain the unsuccessful soaking attempt.

6.3 Investigating Candidate Compound Binding Positions with MD Simulations

In the absence of crystallographic information on the candidate compound binding poses, we used MD simulations to gain insight into the possible conformation of A11 and A12 in the hydrophobic pocket of SAS-6_N. The low affinity of the compounds makes docking positions less likely to reflect possible interactions well. In such cases a refinement of the docking poses by extended atomistic MD simulations can reveal more stable conformations [252].

Structures of A11 and A12 were docked using *Autodock Vina* (see Chapter 2 for methods) into crystallographic structures of *C. elegans*, *C. reinhardtii* and the newly acquired *H. sapiens* SAS-6_N structure (Figure 6.3, 6.4, 6.5 and 6.7). Initially, protein-ligand complexes were simulated for 500 ns. Two repeats of these simulations of 100 ns

length were conducted to gauge the reproducibility of the sampled binding poses. Repeats were simulated for 100 ns, to reduce computational time and accommodate the time frame of this project. Comparable binding positions to the 500 ns simulation were sampled during the 100 ns replicate simulations.

During all simulations conducted with A11 - for *C. elegans*, *C. reinhardtii* and the *H. sapiens* SAS-6_N - the small molecule reoriented in the binding pocket to form interactions with the backbone amides in the α1-α2 connecting loop (Figure 6.3, 6.4 and 6.5). These interactions are mediated by the A11 sulfonamide group. The aromatic ring is positioned towards the centre of the pocket.

In simulations of A11 with *C. elegans* SAS-6_N, the aromatic ring is oriented in a way that positions the chlorine group in the meta position of the ring into the pocket, leaving the opposing ethylether group solvent exposed. This binding position was observed stably in all simulations of A11 and the *C. elegans* and *H. sapiens* SAS-6_N (Figure 6.3 A, B, D and 6.5 A, B).

An alternative conformation of the aromatic ring was observed as the most stable binding position in all simulations with A11 and *C. reinhardtii* SAS-6_N (Figure 6.4 A, B and D) and could also be seen with relative stability during one of the simulations with A11 and *C. elegans* SAS-6_N (Figure 6.3 C). In these cases the ethylether group was positioned towards the centre of the pocket and the chlorine in its stead was solvent exposed (Figure 6.3 C).

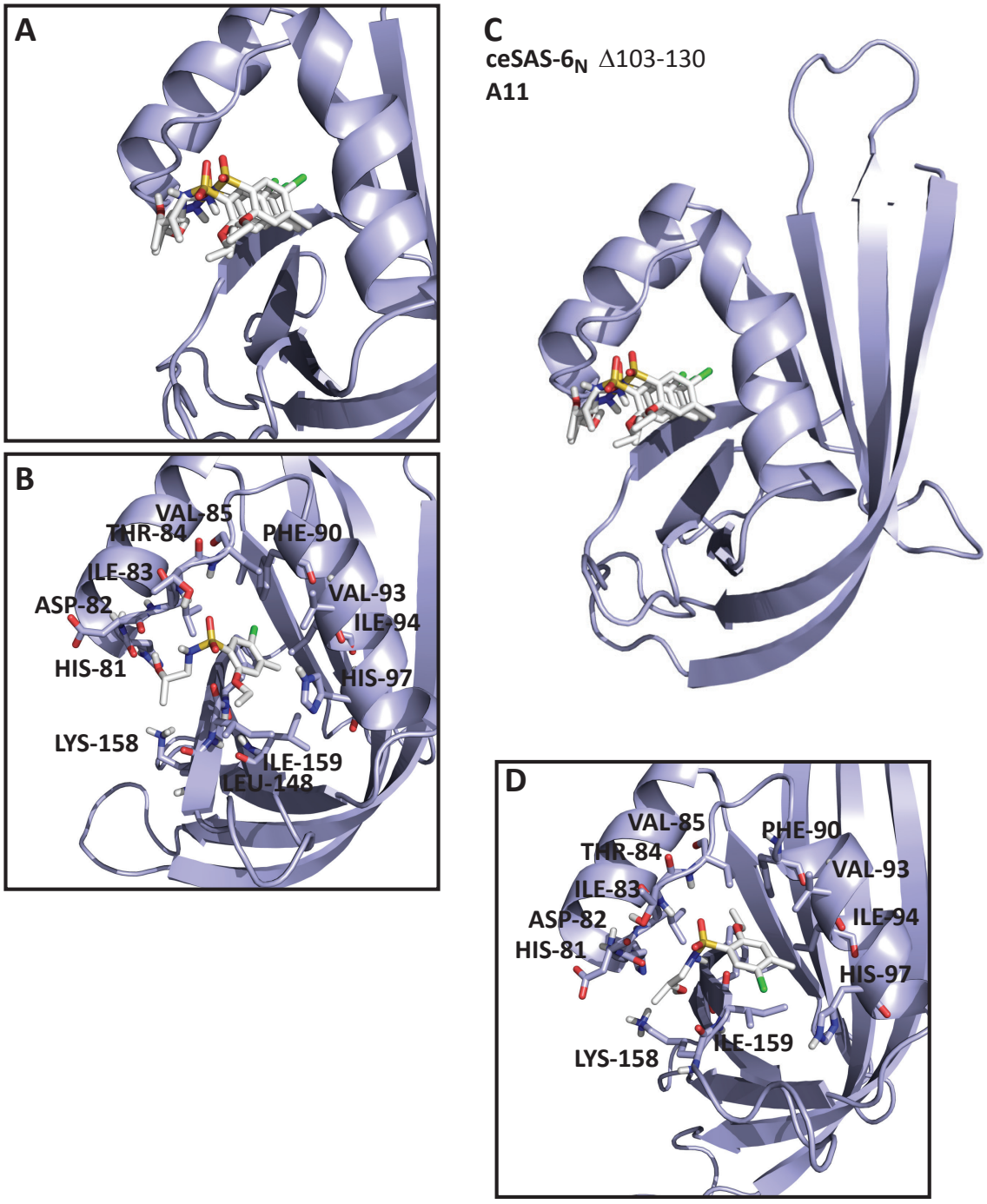


Fig. 6.3: Simulated A11 Binding Position on *C. elegans* SAS-6 N-terminal Domain: (A) Final binding position of A11 from three simulations in the the *C. elegans* SAS-6_N hydrophobic pocket as overlay - zoomed in on pocket. (B) Exemplary A11 binding position from one of three simulations in the *C. elegans* SAS-6_N hydrophobic pocket. (C) Final binding positions of A11 from three simulations in the *C. elegans* SAS-6_N hydrophobic pocket as overlay - full view. (D) Alternate A11 binding position also sampled in one of three simulations in the *C. elegans* SAS-6_N hydrophobic pocket.

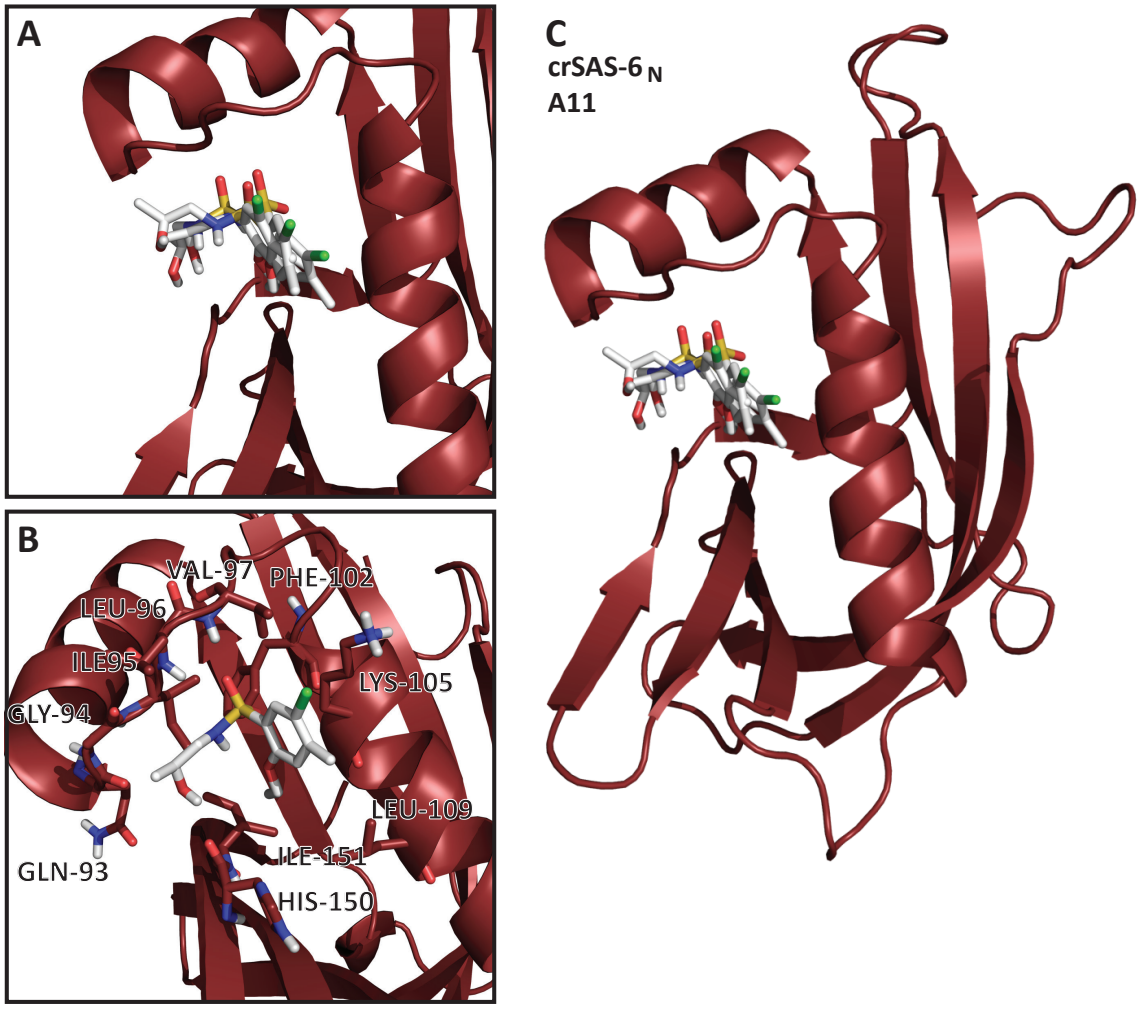


Fig. 6.4: Simulated A11 Binding Position on *C. reinhardtii* SAS-6 N-terminal Domain: (A) Final binding position of A11 from three simulations in the *C. reinhardtii* SAS-6_N hydrophobic pocket as overlay - zoomed in on pocket. (B) Exemplary A11 binding position from one of three simulations in the *C. reinhardtii* SAS-6_N hydrophobic pocket. (C) Final binding position of A11 from three simulations in the *C. reinhardtii* SAS-6_N hydrophobic pocket as overlay - full view.

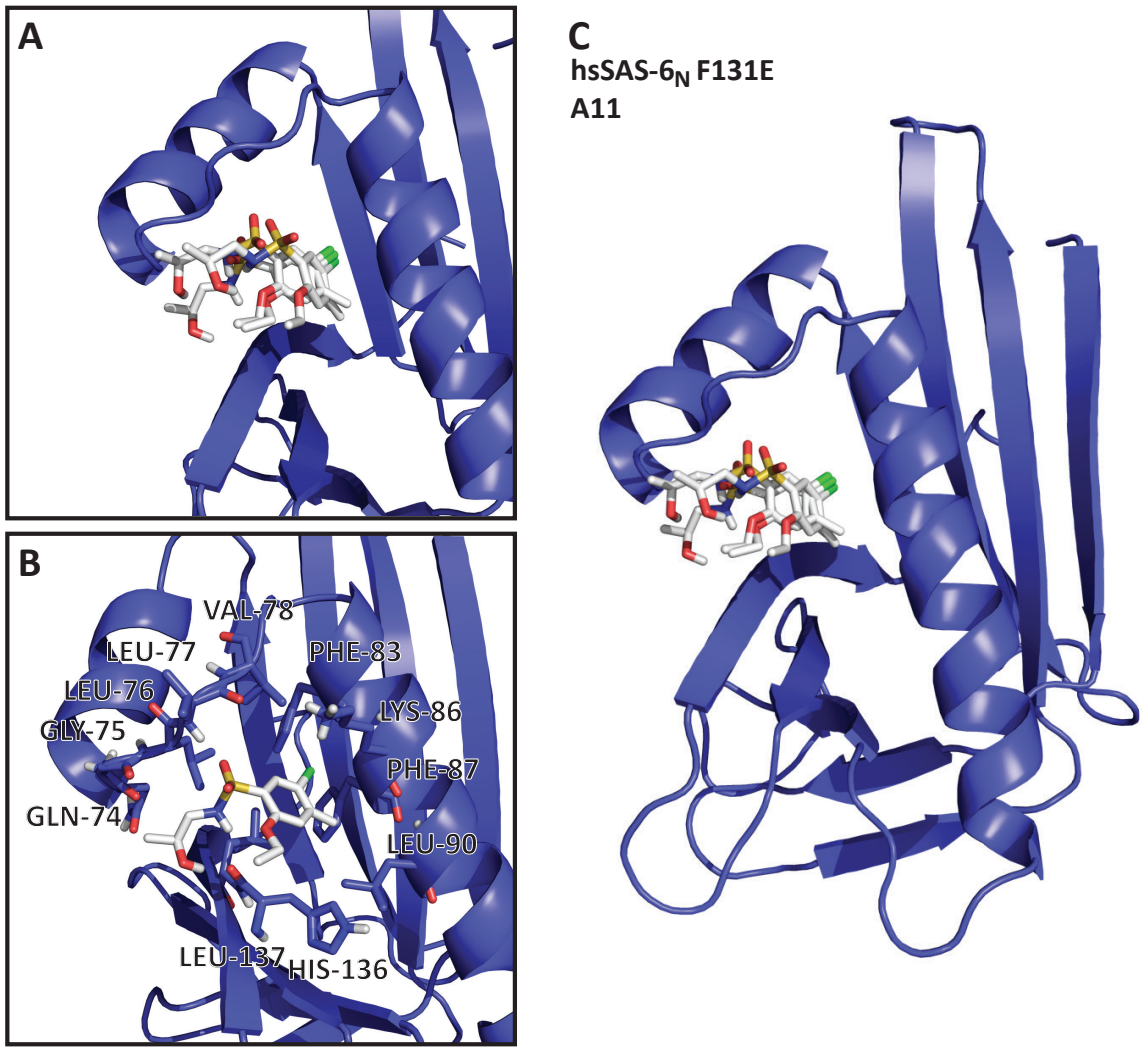


Fig. 6.5: Simulated A11 Binding Position on *H. sapiens* SAS-6 N-terminal Domain: (A) Final binding positions of A11 from three simulations in the *H. sapiens* SAS-6_N hydrophobic pocket as overlay - zoomed in on pocket. (B) Exemplary A11 binding position from one of three simulations in the *H. sapiens* SAS-6_N hydrophobic pocket. (C) Final binding position of A11 from three simulations in the *H. sapiens* SAS-6_N hydrophobic pocket as overlay - full view.

The consistent orientation of the sulfonamide group towards the α1-α2 connecting loop in simulations of all the SAS-6 variants hints at the significance of the sulfonamide group for compound binding (Figure 6.6). Specifically, backbone interactions with three hydrophobic residues (LEU/ILE-THR/LEU-VAL) at the end of the α1-

α2 connecting loop are involved in the sulfonamide interactions. During simulations two more residues in the α1-α2 connector loop were involved in interactions with the A11 molecule. These residues are N-terminally located from the hydrophobic residues stabilising the sulfonamide group of A11 and appear less conserved (HIS/GLN-ASP/GLY). In the vicinity of the two more variable residues (HIS/GLN-ASP/GLY) is the A11 propanol group. Interactions with this group do not appear optimal, as less simulation time is spent in close proximity to residues HIS/GLN-ASP/GLY. The ligand propanol group could be substituted in future to optimise interactions with the species specific α1-α2 connector loop residues, thereby improving the A11 binding affinity and potentially the specificity.

The aromatic group of A11 occupies the centre of the hydrophobic pocket and might aid binding. The orientation of the chlorine and methyl and ethylether groups on the ring appears variable in simulations as mentioned earlier. Two alternate states were observed for comparable extended time periods in simulations, with the chlorine group pointing into the pocket seemingly preferred in simulations of the *C. elegans* and *H. sapiens* SAS-6_N (Figure 6.3 A, B, D and 6.5 A, B, C), whereas the ethylether group occupied the pocket centre in simulations of *C. reinhardtii* SAS-6_N (Figure 6.3 A, B, C). Since the ligand positioning observed in the *C. reinhardtii* SAS-6_N simulations could also be transiently observed in the *C. elegans* SAS-6_N simulations (Figure 6.3 C), modifications of these groups could be tested as a means to improve specificity and affinity. The third site of contacts for the A11 molecule are hydrophobic side-chains originating from the β6-β7 sheet. Contact with these residues could be improved by ring modifications in the ethylether or chloride position. Particularly, a HIS/LYS residue on this sheet could be considered for potential interaction with altered A11-based compounds to improve interactions and specificity.

The binding position of A12 simulated on the *H. sapiens* SAS-6$_N$ shows that the sulfonamide group of A12 forms backbone interactions with residues LEU-LEU-VAL of the α1-α2 loop in a manner similar to A11 (Figure 6.5, 6.6 and 6.7, 6.8).

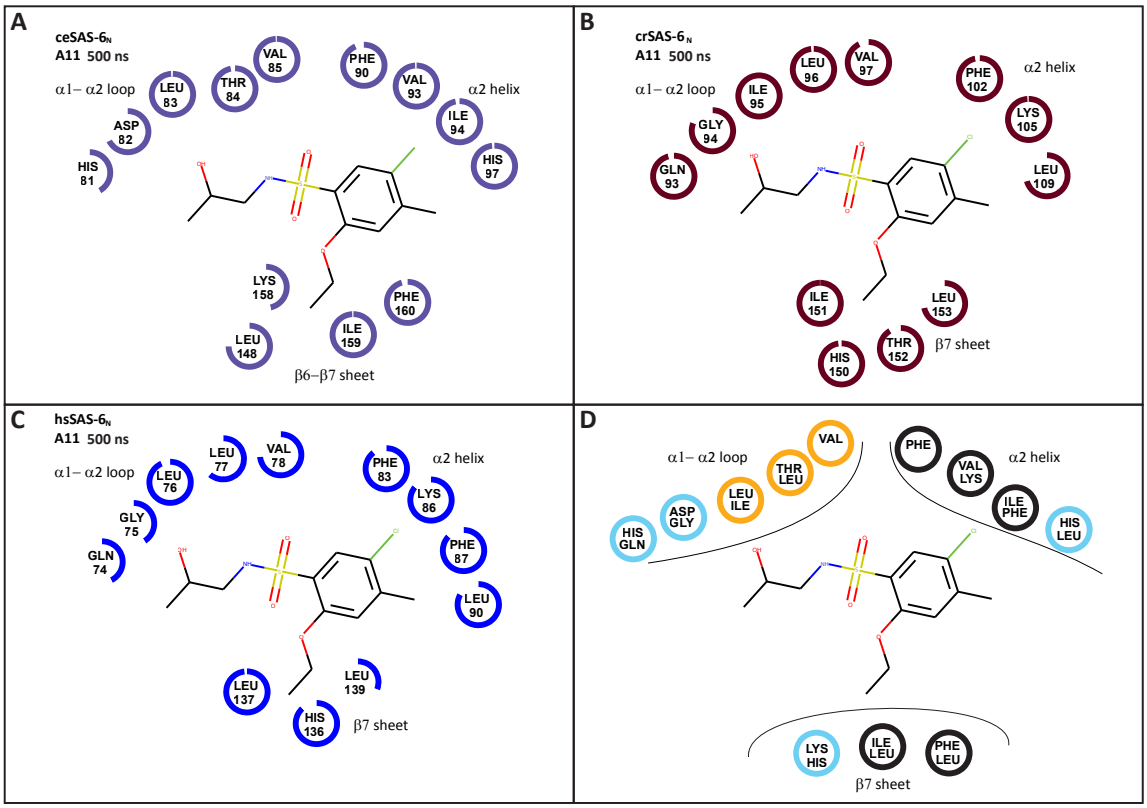


Fig. 6.6: Interaction Time of A11 with Pocket Residues in Simulation: (A) Analysis of time spent by A11 in interaction with residues of *C. elegans* SAS-6$_N$ during 500 ns simulations. (B) Analysis of time spent by A11 in interaction with residues of *C. reinhardtii* SAS-6$_N$ during 500 ns simulations. (C) Analysis of time spent by A11 in interaction with residues of *H. sapiens* SAS-6$_N$ during 500 ns simulations. (D) Depiction of conservation of pocket residues interacting consistently with A11 over (A), (B) and (C). *Black* residues signify conserved residues and residues highlighted in *orange* are crucial for sulfonamide interaction. Residues depicted in *light blue* have less conserved properties and spent less time in interaction with A11.

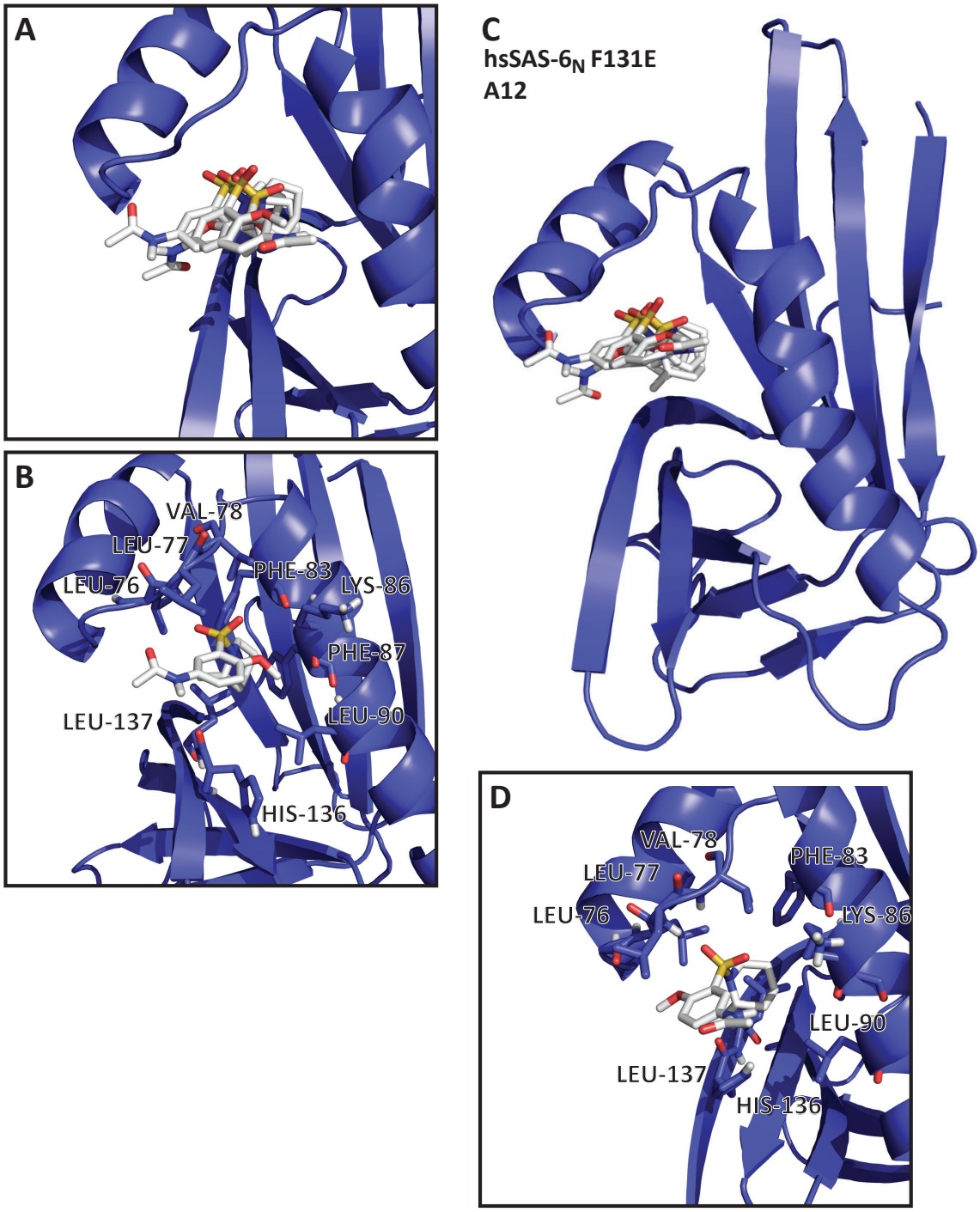


Fig. 6.7: Simulated A12 Binding Position on *H. sapiens* SAS-6 N-terminal Domain: (A) Final binding position of A12 from three simulation in the *H. sapiens* SAS-6_N hydrophobic pocket as overlay - zoomed in on pocket. (B) Exemplary A12 binding position from one of three simulations in the *H. sapiens* SAS-6_N hydrophobic pocket. (C) Final binding position of A12 from three simulation in the *H. sapiens* SAS-6_N hydrophobic pocket as overlay - full view. (D) Alternate A12 binding position also sampled in one of three simulations in the *H. sapiens* SAS-6_N hydrophobic pocket.

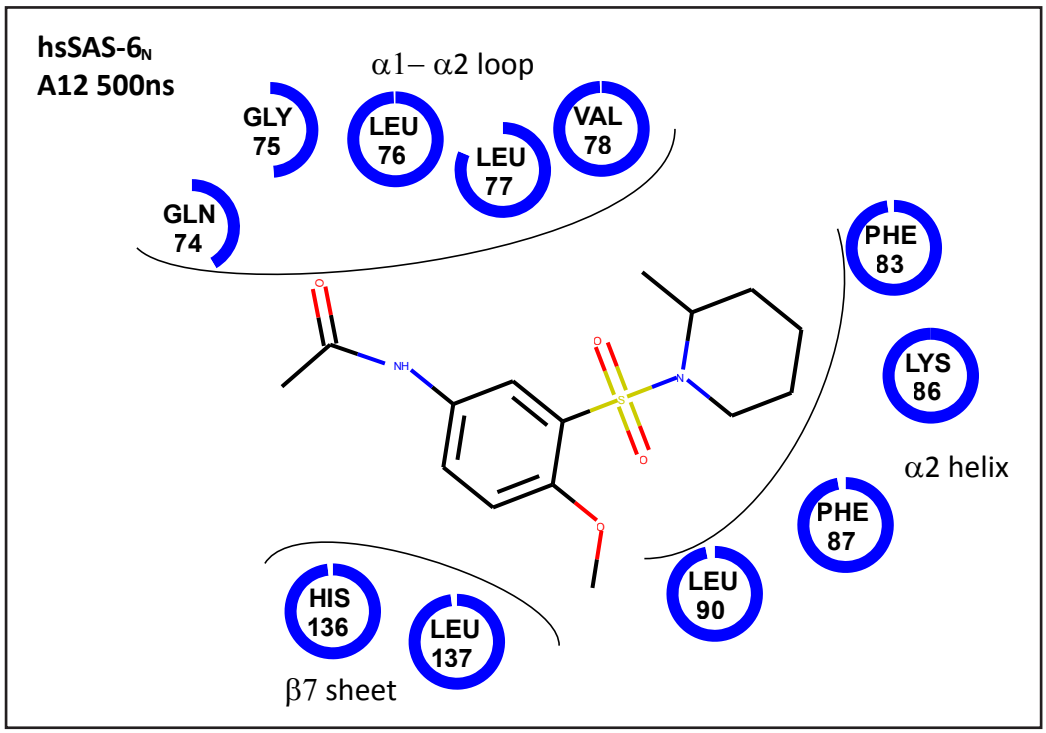


Fig. 6.8: Interaction Time of A12 with Pocket Residues in Simulation: Analysis of time spent by A12 in interaction with resdiues of *H. sapiens* SAS-6_N during 500 ns simulations.

Interestingly, several more sulfonamide containing compounds were included in the initial screening selection of 37 compounds which did not show signs of binding to any of the tested SAS-6 N-terminal domains in an NMR based screen (see Chapter 5) - these and possible implications are further discussed in Chapter 7.2. The specific binding of A12 to only *H. sapiens* SAS-6_N also suggests that the binding of these compounds is not solely due to the sulfonamide group, despite its central role.
Consistently, the methyl-piperidine group is centered towards the centre of the hydrophobic pocket. The aromatic ring of A12 is comparable in position to the A11 propanol group and is close to the less conserved residues in the α1-α2 loop GLN-GLY. The orientation of the aromatic ring is less consistent. In two out of three simulations, the A12 amide group on the aromatic ring points towards the α1-α2 residues (Fi-

gure 6.7 B). In one out of three simulations, the molecule in the most stable position is flipped, so that the methlyether group points towards the loop. These positions could be stabilised through interactions with H136 and L137 on the β7 strand. Modifications of the amide and methylether could improve these interactions and increase the binding affinity.

6.4 Conclusion

Co-crystallisation screens of A11 and A12 with *C. elegans* and *H. sapiens* SAS-6_N yielded protein crystals, but did not provide protein-ligand structures. Data collected from these crystals allowed us to solve the structure of *H. sapiens* SAS-6_N at a resolution of 1.45 Å. This is the first structure of an *H. sapiens* SAS-6 domain.

We performed molecular dynamics simulation of the protein-ligand complex of A11 and *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6_N, as well as A12 and *H. sapiens* SAS-6_N, which allow for an insight into the likely binding position of A11 and A12. This information can be used to optimise ligand binding as further considered in Chapter 7 of this thesis.

Chapter 7

Discussion and Future Directions

"What we do in life echoes in eternity."

The Gladiator

7.1 The Protein-Protein Interface of the SAS-6 N-terminal Domain Dimer

The work presented in this thesis explores aspects of the dimerisation of the N-terminal domain of SAS-6 from different angles. We have studied a loop feature unique to *C. elegans* SAS-6, without which the N-terminal domain of ceSAS-6 can not dimerise (in Chapter 3) and searched for small molecule inhibitors for the N-terminal domain dimerisation (Chapters 4-6). These efforts add to insights of numerous previous studies on SAS-6 structures and biophysical studies into the N-terminal domain dimerisation [93–95, 103, 160, 161]. Here we summarise what our findings indicate about the protein-protein interface and add to what was already known about the dimer structures. We describe the differences between the interfaces and explore the conserved

features in the N-terminal domain dimer interfaces in the SAS-6 variants with which we worked in the research for this thesis.

Our results confirm that the α2-β5 loop is integral to the dimerisation of the *C. elegans* SAS-6 N-terminal domain, but its absence in SAS-6 proteins found in other organism does not hinder SAS-6 N-terminal domain dimerisation. The structures of the Δ 103-130 *C. elegans* SAS-6 N-terminal domain and the other examined SAS-6 orthologues are remarkably similar (as discussed in Chapter 1). However, the amino acid sequences of the SAS-6 N-terminal domains are not as well conserved as their structure. Given that the dimerisation strengths of the studied SAS-6_N variants of *C. elegans*, *C. reinhardtii* and *D. rerio* [94, 103, 160] are comparable - with K_d values of $\sim$ 100 μM - the differences in individual amino acids and their interactions in the dimer interface must compensate for the lack of the loop in non-*C. elegans* SAS-6.

The interface is formed by the α1-α2 helix-turn-helix region, which is the loop connecting β6 and β7 (Figure 7.1 *light pink*), with the conserved key residue at its tip (Figure 7.1 *red*). From our studies (described in Chapter 3), we conclude that loop α2-β5 unique to ceSAS-6, interacts dynamically with residues in the α1-α2 region, thereby extending the dimer interface. Loop residues interacted preferentially with residues in the α1-helix in our simulations (example shown in Figure 7.2 A). The preferentially contacted residues in the α1-helix are not part of the dimer interface observed in the crystal structures of ceSAS-6_N Δ103-130 (PDB code 4G79, 1.8 Å resolution [103]), crSAS-6_N (PDB code 3Q0Y, 2.1 Å resolution [93]) or drSAS-6_N (PDB code 2Y3V, 1.92 Å resolution [94]) (Figure 7.1 A).

