Mechanistic Insights on Protein Aggregation and Phase Separation



Virgínia Casablancas Antràs

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2023

This thesis is dedicated to all my loved ones
who dared not only ask how it was going
but listened to the answer.

Acknowledgements

I would like to first acknowledge my supervisor, Andy Baldwin, for his guidance and support throughout the course of my DPhil. I appreciate his encouragement in branching out of my comfort zone and doing research using a physical chemistry approach, which having trained as a biomedical scientist, was definitely challenging. He trusted me in being able to learn the necessary skills and personally taught me many of them, which I'm grateful for, as well as inspired me with his scientific mindset.

I would next like to thank our industrial collaborators at Eli Lilly, for their contributions to the work presented and their mentorship, and in particular Dr. Gary Sharman, Dr. Suchira Bose and Dr. Haitao Hu, who have acted as my industrial supervisors. I consider myself fortunate to have experienced research at the interface of academia and industry, learning about the unique strengths of each and how to find productive synergies. I would also like to thank the remaining collaborators whose work is included in this thesis (chapter 5): Satoshi Kishigami, James Eaton, Richard Wheeler and Spencer Anthony Cahill. I would like to acknowledge excellent proofreading and feedback by colleagues and friends: Susanne Mesoy, Roi Asor, Joan Claramunt, Abi Turner, Neha Kalmankar, Eugene Lin, James Eaton, Carla Kirschbaum, Gogulan Karunanithy, Edmond Toma, Wiktoria Sadowska, Eric Mates and Grace McDonald.

I am grateful for DPhil funding from the EPSRC (EPSRC and MRC Centre for Doctoral Training in Systems Approaches to Biomedical Science - grant number EP/L016044/1) and Eli Lilly. I would like to acknowledge additional financial support from the COVID extension fund, the ERC under the European Union's Horizon 2020 research and innovation programme (grant agreement no. 101002859) and a Eurofins Foundation award.

I would like to thank the SABS course directors, Garrett Morris and Charlotte Deane, for selecting me for this DPhil programme and coordinating a range of very interesting courses and lab experiences during my first year. In the course of my DPhil, parts of this thesis were presented at the milestones of transfer of status and confirmation of status. I would like to thank Justin Benesch, Grant Richie, Carol Robinson and William Barford for their critical assessment and helpful suggestions in these examinations.

It has been a privilege to be part of the Physical Theoretical Chemistry Laboratory and the Kavli Institute for Nanoscience Discovery, where I have found a collegial supportive environment to discuss my research with others and learn about exciting projects in many distinct fields. I would like to start by thanking and acknowledging former and current lab managers Vicky Cullen and Andy Dolan: their excellent organisation and maintenance of our research space contributed greatly to my experimental results - thank you also for the moral support! I would like to acknowledge the members of the Baldwin lab, past and present, in particular Gogs, James and Charlie for their help with learning NMR, and their friendship. Special mention goes to Satoshi, my "partner in submission": we have really bonded over the challenges and successes of this stage. I also thank the members of the Benesch, Robinson, Rauschenbach, Kukura and Struwe group, past and present. I have always found help and suggestions in my interactions with them, from specific lab procedures to the bigger picture of the research projects discussed. In particular, I'd like to thank Mark Agasid, Jani Bolla, Abraham Oluwole and Dominik Saman for refreshing and improving my protein purification skills at the beginning of my DPhil. And to my angels (Daniel, Hannah, Maya, Neha and Susanne): thank you for your friendship and making lunch a highlight of my day.

Oxford has been a great place to carry out this part of my education, and I have found amazing communities that have enriched my experience and supported me through the challenging stages of my DPhil; special mention goes to my DTC cohort (Conor, Grace, Mihaela and Zuzana), friends at St Anne's College (Brittany, Cassie, Dan, Eirini, Nesrine,Valeria and Vassilios), teammates at the Oxford University Table Tennis team (Ali, Jinlin, Julia and Kin), colleagues at the Society of Spanish Researchers in the UK, and fellow amateur musicians of the Oxford Millenium Orchestra. I have been fortunate to find many friendships along the way, and would like to thank as well Eddie, Ela, Irene, Frankie, Galina, Helena, Pablo, Melissa, Roger and Vicky.

I have been away from my hometown, Sabadell, for the course of this degree, but I have always felt my friends and family very close: I would like to thank them for encouraging me to pursue this adventure and making an effort to stay connected. Gràcies als amics de tota la vida: els xerinolos (Albert, David,Joan,Lluís,Marta, Marina, Miquel,Mireia i Núria), "Els Notes" (Aida,Eric, Maria, Sílvia), les xaxipirulis (Bea, Èlia, Laura, Qian), Àngels,Ari, Joan, Laia,Liz, Montse, Mireia i Neus.

Finally, I'd like to thank my mum, dad and grandma, for their support and interest in my research. Even though my family does not (yet) have any physical chemist in it, I believe that some of the many qualities they have passed to me have been specially instrumental in this DPhil: from my dad, his wide-ranging curiosity and academic rigour; from my mum, her patience and logical reasoning; and last but not least, from my grandma, her personal and professional entrepreneurial spirit.

Contents

List of Figures **xiii**

List of Tables **xvi**

List of Abbreviations **xvii**

1 Background **1**

1.1 Introduction . . . 1

1.2 Protein folding, misfolding and aggregation . . . 1

1.2.1 Functional protein organisation at several scales . . . 2

1.2.2 Protein aggregation and disease . . . 4

Protein aggregates: a structural view . . . 4

Metastability of the amyloid state . . . 6

Kinetics of protein aggregation . . . 7

Challenges in the study of protein aggregation kinetics . . . 8

1.3 Tau biology and pathology . . . 11

1.3.1 Domain organisation of tau . . . 11

1.3.2 Functions of tau . . . 13

1.3.3 Tau pathology . . . 13

1.3.4 Mechanisms of tau aggregation . . . 15

1.3.5 Structural transitions of tau involved in aggregation . . . 20

1.3.6 Aggregation modulating inhibitors of tau for AD . . . 22

Immunotherapy . . . 23

Small molecules . . . 24

Peptides . . . 25

1.3.7 How well do we understand the mechanism of action of tau and of tau-targeting inhibitors? . . . 29

1.4 Protein phase separation . . . 30

1.4.1 Overview of membraneless organelles . . . 30

1.4.2 Importance of disorder for phase separation . . . 32

1.4.3 Functions of membraneless organelles . . . 32

1.4.4 Interactions that drive phase separation . . . 33

1.4.5 Factors that affect phase separation . . . 37
1.4.6 Links between phase separation and disease . . . 39
1.4.7 Can we modulate phase separation? . . . 39
1.5 Nuclear Magnetic Resonance . . . 40
1.5.1 The 1D experiment . . . 40
1.5.2 Diffusion NMR . . . 43
1.5.3 Chemical exchange . . . 45
1.5.4 Applications of solution NMR to the study of protein aggregation 45
Summary of advantages of NMR versus other techniques for the study of protein aggregation . . . 46
1.5.5 Applications of NMR to the study of phase separation . . . 47
1.6 Thesis overview . . . 47

2 R3 as a model system for tau aggregation 48
2.1 Abstract . . . 48
2.2 Introduction . . . 49
2.3 Methods . . . 50
2.3.1 Contributions . . . 50
2.3.2 Materials . . . 50
2.3.3 Thioflavin T monitored kinetic assays . . . 51
2.3.4 NMR . . . 51
Assignment of R3 peptide . . . 51
Experimental kinetic run set-up . . . 51
Aggregation assays . . . 52
Kinetic data processing . . . 52
Diffusion . . . 53
2.3.5 Seeding with aggregated R3 peptide in a cell-based assay . . 54
Seed preparation . . . 54
Cellular assay . . . 54
Immunofluorescence . . . 54
2.4 Results and discussion . . . 55
2.4.1 R3 and R3R4 self aggregate in PBS . . . 55
2.4.2 Aggregation of R3 followed by NMR . . . 56
2.4.3 R3 stock samples are disordered and contain monomeric and dimeric species . . . 56
2.4.4 R3R4 doesn't aggregate in non-shaking conditions . . . 58
2.4.5 The aggregation of R3 in non shaking conditions is reproducible 59
2.4.6 The formation of dimer occurs in oxidising conditions and is essential for primary aggregation of both R3 and R3R4 . . . 61
2.4.7 R3 can seed full length tau in cells . . . 64
2.5 Conclusions . . . 65

3 Development of a modelling strategy **68**
3.1 Abstract . . . 68
3.2 Introduction . . . 70
3.2.1 Model space challenge . . . 70
3.2.2 Model parameter optimisation . . . 72
Numerical integration methods . . . 73
3.3 Methods . . . 75
3.3.1 Contributions . . . 75
3.3.2 Representative model for tests . . . 75
3.3.3 Settings and timing of numerical integration simulations . . 76
3.4 Results . . . 77
3.4.1 Description of the algorithm . . . 77
Example modelling pathways from simulations . . . 81
3.4.2 Evaluation of the performance of several ODE solvers . . . 81
Simulation of kinetic traces . . . 81
Performance of numerical methods in LMA optimisation . . 85
3.4.3 Model fitting and comparison of model's goodness of fit . . . 86
Model fitting . . . 86
When to stop increasing complexity . . . 87
3.4.4 Global optimisation settings . . . 92
Parameter initial guesses . . . 92
Perturbation of parameters in basin-hopping . . . 93
3.5 Conclusions . . . 94

4 Mechanism of aggregation and inhibition of tau R3 **97**
4.1 Abstract . . . 97
4.2 Introduction . . . 98
4.2.1 Inhibition of tau by D-TLKIVW . . . 99
4.2.2 Summary of qualitative findings on R3 aggregation from the previous chapter . . . 99
4.3 Methods . . . 101
4.3.1 Experimental . . . 101
4.3.2 Computational . . . 101
Initial rates fitting . . . 101
Overview of model exploration strategy . . . 102
Model Optimisation . . . 104
4.3.3 Composition of the chemical steps pool . . . 112
Protein-only pool . . . 112
Inhibitor pool . . . 113

4.3.4 Mathematical approximations . . . 113
Single-step steady-state nucleation reaction . . . 113
Mass balance considerations for fragmentation reactions . . . 119
Surface binding of inhibitor . . . 122
Mass balance corrections for elongation reactions in models including surface binding of inhibitor . . . 125
4.4 Results and Discussion . . . 129
4.4.1 Mechanism of R3 aggregation . . . 129
Concentration dependency of the monomer initial decay rates 129
Modelling results . . . 131
Comparison of nucleation hypotheses at two stages of the modelling . . . 137
Discussion of the first dataset winning model . . . 140
Analysis of the second dataset reveals a monomer conformational change prior to dimerisation . . . 144
Discussion of the second dataset model . . . 153
Comparison of the first and second dataset models . . . 153
4.4.2 Mechanism of action of D-TLKIVW on R3 aggregation . . . 155
General features of R3 aggregation in the presence of D-TLKIVW . . . 155
The mechanism of inhibition of D-TLKIVW is multifaceted and affects multiple species in the pathway to aggregates 158
Discussion of the inhibition mechanism . . . 178
4.4.3 Integration of the kinetic model with structural information 180
4.4.4 Discussion of the performance of the modelling algorithm . . 180
4.5 Conclusions . . . 182
4.6 Further work . . . 184

5 Partitioning of small molecules in biomolecular condensates 187
5.1 Abstract . . . 188
5.2 Introduction . . . 189
5.2.1 FUS . . . 189
Overview . . . 189
Phase separation of FUS . . . 190
Modulation of stress granules and FUS condensates by lipoamide 192
5.2.2 DDX4 . . . 193
Features of DDX4 phase separation . . . 193
Chemical environment of DDX4 . . . 194
5.2.3 Motivation for the partitioning experiments into DDX4 *in vitro* condensates . . . 196

5.2.4 Nucleotide effects on phase separation . . . 196
5.2.5 Literature evidence of chemical features influencing nucleotide-condensate interactions . . . 197
Interactions between ATP and phase separating proteins . . 197
Importance of the triphosphate moiety of ATP . . . 198
Importance of the nucleobase moiety of ATP . . . 199
Requirement of the nucleobase of ATP for its effects on proteins 202
What is the mechanism of action of nucleic acids on phase separation? . . . 203
5.3 Contributions . . . 203
5.4 Methods . . . 204
5.4.1 Protein expression . . . 204
5.4.2 Partitioning of small molecules in condensates . . . 205
FUS-Lipoamide . . . 205
Relative partitioning measurements of ATP, cAMP, GTP and cGMP in DDX4^{N1}NoCys . . . 209
5.4.3 Computational analysis . . . 210
5.4.4 Comparison of methods to determine pellet volume . . . 212
5.4.5 Measurements of absolute partitioning of ATP, cAMP, GTP, cGMP and deoxyribonucleosides . . . 213
5.5 Results and Discussion . . . 216
5.5.1 Lipoamide partitions into *in vitro* FUS condensates . . . 216
5.5.2 Development of a semi automated pipeline for analysis of partitioning: test case results . . . 217
5.5.3 Partitioning into DDX4^{N1}NoCys is affected by the nucleobase (G vs A) and the triphosphate vs cyclic moiety of a set of nucleotides . . . 218
5.5.4 Development of a PCV tube based method for determining partition coefficient . . . 222
5.5.5 Comparison of partitioning trends of thiamine, GTP, cGMP, ATP and cAMP in DDX4 depending on construct and pH . 225
5.5.6 Purine deoxyribonucleosides partition more than pyrimidines into DDX4^{N1} WT . . . 227
5.5.7 Rationalisation of the partitioning trends as a reflection of protein-nucleotide interactions in the condensate phase . . . 229
Cation-π interactions . . . 231
π-π interactions . . . 232
Hydrogen bonds . . . 233
Multiple interactions . . . 236

Considerations on the correlations between partitioning trends and interaction trends 236
5.6 Conclusions 237
5.7 Further work 240
Method development 240
Partitioning trends 241

6 Conclusions **243**
6.1 Mechanism of aggregation and inhibition of tau 243
6.2 Probing the chemical environment of *in vitro* biomolecular condensates 245

Appendices

A Appendix **249**
A.1 Full implementation of best R3 + D-TLKIVW mechanistic model . 249
A.2 Fitted parameters for best models 260

References **262**

Abstract

Proteins are essential structural and functional components of cells and act at a range of assembly states, from single proteins to macromolecular complexes. The pathological aggregation of proteins into amyloid fibrils and other morphologies of aggregates is a hallmark of several human diseases, many of which are neurodegenerative disorders like Alzheimer's and Parkinson's and pose a global health burden due the limited therapeutic approaches.

Proteins carry out their function not only in very well defined structured states, which is the basis of the structure-function dogma, but also by exploiting the unique properties of conformationally heterogeneous disordered states. In particular, the recently discovered phenomenon of protein phase separation, in which proteins and nucleic acids form a distinct phase (a membraneless organelle or biomolecular condensate) from the surrounding bulk environment, is emerging as a key compartmentalisation mechanism in cells, alternative to membrane-bound compartments, with a wide range of functions and unique properties.

In this thesis, I present the results of applying biophysical approaches, mainly nuclear magnetic resonance (NMR) and modelling, to the processes of protein aggregation and phase separation, with the aim to gain a better understanding of those at the molecular level. Chapter 1 introduces the topics of study, covering the fundamental aspects as well as the specific systems of study in more detail.

In chapters 2 through 4, I present the application of a recently developed NMR-based chemical kinetics method to the mechanism of aggregation of the protein tau and its inhibition by a peptide. In chapter 2, I present the R3 model system of tau as a biologically relevant model for tau which is amenable to our method. Chapter 3 is concerned with the mathematical modelling challenges and presents further developments to the initial method. Finally, in chapter 4, I present the mechanisms of aggregation and inhibition obtained by applying this modelling strategy to R3 and the peptide D-TLKIVW. The primary aggregation of tau R3 includes dimerisation and the formation of a tetrameric nucleus that elongates into fibrils, which can fragment. We find that D-TLKIVW, originally developed as an elongation blocker, has a more nuanced mode of action than expected and targets in particular the nucleus.

In chapter 5, I focus on the phenomenon of phase separation and in particular the hypothesis that condensates constitute a distinct chemical environment, a

solvent. We investigate this by studying the partitioning of small molecules into condensates of two phase separating proteins, DDX4 and FUS. The results include the development of several NMR-based methods to measure partitioning, as well as reveal chemical trends in the relative partitioning of a set of nucleotides.

List of Figures

1.1 Protein aggregation consists of the conversion of native states into pathological aggregates. 2
1.2 Overview of protein functional organisation through a range of scales. 3
1.3 Protein aggregates from a structural point of view. 5
1.4 Energy landscape of protein folding, unfolding and aggregation. . . 7
1.5 Chemical kinetics modelling of protein aggregation. 8
1.6 Human tau: isoforms and key literature constructs. 12
1.7 Pathological aggregates of the protein tau. 14
1.8 Literature mechanisms for tau aggregation. 18
1.9 Overview of phase separation in biology. 31
1.10 Interactions and factors influencing phase separation. 34
1.11 Overview of NMR. 41
1.12 Diffusion NMR. 44
1.13 Overview of chemical exchange. 46

2.1 Aggregation of R3 and R3R4 followed by thioflavin T fluorescence. . 55
2.2 Aggregation of R3 followed by NMR. 57
2.3 Lack of secondary structure of an R3 sample in non aggregating conditions. 58
2.4 Diffusion analysis of an R3 sample in non aggregating conditions. . 58
2.5 Aggregation of R3R4 with and without shaking. 60
2.6 Comparison of two 100 µM R3 NMR kinetic experiments. 60
2.7 Effect of DTT on the aggregation of R3 and R3R4 aggregation. . . 62
2.8 Seeding of full length tau by R3 aggregates. 65
2.9 R3 as a model system for tau aggregation: summary of features. . . 66

3.1 Mathematical complexity in protein aggregation systems. 71
3.2 Staircase representation of the basin-hopping framework. 73
3.3 Algorithm for mechanistic modelling - Description and example application outcomes. 80
3.4 Comparison of the accuracy of the forward Euler, explicit midpoint and variable step multistep methods. 83

3.5 Accuracy vs calculation time comparison of the forward Euler, explicit midpoint and variable step multistep methods. 84
3.6 Summary of modelling method developments. 95

4.1 D-TLKIVW, a computationally designed inhibitor of tau aggregation. 100
4.2 Pool of steps proposed for the aggregation of R3. 112
4.3 Pool of steps proposed for the inhibition of R3 aggregation by D-TLKIVW. 114
4.4 Free energy landscape for the two-step tetrameric nucleation mechanism. 116
4.5 Example model requiring mass balance corrections. 126
4.6 Initial rates analysis of protein-only data. 130
4.7 Mechanistic model for the aggregation of R3. 132
4.8 Evolution of goodness of fit through the modelling process. 135
4.9 Results of the second pruning stage of the first R3-only dataset. . . 136
4.10 Comparison of nucleation steps at an early and late stage of the modelling process. 139
4.11 Comparison of good nucleation mechanisms. 141
4.12 Mechanistic analysis of the second dataset of R3-only data. 145
4.13 Modelling path for the R3-only second dataset. 149
4.14 First pruning stage for the R3-only second dataset. 151
4.15 Second pruning stage for the R3-only second dataset. 152
4.16 Overlay of ^{1}H amide and methyl regions in the presence or absence of inhibitor. 156
4.17 Overview of modelling results for the action of the D-TLKIVW peptide on R3. 157
4.18 Evolution of modelling process for D-TLKIVW inhibition mechanism - Fits and residuals. 159
4.19 Evolution of modelling process for D-TLKIVW inhibition mechanism - Analysis of binding between small species and inhibitor. 160
4.20 Final individual-step pruning stage of the modelling of inhibition by D-TLKIVW - Fits and residuals. 172
4.21 Final individual-step pruning stage of the modelling of inhibition by D-TLKIVW - Analysis of binding between small species and inhibitor. 173
4.22 Custom testing of mechanistic hypothesis of inhibition by D-TLKIVW - Fits and residuals. 178
4.23 Custom testing of mechanistic hypothesis of inhibition by D-TLKIVW - Analysis of binding between small species and inhibitor. 179
4.24 Structural model for the aggregation of R3. 181
4.25 Overall mechanism of aggregation of R3 and inhibition by D-TLKIVW. 183

5.1 Overview of FUS phase separation. 191
5.2 Overview of DDX4 phase separation. 195
5.3 Overview of the experimental protocol for partition coefficient determination. 206
5.4 Volume fraction measurement. 207
5.5 Measurement of the lipoamide partition coefficient into FUS condensates by NMR. 217
5.6 Adjustments in the semiautomated NMR processing pipeline. 219
5.7 Relative partitioning of guanine and adenine nucleotides into DDX4^{N1}NoCys *in vitro* condensates. 220
5.8 Experimental method developments. 223
5.9 The partitioning trend of a set of aromatic molecules varies with protein construct and pH. 225
5.10 Partitioning of deoxyribonucleosides into DDX4^{N1} WT. 228
5.11 Hydrogen bonding of nucleobases. 235
5.12 Schematic summary of the investigation of the influence of chemical features on the partitioning of nucleotides into DDX4 condensates. . 239

List of Tables

1.1 Comparison of techniques for monitoring protein aggregation 10

3.1 Rate values used for simulations. 77
3.2 *rune* settings tested and corresponding *dt* in the forward Euler method. 77
3.3 Experimental scatter in duplicates of R3-only samples. 91

4.1 Mass balance corrections for elongating steps involving inhibitor surface-bound oligomers. 129
4.2 Modelling results for the R3-only first dataset. 133
4.3 Modelling results for the R3-only second dataset. 147
4.4 Modelling results for the R3 + D-TLKIVW dataset. 166

A.1 Rate constants fitted for the R3-only best model (first dataset) and for the R3 + D-TLKIVW best model. 260
A.2 Rate constants fitted for the R3-only second dataset best model. . 261

List of Abbreviations

1D . . . . . one-dimensional

3R . . . . . three repeat construct

4R . . . . . four repeat construct

AD . . . . . Alzheimer's disease

ADP . . . . . adenosine diphosphate

AMP . . . . . adenosine monophosphate

ATP . . . . . adenosine triphosphate

cAMP . . . . . 3',5'-cyclic adenosine monophosphate

cGMP . . . . . 3',5'-cyclic guanosine monophosphate

CD . . . . . circular dichroism

CSP . . . . . chemical shift perturbation

cryoEM . . . . cryogenic electron microscopy

dA . . . . . deoxyadenosine

dC . . . . . deoxycytodine

dG . . . . . deoxyguanosine

dU . . . . . deoxyuridine

DNA . . . . . deoxyribonucleic acid

DLS . . . . . dynamic light scattering

D-peptide . . . all D-amino acid peptide

DTT . . . . . dithiothreitol

EM . . . . . electron microscopy

FID . . . . . free induction decay

FRET . . . . . fluorescence resonance energy transfer

FUS . . . . . "FUsed in Sarchoma"

GTP . . . . . guanosine triphosphate

IDP . . . . . intrinsically disordered protein

IDR . . . . . intrinsically disordered region

LLPS . . . . . liquid-liquid phase separation

MTBR . . . . microtubule binding repeat domain

NMR . . . . . nuclear magnetic resonance

NoCys . . . . . cysteine-free mutant

ODE . . . . . ordinary differential equation

PBS . . . . . phosphate buffered saline

PC . . . . . partition coefficient

PHF . . . . . paired helical filament

ppm . . . . . parts per million

PTM . . . . . post-translational modification

RNA . . . . . ribonucleic acid

SF . . . . . straight filament

TCEP . . . . . tris(2-carboxyethyl)phosphine

TDP-43 . . . . TAR DNA-binding protein 43

ThT . . . . . thioflavin T

WT . . . . . wild type

1 Background

1.1 Introduction

The nature of this work is fundamentally interdisciplinary as is reflected in this introductory chapter, which covers the basic knowledge and description of research questions for the phenomena of protein aggregation and phase separation, and the fundamental principles of the experimental, mathematical and computational approaches that have been used to model these, in the broad literature and in the context of this detailed study. Within each subtopic, an effort has been made to achieve the right balance between width of scope and detailed description, since all of these fields have a rich literature of research that is beyond the scope of this work. Therefore, this introductory chapter offers in-depth descriptions of the aspects most relevant to the particular focus and methods used in this thesis, and aims to provide a broad description of other aspects of each field.

1.2 Protein folding, misfolding and aggregation

Protein aggregation consists of the conversion of functional native (ordered or disordered) proteins into pathological aggregates of amyloid, native or amorphous structure (see Figure 1.1) [1]. Aggregation represents a failure of proteostasis, the maintenance of proteins in their functional state, which is carried out by a

network of proteins: molecular chaperones, which assist *de novo* folding or refolding, and the ubiquitin-proteasome and autophagy systems, which facilitate targeted degradation of misfolded proteins [2]. Deposits of aggregated proteins constitute hallmarks of many human diseases, most notably devastating neurodegenerative disorders such as Alzheimer's, Parkinson's and Huntington's disease, and other forms of dementia. These disorders pose global health challenges yet have limited therapeutic approaches, with a high failure rate in the drug discovery pipeline [3]. There are many open questions in this field, such as in the mechanisms of formation of the aggregates, the actual cause of the disease (loss of function vs gain of toxicity, and which aggregates are most toxic), and the mechanisms of action of potential therapeutics.

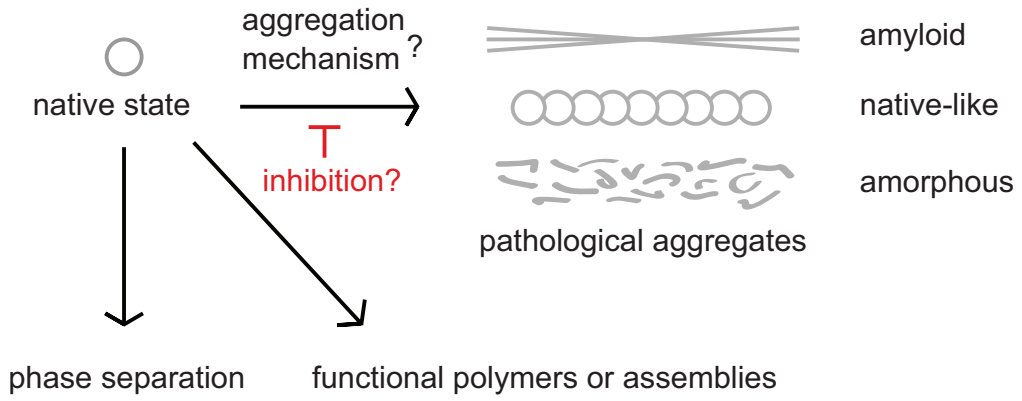


Figure 1.1: Protein aggregation consists of the conversion of native states into pathological aggregates. Schematic of the questions addressed in chapters 2,3 and 4 of this thesis.

In order to better understand what protein aggregation is, it is also helpful to consider what it is not: the wide range of functional protein assemblies through which proteins carry out their functions.

1.2.1 Functional protein organisation at several scales

The structure-function paradigm states that protein function stems from its three-dimensional structure, which is determined by its primary sequence. In this context, proteins can carry out their function in a range of assembly states:

single subunits, small oligomers and large macromolecular complexes or cellular substructures (see Figure 1.2 A) [4][5][6].

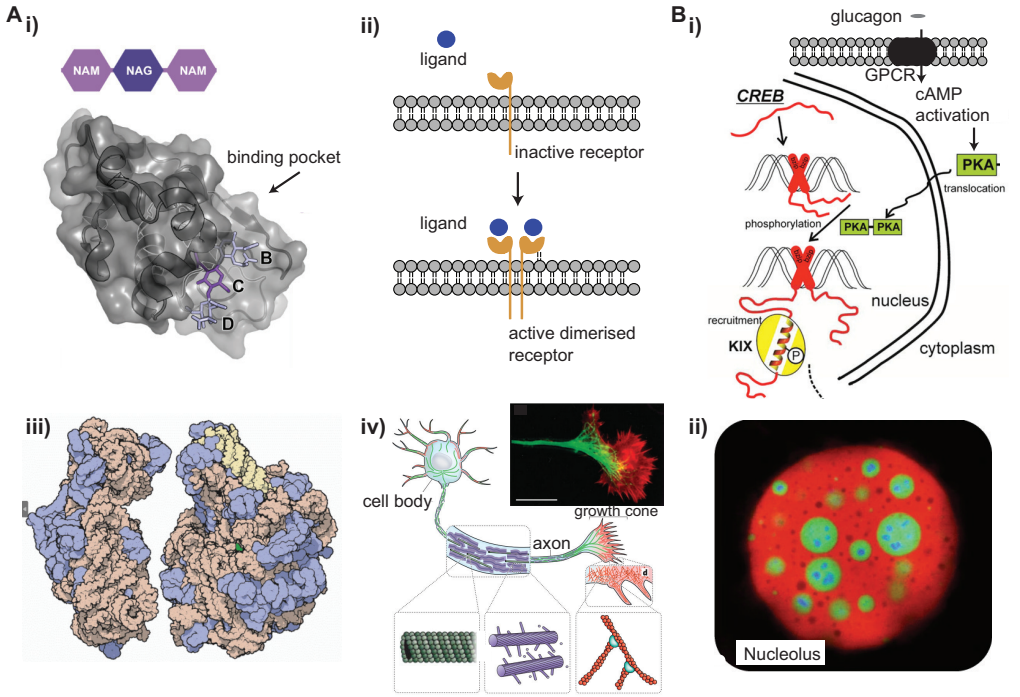


Figure 1.2: Overview of protein functional organisation through a range of scales. A. Functional proteins in a range of (mostly) structured assemblies. (i) X-ray crystal structure of lysozyme, which cleaves carbohydrate chains on the surface of bacteria (taken from [7]). (ii) Schematic representation of the activation of a signalling cascade *via* ligand binding and dimerisation of a cellular receptor. (iii) The ribosome is a macromolecular complex composed of more than 30 proteins, and of RNA, where protein translation occurs (taken from [8]). (iv) The cytoskeleton is a large protein network composed of microtubules (green), actin (purple) and intermediate filaments (red), represented schematically in a neuron. The fluorescence image corresponds to a neuronal growth cone during development (taken from [9]). B. Disordered proteins or regions are also functional, and linked to phase transitions which form membraneless organelles such as the nucleolus. (i) The cyclic-AMP response binding protein CREB and its binding partner both become more folded upon binding, which is triggered by phosphorylation of CREB downstream of GPCR activation (taken from [10]). (ii) The nucleolus is a multilayer membraneless organelle formed by phase separation (taken from [11]).