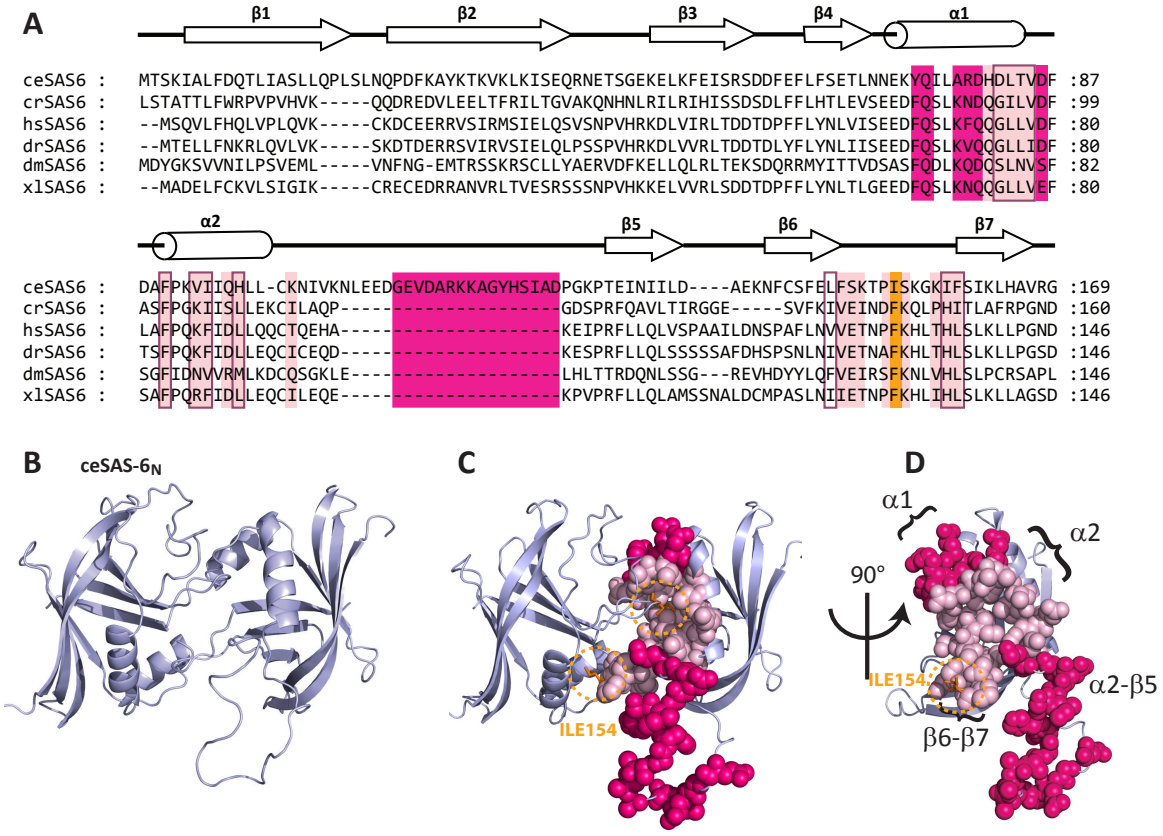


Fig. 7.1: Residues Involved in the SAS-6 N-terminal Dimerisation Interface: (A) Sequence alignment of SAS-6 N-terminal domain sequences. Residues highlighted in *pink* are involved in the interaction of loop α2-β5 within the dimerisation interface in ceSAS-6; residues highlighted in *light pink* are involved in the dimerisation interface; residues with *brown* rim are involved in A11 binding; residue highlighted in *red* is the key residue binding to the hydrophobic pocket in the dimerisation interface. (B) *C. elegans* SAS-6 N-terminal domain dimer structure (PDB code 4G79, 1.8 Å resolution [103]) with model of loop α2-β5. (C) Dimer shown in (B). Isoleucine 154 is shown in *orange* and position highlighted by *orange* circle; residues on one side of the dimer interface are shown as spheres. Residues highlighted in *pink* are involved in the interaction of loop α2-β5 with the dimerisation interface in ceSAS-6; residues highlighted in *light pink* are involved in the dimerisation interface. (D) Monomer of ceSAS-6_N depicted also in (B), rotated 90°.

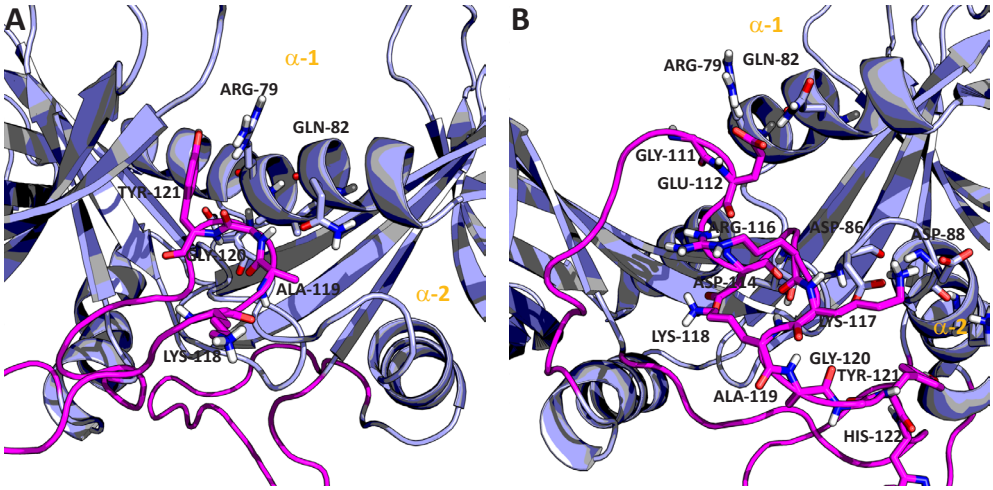


Fig. 7.2: Examples of Loop α2-β5 Interacting with Residues in the α1-α2 Region in the Dimerisation Interface: Example snap shots from MD simulations of ceSAS-6 loop α2-β5 interacting with residues within the N-terminal domain dimer interface. Loop α2-β5 depicted in *pink*. (A) Example 1 - interactions formed between residues in loop α2-β5 and residues in helix α1. (B) Example 2 - interactions formed between residues in loop α2-β5 and residues in α1-α2 region. Simulations described in Chapter 3.

However, dimer interface forming residues in the α1-α2 region were also involved in interactions in the minority of simulations (example shown in Figure 7.2 B and MD analyses described in Chapter 3). Some of the preferentially contacted residues are conserved, but there are differences; the most notable of which is ceSAS-6 A78, which in all other species is a conserved lysine residue. Since this residue is not facing the dimerisation interface, we cannot speculate about the function of the lysine substitution. The type of interaction established between residues in loop α2-β5 and the α1-α2 helix-turn-helix region ranges from side-chain - side-chain to side-chain - backbone and backbone - backbone interactions, and involves both polar and non-polar interactions. Our simulations made clear that the interactions formed are non-static so that at any given time different interactions are established between a different number of amino acids. These factors make it difficult to estimate the potential impact that the sum of these interactions might have on the overall dimerisation affinity of the ceSAS-6_N

domain.

In addition to the unique loop region, the most apparent difference in the dimerisation interface that distinguishes ceSAS-6 is the difference in the hydrophobic residue that connects to the hydrophobic pocket making up most of the dimerisation interface. This key residue is an isoleucine in ceSAS-6, but a phenylalanine in all other species. This difference might well explain some of the reduced binding affinity of the ceSAS-6_N dimer, since we know mutations to tryptophan of this residue strengthen the dimerisation [93, 103]. Interestingly, there is no apparent difference in the pocket size or depth between the isoleucine accommodating and the phenylalanine binding hydrophobic pockets.

There are more differences in the amino acid sequence of the interface residues which might help explain the necessity for the loop for ceSAS-6_N dimerisation. Some of these differences also became apparent in our search for small molecules binding to the hydrophobic pocket. Although no dimer structure for hsSAS-6 exists, from the sequence alignment we know that the drSAS-6_N interface is likely to be similar (Figure 7.1 A). A dimeric crystallographic structure of drSAS-6_N exists (2Y3V) (Figure 7.3 E and F).

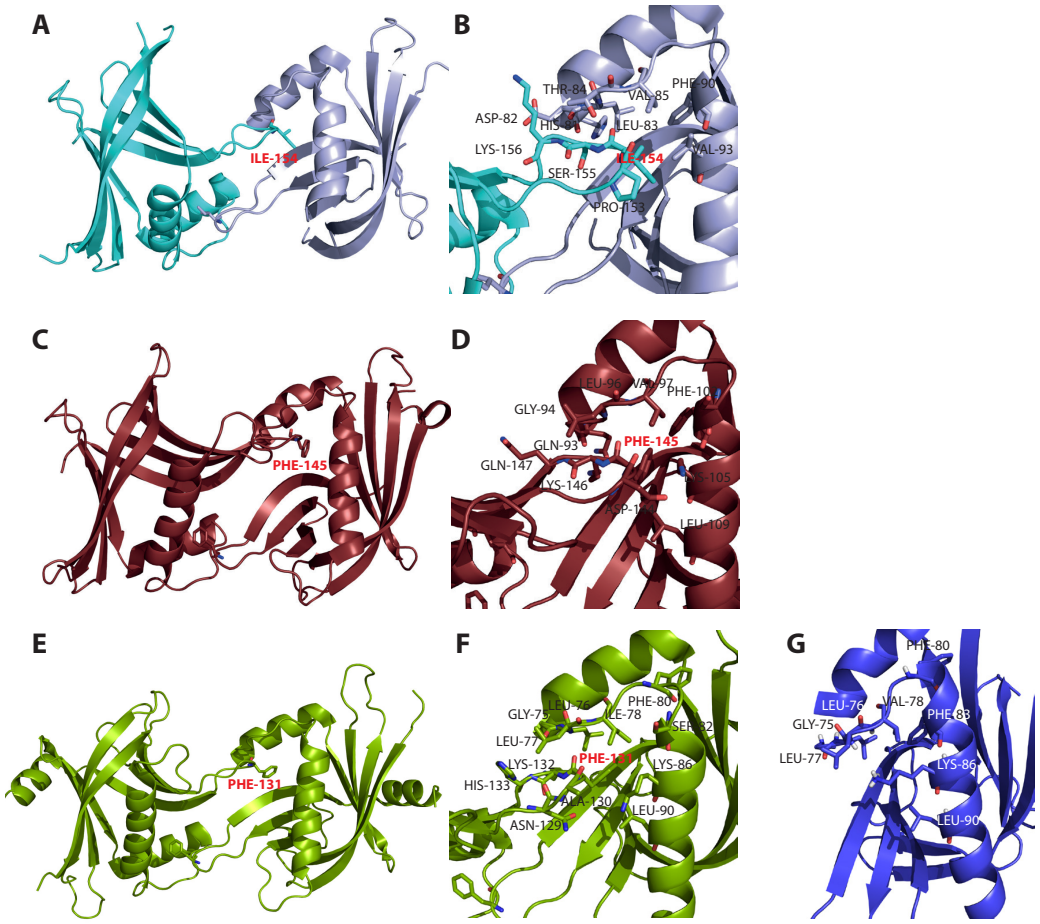


Fig. 7.3: SAS-6 N-dimer Interface Structures (A) *C. elegans* N-dimer as present in structure PDB code 4G79, 1.8 Å resolution [103]. (B) Interface residues in structure PDB code 4G79, 1.8 Å resolution [103]. (C) *C. reinhardtii* N-dimer as present in structure PDB code 3Q0Y, 2.1 Å resolution [93]. (D) Interface residues in structure PDB code 3Q0Y, 2.1 Å resolution [93]. (E) *D. rerio* N-dimer as present in structure PDB code 2Y3V, 1.92 Å resolution [94]. (F) Interface residues in *D. rerio* structure PDB code 2Y3V, 1.92 Å resolution [94]. (G) Monomeric structure of *H. sapiens* SAS-6_N, 1.49 Å resolution (described in Chapter 6).

Given this and our knowledge of the hydrophobic pocket from inhibitor screening, we can hypothesis on what the dimer interface of the *H. sapiens* SAS-6 N-terminal domain dimer must look like and visualise it with the help of the monomeric N-terminal domain structure presented in this study (Figure 7.3 G).

The compound A11 that binds to ceSAS-6_N crSAS-6_N and hsSAS-6_N interacts with residues outlined in brown in Figure 7.1 A. The residues in the interface A11 interacts with, are more similar between crSAS-6 and hsSAS-6 than ceSAS-6. Apart from residue V85 and F90 in ceSAS-6, there are no conserved residues shared between all three species. The patch of amino acids in the hydrophobic pocket responsible for most of the interaction with the key residue I/F and its supporting β-sheet turn also features prominently in the A11 binding. Specifically, these residues are located in the turn between α1 and α2 - DLTV in ceSAS-6, GILV in crSAS-6 or GLLV in hsSAS-6. The valine residue at the end of this turn motif is the before mentioned conserved V85 (ceSAS-6 residue). This residue is positioned in a way to potentially form backbone interactions with the key residue while also allowing for hydrophobic interactions in the pocket. Further residues in this short motif are in close contact with the key residue and the β-turn in which it is contained, but despite the low sequence conservation the interactions formed are amino acid backbone contacts, so that there is little indication for the need for a specific amino acid sequence in this region. The second conserved residue F90 (in ceSAS-6) is part of the centre of the hydrophobic pocket, and might allow aromatic interactions with the key residue in SAS-6_N variants where the key residue is a phenylalanine. This would add about 10 kJ/mol to the interaction [253].

Other contributing factors to the difference in dimerisation interfaces of ceSAS-6, crSAS-6 and hsSAS-6 can be seen in the residues flanking the hydrophobic key residue. For ceSAS-6, the key residue I154 is flanked by P153 which faces residue V93 on the α2-helix. In the case of crSAS-6, the residue N-terminal to key residue F145 is D144 which interacts with K105 - the residue taking the place of the mentioned V93 in ceSAS-6. The interacting residues in hsSAS-6 are slightly shifted in comparison to ceSAS-6 and crSAS-6. Here N129 precedes the key residue F131, and interacts

with the residue in the position of the ceSAS-6 V93 which in this case is residue K86. Notably, in the crSAS-6 and hsSAS-6 dimer these interactions are likely to be charged interactions and therefore stronger compared to the hydrophobic interactions present in the ceSAS-6 dimer interface in the same position ($\sim$ 5.6 kJ/mol compared to $\sim$ 2 kJ/mol) [254].

Clear hydrogen bond and charged interactions are rare across the dimerisation interface and hydrophobic interactions seem to dominate. This leaves very little on which to form a hypothesis about the individual causes of the varying interactions strengths of the N-terminal dimerisation domain of SAS-6 between species. We have to assume it is the accumulation of many small effects that allow SAS-6_N dimerisation without a loop α2-β5 in all species except ceSAS-6, where only the additional cooperativity contribution of the loop flexible residues results in biologically relevant binding affinity.

7.2 Discussion on Next Steps in Structure Based Inhibitor Design Targeting SAS-6 Oligomerisation

In this thesis we describe steps taken to find small molecule inhibitors for SAS-6 oligomerisation meant for use in cell biological experiments to further study centriole duplication and cell cycle processes. *In silico* screening allowed us to select a range of compounds for initial *in vitro* screens (see Chapter 4 and Chapter 5). We showed compound A11 binding to the N-terminal domains of SAS-6 orthologues of three species: *C. elegans*, *C. reinhardtii* and *H. sapiens* and compound A12 binding exclusively to *H. sapiens* SAS-6 (see Chapter 5).

Preliminary NMR experiments suggested A11 and A12 to have affinities to the

N-terminal domain in the millimolar range. Nevertheless, compound A11 was shown to have an impact on ceSAS-6 oligomerisation in AUC experiments (see Chapter 5). Since A11 is the first compound to be identified to bind to several orthologues of SAS-6 and A12 is the first compound shown to bind hsSAS-6 selectively, we propose it would be of interest to optimise binding of these compounds. In addition, the solubility limits of these compounds need to be addressed in compound optimisation. The impact of solubility limits became apparent when conducting NMR experiments at concentrations of and above 6 mM compound and AUC experiments where the DMSO used to aid A11 and A12 solvation influenced SAS-6 oligomerisation, making it impossible to judge the compound effect (see Chapter 5).

Here we discuss what possible modifications of A11 and A12 could be based on structural insights from MD simulations (described in Chapter 6) and suggest what the next experiments on the way towards high affinity inhibitors might look like.

Classical drug discovery pipelines most commonly consist of initial high-throughput screening followed by more focused rounds of screening with compounds selected for common functional features and divers modifications modulating the chemicals properties [255, 256]. Several iterations of screening and compound selection refinement are performed until a high affinity inhibitor is identified. Under ideal circumstances the resulting lead compound has a low micromolar or even nanomolar affinity for the target.

The first step towards a selection of compounds with the potential for higher affinity hits is to understand the structure-activity relationship (SAR) of the compound and the target pocket [257, 258]. With the help of MD simulations we evaluated the potential binding poses of A11 and A12 (Figure 7.4 and Figure 7.5). We concluded that the sulfonamide group interacting with the amino acid backbone of residues just

before the end of loop α1-α2 is essential. (Figure 7.6 A).

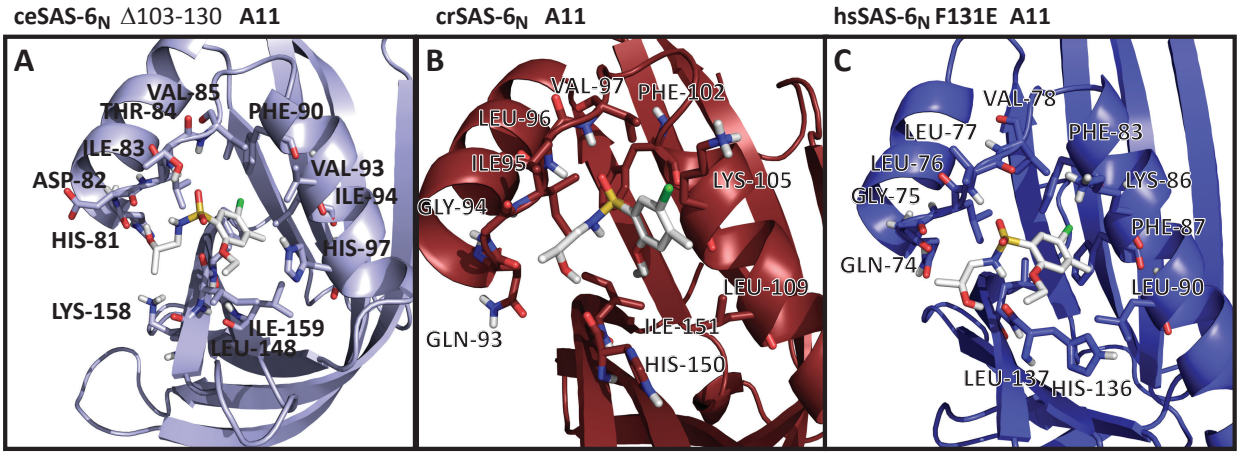


Fig. 7.4: Predicted Binding Poses of A11 on SAS-6 N-terminal Domain: MD refined binding poses of A11 in the hydrophobic pocket of SAS-6$_N$ of (A) *C. elegans* (B) *C. reinhardtii* and (C) *H. sapiens*. For details refer to Chapter 6.

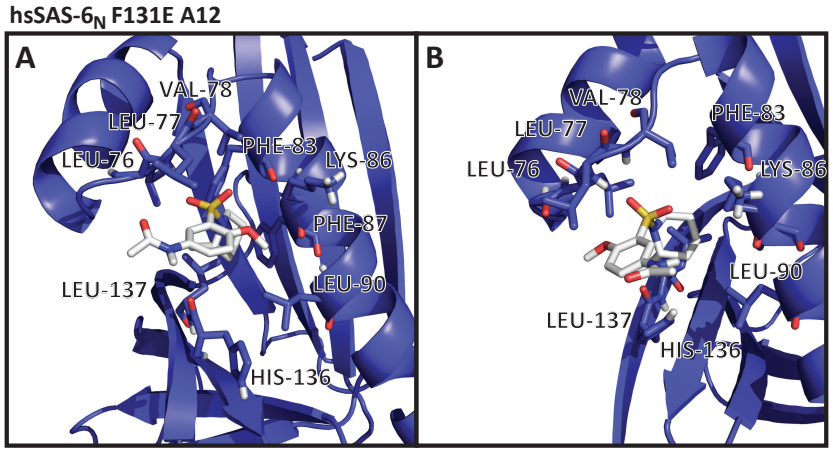


Fig. 7.5: Predicted Binding Poses of A12 on SAS-6 N-terminal Domain: MD refined binding poses of A12 in the hydrophobic pocket of *H. sapiens* SAS-6$_N$. (A) Binding poses found most stable in two of three simulations. (B) Alternative binding pose found in one of three simulations. For details refer to Chapter 6.

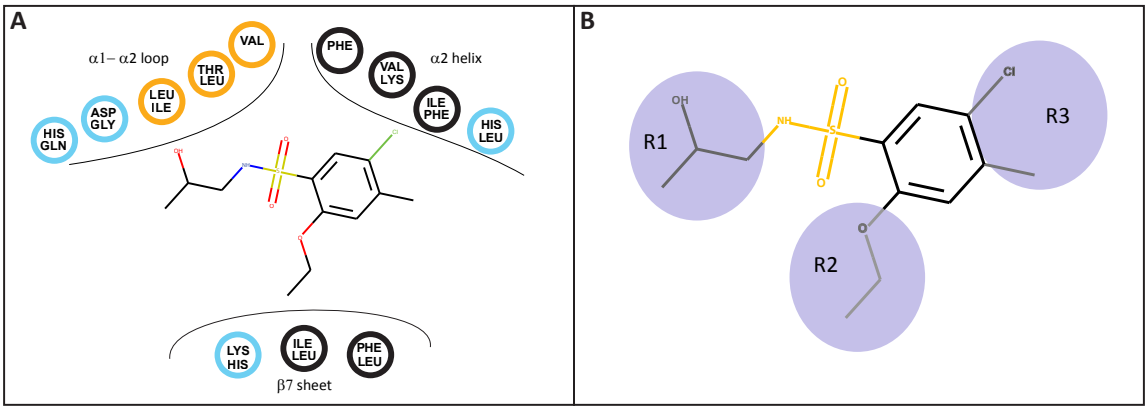


Fig. 7.6: Identifying A11 Components Essential for SAS-6 Binding: (A) Consensus of hydrophobic pocket residues involved in binding A11 as seen in simulations of A11 with *C. elegans*, *C. reinhardtii* and *H. sapiens* SAS-6_N (refer to Chapter 6). Residues with *orange* circle represent consensus sulfonamide binding motif; *black* circle indicates residues with conserved features; *light blue* circle indicates residues with less well conserved characteristics. (B) A11 depicted as a functional core and suggested positions of modification for potential ligand optimisation R1, R2 and R3.

The selection of compounds used in the NMR based screen that produced A11 and A12 also included a number of other sulfonamide containing compounds, which did not show binding to any of the tested SAS-6 orthologues (Figure 7.7). For compound A3 (ZINC72414547), a clear explanation exists for its failure to bind to the SAS-6 N-terminal domain in the screen as it was determined to be insoluble in the aqueous buffer used in NMR experiments. The $x\log P$ of 4.41 assigned to this compound is one of the highest in the screen and in this case clearly reflects the *in vitro* behaviour of the compound quite well. The potential of a similar sulfonamide to bind could instead be tested by using the alternative compound in ZINC41096198 with $x\log P$ of 2.64. This compound contains an isopropanol instead of an isopropanyl group, and a methylether instead of ethylether, making the solubility properties more favourable. This compound would be included in a selection of compounds for the next round of screening.

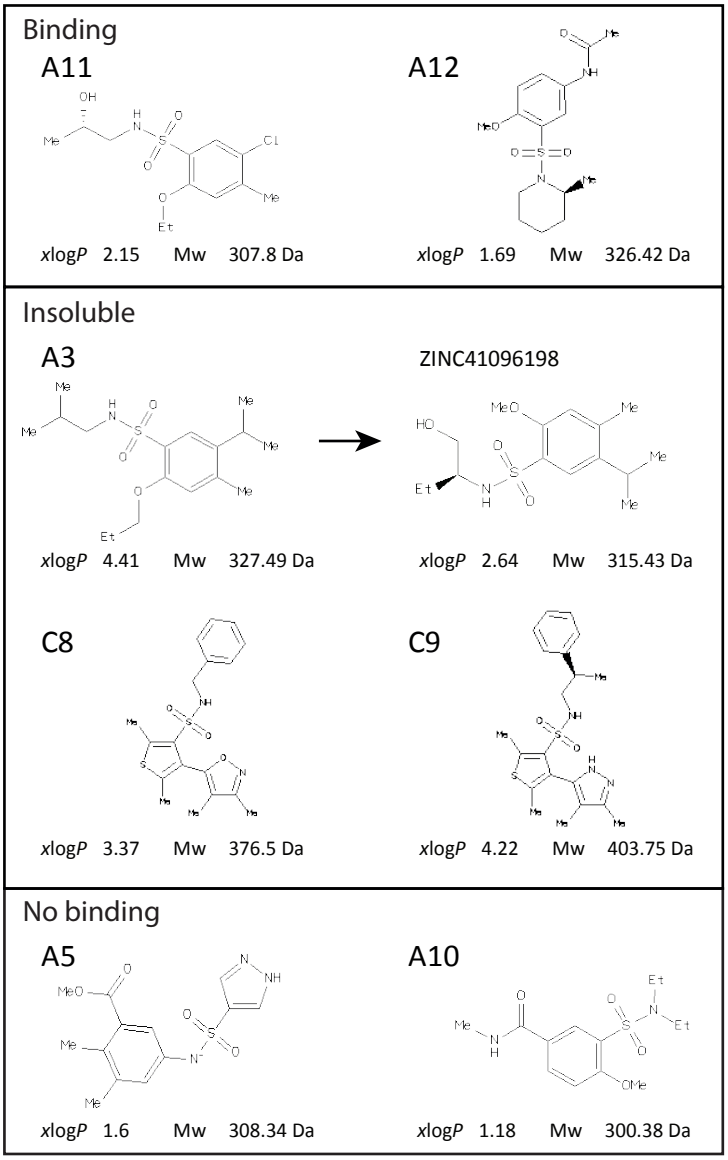


Fig. 7.7: Sulfonamide Containing Compounds in Screening Selection: Depiction of chemical structure of compounds containing sulfonamide groups from compound selection used for *in vitro* screening in Chapter 5. Identified as showing binding to SAS-6_N: A11 and A12; insoluble: A3, C8 and C9; no binding: A5 and A10. Compound ZINC41096198 depicted as a possible alternative for compound A3 for future use.

Similarly, the binding potential of compounds C8 (ZINC64845906) with $x\log P$ of 3.37 and C9 (ZINC33306742) with $x\log P$ of 4.22 could not be assessed due to their low solubility in the NMR buffer. A more soluble alternative for these compounds is

not immediately available. The compounds are more dissimilar to A11 and A12, so that suggestions for suitable alterations are not obvious without a deeper SAR insight into A11 and A12.

Both compounds A5 (ZINC58247642) with $x\log P$ of 1.6 and A10 (ZINC13581301) with $x\log P$ of 1.1 were soluble in NMR buffer. In the *in vitro* screen (see Chapter 5) these compounds did not show clear binding to the hydrophobic target pockets of any of the tested SAS-6 proteins (Figure 7.8, Figure 7.9 and Figure 7.10), despite the presence of sulfonamide derivative groups. Solely in the experiment of crSAS-6_N with A5 can a pattern of possible weak interactions for residues similar to the ones affected by A11 binding be seen (Figure 7.9 A and B). This might be a sign of weak affinity, however the intensity of the observed effect is much weaker than in the A11 experiment, which leads us to categorise this compound as a non-hit.

We know from simulations with A11 and A12 that the sulfonamide group binds to pocket residues backbone in at least two orientations. The ability of A12 to bind to the hsSAS-6 hydrophobic pocket also suggests that the accessibility of the sulfonamide nitrogen is not essential, so that the modifications in A5 and A10 at the sulfonamide nitrogen do not seem to explain their failure to bind.

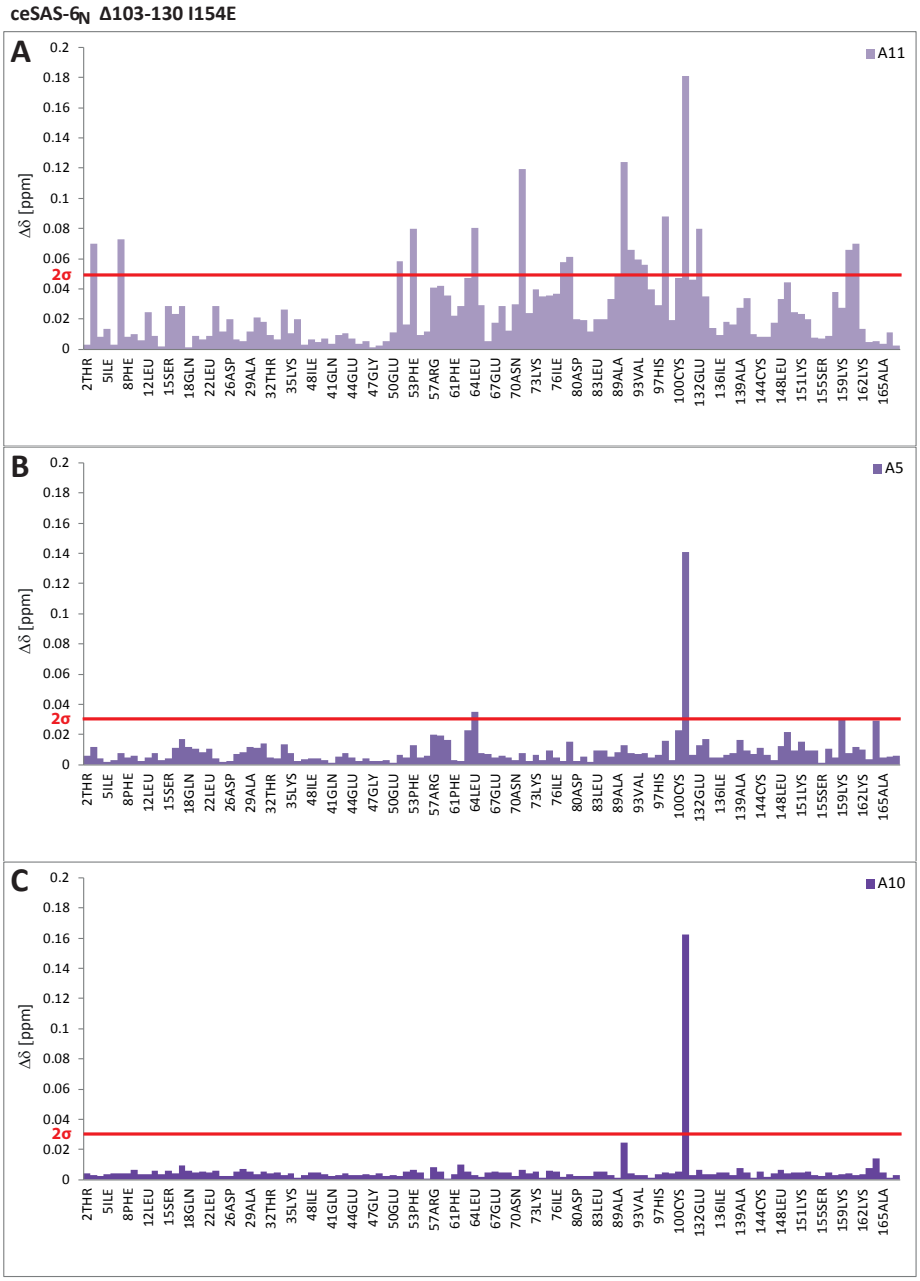


Fig. 7.8: Resonance Differences from NMR Screen with *C. elegans* SAS-6$_N$ for Compounds A11, A5 and A10: (A) Resonance difference seen as effect of addition of 2 mM of compound A11 with $2\sigma = 0.056$ ppm; (B) addition of A5 with $2\sigma = 0.027$ ppm; (C) addition of A10 with $2\sigma = 0.029$ ppm.