More recently, functional disorder has challenged the structure-function dogma, with intrinsically disordered proteins (IDPs) and proteins with intrinsically disordered regions (IDRs)[12]. Around 30 % of proteins are predicted to contain disordered regions, a proportion that increases with organism complexity [13]. These partially or fully disordered protein states are functional, interacting with a wider range of

partners, sometimes folding upon binding[14]. More importantly, IDPs and IDRs are linked to the emerging phenomenon of phase separation [15] [16], which is the object of Chapter 5 of this thesis and described in more detail below. Phase separation leads to the formation of large assemblies of protein and nucleic acids in a condensate phase distinct to the surrounding bulk. Interestingly, there are many links between phase separation and aggregation, which are discussed later.

1.2.2 Protein aggregation and disease

More than 30 peptides/proteins are associated with human disease, wherein their deposits as aggregated proteins constitute a hallmark of the disease [1]. Many protein-aggregation diseases are neurodegenerative. Some examples of the aggregating proteins include amyloid β and tau in Alzheimer's disease (see Figure 1.3 Ai), α-synuclein in Parkinson's disease (see Figure 1.3 Bii) and huntingtin in Huntington's disease. Some aggregating proteins are natively folded in their non-aggregating state, while others are partially or fully intrinsically disordered. There is also a wide range in terms of the genetics of these diseases, with cases at either extreme (mainly hereditary or mainly sporadic) [1].

Protein aggregates: a structural view

Protein aggregates can take several morphologies: amyloid, native-like and amorphous [1]. The amyloid fold is a very well defined structural feature which consists of β sheets arranged with a 4.7 Å longitudinal spacing (along the fibril axis) and 6-11 Å interstrand spacing. These spacings give rise to a characteristic cross-β diffraction pattern (see Figure 1.3 B). Even though amyloid structures have been studied mostly in disease-related proteins, it is important to remember that amyloid does not necessarily imply pathology, since there are examples of functional amyloids in humans and other organisms, such as in melanin production [20], hormone storage [21] and the yeast stress response [22].

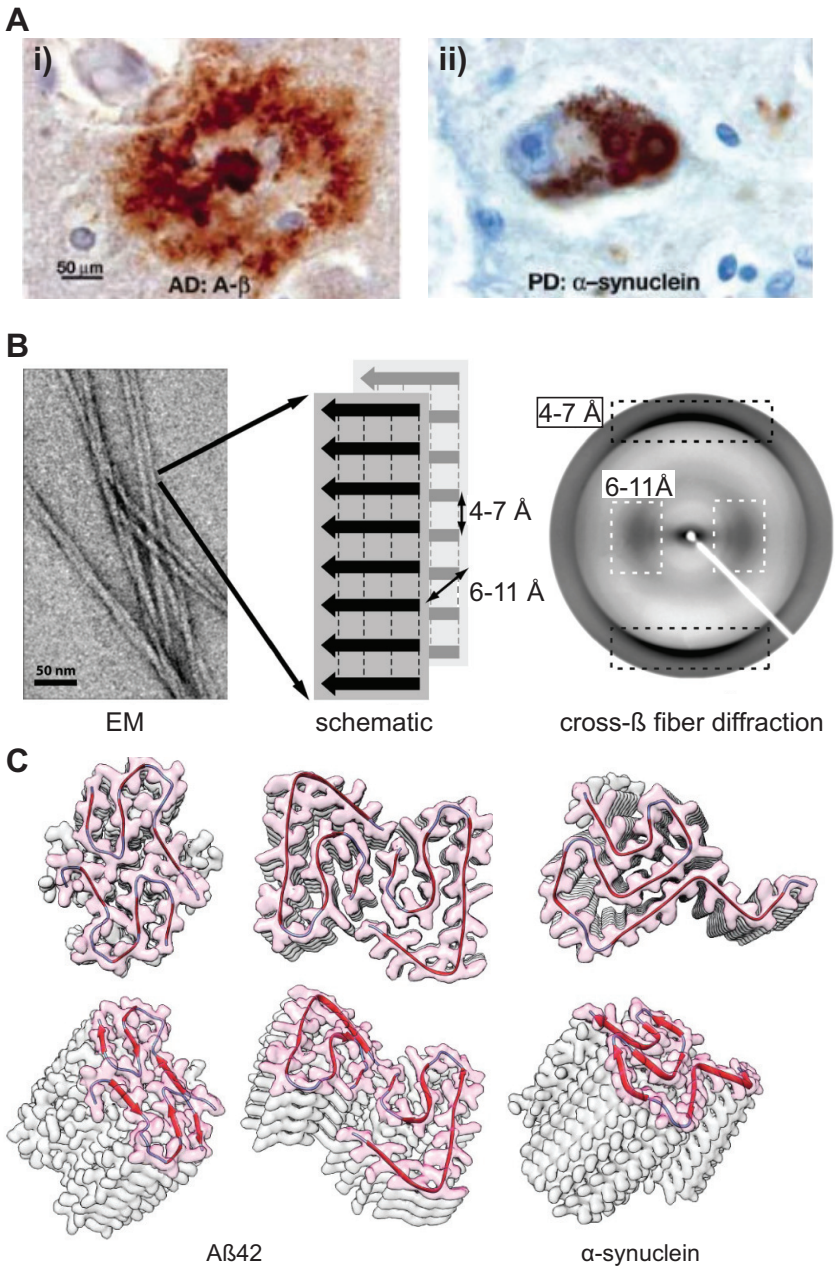


Figure 1.3: Protein aggregates from a structural point of view. A. Example of pathological aggregates in two very severe neurodegenerative diseases: (i) intracellular neurofibrillary tangles of tau (in brown) and extracellular amyloid β plaques (in blue) in Alzheimer's disease. (ii) Inclusions of α synuclein are shown in brown in Parkinson's disease. Images taken from [17]. B. Biophysical description of amyloid structure. Elongated (and often twisted fibrils), easily visible by negative stain EM, are composed of single or multiple protofilaments that have a β-sheet core. The longitudinal and perpendicular spacings between strands, with respect to the growth axis, give rise to distinctive cross-β fiber diffraction patterns. Taken from [18]. C. Structures of the cores of fibrils of amyloid β 42 (two polymorphs) and α-synuclein resolved by cryoEM (taken from [19]).

In contrast with the very well-defined structural amyloid core, amyloid fibrils show considerable polymorphism in terms of the residues forming the β -sheet fold (which usually does not encompass the entire protein, but a central core, leaving a "fuzzy coat" around) as well as the arrangement of protofilaments, for example regarding the degree of twist [1] [23]. Structural aspects of amyloids have been studied using a wide range of biophysical methods, which include circular dichroism, Fourier-transform infrared spectroscopy, Raman spectroscopy and solution NMR (secondary propensities of small species), X-ray crystallography (ordered cores) and electron microscopy (EM) and solid state NMR (oligomers and full fibrils) [24]. Recently, the study of amyloid fibrils has also benefited from the advances in cryogenic electron microscopy (cryoEM) that have revolutionised structural biology, allowing atomic resolution of the fibril cores of both patient derived and recombinant fibrils (see examples for amyloid β 42 and α-synuclein in Figure 1.3 C).

Metastability of the amyloid state

It is perhaps surprising (and worrying) that most proteins can form amyloid structures in the right conditions, *in vitro*. In fact, a systematic study of the amyloid state free energy showed a dependency on the number of residues, in contrast to the free energy of folding versus unfolded states (which depends specifically on the sequence) [25]. The free energy difference between the amyloid and the native state was often higher than the difference between the native and unfolded states; in addition, for most analysed proteins, the critical concentration above which amyloid formation would be favourable was lower than the *in vivo* concentration. These findings led to the conclusion that in physiological conditions the native state is only metastable with respect to the amyloid state (while it is stable with respect to the unfolded state), i.e. the barriers between functional native state proteins and the disease-related amyloid forms they can adopt are kinetic, not thermodynamic. This implies that the role of molecular chaperones is related to slowing down certain transitions (i.e. unfolding) and speeding up others (i.e. refolding) (see Figure 1.4) [2]. At this point it is worth mentioning that the fibril state has sometimes

been considered irreversible, yet there is ever-increasing evidence that chaperone complexes can disaggregate fibrils [26] [27] [28].

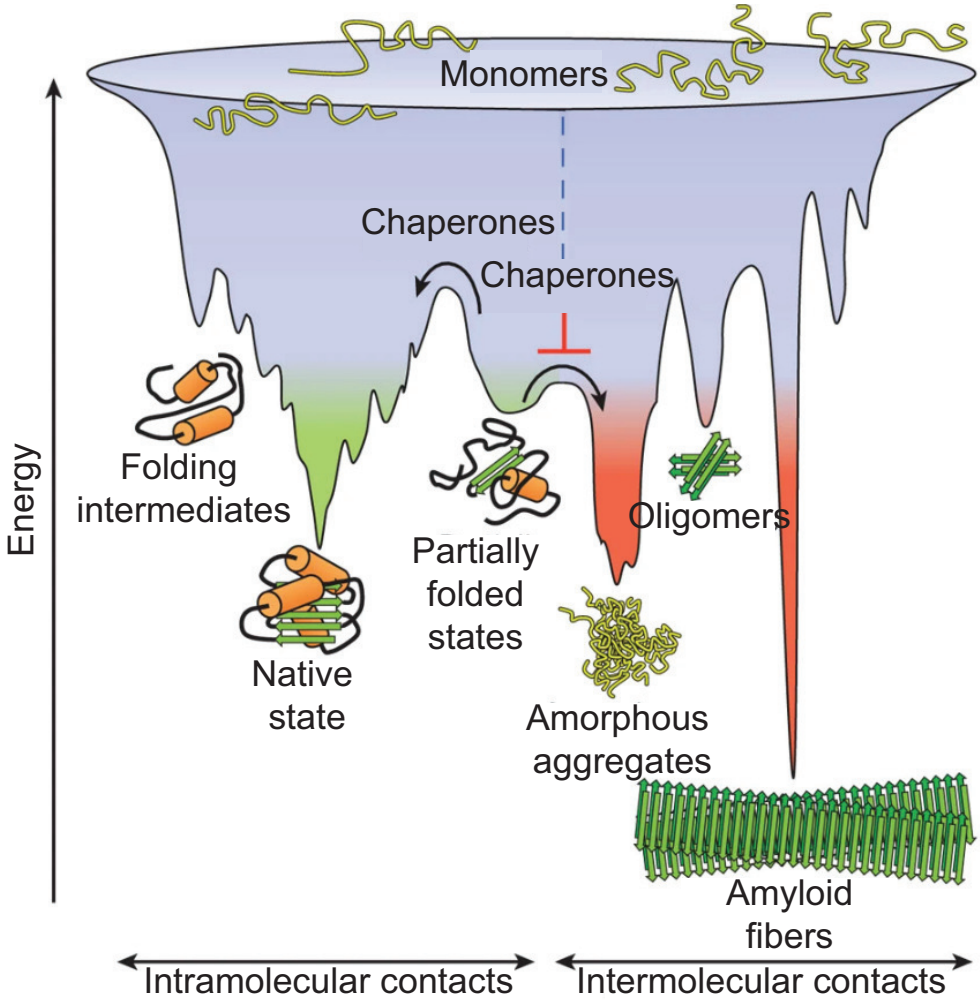


Figure 1.4: Energy landscape of protein folding, unfolding and aggregation. Energetic minima correspond to different states, with kinetic barriers in between. The fact that the amyloid state is the global minimum makes the native state metastable. In this sense, chaperones contribute to proteostasis by raising or lowering some of these kinetic barriers. Adapted from [2].

Kinetics of protein aggregation

There are a few features of protein aggregation kinetics that recur across different proteins [1] [29] [30] (see Figure 1.5). First, aggregation is considered to be a nucleated process, with initial unfavourable rate-limiting steps, which stems from the observation of a lag phase in the appearance of large aggregates. This can include

also conformational changes or unfolding of the monomer. Secondary processes, such as fragmentation and secondary nucleation, have also been considered for many systems. Finally, aggregates can propagate from cell to cell, in what is called *seeding*, which bypasses the need for the slow nucleation stage; this has been observed in neurodegenerative diseases, where aggregate pathology spreads through the brain [31] [32].

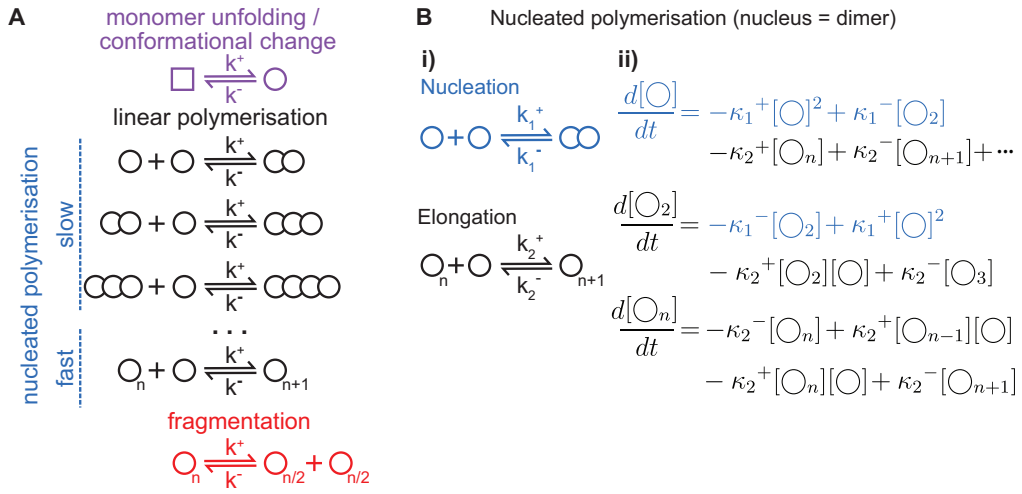


Figure 1.5: Chemical kinetics modelling of protein aggregation. A. Schematic of mechanistic steps proposed for protein aggregation, building on the simplest case of linear polymerisation. B. Mathematical expression of the chemical equilibria of a nucleated polymerisation model. (i) Chemical equilibria. (ii) The system of ordinary differential equations (ODEs) that describes the evolution of the system.

Challenges in the study of protein aggregation kinetics

Protein aggregation systems are highly heterogeneous and fast evolving, which leads to experimental and mathematical challenges. Further, *in vivo* aggregation occurs in the cellular context, where interacting partners of the proteins, as well as the proteostasis network, are likely to influence the process, which leads to biochemical challenges, some unique to this field of research and others intrinsic to *in vitro* work.

Biochemical challenges In order to conduct kinetics experiments, the aggregation assay needs to be reproducible and occur at a suitable timescale. In some cases, the

aggregation propensity of the protein is so high that after recombinant expression it is very difficult to keep it from precipitating, while in others, the opposite problem occurs, with the native full length protein being very stable in *in vitro* conditions. To address the latter, in particular, strategies such as the use of smaller/mutant (more aggregation-prone) constructs and aggregation inducers are common, as well as incubation under agitation conditions for increased reproducibility. Finally, aggregation is usually monitored *in vitro* (a most notable exception is the use of fluorescence to follow aggregation in *Caenorhabditis elegans*, which is transparent), which is often criticised for challenges in reproducibility of patterns of post-translational modifications, unlike *in vivo* studies. At this point I would like to note that in the course of this thesis, and for clarity, I will refer to studies as *in vitro* or "in cells" as separate categories, even though cellular work is technically *in vitro*. Therefore, in this work, *in vitro* refers to *in vitro* studies not carried out in cells, "in cells" refers to *in vitro* cellular work, and *in vivo* refers to work in whole organisms.

Experimental challenges Aggregation systems are very heterogeneous, with many distinct species that are evolving at the same time. For this reason, no single technique is suitable to study the behaviour of all the species [33]. I will consider three techniques to illustrate this in more detail, as well as introduce solution Nuclear Magnetic Resonance (NMR) as a technique with unique advantages that are still relatively under explored in this research area. Table 1.1 summarises the physical basis, advantages and disadvantages of dye-based ensemble fluorescence (the most commonly used technique), dynamic light scattering and NMR. Fluorescence-based methods have been extensively used in the protein aggregation field given the ability of certain dyes (thioflavin S, thioflavin T and derivatives, among others) to recognise amyloid fibrils in histological samples (reviewed in [34]). Their use in kinetics monitoring is also extensive, even though it presents limitations such as the incomplete understanding of the binding mechanism[35], which complicates the interpretation of signal in terms of concentration of particular species, as well as the interference of dyes with the aggregation process or inhibitors [33].

Table 1.1: Comparison of techniques for monitoring protein aggregation

Technique	Physical basis	Advantages	Disadvantages
Fluorescence	Molecular rotor fluorophores emit signal only when their rotation is restricted by binding amyloid structures	Amenable to high throughput Low sample requirement Can be used in cells /tissues / whole organisms	Most commonly used dyes only recognise amyloid structures Requires labelling (potential interference) Unclear signal-to-concentration relationship
Dynamic Light Scattering	Size determines diffusion, which affects light scattering fluctuations differently	Amenable to high throughput Low sample requirement Label free Can detect small species	Large or non-spherical aggregates cause bias Concentration limited (too turbid samples generate secondary scattering)
Solution NMR	Monitors small species, since with growing size transverse relaxation broadens signal beyond detection	Signal directly proportional to concentration Label free Multiple channels available (using isotopic labelling) Structural information	Large sample requirement Low throughput Limited information on large aggregates

While in principle dynamic light scattering would report on all species throughout the size range, this technique suffers from bias from non-spherical aggregates, which masks the signal of small species, as well as turbidity, which constrains the concentrations that can be used. In contrast, solution NMR is very well suited to probe the early smallest species in the aggregation pathway, given the signal-broadening effects of increasing molecular weight on molecular tumbling, usually a limitation in protein studies. The fundamentals of NMR, as well as the specific aspects most relevant to this thesis, are presented in a later section.

Mathematical challenges Chemical kinetics provides a framework for the study of reactions, in terms of rate laws and rate constants, which is applicable also to the field of protein aggregation. In this context, a mechanistic model of aggregation consists of a series of chemical steps such as the ones outlined in Figure 1.5. These can be treated as any chemical reaction and expressed mathematically as a series of ordinary differential equations (ODEs). The systems of ODEs that represent aggregation models are non-linear, due to the presence of bimolecular events, and in most cases an analytical solution cannot be found. Two main approaches have been used: exact numerical integration methods [36] [37] and approximation methods

[38] [39]. The work in this thesis uses an exact numerical approach, in order to exploit the very well-understood nature of the NMR signals and explicitly model the actions of inhibitors, an advantage with respect to existing approximation treatments [40]. This method was developed in our lab by Tkachenko *et al.* , and the proof-of-principle study characterised the mechanism of aggregation of the model protein α-lactalbumin and the inhibition of its nucleation by the molecular chaperone αB-crystallin [37]. A detailed introduction to the specific challenges in applying exact approaches for protein aggregation kinetics is included in Chapter 3, which presents and discusses developments to the existing method.

1.3 Tau biology and pathology

Tau is a protein encoded by the gene MAPT, located in chromosome 17q21.31 [41] (UniProt ID P10636), that was initially identified as a protein necessary for microtubule polymerisation [42]. Tau has taken a prominent place in biological and biomedical research due to its involvement in protein aggregation diseases such as Alzheimer's disease and other neurodegenerative disorders [43].

1.3.1 Domain organisation of tau

Tau is composed of an acidic N-terminal domain, a proline rich region, a a microtubule binding repeat domain (MTBR) and a C-terminal domain (see Figure 1.6). Multiple isoforms arise from alternative splicing of 16 exons [41]. Six isoforms are expressed in the central nervous system, ranging in size between 352 and 441 amino acids, and showing differences in the N-terminal and MTBR domains [44]. The terminology 3R and 4R refers to the presence of three or four repeat domains in the MTBR, respectively. The differences in the N-terminal region consist of the absence (0N) or presence of one (1N) or two (2N) short inserts. In the adult brain, there is a roughly equal composition of 3R and 4R isoforms, while only 3R forms (mainly 0N3R, called "fetal") are expressed in the fetal brain. A larger isoform of tau, called "Big Tau", was discovered in rats [45] and is expressed in the peripheral nervous system. Big Tau contains an insert of 254 amino acids in the N-terminal

domain [46]. The role of this isoform in health and disease is not very clear (see [47] for a comprehensive review). Finally, the largest central nervous system isoform, 2N4R (UniProt ID P10636-8) is often used as the reference sequence for residue numbering, a convention also followed in this work.

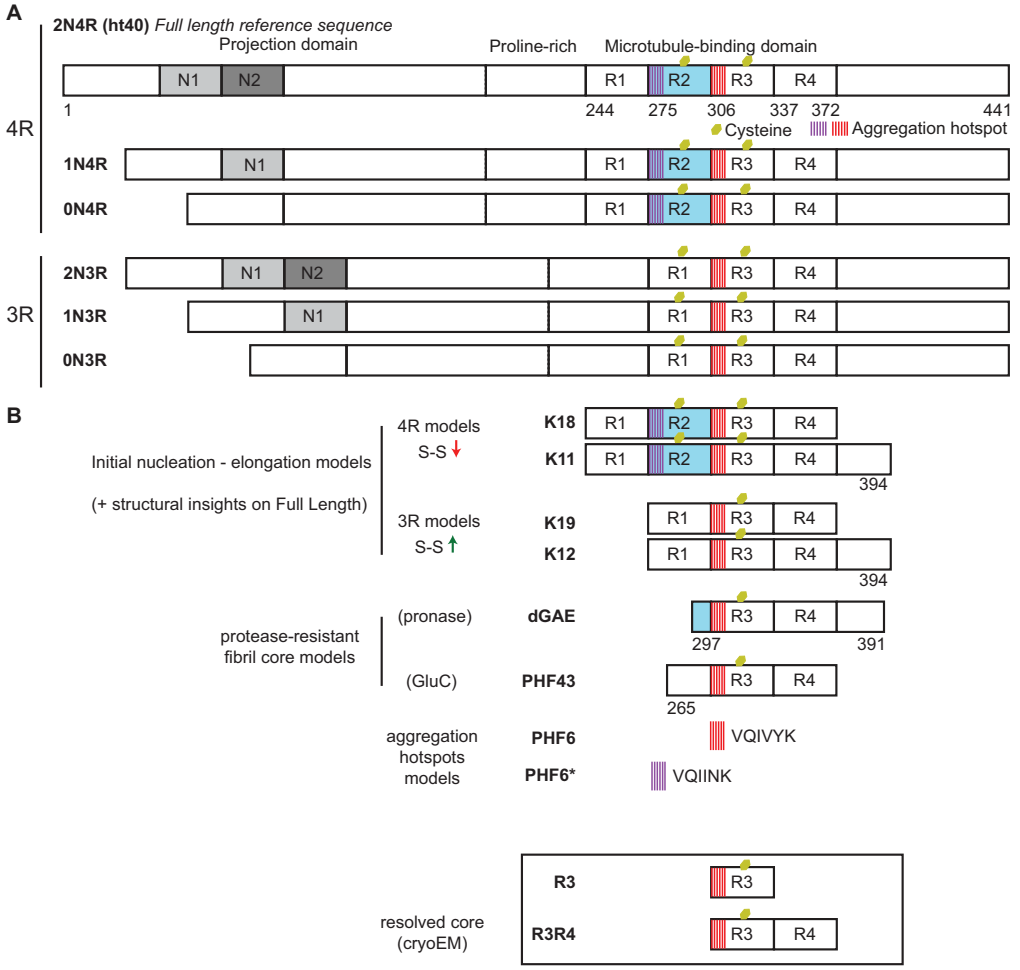


Figure 1.6: Human tau: isoforms and key literature constructs. A. Isoforms in the adult central nervous system. There is a larger isoform, Big Tau, which contains an additional 254 amino acids within the N-terminal domain (not shown). B. Model constructs used to probe tau aggregation mechanisms and structural features, discussed in the text. R3 and R3R4 are the model systems explored in this thesis, in chapter 2. The numbering of all isoforms and constructs refers to the 2N4R isoform, as is consensus in the literature.

Tau is an IDP, with almost no secondary structure [48] [49] and an elongated asymmetric shape in solution [50], yet with long range interactions that indicate a certain degree of global folding [51] [52], an aspect that is discussed below in more detail. Tau presents many sites for post-translational modifications (PTMs)

[53] [54]. Hyperphosphorylated tau is a hallmark of neurofibrillary tangles in AD, and phosphorylation was initially proposed to trigger aggregation by detaching tau from microtubules. However, it is emerging that the effects of phosphorylation on tau function and pathology are much more nuanced [55] [56].

1.3.2 Functions of tau

Tau promotes the formation and stability of microtubules in neuritic processes, mainly axons [57] [58]. Lee *et al.* identified the three repeat sequences in the C-terminal half of a short (fetal) tau construct as the sites of binding to microtubules; hence, this sequence has been referred to as the microtubule-binding domain (MTBR). However, binding to microtubules requires the adjacent proline-rich and C-terminal domains, and is modulated by the N-terminal "projection" domain, which remains disordered in microtubule-bound tau [58]. The N-terminal domain associates with the plasma membrane [59]. Apart from the classical role in microtubule dynamics, other functions have been proposed for tau more recently, such as regulation of axonal transport and a scaffolding role in signaling [60].

1.3.3 Tau pathology

Tau aggregation is a hallmark of many neurodegenerative diseases [61], amongst which Alzheimer's disease (AD) stands out in terms of disease prevalence. AD is the main cause of dementia, with rising prevalence predicted to reach around 150 million people by 2050 [62]. Tauopathies are mostly sporadic, but there are cases of dominant genetic familial forms that lead to frontotemporal dementia and parkinsonism (5% of the cases) and other disorders such as Pick's disease [43]. Based on the grouping of isoforms into 3R and 4R types (see Figure 1.6), tauopathies can be defined as purely 3R (e.g. Pick's disease), purely 4R (e.g. progressive supranuclear palsy) or a combination of 3R and 4R isoforms (AD) [43]. Figure 1.7 shows a few features of tau aggregates, including their spread across the brain in AD and their polymorphism as observed by EM and cryoEM.

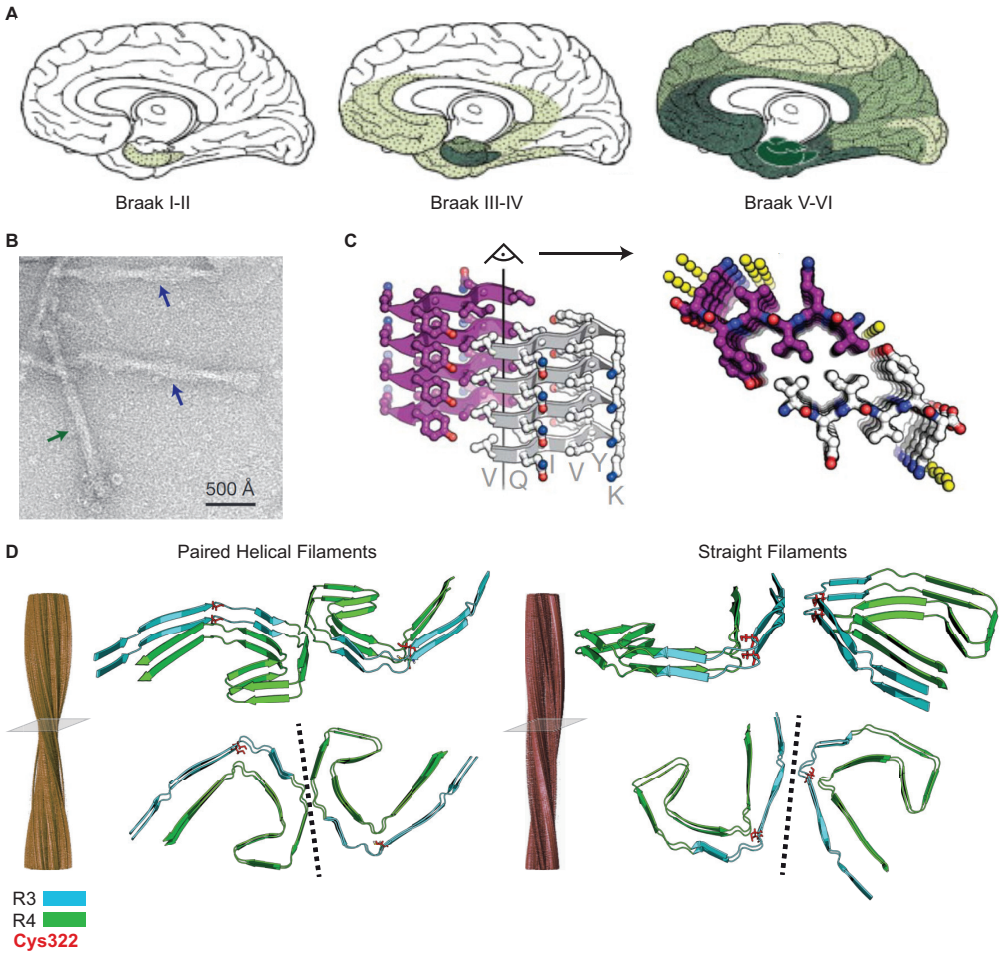


Figure 1.7: Pathological aggregates of the protein tau. A. In AD, tau pathology spreads through the brain with characteristic patterns, which led to the definition of six disease stages by Braak and Braak [63] (taken from [64]). B. Two main polymorphs of tau are found in AD: paired helical filaments (PHFs), indicated by blue arrows, and straight filaments (SFs), indicated by green arrows (taken from [65]). C. The hexapeptide VQIVYK aggregates into fibrils with a "steric zipper" interdigitated β-sheet interface (taken from [66]). D. cryoEM structures of PHFs and SFs, showing a common core encompassing the R3 and R4 repeats plus some adjacent residues, which forms the two polymorphs *via* distinct interfaces (taken from [65]).