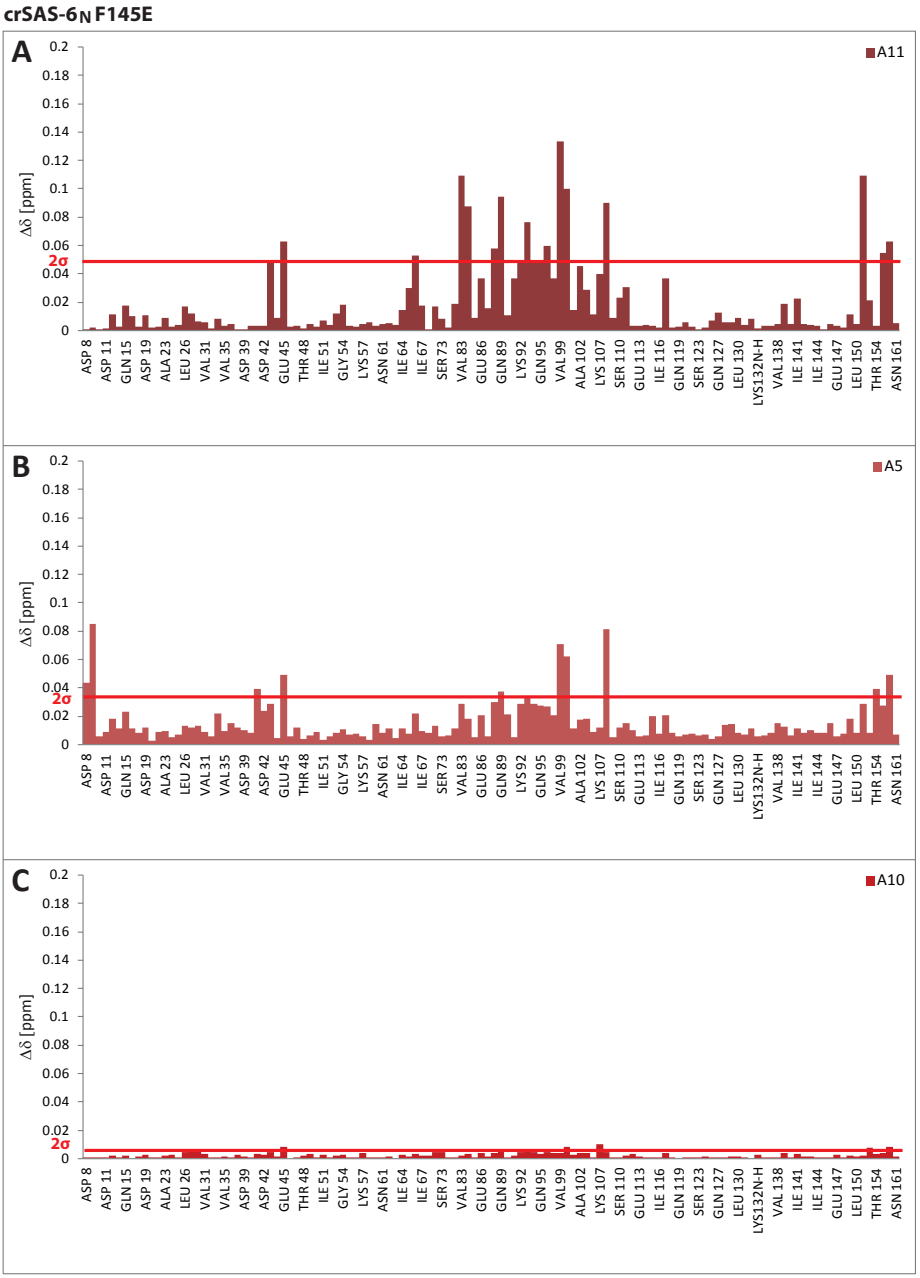


Fig. 7.9: Resonance Differences from NMR Screen with *C. reinhardtii* SAS-6$_N$ for Compounds A11, A5 and A10: (A) Resonance difference seen as effect of addition of 2 mM of compound A11 $2\sigma = 0.054$ ppm; (B) addition of A5 with $2\sigma = 0.029$ ppm; (C) addition of A10 with $2\sigma = 0.004$ ppm.

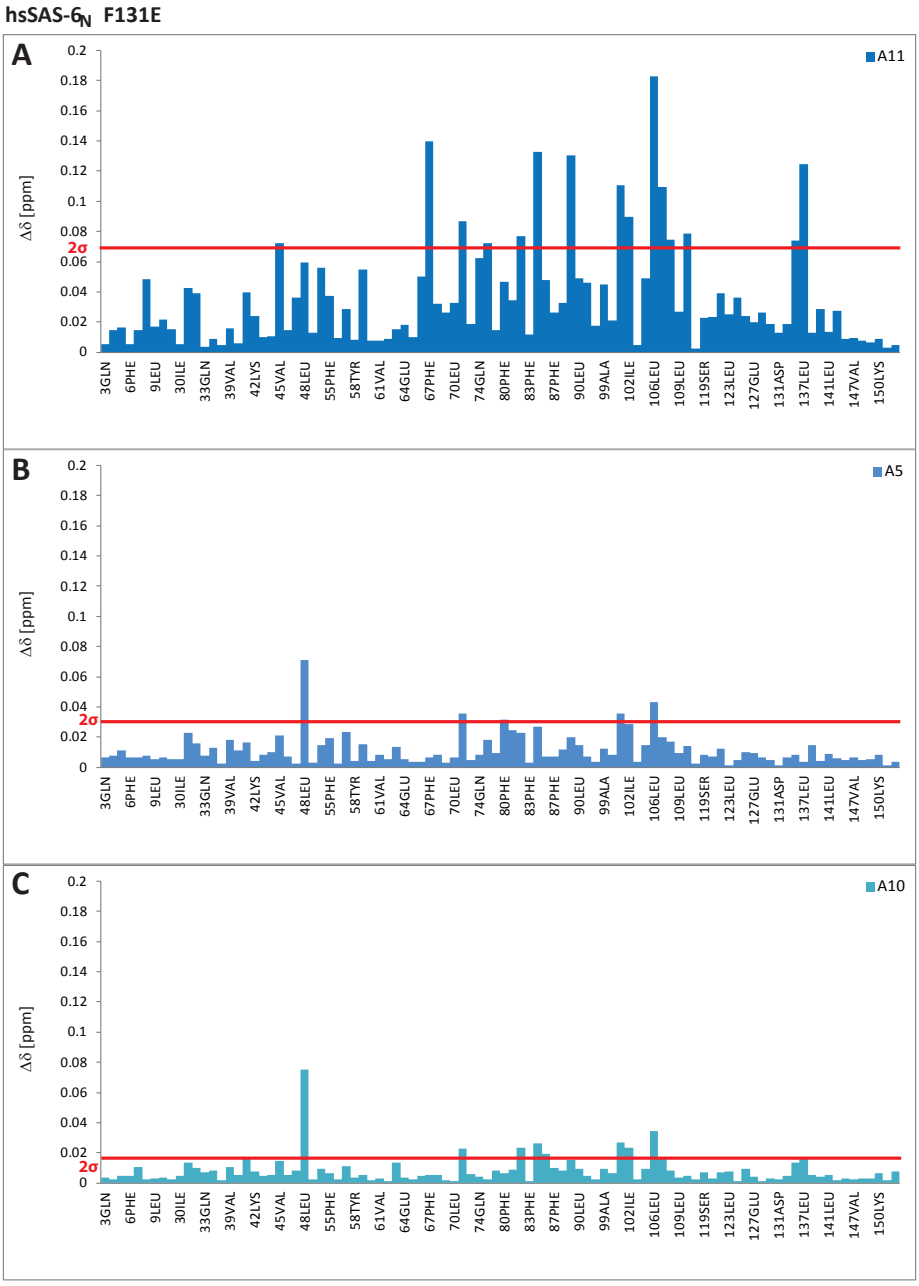


Fig. 7.10: Resonance Differences from NMR Screen with *H. sapiens* SAS-6$_N$ for Compounds A11, A5 and A10: (A) Resonance difference seen as effect of addition of 2 mM of compound A11 with $2\sigma = 0.07$ ppm; (B) addition of A5 with $2\sigma = 0.021$ ppm; (C) addition of A10 with $2\sigma = 0.019$ ppm.

In addition to the interactions mediated by the A11 sulfonamide group, the aromatic group associating with the hydrophobic target pocket consistently reached a

stable position in simulations. Based on these observations we would assume that a selection of compounds based on the binding potential of A11 would consist of compounds containing a core of an accessible sulfonamide with an aromatic ring (Figure 7.6 B). The core structure can be expanded by the addition of groups at position R1 to substitute the isopropanol group, R2 - substituting the ethylether group - and R3 - representing alterations of the aromatic residue and its substituents - to increase affinity and solubility. Alterations in position R1 and R2 are potentially achievable, since ether and sulfonamide groups offer options for synthesis reactions.

Given the vastness of chemical space, the ideal combination of chemical groups to occupy R1, R2 and R3 must theoretically exist. Structure based compound design on the basis of only one iteration of screening however is difficult, given that affinities are still difficult to predict. Nonetheless we can theorise based on chemical logic about the kind of substituents that could benefit interactions in the pocket.

To increase the affinity of a compound to the ceSAS-6 pocket, contacts with H81 could be utilised through addition of a group capable of forming π-π, cation-π or H-π interactions with the H81 residue. This could be aided by the inclusion of a linker at the R1 position [259, 260]. Another option for an R1 substitution would be a positively charged group, with a linker to interact with D82. Alternatively, a R1 substitution could also increase contacts to K158 if negatively charged. Adding charged groups could help with the overall polarity of the compound and therefore solubility. The ideal linker length and chemical group however, would have to be experimentally determined to find if interactions with H81, D82 or K158 lead to increased compound binding.

In the crSAS-6 pocket, G93 is positioned in place of an aspartate in the N-terminal position to the sulfonamide binding motif (L/I-T/L-V). Even though backbone interactions are possible with this residue, interactions with the Q92 side-chain might be

easier to target with the right substituent in the R1 position. Another group that could be used to design contacts would be H150. This could be reached by substituents in either the R1 or R2 positions.

The binding pocket of hsSAS-6 also offers the potential of altering A11 with a substituent to interact with the side-chains of glutamine residue Q74 in the N-terminal position to the sulfonamide binding motif and H136, which are both also present in crSAS-6.

Valuable lessons about the importance of the sulfonamide group and the possibility of substituting the aromatic ring with another ring group to occupy the hydrophobic pocket could be learned from the discovery and analysis of the A12 hsSAS-6 binding. However, there is less information about this compound currently, since only the hsSAS-6 N-terminal domain target pocket is available for study and no comparison of the binding behaviour in other pockets can be made. The factors that make A12 a selective binder to hsSAS-6 are more difficult to determine, because a comparison of A12 binding positions in other species is not possible.

As mentioned previously the likelihood of being able to design a high affinity compound on the information gathered so far is low due to difficulties in predicting the effects that alterations of the ligand might have on *in vitro* affinities. Furthermore chemical synthesis of custom compounds is a long and involved process. Realistically, custom compound synthesis would be considered at a later stage in compound development, after the SAR is better understood and higher affinity has been achieved, so that only a few select modifications are to be made to the compound.

To build an SAR profile, a compound screening selection with diverse substitutions at position R1, R2 and R3 needs to be created and screened for binding. Results from

such a screen would both serve to test the assumptions made about the binding pose based on the analysis of MD simulations as well as deliver information about the impact of the substituents and potentially reveal compounds with improved binding. A selection of compounds that could be used for such SAR screening can be proposed (Figure 7.11 and Figure 7.12). These compounds have both interesting substituents in R1, R2 and alteration of the aromatic R3, which would allow us to test our assumptions about the proposed binding positions and also amount to favourable $x\log P$ values. In addition, the suggested compound selection contains only commercially available compounds, making this a feasible next screening step.

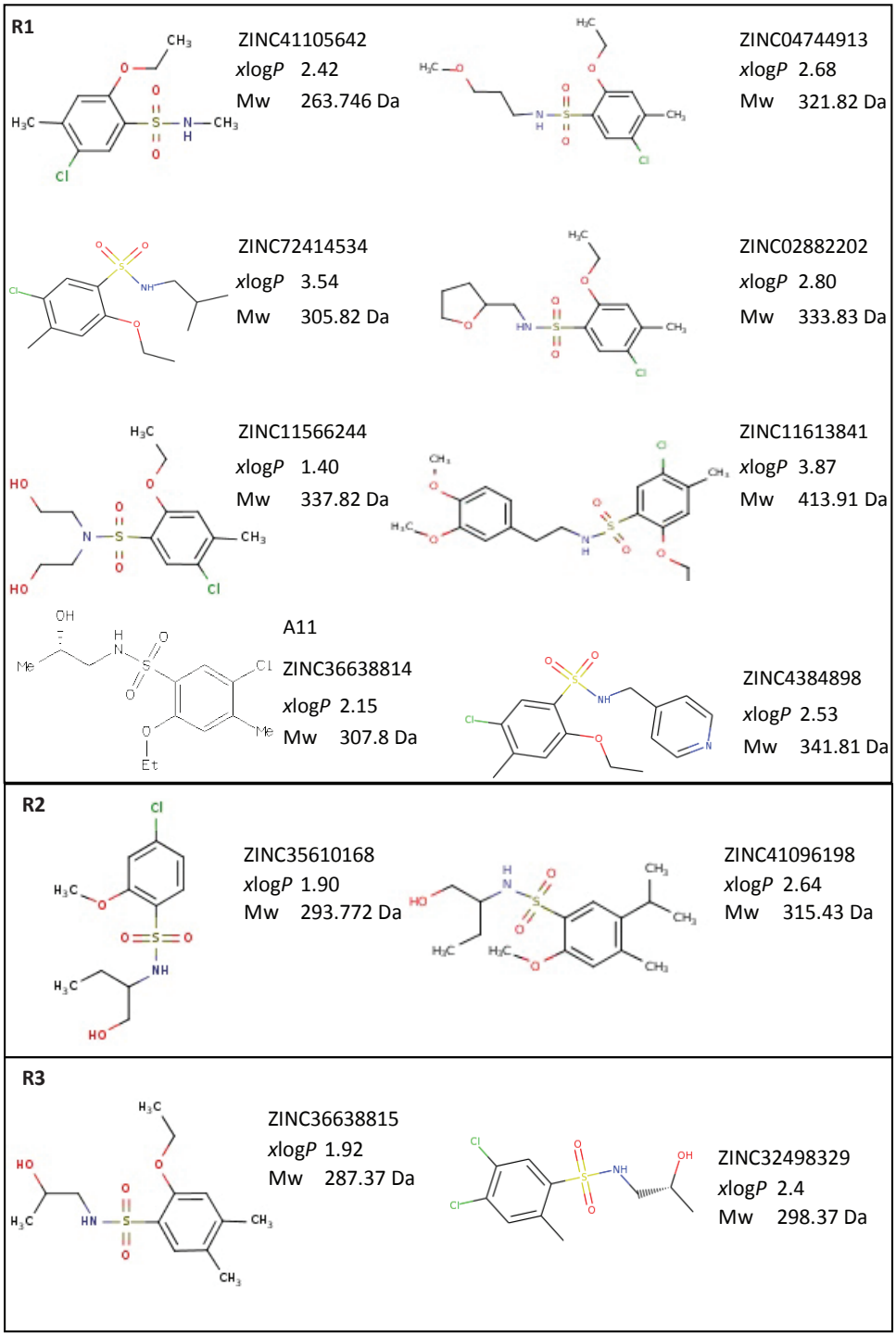


Fig. 7.11: Selection of Commercially Available Compounds for Next Generation Screening Based on A11: (R1) Substitutions of isopropanol group, (R2) substitutions in ethylether group, (R3) modifications of benzyl appendages.

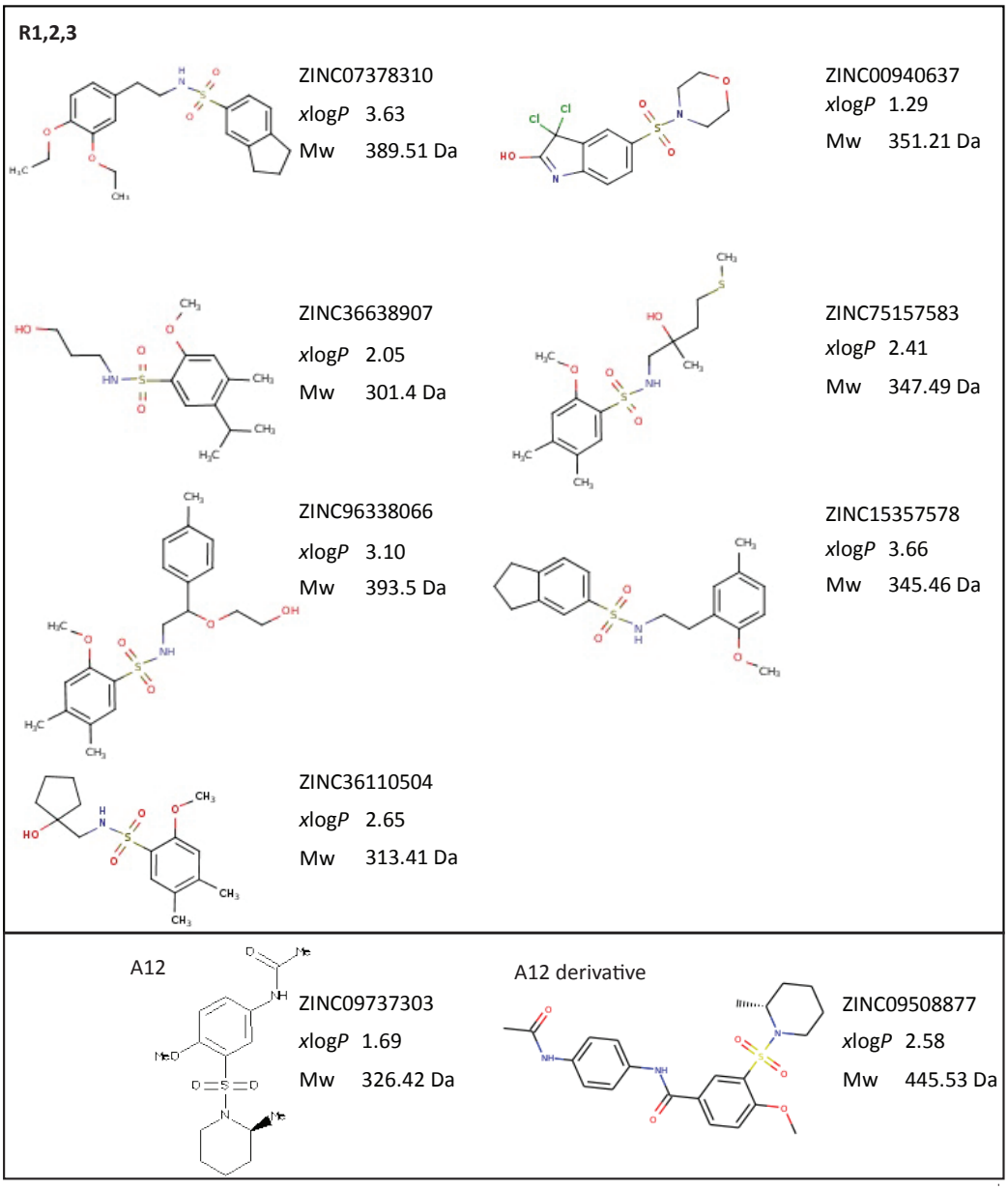


Fig. 7.12: Selection of Commercially Available Compounds for Next Generation Screening Based on A11 and A12: (R1,2,3) More complex modifications of core A11 structure involving regions R1, R2 and R3. (A12 derivative) Commercially available compound with A12 substructure.

7.3 Considerations on Future Use of SAS-6 N-terminal Dimerisation Inhibitors

Hit molecules targeting the SAS-6 N-terminal dimerisation for inhibition have been identified in this study. Further steps in lead compound design are necessary to develop the identified hit compounds further. Here we speculate about the possible uses of a finalised SAS-6 N-terminal domain dimerisation inhibitor.

7.3.1 N-terminal Domain Dimerisation Inhibitors as Research Tool

SAS-6 is assumed to be the main component of the cartwheel, which ensures the nine-fold symmetric formation of the microtubule assembly that is the basis of the centriole [90, 125–127, 261]. Since it has a central role in centriole duplication, several aspects of this fundamental cell cycle process could be examined given a high affinity and specific SAS-6 N-terminal domain dimerisation inhibitor.

Microscopic and tomographic imaging techniques, as well as crystallographic structures and computational models give us insight into the stacking of SAS-6 rings, which are believed to form the cartwheel [77, 81, 83, 91, 92, 94, 95, 158]. The inherent periodicity of the nine-fold spokes, the spacing of the stacks and the joining of SAS-6 spokes in the z-dimension give us a picture of how microtubule symmetry might be controlled by the scaffolding through the SAS-6 oligomer [90, 92, 94]. Even though the protein and its oligomerisation behaviour have been studied extensively in *in vitro* and *in vivo* experiments, open questions about the mechanisms of cartwheel formation and microtubule recruitment remain. SAS-6 oligomerisation inhibitors could be used to study the microtubule assembly in centriole duplication, without the drawbacks

of potentially altered expression levels and leaky expression that occur when genetic techniques are used.

Other aspects of centrosome duplication and cell cycle pathways which could be studied with a SAS-6 oligomerisation inhibitor would be phosphorylation states of proteins in the centriole duplication arrested state, which occurs when SAS-6 oligomers have been eliminated [127, 135]. The advantage of using a small molecule in this context would be that the timing of this blockade could be tightly controlled, which might avoid the alteration of expression levels of other proteins involved, including kinases. Moreover, insights into *de novo* centriole formation could be gained in experiments without altered protein levels.

Techniques conventionally used to study the roles of proteins in their pathways in the context of the cell, include knock-out or knock-down of the protein of interest, as well as the expression of mutant versions of the protein. When used in the context of tightly regulated cell cycle related pathways, these techniques come with a number of restrictions. Proteins essential for cell division such as SAS-6 cannot be knocked-out completely, as the phenotype would be lethal. Instead, a knock-down using, for example, siRNA can be performed [262, 263]. Unfortunately, this technique can lead to unpredictable background expression and is not available in some model organisms. The expression of mutated versions of the protein of interest can involve an alteration of expression levels, which might have influence on the function of the target protein and connected proteins upstream and downstream in the pathway.

The advantages of using a small molecule to inhibit SAS-6 function over modifications performed on a genetic level include control over the time point at which the effect is induced, control over the amount of compound added, and the lack of alteration in expression levels. In addition, a well designed small molecule compound could be used in model systems where the tool box for genetic manipulation is not

easily accessible.

7.3.2 N-terminal Domain Dimerisation Inhibitors as Drug Candidates

The pivotal role of the SAS-6 oligomerisation event in centrosome duplication links SAS-6 function to a number of disease relevant mechanisms. As discussed in the introductory section of this thesis, in the absence of SAS-6 oligomers, centriole duplication is aborted and cell division severely impaired. Inhibiting SAS-6 oligomerisation with an N-terminal domain dimerisation inhibitor could therefore be thought of as a potential way to fight cancer cell proliferation. The close connection of centriole duplication and cell cycle progression brought the centriole pathway controlling kinase Plk-4 into the focus of cancer drug research [153, 154].

The kinase is a more durggable target in the classical sense. It features an accessible activity pocket that can be targeted, and screening libraries and methods for inhibitor selection are tailored to kinase targets, as there is an abundance of experience in the field of drug discovery for kinase targets.

Plk-4 is involved in a number of processes, making side effects in addition to the inhibition of centriole formation highly likely. On the other hand, the inhibition of a purely structural mechanism like the scaffolding provided by the SAS-6 oligomer has the potential for fewer effects on non-targeted pathways. Expression of non-oligomerising SAS-6 (F131E) in human tissue culture showed that cells could still divide in absence of SAS-6 oligomers in a p53 knockout background [135]. However, the impairment of cell division might hypothetically be utilised to support already existing cancer therapy drugs. Thorough research would be needed to look into the potential of these thought experiments.

Another interesting avenue to explore would be the potential use of SAS-6 oligomerisation inhibitors as drugs targeting protozoa related diseases such as malaria and sleeping sickness. Recent studies on SAS-6 oligomerisation defects in *Plasmodium* and *Trypanosomes* showed effects on flagella function, which also lead to parasite cell death and diminished transmission rates [155, 156]. Developing compounds with high specificity to the SAS-6 N-terminal domains would be imperative to avoid side effects by blocking hsSAS-6 oligomerisation. Our discovery of A12 - a hsSAS-6 specific binding compound - shows, that despite the high structural similarity of SAS-6 across species, the design of species specific SAS-6 N-terminal domain dimerisation inhibitors might be possible.

Chapter 8

Conclusion

"Begin at the beginning and go on till you come to the end: then stop."

Lewis Carroll

Alice in Wonderland

This thesis investigated aspects of the making and breaking of the N-terminal domain dimerisation of SAS-6.

Insights into the molecular mechanisms involved in the dimerisation of this domain contribute to our understanding of SAS-6 oligomerisation, cartwheel formation and ultimately centriole duplication. The elongated loop α2-β5 unique to *C. elegans* SAS-6 was found to be necessary for the N-terminal domain dimerisation and directly involved in interactions within the dimer interface. A hypothesis about the mechanisms of the loop α2-β5 involvement was formed, which helps explain the influence of loop α2-β5. Furthermore, the first structure of the *H. sapiens* SAS-6 N-terminal domain adds to our knowledge about structural conservation of SAS-6 and delivers insights valuable for future research into centriole duplciation in *H. sapiens*.

The prospects of breaking SAS-6 dimerisation with the use of a small molecule were explored. The oligomerisation of SAS-6 is a potential target for inhibition from

both a scientific as well as a medical perspective. The first inhibitory small molecule compound for the *C. elegans* SAS-6 oligomerisation was identified. This compound is also the first compound to show binding to the N-terminal domain dimerisation interface of several organisms - *C. elegans*, *C. reinhardtii* and *H. sapiens*. The identification of another compound binding to only *H. sapiens* shows that the small differences between the structurally similar dimerisation interfaces can be used to obtain selective inhibitors.

This study, in conclusion, has furthered our understanding of a fundamental mechanism of centriole duplication and has contributed to the search for scientific tools and potential drug candidates targeting SAS-6.

References

[1] U. Scheer, "Historical roots of centrosome research: discovery of Boveri's microscope slides in Wurzburg," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130469–20130469, Aug. 2014.

[2] J. Azimzadeh and M. Bornens, *The centrosome in evolution. In Centrosomes in development and disease*, vol. pp. 93-122. Wiley-VCH, 2005.

[3] T. Boveri, "Zellen-studien: Ueber die Natur der Centrosomen," *Zellenstudien*, pp. 1–255, 1900.

[4] T. Boveri, "Concerning the origin of malignant tumours by Theodor Boveri. Translated and annotated by Henry Harris," *Journal of Cell Science*, vol. 121, no. Supplement 1, p. 1, 2007.

[5] Z. Carvalho-Santos, P. Machado, P. Branco, F. Tavares-Cadete, A. Rodrigues-Martins, J. B. Pereira-Leal, and M. Bettencourt-Dias, "Stepwise evolution of the centriole-assembly pathway," *Journal of Cell Science*, vol. 123, pp. 1414–1426, May 2010.

[6] M. E. Hodges, N. Scheumann, B. Wickstead, J. A. Langdale, and K. Gull, "Reconstructing the evolutionary history of the centriole from protein components," *Journal of Cell Science*, vol. 123, pp. 1407–1413, May 2010.

[7] J. Azimzadeh, "Exploring the evolutionary history of centrosomes," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130453–20130453, Aug. 2014.

[8] C. Katstros, D. Karyophyllis, and B. Galatis, "Cytoskeleton and morphogenesis in brown algae," *Annals of Botany*, vol. 97, pp. 679–693, 2006.

[9] J. V. Kilmartin, "Lessons from yeast: the spindle pole body and the centrosome," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1650, pp. 1–7, 2014.

[10] C. Daunderer, M. Schliwa, and R. Gräf, "Dictyostelium discoideum: A promising centrosome model system," *Biology of the Cell*, vol. 91, pp. 313 – 320, 1999.

[11] A. De Martino, A. Amato, and C. Bowler, "Mitosis in diatoms: rediscovering an old model for cell division," *BioEssays*, vol. 31, pp. 874–884, 2009.

[12] S. K. Dutcher and E. T. O'Toole, "The basal bodies of Chlamydomonas reinhardtii," *Cilia*, vol. 5, no. 1, p. 18, 2016.

[13] I. Schulz, A. Erle, R. Gräf, A. Krüger, H. Lohmeier, S. Putzler, M. Samereier, and S. Weidenthaler, "Identification and cell cycle-dependent localization of nine novel, genuine centrosomal components in Dictyostelium discoideum," *Cell Motility and the Cytoskeleton*, vol. 66, 2009.

[14] W. F. Marshall, "Centriole evolution," *Current Opinion in Cell Biology*, vol. 21, pp. 14–19, 2009.

[15] Z. Carvalho-Santos, J. Azimzadeh, J. B. Pereira-Leal, and M. Bettencourt-Dias, "Evolution: Tracing the origins of centrioles, cilia, and flagella.," *The Journal of Cell Biology*, vol. 194, pp. 165–175, Aug. 2011.

[16] J. Azimzadeh and W. F. Marshall, "Building the centriole.," *Current Biology*, vol. 20, pp. R816–25, Oct. 2010.

[17] M. Bornens, "Centrosome composition and microtubule anchoring mechanisms.," *Current Opinion in Cell Biology*, vol. 14, pp. 25–34, Mar. 2002.

[18] M. Bornens, "The centrosome in cells and organisms.," *Science*, vol. 335, pp. 422–426, Feb. 2012.

[19] G. Sluder, "One to only two: a short history of the centrosome and its duplication," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130455–20130455, Aug. 2014.

[20] C. Arquint, A. M. Gabryjonczyk, and E. A. Nigg, "Centrosomes as signalling centres," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130464–20130464, Aug. 2014.

[21] J. Reina and C. Gonzalez, "When fate follows age: unequal centrosomes in asymmetric cell division," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130466–20130466, Aug. 2014.

[22] P. Gönczy, "Mechanisms of asymmetric cell division: flies and worms pave the way," *Nature Reviews Molecular Cell Biology*, vol. 9, pp. 355–366, June 2008.

[23] A. Noatynska, N. Tavernier, M. Gotta, and L. Pintard, "Coordinating cell polarity and cell cycle progression: what can we learn from flies and worms?," *Open Biology*, vol. 3, p. e130083, Sept. 2013.

[24] L. Fletcher and R. J. Muschel, "The centrosome and the DNA damage induced checkpoint.," *Cancer Letters*, vol. 243, pp. 1–8, Dec. 2006.

[25] J. M. Kollman, A. Merdes, L. Mourey, and D. A. Agard, "Microtubule nucleation by gamma-tubulin complexes," *Nature Reviews Molecular Cell Biology*, vol. 12, pp. 709–721, Nov. 2011.

[26] J. B. Woodruff, O. Wueseke, and A. A. Hyman, "Pericentriolar material structure and dynamics," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130459–20130459, Aug. 2014.

[27] C. B. O'Connell and A. L. Khodjakov, "Cooperative mechanisms of mitotic spindle formation," *Journal of Cell Science*, vol. 120, pp. 1717–1722, June 2007.

[28] C. Klotz, M. C. Dabauvalle, M. Paintrand, T. Weber, M. Bornens, and E. Karsenti, "Parthenogenesis in Xenopus eggs requires centrosomal integrity.," *The Journal of Cell Biology*, vol. 110, pp. 405–415, Mar. 1990.

[29] M. Kirkham, T. Müller-Reichert, K. Oegema, S. Grill, and A. A. Hyman, "SAS-4 is a c. elegans centriolar protein that controls centrosome size.," *Cell*, vol. 112, pp. 575–87, Mar. 2003.

[30] N. R. Stevens, A. A. A. S. F. Raposo, R. Basto, D. S. Johnston, and J. Raff, "From stem cell to embryo without centrioles," *Current Biology*, pp. 1–6, Oct. 2007.

[31] A. Rodrigues-Martins, M. Riparbelli, G. Callaini, D. M. Glover, and M. Bettencourt-Dias, "From centriole biogenesis to cellular function: centrioles are essential for cell division at critical developmental stages.," *Cell Cycle*, vol. 7, pp. 11–6, Feb. 2008.

[32] A. Khodjakov, R. W. Cole, B. R. Oakley, and C. L. Rieder, "Centrosome-independent mitotic spindle formation in vertebrates.," *Current Biology*, vol. 10, pp. 59–67, Feb. 2000.

[33] T. Ohba, M. Nakamura, H. Nishitani, and T. Nishimoto, "Self-organization of microtubule asters induced in xenopus egg extracts by GTP-bound Ran," *Science*, vol. 284, no. 5418, pp. 1356–1358, 1999.

[34] P. Kalab, K. Weis, and R. Heald, "Visualization of a Ran-GTP gradient in interphase and mitotic xenopus egg extracts," *Science*, vol. 295, no. 5564, pp. 2452–2456, 2002.

[35] R. E. Carazo-Salas, O. J. Gruss, I. W. Mattaj, and E. Karsenti, "Ran-GTP coordinates regulation of microtubule nucleation and dynamics during mitotic-spindle assembly," *Nature Cell Biology*, vol. 3, pp. 228–234, 2001.

[36] P. Kalab and R. Heald, "The RanGTP gradient - a GPS for the mitotic spindle," *Journal of Cell Science*, vol. 121, pp. 1577–1586, June 2008.

[37] S. L. Prosser and L. Pelletier, "Mitotic spindle assembly in animal cells: a fine balancing act," *Nature Reviews Molecular Cell Biology*, vol. 18, pp. 187–201, Feb. 2017.

[38] A. J. Holl, A. J. Holland, D. W. Clevel, and D. W. Cleveland, "Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis," *Nature Reviews Molecular Cell Biology*, vol. 10, pp. 478–487, Aug. 2009.

[39] D. L. Ringo, "Flagellar motion and fine structure of the flagellar apparatus in chlamydomonas.," *The Journal of Cell Biology*, vol. 33, pp. 543–71, July 1967.

[40] P. Guichard and P. Gönczy, "Basal body structure in Trichonympha," *Cilia*, vol. 5, no. 1, p. 9, 2016.

[41] T. D. Loreng and E. F. Smith, "The central apparatus of cilia and eukaryotic flagella.," *Cold Spring Harbor Perspectives in Biology*, vol. 9, p. a028118, Mar. 2017.

[42] H. Lodish, A. Berk, S. L. Zipursky, P. Matsudaira, and a. J. D. David Baltimore, *Molecular Cell Biology*, vol. Section 19.4, Cilia and Flagella: Structure and Movement. New York: W. H. Freeman, 4th ed., 2000.

[43] S. Nonaka, Y. Tanaka, Y. Okada, S. Takeda, A. Harada, Y. Kanai, M. Kido, and N. Hirokawa, "Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein.," *Cell*, vol. 95, pp. 829–837, Dec. 1998.

[44] E. Lauga, E. Lauga, T. R. Powers, and T. R. Powers, "The hydrodynamics of swimming microorganisms," *Reports on Progress in Physics*, vol. 72, p. 096601, Oct. 2009.

[45] H. Praetorius and K. R. Spring, "Bending the MDCK cell primary cilium increases intracellular calcium," *The Journal of Membrane Biology*, vol. 184, pp. 71–9, Nov. 2001.

[46] Z. Li, S. Wakao, B. B. Fischer, and K. K. Niyogi, "Sensing and responding to excess light," *Annual Review of Plant Biology*, vol. 60, pp. 239–260, July 2009.

[47] M. V. Nachury, "How do cilia organize signalling cascades?," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130465–20130465, Aug. 2014.

[48] S. C. Goetz and K. V. Anderson, "The primary cilium: a signalling centre during vertebrate development," *Nature Reviews Genetics*, vol. 11, pp. 331–344, June 2010.

[49] J. Lüders and T. Stearns, "Microtubule-organizing centres: a re-evaluation.," *Nature Reviews. Molecular Cell Biology*, vol. 8, pp. 161–7, Mar. 2007.

[50] E. A. Nigg and J. W. Raff, "Centrioles, centrosomes, and cilia in health and disease.," *Cell*, vol. 139, pp. 663–78, Dec. 2009.

[51] S. A. Godinho and D. Pellman, "Causes and consequences of centrosome abnormalities in cancer," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130467–20130467, Aug. 2014.

[52] E. Nigg, "Centrosome aberrations: Cause or consequence of cancer progression?," *Nature Reviews Genetics*, vol. 3, pp. 815–815, Nov. 2002.

[53] D. Zyss and F. Gergely, "Centrosome function in cancer: guilty or innocent?," *Trends in Cell Biology*, vol. 19, pp. 334–46, Aug. 2009.

[54] J. Y. Chan, "A clinical overview of centrosome amplification in human cancers.," *International Journal of Biological Sciences*, vol. 7, pp. 1122–1144, Feb. 2011.

[55] G. A. Pihan, J. Wallace, Y. Zhou, and S. J. Doxsey, "Centrosome abnormalities and chromosome instability occur together in pre-invasive carcinomas," *Cancer Research*, vol. 63, pp. 1398–1404, Mar. 2003.

[56] J. Y. Chan, "A clinical overview of centrosome amplification in human cancers.," *International Journal of Biological Sciences*, vol. 7, pp. 1122–1144, 2011.

[57] E. A. Nigg, "Origins and consequences of centrosome aberrations in human cancers.," *International Journal of Cancer*, vol. 119, no. 12, pp. 2717–23, 2006.

[58] P. Gönczy, "Centrosomes and cancer: revisiting a long-standing relationship.," *Nature Review*, vol. 15, pp. 639–652, Nov. 2015.

[59] K. F. O'Connell, C. Caron, K. R. Kopish, D. D. Hurd, K. J. Kemphues, Y. Li, and J. G. White, "The C. elegans zyg-1 gene encodes a regulator of centrosome duplication with distinct maternal and paternal roles in the embryo.," *Cell*, vol. 105, pp. 547–558, June 2001.

[60] M. M. Lettman, Y. L. Wong, V. Viscardi, S. Niessen, S.-H. Chen, A. K. Shiau, H. Zhou, A. Desai, and K. Oegema, "Direct binding of SAS-6 to ZYG-1 recruits SAS-6 to the mother centriole for cartwheel assembly.," *Developmental Cell*, vol. 25, pp. 284–298, June 2013.

[61] M. Bettencourt-Dias, A. Rodrigues-Martins, L. Carpenter, M. Riparbelli, L. Lehmann, M. K. Gatt, N. Carmo, F. Balloux, G. Callaini, and D. M. Glover, "SAK/PLK4 is required for centriole duplication and flagella development.," *Current Biology*, vol. 15, no. 24, pp. 2199–2207, 2005.

[62] N. Dzhindzhev, G. Tzolovsky, Z. Lipinszki, S. Schneider, R. Lattao, J. Fu, J. Debski, M. Dadlez, and D. Glover, "Plk4 phosphorylates Ana2 to trigger Sas6 recruitment and procentriole formation," *Current Biology*, vol. 24, pp. 2526–2532, Nov. 2014.

[63] M. Kim, C. Fong, and M.-F. B. Tsou, "Centriole duplication: When PLK4 meets Ana2/STIL," *Current Biology*, vol. 24, pp. R1046 – R1048, Dec. 2014.

[64] T. Nakamura, H. Saito, and M. Takekawa, "SAPK pathways and p53 cooperatively regulate PLK4 activity and centrosome integrity under stress," *Nature Communications*, vol. 4, p. 1775, May 2013.

[65] Ö. Serçin, J.-C. Larsimont, A. E. Karambelas, V. Marthiens, V. Moers, B. Boeckx, M. L. L. Mercier, D. Lambrechts, R. Basto, M. L. L. Mercier, and C. Blanpain, "Transient PLK4 overexpression accelerates tumorigenesis in p53-deficient epidermis," *Nature Cell Biology*, vol. 18, pp. 100–110, Dec. 2015.

[66] P. A. Futreal, L. Coin, M. Marshall, T. Down, T. Hubbard, R. Wooster, N. Rahman, and M. R. Stratton, "A census of human cancer genes," *Nature Reviews Cancer*, vol. 4, pp. 177–183, Feb. 2004.

[67] M. S. Lawrence, P. Stojanov, C. H. Mermel, L. A. Garraway, T. R. Golub, M. Meyerson, S. B. Gabriel, E. S. Lander, and G. Getz, "Discovery and saturation analysis of cancer genes across 21 tumor types.," *Nature*, vol. 505, pp. 495–501, Jan. 2014.

[68] T. I. Zack, S. E. Schumacher, S. L. Carter, A. D. Cherniack, G. Saksena, B. Tabak, M. S. Lawrence, C.-Z. Zhang, J. Wala, C. H. Mermel, C. Sougnez, S. B. Gabriel, B. Hernandez, H. Shen, P. W. Laird, G. Getz, M. Meyerson, and R. Beroukhim, "Pan-cancer patterns of somatic copy-number alteration," *Nature Genetics*, vol. 45, pp. 1134–1140, Feb. 2013.

[69] W. Y. Tsang, A. Spektor, S. Vijayakumar, B. R. Bista, J. L. andf Irma Sanchez, S. Duensing, , and B. D. Dynlacht, "Cep76, a centrosomal protein that specifically restrains centriole re-duplication," *Developmental Cell*, vol. 16, pp. 649–660, June 2009.

[70] M. Bettencourt-Dias, F. Hildebrandt, D. Pellman, G. Woods, and S. A. Godinho, "Centrosomes and cilia in human disease.," *Trends in Genetics*, vol. 27, pp. 307–315, Sept. 2011.

[71] K. Mikule, B. Delaval, P. Kaldis, A. Jurcyzk, P. Hergert, and S. Doxsey, "Loss of centrosome integrity induces p38-p53-p21-dependent G1-S arrest," *Nature Cell Biology*, vol. 9, pp. 160–170, Feb. 2007.

[72] Y. Uetake, J. Lončarek, J. J. Nordberg, C. N. English, S. La Terra, A. Khodjakov, and G. Sluder, "Cell cycle progression and de novo centriole assembly after centrosomal removal in untransformed human cells," *The Journal of Cell Biology*, vol. 176, pp. 173–182, Jan. 2007.

[73] T. Zarubin and J. Han, "Activation and signaling of the p38 MAP kinase pathway," *Cell Research*, vol. 15, pp. 11–18, Jan. 2005.

[74] M. Fliegauf, T. Benzing, and H. Omran, "When cilia go bad: cilia defects and ciliopathies," *Nature Reviews Molecular Cell Biology*, vol. 8, pp. 880–893, Dec. 2007.

[75] P. L. Chavali, M. Pütz, and F. Gergely, "Small organelle, big responsibility: the role of centrosomes in development and disease," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 369, pp. 1–12, July 2014.

[76] A. M. Waters and P. L. Beales, "Ciliopathies: An expanding disease spectrum.," *Pedriatric Nephrology*, vol. 26, pp. 1039–1056, July 2011.

[77] M. Winey and E. O'Toole, "Centriole structure," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130457–20130457, Aug. 2014.

[78] P. T. Conduit, A. Wainman, and J. W. Raff, "Centrosome function and assembly in animal cells," *Nature Reviews Molecular Cell Biology*, vol. 16, pp. 611–625, Sept. 2015.

[79] J. Azimzadeh and M. Bornens, "Structure and duplication of the centrosome," *Journal of Cell Science*, vol. 120, no. 13, pp. 2139–2142, 2007.

[80] M. Paintrand, M. Moudjou, H. Delacroix, and M. Bornens, "Centrosome organization and centriole architecture: Their sensitivity to divalent cations," *Journal of Structural Biology*, vol. 108, pp. 107–128, 1992.

[81] P. Guichard, V. Hachet, N. Majubu, A. Neves, D. Demurtas, N. Olieric, I. Fluckiger, A. Yamada, K. Kihara, Y. Nishida, S. Moriya, M. Steinmetz, Y. Hongoh, and P. Gönczy, "Native architecture of the centriole proximal region reveals features underlying its 9-fold radial symmetry," *Current Biology*, vol. 23, pp. 1620–1628, Sept. 2013.

[82] S. Li, J.-J. Fernandez, W. F. Marshall, and D. A. Agard, "Three-dimensional structure of basal body triplet revealed by electron cryo-tomography," *The EMBO Journal*, vol. 31, pp. 552–562, 2011.

[83] P. Guichard, "Cartwheel architecture of Trichonympha basal body," *Science*, vol. 337, pp. 553–553, Sept. 2012.

[84] M. Hirono, "Cartwheel assembly," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, pp. 20130458–20130458, Aug. 2014.

[85] L. Pelletier, E. O'Toole, A. Schwager, A. A. Hyman, and T. Müller-Reichert, "Centriole assembly in Caenorhabditis elegans," *Nature*, vol. 444, pp. 619–623, Dec. 2006.

[86] S. Geimer and M. Melkonian, "The ultrastructure of the Chlamydomonas reinhardtii basal apparatus: identification of an early marker of radial asymmetry inherent in the basal body," *Journal of Cell Science*, vol. 117, pp. 2663–2674, 2004.

[87] I. R. Gibbons and A. V. Grimstone, "On flagellar structure in certain flagellates," *Journal of Biophysical and Biochemical Cytology*, vol. 7, pp. 697–716, July 1960.

[88] P. Guichard, V. Hamel, A. Neves, and P. Gönczy, *Isolation, cryotomography, and three-dimensional reconstruction of centrioles*, vol. 129 of *Methods in Cell Biology*, ch. 11, pp. 191 – 209. Academic Press, 2015.

[89] M. Hiraki, Y. Nakazawa, R. Kamiya, and M. Hirono, "Bld10p constitutes the cartwheel-spoke tip and stabilizes the 9-fold symmetry of the centriole.," *Current Biology*, vol. 17, pp. 1778–83, Nov. 2007.

[90] Y. Nakazawa, M. Hiraki, R. Kamiya, and M. Hirono, "SAS-6 is a cartwheel protein that establishes the 9-fold symmetry of the centriole.," *Current Biology*, vol. 17, no. 24, pp. 2169–74, 2007.

[91] M. Jerka-Dziadosz, D. Gogendeau, C. Klotz, J. Cohen, J. Beisson, and F. Koll, "Basal body duplication in Paramecium: the key role of Bld10 in assembly and stability of the cartwheel.," *Cytoskeleton*, vol. 67, pp. 161–71, Apr. 2010.

[92] A. Schuldt, "Cytoskeleton: SAS-6 turns a cartwheel trick," *Nature Reviews Molecular Cell Biology*, vol. 12, pp. 137–137, Apr. 2011.

[93] D. Kitagawa, I. Vakonakis, N. Olieric, M. Hilbert, D. Keller, V. Olieric, M. Bortfeld, M. C. Erat, I. Flückiger, P. Gönczy, and M. O. Steinmetz, "Structural basis of the 9-fold symmetry of centrioles," *Cell*, vol. 144, pp. 364–375, Mar. 2011.

[94] M. van Breugel, M. Hirono, A. Andreeva, H. aki Yanagisawa, S. Yamaguchi, Y. Nakazawa, N. Morgner, M. Petrovich, I.-O. Ebong, C. V. Robinson, C. M. Johnson, D. Veprintsev, and B. Zuber, "Structures of SAS-6 suggest its organization in centrioles.," *Science*, vol. 331, pp. 1196–1199, Apr. 2011.

[95] M. van Breugel, R. Wilcken, S. H. McLaughlin, T. J. Rutherford, and C. M. Johnson, "Structure of the SAS-6 cartwheel hub from," *eLife*, vol. 3, p. e01812, Jan. 2014.

[96] V. Hamel, M. L. L. Guennec, N. Banterle, I. Iacovache, V. Nem&ccaron, V. Nem, &iacute, ccaron, kov&aacute, I. Fl&uuml, iacute, kov, aacute, I. Fl, D. L&eacute, uuml, ckiger, K. N. Goldie, H. Stahlberg, P. G&ouml, D. L, eacute, vy, B. Zuber, P. Guichard, P. G, ouml, nczy, V. Nemčíková, I. Flückiger, D. Lévy, and P. Gönczy, "Cell-free reconstitution reveals centriole cartwheel assembly mechanisms," *Nature Communications*, vol. 8, p. 14813, Apr. 2017.

[97] A. Dammermann, P. S. Maddox, A. Desai, and K. Oegema, "SAS-4 is recruited to a dynamic structure in newly forming centrioles that is stabilized by the gamma-tubulin–mediated addition of centriolar microtubules," *The Journal of Cell Biology*, vol. 180, pp. 771–785, Mar. 2008.

[98] Y.-C. Lin, C.-W. Chang, W.-B. Hsu, C.-J. C. Tang, Y.-N. Lin, E.-J. Chou, C.-T. Wu, and K. T. Tang, "Human microcephaly protein CEP135 binds to hSAS-6 and CPAP, and is required for centriole assembly," *The EMBO Journal*, vol. 32, pp. 1141–1154, Apr. 2013.

[99] A. D. Dingle and C. Fulton, "Development of the flagellar apparatus of Naegleria," *The Journal of Cell Biology*, vol. 31, pp. 43–54, 1966.

[100] E. Stubblefield and B. R. Brinkley, *Formation and fate of cell organelles*, vol. 1, ch. Architecture and function of the mammalian centriole., pp. 175–218. Academic Press, 1967.

[101] P. Strnad, S. Leidel, T. Vinogradova, U. Euteneuer, A. Khodjakov, and P. Gönczy, "Regulated HsSAS-6 levels ensure formation of a single procentriole per centriole during the centrosome duplication cycle.," *Developmental Cell*, vol. 13, pp. 203–13, Sept. 2007.

[102] M. Delattre, C. Canard, and P. Gönczy, "Sequential protein recruitment in C. elegans centriole formation.," *Current Biology*, vol. 16, pp. 1844–9, Oct. 2006.

[103] M. Hilbert, M. C. Erat, V. Hachet, P. Guichard, I. D. Blank, I. Fluckiger, L. Slater, E. D. Lowe, G. N. Hatzopoulos, M. O. Steinmetz, P. Gonczy, and I. Vakonakis, "Caenorhabditis elegans centriolar protein SAS-6 forms a spiral

that is consistent with imparting a ninefold symmetry," *Proceedings of the National Academy of Sciences*, vol. 110, pp. 11373–11378, Aug. 2013.

[104] D. A. Brito, S. M. Gouveia, and M. Bettencourt-Dias, "Deconstructing the centriole: structure and number control," *Current Opinion in Cell Biology*, vol. 24, pp. 4–13, Mar. 2012.

[105] P. Strnad and P. Gönczy, "Mechanisms of procentriole formation.," *Trends in Cell Biology*, vol. 18, pp. 389–396, Sept. 2008.

[106] G. Goshima, R. Wollman, S. S. Goodwin, N. Zhang, J. M. Scholey, R. D. Vale, and N. Stuurman, "Genes required for mitotic spindle assembly in Drosophila S2 cells," *Science*, vol. 316, pp. 417–421, May 2007.

[107] J. Dobbelaere, F. Josué, S. Suijkerbuijk, B. Baum, N. Tapon, and J. Raff, "A genome-wide RNAi screen to dissect centriole duplication and centrosome maturation in drosophila.," *PLoS Biology*, vol. 6, p. e224, Oct. 2008.

[108] F. Balestra, P. Strnad, I. Flückiger, and P. Gönczy, "Discovering regulators of centriole biogenesis through siRNA-based functional genomics in human cells," *Developmental Cell*, vol. 25, pp. 555–571, June 2013.

[109] M.-F. B. Tsou, W.-J. Wang, K. A. George, K. Uryu, T. Stearns, and P. V. Jallepalli, "Polo kinase and separase regulate the mitotic licensing of centriole duplication in human cells," *Developmental Cell*, vol. 17, pp. 344–354, Sept. 2009.

[110] G. Cabral, S. S. Sans, C. Cowan, and A. Dammermann, "Multiple mechanisms contribute to centriole separation in C. elegans," *Current Biology*, vol. 23, pp. 1380–1387, July 2013.

[111] P. Gönczy, C. Echeverri, K. Oegema, A. Coulson, S. J. M. Jones, R. R. Copley, J. Duperon, J. Oegema, M. Brehm, E. Cassin, E. Hannak, M. Kirkham, S. Pichler, K. Flohrs, A. Goessen, S. Leidel, A.-M. Alleaume, C. Martin, N. Özlü, P. Bork, and A. A. Hyman, "Functional genomic analysis of cell division in C. elegans using RNAi of genes on chromosome III," *Nature*, vol. 408, pp. 331–336, Nov. 2000.

[112] K. F. O'Connell, K. N. Maxwell, and J. G. White, "The spd-2 gene is required for polarization of the anteroposterior axis and formation of the sperm asters in the Caenorhabditis elegans zygote.," *Developmental Biology*, vol. 222, pp. 55–70, July 2000.

[113] C. A. Kemp, K. R. Kopish, P. Zipperlen, J. Ahringer, and K. F. O'Connell, "Centrosome maturation and duplication in C. elegans require the coiled-coil protein SPD-2.," *Developmental Cell*, vol. 6, pp. 511–23, May 2004.

[114] L. Pelletier, N. Ozlü, E. Hannak, C. Cowan, B. Habermann, M. Ruer, T. Müller-Reichert, and A. A. Hyman, "The caenorhabditis elegans centrosomal protein SPD-2 is required for both pericentriolar material recruitment and centriole duplication.," *Current Biology*, vol. 14, pp. 863–73, June 2004.

[115] K. F. Sonnen, A. M. Gabryjonczyk, E. Anselm, Y. D. Stierhof, and E. A. Nigg, "Human Cep192 and Cep152 cooperate in Plk4 recruitment and centriole duplication," *Journal of Cell Science*, vol. 126, pp. 3223–3233, Aug. 2013.

[116] E. Firat-Karalar, N. Rauniyar, J. Yates, and T. Stearns, "Proximity interactions among centrosome components identify regulators of centriole duplication," *Current Biology*, vol. 24, pp. 664–670, Apr. 2014.

[117] P. T. Conduit, J. H. Richens, A. Wainman, J. Holder, C. C. Vicente, M. B. Pratt, C. I. Dix, Z. A. Novak, I. M. Dobbie, L. Schermelleh, and J. W. Raff, "A molecular mechanism of mitotic centrosome assembly in *Drosophila*," *eLife*, vol. 3, p. e03399, Aug. 2014.

[118] M. Delattre, S. Leidel, K. Wani, K. Baumer, J. Bamat, H. Schnabel, R. Feichtinger, R. Schnabel, and P. Gönczy, "Centriolar SAS-5 is required for centrosome duplication in C. elegans," *Nature Cell Biology*, vol. 6, pp. 656–664, Aug. 2004.

[119] J. Vulprecht, A. David, A. Tibelius, A. Castiel, G. Konotop, F. Liu, F. Bestvater, M. S. Raab, H. Zentgraf, S. Izraeli, and A. Kramer, "STIL is required for centriole duplication in human cells," *Journal of Cell Science*, vol. 125, pp. 1353–1362, May 2012.

[120] N. R. Stevens, J. Dobbelaere, K. Brunk, A. Franz, and J. W. Raff, "Ana2 is a conserved centriole duplication factor," *The Journal of Cell Biology*, vol. 188, pp. 313–323, Mar. 2010.

[121] M. Ohta, T. Ashikawa, Y. Nozaki, H. Kozuka-Hata, H. Goto, M. Inagaki, M. Oyama, and D. Kitagawa, "Direct interaction of Plk4 with STIL ensures formation of a single procentriole per parental centriole," *Nature Communications*, vol. 5, p. 5267, Nov. 2014.

[122] T. C. Moyer, K. M. Clutario, B. G. Lambrus, V. Daggubati, and A. J. Holland, "Binding of STIL to Plk4 activates kinase activity to promote centriole assembly," *The Journal of Cell Biology*, vol. 209, pp. 863–878, July 2015.

[123] C. Arquint, A.-M. Gabryjonczyk, S. Imseng, R. Böhm, E. Sauer, S. Hiller, E. A. Nigg, and T. Maier, "STIL binding to polo-box 3 of PLK4 regulates centriole duplication.," *eLife*, vol. 4, p. e07888, Aug. 2015.

[124] A.-S. Kratz, F. Bärenz, K. T. Richter, and I. Hoffmann, "Plk4-dependent phosphorylation of STIL is required for centriole duplication.," *Biology Open*, vol. 4, pp. 370–377, Mar. 2015.

[125] A. Dammermann, T. Müller-Reichert, L. Pelletier, B. Habermann, A. Desai, and K. Oegema, "Centriole assembly requires both centriolar and pericentriolar material proteins.," *Developmental Cell*, vol. 7, no. 6, pp. 815–29, 2004.

[126] S. Leidel, M. Delattre, L. Cerutti, K. Baumer, and P. Gönczy, "SAS-6 defines a protein family required for centrosome duplication in C. elegans and in human cells.," *Nature Cell Biology*, vol. 7, pp. 115–25, Mar. 2005.

[127] A. Rodrigues-Martins, M. Bettencourt-Dias, M. Riparbelli, C. Ferreira, I. Ferreira, G. Callaini, and D. M. Glover, "DSAS-6 organizes a tube-like centriole precursor, and its absence suggests modularity in centriole assembly.," *Current Biology*, vol. 17, pp. 1465–1472, Oct. 2007.

[128] S. Leidel and P. Gönczy, "SAS-4 is essential for centrosome duplication in C. elegans and is recruited to daughter centrioles once per cell cycle.," *Developmental Cell*, vol. 4, pp. 431–9, Apr. 2003.

[129] L. Y. Hung, C. J. Tang, and T. K. Tang, "Protein 4.1 R-135 interacts with a novel centrosomal protein (CPAP) which is associated with the gamma-tubulin complex.," *Molecular and Cellular Biology*, vol. 20, pp. 7813–7825, Nov. 2000.

[130] K. Sugioka, D. R. Hamill, J. B. Lowry, M. E. McNeely, M. Enrick, A. C. Richter, L. E. Kiebler, J. R. Priess, and B. Bowerman, "Centriolar SAS-7 acts upstream of SPD-2 to regulate centriole assembly and pericentriolar material formation.," *eLife*, vol. 6, p. e20353, Feb. 2017.

[131] M. Ohta, A. Desai, and K. Oegema, "How centrioles acquire the ability to reproduce.," *eLife*, vol. 6, p. e25358, Apr. 2017.

[132] W. F. Marshall, Y. Vucica, and J. L. Rosenbaum, "Kinetics and regulation of de novo centriole assembly. implications for the mechanism of centriole duplication.," *Current Biology*, vol. 11, pp. 308–317, Apr. 2001.

[133] A. Khodjakov, C. L. Rieder, G. Sluder, G. Cassels, O. Sibon, and C.-L. Wang, "De novo formation of centrosomes in vertebrate cells arrested during S phase," *The Journal of Cell Biology*, vol. 158, pp. 1171–1181, Oct. 2002.

[134] S. L. Terra, C. N. English, P. Hergert, B. F. McEwen, G. Sluder, and A. Khodjakov, "The de novo centriole assembly pathway in HeLa cells," *The Journal of Cell Biology*, vol. 168, pp. 713–722, Mar. 2005.

[135] W.-J. Wang, D. Acehan, C.-H. Kao, W.-N. Jane, K. Uryu, and M.-F. B. Tsou, "De novo centriole formation in human cells is error-prone and does not require SAS-6 self-assembly.," *eLife*, vol. 4, p. e10586, Dec. 2015.

[136] N. S. Dzhindzhev, Q. D. Yu, K. Weiskopf, G. Tzolovsky, I. Cunha-Ferreira, M. Riparbelli, A. Rodrigues-Martins, M. Bettencourt-Dias, G. Callaini, and D. M. Glover, "Asterless is a scaffold for the onset of centriole assembly," *Nature*, vol. 467, pp. 714–718, Oct. 2010.

[137] J. Kleylein-Sohn, J. Westendorf, M. L. Clech, R. Habedanck, Y.-D. Stierhof, and E. A. Nigg, "Plk4-induced centriole biogenesis in human cells," *Developmental Cell*, vol. 13, pp. 190–202, Aug. 2007.

[138] N. Peel, N. R. Stevens, R. Basto, and J. W. Raff, "Overexpressing centriole-replication proteins in vivo induces centriole overduplication and de novo formation," *Current Biology*, vol. 17, pp. 834–843, 2007.

[139] X. Yan, H. Zhao, and X. Zhu, "Production of basal bodies in bulk for dense multicilia formation.," *F1000Research*, vol. 5, pp. 1533 1–7, Feb. 2016.

[140] S. P. Sorokin, "Reconstructions of centriole formation and ciliogenesis in mammalian lungs," *Journal of Cell Science*, vol. 3, pp. 207–230, 1968.

[141] H. Zhao, L. Zhu, Y. Zhu, J. Cao, S. Li, Q. Huang, T. Hu, X. Yan, and X. Zhu, "The Cep63 paralogue Deup1 enables massive de novo centriole biogenesis for vertebrate multiciliogenesis," *Nature Cell Biology*, vol. 15, pp. 1434–1444, Nov. 2013.

[142] D. K. Dehring, E. Vladar, M. Werner, J. W. Mitchell, P. Hwang, and B. J. Mitchell, "Deuterosome-mediated centriole biogenesis," *Developmental Cell*, vol. 27, pp. 103–112, Oct. 2013.

[143] J.-H. Sir, A. R. Barr, A. K. Nicholas, O. P. Carvalho, M. Khurshid, A. Sossick, S. Reichelt, C. D. Santos, G. C. Woods, and F. Gregerly, "A primary microcephaly protein complex forms a ring around parental centrioles," *Nature Genetics*, vol. 43, pp. 1147–1153, Oct. 2011.

[144] A. A. al Jord, A. al Jord, A.-I. Lemaître, N. Delgehyr, M. Faucourt, N. Spassky, and A. Meunier, "Centriole amplification by mother and daughter centrioles differs in multiciliated cells," *Nature*, vol. 516, pp. 104–107, Oct. 2014.

[145] D. Keller, M. Orpinell, N. Olivier, M. Wachsmuth, R. Mahen, R. Wyss, V. Hachet, J. Ellenberg, S. Manley, and P. Gönczy, "Mechanisms of HsSAS-6 assembly promoting centriole formation in human cells.," *The Journal of Cell Biology*, vol. 204, pp. 697–712, Apr. 2014.

[146] C.-J. C. Tang, S.-Y. Lin, W.-B. Hsu, Y.-N. Lin, C.-T. Wu, Y.-C. Lin, C.-W. Chang, K.-S. Wu, and K. T. Tang, "The human microcephaly protein STIL interacts with CPAP and is required for procentriole formation," *The EMBO Journal*, vol. 30, pp. 4790–4804, Dec. 2011.

[147] R. Qiao, G. Cabral, M. M. Lettman, Alex, er Dammermann, and G. Dong, "SAS-6 coiled-coil structure and interaction with SAS-5 suggest a regulatory mechanism in C. elegans centriole assembly," *The EMBO Journal*, vol. 31, pp. 4334–4347, Nov. 2012.

[148] M. A. Cottee, N. Muschalik, S. Johnson, J. Leveson, J. W. Raff, and S. M. Lea, "The homo-oligomerisation of both Sas-6 and Ana2 is required for efficient centriole assembly in flies.," *eLife*, vol. 4, p. e07236, June 2015.

[149] D. Kitagawa, C. Busso, I. FlUckiger, and P. GOnczy, "Phosphorylation of SAS-6 by ZYG-1 is critical for centriole formation in C. elegans embryos," *Developmental Cell*, vol. 17, pp. 900–907, Dec. 2009.

[150] C. M. Pfleger and M. W. Kirschner, "The KEN box: an APC recognition signal distinct from the D box targeted by Cdh1.," *Genes & Development*, vol. 14, pp. 655–665, Apr. 2000.

[151] A. J. Holland, D. Fachinetti, Q. Zhu, M. Bauer, I. M. Verma, E. A. Nigg, and D. W. Cleveland, "The autoregulated instability of Polo-like kinase 4 limits centrosome duplication to once per cell cycle," *Genes & Development*, vol. 26, pp. 2684–2689, Dec. 2012.

[152] H. Bazzi and K. V. Anderson, "Acentriolar mitosis activates a p53-dependent apoptosis pathway in the mouse embryo," *Proceedings of the National Academy of Sciences*, vol. 111, pp. E1491–E1500, May 2014.

[153] X. Liu, "Targeting Polo-like kinases: A promising therapeutic approach for cancer treatment," *Translational Oncology*, vol. 8, pp. 185–195, July 2015.

[154] S. T. Sredni, M. Suzuki, J.-P. Yang, J. Topczewski, A. W. Bailey, T. Gokirmak, J. N. Gross, A. de Andrade, A. Kondo, D. R. Piper, and T. Tomita, "A functional screening of the kinome identifies the Polo-like kinase 4 as a potential therapeutic target for malignant rhabdoid tumors, and possibly, other embryonal tumors of the brain.," *Pediatric Blood & Cancer*, vol. 64, p. e26551, May 2017.

[155] S. R. Marques, C. Ramakrishnan, R. Carzaniga, A. M. Blagborough, M. J. Delves, A. M. Talman, and R. E. Sinden, "An essential role of the basal body protein SAS-6 in Plasmodium male gamete development and malaria transmission.," *Cellular Microbiology*, vol. 17, pp. 191–206, Mar. 2015.

[156] H. Hu, Y. Liu, Q. Zhou, S. Siegel, and Z. Li, "The centriole cartwheel protein SAS-6 in trypanosoma brucei is required for probasal body biogenesis and flagellum assembly," *Eukaryotic Cell*, vol. 14, pp. 898–907, Sept. 2015.

[157] J. Gopalakrishnan, P. Guichard, A. H. Smith, H. Schwarz, D. A. Agard, S. Marco, and T. Avidor-Reiss, "Self-assembling SAS-6 multimer is a core centriole building block," *Journal of Biological Chemistry*, vol. 285, pp. 8759–8770, Apr. 2010.

[158] H. C. R. Klein, H. C. Klein, P. Guichard, V. Hamel, P. Gönczy, and U. S. Schwarz, "Computational support for a scaffolding mechanism of centriole assembly," *Scientific Reports*, vol. 6, p. 27075, July 2016.

[159] P. Gönczy, "Towards a molecular architecture of centriole assembly," *Nature Reviews Molecular Cell Biology*, vol. 13, pp. 425–435, July 2012.

[160] M. Hilbert, A. Noga, D. Frey, V. Hamel, P. Guichard, S. H. W. Kraatz, M. Pfreundschuh, S. Hosner, I. Flückiger, R. Jaussi, M. M. Wieser, K. M. Thieltges, X. Deupi, D. J. Müller, R. A. Kammerer, P. Gönczy, M. Hirono, M. O. Steinmetz, and S. H. Kraatz, "SAS-6 engineering reveals interdependence between cartwheel and microtubules in determining centriole architecture," *Nature Cell Biology*, vol. 18, pp. 393–403, Apr. 2016.

[161] M. A. Cottee, N. Muschalik, S. Johnson, J. Leveson, J. W. Raff, and S. M. Lea, "The homo-oligomerisation of both Sas-6 and Ana2 is required for efficient centriole assembly in flies.," *eLife*, vol. 4, p. e07236, June 2015.

[162] T. Ishida and K. Kinoshita, "Prediction of disordered regions in proteins based on the meta approach," *Bioinformatics*, vol. 24, pp. 1344–1348, June 2008.

[163] S. Mezulis, C. M. Yates, M. N. Wass, M. J. E. Sternberg, and L. A. Kelley, "The Phyre2 web portal for protein modeling, prediction and analysis," *Nature Protocols*, vol. 10, pp. 845–858, May 2015.

[164] A. Lupas, M. van Dyke, and J. Stock, "Predicting coiled coils from protein sequences," *Science*, vol. 252, pp. 1162–1164, June 1991.