Full length isoforms are highly soluble and it is difficult to find conditions for aggregation *in vitro*. For this reason, shorter constructs, mainly of the MTBR, have been routinely used, since they aggregate on a faster timescale [67]. Also, cofactors (also referred to as inducers) have been used, such as heparin [68], polyglutamate [69], arachidonic acid [70] and RNA [71]. Studies with constructs and in the presence of cofactors have led to a few features of aggregation of tau being unveiled.

The aggregation propensity of tau is located in the MTBD, indicating a

competition between microtubule binding and aggregation [67] [72]. In particular, two hexapeptide motifs, PHF6 (tau306-311) and PHF6* (tau275-280), located at the beginning of repeats R2 and R3, have the highest β-sheet propensities of the MTBR, and disease mutations that affect their secondary structure propensity have effects on the aggregation kinetics *in vitro* [72] [73]. The study from Von Bergen *et al.* was also seminal in unambiguously detecting β structure upon heparin-induced aggregation of constructs K19, PHF43 (resulting from the digestion of AD Paired Helical Filaments - PHFs - by GluC), and PHF6 (^{306}VQIVYK^{311}), which resolved the controversy of whether there was β structure at the core of tau aggregates [72].

The discovery of the aggregation hotspots enabled the initial atomic resolution structures of tau fibrils using microcrystals of VQIVYK [66], which revealed interdigitating β-sheets (see Figure 1.7 C). However, a true revolution in the structural biology study of tau came from the cryoEM resolved structures of AD brain derived PHF and straight filament (SF) fibrils [65] (see Figure 1.7 D). In the last few years, a wealth of structures of tau fibrils in distinct diseases have appeared [74] [75] [76] [77]. These have added another dimension to the existing debate regarding the biological relevance of aggregation studies using cofactors. The finding that the structure of recombinant fibrils formed with heparin was different to that of AD fibrils [76] questioned the use of cofactors in aggregation studies and led to a move towards constructs that aggregate in the absence of any external agents [77] [78] [79]. However, other studies argue that the cofactors used *in vitro* have biological relevance [75].

1.3.4 Mechanisms of tau aggregation

The mechanism of aggregation of tau into amyloid fibrils has been an ongoing topic of study for the last fifty years, with some emerging features, including the existence of a lag phase in the formation of amyloid aggregates and structural heterogeneity at the level of the monomer. Comparing and integrating the results of multiple studies is complicated by the large variability in experimental conditions of the aggregation reaction, which include: 1) which isoform or repeat region

construct was used (and any mutations, either disease-associated or at the cysteines in order to affect disulfide bond formation and/or introduce fluorophore labelling sites), 2) which cofactor, if any, was used, and at which ratio with respect to the protein, 3) the buffer composition in terms of salt, pH, and redox state, 4) whether incubation was quiescent or with agitation, and 5) which methods were used to monitor the reaction and model the kinetics.

In this section I present an overview of several kinetic models proposed for tau aggregation, from early seminal work to more recent models derived from experiments in cofactor-free settings. The latter are most relevant to the results presented in Chapter 4 using a cofactor-free model described in Chapter 2. I also review the evidence for conformational changes at the level of monomeric tau.

The earliest mechanistic descriptions were proposed in the late 1990s and differed in the redox conditions of the assay, the inducer of aggregation and the conclusions regarding the limiting step [70] [80] [81]. Following their development of a thioflavin S fluorescence monitored assay for full length of MTBR tau [69], Friedhoff *et al.* proposed a nucleated polymerisation mechanism with a dimer as the species involved in nucleation and elongation [80]. One of the characteristic features of the nucleation-dependent polymerisation model, an initial lag-phase, could be bypassed by a large enough protein concentration or a seeded assay (using preformed aggregates, called seeds, in addition to dimers) [80]. Kinetically, seeding manifested as a linear dependency of maximal velocities on the concentration of seeds in conditions of constant protein concentration and vice versa [80]. Their kinetic model described the aggregation pathway as an initial equilibrium of monomer (P_1) and oxidised dimer (P_2), a nucleation reaction involving n dimers and elongation by addition of dimer:

$$\mathrm{P}_1 + \mathrm{P}_1 \rightleftharpoons \mathrm{P}_2 \tag{1.1}$$

$$\mathrm{nP}_2 \xrightarrow{k_{nuc}} \mathrm{P}_{2n} \tag{1.2}$$

$$\mathrm{P}_{2n} \xrightarrow{k_{el}} \mathrm{P}_{2n+2} \tag{1.3}$$

The nucleation parameter n was fitted to a value of 4-7 dimers depending on the cofactor (4.4 for polyglutamate and 6.2-7.2 for heparin), and the dissociation rates were neglected as indicated in the equations. The model was experimentally tested with primary (for the nucleation reaction) and seeded (for the elongation reaction) experiments under oxidising conditions, using the 3R constructs K12 and K19, in the presence of either polyglutamate or heparin. The dimerisation reaction was not investigated kinetically, since dimers were allowed to form and isolated by size exclusion chromatography prior to all reactions. In these conditions, the aggregates formed were predominantly of the PHF morphology.

In contrast, King *et al.* proposed that tau nucleated through a ligand-mediated mechanism in the presence of arachidonic acid, with the binding of the inducer as the obligate first step for a rate-limiting conformational change [70]. This model was based on experiments performed in reducing conditions with full-length 2N4R tau, which predominantly formed SF fibrils. The lack of change in the potency of arachidonic acid when monomeric tau concentration was varied favoured a ligand-mediated mechanism (binding of the inducer followed by a conformational change) over a ligand-facilitated mechanism where the ligand would stabilise pre-existing oligomers. Polymerisation yielded SF filaments that would over time attain the morphology of PHFs. This was mistakenly attributed to self-association since at the time it was thought that SFs only contained one protofilament, which has been since proven incorrect by cryoEM [65].

These contrasting models and the distinct assay conditions were partially reconciled in an extensive comparative study by Barghorn *et al.* [81]. First, regarding the differences between using heparin and arachidonic acid as inducers, the authors noted that above the critical micellar concentration arachidonic acid would function as a polyanion, with the carboxylic acid groups pointing outwardly. Second, using either inducer, the 3R constructs tended to form PHFs, while the 4R would form SF (which could over time convert into PHFs, as discussed above). The unified model proposed by Barghorn *et al.* is shown in Figure 1.8 A.

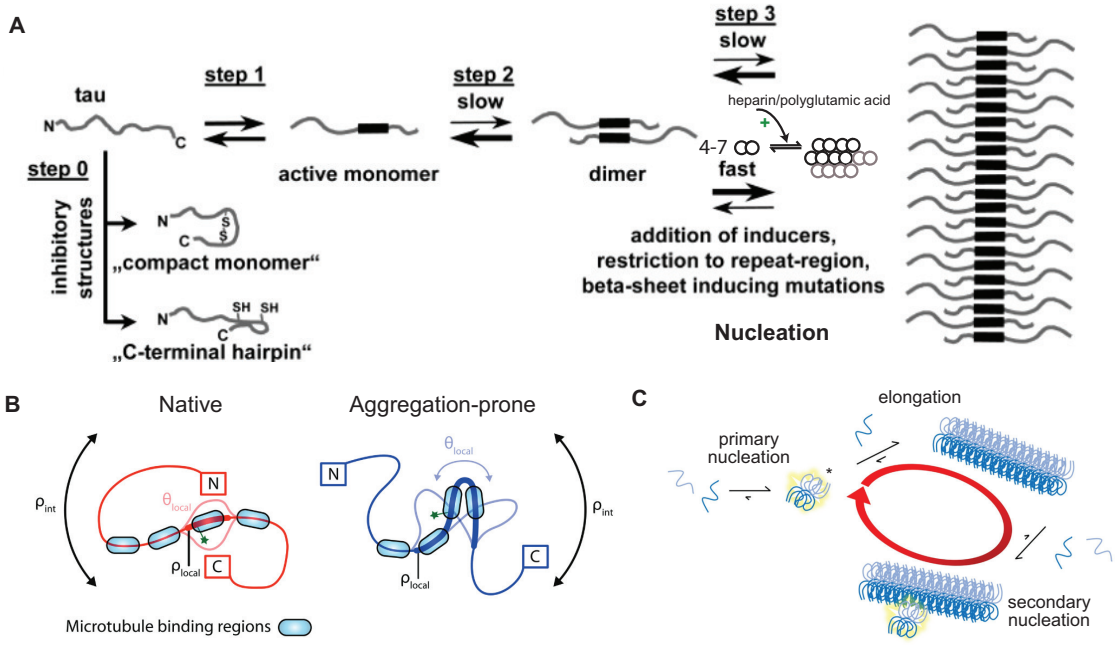


Figure 1.8: Literature mechanisms for tau aggregation. A. Unified model by Barghorn *et al.* , adapted from [81] and including the proposed nucleation mechanism for air-oxidised K19 dimers [80]. B. Structural models for the two families of full length tau monomer conformers identified using single molecule anisotropy measurements and characterised by anisotropy decay analysis (taken from [82]). The conformer families differ in the degree of overall compaction and the local rigidity of the MTBR; the aggregation-prone conformation (in blue) shows a loss of long range contacts between the termini and the MTBR, yet an increase in ridigity of the MTBR, similar to the effects measured by single molecule FRET upon addition of heparin [52]. C. Mechanism modelled for the cryoEM structure core fragment 304-380 (C322S) in quiescent conditions without inducers (taken from [78]).

The model included conformational changes of monomer, on and off pathway, on-pathway dimerisation followed by a less-precisely defined stage of nucleation, and elongation events, overall made faster by the presence of inducers or certain mutations. Both the conformational change of monomer and the dimerisation reaction were modelled as unfavourable (in terms of the rate constants) due to the very slow aggregation of full length constructs. The dimerisation equilibrium was studied in detail for 3R vs 4R MTBR constructs and found to be much more influenced by the redox state in the presence of heparin than with arachidonic acid. This difference was one of the main points of discrepancy prevailing at the end of this study. In aggregation reactions with heparin and under oxidative conditions, dimerisation of 3R constructs was strongly favoured and led to aggregation, while

the 4R constructs would be split roughly equally between an extended monomeric conformation and a more compact form, and aggregation would be impaired. This was attributed to an intermolecular disulfide bridge between two Cys322 residues, for 3R forms, while in the 4R construct an intramolecular Cys291-Cys322 bridge would lock the monomer into one of the off-pathway conformations. In the absence of any cofactor, and in reducing conditions, non covalent dimers of either 3R or 4R constructs form in a small percentage, which would be enough for aggregation to proceed when inducers are present. Overall, differences in the relative magnitude of rate constants would account for the observed rate-limiting steps in different experimental setups. Regarding the distinct effects of redox conditions on 3R vs 4R constructs, a recent study of Al-Hilaly *et al.* probed the dGAE construct, which only contains one cysteine, in a cofactor free assay. This study showed that reducing conditions, or the Cys322Ala mutation, favoured aggregation [79]. This indicates there are more nuanced sequence features relevant for understanding the relationship between disulfide bond formation and aggregation.

In support of the conformational change rate-limited model, Chirita *et al.* showed that for the 2N4R full length isoform, in reduced quiescent conditions and induction with carboxylate-substituted polystyrene microspheres, the initial velocities of thioflavin fluorescence appearance were first order with respect to monomeric concentration during the lag time before fibrils were visible by EM [83]. Further, there was a critical concentration of 2 µM for fibrillisation, similar to previous literature values, but not for the appearance of fluorescence, which would reflect the formation of the monomeric intermediate. They proposed that anionic inducers stabilise a pre-molten globule state, a partially folded conformation with higher β-sheet content, which precedes elongation into fibrils.

One mechanism of particular interest for the context of this work is that of the construct R3R4 (304-380) in the absence of cofactors and in quiescent conditions, at pH 8 [78]. The construct encompasses the resolved core of the cryoEM structure for AD fibrils [65] plus two residues in each direction, yet with the mutation Cys322Ser that abolishes the ability to dimerise through a disulfide bond at that position. As

described in the general method publication [39], the aggregation reactions were monitored by fluorescence using the dye X34 and fitted using integrated rate laws derived by approximations. A good fit for either primary aggregation or seeded aggregation was not possible with a model including only nucleation and elongation, while similar fits were achieved by the addition of a secondary event, either secondary nucleation or fragmentation, the model with secondary nucleation having a 5% smaller χ^2. Seeded assays with ^{15}N labelling of the seeds, followed by isolation of oligomers by centrifugation and size exclusion chromatography, showed that less than 0.035% of oligomers originated from the seeds (compared to the theoretical 17% maximum possible given the mixture of seeds and unlabelled monomer), which led the authors to conclude that secondary nucleation of monomers was the main secondary process. It is worth pointing out that the primary nucleus could actually correspond to an initial conformational change of the monomer, since the parameter nc (how many monomers make the nucleus) was fitted to 1 in the best fit. In contrast, in the best fit secondary nucleation was fitted as requiring two monomers to nucleate on the surface of fibrils.

1.3.5 Structural transitions of tau involved in aggregation

The conformational landscape of tau has been probed using several ensemble and single-method spectroscopic methods. However, in many studies it was not possible to determine the order of events between the conformational change and the formation of small oligomers. Early work showed that upon aggregation tau acquired local β-sheet structure within a globally random-coil structure [72], which was observable by CD and FTIR in shorter constructs (but masked in the longest isoforms) and has abundantly been seen by cryoEM in recent years. This transition was considered to occur post-dimerisation, in the nucleation stage [84], but this could be due to lack of detectability; following a complete assignment of the 441 residue full length longest isoform, NMR showed β-sheet propensity regions in monomeric tau [85]. The propensity localised to stretches in the repeat regions with populations between 12 and 25 %, and with the R2 and R3 hexapeptide motif regions having

the highest propensity, in agreement with the identification of those as hotspots [72]. In line with this, another ensemble study, this time using double electron electron resonance (DEER), was able to detect an extension of the hexapeptide regions in soluble species upon induction of aggregation by heparin; this occured on a faster timescale than the appearance of fibrils but it was not possible to resolve whether the conformational transition preceded or followed the formation of oligomers[86].

In terms of large conformational rearrangements, the full length protein has been shown to be disordered and elongated, yet a seminal study using ensemble FRET detected long range contacts between the N-terminal domain and the MTBR, leading to the so-called "paper clip" conformation model [51]. A certain degree of folding had already been suggested since epitopes for antibodies that detected early conformational changes were spread across sequences far away in the primary sequence [87]. In parallel, the NMR study by Mukrasch *et al.* used Paramagnetic Relaxation Enhancement (PRE) restraints to model the conformational ensemble of monomeric tau; if only allowed to fit one conformation, a quite compact conformation was found, yet modelling an ensemble of 30 conformers (considered by the authors to be an underestimation) led to the appearance of both compact and more extended conformations [85].

In a seminal single-molecule FRET study of tau, Elbaum-Garfinkle *et al.* showed a structural rearrangement induced by the high affinity (nM) binding to heparin [52]. Single molecule FRET data of 2N4R full length and eight truncated constructs, in conditions with heparin, could be fitted to a two-state model (and not a continuous transition expected from charge screening), in which heparin reduced long-range interactions between the termini and between each terminus and the MTBR, leading to an overall more extended conformation, yet with a more compacted MTBR. Some truncated constructs also had a compact MTBR, which would mean both the presence of heparin and truncation can shift the equilibrium towards more compact aggregation-prone conformations. In that study, the heparin-free state was described as S-shaped. Importantly, the construct used did not have any cysteines able to form disulfide bonds, since they had been used for fluorophore labelling.

Finally, a recent single molecule study measuring fluorescence anisotropy of trapped monomeric 2N4R full length tau also showed a bimodal distribution with conversion in the timescale of tens of minutes, which the authors interpreted as two families of conformations [88]. A further study using time-resolved anisotropy decay analysis, and the same trapping, revealed one of the two families to have a more rigidly structured MTBR domain, reflected in lower local rotational motion, that is able to sample more conformations (i.e. has larger local anisotropy)[82] (see Figure 1.8), indicating the free protein is already sampling the conformations observed upon heparin addition [52].

From the kinetic point of view, and further to the above-mentioned studies that had detected a kinetic fingerprint of an early conformational change [70][83], a recent kinetic study proposed a structural rearrangement at the level of small oligomers which would render them elongation competent. Shammas *et al.* [89] proposed a nucleation-conversion-elongation model based on studies of a K18 construct without free cysteines (C291A, C322A, I260C for labelling with fluorophore *via* maleimide chemistry), using a single molecule FRET technique to distinguish monomers from oligomers and applied previously to study α-synuclein [90]. Their conditions were pH 7 and a molar ratio of 1:4 tau:heparin (generally considered the optimum ratio since early studies [80]). Fitting to a nucleation-elongation mechanism, considering two species, monomer and oligomers, yielded a poorer fit than when considering a more complex model with monomer, non-growth competent oligomers and growth competent fibrils. In the best fitting model, there is a reaction of conversion of oligomers into fibrils that is first order with respect to monomer, as well as a nucleation reaction which is assumed to be a dimerisation reaction. In both models, the heparin effects were subsumed in the nucleation rate constant (i.e. not accounted for explicitly).

1.3.6 Aggregation modulating inhibitors of tau for AD

Therapeutics development for AD was initially focused on amyloid β, following the amyloid hypothesis according to which the initial pathogenic event was the

formation of amyloid β plaques, which would trigger cellular dysfunction phenotypes, amongst which are the formation of tau neurofibrillary tangles [91]. The continued failures in clinical trials, together with evidence challenging the amyloid hypothesis, such as the correlation of neuronal loss with tau deposits but not amyloid β plaques [92] and the better prediction of mild AD to severe AD transitions based on tau imaging compared to amyloid β measures [93], have led to a shift towards tau-targeting therapeutics, which will also benefit tauopathies without amyloid β pathology [94].

Below I focus in particular in therapeutic strategies directly targeting tau, rather than affecting it by reducing its expression or by targeting enzymes that regulate tau post-translational modification [95]. A variety of therapeutic approaches targeting tau aggregation are under preclinical and clinical study; most of them use small molecules [96] or immunotherapeutics [97][98], and emerging new approaches include RNA based therapies (in clinical trials) [99][100] and peptide therapies (in preclinical stage) [101].

A discussion on the most prominent small molecule inhibitor, methylene blue, is used to illustrate the lack of detailed understanding of the mechanism of action of potential therapeutics, even as they progress along the clinical trials path. Finally, since in Chapter 4 I present the mechanistic analysis of the inhibition by an all D amino acid peptide, I include an overview of peptide therapeutics and in particular early stage (preclinical) D-peptide candidates for tau targeting.

Immunotherapy

There has been extensive work on developing immunotherapeutics that target tau, using either passive (antibodies) or active (vaccines) approaches. Tau antibodies target tau in either monomeric or aggregated forms, with specific PTMs or conformational changes and throughout the distinct domains of the protein. Four clinical candidates have recently failed in phase I or II clinical trials, but there are some with ongoing trials [98][102]. There are also two active immunotherapeutics for tau: AADvac1 and ACI-35. AADvac1 targets the C terminal sequence of repeat R2 [103] [104], and has

successfully completed phase I and II trials [105].ACI-35 targets a phosphorylated sequence in the C terminal region [106]; following initial low immunogenecity in phase I trials, one of the redesigned versions has been selected for further development based on preliminary results from a phase 1b/2a trial [107] [108]. Investment in tau immunotherapeutics is likely to increase following the (controversial) approval of aducanumab in the US as the first amyloid-β targeting antibody for AD [109] [110].

Small molecules

Several small molecules have been identified *in vitro*[111]. Methylene blue (MB) and its derivatives have been the most studied since the finding that it made the core fragment of PHFs isolated from AD brain tissue susceptible to proteolytic cleavage by pepsin, which it would otherwise be resistant to [112]. MB is a redox-cycling phenothiazine initially used for the treatment of malaria, in the late 19th century, and currently indicated for acquired methemoglobinemia but actively studied for other indications [113]. MB has been shown to inhibit tau aggregation *in vitro* [112][114], and mitigate tauopathy *in vitro* and *in vivo* through induction of autophagy [115].

MB has been repurposed as a neurodegeneration therapeutic by the company TauRx and tested in clinical trials in two different formulations: a phase II trial of methylthioninium chloride (MTC; Rember™) showed cognitive improvements in patients with moderate AD[116], in patients receiving the second highest dose. The lack of response to the highest dose prompted a new formulation as the reduced form leuco-methylthioninium (LMTX trademark), with better pharmacokinetic properties [117]. LMTX failed to show efficacy in two phase III trials (Gauthier 2016), yet another related compound, Hydromethylthionine mesylate, has been recently tested in a phase III trial with positive (yet controversial, due to *a posteriori* changes in the analysis of the data) results, following which TauRx announced they would pursue FDA approval [118].

The mechanism of action of MB on tau aggregation has been under some controversy. MB was proposed by two studies to act through a redox cycling mechanism that promoted disulfide bridges, but has since been shown to inhibit

aggregation under reducing conditions as well. Briefly, Crowe *et al.* [114] showed that both methylene blue and another set of inhibitors (aminothienopyridazines, or ATPZ) promoted the formation of disulfide bridges only under oxidising conditions, and had different effects on the heparin-induced aggregation of the constructs K18 and K19. These compounds prevented the aggregation of the 4R K18, which can form the compact monomer form through an intramolecular disulfide bridge; however, they did not affect the aggregation of K19, instead they shifted K19's monomer-dimer equilibrium towards dimer in a similar manner to air oxidation, indicating the formation of the intermolecular disulfide bridge previously identified[119]. MB showed non-specific cysteine oxidation of a 10mer peptide into an intramolecular disulfide bridge, and a redox-cycling mode of action was suggested. Similar observations were made using NMR on the 2N4R full-length isoform with Azure A and B, which are derivative methylated forms of MB and also exist in oxidised and reduced form [120]. However in that case, using higher concentrations of MB, the aggregation of the 3R K19 construct was also inhibited by MB (with a lower potency compared to K18) both in oxidising conditions and reducing conditions (up to 5 mM DTT) but to a lesser extent in the latter, in which aggregation itself was impaired. Mass spectrometry showed evidence of the azure compounds converting the sulfhydryl groups of cysteines into sulfenic, sulfinic and sulfonic acid.

Al-Hilaly *et al.* proposed a different mechanism based on the findings that MB exists predominantly in the reduced form (LMT) in reducing conditions, and that it inhibited aggregation of the construct dGAE more strongly in reducing vs oxidising conditions [121]. Therefore LMT would be expected to be active in the natural reducing environment of cells, which was questioned by Crowe *et al.* [114]. This would seem to be in agreement with current clinical trials using the stabilised reduced form.

Peptides

Peptides (2-100 amino acids) have become promising candidates for protein-protein interaction disruption, thanks to their distinct properties compared to small

molecules and antibodies, and there are currently 60 peptide therapeutics in use, mostly for metabolic diseases and cancer, as well as hundreds in clinical trials [122]. Their advantages include a very high control in their production, using chemical synthesis, as well as high biological specifity and potentially low immunogenicity, while their main drawbacks are instability due to serum proteolysis and fast kidney clearance as well as low penetration through the blood-brain barrier [123]. However, a variety of strategies are under development that address these limitations.

The usage of all D-amino acid peptides (D-peptides) is one of the main strategies employed to reduce the protease susceptibility of peptides [124], together with the identification and replacement of possible cleavage sites and chemical modifications to enhance the degree of secondary structure of the peptides [125]. D-peptides could have better oral absorption, which is an important area of development in peptide therapeutics, since most are injectables.

D-amino acid peptide development There are several established methods for the development of all D-amino acid peptides, mainly retro-inverse synthesis and more recently mirror image phage display [124]. The starting point for retro-inverse synthesis is the L-peptide of interest; by assembling the D-amino acids in the reverse order to the L-counterpart reference peptide [126], a potentially equivalent all D-amino acid peptide is generated. The original L-peptide and the retro-inverse peptide have topologically equivalent side chains, with the backbone reversed. However, this approach does not guarantee that the D-peptide will maintain the chemical interactions and/or biological function [127].

Phage-display [128][129] is the main combinatorial screening method for protein-protein interactions, with great relevance in drug discovery pipelines for antibodies and peptides [130]. Libraries of peptides/proteins are fused to a coat protein encoding gene from filamentous phage, and a ligand of interest is immobilised to a surface and exposed to the phage library. Selection, consisting of several rounds of screening, is based on the presence of the interaction between the ligand and one or more displayed peptides/proteins; the sequence of the final best hit(s) is

then determined by sequencing. This approach is becoming even more powerful by the development of methods that allow cyclisation of peptides, post translational modifications and other chemical modifications, all of which expand the chemical space probed by the libraries.

One particular development useful for D-amino acid therapeutics came in the form of mirror image phage display [131], in which the target is synthesized using D-amino acids (for large proteins, using native chemical ligation [132]), and the library contains L-amino acid peptides. Once a peptide has been selected, its D-enantiomeric form is synthesized and therefore the symmetry of the binding interface is maintained.

There are also many computational methods for peptide design which have spun out from the field of small molecule docking and which can be classified into three categories in increasing order of difficulty: template-based docking, local-docking and global docking [133], depending on the amount of information (atomic resolution structure of the target and/or target-peptide complex, location of binding site, or none, respectively) available about the interaction. Both computational and experimental screening methods have been applied to the development of tau targeting D-peptides [101].

Computationally-designed tau-targeting D-peptides Sievers *et al.* used a mirror image template-based docking method, starting with the inverted stereochemistry structure of an X-ray resolved structure of VQIVYK fibrils, which showed a "steric zipper" [66] [134]. This inhibitor and its design are discussed in Chapter 4, where its mechanism of action on a model system of tau is analysed.

Furthermore, the same group developed a longer 10mer peptide, W-MINK (DVWMINKKRK), by applying the same computational strategy to the structure of fibrils containing the VQIINK sequence (beginning of the R2 repeat), which formed a stronger zipper interface in terms of larger buried surface area - this inhibitor had a lower IC50 for 2N4R tau compared to D-TLKIVW[135]. Finally, another computational design study, this time targeted at residues 16-22 of amyloid

β, led to the development of D1 (D-LYIWVQ), D1b (D-LIYWIWRT) and D1d (D-LYIWIQKT), which not only prevented the aggregation of amyloid β but also impaired the ability of preformed amyloid β seeds to cross-seed intracellular (K18) tau in cells, as well as primary and seeded aggregation of 2N4R tau [136]. The effects were specific for amyloid β and tau, since other amyloid forming proteins were not inhibited. Tau binding was mapped to the VQIINK and VQIVYK segments of tau, suggesting structural similarity between the tau and amyloid β cores.

Experimentally selected tau-targeting D-amino acid peptides Experimental approaches have mostly used phage display. Using mirror-image phage display based on fibrils made of D-VQIVYK, Dammers *et al.* selected nine peptides to compare with D-TLKIVW as the standard positive control in *in vitro* aggregation assays with 2N4R full length tau, using arachidonic acid as the inducer [137]. Of those, five were active as inhibitors of the three repeat construct K19, with TD28 having the lowest IC50 at 8 µM. However, even though they were able to penetrate cells, cytotoxic effects were observed. From a structural point of view, it's interesting to note that *in silico* modelling revealed similarities in the binding of TD28 and D-TLKIVW to VQIVYK: stabilising hydrogen bonds with the backbone of VQIVYK, an interaction between a central positively charged residue and Q307 and a steric clash that would block elongation (the D-TLKIVW case is described in Chapter 4).

In a similar study, the peptide p-NH (D-NITMNSRRRRNH) was found to inhibit VQIVYK fibrils [138]. Seven candidates were selected by phage display, and from three of those being carried forward due to higher affinity to VQIVYK fibrils in ELISA assays; from the final three peptides, p-NH (but not its isomer L-p-NH) was inhibitory of primary aggregation and also disassembled preformed aggregates, while the other two actually promoted aggregation. p-NH was able to enter cells and caused site specific reduction of tau phosphorylation, as well as reduced neurofibrillary tangles. Importantly, these effects were recapitulated in transgenic mice expressing the human P301S tau gene, using intranasal administration (which bypasses the

blood-brain barrier), and cognitive improvements were detected, which makes this peptide the most promising candidate for clinical trials.