[165] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, pp. 1792–1797, Apr. 2004.

[166] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. S&ouml, J. S, ouml, ding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega," *Molecular Systems Biology*, vol. 7, p. e539, Nov. 2011.

[167] M. W. Davis, "ApE -a plasmid editor v.2," 2013.

[168] Hoogland, Gattiker, Duvaud, Wilkins, Appel, and B. A. Gasteiger, "ProtParam."

[169] E. Krissinel and K. Henrick, "Inference of macromolecular assemblies from crystalline state.," *Journal of Molecular Biology*, vol. 372, pp. 774–97, Oct. 2007.

[170] J. L. Cole, J. W. Lary, T. Moody, and T. M. Laue, "Analytical ultracentrifugation: Sedimentation velocity and sedimentation equilibrium," *Methods in Cell Biology*, vol. 84, pp. 143–179, Feb. 2008.

[171] P. Schuck, "Size-distribution analysis of macromolecules by sedimentation velocity ultracentrifugation and lamm equation modeling.," *Biophysical Journal*, vol. 78, pp. 1606–1619, Apr. 2000.

[172] P. Schuck, M. A. Perugini, N. R. Gonzales, G. J. Howlett, and D. Schubert, "Size-distribution analysis of proteins by analytical ultracentrifugation: strategies and application to model systems.," *Biophysical Journal*, vol. 82, pp. 1096–111, Mar. 2002.

[173] F. Delaglio, S. Grzesiek, G. W. Vuister, G. Zhu, J. Pfeifer, and A. Bax, "NMRPipe: a multidimensional spectral processing system based on UNIX pipes.," *Journal of Biomolecular NMR*, vol. 6, pp. 277–93, Dec. 1995.

[174] T. D. Goddard and D. G. Kneller, "Sparky 3."

[175] M. P. Williamson, "Using chemical shift perturbation to characterise ligand binding," *Progress in Nuclear Magnetic Resonance Spectroscopy*, vol. 73, pp. 1–16, Sept. 2013.

[176] A. Bahrami, A. H. Assadi, J. L. Markley, and H. R. Eghbalnia, "Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein nmr spectroscopy.," *PLoS Computational Biology*, vol. 5, p. e1000307, Apr. 2009.

[177] D. S. Wishart, B. D. Sykes, and F. M. Richards, "The chemical shift index: a fast and simple method for the assignment of protein secondary structure through nmr spectroscopy.," *Biochemistry*, vol. 31, pp. 1647–1651, Mar. 1992.

[178] D. S. Wishart and B. D. Sykes, "The 13C chemical-shift index: a simple method for the identification of protein secondary structure using 13C chemical-shift data.," *Journal of Biomolecular NMR*, vol. 4, pp. 171–80, Apr. 1994.

[179] M. P. Williamson, *Applications of the NOE in Molecular Biology - Chapter 3*, vol. 65. Annual Reports on NMR Spectroscopy, Feb. 2009.

[180] G. Winter, "XIA2 an expert system for macromolecular crystallography data reduction," *Journal of Applied Crystallography*, vol. 43, pp. 186–190, Dec. 2009.

[181] Winn, Ballard, C.C., Cowtan, K.D., Dodson, E.J., Emsley, P., Evans, P.R., Keegan, R.M., Krissinel, E.B., Leslie, A.G.W., McCoy, A., McNicholas, S.J., Murshudov, G.N., Pannu, N.S., Potterton, E.A., Powell, H.R., Read, R.J., Vagin, Wilson, and K.S., "Overview of the CCP4 suite and current developments," *Acta Crystallographica Section D Biological Crystallography*, vol. 67, pp. 235–242, Apr. 2011.

[182] McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, Storoni, L.C., Read, and R.J., "Phaser crystallographic software," *Journal of Applied Crystallography*, pp. 1–17, Aug. 2007.

[183] Adams, P.D., Afonine, P.V., Bunkoczi, G., Chen, V.B., Davis, I.W., Echols, N., Headd, J.J., Hung, L.-W., Kapral, G.J., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Oeffner, R., Read, R.J., Richardson, D.C., J.S., Terwilliger, T.C., Zwart, and P.H., "PHENIX: a comprehensive python-based system for macromolecular structure solution," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 213–221, Feb. 2010.

[184] Emsley, P., Lohkamp, B., Scott, W.G., Cowtan, and K., "Features and development of Coot," *Acta Crystallographica Section D Biological Crystallography*, vol. 66, pp. 486–501, Apr. 2010.

[185] Sheldrick, G. M. Sheldrick, and G.M., "A short history of SHELX," *Acta Crystallographica Section A*, vol. 64, pp. 112–122, Jan 2007.

[186] L. Palatinus and G. Chapuis, "SUPERFLIP - a computer program for the solution of crystal structures by charge flipping in arbitrary dimensions," *Journal of Applied Crystallography*, vol. 40, pp. 786–790, Aug. 2007.

[187] A. Sali and T. L. Blundell, "Comparative protein modelling by satisfaction of spatial restraints.," *Journal of Molecular Biology*, vol. 234, pp. 779–815, Dec. 1993.

[188] N. Eswar, B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M.-Y. Shen, U. Pieper, and A. Sali, "Comparative protein structure modeling using modeller.," *Current Protocols in Bioinformatics*, vol. Chapter 5, p. Unit 5.6, Nov. 2006.

[189] A. Fiser, R. D. K, and A. Sali, "Modeling of loops in protein structures.," *Protein Science : a Publication of the Protein Society*, vol. 9, pp. 1753–1773, Oct. 2000.

[190] O. Trott and A. J. Olson, "AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.," *Journal of Computational Chemistry*, vol. 31, pp. 455–461, Feb. 2010.

[191] G. M. Morris, R. Huey, W. Lindstrom, M. F. Sanner, R. K. Belew, D. S. Goodsell, and A. J. Olson, "AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.," *Journal of Computational Chemistry*, vol. 30, no. 16, pp. 2785–91, 2009.

[192] N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch, and G. R. Hutchison, "Open Babel: An open chemical toolbox," *Journal of Cheminformatics*, vol. 3, p. 33, Feb. 2011.

[193] G. van Rossum and F. L. D. Jr., *The Python Language Reference Manual (version 3.2)*, vol. 1. 2014.

[194] J. J. Irwin, T. Sterling, M. M. Mysinger, E. S. Bolstad, and R. G. Coleman, "ZINC: A free tool to discover chemistry for biology," *Journal of Chemical Information and Modeling*, vol. 52, pp. 1757–1768, Aug. 2012.

[195] C. A. Lipinski, "Rule of five in 2015 and beyond: Target and ligand structural limitations, ligand chemistry structure and drug discovery project decisions," *Advanced Drug Delivery Reviews*, vol. 101, pp. 34–41, July 2016.

[196] T. Kaserer, K. R. Beck, M. Akram, A. Odermatt, D. Schuster, and K. Beck, "Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases," *Molecules*, vol. 20, pp. 22799–22832, Dec. 2015.

[197] J. W. Ponder and D. A. Case, "Force fields for protein simulations.," *Advances in Protein Chemistry*, vol. 66, pp. 27–85, Feb. 2003.

[198] A. R. Leach, "Molecular modelling : Principles and applications," *Pearson Education*, vol. 2, pp. 1–773, Feb. 2011.

[199] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *The Journal of Chemical Physics*, vol. 103, pp. 8577–8593, Nov. 1995.

[200] K. Lindorff-Larsen, S. Piana, K. Palmo, P. Maragakis, J. L. Klepeis, R. O. Dror, and D. E. Shaw, "Improved side-chain torsion potentials for the Amber ff99SB protein force field.," *Proteins*, vol. 78, pp. 1950–1958, July 2010.

[201] K. Lindorff-Larsen, P. Maragakis, S. Piana, M. P. Eastwood, R. O. Dror, and D. E. Shaw, "Systematic validation of protein force fields against experimental data.," *PloS One*, vol. 7, p. e32131, Feb. 2012.

[202] W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein, "Comparison of simple potential functions for simulating liquid water," *The Journal of Chemical Physics*, vol. 79, pp. 926–935, Oct. 1983.

[203] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M. R. Shirts, J. C. Smith, P. M. Kasson, D. van der Spoel, B. Hess, and E. Lindahl, "GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit," *Bioinformatics*, vol. 29, pp. 845–854, May 2013.

[204] M. J. Abraham, T. Murtola, Rol, Schulz, S. Páll, J. C. Smith, B. Hess, and E. Lindahl, "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers," *SoftwareX*, vol. 1-2, pp. 19–25, Sept. 2015.

[205] C. Kutzner, S. Páll, M. Fechner, A. Esztermann, B. L. de Groot, and H. Grubmüller, "Best bang for your buck: GPU nodes for GROMACS biomolecular simulations.," *Journal of Computational Chemistry*, vol. 36, pp. 1990–2008, Nov. 2015.

[206] A. K. Mazur, "Common molecular dynamics algorithms revisited: Accuracy and optimal time steps of Stroemer-leapfrog integrators," *Journal of Computational Chemistry*, vol. 136, pp. 354–365, Sept. 1997.

[207] B. Hess, H. Bekker, H. J. Berendsen, and J. G. Fraaije, "LINCS: A linear constraint solver for molecular simulations," *Journal of Computational Chemistry*, vol. 18, pp. 1463–1472, Sept. 1997.

[208] H. J. Berendsen, J. P. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak, "Molecular dynamics with coupling to an external bath," *The Journal of Chemical Physics*, vol. 81, pp. 3684–3690, Nov. 1984.

[209] H. Eslami, F. Mojahedi, and J. Moghadasi, "Molecular dynamics simulation with weak coupling to heat and material baths.," *The Journal of Chemical Physics*, vol. 133, p. 084105, Sept. 2010.

[210] S. A. Adcock and J. A. McCammon, "Molecular dynamics: survey of methods for simulating the activity of proteins.," *Chemical Reviews*, vol. 106, pp. 1589–615, June 2006.

[211] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople, "Gaussian 03," *Gaussian, Inc., Wallingford, CT*, 2004.

[212] L. Domicevica, T. Paramo, T. Newport, and P. C. Biggin, "LINTools - dynamic representation of protein-ligand interactions.," *Zenodo*, vol. 1, p. 1, 2016.

[213] G. A. Tribello, M. Bonomi, D. Branduardi, C. Camilloni, and G. Bussi, "PLUMED 2: New feathers for an old bird," *Computer Physics Communications*, vol. 185, pp. 604 – 613, Nov. 2013.

[214] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, pp. 33–38, Feb. 1996.

[215] C. Arquint, K. F. Sonnen, Y. D. Stierhof, and E. A. Nigg, "Cell-cycle-regulated expression of STIL controls centriole number in human cells," *Journal of Cell Science*, vol. 125, pp. 1342–1352, May 2012.

[216] S. Leidel and P. Gönczy, "Centrosome duplication and nematodes: recent insights from an old relationship.," *Developmental Cell*, vol. 9, pp. 317–25, Oct. 2005.

[217] T. Kaletta and M. O. Hengartner, "Finding function in novel targets: C. elegans as a model organism," *Nature Reviews Drug Discovery*, vol. 5, pp. 387–399, May 2006.

[218] T. Müller-Reichert, G. Greenan, E. O?Toole, and M. Srayko, "The elegans of spindle assembly," *Cellular and Molecular Life Sciences*, vol. 67, pp. 2195–2213, Aug. 2010.

[219] A. L. Hopkins and C. R. Groom, "The druggable genome," *Nature Reviews Drug Discovery*, vol. 1, pp. 727–730, Oct. 2002.

[220] Harris, J. P. Overington, S. Harris, Susanne, B. Al-Lazikani, and A. L. Hopkins, "How many drug targets are there?," *Nature Reviews Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006.

[221] F. Pammolli, L. Magazzini, and M. Riccaboni, "The productivity crisis in pharmaceutical R&D," *Nature Reviews Drug Discovery*, vol. 10, pp. 428–438, July 2011.

[222] J. Busch, "Bacteria on the rise: The fight against antibiotic resistance," *Science Inovation Union (science-union.org)*, vol. 1, pp. e1–2, Nov. 2016.

[223] J. A. Wells and C. L. McClendon, "Reaching for high-hanging fruit in drug discovery at protein–protein interfaces," *Nature*, vol. 450, pp. 1001–1009, Dec. 2007.

[224] P. Buchwald, "Small-molecule protein-protein interaction inhibitors: therapeutic potential in light of molecular size, chemical space, and ligand binding efficiency considerations.," *IUBMB Life*, vol. 62, pp. 724–31, Nov. 2010.

[225] M. Arkin, Y. Tang, and J. Wells, "Small-molecule inhibitors of protein-protein interactions: Progressing toward the reality," *Chemistry & Biology*, vol. 21, pp. 1102–1114, Oct. 2014.

[226] M. Bakail and F. Ochsenbein, "Targeting protein-protein interactions, a wide open field for drug design," *Comptes Rendus Chimie*, vol. 19, pp. 19–27, Apr. 2016.

[227] D. E. Scott, A. R. Bayly, C. Abell, and J. Skidmore, "Small molecules, big targets: drug discovery faces the protein-protein interaction challenge," *Nature Reviews Drug Discovery*, vol. 15, pp. 533–550, May 2016.

[228] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney, "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings.," *Advanced Drug Delivery Reviews*, vol. 46, pp. 3–26, Apr. 2001.

[229] A. Winter, A. P. Higueruelo, M. Marsh, A. Sigurdardottir, W. R. Pitt, and T. L. Blundell, "Biophysical and computational fragment-based approaches to targeting protein-protein interactions: applications in structure-guided drug discovery," *Quarterly Reviews of Biophysics*, vol. 45, pp. 383–426, Dec. 2012.

[230] A. G. Coyne, D. E. Scott, and C. Abell, "Drugging challenging targets using fragment-based approaches," *Current Opinion in Chemical Biology*, vol. 14, pp. 299–307, June 2010.

[231] M. Congreve, R. Carr, C. Murray, and H. Jhoti, "A 'rule of three' for fragment-based lead discovery?," *Drug Discovery Today*, vol. 8, pp. 876–7, Nov. 2003.

[232] C. Sheng, G. Dong, Z. Miao, W. Zhang, and W. Wang, "State-of-the-art strategies for targeting protein-protein interactions by small-molecule inhibitors," *Chemical Society Reviews*, vol. 44, pp. 8238–8259, Feb. 2015.

[233] B. C. Doak, R. S. Norton, and M. J. Scanlon, "The ways and means of fragment-based drug design," *Pharmacology & Therapeutics*, vol. 167, pp. 28–37, Nov. 2016.

[234] R. B. Jacob, T. Andersen, and O. M. McDougal, "Accessible high-throughput virtual screening molecular docking software for students and educators.," *PLoS Computational Biology*, vol. 8, p. e1002499, June 2012.

[235] S. P. Leelan, S. P. Leelananda, a, and S. Lindert, "Computational methods in drug discovery," *Beilstein Journal of Organic Chemistry*, vol. 12, pp. 2694–2718, Feb. 2016.

[236] T. T. Talele, S. A. Khedkar, and A. C. Rigby, "Successful applications of computer aided drug discovery: moving drugs from concept to the clinic.," *Current Topics in Medicinal Chemistry*, vol. 10, pp. 127–141, Feb. 2010.

[237] S. Forli, "Charting a path to success in virtual screening," *Molecules*, vol. 20, pp. 18732–18758, Nov. 2015.

[238] R. Sable and S. Jois, "Surfing the protein-protein interaction surface using docking methods: Application to the design of PPI inhibitors," *Molecules*, vol. 20, pp. 11569–11603, July 2015.

[239] S. Betzi, A. Restouin, S. Opi, S. T. Arold, I. Parrot, F. Guerlesquin, X. Morelli, and Y. Collette, "Protein protein interaction inhibition (2p2i) combining high throughput and virtual screening: Application to the hiv-1 nef protein.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19256–61, 2007.

[240] L. A. Emert-Sedlak, H. M. Loughran, H. Shi, J. L. 3rd, S. T. Shu, J. Zhao, B. W. Day, J. E. Wrobel, A. B. Reitz, and T. E. Smithgall, "Synthesis and evaluation of orally active small molecule HIV-1 Nef antagonists.," *Bioorganic & Medicinal Chemistry Letters*, vol. 26, pp. 1480–1484, Apr. 2016.

[241] S. H. Joo, "Cyclic peptides as therapeutic agents and biochemical tools.," *Biomolecules & Therapeutics*, vol. 20, pp. 19–26, Feb. 2012.

[242] S. Zhang, K. Kumar, X. Jiang, A. Wallqvist, and J. Reifman, "DOVIS: an implementation for high-throughput virtual screening using AutoDock.," *BMC Bioinformatics*, vol. 9, p. 126, Mar. 2008.

[243] T. A. Bikowski, W. Jiang, B. Roux, W. F. Anderson, and A. Joachimiak, "Virtual high-throughput ligans screening," *Structural Genomics and Drug Discovery*, vol. 1140, pp. 251–261, Feb. 2014.

[244] S. R. Ellingson, J. C. Smith, and J. Baudry, "Vinampi: facilitating multiple receptor high-throughput virtual docking on high-performance computers.," *Journal of Computational Chemistry*, vol. 34, pp. 2212–2221, Oct. 2013.

[245] E. Glaab, "Building a virtual ligand screening pipeline using free software: a survey," *Briefings in Bioinformatics*, vol. 17, pp. 352–366, Apr. 2016.

[246] J. Biesiada, A. Porollo, P. Velayutham, M. Kouril, and J. Meller, "Survey of public domain software for docking simulations and virtual screening.," *Human Genomics*, vol. 5, pp. 497–505, Aug. 2011.

[247] Z. Wang, H. Sun, X. Yao, D. Li, L. Xu, Y. Li, S. Tian, and T. Hou, "Comprehensive evaluation of ten docking programs on a diverse set of protein-ligand complexes: the prediction accuracy of sampling power and scoring power," *Physical Chemistry Chemical Physics*, vol. 18, pp. 12964–12975, Feb. 2016.

[248] S. B. Shuker, P. J. Hajduk, R. P. Meadows, and S. W. Fesik, "Discovering high-affinity ligands for proteins: SAR by NMR.," *Science*, vol. 274, pp. 1531–1534, Dec. 1996.

[249] P. Kuhn, K. Wilson, M. G. Patch, and R. C. Stevens, "The genesis of high-throughput structure-based drug discovery using protein crystallography.," *Current Opinion in Chemical Biology*, vol. 6, pp. 704–10, Nov. 2002.

[250] C. J. Manly, J. Chandrasekhar, J. W. Ochterski, J. D. Hammer, and B. B. Warfield, "Strategies and tactics for optimizing the hit-to-lead process and beyond–a computational chemistry perspective.," *Drug Discovery Today*, vol. 13, pp. 99–109, Mar. 2008.

[251] I. Müller, "Guidelines for the successful generation of protein-ligand complex crystals.," *Acta Crystallographica. Section D, Structural Biology*, vol. 73, pp. 79–92, Mar. 2017.

[252] H. Alonso, A. A. Bliznyuk, and J. E. Gready, "Combining docking and molecular dynamic simulations in drug design," *Medicinal Research Reviews*, vol. 26, pp. 1–38, Aug. 2006.

[253] C. Bissantz, B. Kuhn, and M. Stahl, "A medicinal chemist's guide to molecular interactions.," *Journal of Medicinal Chemistry*, vol. 53, pp. 5061–5084, Aug. 2010.

[254] J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, vol. 1. W H Freeman, 8 ed., 2015.

[255] M. R. Arkin and J. A. Wells, "Small-molecule inhibitors of protein–protein interactions: progressing towards the dream," *Nature Reviews Drug Discovery*, vol. 3, pp. 301–317, May 2004.

[256] H. Zhao and Z. Guo, "Medicinal chemistry strategies in follow-on drug discovery.," *Drug Discovery Today*, vol. 14, pp. 516–22, June 2009.

[257] P. L. Kastritis and A. M. Bonvin, "On the binding affinity of macromolecular interactions: daring to ask why proteins interact," *Journal of the Royal Society Interface*, vol. 10, p. e20120835, Dec. 2012.

[258] Z. Yu, J. P. D. Veldhoven, J. Louvel, I. M. E. 't Hart, M. B. Rook, M. A. G. Heyden, L. H. Heitman, A. P. IJzerman, J. P. van Veldhoven, I. M. Hart, and M. A. van der Heyden, "Structure–affinity relationships (SARs) and structure–kinetics relationships (SKRs) of K 11.1 blockers," *Journal of Medicinal Chemistry*, vol. 58, pp. 5916–5929, Sept. 2015.

[259] E. Cauet, M. Rooman, R. Wintjens, J. Lievin, and C. Biot, "Histidine aromatic interactions in proteins and protein ligand complexes : A quantum chemical study of X-ray and modeling," *Journal of Chemical Theory and Computation*, vol. 1, pp. 472–483, June 2005.

[260] S.-M. Liao, du Qi-Shi, J.-Z. Meng, Z.-W. Pang, and R.-B. Huang, "The multiple roles of histidine in protein interactions," *Chemistry Central Journal*, vol. 7, p. 44, Feb. 2013.

[261] M. A. Cottee, J. W. Raff, S. M. Lea, and H. Roque, "SAS-6 oligomerization: the key to the centriole?," *Nature Chemical Biology*, vol. 7, pp. 650–653, Oct. 2011.

[262] A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello, "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans.," *Nature*, vol. 391, pp. 806–811, Mar. 1998.

[263] S. E. Mohr, J. A. Smith, C. E. Shamu, R. A. Neumüller, and N. Perrimon, "RNAi screening comes of age: improved techniques and complementary approaches," *Nature Reviews Molecular Cell Biology*, vol. 15, pp. 591–600, Sept. 2014.

[264] K. F. Sonnen, L. Schermelleh, H. Leonhardt, and E. A. Nigg, "3D-structured illumination microscopy provides novel insight into architecture of human centrosomes," *Biology Open*, vol. 1, pp. 965–976, Nov. 2012.

[265] A. Kumar, S. C. Girimaji, M. R. Duvvari, and S. H. Blanton, "Mutations in STIL, encoding a pericentriolar and centrosomal protein, cause primary microcephaly," *American Journal of Human Genetics*, vol. 84, pp. 286–290, Mar. 2009.

[266] S. Izraeli, L. A. Lowe, V. L. Bertness, D. J. Good, D. W. Dorward, I. R. Kirsch, and M. R. Kuehn, "The SIL gene is required for mouse embryonic axial development and left-right specification.," *Nature*, vol. 399, pp. 691–694, July 1999.

[267] A. Castiel, M. M. Danieli, A. David, S. Moshkovitz, P. D. Aplan, I. R. Kirsch, M. Brandeis, A. Kramer, and S. Izraeli, "The stil protein regulates centrosome integrity and mitosis through suppression of chfr," *Journal of Cell Science*, vol. 124, pp. 532–539, Feb. 2011.

[268] K. L. Pfaff, C. T. Straub, K. Chiang, D. M. Bear, Y. Zhou, and L. I. Zon, "The zebra fish cassiopeia mutant reveals that sil is required for mitotic spindle organization," *Molecular and Cellular Biology*, vol. 27, pp. 5887–5897, Aug. 2007.

Appendix A

Chapter 3

A.1 Observations on the Role of S123 on Loop 103-130 Interactions from Molecular Dynamics Simulation

The pathway controlling kinase ZYG-1 has been shown to interact with SAS-6 . However only conflicting evidence exists on SAS-6 phosphorylation. Residue S123 has been suggested to be a potential site of phosphorylation, the modification of which was hypothesised to be necessary [149]. In this study, phosphorylation specific antibodies detected GFP-SAS-6, but not GFP-SAS-6 S123A *in cellulo*. Furthermore, GFP-SAS-6 S123D - used as a phosphomimicking mutation - seemed to rescue a ZYG-1 siRNA induced failure of centriole duplication in *C. elegans* embryo cells. Despite this evidence, in experiments with more tightly controlled expression levels, SAS-6 S123A was able to rescue a *sas6* deletion phenotype, contradicting the reported necessity of S123 phosphorylation [60].

Work on this residue has given rise to constructs of SAS-6 with a S123E mutation, which seem to have better stability during protein purification compared to wild type

constructs and are therefore preferentially used for biophysical experiments. An interest in the nature of mutations of S123 led to a series of computational experiments into the behaviour of loop 103-130 with S123A, S123E and S123D mutation. The data discussed here is preliminary in nature and is presented here for reasons of completeness of documentation, as no conclusion could be drawn.

One of the loop 103-130 models described in Chapter 3 was used for 50 ns MD simulations with mutations S123A, S123E and S123D (Figure A.1). In all simulations loop 103-130 assumed a position closer to the dimerisation interface and contacted the helix region α1-α2. This is comparable to the behaviour of the loop region in simulations with the wild type loop (Chapter 3).

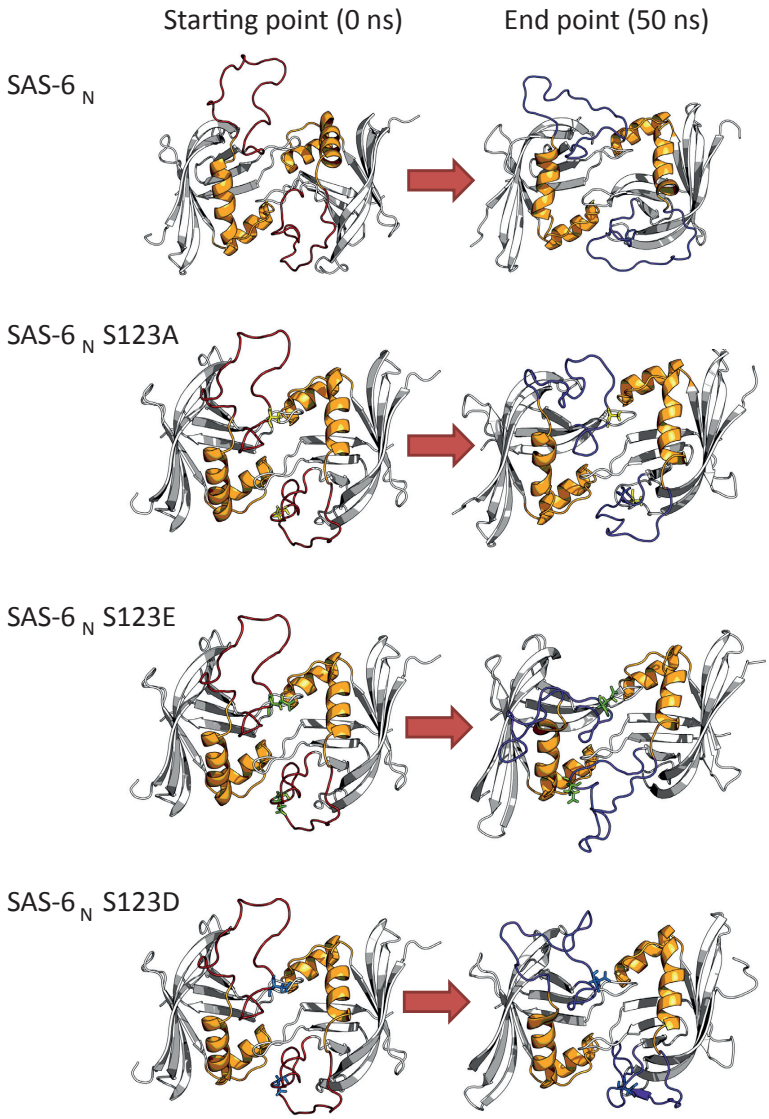


Fig. A.1: Model of SAS-6 N-terminal Domain Loop (S123A, S123E and S123D) 103-130 in MD: Structure of the *C. elegans* SAS-6_N dimer PDB code 4G79 (1.8 Å resolution) [103] incorporated in three models of loop 103-130 in SAS-6 dimer before and after 50 ns of MD. Loops in dimer were modelled individually. Residues are highlighted in *red* for the starting point and *blue* for the end point of the simulation. Residues of helix region α1-α2 are highlighted in *orange*. Residue S123A is highlighted in *yellow*, residue S123E is highlighted in *green* and residue S123D is highlighted in *light blue*.

The analysis of the time residues of loop 103-130 and α1-α2 spent in contact reveals no obvious differences in the α1-α2 residues contacted compared to the previously

discussed results for the wild type loop (Chapter 3) (Figure A.2). Specifically, the point mutation S123A overall recreated the contact pattern of the wild type simulations. For mutants S123E and S123D a decreased affinity for the α1 helix residues Q75 and I76 could be assumed. Since only one starting position for loop 103-130 was used, this minor effect could be due to unfavourable sampling. The region of interacting residues on loop 103-130 for both S123E and S123D is notably smaller, than in wild type and S123A simulations. This might be a first indication of altered loop properties.

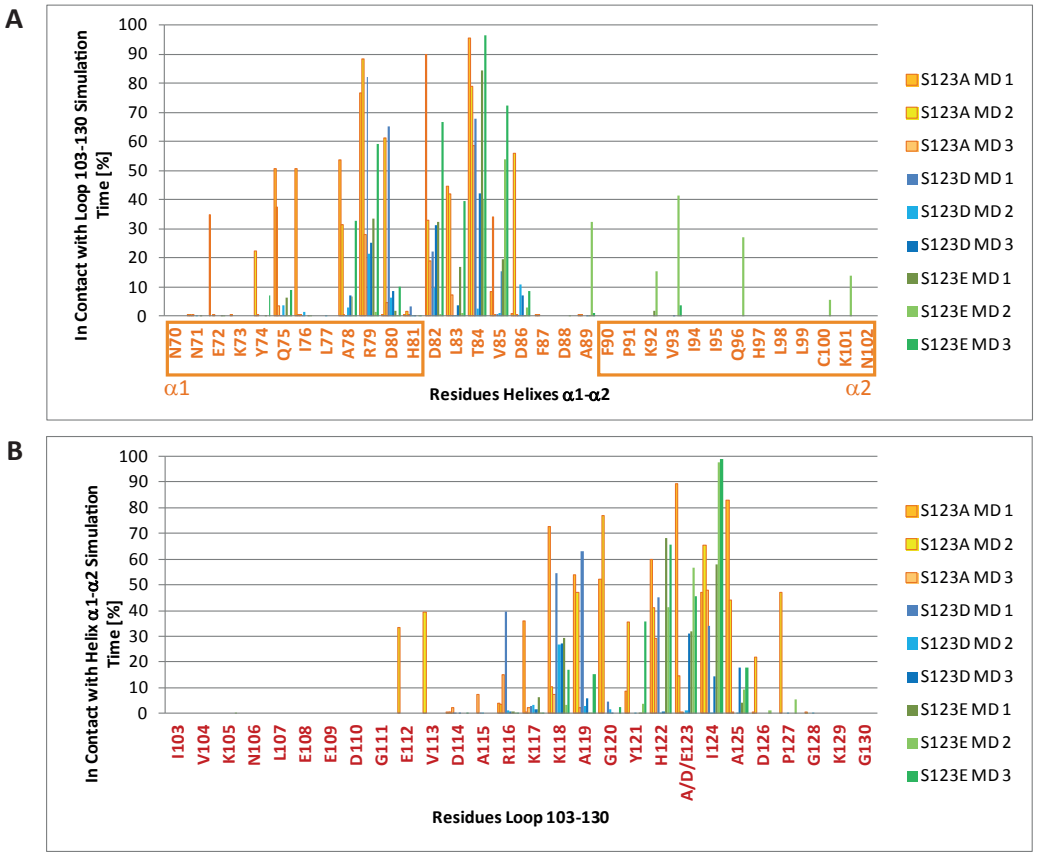


Fig. A.2: Analysis of Interactions Formed During MD between Loop 103-130 and α1-α2 (S123A, S123E and S123D): Over 50 ns simulations conducted on ceSAS-6$_N$ with loop model containing S123A, S123E or S123D - three repeats. (A) Time spent by residues in region α1-α2 in contact with residues of loop 103-130. (B) Time spent by residues in loop 103-130 in contact with residues of region α1-α2.

A.1 Observations on the Role of S123 on Loop 103-130 Interactions from Molecular Dynamics Simulation

For SAS-6_N construct S123E SMD simulations with applied pulling force were performed in a similar manner to the simulations described in Chapter 3. One crucial difference in the set-up of these experiments is that as starting point for this simulation, the end point of a MD simulation with S123E integrated was chosen. For earlier SMD simulations with loop region 103-130, the same starting position of the loop was used to compare I154 , I154W and I154E. As a control experiment, an SMD simulation with the loop 103-130 starting position used for the earlier experiments and inserted S123E mutation would have to be performed.

In the starting position used for the simulation of SAS-6_N S123E, one of both loops was already detached from the dimerisation interface (Figure A.3). The second loop detaches after 15 ns. The work required for pulling of this construct is comparable to work required to pull apart SAS-6_N I154E (analysis shown in Figure A.4 together with analysis of other SAS-6_N SMD simulations already shown in Chapter 3). This is likely to be an artifact of the chosen starting position, with only one attached loop.

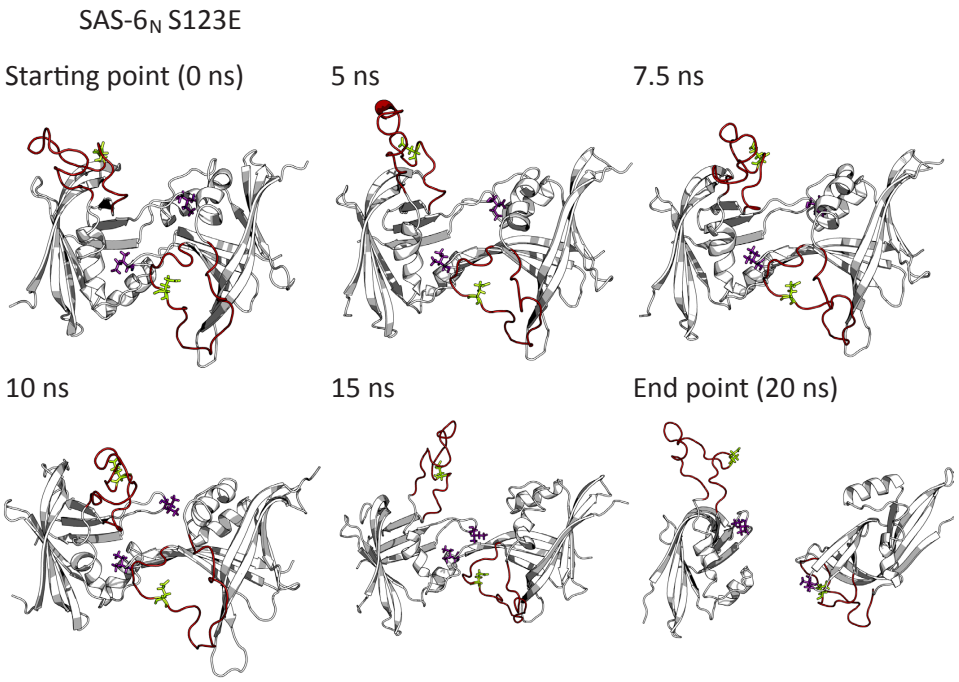


Fig. A.3: SAS-6$_N$ Constructs with Modelled Loop 103-130 S123E During SMD Simulations: Stages of separation of SAS-6$_N$ S123E dimer during 20 ns SMD simulation. Loop 103-130 is highlighted in *red*, residue I154 in *purple* and S123E in *green*.

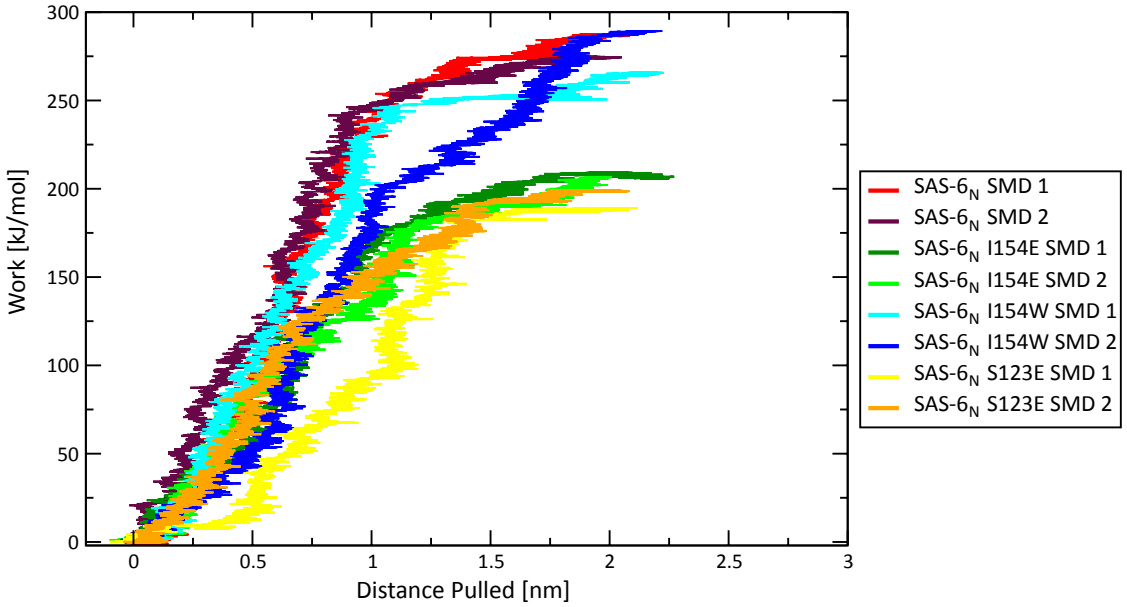


Fig. A.4: Work Required for SAS-6$_N$ Dimer Separation in SMD: Analysis of two repeats of simulations of SAS-6$_N$ (*red* and *dark red*), SAS-6$_N$ I154E (*green* and *dark green*), SAS-6$_N$ I154W (*blue* and *dark blue*) (as shown in Chapter 3) and additional analysis of two repeats of simulations of SAS-6$_N$ S123E (*yellow* and *orange*).

Appendix B

Chapter 4

B.1 Virtual Screening Pipeline Script

B.1.1 split-mol2-3.0.py

```
#!/usr/bin/env python3.2
#
#This program has been adapted for python3 and is used to
split ZINC database mol2 files containing several entries into
individual files and write them into a folder hierarchy that
follows the ZINC ID like so: entry ZINC12345678 ->
workingdirectory/1/2/3/4/5/6/7/8/ZINC12345678.mol2
#Already existing IDs will result in _2.mol2 entries.
#command line requires -i filename arguemnt.
import os, sys


if __name__ == '__main__':
    import sys
    import getopt
    import re


    def usage():
        #"Print helpful, accurate usage statement to stdout."
        print ("Usage: split_mol2_3.0.py -i filename")
        print ()
        print ("    Description of command...")
        print ("         -i     multi_mol2_filename")
        print ("    Optional parameters:")
        print ("        [-v]    verbose output")



    # process command arguments
    try:
        opt_list, args = getopt.getopt(sys.argv[1:], 'i:vh')
    except getopt.GetoptError as msg:
        print ('split_multi_mol2_file.py: %s' %msg)
        usage()
        sys.exit(2)

    # initialize required parameters
    #-i: multi_mol2_filename
    multi_mol2_filename =  None
    verbose = False

    #'i:vh'
    for o, a in opt_list:
        #print "o=", o, " a=", a
        if o in ('-i', '--i'):
            multi_mol2_filename = a
            if verbose: print ('set multi_mol2_filename to ',
a)
        if o in ('-v', '--v'):
            verbose = True
            if verbose: print ('set verbose to ', True)
        if o in ('-h', '--'):
            usage()
            sys.exit()
```

```
    if not  multi_mol2_filename:
        print ('split_multi_mol2_file: multimol2 file name
must be specified.')
        usage()
        sys.exit()


    #step one: open multiple mol2 file and get all the lines
    fptr = open(multi_mol2_filename)
    alllines = fptr.readlines()
    fptr.close()
    #step two: set up counter for filenames, molecule counter
and flag
    molctr = 0
    #inmol = 0
    #step three: process alllines
    in_molecule = False
    for i in range(len(alllines)):
        line = alllines[i]
        #optr.write(line)
        if line.find("@<TRIPOS>MOLECULE")==0: # check for
beginning of mol
            if verbose: print ('found beginning of molecule ',
molctr)
            if in_molecule:
                optr.close()
                if verbose: print ("closed ", filename)
            else:
                in_molecule = True
            zid = alllines[i+1].strip()
            filename = zid + '.mol2'
            #get id number of molecule for folder hierarchy
            filenumber = re.findall(r'\d+', zid)
            print (filenumber)
            #make a list of numbers
            realnumber = str(filenumber[0])
            print (realnumber)
            la = list(realnumber)
            print (la)
            #map(int,str(12345)) for list of integers
            #make list of only directory names in current
working directory and define home directory for return
            home = os.getcwd()
            print(home)
            for h in range(0, 8):
             directorylist = list()
             for x in os.listdir(os.getcwd()):
              if os.path.isfile(x): pass
              else: directorylist.extend(x)
             #print(directorylist)
            # compare filename to directories list, create if
not existent and change into number h layer of directory)
             if not la[h] in directorylist :
os.makedirs(la[h])
             os.chdir(la[h])




              #print (os.getcwd())


             #directories= os.walk(os.getcwd())
             #print (directories)

             if os.path.exists(filename):
              pass
                 #for i in range(1, 100):
                     #filename = zid + "_" + str(i) + ".mol2"
                     #if not os.path.exists(filename):
                         #break


             optr = open(filename, 'w')
             molctr += 1
         optr.write(line)
         #go back to home directory
         os.chdir(home)
         #print ('now we are home', os.getcwd())
     if verbose: print ('split %s into %d files' %(multi_mol2
  _filename, molctr))
```

B.1.2 pdbqt-converter.py

```
#!/usr/bin/env python3.2
#
#This script converts all .mol2 files in all subdirectories
into .pdbqt files using open babel.
import os, sys



if __name__ == '__main__':
    import sys
    import getopt
    import re

    from itertools import chain
    import subprocess

#start by getting a list of all directories in the current
working directory
    directorylist = list()
    for x in os.listdir(os.getcwd()):
              if os.path.isfile(x): pass
              else: directorylist.extend(x)
    #print(directorylist)

#use the os.walk modlue to go through all the directories a
pathname is needed, so create a list of complete pathnames for
the directories found
    paths = list()
    mol2files = list()

    for h in range(len(directorylist)):
     paths.append(os.getcwd() + '/' + directorylist[h])
    #print(paths)

#execute the os.walk over the path list and iterate through
them.
#files is a list starting with start folder path and then
nested information about found items
#filenamemol2 is list of each individual item in files (my
take on removing the nested list problem)
#extract all files .mol2 and put in last list mol2files
(filenames only no path information)
    for files in chain.from_iterable(os.walk(path) for path in
paths):
     #print(files)
     filenamemol2=[j for i in files for j in i]
     for y in range(len(filenamemol2)):
      if filenamemol2[y].endswith(".mol2"):
       mol2files.append(filenamemol2[y])
    #print(mol2files)

#make list of filenames with .pdbqt
    pdbqtfiles=list()
    for x in range(len(mol2files)):
     outputformat=mol2files[x][:-5] + ".pdbqt"
     #print(outputformat)
```

```
     pdbqtfiles.append(outputformat)
    #print(pdbqtfiles)

#make path for each mol2file for use in openbabel command
line, first string of numbers into list of single numbers,
then join with slash, then write in list
    teilpath=list()
    for z in range(len(mol2files)):
     filenumber = re.findall(r'\d+', mol2files[z][:-5])
     #print (filenumber)
     realnumber = str(filenumber[0])
     #print (realnumber)
     singlenumbers=list(realnumber)
     joinnumbers= "/".join(singlenumbers)
     teilpath.append("/"+joinnumbers+"/")
    #print(teilpath)

    filepaths=list()
#put root infront and make list
    for w in range(len(teilpath)):
     filepaths.append(os.getcwd() + teilpath[w])
    #print(filepaths)

#combine with filename and run babel via linux command
    molctr=0
    doublectr=0
    for v in range(len(mol2files)):
     fileinput= filepaths[v] + mol2files[v]
     fileoutput= filepaths[v] + pdbqtfiles[v]
     if os.path.exists(fileoutput): doublectr += 1
     else: subprocess.call(["babel", "-i", "mol2", fileinput,
"-o", "pdbqt", fileoutput])
    #add 1 to molecule count
     molctr += 1
    print ('converted %s molecules to pdbqt with openbabel, %s
of which already existed in the right format' %(molctr,
doublectr))
```

B.1.3 docking.py

```
#!/usr/bin/env python3.2
#
# This program is used to dock multiple .pdbqt ligands to one
receptor.pdbqt using autodock vina. The receptor.pdbqt and
config.txt are required to be in working directory. Ligand
files have to have ZINC*.pdbqt format and need to be located
in subdirectories. 'config.txt' and receptor.pdbqt are
necessary in working directory.
#Errors might occur if run is repeated, in this case please
change receptor name in config file.

import os, sys



if __name__ == '__main__':
    import sys
    import getopt
    import re

    from itertools import chain
    import subprocess
    import fileinput

#check if required config.txt is in working directory
    if not os.path.exists('config.txt'):
     print('config.txt is missing in current working
directory')
     sys.exit()
#get name of target molecule from config file
    regexp = re.compile(r'receptor = (.*)$')
    receptor = []
    with open('config.txt') as f:
       for line in f:
         match = regexp.match(line)
         if match:
          receptor.append(match.group(1))
    receptor_dock = receptor[0]
    receptor_name = receptor_dock[:-6]
#check if required receptor.pdbqt is in working directory
    if not os.path.exists(receptor_dock):
     print('receptor.pdbqt is missing in current working
directory')
     sys.exit()

#start by getting a list of all directories in the current
working directory
    directorylist = list()
    for x in os.listdir(os.getcwd()):
              if os.path.isfile(x): pass
              else: directorylist.extend(x)
    #print(directorylist)
#to use the os.walk modlue to go through all the directories a
pathname is needed, so create a list of complete pathnames for
the directories found
    paths = list()
```

```
    pdbqtfiles = list()

    for h in range(len(directorylist)):
     paths.append(os.getcwd() + '/' + directorylist[h])
    #print(paths)
# execute the os.walk over the path list and iterate through
them.
#files is a list starting with start folder path and then
nested information about found items
#filenamepdbqt is list of each individual item in files (my
take on removing the nested list problem
#extract all files .pdbqt and put in last list pdbqtfiles
(filenames only no path information)
    for files in chain.from_iterable(os.walk(path) for path in
paths):
     #print(files)
     filenamepdbqt=[j for i in files for j in i]
     for y in range(len(filenamepdbqt)):
      if filenamepdbqt[y].endswith(".pdbqt"):
       if filenamepdbqt[y].startswith('ZINC'):
        pdbqtfiles.append(filenamepdbqt[y])
    #print(pdbqtfiles)

#make list of filenames with .txt for log
    logfiles=list()
    for x in range(len(pdbqtfiles)):
     outputformat=pdbqtfiles[x][:-6] + ".txt"
     #print(outputformat)
     logfiles.append(outputformat)
    #print(pdbqtfiles)

#make path for each pdbqtfile paths, first string of numbers
into list of single numbers, then join with slash, then write
in list
    teilpath=list()
    for z in range(len(pdbqtfiles)):
     filenumber = re.findall(r'\d+', pdbqtfiles[z][:-5])
     #print (filenumber)
     realnumber = str(filenumber[0])
     #print (realnumber)
     singlenumbers=list(realnumber)
     joinnumbers= "/".join(singlenumbers)
     teilpath.append("/"+joinnumbers+"/")
    #print(teilpath)
    filepaths=list()
#put root infront and make list
    for w in range(len(teilpath)):
     filepaths.append(os.getcwd() + teilpath[w])
    #print(filepaths)

#set a molecule counter for result line
    molctr=0

#combine with filename and run vina via linux command
    for v in range(len(pdbqtfiles)):
#get current entry to config file and change config file for
```

```
ligand_pdbqt path and output all_pdbqt path
     inputfile= filepaths[v] + pdbqtfiles[v]
     filelog= filepaths[v] + 'log_'+ receptor_name + '_' +
logfiles[v]
     newall= filepaths[v] + 'all_' + receptor_name + '_' +
pdbqtfiles[v]

     regexp = re.compile(r'ligand = (.*)$')
     result = []
     with open('config.txt') as f:
        for line in f:
          match = regexp.match(line)
          if match:
           result.append(match.group(1))
     #print(result)
     oldconfig = result[0]
     #print(oldconfig)
     for line in fileinput.input(['config.txt'],
inplace=True):
      sys.stdout.write(line.replace(oldconfig, inputfile))

     regexp = re.compile(r'out = (.*)$')
     all_dock = []
     with open('config.txt') as f:
        for line in f:
          match = regexp.match(line)
          if match:
           all_dock.append(match.group(1))
     #print(all_dock)
     oldall = all_dock[0]
     #print(oldall)

     for line in fileinput.input(['config.txt'],
inplace=True):
      sys.stdout.write(line.replace(oldall, newall))
#add 1 to molecule count
     molctr += 1
#if logfile already exsitst do not overwrite
     if os.path.exists(filelog):
      print('log files for this docking process already exist,
please change receptor name in the config.txt if you wish to
continue')
      pass
     else: subprocess.call(["vina", "--config", 'config.txt',
"--log", filelog])

    print ('docked %s molecules to %s' %(molctr,
receptor_dock))
```

B.1.4 docking-parallel.py

```
#!/usr/bin/env python3.2
#
# This program is used to dock multiple .pdbqt ligands to one
receptor.pdbqt using autodock vina. The receptor.pdbqt and
config.txt are required to be in working directory. Ligand
files have to have ZINC*.pdbqt format and need to be located
in subdirectories. 'config.txt' and receptor.pdbqt are
necessary in working directory.
#Errors might occur if run is repeated, in this case please
change receptor name in config file.

import os, sys
import shutil #used to reproduce config file
def dock(v):
#combine paths with filename, test if config folder exists and
make specific config files from original config.txt in config
folder and run vina via linux command
    #for v in range(len(pdbqtfiles)):
 #make config folder if does not exist
     if not os.path.exists('config'):
      os.makedirs('config')
 #get current entry to config file and change config file for
ligand_pdbqt path and output all_pdbqt path
     inputfile= filepaths[v] + pdbqtfiles[v]
     filelog= filepaths[v] + 'log_'+ receptor_name + '_' +
logfiles[v]
     newall= filepaths[v] + 'all_' + receptor_name + '_' +
pdbqtfiles[v]
 #copy original config file into config filder and rename
accoring to molecule
     shutil.copyfile('config.txt',
'config/config'+pdbqtfiles[v]+'.txt')
 #change ligand and outfile
     regexp = re.compile(r'ligand = (.*)$')
     result = []
     with open('config/config'+pdbqtfiles[v]+'.txt') as f:
        for line in f:
          match = regexp.match(line)
          if match:
           result.append(match.group(1))
     #print(result)
     oldconfig = result[0]
     #print(oldconfig)
     for line in
fileinput.input(['config/config'+pdbqtfiles[v]+'.txt'],
inplace=True):
      sys.stdout.write(line.replace(oldconfig, inputfile))

     regexp = re.compile(r'out = (.*)$')
     all_dock = []
     with open('config/config'+pdbqtfiles[v]+'.txt') as f:
        for line in f:
          match = regexp.match(line)
          if match:
           all_dock.append(match.group(1))
     #print(all_dock)
```

```
     oldall = all_dock[0]
     #print(oldall)

     for line in
fileinput.input(['config/config'+pdbqtfiles[v]+'.txt'],
inplace=True):
      sys.stdout.write(line.replace(oldall, newall))
 #add 1 to molecule count
     #molctr += 1
 #if logfile already exsitst do not overwrite
     if os.path.exists(filelog):
      print('log files for this docking process already exist,
please change receptor name in the config.txt if you wish to
continue')
      pass
     else: subprocess.call(["vina", "--config",
'config/config'+pdbqtfiles[v]+'.txt', "--log", filelog])

if __name__ == '__main__':
    import sys
    import getopt
    import re

    from itertools import chain
    import subprocess
    import fileinput
    from joblib import Parallel, delayed #used to parallelise

#check if required config.txt is in working directory
    if not os.path.exists('config.txt'):
     print('config.txt is missing in current working
directory')
     sys.exit()
#get name of target molecule from config file
    regexp = re.compile(r'receptor = (.*)$')
    receptor = []
    with open('config.txt') as f:
       for line in f:
         match = regexp.match(line)
         if match:
          receptor.append(match.group(1))
    receptor_dock = receptor[0]
    receptor_name = receptor_dock[:-6]
#check if required receptor.pdbqt is in working directory
    if not os.path.exists(receptor_dock):
     print('receptor.pdbqt is missing in current working
directory')
     sys.exit()

#start by getting a list of all directories in the current
working directory
    directorylist = list()
    for x in os.listdir(os.getcwd()):
              if os.path.isfile(x): pass
              else: directorylist.extend(x)
    #print(directorylist)
```

```
#to use the os.walk modlue to go through all the directories a
pathname is needed, so create a list of complete pathnames for
the directories found
    paths = list()
    pdbqtfiles = list()

    for h in range(len(directorylist)):
     paths.append(os.getcwd() + '/' + directorylist[h])
    #print(paths)
# execute the os.walk over the path list and iterate through
them.
#files is a list starting with start folder path and then
nested information about found items
#filenamepdbqt is list of each individual item in files (my
take on removing the nested list problem
#extract all files .pdbqt and put in last list pdbqtfiles
(filenames only no path information)
    for files in chain.from_iterable(os.walk(path) for path in
paths):
     #print(files)
     filenamepdbqt=[j for i in files for j in i]
     for y in range(len(filenamepdbqt)):
      if filenamepdbqt[y].endswith(".pdbqt"):
       if filenamepdbqt[y].startswith('ZINC'):
        pdbqtfiles.append(filenamepdbqt[y])
    #print(pdbqtfiles)

#make list of filenames with .txt for log
    logfiles=list()
    for x in range(len(pdbqtfiles)):
     outputformat=pdbqtfiles[x][:-6] + ".txt"
     #print(outputformat)
     logfiles.append(outputformat)
    #print(pdbqtfiles)

#make path for each pdbqtfile paths, first string of numbers
into list of single numbers, then join with slash, then write
in list
    teilpath=list()
    for z in range(len(pdbqtfiles)):
     filenumber = re.findall(r'\d+', pdbqtfiles[z][:-5])
     #print (filenumber)
     realnumber = str(filenumber[0])
     #print (realnumber)
     singlenumbers=list(realnumber)
     joinnumbers= "/".join(singlenumbers)
     teilpath.append("/"+joinnumbers+"/")
    #print(teilpath)
    filepaths=list()
#put root infront and make list
    for w in range(len(teilpath)):
     filepaths.append(os.getcwd() + teilpath[w])
    #print(filepaths)

#set a molecule counter for result line
    #molctr=0
```

```
 #execute function defined earlier (makes new config file and
 calls vina) in parallel. define level of parallelisation in
 n_jobs.
     Parallel(n_jobs=8)(delayed(dock)(v) for v in
 range(len(pdbqtfiles)))
     #print ('docked %s molecules to %s' %(molctr,
 receptor_dock))
```

B.1.5 scoring.py

```
#!/usr/bin/env python3.2
#
#This program finds results from autodock vina runs in format
('all_'+receptor-file_+ZINC12345678.pdbqt) in subfolders,
extracts vina score and produces a sorted list of user defined
length given as argument in command line. In addition a
config.txt file is needed in working directory.

import os, sys



if __name__ == '__main__':
    import sys
    import getopt
    import re

    from itertools import chain
    import subprocess
    import fileinput
    import argparse
    #import natsort
#start by getting a list of all directories in the current
working directory
    directorylist = list()
    for x in os.listdir(os.getcwd()):
              if os.path.isfile(x): pass
              else: directorylist.extend(x)
    #print(directorylist)
#get name of target molecule from config file
    regexp = re.compile(r'receptor = (.*)$')
    receptor = []
    with open('config.txt') as f:
       for line in f:
         match = regexp.match(line)
         if match:
          receptor.append(match.group(1))
    receptor_dock = receptor[0]
    receptor_name = receptor_dock[:-6]
#to use the os.walk modlue to go through all the directories a
start pathname is needed, so create a list of complete
pathnames for the directories found, but only for directories
0,1,2,3,4,5,6,7,8,9 (results folder in same directories are
being ignored)
    paths = list()
    all_pdbqtfiles = list()

    for h in range(len(directorylist)):
     if (directorylist[h].startswith('0') or
directorylist[h].startswith('1') or
directorylist[h].startswith('2') or
    directorylist[h].startswith('3') or
directorylist[h].startswith('4') or
    directorylist[h].startswith('5') or
directorylist[h].startswith('6') or
    directorylist[h].startswith('7') or
```

```
directorylist[h].startswith('8') or
    directorylist[h].startswith('9')):
      if os.getcwd() + '/' + directorylist[h] not in paths:
       paths.append(os.getcwd() + '/' + directorylist[h])
    #print(paths)
# execute the os.walk over the path list and iterate through
them.
#files is a list starting with start folder path and then
nested information about found items
#filenameall is list of each individual item in files (my take
on removing the nested list problem
#extract all files beginning with all_receptorname and put in
last list all_pdbqtfiles (filenames only no path information)
    for files in chain.from_iterable(os.walk(path) for path in
paths):
     #print(files)
     filenameall=[j for i in files for j in i]
     for y in range(len(filenameall)):
      #for y in range(len(filenameall)):
      if (filenameall[y].endswith("_1.pdbqt")or
filenameall[y].endswith("_2.pdbqt") or
filenameall[y].endswith("_3.pdbqt") or
filenameall[y].endswith("_4.pdbqt") or
filenameall[y].endswith("_5.pdbqt")or
filenameall[y].endswith("_6.pdbqt")or
filenameall[y].endswith("_7.pdbqt")or
filenameall[y].endswith("_8.pdbqt")or
filenameall[y].endswith("_9.pdbqt")or
filenameall[y].endswith("_10.pdbqt")or
filenameall[y].endswith("_11.pdbqt")or
filenameall[y].endswith("_12.pdbqt")or
filenameall[y].endswith("_13.pdbqt")or
filenameall[y].endswith("_14.pdbqt")or
filenameall[y].endswith("_15.pdbqt")or
filenameall[y].endswith("_16.pdbqt")or
filenameall[y].endswith("_17.pdbqt")or
filenameall[y].endswith("_18.pdbqt")or
filenameall[y].endswith("_19.pdbqt") or
filenameall[y].endswith('~')) :
       pass
      elif filenameall[y].startswith("all_"+receptor_name):
       all_pdbqtfiles.append(filenameall[y])
    #print(all_pdbqtfiles)

#make path for each all_pdbqtfile paths, first get ZINC ID
then make string of numbers into list of single numbers, then
join with slash, then write in list
    ZINC=list()
    teilpath=list()
    for z in range(len(all_pdbqtfiles)):
     extract = all_pdbqtfiles[z].split('_')
     #print(extract)
     for i in range(len(extract)):
      if extract[i].startswith('ZINC'):
ZINC.append(extract[i])
    print(ZINC)
```

```
    for b in range(len(ZINC)):
     filenumber = re.findall(r'\d+', ZINC[b][:-5])
     #print (filenumber)
     realnumber = str(filenumber[0])
     #print (realnumber)
     singlenumbers=list(realnumber)
     joinnumbers= "/".join(singlenumbers)
     teilpath.append("/"+joinnumbers+"/")
    #print(teilpath)

#put root infront and make list all_files which contains the
paths to all the all_pdbqt files
    filepaths=list()
    all_files=list()
    for w in range(len(teilpath)):
     filepaths.append(os.getcwd() + teilpath[w])
    for c in range(len(filepaths)):
     all_files.append(filepaths[c] + all_pdbqtfiles[c])
    #print(filepaths)
    #print(all_files)

#open and read files
    all_firsts=list()
    for d in range(len(all_files)):
     fptr = open(all_files[d])
     alllines = fptr.readlines()
     fptr.close()
#find score and molecule number
#put score and molecule number in list then list
     nameline= alllines[2]
     scorelist=list()
     scorelist.append(nameline.split()[3])
     scoreline= alllines[1]
     #print(scoreline.split("      "))
     negative = scoreline.split("      ")[1]
     if negative.find("-")==0: #only negative results
      scorelist.append(negative)
     else:
      scorelist.append("0.0")
     #print(scorelist)
     all_firsts.append(scorelist)
    #print(all_firsts)
#sort list of lists for score
    sorted_list = sorted(all_firsts, key=lambda first:
first[1] , reverse= True)
    #print(sorted_list)

#print list of litst to result summary file
    user_input = sys.argv[1]
    #user_input2 = sys.argv[2]
    #user_input = input("Some input please: ")
    if int(user_input)>len(sorted_list):
     print('Insufficient data available, please choose a
smaller number of top results.')
    else:
# set up result_folder
```

```
     result_folder='Top_'+user_input+'_results_'+receptor_name
     if not os.path.exists(result_folder):
      os.makedirs(result_folder)

#write result summary file in result_folder
    #optr = open(result_folder+'/'+result_folder + '.txt',
'w+')
            #molctr += 1
     string_list=list()
     for i in range(len(sorted_list)):
      string_list.append(' '.join(sorted_list[i]))
      #print (string_list)
     for x in range(0,int(user_input)):
      line= string_list[x]
      with open(result_folder+'/'+result_folder + '.txt',
mode='a+', encoding='utf-8') as myfile:
        myfile.write(str(x) +' '+ line + '\n')

#write top results in individual .pdbqt files
     numbers_list=list(range(1,int(user_input)+1))
     string_numbers=list()
     #print(numbers_list)
     for i in range(len(numbers_list)):
      la=str(numbers_list[i])
      filename= result_folder+'/'+'Top_' + la + '_' +
sorted_list[i][0]+'.pdbqt'
      #print(filename)

# paths to top files
      for b in range(len(all_files)):
       if all_files[b][-18:].find(sorted_list[i]
[0]+'.pdbqt')==0:
        print(all_files[b])
        fptr = open(all_files[b])
        alllines = fptr.readlines()
        fptr.close()
        for e in range(len(alllines)):
         line = alllines[e]
         if line.find("MODEL 2")==0:break
         else:
          if not line.startswith("END"):
           optr = open(filename, 'a+')
           optr.write(line)
```

Appendix C

Chapter 5

C.1 Complete NMR Screening Results for Compound Selection

The full results of the NMR-based screens described in Chapter 5 are shown as chemical shift perturbations. These results are shown here for completeness.