Mirror-image phage display has also been applied to VQIINK fibrils which identified MMD3 and MMD3rev (the retro inverse version), which were able to inhibit fibrillisation of the VQIINK (but not VQIVYK) fragment as well as the 2N4R full-length isoform, by promoting the formation of amorphous aggregates [139]. *In silico* modelling showed similar interaction features with the peptides developed by Seidler *et al.* [135].

1.3.7 How well do we understand the mechanism of action of tau and of tau-targeting inhibitors?

Neurodegenerative diseases associated with protein aggregation have the highest failure rates in the drug discovery pipeline. This is not surprising given the many open questions regarding how proteins evolve from their native, folded or disordered, states into pathological aggregates. Without a detailed mechanistic understanding of the aggregation pathways of proteins, there is little solid basis upon which to decipher the mechanism of action of a potential therapeutic [140]. In the specific case of Alzheimer's disease, the earlier therapeutic focus on amyloid β, consequence of the amyloid hypothesis [141], has been shifting towards tau [94], given evidence that tau pathology correlates better with disease severity [92]. The long-standing controversy on the mechanism of MB on tau aggregation, discussed above, is an example of the gaps in our understanding of tau therapeutics, as well as how investigation of a detailed kinetic mechanism of action often follows, rather than precedes, the start of clinical trials of potential therapeutics.

The study of tau aggregation *in vitro* is hindered by the high solubility of the full length protein, requiring the use of model systems and a variety of aggregation inducing cofactors and conditions as discussed in the previous section. There is a consensus that rate-limiting nucleation events initiate aggregation, yet the nature of the aggregation nucleus has been ill-defined in the proposed mechanisms, and studies investigating conformational heterogeneity of monomeric tau and conformational

changes in the early stages of aggregation have struggled placing the rearrangements precisely within the overall mechanism. Further, there is also conflicting evidence regarding the influence of redox conditions on aggregation.

The aim of the first project presented in this thesis was to improve our understanding of the mechanism of aggregation of the protein tau, involved in AD and neurodegenerative disorders, by using a suitable model system, as well as its inhibition by a literature peptide (see Figure 1.1). We tackled these aims within the framework of NMR-based chemical kinetics, in which a mechanism of aggregation is modelled explicitly as a set of chemical reactions, and multiple mechanisms are optimised and ranked against NMR kinetic traces [37]. This method is particularly well suited to probe the early stages of aggregation and explicitly account for inhibitor interactions.

More detailed introductions to the technical aspects are provided at the end of this introductory section (NMR) and in the introduction of chapter 3 (mathematical modelling), and the results of this project are presented in chapters 2, 3 and 4.

1.4 Protein phase separation

1.4.1 Overview of membraneless organelles

The phenomenon of phase separation is a recent discovery in the field of cell biology [15], even if the membraneless organelles (MLOs) it gives rise to (also referred to as biomolecular condensates) have been observed for a much longer time using light microscopy. In this context, liquid-liquid phase separation (LLPS) involves the formation of condensates which behave like oil droplets in water, for which they are known as "liquid-like". MLOs form through weak multivalent protein-protein interactions that under certain conditions become stronger than interactions with the solvent, water in this case. Figure 1.9 shows an overview of the MLOs present in eukaryotic cells and the main physicochemical properties of these compartments.

MLOs differ from membrane enclosed compartments because they form and dissolve upon stimuli[144], and they are more permeable[145][146]. Three key properties were initially proposed to define an MLO: rapid internal rearrangement,

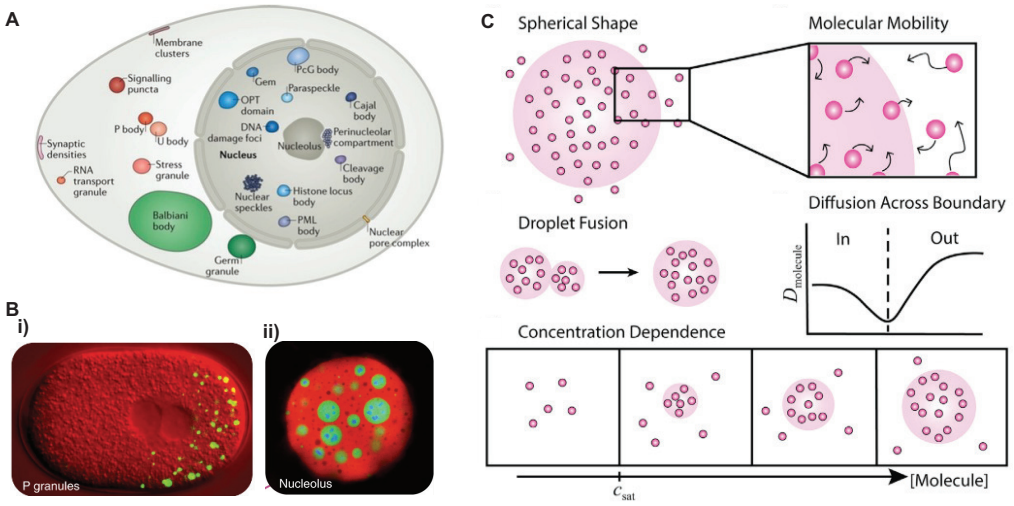


Figure 1.9: Overview of phase separation in biology. A. Diagram of MLOs identified in eukaryotic cells to arise from LLPS. Taken from [16]. B. Examples of MLOs. (i) Germline P granules (taken from [142]). (ii) The nucleolus, which is a multilayered MLO (taken from [11]). C. Properties of MLOs (taken from [143]).

spherical shape (due to surface tension) and the ability to coalesce (two droplets fuse into one)[147]. The internal rearrangement has been studied with fluorescent recovery after photobleaching (FRAP) experiments, in which the signal of a fluorescent-tagged droplet component is bleached and the signal recovery time reports on diffusion and exchange between the condensate and dilute phase[142] [148]. Two more properties have been added to the list of requirements: slowed diffusion at the interface between the condensate and dilute phases (due to the free energy cost of leaving the condensate phase)[149] and concentration dependence of the formation and size of condensates[150][151].

Finally, a recent review suggested that from a biological perspective the challenging distinction between "true" condensates and dynamic hubs (assemblies with rapidly exchanging multivalent interactions) might be possible based on the criterion of whether there is an internal environment in which the protein (and RNA) constitutes a distinct solvent to the outer bulk environment [152]. The hypothesis of condensates being a distinct chemical solvent environment is one of the focuses of chapter 5.

1.4.2 Importance of disorder for phase separation

Phase separation and protein intrinsic disorder are intricately linked: the proteome of MLOs is enriched in proteins containing IDRs[153] and several features of those are beneficial for phase separation. These include their conformational flexibility, modularity, multivalency, low sequence complexity and accessibility to PTMs [154], as well as the ability to interact promiscuously with other IDRs/IDPs or folded proteins / domains [144]. It has been shown that the dimensions of disordered proteins correlate with phase separation boundaries; indeed, more compact proteins have a lower critical temperature (below which they phase separate) [155]. Also, several disordered domains of phase separating proteins have been shown to remain disordered in the condensate phase [156][157][158][159]. Both folded proteins/domains and IDP/IDRs are important for phase separation, and most phase separating proteins contain both disordered regions and folded domains, the latter including RNA binding domains, oligomerisation domains and specific functional domains [144].

1.4.3 Functions of membraneless organelles

Many compartments of eukaryotic cells which arise by phase separation are now referred to as MLOs, in contrast to the standard membrane-enclosed compartments such as mitochondria, the nucleus and the ER. Some examples are P (germ) granules [142] (Figure 1.9 Bi), the nucleolus [160] (Figure 1.9 Bii), Cajal bodies [161] [162], P bodies [163] and stress granules [147]. MLOs contain tens to hundreds of different proteins, some of which are constitutive ("scaffolds") while others, the majority of components, are only transiently present ("clients"), and often also RNA [164][16]. Some MLOs have multiple subcompartments; in the case of the nucleolus, *in vitro* reconstitution showed that RNA-binding domains drive the specificity of the immiscible subcompartments, suggesting a role for the substructures in facilitating sequential enzymatic reactions in the RNA processing pathway [165][11].

Most functions associated with MLOs concern the regulation of key processes involving nucleic acids, including chromatin organisation [166], DNA repair [167],

transcription [168][169], RNA splicing [170] and translation [163]. More recently, phase separation has been linked to quality control processes such as proteosomal degradation [171] and autophagy [172], as well as cellular signaling [173].

Recent progress suggests phase separation is a conserved mechanism for cellular organisation and also relevant for bacterial cells [174], even if their proteome has a much lower content of predicted disordered regions [13]. The characterisation of these putative MLOs is hindered by their much smaller size with respect to eukaryotes', for which reason super resolution microscopy approaches are increasingly used to extend conventional *in vitro* studies to the cellular setting [143]. A few examples of proteins/systems with *in vitro* and cellular evidence of phase separation, as comprehensively reviewed by Azaldegui *et al.* , are: mRNA decay mediating BR-bodies [175], RNA polymerase clusters [176], the nucleoid protein Dps [177], the asymmetric pole constituent PopZ [178] and the chromosome segregating ParABS partition system [179] [143].

From a physicochemical mechanistic point of view, modulating chemical reactions [180] [181], reducing cellular noise through buffering of biomolecular concentrations [182] [183] and cellular sensing [184] have been proposed as general functions of MLOs. The fast, adaptable, reversible switch-like behaviour of MLOs provides unique advantages with respect to membrane-bound compartments [185]. However, membrane-bound compartments also confer unique properties such as higher stability and the ability to isolate cytotoxic species and maintain ion gradients, which suggests complementary strengths of the two compartmentalisation strategies [16].

1.4.4 Interactions that drive phase separation

Phase separation is driven by multivalent weak interactions, mainly electrostatic charge-charge interactions, aromatic interactions (cation-π or π-π), and dipole-dipole interactions [186]; a role for hydrogen bonding and hydrophobic contacts has also been proposed [144]. Different interactions have distinct salt dependencies, as shown in Figure 1.10.

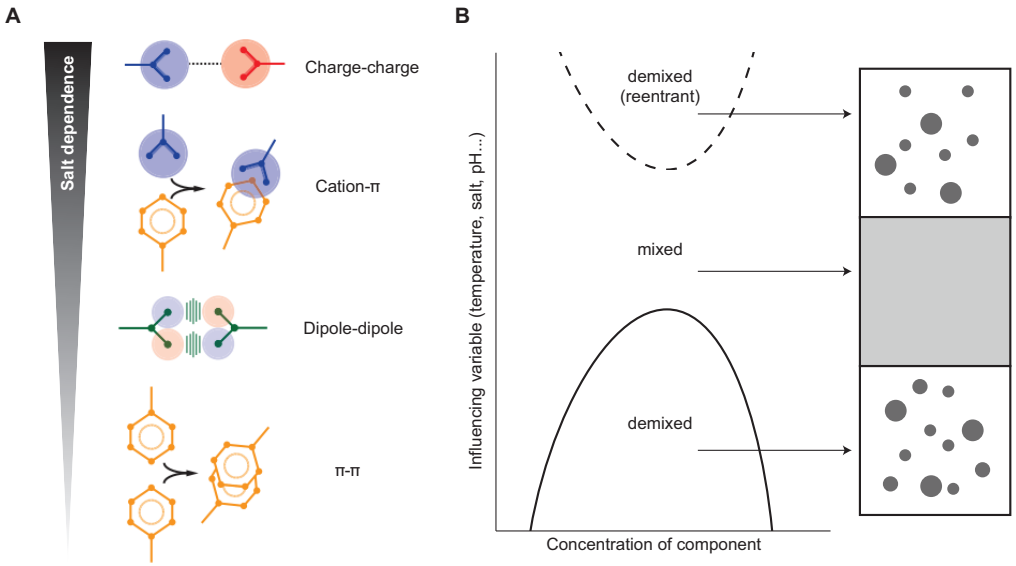


Figure 1.10: Interactions and factors influencing phase separation. A. Taken from [186]. B. Example phase diagram showing phase transition boundaries between mixed and demixed states, showing condensates as circles given their 3D spherical shape due to surface tension. In this system, we would refer to the phase separation occurring above the upper boundary (dashes) as a reentrant phase transition.

Charge-charge Phase separation-driving charge-charge interactions, also commonly referred to as electrostatic interactions, can occur in the context of a single protein containing alternating charges (such as DDX4[148], in which charge scrambling abrogates phase separation) or two oppositely charged biomolecules (such as the coacervation of negatively charged RNA and positively charged tau[187]). Regarding the former, charge-patterning based metrics have been developed to predict phase separation [188].

π interactions π interactions arise from sp^2-hybridised atoms such as those in aromatic compounds, and in the context of protein interactions mainly involve the rings of aromatic residues. π-π interactions occur between two π systems, in a perpendicular or parallel (planar), often off-centered, arrangement. Cation-π interactions occur between a π system and a cation.

Cation-π interactions are abundant in protein structures, with an average of 1 interaction per 77 residues, and specifically 25 % of tryptophan residues being part of one [189]. They are also frequent in protein-ligand recognition, such as in the

paradigmatic example of nicotine binding to the acetylcholine receptor of the brain but not muscle, which is determined by a cation-π interaction in the former case [190], and in particular in protein-DNA/RNA binding [191][192]. Cation-π interactions can be probed by progressive fluorination of the π system, which inverts the quadrupole moment [193], which has also been applied to phase separation systems [148]. On the cation side, methylation of arginine side chains impairs phase separation [148] [194], indicating the role of the interaction as well as revealing PTM-based modulation.

π-π stacking interactions have a stabilising effect on the DNA double helix, potentially contributing more to its stability than hydrogen bond mediated base-pairing [195], and are very prevalent in protein-nucleic acid complexes [196][197]. In the sticker-spacer model proposed based on the analysis of FUsed in Sarchoma (FUS) family proteins, π-π interactions between tyrosines were highlighted as a sticking feature [198]. A recent study by Vernon *et al.* developed a predictor of phase separation based on π-π frequencies, and suggested (based on their results) that the importance of π-π contacts in phase separation has hitherto been underestimated. The authors highlighted that π systems are not only found in aromatic amino acids (tyrosine, tryptophan, phenylalanine, histidine) but also on the side chains of glutamine, asparagine, glutamic acid, aspartic acid and arginine, as well as in peptide backbone amide groups; in fact, planar contacts involving the backbone amide were the most predominant contact in their analysis of π-π in the folded proteome.

Dipole-dipole In the sticker-spacer model, spacers are enriched in glycine, serine and glutamine, and these residues are proposed to control the material properties of the condensate phase [198]. Glutamine in particular contributes to hardening, the progressive loss over time of liquid-like properties which has been seen for MLOs and *in vitro* condensates, which is system dependent and often affected by mutations. This was demonstrated by lack of hardening in a glutamine to glycine FUS variant [198]. In a study of FUS low complexity (LC) domain (QGSY-rich), NMR Nuclear Overhauser Effects (NOE), an NMR signature of through-space

contacts, were detected for intermolecular interactions involving glutamine side chain amide groups and methylene groups [159].

Hydrogen bonds Hydrogen bonds are thought to contribute to phase separation, since most amino acids contain donors and acceptors. Further, hydrogen bonds are key interactions between proteins and nucleic acids, which suggests a role in protein-nucleic acid interactions in condensates [144]. In the study by Murthy *et al.* mentioned above, molecular dynamics simulations showed hydrogen bonds by glutamine (mainly), tyrosine and serine, and the introduction of additional glutamine residues stabilised the phase separated state [159].

The role of hydrophobic residues IDPs/IDRs contain fewer amino acids with hydrophobic side chains than folded proteins, yet there is some evidence for their contribution to LLPS. Helical leucine-rich motifs were shown by NMR to contribute to the multivalent network of P body mRNA decay factors in a phase separating system reconstituted *in vitro* [199]. In another case, the transient formation of an α helix led to interhelix contacts necessary for phase separation, and a mutation enhancing the hydrophobicity of the helix enhanced the phase separation propensity [200].

The LLPS promoting effects of salt for some proteins, such as the FUS LC domain (discussed in the following section), also suggest a role for hydrophobic interactions. The importance of hydrophobic contacts was probed for full length FUS and a PR repeat peptide, which show phase separation at low and high salt concentrations (see a reentrant phase separation diagram in Figure 1.10 B, dashed line). In the high salt regime, only 1,6-hexanediol (which disrupts hydrophobic interactions), but not polyRNA or ATP, was able to inhibit phase separation, while all three could disrupt it at low salt conditions, indicating a role for hydrophobic contacts throughout the range of salt concentrations tested [201].

Finally, it is important to note that in many cases it is very challenging to disentangle the contributions of distinct interactions to phase separation, since there are many amino acids able to engage in more than one interaction [159].

1.4.5 Factors that affect phase separation

Phase separation is modulated by a range of environmental stimuli (salt, temperature, pH), as well as protein concentration and PTMs.

Salt Salt can promote [200][157] or hinder [148][202] phase separation; in a few proteins for which it had been shown to be inhibitory, reentrant phase separation was observed in the molar regime (see Figure 1.10 B) [201]. In the case of FUS low complexity domain, for which salt enhances phase separation, a series of anionic Hofsmeister salts (but not cationic) in a range of 0-1 M were shown to affect phase separation in an analogous way to their effect on protein stability, with kosmotropes (which salt-out proteins) enhancing it and chaotropes inhibiting it [159]. In a high salt regime (above 0.5M), the reentrant behaviour of full length FUS and a PR-repeat peptide was also affected by a series of cationic Hofmeister salts in the same way as protein stability, with higher salt concentration required to induce phase separation for the salts later in the series (more chaotropic) [201].

Multivalent salts present different effects: trivalent cations induce reentrant PS behaviour for acidic proteins [203]; divalent cations have opposing effects on heterotypic peptide-RNA and homotypic RNA condensates, inhibiting the phase separated state of the former and enhancing the latter [204]; and divalent $CaCl_2$ has a neutralising effect on negatively charged DDX4, rather than a charge-screening effect (such as that of NaCl), and promotes LLPS [205].

Temperature Proteins can exhibit phase separation with a lower critical solution temperature (LCST) or an upper critical solution temperature (UCST), depending on whether an increase of temperature stabilises or destabilises the demixed phase separated state, respectively, and in some cases show reentrant phase behaviour [206][144]. UCST phase transitions are explained *via* the increased entropic cost of chain collapse at higher temperatures, which disfavours the demixed state; in the case of LCST, differences in free energy of solvation between amino acid types have been proposed. Indeed, the amino acid composition can be used to predict

which temperature-dependent behaviour a protein will show, with sequences rich in polar and aromatic amino acids tending to show UCST transitions while more hydrophobic sequences have LCST behaviour [207].

pH There are theoretical descriptions of how pH would affect phase separation, especially depending on the content of charged residues and net charge [206] [208]. For the negatively charged $DDX4^{N1}$ domain (the object of study in chapter 5), the LLPS stimulatory effect of lowering the pH could be reproduced by reducing the negative net charge of DDX4 constructs [205], which is expected for anionic polyelectrolytes [206]. Evidence of physiological regulation of LLPS *via* pH is limited, but includes stress granule formation in budding yeast as a response to a decline in cytosolic pH [209], the acidic pH-induction of phase separation of a proinsulin carrier protein [210] and the acidic pH-dependent decondensation of keratohyalin granules which contributes to the formation of the skin surface barrier [211].

Post-translational modifications LLPS is regulated by several PTMs, most notably arginine methylation and phosphorylation (mainly of serine, but also threonine and tyrosine) [212]. Arginine methylation is inhibitory *in vitro* [148] [194] [213], which supports the role of cation-π interactions for LLPS; however, in cellular studies of RNP granules methylation has also been shown to promote LLPS and/or the localisation of proteins into certain MLOs. Phosphorylation has been shown to both promote and inhibit LLPS *in vitro* and in cells.

RNA Many MLO components contain RNA-binding domains, and RNA has been shown to promote LLPS, either specifically [214] or non-specifically [157] [215]. The action of RNA is non monotonic, since it stimulates LLPS at low concentrations and inhibits it at the higher concentration regime [157] [214] (the phase diagram would resemble Figure 1.10 B, with the intermediate state being demixed). This has been attributed to a mechanism of charge inversion (which has also been observed for single stranded DNA) [216]. In addition to effects on the phase separation

boundaries, RNA also tunes the physicochemical properties of MLOs, such as viscosity, both increasing it [214] or decreasing it [202].

1.4.6 Links between phase separation and disease

Phase separation has been linked to several diseases, mainly in neurodegeneration, cancer [217] and infectious diseases [218]. Increasingly, proteins whose aggregation is a hallmark of neurodegenerative diseases have been observed to phase separate, such as tau [219], α-synuclein[220], TDP43 and FUS [221]. *In vitro* studies have shown that phase separation can be followed by a process of maturation, sometimes called hardening, in which the liquid-like properties of the condensates are gradually lost and eventually macroscopic amyloid or amorphous aggregates are formed [16]. In the case of tau, recombinant phosphorylated 2N4R full length tau and brain purified phosphorylated tau formed droplets (with a crowding agent) that rapidly transitioned to a gel state and over time aggregated into amyloid fibrils [219]. Further, in condensates of the tau K18 construct (see Figure 1.6), K18 stayed largely disordered yet showed propensity for β-structure in the hexapeptides and KXGS motifs upon LLPS, suggesting a common structural transition in the LLPS and aggregation of tau [222].

A similar process was observed for Parkinson's related α synuclein, with LLPS proposed to nucleate its aggregation [220]. ALS-linked FUS G156E mutant showed accelerated hardening (shown by reduced coalescence) and formation of fibrous amorphous aggregates, compared to wildtype [147]. ALS-linked TDP-43 mutations disrupted the transient α-helical structure that is necessary for the LLPS of TDP-43's low complexity domain [200].

1.4.7 Can we modulate phase separation?

Targeting phase separation presents unique challenges, because both disordered proteins and biomolecular condensates lie outside what is considered to be *druggable* [223] [224]. One way around this would be to target canonical drug targets, i.e. the enzymes that regulate phase separation and/or client protein partitioning

into condensates [223]. The alternative, targeting the physical chemistry of phase separation, is a promising new avenue that can have distinct strategies, including the selective targeting of transient conformations of disordered proteins and directly targeting the phase separated state using small molecules. Regarding the latter, there is already evidence of small molecules modulating phase separation: a paradigmatic example is 1,6-hexanediol, albeit non-specifically [225].

Small molecules could target the phase separated state by partitioning into it - indeed, the condensate phase has been postulated as a distinct chemical environment to the bulk [148] [152]. The partitioning of small molecules into condensates of two model proteins is the object of study of chapter 5.

1.5 Nuclear Magnetic Resonance

Nuclear Magnetic Resonance (NMR) is a spectroscopic technique based on the alignment of the nuclear spins of certain atoms (usually ^{1}H, ^{15}N, ^{13}C and ^{19}F in biological systems) with an external field. It is extensively used in chemical and biological research due to its sensitivity to the local environment of the nuclei, which provides information on the structure and dynamics of the molecules studied.

1.5.1 The 1D experiment

The basic NMR experiment is a one-dimensional (1D) pulse-acquire (shown in Figure 1.11 in a vector model representation).

Initial state In the beginning of the experiment, because of the presence of an external field (B_0), the nuclear spins of the sample align with or against (α and β states, respectively, for ^{1}H) B_0, and the difference in populations gives an overall net magnetisation aligned with B_0, referred to as z in the vector model.

Pulse application A radiofrequency pulse is applied for a certain duration to get the spins of choice (e.g. ^{1}H) to rotate from the z axis to the transverse (xy) plane.

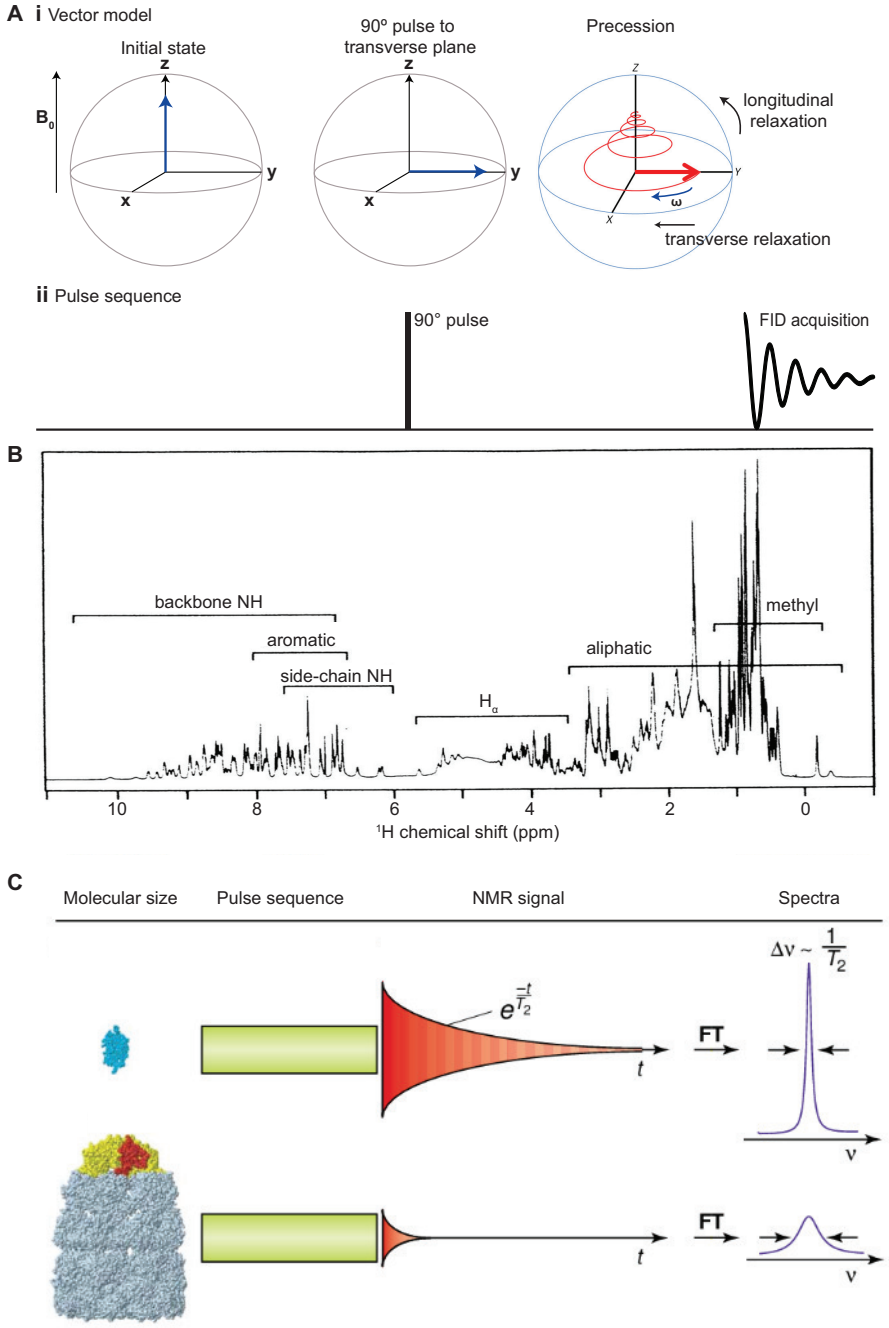


Figure 1.11: Overview of NMR. A. Vector model (top) and pulse sequence (bottom) descriptions of the 1D pulse-acquire experiment. B. Typical ^{1}H spectrum for a folded protein, showing distinct proton environments. Taken from [226]. C. Transverse relaxation causes the broadening of signals as proteins (or complexes) become larger. Taken from [227].

Precession In the transverse plane, magnetisation starts to rotate. This is called *precessing* and it happens at a frequency characteristic of the atom, called the Larmor frequency. Crucially, though, small magnetic fields arising from nearby electrons affect the Larmor frequency, leading to the appearance of the chemical shift, which is the ensemble-averaged deviation (from the Larmor frequency) of the precessing frequency of sets of atoms experiencing a similar electronic environment. We can imagine this using the vector model as several vectors rotating around the z axis, with marginally different frequencies (in the parts per million scale). The relationship between the electronic environment and the precession frequency is one of the aspects that make NMR such a powerful structural technique.

In the case of proteins, for example, distinct chemical shift regions exist for the protons in the amide peptide bond between amino acids, those in aromatic residues, and aliphatic carbons (and CH_3 groups specifically), as shown in Figure 1.11 B for the folded protein ubiquitin. In the amide area, more dispersion of the chemical shifts is observed for folded proteins, with more distinct environments, than for IDPs or unfolded proteins. Changes in the local environment of an active nucleus, for example due to the presence of another molecule nearby, perhaps a ligand, or the sampling of multiple structural states, lead to changes in chemical shift that are thus very informative of structure and dynamics. Further, when magnetisation is transferred between multiple nuclei, in multidimensional NMR experiments, the chemical shifts of ^{1}H, ^{13}C and ^{15}N have signatures of the secondary structure [228].

Acquisition Finally, in the acquisition step, the oscillating magnetisation generate electric currents in a coil, which are recorded. The resulting signal, called the free induction decay (FID), is the sum of all oscillating fields within the system, and needs to be converted to frequency space (to yield the chemical shifts), which is done by Fourier transformation.

The ^{1}H 1D experiment is the basis of the monitoring of aggregation kinetics presented in later chapters. The intensity of the signal is directly proportional to the concentration of spins (and of biomolecules, by proxy), which makes this signal

an excellent observable for quantitative modelling. For a better understanding of this particular application of NMR, it is necessary to discuss the phenomenon of relaxation, a key aspect of NMR. There are two main types of relaxation: longitudinal relaxation, in which the α and β states return to their equilibrium values; and transverse relaxation, a loss of coherence between spins. Due to relaxation, the magnetisation slowly rotates back to the z axis, as well as declines in intensity in the xy plane. Transverse relaxation is proportional to molecular size, which leads to very broad, ultimately disappearing, signals (see Figure 1.11 C). Transverse relaxation poses limits on the size of proteins that can be studied by NMR, for which reason an area of very active method development is that of techniques pushing the size limit of NMR, for example the category of TROSY techniques, by understanding the sources of relaxation and exploiting circumstances in which they cancel each other out [227].

Transverse relaxation is thus often regarded as a negative limiting aspect of biomolecular NMR. However, returning to the study of protein aggregation, there are advantages to this feature. The disappearance of signal from large molecular species leads to only a few observables, which can be well understood and followed. Which specific small species are visible in a kinetic experiment will depend on the size of the monomer, the concentrations of small oligomers and the sensitivity of the spectrometer.

1.5.2 Diffusion NMR

Diffusion NMR experiments combine radiofrequency pulses as described above with the application of gradients along the z axis [229]. In Figure 1.12, we can see that the consequence of applying the gradient is that the spins, represented as net vectors, experience a rotation that depends on their position along the z axis (referred to as "encoding"). Following that, there is a delay (Δ), after which a gradient with opposite effects is applied. In the case that there were no movements during the delay time, the original vectors would be restored to their initial intensity (ignoring relaxation).

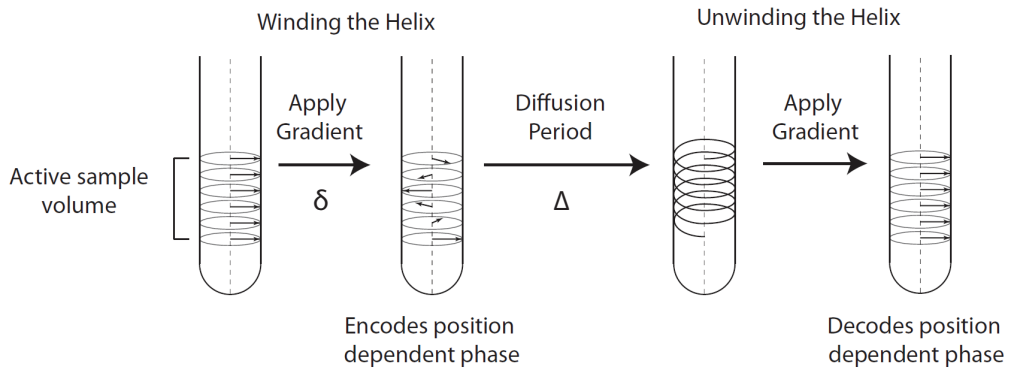


Figure 1.12: Diffusion NMR. The main stages of a diffusion experiment are the dephasing of magnetisation in a position-dependent manner along the z axis, a delay period in which molecules diffuse, and a reverse rephasing of the magnetisation which restores the signal of the molecules that haven't moved in between the gradients. Taken from [230].

In contrast, this experiment is based on the fact that during the delay, biomolecules are expected to diffuse in the solution due to Brownian motion. When the second gradient is applied, some of the magnetisation will not be recovered. Mathematically, this is expressed by the Stejskal-Tanner equation, which states that the signal S at the end of the experiment varies as:

$$S = S_0 e^{-\gamma^2 \delta^2 g^2 (\Delta - \frac{\delta}{3}) D} \tag{1.4}$$

where γ: gyromagnetic ratio, δ: duration of the gradient pulse, g: gradient strength, Δ: duration of the diffusion delay and D: diffusion coefficient.

A series of 1D spectra are acquired at varying gradient strengths, with δ and Δ calibrated to give almost (but not complete) attenuation at the highest gradient strength, and the decay in intensity is fitted to the above equation to yield the diffusion coefficient. The calibration parameters depend on the size of the molecules studied; a protein will require a longer delay and a stronger gradient for the signal to be attenuated, in comparison with a small molecule. However, even if neglected in equation 1.5.2, relaxation is present during the experiment and will reduce the signal. For this reason, it is common to study the diffusion of proteins using a particular type of diffusion experiment, called the pulsed field gradient stimulated echo, in which the magnetisation is stored along the z axis during the diffusion

delay (the longest part of the experiment), which prevents the stronger decay of the signal by transverse relaxation.

Diffusion NMR is thus informative on the molecular size of the biomolecules in the sample. In the case of this thesis, as seen in Chapter 2, we applied it to understanding the identity of the small species observed during aggregation reactions.

1.5.3 Chemical exchange

As discussed above, the chemical shifts of magnetic nuclei within a molecule are very sensitive to the local electronic environment. As a consequence, changes in the environment such as those caused by a binding event cause chemical shift perturbations that can often be observed and quantified in the form of dissociation constants. These are referred to as chemical exchange, given that there are multiple states of the molecule (e.g. free and bound). This is illustrated in Figure 1.13. Depending on the timescale of exchange between the populations, the chemical shifts of both states (slow exchange) or an intermediate chemical shift (fast exchange) are observed. In the intermediate regime, the signals become broadened and hard to detect. Further to this illustration, experiments that probe exchange can be used to obtain structural and dynamical information on high energy states ("excited" states) that are invisible because they exist sparsely and transiently, yet are essential for biochemical function [231].

1.5.4 Applications of solution NMR to the study of protein aggregation

Solution NMR has been extensively applied to the study of protein aggregation, in particular to probe the conformation(s) of monomeric species and detect interactions with large aggregates through exchange methods [233]. However, and perhaps surprisingly given the degree of sophistication in the approaches employed, only recently has NMR started to be used for kinetic monitoring, which as discussed above requires time-series of the relatively "simpler" experiments. These studies include the monitoring of the prion protein [234], amyloid β 40 [235] and 42 [236]

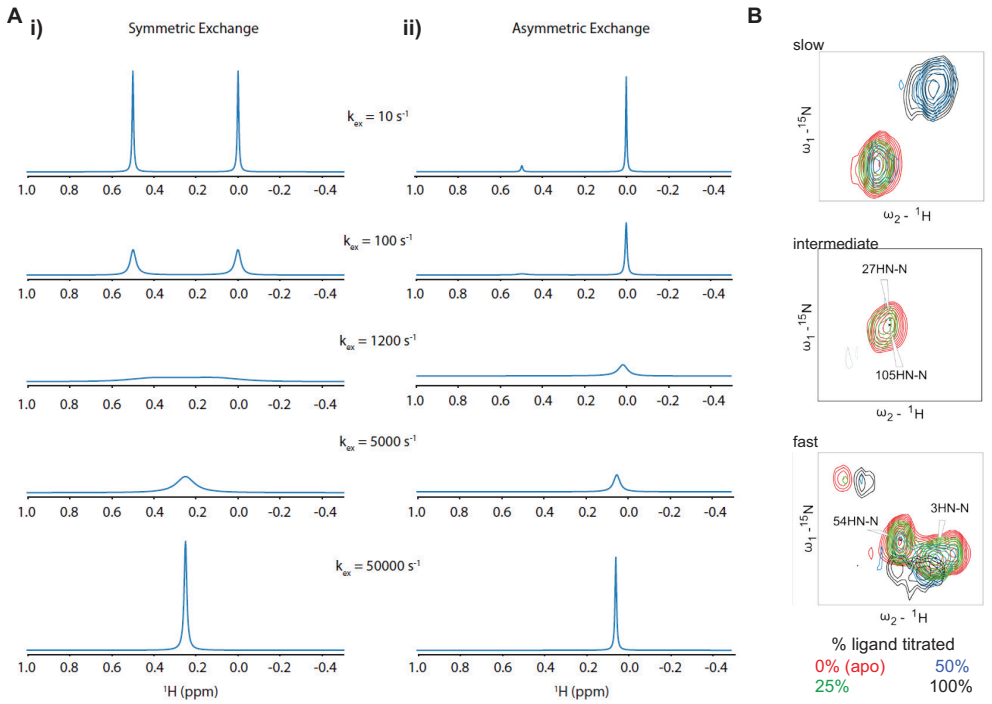


Figure 1.13: Overview of chemical exchange. A. Example of the effects of symmetrical (i) and asymmetrical (ii) exchange on peak shape and chemical shift, depending on the timescale of exchange. In a slow exchange regime, the NMR signals correspond to the two states, in a proportion weighted by the populations. If the exchange is fast, the peak position is located in an intermediate chemical shift position. In an intermediate regime, the signals might be broadened beyond detection. The two states have the same relaxation rates. Taken from [230]. B. Examples of exchange regimes in the amide peaks upon titration of the constant region of immunoglobulin E with a specific antibody. Taken from [232].

[237], human islet amyloid polypeptide [237] and hungtingtin [238], in addition to the study from our lab already discussed [37].

Summary of advantages of NMR versus other techniques for the study of protein aggregation

Table 1.1 shows a comparison of NMR and two other techniques regarding their use for monitoring kinetics of aggregation. Importantly for this thesis, NMR has the unique advantage of providing information on the small species, and therefore the early events in the aggregation pathway. Also, related to the mathematical fitting aspect of kinetics modelling, there is a direct relationship between the signal and the concentration which is very suited to our approach of using exact numerical integration methods.

1.5.5 Applications of NMR to the study of phase separation

Solution NMR is uniquely suited to the study of the conformations and dynamics of disordered proteins, particularly in the context of phase separation [239], in comparison with other classical structural biology techniques such as X-ray crystallography and cryoEM. As discussed in the section above, NMR has shown that several proteins remain disordered during phase separation, as well as detected interactions and motions in the condensate phase [156] [157].

1.6 Thesis overview

This thesis is organised in four chapters. In the first three, I present the results of studying the mechanism of aggregation and inhibition of the protein tau using NMR and chemical kinetics:

- Chapter 2 covers the process of validating R3 as a model system for tau, in terms of biological relevance, reproducibiity of the kinetics and suitability to NMR based monitoring.
- Chapter 3 discusses the mathematical challenges of modelling protein kinetics using exact approaches and presents developments to an existing method.
- Chapter 4 shows the results of applying the modelling strategy described in chapter 3 to the tau model R3 and its inhibition by a literature peptide.

In chapter 5, I focus on the emerging phenomenon of phase separation and in particular on the hypothesis of condensates constituting distinct solvent environments. I present the development of NMR-based partition coefficient measurements and partitioning trends of nucleotides into a model system of a biomolecular condensate.

2

R3 as a model system for tau aggregation

2.1 Abstract

In this chapter, we have assessed two potential tau constructs as model systems amenable to NMR-based chemical kinetics studies of aggregation mechanism and inhibition by small molecules. The constructs tested were the third repeat of the microtubule domain, "R3", and the third and fourth repeats, "R3R4". Both contain one of the aggregation hotspots, the hexapeptide ^{306}VQIVYK^{311}, while R3R4 corresponds to the resolved core in cryoEM structures of Alzheimer's brain fibrils. The two constructs showed many similarities: they reproducibly aggregated in a timescale of days, in a cofactor-free way, and aggregation was inhibited by literature peptide D-TLKIVW. However, R3R4 aggregation was very slow in the absence of agitation, which hinders the temporal resolution that can be achieved using NMR due to the need to shuttle the sample between the spectrometer and a shaking device. For this reason, we focused on R3 as a potential system. We could detect two species with distinct chemical shift for the two histidine residues, and which corresponded to monomer and dimer as assessed by diffusion NMR. Dimerisation and aggregation were inhibited in the presence of DTT, suggesting an on pathway role for a disulfide linked dimer at position Cys322; the need for oxidising conditions was bypassed in the case of seeded elongation. Further, R3 aggregates acted as seeds for aggregation

of full length tau in HEK293 cells. Overall, R3 fulfilled the requirements for a model system that we set out to meet, and a dataset of aggregation curves was acquired by NMR to enable chemical kinetics fitting to models.

2.2 Introduction

Aggregation studies of tau often require the use of constructs, due to the low aggregation propensity of the full length wildtype protein, as well as inducers and agitation conditions [69] [81] [240], the latter of which are especially expected to affect the kinetic mechanism. The finding that heparin-induced recombinant fibrils diverged in structure from Alzheimer's brains fibrils [76] has prompted the search for cofactor-free aggregation models [78] [79]. Additionally, fluorescence dye based monitoring of aggregation, though commonly used, possesses some limitations in terms of the difficulty to interpret the signal in terms of protein concentration, the lack of signal from non-amyloid species preventing detection of early binding/nucleation events, and the requirement of potentially interfering dyes [33]. Recently, NMR has been used to monitor aggregation kinetics [37] [238], with the advantages of being a label-free method with a direct signal-to-concentration relationship, which allows the observation of early events in the aggregation process.

For the above mentioned reasons, we chose NMR as the method for monitoring kinetics of aggregation of tau, and aimed to find a suitable tau model system. Such a system would be defined by:

- Aggregation on a suitable timescale for the experimental set-up
- Reproducible kinetics
- Aggregation in the absence of cofactors or agitation
- Biological relevance

Regarding the biological relevance, we chose two constructs to test given their importance within the full length tau protein isoforms. Briefly, as discussed in the introduction, the microtubule binding domain aggregates faster than the full length

protein, and in particular two sequences of six residues, at the beginning of repeats 2 and 3, constitute hotspots for beta-sheet formation and aggregation. Repeat 3 (tau306-331), unlike repeat 2, is present in all isoforms, and Alzheimer's disease has a mixed content of 3R isoforms (lacking R2) and 4R isoforms (containing all four repeats) (see Figure **1.6**). Further, the repeat 3 had already been found to self-aggregate in the absence of heparin and shaking, even if more slowly, and form amyloid as detected by fluorescence of thioflavin T [241]. The other construct tested, called "R3R4" (tau306-378), encompasses repeats 3 and 4, plus an additional 6 C terminal residues, and was chosen since it corresponded to the protofilament core in Alzheimer's patient brain derived fibrils as resolved by cryoEM [65].

2.3 Methods

2.3.1 Contributions

All experimental procedures were carried out at Eli Lilly facilities (Erl Wood Manor, UK) by our collaborators Gary Sharman and Sylvain Demanze (NMR diffusion experiments and NMR-monitored kinetic runs, including sample preparation, set-up and analysis), Suchira Bose, Annalisa Cavallini, Caroline Kerridge and Sarah Pearce (ThT aggregation assays and cellular seeding assays). Results implications and project directions were discussed with all above.

2.3.2 Materials

The following antibodies were kind gifts from Peter Davies (Albert Einstein College of Medicine, New York): MC1 (Jicha et al., 1997) and PG5 (Ser(P)409(Jicha et al., 1999). R3 and R3R4 peptides were purchased from CPC Scientific. These have the following sequences:

R3:

$$^{306}VQIVYKPVDLSKVTSKCGSLGNIHHK^{331}$$

R3R4:

$^{306}VQIVYKPVDLSKVTSKCGSLGNIHHKPGGGQ$

$VEVKSEKLDFKDRVQSKIGSLDNITHVPGGGNKKIETHLLTF^{372}$

The numbering refers to the 2N4R isoform, as per convention (UniProt ID P10636-8).

2.3.3 Thioflavin T monitored kinetic assays

R3 or R3R4 were resuspended in phosphate buffered saline (PBS) buffer containing 15 µM ThT in a 384 well plate., using dilutions to yield the desired concentrations of peptide. Triplicate experiments for each concentration were performed. Samples were incubated at 37ºC, with shaking at 1200 rpm, and the Thioflavin T fluorescence measured at various time points using an Envision plate reader at ex/em 405/525nm. The plate was incubated until the Thioflavin T signal plateaued.

2.3.4 NMR

Assignment of R3 peptide

R3 was assigned in water (under conditions in which it didn't rapidly aggregate) at 25 ºC.

Experimental kinetic run set-up

A 1 mM solution of R3 peptide was prepared in pure water (to prevent aggregation). For reproducibility, the solution was aliquoted into Eppendorf tubes (54 µl per tube tube), and the aliquots frozen at -80 ºC. Aliquots were removed from the freezer as needed and thawed. 492 µl of 50mM PBS, pH 7.4, was added along with 5.4 µl of DMSO (control) or inhibitor stock in DMSO for inhibition studies. The sample was mixed, transferred to a 5mm NMR tube and placed into the NMR spectrometer with temperature control pre-set to 37 ºC. The delay between mixing R3 with buffer and the first NMR data point was recorded.

R3R4 assays A similar protocol was followed for the R3R4 kinetic assay in the absence of shaking. For the assay with shaking, in between the NMR acquisition points the sample was transferred to a microcentrifuge tube and shaken at 1200 rpm in a Block Thermal Mixer (Thermo Scientific) at 37 ºC.

Assessment of DTT effects To test the effects of DTT on primary and seeded aggregation, two experimental variations were performed. In the first sample, (primary aggregation), an aggregation sample of 100 µM R3 was as described above, but with PBS buffer containing 1 mM DTT. In the second sample (seeded aggregation), a thawed fresh R3 aliquot was added to the recovered sample of a run with 90 µM R3, which contained aggregates that act as seeds, and DTT was added to 1 mM final concentration.

Aggregation assays

NMR data were recorded on a Bruker Avance III 600 MHz instrument. Samples were shimmed with topshim and tuned using automatic tuning and matching. Spectra were acquired using composite pulse presaturation solvent suppression. Spectra were all acquired under identical conditions of 1024 scans, each experiment taking approximately 20 minutes.

Kinetic data processing

Data were processed using MNova software. They were imported into a stacked plot with 1:10 decimation. The same manual phase correction was applied to all spectra, followed by automatic spline baseline correction. Relevant peaks were identified by a window of chemical shift, and intensities extracted using a MNova script (Mestrelab). The previously measured delay was added as an offset to the time dimension.

Diffusion

Set-up and analysis The Bruker pulse sequence "ledbpgppr2s" (stimulated echo, with water presaturation) was used with diffusion delay Δ of 0.2 s and gradient pulse length δ of 4 ms. Data analysis was done using MNova (Mestrelab).

Hydrodynamic radii relationships with diffusion coefficients The relationship between the diffusion coefficient and the hydrodynamic radius (the radius of the corresponding hard sphere which diffuses at the same rate as a particular molecule) can be used to interpret the diffusion data on aggregating species in terms of their oligomer state, by using the appropriate power-law scaling factor for the case of an intrinsically disordered protein. According to the Stokes-Einstein equation:

$$D = \frac{k_B T}{6\pi\mu R_h} \tag{2.1}$$

where D: diffusion coefficient, k_B: Boltzmann constant, T: temperature (in K), μ: solvent viscosity and R_h: hydrodynamic radius.

If we consider multiple oligomeric states within an experimental sample, all parameters in the Stokes-Einstein equation have the same values apart from the R_h. R_h depends on the number of amino acids and the degree of compaction of the protein according to the following power-scaling law, established empirically[242]:

$$R_h = R_0 N^v \tag{2.2}$$

where N is the length of the sequence and R_0 and v are constants fitted from sets of proteins, and which vary for folded (0.29) unfolded/chemically denatured (0.55-0.57) and intrinsically disordered (0.51) proteins [242] [243].

Therefore we can express the ratio of diffusion coefficients of two species 1 and 2 as:

$$\frac{D_2}{D_1} = \frac{R_{h1}}{R_{h2}} = \left(\frac{N_1}{N_2}\right)^v \tag{2.3}$$

For a scaling v of 0.51, since Tau is an IDP, assuming the second species is a dimer and replacing $N1/N2$ by 0.5, we obtain a value of 0.707 for the ratio of diffusion coefficients, which is reported in Figure 2.4.

2.3.5 Seeding with aggregated R3 peptide in a cell-based assay

Seed preparation

Lyophilised R3 peptide was resuspended in PBS containing 1 mM DTT and 15 µM Thioflavin T in a 384 well plate. Samples were incubated at 37ºC, shaking at 1200rpm and the Thioflavin T fluorescence measured at various time points using an Envision plate reader at ex/em 405/525nm. The plate was incubated until the Thioflavin T signal plateaued. Aliquots were stored at -80ºC until further use in the cell based assay.

Cellular assay

HEK293T cells expressing human 1N4R P301S Tau were used to assess the ability of the R3 fibrils to induce aggregation in a model system [244]. Cells were grown in 96 well poly-D-lysine coated plates and induced to express 1N4R P301S Tau with 1 µg/ml tetracycline. Subsequently cells were incubated with R3 aggregates (1, 0.5, 0.25 and 0.125 uM) diluted in complete cell media (DMEM plus 10% fetal calf serum). After 3 hours, the inoculum was removed, washed with PBS to remove exogenous aggregates, and replaced with fresh complete medium. Cells were maintained for a further 3 days at 37 ºC and 5% CO2 in a humidified incubator. Cells were fixed with Glycofix containing 1% Triton X-100 for 20 minutes and intracellular Tau inclusions detected by immunofluorescence.

Immunofluorescence

Fixed cells were washed thoroughly with PBS and blocked with Odyssey blocking buffer prior to incubation with primary antibodies. The antibodies used were MC1 (1:8500) with anti-IgG1 conjugated to Alexa Fluor 568 (1:1000) and PG5 (1:1000) with anti-IgG3 conjugated to Alexa Fluor 488 (1:1000). Washings were carried

out with PBS and antibody solutions prepared with Odyssey blocking buffer. Hoechst staining was used to visualize nuclei. Confocal images were acquired using an Olympus Fluoview 1000 with a 60x oil objective. An Acumen imager was used for high-content imaging.

2.4 Results and discussion

2.4.1 R3 and R3R4 self aggregate in PBS

The constructs R3 and R3R4 aggregated reproducibly in PBS, without the addition of cofactors, as can be seen by monitoring ThT fluorescence (Figure 2.1). There was an initial lag phase, corresponding to early nucleation events (longer for R3R4), followed by a rapid increase in signal and a plateau.

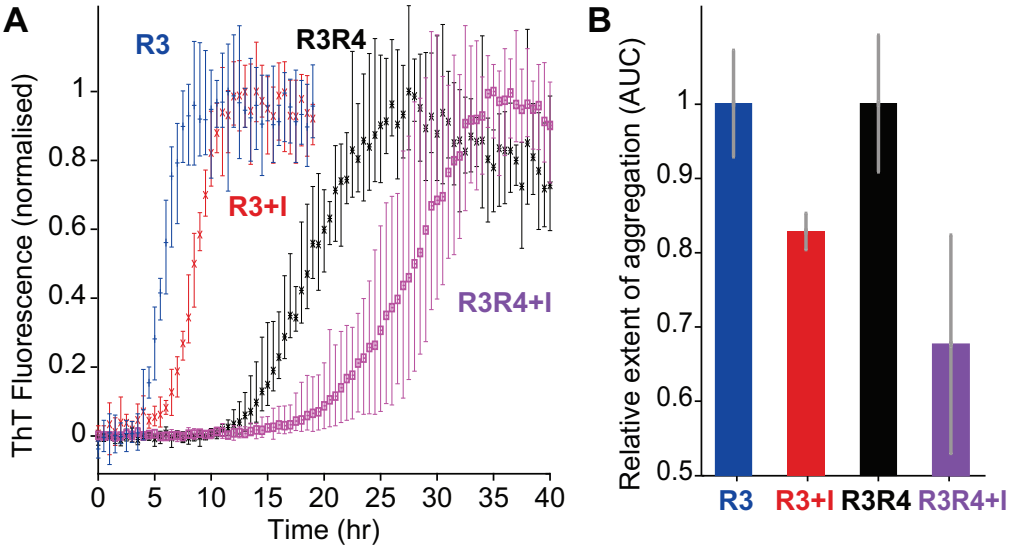


Figure 2.1: Aggregation of R3 and R3R4 followed by thioflavin T fluorescence. A. Kinetic traces of R3 and R3R4 aggregating under shaking conditions, at a concentration of 100 µ, also showing the effects of D-TLKIVW (R3/R3R4 + I traces). B Quantification of the extent of inhibition, assessed by calculating the average area under the curve (AUC) (n=6) of the traces (normalised to the highest point) with respect to the values without inhibitor. The colours of the traces correspond to those in A. Raw data provided by Eli Lilly.

The literature inhibitor D-TLKIVW, found to inhibit aggregation of the K12 and K18 constructs, was inhibitory for both constructs, to a higher extent for R3R4 as can be seen by a longer lag phase before the start of the increase in fluorescence.

2.4.2 Aggregation of R3 followed by NMR

The fragment R3 readily aggregates without shaking, as can be seen in Figure 2.2. From a qualitative analysis of the NMR data, it is clear that there are two species visible, indicated by two very distinct kinetic patterns: one in which the signal decays from the beginning, and another in which there is a build-up and decline of the signal. These patterns can be seen in both amide and methyl regions and correspond to a monomeric and dimeric species as discussed in the next section. For each species, there is excellent agreement in the evolution of intensity of the doublets. The agreement between amide and methyl regions is better for the dimer peaks, while the methyl region monomer peaks have a slightly higher intensity than the amide peaks. This is probably due to the fact that the methyl area has more overlap in amino acid resonances; in contrast, the peaks in the aromatic region can only correspond to histidines($\sim$7.8 and 7 ppm) and one tyrosine (doublets at $\sim$ 7.1 and 6.9 ppm). In particular, the four peaks in the 7.77-7.83 ppm region (see Figure 2.2 B) correspond to one of the imidazole protons of histidine residues H329 and H330, with the main peak and a downfield shoulder corresponding to the dimer for each histidine. The 7.8 ppm was chosen to monitor monomer intensity and the 7.77 for dimer, respectively, for kinetic traces of a larger dataset.

Further, we can see there is a good match between the timescale of aggregation and the experimental set-up. Each data point corresponds to 20 minutes of acquisition, and there are no steep changes in signal intensity where data is not available.

2.4.3 R3 stock samples are disordered and contain monomeric and dimeric species

The fragment R3 was investigated by NMR in non aggregating conditions. First, the assignment showed that there was no secondary structure, as indicated by very small deviations for random coil values for most amino acids (Figure 2.3).