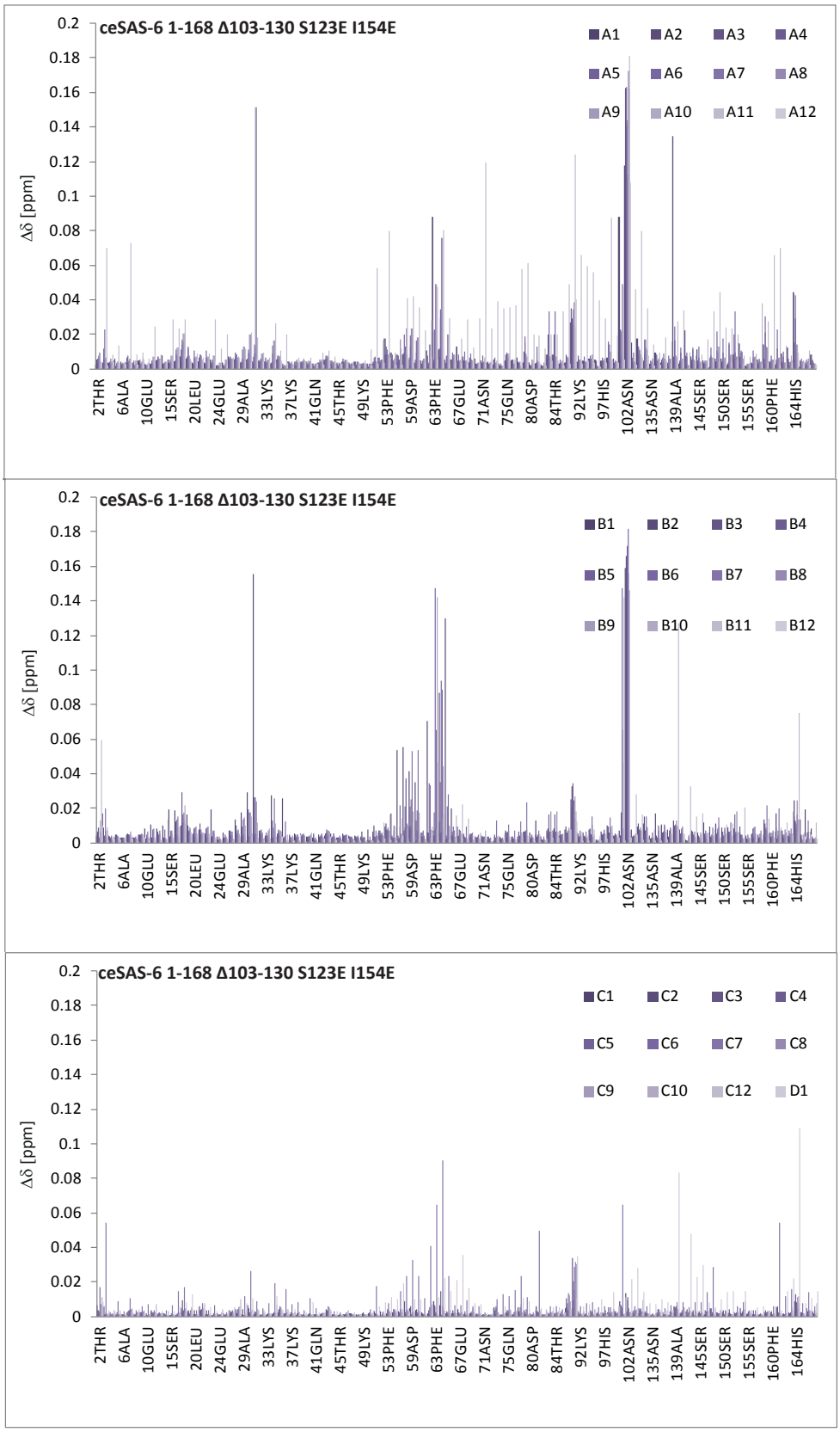


Fig. C.1: NMR Screening Results for Full Compound Selection - ceSAS-6

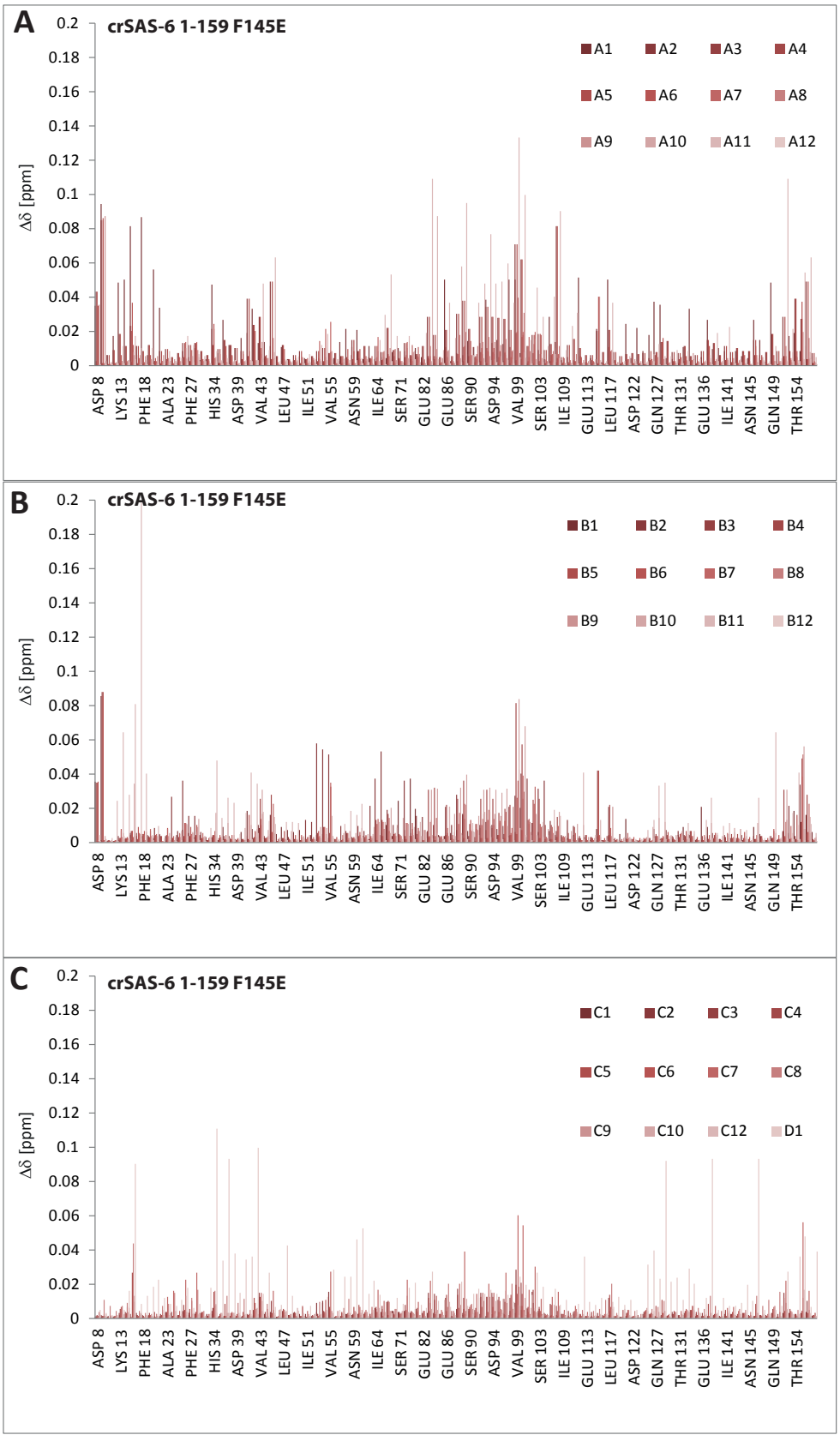


Fig. C.2: NMR Screening Results for Full Compound Selection - crSAS-6

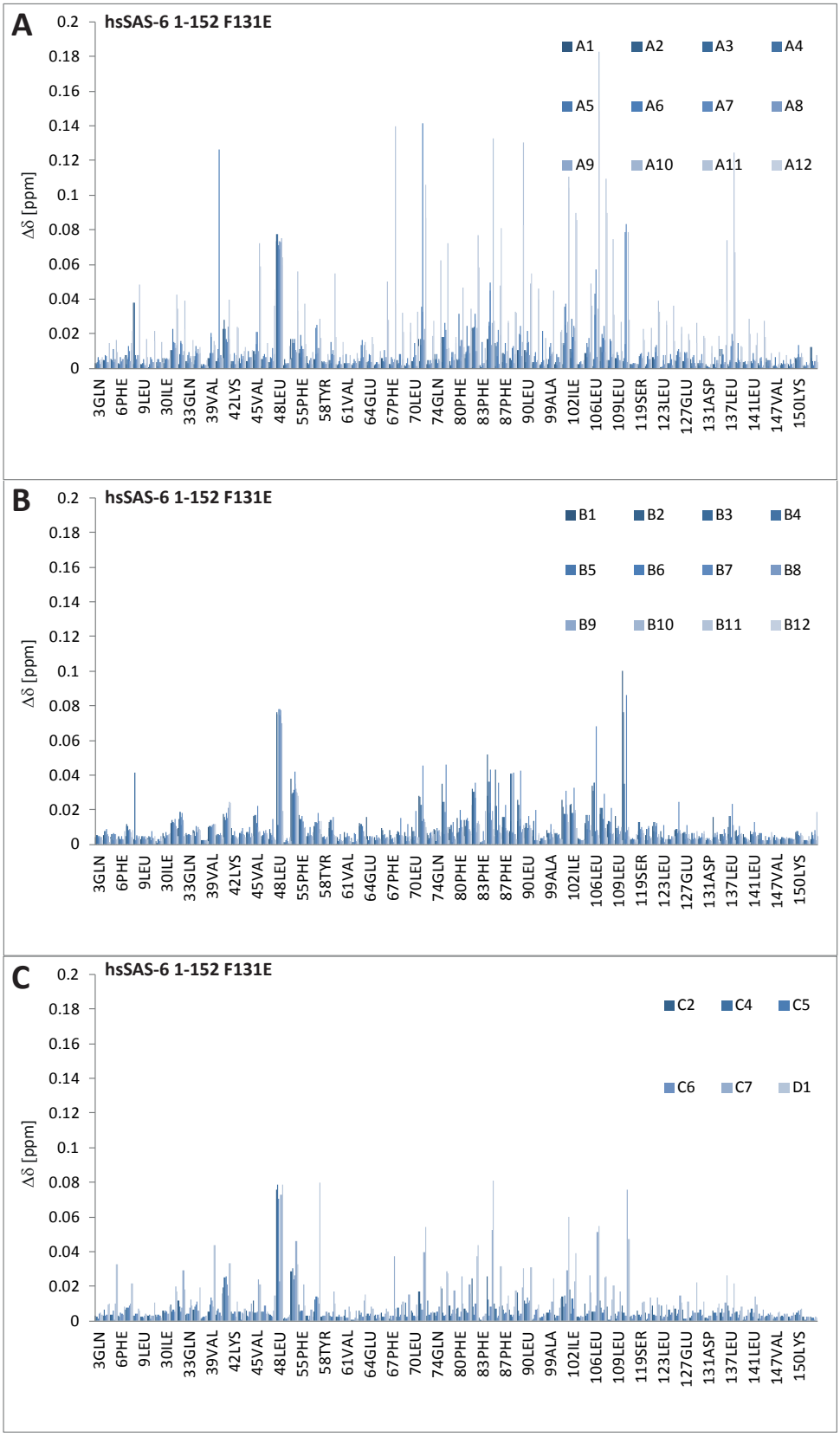


Fig. C.3: NMR Screening Results for Full Compound Selection - hsSAS-6

C.2 Small Molecule Identity

Compounds A11 and A12 were analysed structurally to verify the identity of commercially acquired compounds.

C.2.1 NMR Assignments for A11 and A12

Using 1D ^{1}H and 2D ^{13}C-^{1}H HSQC and HMBC NMR experiments, the composition of compounds A11 and A12 was studied. These experiments allow us to detect chemical groups and their bonds present in the compound structures. Figure C.4 and C.5 show the predicted and measured ^{1}H NMR spectra for A11 and A12. The predictions from ChemNMR (Perkin Elmer) based on the structure of the compounds correspond well with the measured spectra. 2D ^{13}C-^{1}H HSQC and HMBC NMR spectra were measured for A11 (Figure C.6) and A12 (Figure C.7), using natural abundance ^{13}C. Resonances can be assigned to individual carbon atoms in the chemical structures and compared to predicted values from ChemNMR. All carbon atoms present in the A11 and A12 structure could be assigned in the measured spectra.

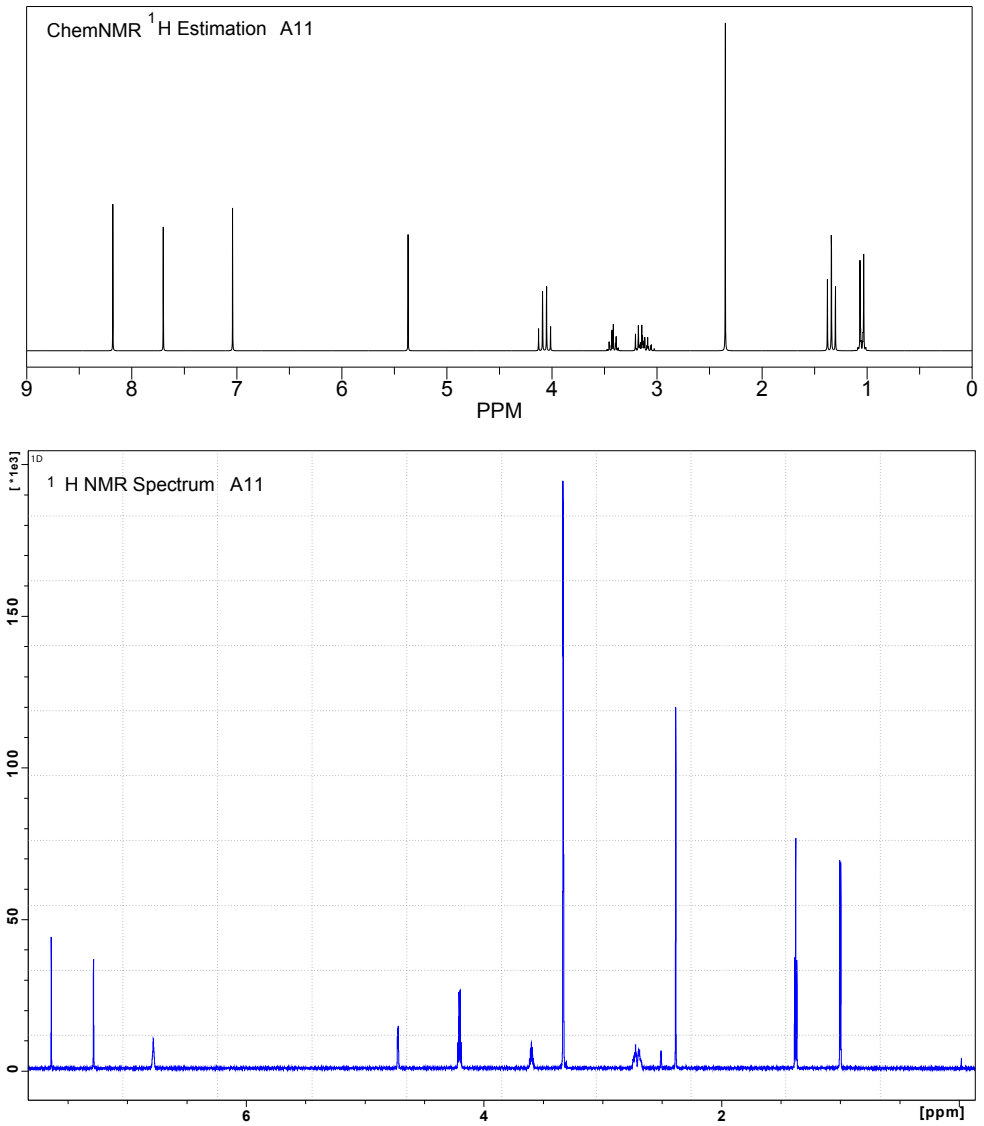


Fig. C.4: Small Molecule NMR Analysis of A11 - 1D NMR ^{1}H NMR

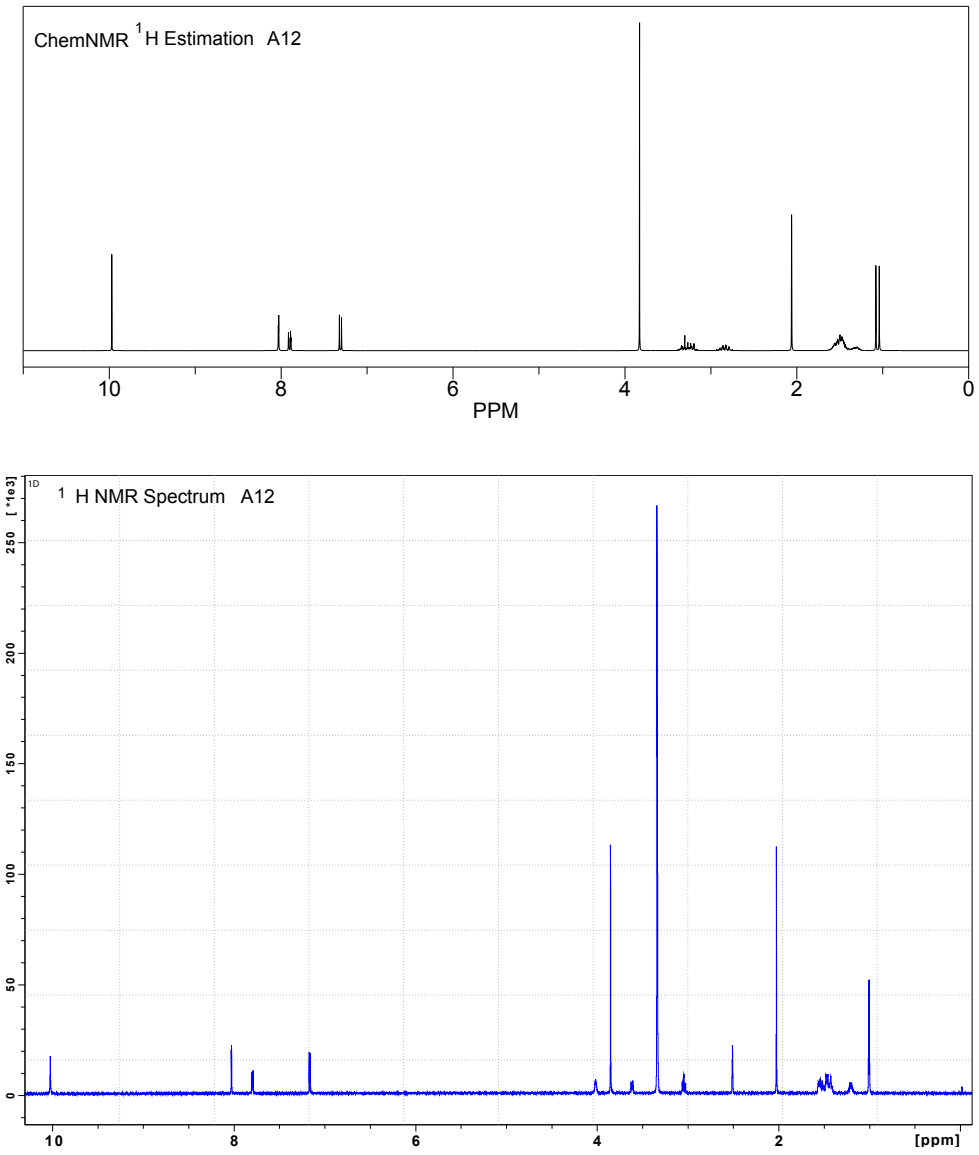


Fig. C.5: Small Molecule NMR Analysis of A12 - 1D NMR ^{1}H NMR

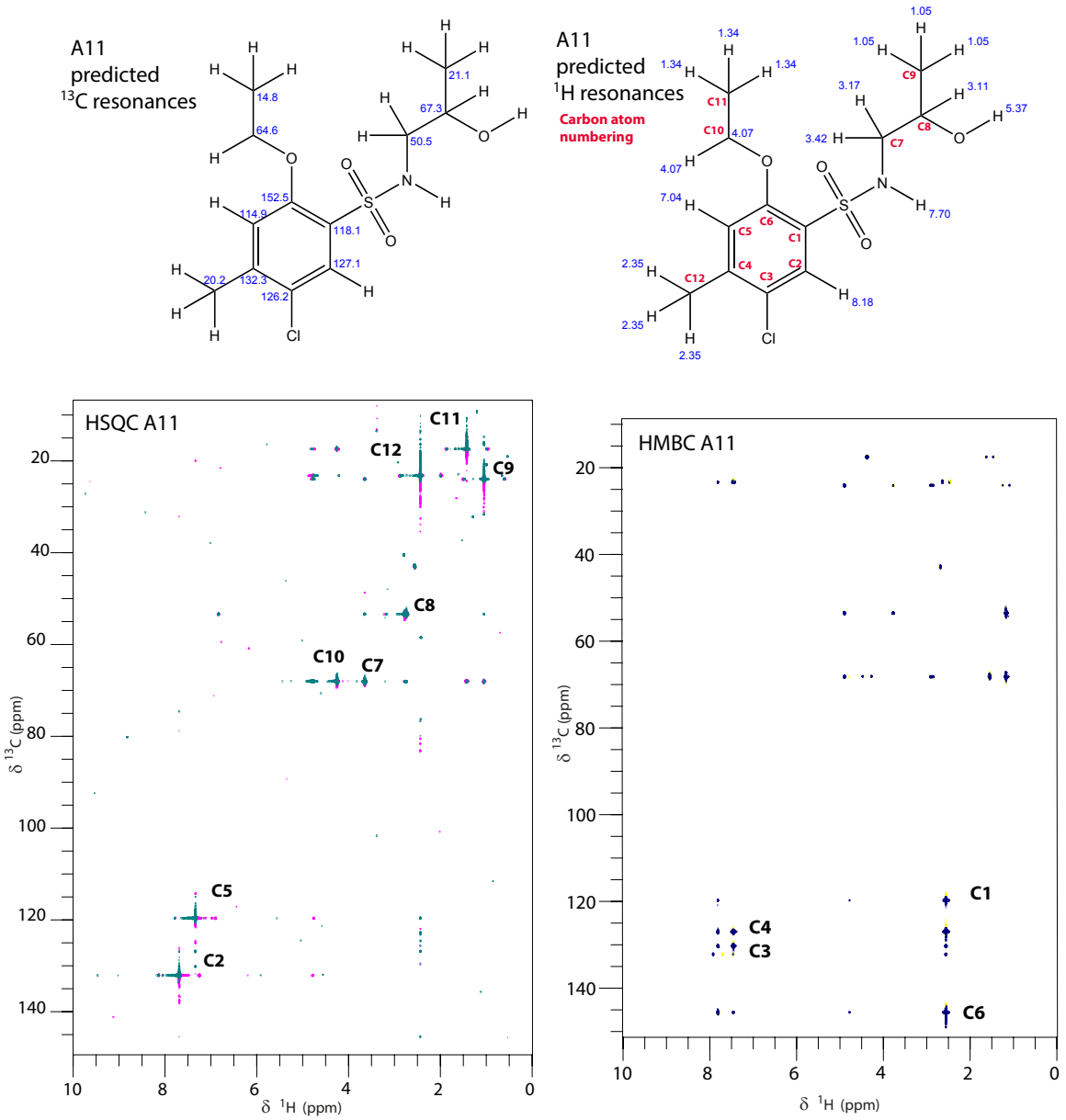


Fig. C.6: Small Molecule NMR Analysis of A11 - 2D NMR ^{13}C-^{1}H HSQC and HMBC

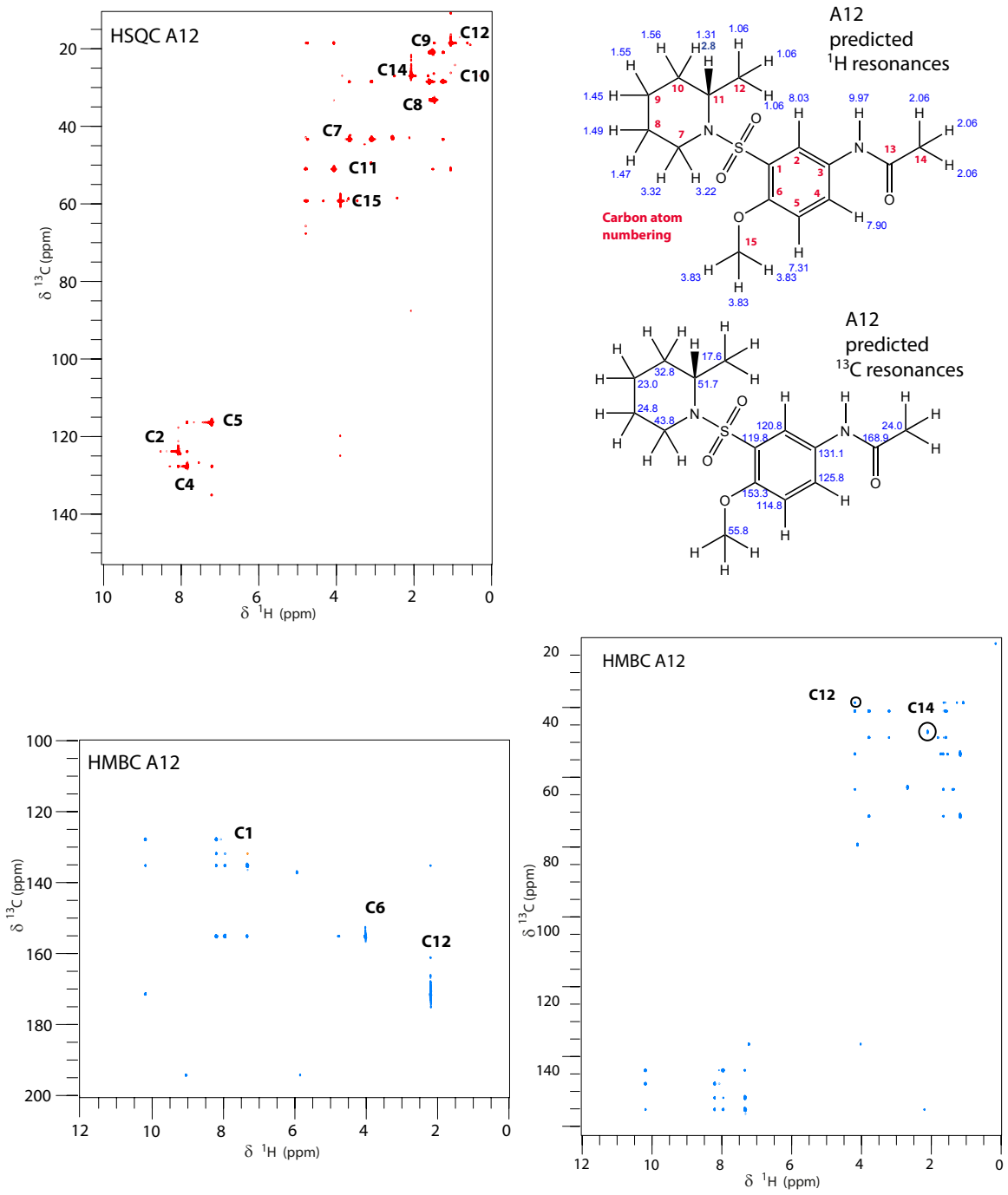
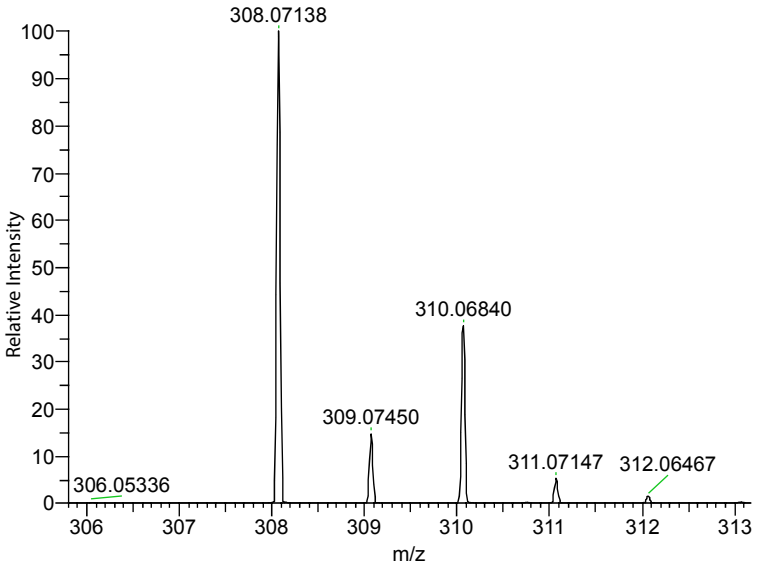
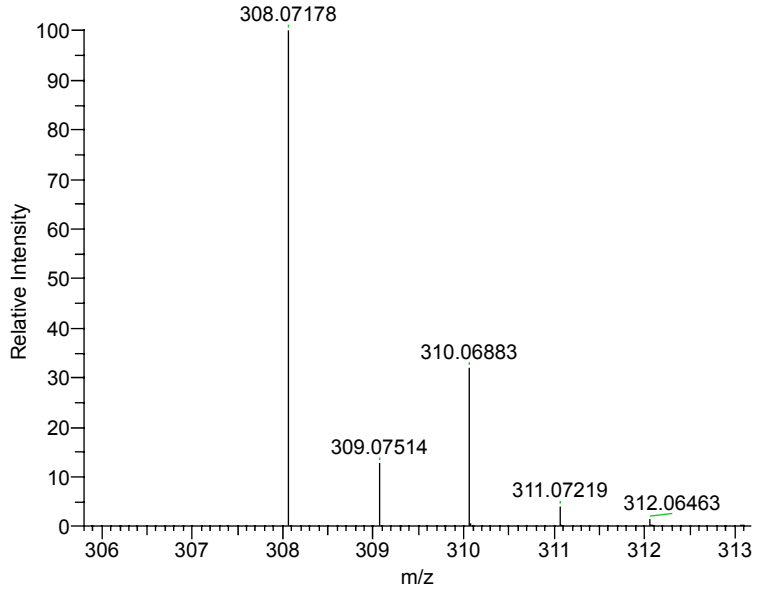


Fig. C.7: Small Molecule NMR Analysis of A12 - 2D NMR ^{13}C-^{1}H HSQC and HMBC

C.2.2 Mass Spectrometry Results for A11 and A12

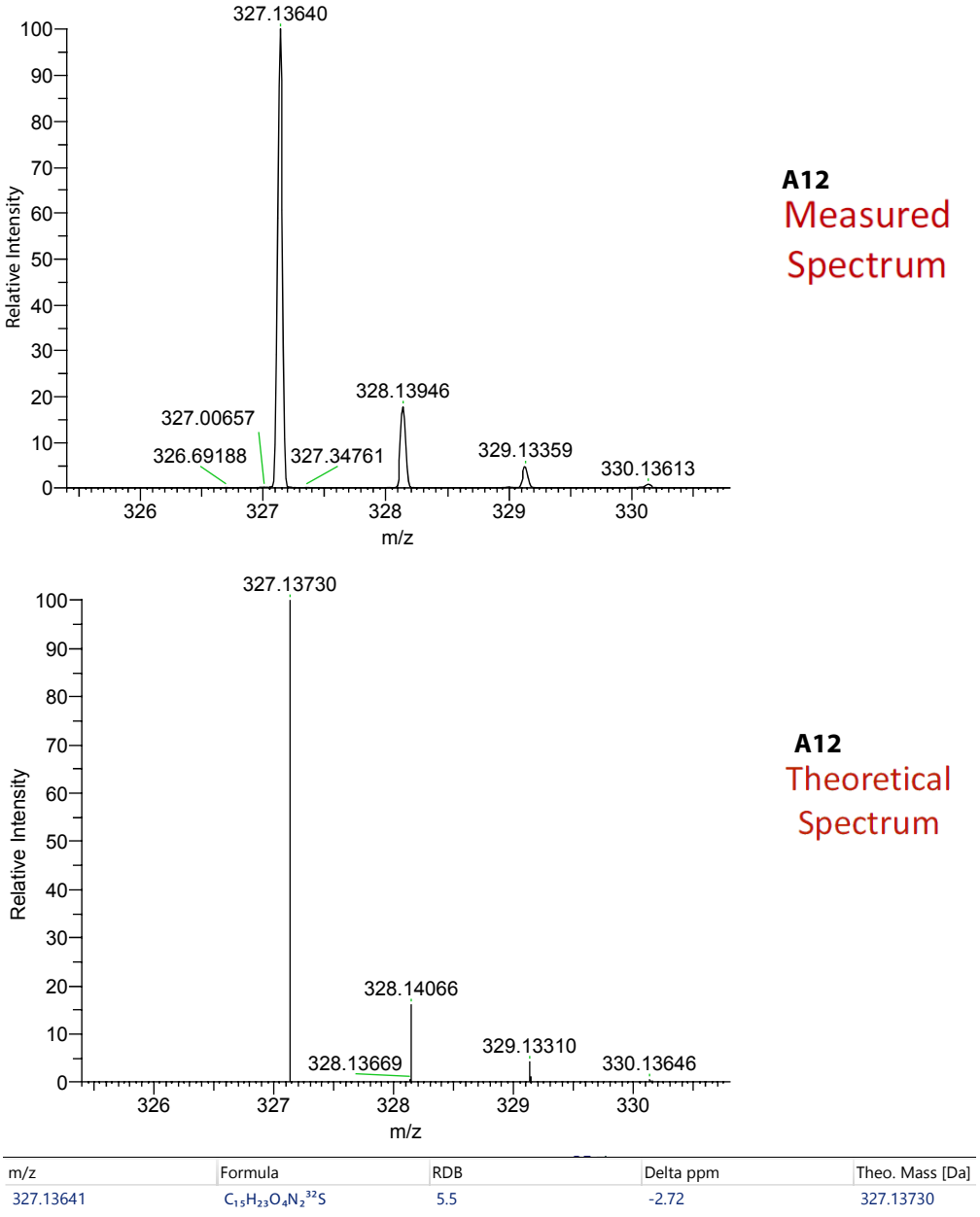
Using NMR all carbon atoms present in A11 and A12 were identified to be present. Mass spectrometry results shown in Figures C.8 and C.9 confirm the theoretical molecular mass of the compounds. This confirms that the compounds have the expected molecular formulae and accounts for the presence and identity of the expected non carbon atoms in the molecules. Mass spectrometry on A11 and A12 was conducted and analysed as part of the Mass Spectrometry Service at the Chemistry Department, University of Oxford by Dr. James Wickens.





m/z	Formula	RDB	Delta ppm	Theo. Mass [Da]
308.07138	$C_{12}H_{19}O_4N^{35}Cl^{32}S$	3.5	-1.31	308.07178

Fig. C.8: Mass Spectrometry Result for A11



m/z	Formula	RDB	Delta ppm	Theo. Mass [Da]
327.13641	$C_{15}H_{23}O_4N_2{}^{32}S$	5.5	-2.72	327.13730

Fig. C.9: Mass Spectrometry Result for A12

C.2.3 Small Molecule X-ray Crystallography of A11 and A12

Small molecules A11 and A12 were crystallised from 2 mM compound solution in 20 mM Tris pH 7.5 5 % DMSO in all PCT conditions (0.1 M TRIS hydrochloride pH 8.5, 2.0 M ammonium sulfate; 0.1 M TRIS hydrochloride pH 8.5, 1.0 M ammo-

nium sulfate; 0.1 M TRIS hydrochloride pH 8.5, 0.2 M magnesium chloride hexahydrate, 30 % w/v PEG 4,000; 0.1 M TRIS hydrochloride pH 8.5, 0.2 M magnesium chloride hexahydrate, 15 % w/v PEG 4,000). Needle and plate-like crystals formed after minutes of inoculation at 18 °C. X-ray diffraction data were collected at the Soleil Synchrotron, Saint-Aubin by Dr. Ioannis Vakonakis to 0.9 Å (A11) and 0.93 Å (A12) resolution. The data were phased using SHELX and SUPERFLIP software [186] by Dr. Pierre Legrand (Soleil Synchrotron, Saint-Aubin). In the interest of time, model building was not undertaken. However, the high resolution made it possible to identify individual atoms in the density visually. In combination with the previously described NMR and MS data, we are confident the compounds used are indeed of the assumed molecular structure. The molecule models depicted in Figure C.10 were positioned into the density manually for purposes of visualisation only and do not represent a structure model. Compound A11 crystallised in a $P2_1\ 2_1\ 2_1$ space group with one molecule per asymmetric unit and 4 asymmetric units per unit cell (unit cell dimensions [Å]: 8.84 12.84 24.87) (Figure C.10 A). Compound A12 crystallised from a racemic mixture with space group P -1 (P1 bar) with one molecule per asymmetric unit and two asymmetric units per unit cell (unit cell dimensions [Å]: 8.58 8.64 11.34).