The oligomeric nature of the species yielding the two traces was investigated by diffusion NMR using a sample in water (non-aggregating conditions), corresponding

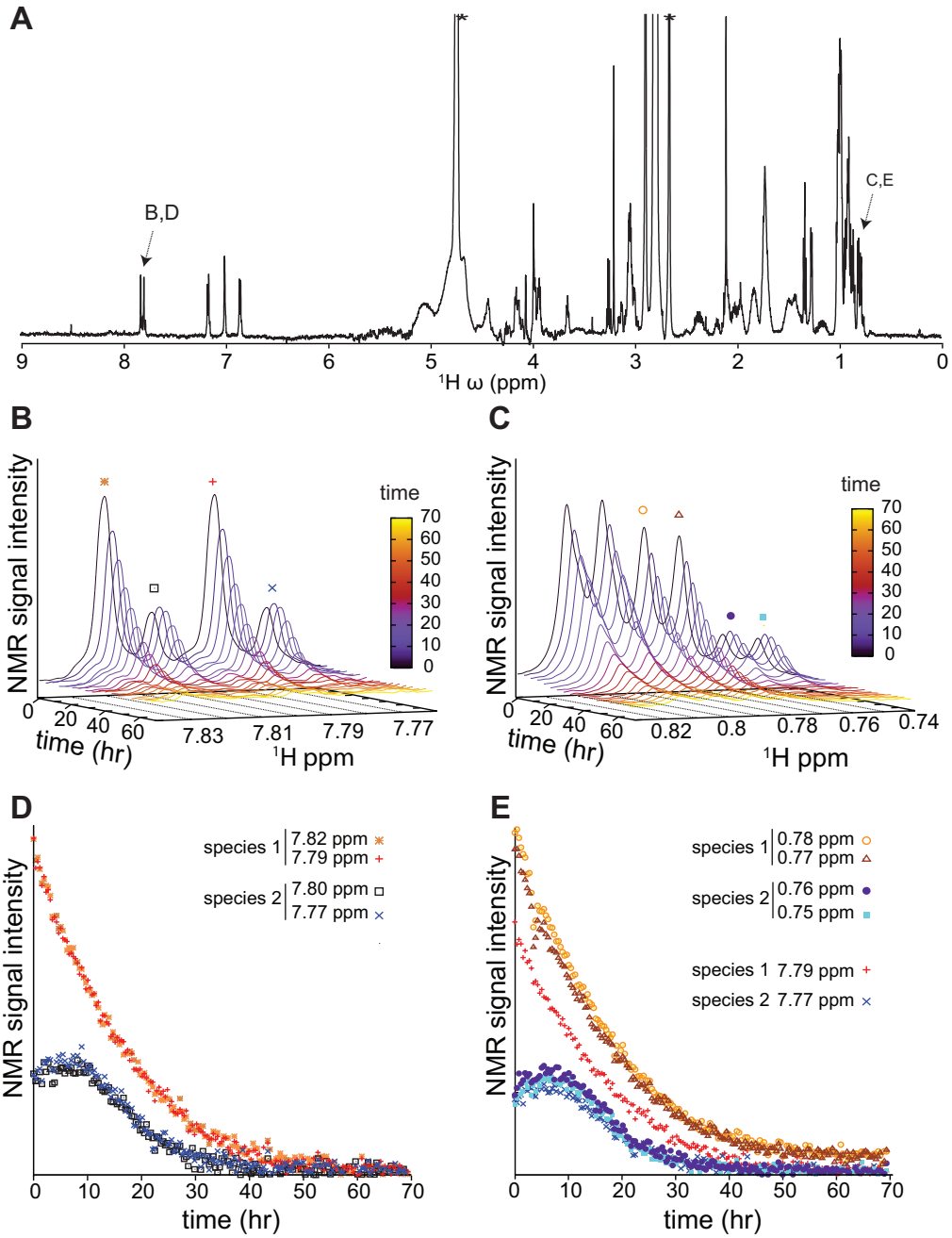


Figure 2.2: Aggregation of R3 followed by NMR. A. ^{1}H spectrum of R3 peptide. B, C. Waterfall plots of ^{1}H spectra during a typical aggregation run of the R3 fragment (B: aromatic region, C: methyl region). D,E. Kinetic traces corresponding to the ppm resonances shown in B and C, quantified as the intensity of the highest point within a small range around the peak maximum at t=0h. Raw data provided by Eli Lilly.

to the initial equilibrium (Figure 2.4). The diffusion coefficients were fitted to $1.48 \cdot 10^{-6}$ cm^{2}s^{-1} for species 1 and $1.03 \cdot 10^{-6}$ cm^{2}s^{-1} for species 2. By considering the power-scaling law that links the hydrodynamic radius and length of sequence,

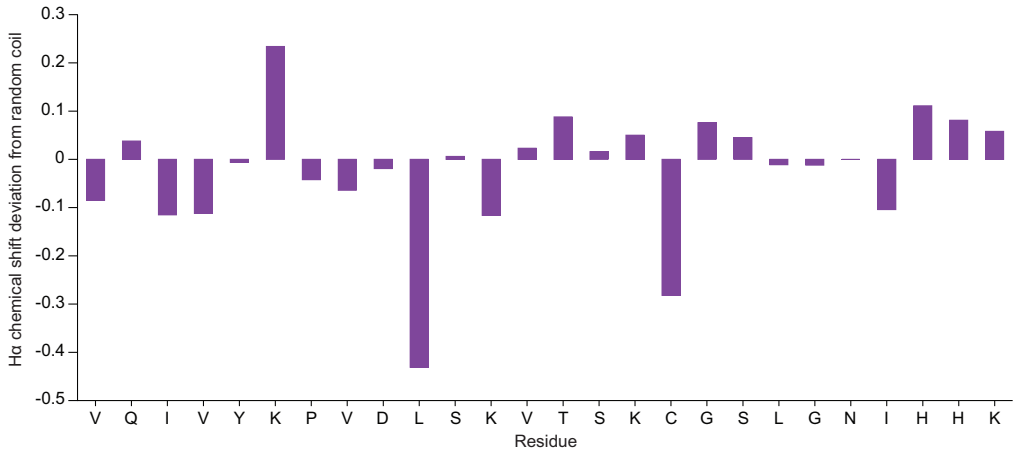


Figure 2.3: Lack of secondary structure of an R3 sample in non aggregating conditions. The deviation shifts were calculated by comparing to reference literature values. Data and analysis provided by Eli Lilly.

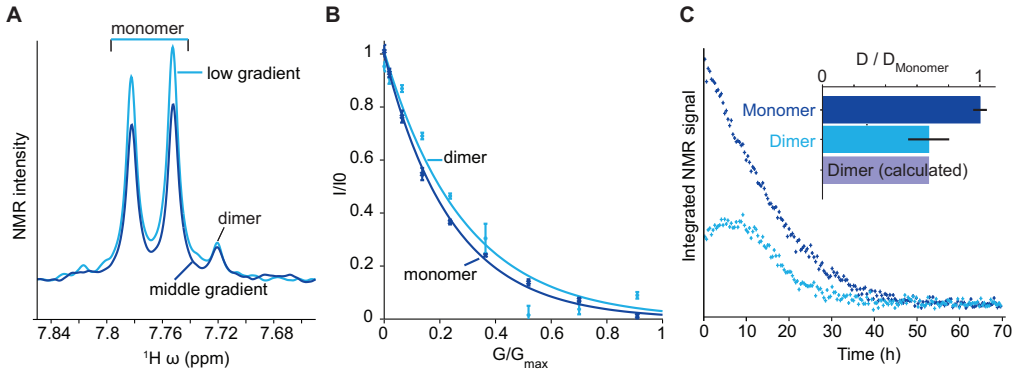


Figure 2.4: Diffusion analysis of R3 in non aggregating conditions. A. NMR spectra of the aromatic histidine resonances, showing a faster decay of monomer signal compared to dimer signal, as expected from a smaller species that has diffused further away from the initial position during the delay time before the signal is refocused. B. Exponential fitting of the signal decay as a function of relative gradient strength. C. Measured diffusion coefficients for monomer and dimer compared to the diffusion coefficient expected for a dimer of an IDP (see Methods). Raw data provided by Eli Lilly.

in particular with scaling for IDPs, we found that the ratio of diffusion coefficients agreed with a monomeric species for the decaying resonances in Figure 2.2 and dimeric species for the trace that builds up and decays.

2.4.4 R3R4 doesn't aggregate in non-shaking conditions

In the case of R3, both monomer and dimer traces substantially decay over time to a plateau close to zero, indicating that almost all the small species present initially have been incorporated into large oligomers or fibrils that are not visible

by NMR due to their large size and relaxation properties. For R3R4 (Figure 2.5), similar monomer/dimer features are observed but the monomer and dimer plateau at around 25% and 90% of their maximum intensity, suggesting they have reached an equilibrium but are not progressing into larger aggregates at a significant rate that would correspond to the quick consumption of monomer expected for elongation reactions. This is in line with findings by Carlomagno *et. al*, who showed aggregation of this construct in the absence of cofactors only with agitation, also at 37 ºC (even if the temperature requirement was more flexible and aggregation also happened at room temperature) [245].

We next explored a hybrid shaker-NMR assay for R3R4 in which the sample was manually transferred from a shaker to the spectrometer at several points in the reaction, which can be seen in the trace with five datapoints in Figure 2.5. As expected, the timescale of decay of the total signal (∼30h) matched the build-up of fluorescence in Figure 2.1.

However, given the experimental challenges in running this hybrid assay over several days, plus the expected larger variability and lower temporal resolution, we decided this was not a suitable system to acquire a dataset for quantitative, modelling and chose to focus on R3.

2.4.5 The aggregation of R3 in non shaking conditions is reproducible

We had obtained good reproducibility in ThT monitored aggregation assays of R3 (see Figure 2.1), however it was necessary to verify that this would also be the case in non-shaking conditions and in a different experimental set-up, given how sensitive aggregation experiments can be to many sources of experimental variability. Figure 2.6 shows that there were small variations between two replicates of 100 µM, as expected, but the reproducibility was very good in the context of aggregation kinetics studies in the literature.

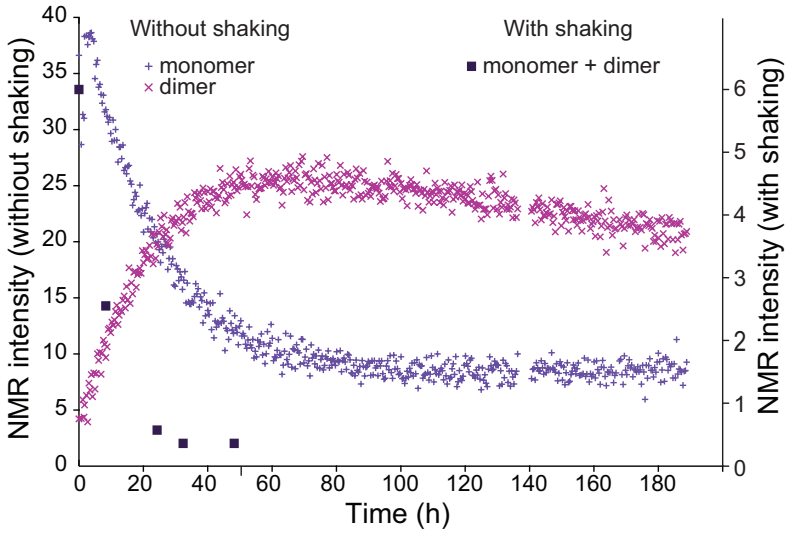


Figure 2.5: Aggregation of R3R4 with and without shaking. NMR monitored aggregation assay of 100 µM R3R4 fragment (red: monomer, blue: dimer). Without shaking, dimer is formed but aggregation does not proceed. With shaking, the decay of total intensity takes around 30h, similar to the build-up in Figure 2.1. Raw data provided by Eli Lilly.

.

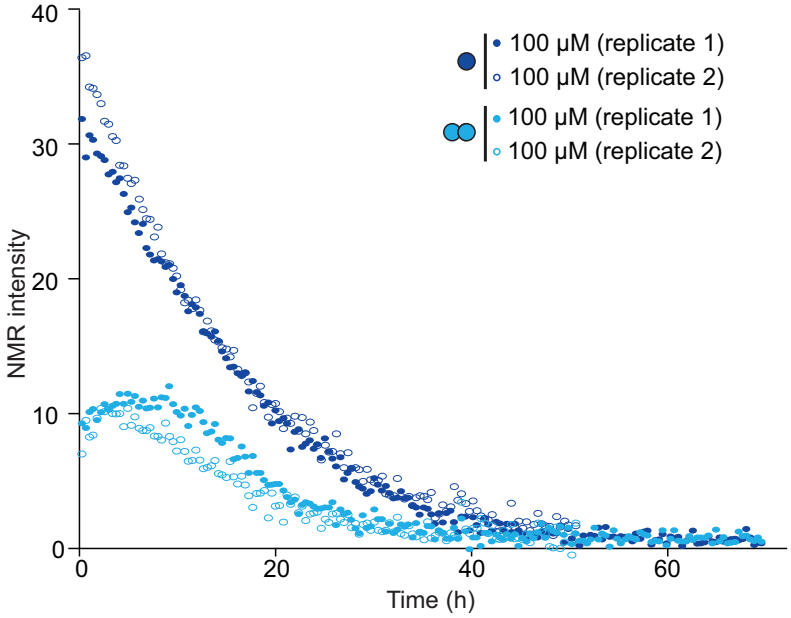


Figure 2.6: Comparison of two 100 µM R3 NMR kinetic experiments. Two replicates of an aggregation experiment are shown overlaid. Raw data provided by Eli Lilly.

.

2.4.6 The formation of dimer occurs in oxidising conditions and is essential for primary aggregation of both R3 and R3R4

Given the literature concerning dimerisation via the cysteine residue in R3 (Cys322), we investigated the effects of a reducing environment on dimerisation and overall aggregation by using dithiothreitol (DTT). When DTT was added to the resuspension buffer, the amount of dimeric species present initially (approximately 15 minutes after onset of aggregation) was much lower and further decreased with time, indicating the initial pre-equilibrium of monomer-dimer is disrupted by DTT (see Figure 2.7 A). Additionally, aggregation was greatly abrogated, with the monomer trace staying at around 80 % of the initial value after 65 h (compare with an almost complete decay in panels D and E in Figure 2.2). Consistently, the monomeric trace decay was much less pronounced, since levels of monomer stayed around 80% of the initial values. We then tested whether the oxidising environment was required at all stages of aggregation by performing a seeded assay with preformed R3 aggregates, in the presence of DTT. In that case, there was no detectable dimer present, but the decay of monomer proceeded much faster, plateauing earlier than primary aggregation without DTT. This is consistent with seeded aggregation bypassing nucleation events, and indicates that elongation is not dependent on either a substantial amount of free dimer nor oxidising conditions.

We also tested the effects of DTT on the longer R3R4 model system, using a fluorescence based assay with shaking, and found similar inhibition of aggregation to that observed with R3 (see Figure 2.7 B).

R3 and R3R4 behave similarly to literature 3R constructs in heparin-induced aggregation reactions In the literature, dimers (and higher order) species have been observed under oxidising conditions by native gel electrophoresis for 3R constructs (lacking the second repeat, e.g. K12), which only contain one cysteine (Cys322), while in the case of 4R constructs (e.g.K11) two bands corresponding to different degrees of compaction of the monomer were observed [119]. When

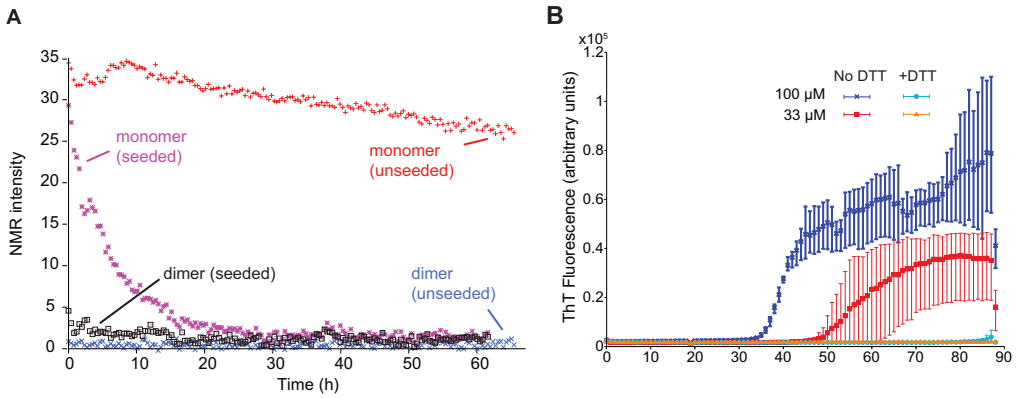


Figure 2.7: Effect of DTT on the aggregation of R3 and R3R4. NMR monitored aggregation assay of 100 µM R3 fragment with 1 mM DTT (red: monomer, blue: dimer) and with addition of preformed aggregates to seed the reaction (purple: monomer, black: dimer). ThT monitored aggregation assay of R3R4 at concentrations of 33 and 100 µM R3R4, with or without DTT. Raw data provided by Eli Lilly.

the cysteine was mutated to alanine, or under reducing conditions, the dimer band disappeared, confirming the presence of an intermolecular disulfide bond. In the 4R K11 construct, the compact monomer was the predominant species under oxidising conditions, while a reducing environment led to a majoritary extended conformation, suggesting an alternative compact conformation held together by an intramolecular disulfide bond. The aggregation of the two constructs followed opposite trends with redox conditions; K12 aggregated only in oxidising conditions and was inhibited by reducing agents or the mutation Cys322Ala, while K11 aggregated only in reducing conditions or when one of the cysteines (Cys291, in repeat 2) was mutated to alanine. These distinct effects of the redox environment on 3R versus 4R constructs were only observed when using heparin as the inducer, but not with arachidonic acid [**Barghorn2000**].

In the course of this PhD project, a few studies have reported on redox conditions effects for similar model systems to ours and in a cofactor-free manner, including the closely related entire R3 repeat(tau306-336, containing an additional 5 C terminal residues)[246] [247], and for the same R3R4 cryoEM core sequence[245], as well as a 3R construct (dGAE;297-391)[79].

For the 306-336 R3 construct, aggregation was faster in oxidising conditions than using a reducing agent or mutating cysteine 322 to serine; furthermore, chemical

disulfide bond formation prior to aggregation greatly accelerated the kinetics (yet reduced seeding efficiency to much lower levels, lower than the aggregates formed in reduced conditions which had much slower aggregation as measured by fluorescence) [246].

Annadurai *et al.* reported that R3 but not R2 was able to self-aggregate without heparin, but that the aggregation of the C322A mutant, unable to form disulfide bonds, was rescued by heparin; the effects of reducing agents were not explored further [247]. Both R2 and R3 formed dimers in the presence of heparin, but not at time 0. Their assay set-up method was similar to ours (resuspension of lyophilised peptides in water, stored at -80ºC), however the lack of timepoints between 0 and 24 hours and detail regarding the use of native or SDS denatured elecrophoresis prevents further discussion of this potential difference in monomer-dimer pre-equilibrium populations.

In contrast, in the case of R3R4, our results differ with those of Carlomagno *et al.*, given that they performed aggregation in reducing conditions (1 mM TCEP) and it proceeded with very similar kinetic features to ours (in terms of time to plateauing of fluorescence signal and disappearance of most of the protein from the supernatant after centrifugation)[245]. This is surprising given that the buffer conditions were not very different (PBS, pH 7.4 in our assay; 10 mM HEPES, pH 7.4, 100 mM NaCl in theirs) and the temperature was 37ºC in both cases, suggesting a dependency on buffer and/or salt. Finally, another proof that the earlier 3R vs 4R explanation of redox effects might be too simplistic came from studies with the dGAE construct (tau297-391), which starts with the last R2 4 residues, and includes 20 residues past repeat 4, and is one of the protease(pronase)-resistant constructs derived from Alzheimer's paired helical fibrils [248]. The aggregation of dGAE was qualitatively found to speed up in the presence of DTT and even further in the mutant Cys322Ala, which led the authors to suggest there may be a competition between on-pathway and off-pathway dimers[79]. As a follow up to that finding, an experimental screen of aggregation conditions of dGAE and similar constructs, combined with cryoEM resolution of the structures, was carried out in

always reducing conditions (10 mM DTT) and found to yield fibrils with the same structural cores as *in vivo* fibrils in AD and other neuropathies [77].

In our case, both R3 and R3R4 are behaving like a 3R construct in the classical sense, R3 in agreement with the closely related 306-336 construct [246]. The observation of a redox-dependent monomer-dimer pre-equilibrium of R3 is in agreement with literature, with the caveat that we cannot compare the timescale of the monomer-dimer conversion in our data the previous studies [119] [246] due to lack of experimental details.

Why is the R3R4 dimer not progressing to aggregation in the absence of shaking? Carlomagno *et al.* reported that R3R4 did not form fibrils (visible either by EM or thioflavin fluorescence) in the absence of shaking (but with reducing agent, as discussed above). In our case, under oxidising conditions and in the absence of shaking, we have observed NMR signatures of both monomer and dimer populations. 2.5. We speculate that either R3R4 can form both on-pathway and off-pathway (disulfide-linked) dimers with the equilibrium affected by agitation, or that there are higher conformational constraints (compared to R3) for on-pathway dimers to proceed in the next steps towards fibrillisation and which are removed by agitation. Further work would be required to investigate this behaviour.

2.4.7 R3 can seed full length tau in cells

In a study exploring the seeding capability of R3 related constructs as a proxy for biological relevance, it had been found that the entire repeat (306-336) had strong seeding ability in a HEK cell line stably expressing the four repeat construct (244-372; with mutations P301L and V337M), with 50% cells showing aggregates, but that shorter constructs including our model system (306-331 and 306-326) also had seeding ability, with 10% of cells showing aggregation[246]. On the other hand, smaller constructs, most notably including the hexapeptide ^{306}VQIVYK^{311}, did enter the cells but failed to seed the microtubule binding domain.

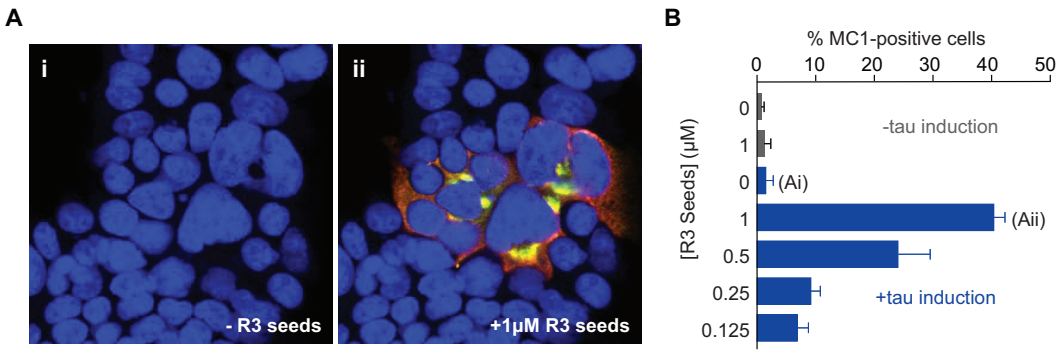


Figure 2.8: Seeding of full length tau by R3 aggregates. A. Representative seeding assays with HEK293 cells expressing human P301S 1N4R tau. (i) Negative control without seeds. (ii) Seeding with 1 µM R3 seeds). B. MC1-positive tau inclusions measured by high content imaging showing dose-dependent inclusion formation with R3 aggregates upon induction of tau expression. Data and analysis provided by Eli Lilly.

We confirmed this capacity for the R3 aggregates formed in our experimental conditions, using a well-established cellular assay with transiently expressed full length (1N4R) P301L tau [244]. HEK293T cells overexpressing tau were exposed to pre-formed R3 amyloid fibrils for 3 days, resulting in MC-1 positive (red) and PG5 positive (green) tau inclusions which can be seen in Figure 2.8 Aii as merged yellow signal. Staining for these antibodies indicates misfolded tau (MC-1) and the presence of hyperphosphorylated aggregates (PG5), both of which are toxic. Seeding of full length tau was concentration dependent and did not occur without induction of tau expression (see Figure 2.8 panel B) or by expression of tau without addition of seeds.

Interestingly, both R2 and R3 aggregates formed in the presence of heparin have been shown to have cell seeding capability, but that was only the case for R3 (and not the R3 C322A mutant) for heparin-free seed formation [247].

Overall, our results and those of others highlights the relevance of the R3-derived construct 306-331 in the cellular context of longer constructs (such as a microtubule binding domain construct) and full length.

2.5 Conclusions

In this chapter, we aimed to test two potential model systems for tau aggregation, given the low aggregation propensity of full length tau, in cofactor-free and shaking

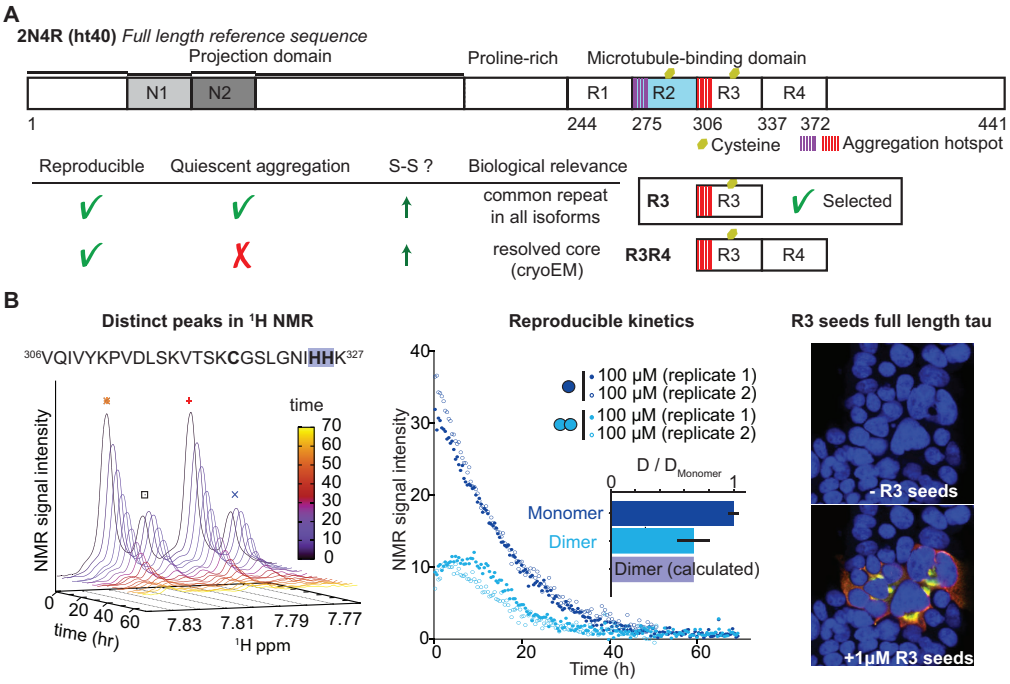


Figure 2.9: R3 as a model system for tau aggregation: summary of features A. Structure of full length tau (2N4R) and schematic representation of constructs R3 and R3R4, explored in this chapter as potential model systems for aggregations, as well as main findings for each construct. B. Features of R3 aggregation. R3 aggregation can be monitored by following the aggregation intensity of the C-terminal histidine residues, which have distinct chemical shift signatures for monomer and dimer oligomeric states as probed by diffusion NMR. The aggregation of R3 is reproducible in those conditions, and produces aggregates that can seed the aggregation of endogenous full length tau in cells, an indication of biological relevance.

free conditions (see Figure 2.9).

Both the R3-derived construct "R3" and the related longer sequence, "R3R4", which forms the structurally resolved core of Alzheimer's brains derived fibrils, showed many similarities in terms of aggregation timescale (order of days), formation of a dimer in early stages of aggregation and primary aggregation dependence on an oxidising environment. However, R3R4 required vigorous shaking for aggregation to proceed to fibrils, for which reason it was discarded as a model.

R3 exists in a pre-equilibrium of monomer and dimer species in water, as observed by diffusion NMR, and which is quickly shifted towards monomer by the addition of the reducing agent DTT, indicating the dimer is disulfide bonded. The dimer is on-pathway to nucleation, since primary aggregation (but not seeded aggregation,

as a proxy for elongation) require an oxidising environment. The monomer and dimer species of R3 have distinct chemical shift signatures in the two histidine residues, which allows us to monitor their concentration separately and with good reproducibility. Finally, the R3 construct was capable of seeding full length tau aggregation in a well-established assay, similar to literature results.

Overall, we have established an R3-derived construct as a suitable model system for NMR-based quantitative chemical kinetics modelling of tau aggregation, which aggregates reproducibly in the absence of cofactors or agitation, and which is biologically relevant.

3

Development of a modelling strategy

Contents

3.1 Abstract 68
3.2 Introduction 70
3.2.1 Model space challenge 70
3.2.2 Model parameter optimisation 72
3.3 Methods 75
3.3.1 Contributions 75
3.3.2 Representative model for tests 75
3.3.3 Settings and timing of numerical integration simulations 76
3.4 Results 77
3.4.1 Description of the algorithm 77
3.4.2 Evaluation of the performance of several ODE solvers . 81
3.4.3 Model fitting and comparison of model's goodness of fit 86
3.4.4 Global optimisation settings 92
3.5 Conclusions 94

3.1 Abstract

The modelling of protein aggregation in a chemical kinetics approach presents a two-layer optimisation challenge. First, the model space, consisting of all possible chemical interactions and conversions, is very large, given that aggregates reach lengths of µm, corresponding to 10^3-10^4 subunits. A common approach is to sample subsets of potential chemical reactions into proposed models. Second, within a

specific aggregation model, there is a multidimensional rugged energy space with local minima, which requires a suitable optimisation strategy. In this chapter, I present method developments in both layers, building on previous work from our lab applied to a different aggregating protein. We define an algorithm for sampling of the model space, based on a systematic gradual increase of model complexity, with inheritance of fitted parameters from simpler models in the more complex models, followed by a reverse pruning procedure. The goal of this strategy is to find the minimally complex model, in number of parameters, which can explain the kinetic features observed. Further, we use an intermediate complexity model and the data obtained for tau R3 to investigate three numerical integration methods in terms of accuracy vs computing time, since numerical integration constitutes the backbone of the optimisation method used. The results confirmed the suitability of using the simplest of the methods, explicit/forward Euler, for the optimisation of each aggregation model, which is accurate enough when dt is adequate for the rates of evolution of the system. We decided to incorporate a result verification check, after optimisation of parameters, comparing the χ^2 with the value obtained using the more accurate linear multistep predictor-corrector method. This updated modelling strategy is applied to the biological system of tau R3 and the inhibitor D-TLKIVW in the following chapter.

3.2 Introduction

The modelling of protein aggregation kinetics with the chemical kinetics framework presents several challenges. Aggregates reach large sizes in terms of number of individual protein units, which results in a large number of potential chemical reactions. Also, the resulting systems of differential equations (ODEs) are non-linear, due to bimolecular interactions required for the growth of aggregates. What follows is a detailed discussion of the challenges in terms of the amount of potential reactions (model space challenge) and the difficulty of optimising non-linear systems (optimisation challenge), in order to contextualise the method developments in both areas that are shown in the Results section.