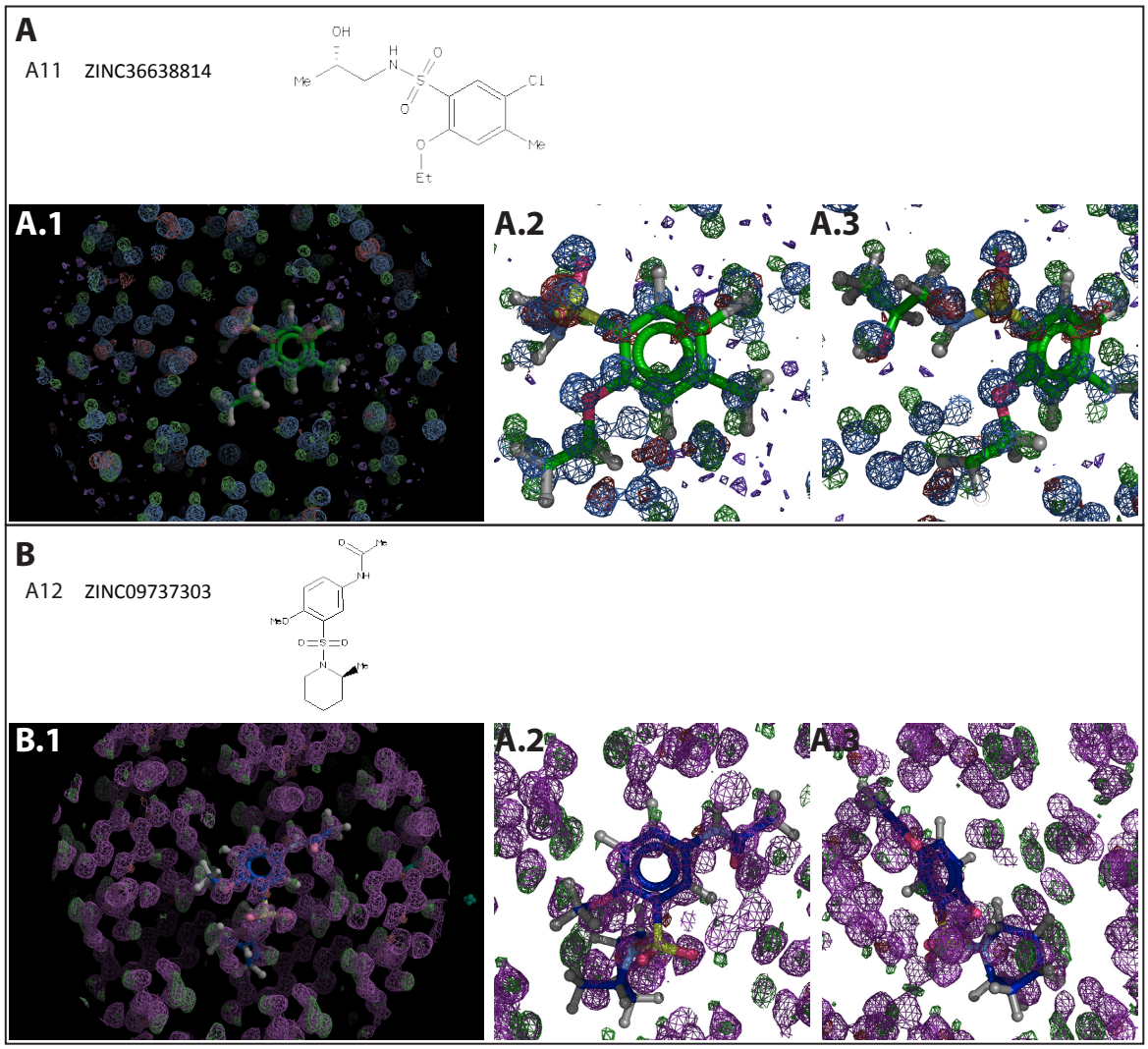


Fig. C.10: Small Molecule Crystallography A11 and A12: (A) Views of A11 compound crystallographic densities with manually added model of A11 for illustration. (B) Views of A12 compound crystallographic densities with manually added model of A12 for illustration.

Appendix D

Chapter 6

D.1 Crystal Morphology and Crystal Packing of hsSAS-6$_N$ F131E

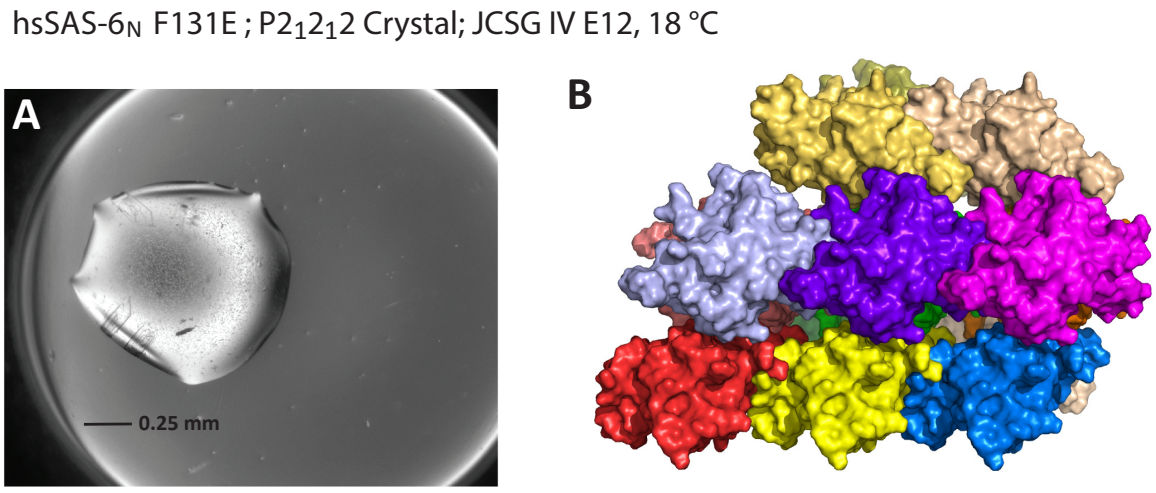


Fig. D.1: Crystal Morphology and Crystal Packing of hsSAS-6$_N$ F131E: (A) hsSAS-6$_N$ F131E crystal in screening condition JSCG IV E12 after 5 days at 18 °C. (B) hsSAS-6$_N$ F131E crystal packing with P2$_1$2$_1$2 space group.

Appendix E

Additional Work on Centriolar Protein STIL

In this section preliminary results on a separate centriole duplication related project are briefly described, which were generated as part of the work conducted in the pursuit of the degree of Doctor of Philosophy at the University of Oxford, by the author. This work is included for the sake of completeness and to create a record of the progress made.

E.1 Investigation into the Phosphorylation Dependent Interaction of STIL and the SAS-6 Coiled-Coil Domain

E.1.1 Introduction

The human protein STIL is essential for centriole duplication, but is poorly conserved between species [119]. STIL is considered to be the putative functional equivalent of SAS-5 (in *C. elegans*), and Ana2 in *D. melanogaster* [118, 120] (Figure E.1). Sequence similarity between these proteins is low and their molecular functions have only partially been determined. STIL, SAS-5 and Ana2 are essential for centriole formation

[118–120]. The proteins all localise to the centriole in a comparable manner [264], and they all contain a small coiled-coil motif and share interaction partners. Experiments using fluorescence recovery after photobleaching (FRAP) suggested that SAS-5 and STIL do not permanently localise at the forming centriole and therefore may not be an integral structural component [118, 119], but rather might have a shuttling function or involvement in regulatory processes. Mutations truncating the STIL protein in humans can be directly linked to autosomal recessive primary microcephaly, a neurodevelopmental disorder characterised by reduced head circumference and brain size [265]. Work in mouse and zebra fish models confirms that STIL defects affect cell numbers during neuronal development [266–268], which might be a connection to the brain development in microcephaly. Despite its medical relevance, our understanding of STIL structure, function and regulation remains limited.

STIL localises to the centriole by interacting with SAS-6, and experimental evidence for an interaction between STIL and SAS-6 suggests that STIL interacts with the SAS-6 coiled-coil region through the STAN motif [119]. Furthermore, it was recently found that the STIL STAN-SAS-6 interaction might be regulated by Plk4 [121–124]. Several phosphorylation sites on STIL - some of which are in the STAN motif - have been identified (Figure E.2). However, the exact phosphorylation state of STIL *in vivo* linked to Plk4 activity and the effect of phosphorylation in the STAN motif are not yet clear. Some predicted phosphorylation sites could be found in *in vitro* kinase assays and detected on extracted STIL, however published studies could not reach a consensus on the physiologically relevant phosphorylation sites. Plk4 inhibition or depletion lead to loss of STIL localistion to centrioles [122]. Mutation of five of the identified Plk4 phosphorylation sites in the STIL STAN domain to alanine residues also diminished STIL localisation to the centriole. However, the STAN motif was shown to not be required for STIL localisation to the centriole, as STILΔSTAN constructs still success-

fully localised at the centriole. These observations would suggest that the presence of the STAN motif in some manner inhibits the association of STIL with the centriole and that the phosphorylation overcomes this inhibition. STIL constructs with point mutations replacing putative Plk4 phosphorylation sites with glutamate residues to mimic phosphorylation were used to investigate the interaction and the relevance and function of phosphorylated STIL. Specifically, residues S1108 and S1116 were shown to be influential, as alanine mutations of these sites resulted in reduced centrosome amplification. The introduction of an S1108D mutation rescued this phenotype, however S1116D did not [122]. In co-immunoprecipitation experiments in this study a S1108A mutation was not identified as crucial for SAS-6 binding [121]. In another mutational scan using alanine substitution, S1061 and S1111 showed reduced centriole duplication, whereas a phosphomimicking mutation showed over amplification of centrioles in human cell culture [121].

This conflicting evidence on the effect and influence of individual phosphorylation sites on the STIL-SAS-6 binding lead us to believe that more investigation is needed. We aimed to investigate the STAN-SAS-6 interaction and find a biophysical assay to quantify the effects that specific phosphomimicking mutations would have on this interaction *in vitro*.

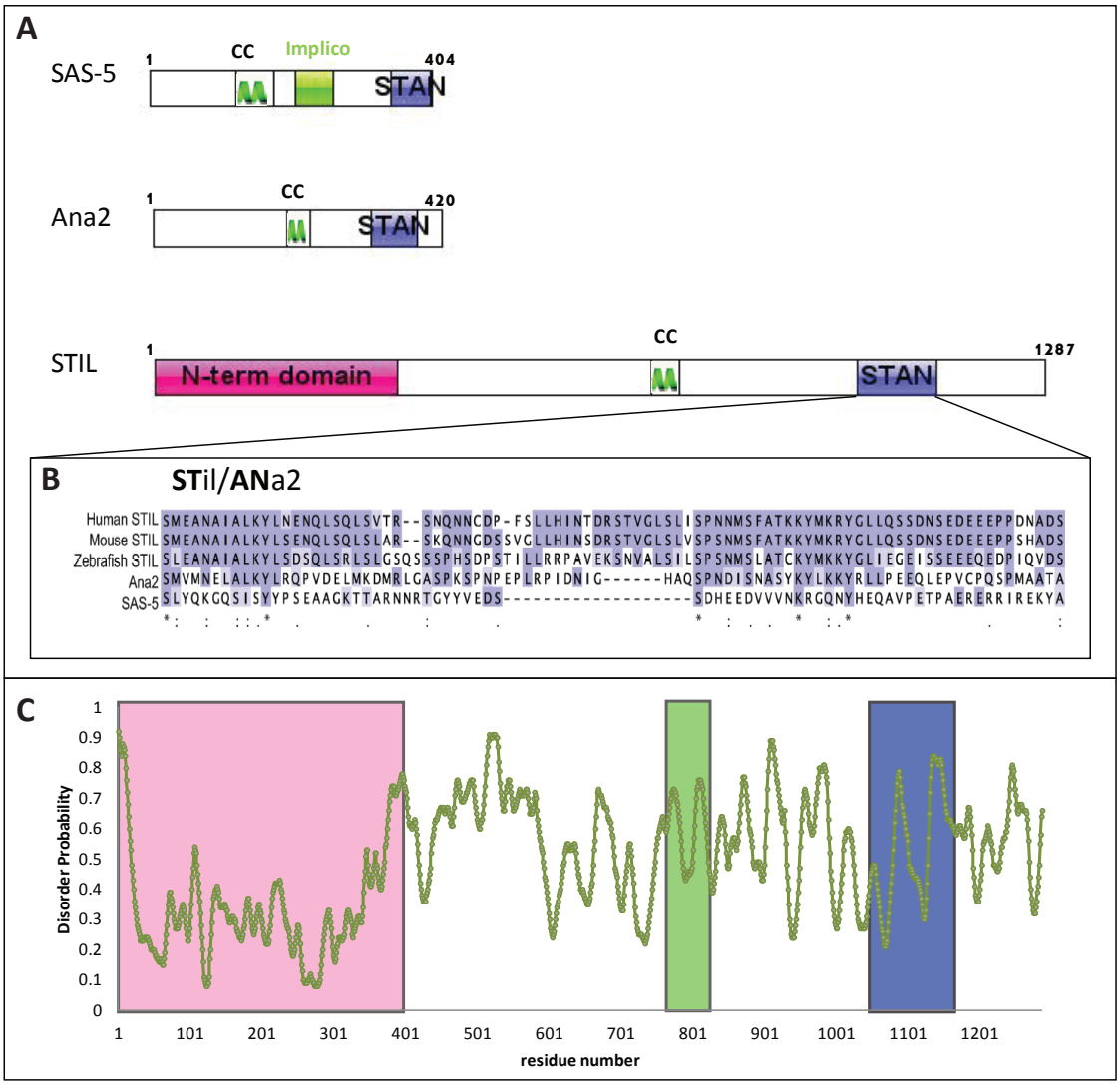


Fig. E.1: STIL Architecture and Conservation: (A) Schematic representation of STIL, SAS-5 and Ana2 architecture. Highlighted in *pink* is the folded N-terminal domain; highlighted in *green* is the short coiled-coil domain; highlighted in *blue* is the conserved STAN motif. (B) Sequence alignment of STAN domains. (C) Disorder prediction for hsSTIL.

S/T/Y in STIL STAN 1051-1147	S1051	S1054	S1061	Y1071	S1078	S1081	T1083	S1085	S1094	T1100	S1103	S1108	S1111	S1116	T1119	Y1122	Y1126	S1131	S1132	S1135	S1147
Predicted Plk Phosphorylation Sites	x	x			x	x	x				x	x	x		x			x	x	x	x
Collective Experimental Evidence for Effect on SAS-6 Interaction			X		X							X	X	X	X				X	X	
Construct 2E			x										x								
Construct 2E2												x	x	x							

Fig. E.2: Possible Phosphorylation Sites in STIL STAN: List of all serine, threonine and tyrosine residues in the hsSTIL STAN motif. Predicted Plk phosphorylation sites are indicated, as well as experimentally detected phosphorylation sites [121–124]. Residues mutated in construct STIL STAN 2E are highlighted in *green*. Residues mutated in construct STIL STAN 2E2 are highlighted in *yellow*.

E.1.2 STIL STAN and SAS-6 Coiled-Coil Constructs and Purification

Constructs of human STIL STAN were designed that were suitable for NMR and FP experiments (Table E.1]. Due to low expression yields of previous constructs, the genes were commercially codon optimised (Eurofins) for expression in *E. coli*. To ensure site directed labeling for FP experiments, residue C1090 was mutated to glycine. Two phosphomimicking constructs were designed to test if a difference in interaction with SAS-6 could be detected *in vitro*. Construct STIL STAN 2E incorporates two mutations of residues shown to have a positive effect on SAS-6 binding in phosphomimicking substitution (S1061E and S1111E). Construct 2E2 contains two phosphomimcking mutations of residues for which conflicting evidence exists (S1108E and S1116E). Two residues were chosen to increase potentially small effects.

Table E.1: STIL STAN and SAS-6$_{CC}$ Constructs

Protein Constructs				
Construct	Organism of Origin	Plasmid	Tag and Cleavage Site	Reference
STIL STAN wt (hsSTIL 1051-1147 C1090G)	*H. sapiens*	pfloat	His,HRV 3C	NEW
STIL STAN 2E (hsSTIL 1051-1147 C1090G S1061E S1111E)	*H. sapiens*	pfloat	His,HRV 3C	NEW
STIL STAN 2E2 (hsSTIL 1051-1147 C1090G S1108E S1116E)	*H. sapiens*	pfloat	His,HRV 3C	NEW
SAS-6 FL CC F145E (hsSAS-6 155-484)	*H. sapiens*	pHis	His, HRV 3C	
SAS-6 CC1 (hsSAS-6 155-271)	*H. sapiens*	pfloat	His,HRV 3C	NEW
SAS-6 CC3 (hsSAS-6 211-312)	*H. sapiens*	pfloat	His,HRV 3C	NEW
SAS-6 CC2 (hsSAS-6 310-439	*H. sapiens*	pfoalt	His,HRV 3C	NEW
SAS-6 CC4 (hsSAS-6 373-484)	*H. sapiens*	pfoalt	His,HRV 3C	NEW

The STIL STAN constructs were expressed with a histidine tag and were purified under denaturing condition in PBS with 8 M Urea and 500 mM NaCl during affinity chromatography. Talon metal affinity resin was incubated with the protein containing supernatant for 1 h at room temperature. The resin was then washed with 2 CV of PBS 8 M urea and 500 mM NaCl. The wash was repeated 10 times. Unspecific binding was removed with five pre-elution steps with 2 CV buffer containing 20 mM imidazole. Protein was eluted in 5 steps of 1 CV each using buffer containing 500 mM imidazole.

Protein containing fractions were identified by SDS PAGE and pooled and dialysed into refolding buffer (PBS, 150 mM NaCl, 1 mM DTT and 1 mM EDTA pH 7) at 4 °C over night. STIL STAN constructs were further purified on Superdex 75 column equilibrated in PBS with 150 mM NaCl, 1 mM DTT and 1 mM EDTA pH 7. The protein containing fractions were then concentrated for further use and concentration determined by a NanoDrop spectrophotometer at 280 nm wavelength. An additional step of purification was necessary for STIL STAN constructs. Ion exchange chromatography with a HiTrap Q column was used with 20 mM Tris and 150 mM NaCl equilibration buffer. Separation was achieved with a linear gradient from equilibration buffer to 20 mM Tris with 1 M NaCl.

SAS-6 coiled-coil constructs were purified according to the protein purification protocol described in Chapter 2. Due to the elongated shape of coiled-coil proteins, a Superose 6 column was used for SEC of the full length SAS-6 coiled-coil construct.

E.1.3 Localising STIL STAN Interaction Region on SAS-6 Coiled-Coil Domain

To test for the presence of an interaction between STIL and SAS-6, ^{15}N HSQC experiment with 30 µM ^{15}N STIL STAN 2E were conducted. The 2E construct was chosen for these experiments, as yield of wild type STIL STAN did not produce the protein quantities needed in the minimal medium required for the production of labeled protein. The spectrum of STIL STAN 2E shows a low range of resonances, characteristic of unfolded protein (Figure E.3 A). Upon addition of 53 μM SAS-6 FL CC, the spectrum shows a significant reduction in detected resonances (Figure E.3 B). Amino acids formerly flexible enough to produce measurable resonances now have transitioned into a different chemical exchange regime. This structural change indicates an interaction or restructuring of the region.

Shorter constructs of the long SAS-6 coiled-coil domain were produced to narrow down the binding site of STIL STAN on SAS-6. The coiled-coil constructs were designed using predictions of the α-helix heptad repeats, to ensure coiled-coil formation. Constructs SAS-6 CC1, CC2, CC3 and CC4 span the coiled-coil domain, with the CC1 construct incorporating the N-terminal boundary of the full length coiled-coil domain and CC4 ending with the C-terminal boundary of the SAS-6 coiled-coil domain. Some overlap between the construct was included. The constructs were characterised using CD measurements and dimer formation of the coiled-coil fragments was tested with SEC MALS.

The SAS-6 CC constructs 1-4 were tested for interaction with ^{15}N STIL STAN 2E in HSQC experiments as described above. Whereas SAS-6 CC1 to 3 did not cause changes in the resonances detected (Figure E.3 C, D and E), the addition of SAS-6 CC4 (residue 373-484) caused a similar change in resonances as did the addition of full length coiled-coil (Figure E.3 F). This suggests that the interaction of STIL STAN with the SAS-6 coiled-coil domain happens in the N-terminal fragment of this domain, specifically between hsSAS-6 residues 373-484.

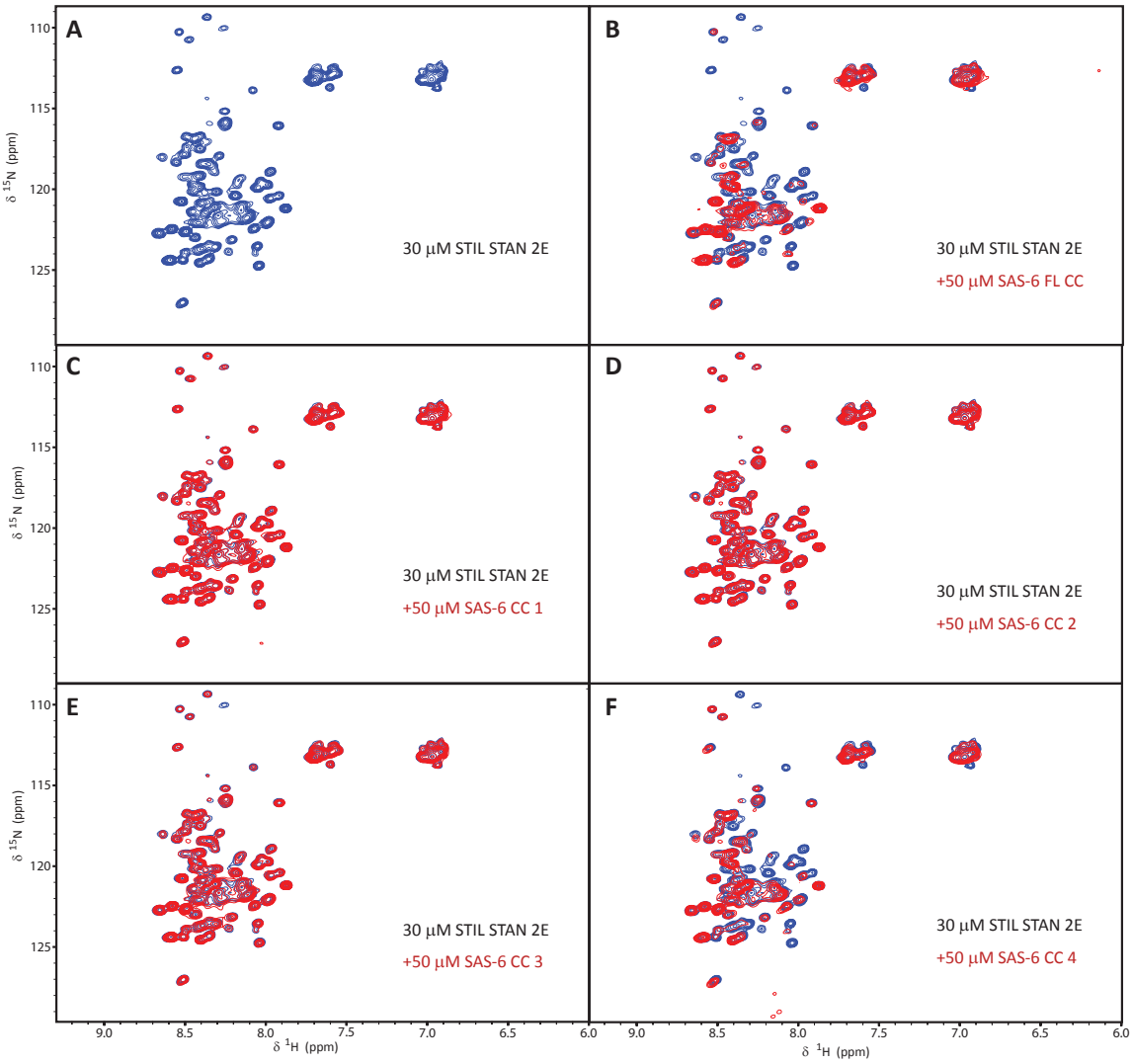


Fig. E.3: ^{15}N ^{1}H HSQC NMR - STIL STAN and SAS-6 Coiled-Coil Interaction: (A) NMR spectra of 30 μM ^{15}N labeled STIL STAN 2E *blue*. Overlay with *red*: (B) added 50 μM SAS-6 FL CC, (C) added 50 μM SAS-6 CC1, (D) added 50 μM SAS-6 CC2, (E) added 50 μM SAS-6 CC3, (F) added 50 μM SAS-6 CC4.

E.1.4 Development of Quantitative Interaction Assays for STIL STAN Phosphomimetic Constructs with SAS-6

An FP based interaction assay with 5-FAM labeled STIL STAN constructs was conducted to quantify the strength of the STIL STAN-SAS-6 CC4 interaction and to

compare the effect of the phosphomimicking mutant test constructs (Figure E.4). For the assay, a titration series of SAS-6 FL CC, SAS-6 CC1, SAS-6 CC4 and an unrelated protein for negative control (spectrin repeats 11-13) were added to the labeled STIL STAN proteins. The signal intensity in the experiments testing binding of STIL STAN wt construct was low, due to low purification yields (Figure E.4 A). The errors attached to data from these experiments are higher than in experiments with STIL STAN 2E and STIL STAN 2E2, which gave rise to a poor fit. The K_d for the interactions tested can only be estimated to be in the range of $\sim$ 20 - 70 μM for the interaction with SAS-6 FL CC and SAS-6 CC4. The purification of this protein would need to be further optimised and the experiment repeated to determine a reliable K_d value. There was no binding detected between STIL STAN wt and SAS-6 CC1 or in the negative control.

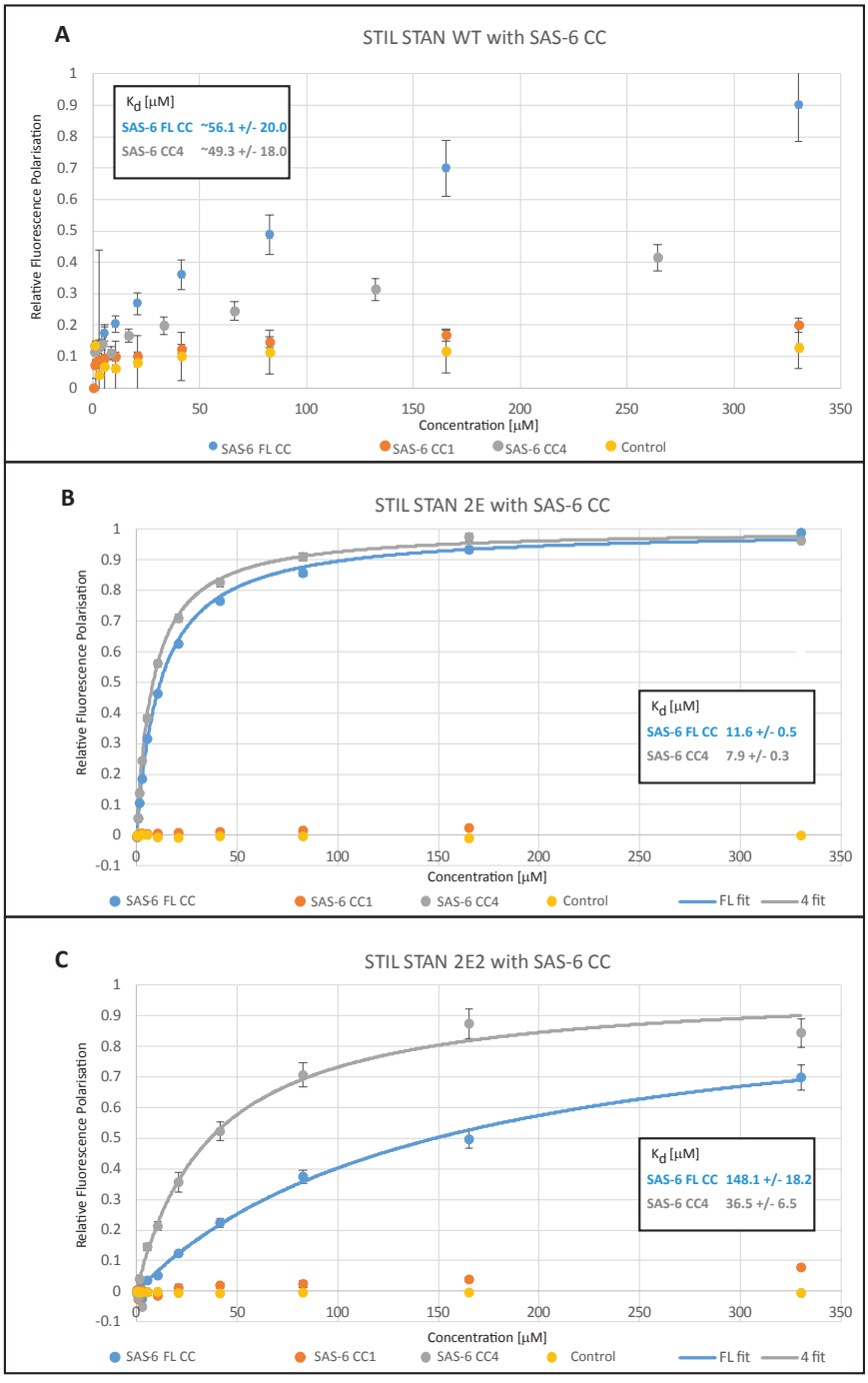


Fig. E.4: Interaction Assay for STIL STAN and SAS-6 Coiled-Coil: FP based interaction assay with fluorescence labeled STIL STAN constructs and titration series of SAS-6 FL CC (*light blue*, SAS-6 CC1 (*orange*), SAS-6 CC4 (*grey*) and control protein (spectrin repeats 11-13) for negative control (*yellow*). (A) Experiments conducted with STIL STAN wt construct, K_d only estimated. (B) Results shown for experiments with STIL STAN 2E construct. (C) Results shown for experiments with STIL STAN 2E2 construct.

Results from experiments conducted with STIL STAN 2E show clear interaction (Figure E.4 B). The titration fit performed suggests a K_d of 11.6±0.5 μM for the interaction with SAS-6 FL CC and 7.9±0.3 μM for the interaction with SAS-6 CC4. The results for STIL STAN 2E2 suggest a K_d of 148.1±18.2 μM for the interaction with SAS-6 FL CC and 36.5±6.5 μM for the interaction with SAS-6 CC4 (Figure E.4 C). As expected, no binding was detected between either STIL STAN 2E or STIL STAN 2E2 and SAS-6 CC1, or the negative control protein. In our experiments all constructs showed better binding to SAS-6 CC4 than to SAS-6 FL CC, which might have something to do with the mechanism of binding. This would need to be further examined. The binding of STIL STAN 2E to both SAS-6 FL CC and SAS-6 CC4 was strongest, compared to STIL STAN 2E and STIL STAN 2E2. Even with the range of K_d estimated for STIL STAN wt binding to SAS-6 FL CC and SAS-6 CC4, it appears that the phosphomimicking mutation in positions S1061E S1111E (2E) has a strengthening effect on the interaction with SAS-6, whereas the S1108E S1116E (2E2) mutation had a weakening effect on binding. This might explain some of the obsevations made with phosphomimicking mutations at these positions in *in cellulo* experiments mentioned earlier, where S1061D showed improved localisation of STIL to the centriole even in the absence of Plk4, whereas S1108A and S1108D showed the same lack of STIL localisation. This could suggest that the chemical difference between true phosphorylation and phosphomimicking mutations is too significant to replicate the interactions that phophorylated STIL would form. If the observed difference in binding behaviour of the site specifically mutated constructs is not an artifact, this could hint to a tunable binding between SAS-6 and STIL, regulated by the phsophorylation state of STIL. Further experiments would be needed to study the observed phenomena.

E.2 Conclusion

Our experiments showed that the *H. sapiens* STIL STAN domain interacts with a C-terminal region of the *H. sapiens* SAS-6 coiled-coil domain. Specifically, that the interaction region on the SAS-6 coiled-coil is located between residues 373-484. Interaction assays based on FP confirmed the binding of STIL STAN to the full length SAS-6 coiled coil and SAS-6 coiled-coil fragment 373-484. None of the tested STIL STAN constructs showed binding to SAS-6 CC1; all of them showed binding to SAS-6 FL CC and SAS-6 CC4, which supports the observations made in NMR experiments. We found that the effect of phosphomimicking mutations on the binding of STIL STAN to SAS-6 CC is site specific and that using FP based assays, the effect could be quantified. These are merely preliminary results and thorough research and optimisation are needed to unravel the many implications that these observations could have.