3.2.1 Model space challenge

The experimentally observed length of amyloid fibrils is in the µm range [23]. Considering the spacing of 4.7 Åcharacteristic of the amyloid fold, the range of maximum protein subunits per fibrils would be in a range of $2 \cdot 10^3$ (1 µm) to $2 \cdot 10^4$ (10 µm). In the most exhaustive chemical description, all species would be expected to interact with each other, for example by binding into larger species or breaking apart. Only considering bimolecular binding reactions, the number of reaction rates (forward and backward) to fit corresponds to n^2 (where n is the number of subunits in the largest fibril); in addition, we could have n rates for unimolecular events such as structural conversion (see representation in Figure 3.1). In the case of structural conversion, this opens up another n^2 space of bimolecular binding events, since it is reasonable to assume that binding rates depend on conformation. Overall, this complexity would be very hard to tackle computationally. We would also risk running into an overfitting situation, having more parameters to fit than datapoints. For this reason, most aggregation models effectively group subsets of rates into single forward and backward rates as well as set other rates to 0. For example, in the case of linear polymerization, all P_i species elongate by P_1 (monomer) (see panel B i)) with equal forward and backward rates, and all other rates of interactions of P_i with $P_i (i \neq 1)$ are set to 0. In a slightly more complex nucleated model, the

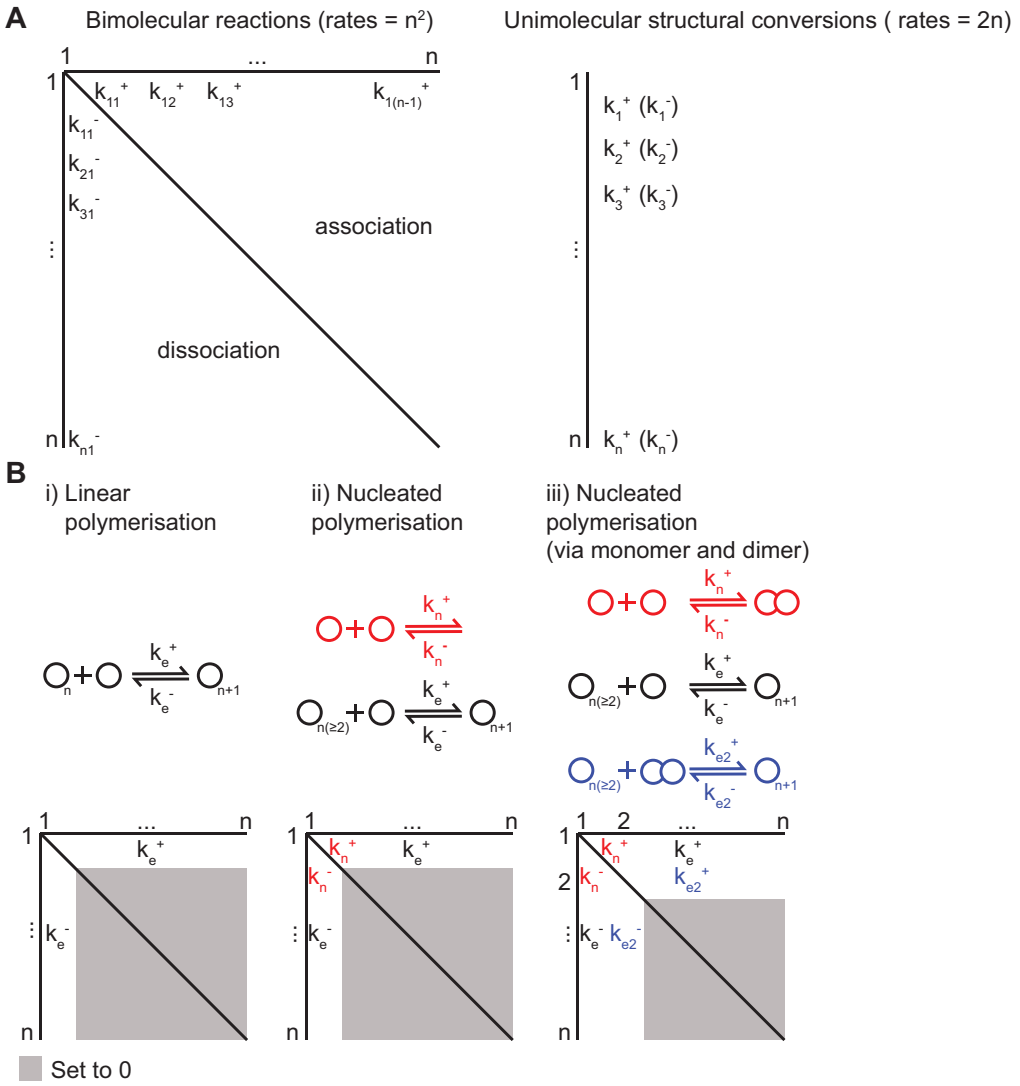


Figure 3.1: Mathematical complexity in protein aggregation systems. A. Representation of the possible unique reaction rates space in an aggregating system with maximum length n subunits of protein, considering bimolecular binding events (rates scaling with n^2) and some unimolecular events (rates scaling with n). Note that in the case of unimolecular structural changes this could open up another n^2 space of interactions. B. Illustration of three models that simplify the complete model space by grouping sets of elementary reactions to have the same rates, and by setting the rate of sets of events to 0.

equilibrium of early species would have distinct rates (see panel B ii)). The latter two approaches were considered by Oosawa in his seminal work on actin polymerisation, which was fitted to a nucleated model with a trimeric nucleus [249]. As another example, adding the chemical step of P_i elongating by addition of P_2 (dimer) would incorporate some of the reactions whose rates were set to 0 in the previous two models, with distinct (also grouped) forward and backward rates (see B iii)).

In order to address this layer of optimisation, i.e. how to choose the reactions to include, and any grouping of rates, in the Results section we define an algorithm for model space search based on a systematic gradual increase of complexity, with inheritance of chemical steps and fitted rates.

3.2.2 Model parameter optimisation

Aggregation models pose a multidimensional optimization challenge, especially those including several primary and secondary events. In a rugged energy landscape, there are multiple local minima which can prevent the optimization routine from reaching the global minimum. Previous work of our lab focused on developing a model optimiser for the application to the aggregation mechanism of α lactalbumin and the effects of a chaperone protein [37]. Tkachenko *et al.* used a global algorithm inspired in the basin-hopping algorithm, using the Levenberg-Marquadt algorithm (LMA) as the local optimisation method [37]. The LMA is well suited for non-linear least square fitting problems, such is the case for chemical mechanisms involving many bimolecular reactions. It is an iterative method based on computing the gradient of the function to minimise with respect to its parameters, around a trust region. However, if there are local minima, it is very likely that LMA will not converge on the true global minimum unless the starting parameter guesses are close enough to the solution. It would be inefficient to use a brute-force grid search method for such systems, given the multidimensionality (number of rates) and the range of rate values to explore. Instead, Tkachenko *et al.* used a global search method inspired by the framework of basin-hopping [250], which has been applied to a range of biological optimisation problems [251]. Further, in the results section, we describe two strategies for parameter value initialisation.

Basin-hopping couples local searches (using any suitable method) with the introduction of perturbations to the parameters which bring the overall optimisation to higher energy levels, some of which will lead to neighbouring lower local minima and eventually the global minimum. This is represented in two dimensions as a staircase of energy wells (see Figure 3.2).

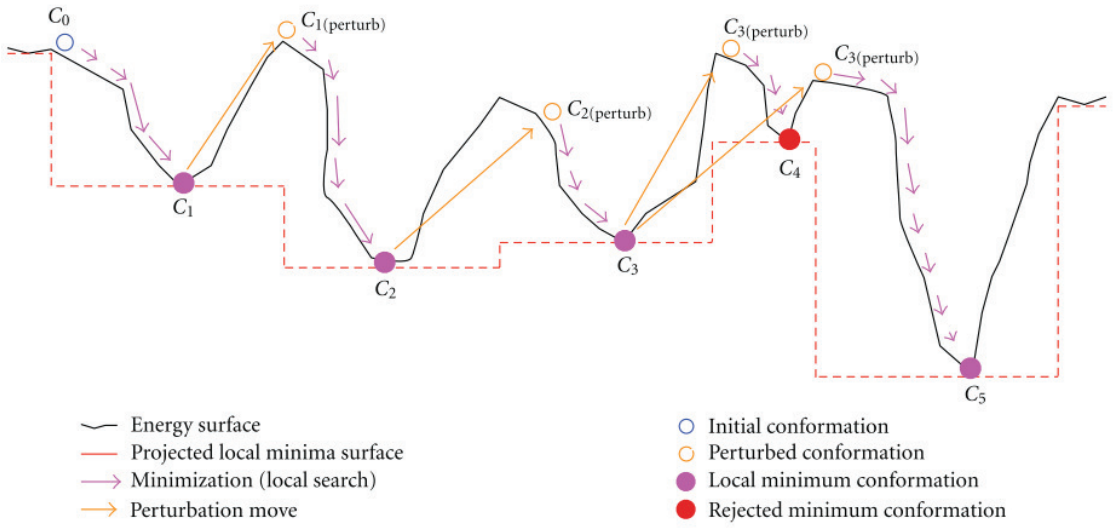


Figure 3.2: Staircase representation of the basin-hopping framework. Taken from [251].

In basin hopping, an important parameter is the amount of perturbation applied to the parameters. If the change is too small, the new state of the system will be within the same energy well and the local search will return the previous local minimum. In the other extreme, if the parameters change too much, the new position might be very far away (in energy space) from the current position. In the latter case, this framework degenerates into a multistart method, losing the connection with the previous state[251]. In the result section, we describe our approach with regards to setting the amount of perturbation.

At the lowest layer of the optimisation, in any local search strategy (such as LMA, used here), there is a need to choose a method for numerical integration of the systems of ODEs. In the next section, we describe three methods of varying accuracy whose performance is discussed in the Results section.

Numerical integration methods

Let's consider a chemical reaction of the form:

$$\mathrm{P}_1 + \mathrm{P}_1 \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{P}_2 \tag{3.1}$$

The rates of change of each species are given by:

$$\frac{\mathrm{d[P_1]}}{\mathrm{dt}} = -2\mathrm{k_+[P_1]^2}+2\mathrm{k_-[P_2]} \tag{3.2}$$

$$\frac{\mathrm{d[P_2]}}{\mathrm{dt}} = \mathrm{k_+[P_1]^2} \quad -\mathrm{k_-[P_2]} \tag{3.3}$$

In a more general sense, the evolution of a species, for example P_1, is defined as a function of time $y(t)$. Given the state of the system y_n at time n, we need to calculate the state of the system at time y_{n+1}, after a time interval (dt above). This is done differently depending on the method. In the simplest method, developed by Euler in the 18^{th} century and referred to as "forward/explicit Euler", this approximation is made:

$$y(t+h) \approx y(t) + h(f(t, y(t)) \tag{3.4}$$

where $f(t, y(t))$ is the gradient at time t and h is the time interval (dt above). Therefore the value at a time *n+1* would be calculated as:

$$y_{n+1} = y_n + hf(t_n, y_n) \tag{3.5}$$

Because the value y_{n+1} is defined in terms of known values, it is a type of *explicit* method. The global error accumulated is on the order of h, which is why the method is called *first-order*. In order to improve the accuracy of integration, two broad types of methods exist: methods in which more points in the interval h are used, often referred to as Range-Kutta methods, and multistep methods in which the value at y_{n+1} depends not only on y_n but more values in the past (e.g. y_{n-1}). Further to that, methods have been developed that use a variable step size (variable h) which is adjusted based on a computation of a local error.

We have compared forward Euler to a second order Range-Kutta method and a multistep method with variable step size (with an order ranging between 1 and 12):

1. **Explicit midpoint method**. The value of the function at time t_{n+1} is calculated as:

$$y_{n+1} = y_n + hf(t_n + \frac{h}{2}, y_n + \frac{h}{2}f(t_n, y_n)) \tag{3.6}$$

This requires two computations: 1) a calculation of the state of the system at the midpoint ($y_{n+1/2}$, at time $t_{n+h/2}$), and 2) the calculation of the state at time t_{n+h} using the gradient at the intermediate point to evolve y_n to y_{n+1}. The error is on the order of h^2.

2. **Variable step multistep methods.** This method was implemented from the GSL2.7 library using the step type *msadams*. It is a multistep method, which considers more positions other than the immediately past one. This particular implementation uses an explicit predictor method (Adams-Bashforth) and an implicit corrector method (Adams-Moulton). It is also a variable step size method in which the size of the step is automatically modified based on a specified relative error and absolute error, from an initial *dt*. The step size is adjusted to be accurate enough and also increased if the error is much below the tolerance, providing an optimal computational cost.

3.3 Methods

3.3.1 Contributions

The back-end C++ optimisation code used in this thesis was written by Andy Baldwin.

3.3.2 Representative model for tests

A model of intermediate complexity, with six rates, was chosen as a representative case to explore the trade-off between accuracy and computation time of three numerical integration methods. The model included off-pathway dimerisation (forming $P_{2,off}$), on-pathway dimer nucleation (forming $P_{2,on}$) and elongation by addition of monomer (from $P_{2,on}$). The chemical equilibria of this model

were as follows:

$$P_1 + P_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} P_{2,\text{off}} \tag{3.7}$$

$$P_1 + P_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} P_{2,\text{on}} \tag{3.8}$$

$$P_{2,\text{on}} + P_1 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} P_3 \tag{3.9}$$

$$P_{i\geq 3} + P_1 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} P_{i+1} \tag{3.10}$$

The parameters were set at the values shown in Table 3.1. The maximum length of fibrils was set to 240. The simulations were done using three datasets, with protein concentrations of 400 µM, 100 µM and 50 µM, respectively.

3.3.3 Settings and timing of numerical integration simulations

In our implementation, *dt* was set as the result of dividing the total time by the setting *rune*, independently for each protein concentration. It follows that, in the case of the forward Euler method, the number of function calls was directly proportional to the parameter *rune*, and in particular three-fold since we used three datasets (50 µM, 100 µM and 400 µM).

Table 3.2 shows the *dt* values for the *rune* settings tested. For the variable step multistep method, initial *dt* was set to 10^{-3} h (0.27 s) and the relative and absolute error tolerances were both set to 10^{-6}.

Since the total time of the experiment was different depending on the protein concentration ($\sim$ 89h for 50 µM, $\sim$ 69h for 100 µM and $\sim$ 19h for 400 µM), so was *dt*. This is intended, since the rates of evolution of species will be faster for more concentrated samples.

In the forward Euler method, the sampling of *dt* was in the range of hours (for rune = 75, *dt* = 0.25, 0.92 and 1.28 h for 400 µM, 100 µM and 50 µM, respectively) and milliseconds (for *rune* $7.5 \cdot 10^6$, *dt* = 0.009, 0.03 and 0.05 s).

Each simulation was run three times and timed using Bash's "time" command upon calling the C++ executable. The time reported in the results corresponds to the average ± standard deviation in "user time" (the cpu time spent on running

Table 3.1: Rate values used for simulations. The rates correspond to the chemical equilibria above, plus two non-rate parameters: *srac* (signal to concentration ratio) and *frac* (initial monomer fraction /(monomer+dimer)).

k_1	$9.7 \cdot 10^{-5}$
k_{-1}	$4.5 \cdot 10^{-7}$
k_2	$1.0 \cdot 10^{-4}$
k_{-2}	$1.0 \cdot 10^{-6}$
k_3	$1.0 \cdot 10^{-4}$
k_{-3}	$1.0 \cdot 10^{-6}$
srac	0.5
frac	0.9998

Table 3.2: *rune* settings tested and corresponding *dt* in the forward Euler method.

	dt for forward Euler (s)		
rune	400 µM	100 µM	50 µM
75	912	3312	4608
200	342	1242	1728
400	171	621	864
750	91.2	331.2	460.8
$7.5 \cdot 10^3$	9.12	33.12	46.08
$7.5 \cdot 10^4$	0.912	3.312	4.608
$7.5 \cdot 10^5$	0.0912	0.3312	0.4608
$7.5 \cdot 10^6$	0.00912	0.03312	0.04608

the code). The reported χ^2 are cumulative from all the concentrations used in the simulation. The χ^2 error was calculated as:

$$\chi^2_{err}(\%) = \frac{|\chi^2 - \chi^2_{true}|}{\chi^2_{true}} 100 \tag{3.11}$$

3.4 Results

3.4.1 Description of the algorithm

Figure 3.3 shows a general algorithm designed to approach the top layer (i.e. which chemical reactions are included in the model) of this optimisation problem, which can be applied to modelling the mechanism of aggregation of a protein and also the effects of an inhibitor, in distinct stages as will be explained below.

The inputs of the process are the dataset of kinetic traces and the "pool" of available chemical steps. The pool is designed manually based on the reactions

proposed in the literature, as well as experimental insights relevant to the specific case (i.e. which particular species have been observed to appear). The general strategy is to build models of increasing complexity by combining the existing best fitting model ("parent model") with other chemical reactions available in the pool. In the case of a protein only mechanism, the initial model is fixed to be linear polymerisation, since an essential aspect of any good aggregation model is that it needs to be able to generate long aggregates, as observed experimentally.

The iterative process of building complexity goes in generations, each consisting of a set of "child models" that all have the parent model steps in common. The same generation "sibling" models get optimised in parallel, and the model with the lowest χ^2, the winning model, is chosen as the new parent. The steps that did not get selected in the winning model go back into the pool and are available for the following round of model building. It's important to note that in the simplest single-pathway implementation, which is the one used by default to avoid too much complexity, there can exist mutually incompatible steps, for example a nucleation reaction that converts two monomers into a into a dimeric nucleus (that is elongation competent) and a nucleation reaction that generates a trimeric nucleus from a monomer and a dimer. This is taken into account by removing from the pool any steps incompatible with the new parent model.

After each generation is complete, and before starting a new one, there is a set of stop conditions to consider: 1) whether the winning model is significantly better than the previous parent model (to avoid overfitting), 2) whether the error in the fit is still higher than the inherent data variability, (if available) and 3) whether there are any remaining compatible steps in the pool. Steps 1 and 2 are designed to avoid overfitting. If the data was still fitting poorly after reaching condition 3, it would be necessary to either enlarge the initial pool or consider parallel pathways, as was done in the original study [37].

Finally, once the building complexity stage had been completed, there would be a pruning stage, essentially a reversal of the previous process, as a second-pass for avoiding overfitting. In a similar generation-by-generation process, child models

would be created in which one of the chemical steps has been removed. The models would then be optimised and compared with the parent with the same criteria as stop conditions 1 and 2. This would allow us to confidently assess that all the chemical steps included in the final model are needed for the goodness of fit, regardless of the order in which they were incorporated into the parent model. It is possible to envision that at an early stage, in a poorly fitting model, introducing certain reactions could lead to a substantial drop in χ^2 and the acceptance of the steps, but that once the entire system of chemical reactions had reached a global minimum, some of the steps might be partially or fully redundant. This entire process is to be applied twice in the case of the modelling of inhibition. Initially, the data with only protein would be modelled, to achieve a well-fitting mechanism. Afterwards, the chemical reactions and rates (with values fixed at the values fitted) of the winning model would be the constant backbone of all the inhibition models. The design of the inhibitor pool would be informed by the mechanism of the protein, in terms of which species seem to have a specific relevance (i.e. the nucleus). In this way, we can achieve explicit mechanistic modelling of both protein and protein-inhibitor interactions, one of the strengths of using a numerical integration framework.

In a way, this algorithm resembles, at the top layer level, the gradient descent - hopping algorithm that is implemented to optimise individual models. The gradient descent part would correspond to testing several child models in parallel and progressing in the direction of greatest improvement. The hopping/perturbation aspect would correspond to the initialisation of some of the optimisation runs with naive initial guesses, since that moves them away from a local minimum, however not too far away considering the model space, since most of the chemical steps are the same as those of the related parent which after optimisation had achieved that local minimum.

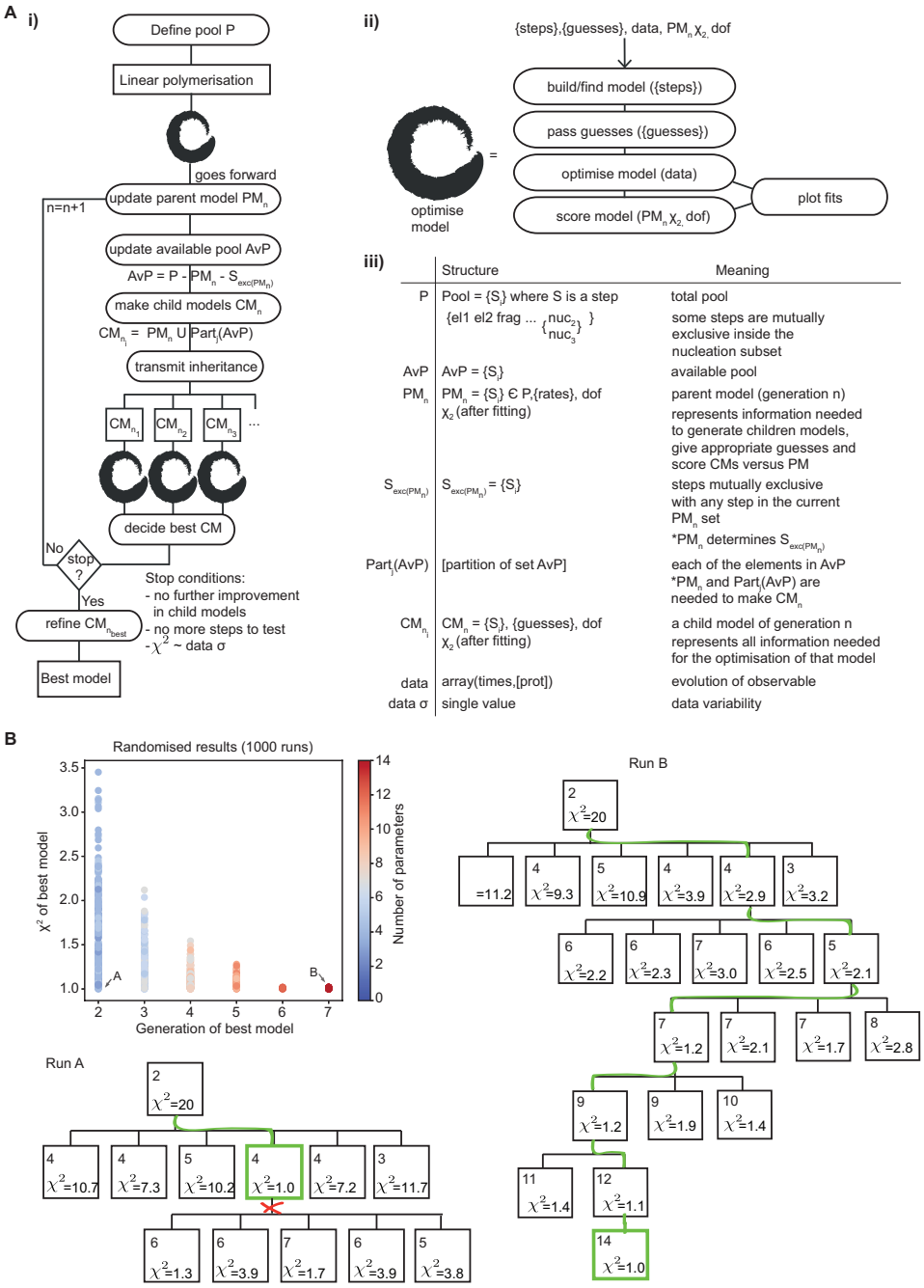


Figure 3.3: A. Schematic of the algorithm flow. (i) Overview of the generation-based model building strategy. (ii) Optimisation steps. (iii) Detailed description of the structures in (i). B. Illustration of the types of pathways taken by the algorithm in A, by repeating the process 1000 times using a pool of seven chemical steps, and a pseudorandom number generator (sampling a uniform distribution from 1 to $50/n^2$, where n is the current generation) at the model optimisation stage. The graph shows the correlation between best model χ^2 and the generation the modelling routine stopped in a series of 1000 runs. The trees depict two specific runs that finished with $\chi^2 = 1.0$ (run A, left), at generation number 2, and another that also finished with $\chi^2 = 1.0$, but at generation number 7 (run B, right).

Example modelling pathways from simulations

Figure 3.3 shows the results of running the modelling algorithm, in build-up stage only, with a pseudorandom number generator as the χ^2 of optimisation. Two examples of pathways are shown to illustrate the process. In the left shorter path, the stop condition of the goodness of fit not improving versus the parent model was met by a model of the 2nd generation, and therefore the algorithm stopped at that point. In contrast, the path on the right exhausted the pool of steps, since there were seven (arbitrary) steps in the pool and it stopped after generation 7 (with only one model left to test) gave a lower χ^2 than the parent. The specific implementation of generating a random χ^2 was designed to reflect a real scenario by sampling χ^2 from a uniform distribution between 1.0 and $50/n^2$, n being the current generation. The dependency on n meant that the higher the generation, the lower the maximum bound of the distribution sampled, which was a way to mimic the expectation that more complex models are expected to fit the data better.

In Figure 3.3 A ii), we can see the lower level of optimisation challenge described in the introduction, the actual optimisation of parameters of a given model. In the next section, we describe the results of comparing three numerical integration methods to be used in the LMA local search.

3.4.2 Evaluation of the performance of several ODE solvers

Simulation of kinetic traces

Assessment of integration accuracy To test the performance of the different numerical integration methods, a set of rates and a model of moderate complexity were selected. To get a reference value for 'accurate', a series of long simulations with short dt were run using the various methods. Confidence was achieved when all methods provided essentially identical descriptions of the kinetics, in this limit (see Figure 3.4). For the purposes of modelling, we sought an optimal method that provided a result with sufficient accuracy in a minimum amount of computational time. To this end, the deviation of the simulation compared to the known correct result was measured and we defined a maximum error threshold of 0.5%. Figure

3.4 shows how all the methods converged to the same value of χ^2 (472.246), which for the Forward Euler and the Explicit midpoint method was achieved at 400 function calls. Using fewer function calls (i.e. too large *dt*) manifested in non-physical discontinuities in the simulated traces, as shown in Figure 3.4 B i and ii. Also, as expected, the Explicit Midpoint method had a smaller error for the same setting of *rune* (with twice as many function calls), which is seen more clearly in the following figure.

Assessment of computational time cost In order to obtain a reasonable assessment of the durations of the different algorithms, the total computation time was measured (see Figure 3.5).

In the case of high values of the *rune* setting ($>10^5$), this is directly measuring the linear dependence expected on the number of core evolution function calls. When this value was smaller, there were deviations from linearity, indicating that ancillary functions of the program are contributing significantly to the measured time. Nevertheless, this simple approach reveals the trade-off between computational time and accuracy between the three methods.

We were interested in comparing the time taken by the variable step multistep method, given these types of methods are designed to find the optimal balance between sampling accurately and sparing resources, and found that it did an excellent job in this case. The variable step multistep method took a similar time to the other two methods with *rune* settings in the 750-7500 range and achieved the same true χ^2 as the two other methods at their largest values of *rune*. Based on this, we have shown the result of the variable step multistep method interpolated at *rune* of 1500, as a high-end conservative estimation (see Figure 3.5 B, inset). At that range of *rune*, using the Forward Euler or the Midpoint method the χ^2 had not completely converged to the true value, indicating that the variable step multistep method performed better (see Figure 3.5 A, inset).

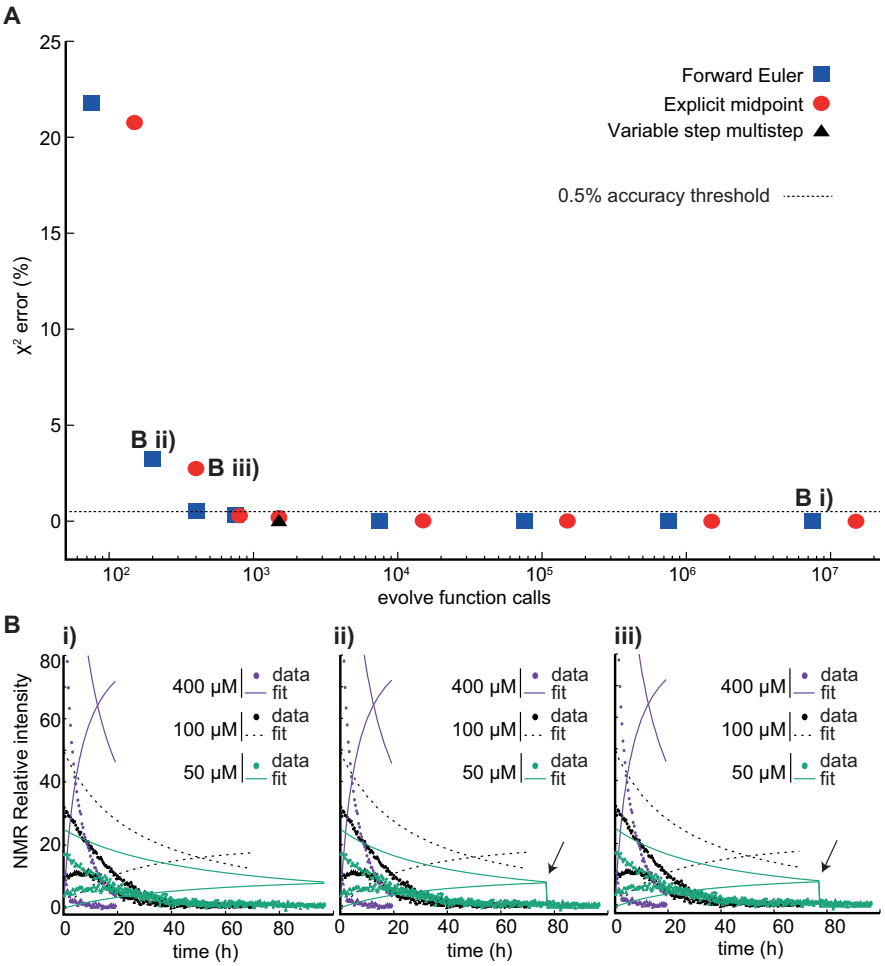


Figure 3.4: Comparison of the accuracy of the forward Euler, explicit midpoint and variable step multistep methods. A. Accuracy of integration depending on *evolve* function calls (shown in log space). For forward Euler, this corresponds to how many times the calculating function is called (for each protein concentration dataset) and to the setting *rune* ($\frac{1}{dt}$), which varies for each concentration. For the explicit midpoint method, the number of function calls per dataset is twice the *rune* value. *rune* is not a relevant parameter for the variable step multistep method; the value is shown interpolated at *rune*=1500, as a high-bound estimate based on measurements of calculation time (shown in the next figure). The χ^2 error was calculated relative to the true χ^2, obtained with the variable step multistep method and the other two methods using a very large *rune* setting, and the threshold of 0.5 % accuracy is shown as a dashed horizontal line. B. Non-physical mass discontinuities arise when using a dt that is not small enough for the rates of evolution of the system of ODEs. (i) Reference simulated traces obtained using the Forward Euler method with $dt < 1ms$. (ii) Simulated traces using the Forward Euler and $dt = 0.5 - 2s$. (iii) Simulated traces using the Explicit Midpoint method with $dt = 0.5 - 2s$. Mass discontinuities in the 50 μM simulated trace are indicated with arrows.

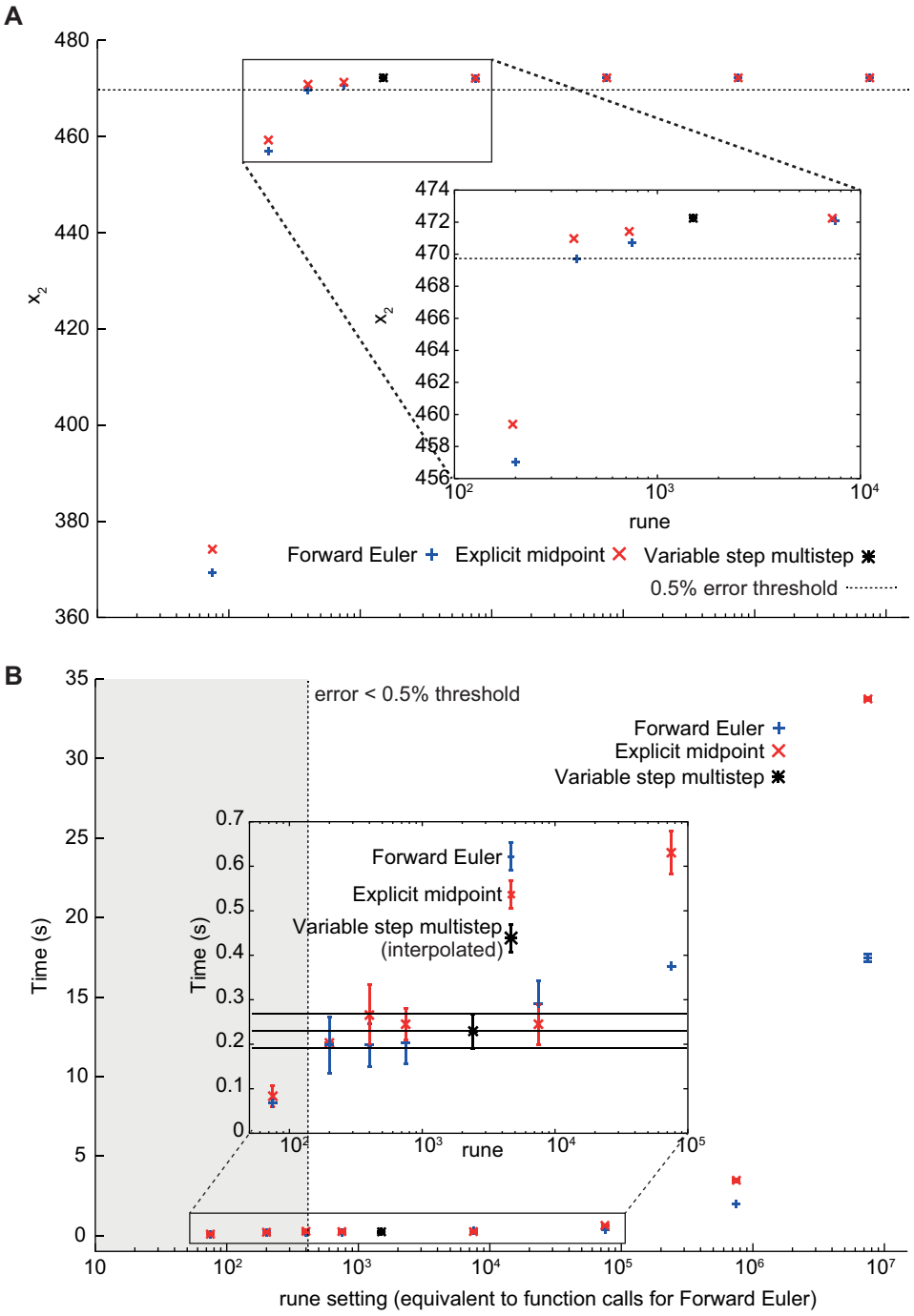


Figure 3.5: Accuracy vs calculation time comparison of the forward Euler, explicit midpoint and variable step multistep methods. A. Accuracy of integration depending on the setting *rune* ($\frac{1}{dt}$) for forward Euler and the explicit midpoint method. Since the setting of *rune* doesn't apply to the variable step multistep solver, that datapoint was interpolated on the x axis based on the calculation time taken, shown in B. A horizontal dashed line represents the threshold of accuracy in which the error is less than 0.5 % compared to the "true" χ^2 at which all methods converge. B. Calculation time vs setting of the parameter *rune* (plotted in log space). The same *rune* setting results in comparatively twice as many function calls for the explicit midpoint method compared to forward Euler. The non-shaded area corresponds to results with < 0.5 % error from the true χ^2.

Performance of numerical methods in LMA optimisation

Following the results that for this test model, the variable step multistep method had performed more efficiently (and not just more accurately) than the other two methods, we tested its use in LMA optimisation. However when we used this method empirically as part of optimisation calculations, frequently it was found not to progress to the local minima in a number of cases; the precise reasons for this were unclear. The algorithm used by the variable step multistep approach to determine the 'optimum' time effectively assumes that all rates being optimised contribute similarly to the overall evolution. In our case, some rates such as the nucleation rates, have extremely strong effects on the overall curves generated, and so we suspect that this is not working well with the heuristics in the variable rate algorithm. To address this, we used the Forward Euler with a manually judged value of dt ($rune = 5 \cdot 10^4$) for the optimisation runs, well within the 0.5 % acceptable error threshold (in this instance), and with a relatively small calculation time. The calculation time for each simulation in LMA would then be in the order of seconds (see Figure 3.5 B, upper *rune* range shown in the inset); in the real implementation, the maximum length of oligomers was set to 800 instead of 240, around three-fold higher.

At the end of each optimisation run, we compared the obtained χ^2_{Euler} with the true χ^2 obtained from the variable step multistep method. If the curves were essentially identical (the difference in χ^2 was below a 0.5% threshold), we concluded that the Euler method and the choice of time step were well set-up, and the result was approved. In the event of a deviation in this test, where the difference was larger than 0.5 %, the result was flagged for manual interrogation. Over the course of modelling three datasets (shown in Chapter 4), we observed around 5% of the runs failing the test. However, for every model several optimisation runs were run (details below in section 3.4.4) and for all models there were always multiple results that passed all the checks, which led us to conclude that in our case, the settings chosen for the Forward Euler method are an adequate trade-off between accuracy and computational resources.

In this section, we considered the tradeoff of accuracy versus computational time of the numerical integration backbone of the model parameters optimisation. We decided that the optimal balance would be to maintain the use of the fastest Forward Euler method and introduce a single final "expensive" χ^2 accuracy check to validate the optimisation result.

3.4.3 Model fitting and comparison of model's goodness of fit

Model fitting

Minimisation We have performed a non linear least squares minimisation procedure on each model proposed, in which the local search algorithm (LMA) minimises the average error per point ($\overline{\chi^2}_{RSS}$), defined as the Residual Sum of Squares divided by datapoints:

$$\overline{\chi^2}_{RSS} = \frac{\sum_{i=1}^{n}(y_{i,exp} - y_{i,sim})^2}{n_p} = \frac{RSS}{n_p} \tag{3.12}$$

where $y_{i,exp}$ represents each datapoint, $y_{i,sim}$ represents each simulated value and n_p is the number of datapoints.

Note that the true χ^2 would be:

$$\chi^2 = \frac{\sum_{i=1}^{n}(y_{i,exp} - y_{i,sim})^2}{\sigma_i^2} \tag{3.13}$$

were σ_i represents the error of point i. The errors are not always known precisely, and such is our case, which is discussed below in section 3.4.3.

Regardless, within each model optimisation, the result of the minimisation does not depend on which metric (RSS, the true χ^2 or $\overline{\chi^2}_{RSS}$ as defined above) is used, since the errors (known or not) and number of datapoints are constant within the dataset.

Comparison of relative goodness of fit between models In terms of comparing models, though, one would aim to compare the true χ^2 and also take into account their complexity in terms of degrees of freedom.

However, given that the errors would be constant within the dataset, not knowing them precisely does not affect comparing the relative goodness of models, and so we have used the $\overline{\chi^2}_{RSS}$.

In the case of child models within a generation, their complexity is equal (within 1 parameter difference, in case some contain steps that consist of only a forward reaction, i.e. fragmentation); therefore we chose as winner the model with the lowest $\overline{\chi^2}_{RSS}$.

The next section explains how we approached comparing models of distinct complexity, in terms of evaluating the winner child of a generation with respect to its parent and avoiding overfitting.

When to stop increasing complexity

As part of the algorithm depicted in Figure 3.3, there is a "STOP?" decision point at the end of each generation, in which the winning child model of the generation needs to be evaluated. This assessment is done to avoid overfitting, a problem we are critically aware of. Certainly, we would expect that as the number of fitting parameters in a model becomes large, the quality of the fit would improve and the χ^2 would decrease. Overfitting would mean that we are effectively fitting noise, thereby increasing the complexity of the model without any gain in mechanistic insight.

We considered several ways to address this problem. The first is a statistical F-test. This allows us to estimate if the improvement in a fit is statistically justified versus the increased number of degrees of freedom. In our case, it is not obvious what to use as a *true* degree of freedom. Naively we would have thought we could take the number of data points as the degree of freedom. When we did this, however, because we have so many time points measured in our experiments, (in the range of 10^3, compared 4-30 parameters), extremely small improvements in χ^2 became justified by this method. Therefore we concluded the F-test did not

address our concerns about overfitting. Instead we relied on three criteria: 1) visual improvement of the fit, 2) loss of goodness of fit from removal of each component, and 3) comparison with experimental variability.

1. Visual improvement of the fit A model of increased complexity (a "child") was accepted as the new parent model if the improvement in $\overline{\chi^2}_{RSS}$ was visible when inspecting the traces and fit curves. Over the course of the project, we arrived at a working value of detectable change in $\overline{\chi^2}_{RSS}$ to be around 0.1, and this cut-off acted as a reasonable 'barometer' of significant improvement.

2. Loss of goodness of fit from removal of each component At the end of the complexity-building stage, when the additional complexity was not justified (according to criteria 1), the final model was tested in terms of all the components being needed for the fit. For this, a systematic removal of each step (referred to as a *pruning* stage) was carried out from the final model. Simpler models (each lacking one component of the final model) were optimised (with the same global optimisation settings - see below) and compared to the final model. The final model was approved if it was better than each of the pruned models according to the cut-off in 1), meaning there was a significant loss of goodness of fit from removing any of its components.

Otherwise, if the final model failed the test against a simpler model, that simpler model was taken as the updated final model and the process of pruning was performed again until a model passed all the tests. The final models presented in Chapter 4 passed this two-step test, and the process is discussed in detail there.

3. Comparison with experimental variability Any experimental measurement carries an error, and this needs to be accounted for when modelling data; a good understanding of the experimental error allows fitting the data to a precision equivalent to the uncertainty in the measurements, avoiding both underfitting (not explaining all the features of the data) and overfitting (fitting noise). In the case of our experiment, there are three main sources of error:

1. Volume and weight handling errors that affect the concentration of protein and small molecules. Since the reactions are sensitive to concentration, small handling errors might lead to large differences in the kinetic traces. This introduces random (the making of each high concentration stock and each sample) and systematic errors (e.g. making samples by serial dilutions from two slightly different concentration stocks).

2. Dead-time. There is dead time between mixing reagents in aggregating conditions and the first recorded datapoint. This could be random or systematic.

3. Intensity recorded by the NMR spectrometer. This would be random and include the intrinsic error of a single measurement as well as any spectrometer drifts during the course of the experiment (which last days).

The intensity fluctuation of a single measurement in identical conditions is very low (0.1 %), therefore this error can be neglected in magnitude in comparison with the others. There were steps taken to minimise errors in handling and dead-time: when possible, the samples were made from serial dilutions from a single high concentration stock (to reduce systematic errors), and the dead-time was kept to a minimum of around 15 minutes between the mixing and the first datapoint being acquired. Within the modelling aspect, to take into account some of the variability, we established a process called "polishing", in which we took the final fitted parameters of each optimisation (see below for full details) and inputted them as initial guesses for more optimisation runs in which we additionally allowed the parameters "*srac*" (NMR signal / concentration) and "*frac*" (initial monomeric fraction) to vary specifically per sample. srac in particular has the effect of treating the sample as a slightly different sample concentration, therefore we only carried out these additional adjustment optimisation runs (after which the $\overline{\chi^2}_{RSS}$ was taken as the goodness of fit for that model, for comparisons described above) in late generations, i.e. to models that had achieved good fits with common frac

and srac parameters across the multiple concentrations in the dataset, therefore preserving the global fit aspect of the procedure.

The best way to obtain good estimates of the error per point would be to acquire all the data in replicates, where three is a good standard in the literature. However, this was not possible in this study due to the expense of both sample and spectrometer time. Instead, we acquired several samples in duplicate in the protein-only case, which we used to obtain an estimate of the average error per point as described below.

We recall the expression of the true χ^2 (3.4.3), which can be expressed in terms of a single $\overline{\sigma}^2$ in the limit where the errors are the same at each point:

$$\chi^2 = \frac{RSS}{\overline{\sigma}^2} \tag{3.14}$$

It is a commonly used criteria for a model to be a good fit to data, and with errors appropriately estimated, that the reduced χ^2 (χ^2_{red}) is approximately equal to 1 [252]:

$$\chi^2_{red} = \frac{\chi^2}{dof} = 1 \tag{3.15}$$

Given equation 3.4.3, then $\frac{RSS}{dof}$ can be related to $\overline{\sigma}^2$ as:

$$\overline{\sigma}^2 = \frac{RSS}{dof} \tag{3.16}$$

where dof is the degrees of freedom. In our case, $\frac{RSS}{dof}$ can be approximated as $\overline{\chi^2}_{RSS}$ (equation 3.4.3) since we are in the limit where the number of points greatly exceeds the number of parameters, and therefore the degrees of freedom are approximately the number of datapoints. Therefore, we can compare the *expected* scatter from a fit, as measured by $\overline{\sigma}$, to the *measured* scatter that we get when we perform repeated measurements at a given set of concentrations.

This allows us to define a third final criterion by which we can stop fitting. When the final $\overline{\chi^2}_{RSS}$ value can be translated into an average scatter ($\overline{\sigma}$) that is comparable to what we see from experimental repeats, then we can reasonably

Table 3.3: Experimental scatter in duplicates of R3-only samples. The average scatter was calculated from two traces in each case, separately for monomer and dimer traces, and is reported as an average error per sample in units of NMR signal (arbitrary) and in percentage error with respect of the highest averaged monomer signal (t=0h) for each sample. From this analysis, we take 3 % as a reference value for experimental error.

sample	monomer $\overline{\sigma}$	dimer $\overline{\sigma}$	average $\overline{\sigma}$ (% of maximum signal)	Time between repeats
100 µM (H_2O)	0.55	0.58	0.57 (1.5)	Days
94 µM (D_2O)	2.62	1.86	2.24 (3.88)	Months
188 µM (D_2O)	5.82	1.94	3.88 (2.9)	Months

state that we are explaining our data to within the precision we expect from repeated measurements, and thus conclude it is unreasonable to continue to increase the complexity of our models.

The duplicates of the sample of 100 µM in 90 % H_2O buffer have an average $\overline{\sigma}$ of 0.6 NMR signal units in both the monomer and dimer traces, which considering the maximum measured signal (35 units) would give an average error per point of 1.5 %. With the same logic, the $\overline{\sigma}$ measured from two sets of duplicate measurements in the 90 % D_2O buffer dataset was in the order of 3 % (see Table 3.3).

The samples in D_2O buffer were made from different high concentration stocks, therefore we consider they reflect a more realistic measure of experimental variability than the replicates of 100 µM in H_2O buffer, which represent the best case scenario performing repeats with the same initial stocks and closely in time (which could affect a range of other relevant variables). Overall, from this analysis we take an average error of 3 % as the estimate for kinetic runs in our datasets.

Looking ahead at the results presented in Chapter 4, our final best models in each dataset have fitted $\overline{\chi^2}_{RSS}$ values that, when considered relative to the maximum signal intensity of the samples the errors were estimated from, correspond to expected scatter around 3 %: the first dataset of R3-only was fitted to $\overline{\chi^2}_{RSS} = 1.2$ (equivalent average error of 3.1 %), the dataset with inhibitor to $\overline{\chi^2}_{RSS} = 1.4$(equivalent to 4%), and the second dataset (in D_2O) to $\overline{\chi^2}_{RSS} = 3.1$ (equivalent to 2.4 %.)

Finally, given the estimation of the error as 3 %, we can relate the 0.1 cut-off defined empirically for a fit to be visually better than another to detecting around a 0.3 % of average fit improvement - typically, though, the improvement

was visible in a particular time range and sometimes only in some of the samples, so the average can be misleading.

Overall, we ascertain that the aggregation models that are our best fits are far more complex than any model tried previously in the literature. However, the application of three criteria to prevent overfitting (visual improvement of fits, worsening of fits upon removal of each component, and consideration of experimental variability) justifies considering our models minimally complex. In principle we could, following the F-test and the current estimate of experimental error, continue to generate models of even higher complexity. The specific gains however will be modest and so the subjective-seeming criteria of 'we need to see the improvement by eye to believe it' is perhaps the most powerful, together with a conservative approach with regards to how precise an estimate of experimental variability we have.

3.4.4 Global optimisation settings

Parameter initial guesses

Local search algorithms such as LMA are very sensitive to initial conditions, and in a space with multiple minima only find the global minimum if the starting conditions are close enough. For this reason, Tkachenko *et al.* incorporated a basin-hopping like framework to perturb the parameters found after each local search so that the perturbed state would lie outside the local minimum well, therefore paving a path towards the global minimum [37]. As described above, we have designed a global search algorithm one layer above, which can be thought of as basin-hopping but in terms of exploring the model space. Each child model is a perturbed model with respect to its parent, with a larger chemical space of reactions. We wanted to preserve local information not only in terms of the set of reactions but also in the parameter values, which led to the "inherited" guess mode of initialising child model's parameters at the fitted values for the same shared steps in the parent.

This could be beneficial in terms of starting the optimisation in a good area both regarding model space and parameter space. However, it is also possible that upon introduction of additional chemical steps, the now true global minimum might

lie far away in parameter space of (some) existing parameters of the simpler model. For this reason, we have designed the "naive" guess mode, in which all parameters are initialised at a low value (10^{-9}), therefore preserving the local environment in terms of model space but clearing the prior information regarding parameter space.

We have optimised each model using both guess initialisation modes and observed either guess mode performing better in a model specific way (see Results of Chapter 4). For this reason we performed optimisations in both initialisation settings in parallel. The only exception is for polishing runs: in those we start from an already fitted model, and allow signal-to-concentration, and initial monomer fraction parameters, to vary independently for each concentration, in order to allow small adjustments equivalent to experimental variability. For this case we only used inherited naive guesses, because the chemical steps composition of the model is not changing, and are by design polishing the result further within a minimum that we consider the global minimum for that model.

Perturbation of parameters in basin-hopping

As mentioned in the introduction, in a basin-hopping framework the amount of perturbation of the parameters after each local search is an important setting. If the perturbation is too small, it is likely the new state will be within the same local minimum, and the LMA will converge back to the same state. If perturbation is too large, the information about the local environment of good parameter space is lost. As in the original method by Tkachenko *et al.*, perturbation was applied for each parameter by replacing the initial parameter value (p_{start}) by a value (p_{final}) drafted from a lognormal distribution with θ (location parameter) = p_{start} and σ as customly defined, where varying σ effectively changes the amount of perturbation; the use of a lognormal distribution was meant to keep parameter values positive [37].

In the original method implementation a few values of σ were tested and one was chosen for the entire modelling process. However, in this case, throughout the project we have found that we couldn't define a single setting that worked for all cases. For this reason, we chose to optimise each model, for each initialisation

mode (naive and inherited) with four σ settings. σ values were 0.01, 0.05, 0.1 and 0.2 for optimisations without polishing (see above, and details in Chapter 4), and 0.01,0.05 and 0.1 for optimisations with polishing.

Overall, the optimisation of a model without polishing consisted of 8 optimisation runs, combining the two initialisation modes and the four settings of perturbation, while the optimisation in polishing mode had inherited guesses and was repeated at 3 settings of perturbation, giving a total of 24 optimisation runs for the more complex models. In the modelling of distinct datasets we have observed models in which all settings led to a very similar fit, and others in which distinct settings yielded the best results. We have not explored any trends further, yet it is clear from our experience that this strategy of using a range of optimisation settings was instrumental to achieve the good mechanistic model fits for this particular system of study, and which are presented in Chapter 4.

3.5 Conclusions

This chapter described developments in several aspects of a generalisable modelling strategy for chemical kinetics modelling of protein aggregation mechanisms, and which is applied to real data in the following chapter. The developments affected all three layers of the modelling challenge, the search within model space (see 3.6 A), the basin-hopping like global parameter search and the lowest layer, the numerical integration of systems of ODEs (see Figure 3.6 B). Additionally, we defined three criteria by which to evaluate models of higher complexity, with the aim to avoid overfitting.

The strategy for model search starts by creating a pool of chemical steps and optimising the simplest model, linear polymerisation (see Figure 3.1 B i)). From there, an iterative generation making procedure is carried out until a good enough model is found and the stop criteria is met. In each generation, additional steps from the pool are added to the existing best ("parent") model, one at a time, creating "sibling" models of similar complexity. After optimisation of each child model, three criteria are applied to evaluate whether the increased complexity is justified.

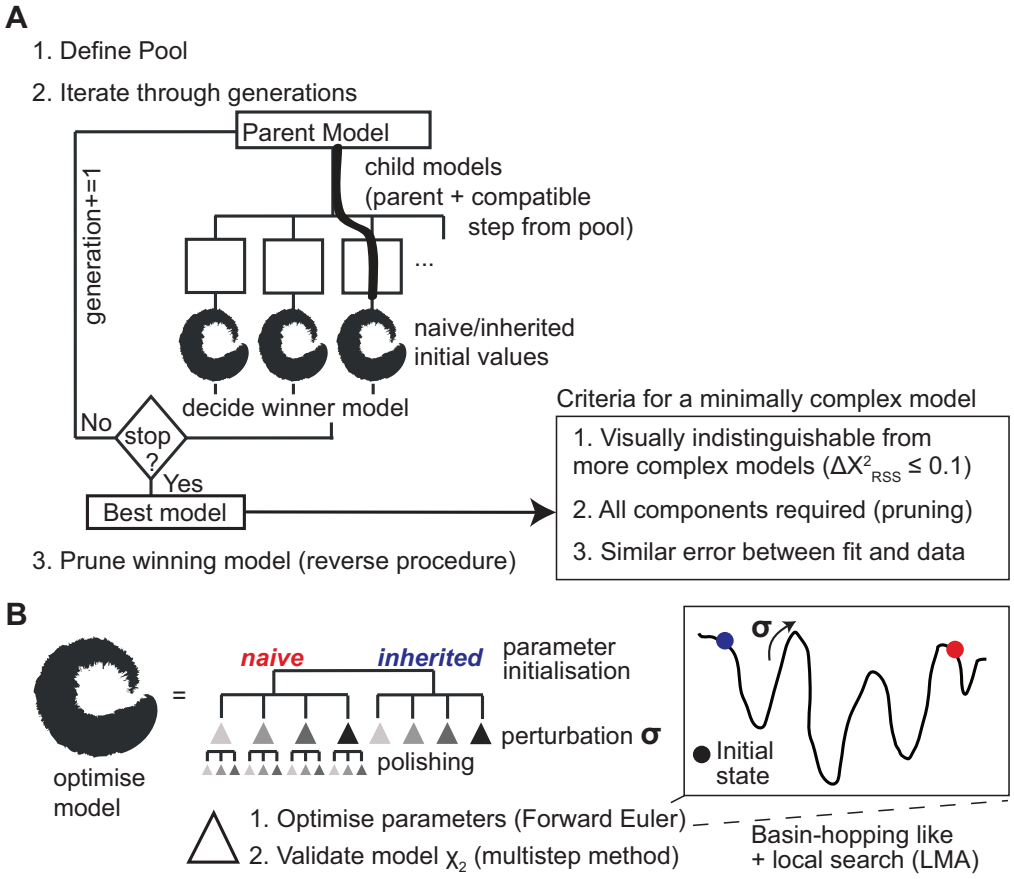


Figure 3.6: Summary of modelling method developments. A. Systematic model space search strategy, consisting of building models with increased complexity from a set of chemical reactions (pool). B. The optimisation of each model requires establishing initial guesses and is done using the faster forward Euler method, with a suitably small dt, followed by a validation step using a higher order more accurate multistep method.

Regarding the optimisation process, each model is optimised in several parallel runs, combining two strategies of parameter value initialisation with several degrees of perturbation between local search. Based on the tests of numerical integration methods, we decided to carry on with using forward Euler and introduce a result validation step in which we would compare the obtained χ^2 with the value obtained with the more accurate multistep method msadams from the GSL2.7 library. Forward Euler should be implemented with a sampling value *rune* (dt = experiment length / *rune*) in the right range, for accuracy purposes. We have chosen *rune* at $5 \cdot 10^4$, a middle point of those tested and which gave values well within the acceptable error (set to 0.5%) for a model of intermediate complexity with the

rates tested here (see Figure 3.4). We believe this approach represents an optimal balance of accuracy versus computational cost, as is shown to be the case in the full results of modelling shown in the following chapter (given the large majority of individual optimisation runs passed the χ^2_{true} comparison test). However, dt should be adjusted for the modelling of distinct datasets, in particular if aggregation proceeds significantly faster.

Overall, we have shown further developments to the existing method from our lab [37] by addressing aspects in all optimisation layers and developing stringent criteria to prevent overfitting. We conclude that it was far more challenging than expected to produce a reliable optimisation strategy. Nevertheless, when implemented in this way, our multilayered optimisation method allows us to 1) achieve a simulation at excellent numerical integration accuracy as measured by the comparison to the variable step multistep method, 2) find a global minimum in parameter space (using a combination of variable parameter initialisation modes and perturbation amounts in the basin-hopping method) and 3) find a minimally complex global minimum in model space (using the top-level model search strategy) from tens to hundreds of satisfactorily optimised models, for each dataset, as will be shown in the next chapter.

All models are wrong, but some are useful.

George E.P. Box

4

Mechanism of aggregation and inhibition of tau R3

Contents

4.1 Abstract 97
4.2 Introduction 98
4.2.1 Inhibition of tau by D-TLKIVW 99
4.2.2 Summary of qualitative findings on R3 aggregation from the previous chapter 99
4.3 Methods 101
4.3.1 Experimental 101
4.3.2 Computational 101
4.3.3 Composition of the chemical steps pool 112
4.3.4 Mathematical approximations 113
4.4 Results and Discussion 129
4.4.1 Mechanism of R3 aggregation 129
4.4.2 Mechanism of action of D-TLKIVW on R3 aggregation 155
4.4.3 Integration of the kinetic model with structural information 180
4.4.4 Discussion of the performance of the modelling algorithm 180
4.5 Conclusions 182
4.6 Further work 184

4.1 Abstract

The mechanism of aggregation of tau is thought to proceed via a nucleation mechanism, yet the nature of the nucleus is ill-defined. In this chapter, we used

NMR-monitored kinetic traces of monomer and dimer concentration of tau's model system R3 (established in chapter 2) to numerically optimise models of protein aggregation, as well as explicit models of inhibition by literature peptide D-TLKIVW. The modelling process followed the algorithm abstractly presented in chapter 3 and was performed as a global fit, using reactions at multiple concentrations of protein and/or inhibitor. The early aggregation steps in the pathway of tau R3 aggregation corresponded to dimerisation and nucleation of a tetrameric nucleus from a dimer and two monomers; an initial monomeric unimolecular step was also required in the case of aggregation in D_2O based buffer. Elongation proceeded via addition of monomer and dimer, with distinct rates, and a secondary process, fragmentation, was also important for the goodness of fit. Importantly, the mechanism of tetrameric nucleation could be kinetically distinguished as a better fit than related mechanisms. In the second stage of the modelling process, the inhibitor D-TLKIVW was found to have a multifaceted mechanism of action on tau, involving binding to small species, nucleus disruption and surface binding to fibrils. Systematic pruning stages were performed for both modelling processes to avoid overfitting and allow a more nuanced understanding of the relative importance of individual chemical steps within the mechanisms. Further work could involve exploring structural aspects of the monomer and dimer species observed, as well as their interactions with inhibitor predicted by the model. Finally, we discuss the implications of this mechanistic model in relationship with atomic resolution cryoEM structures of AD brain derived tau fibrils.

4.2 Introduction

This chapter concerns the kinetic modelling of the aggregation of tau's R3 fragment, the model system discussed in the previous chapter, and the inhibitory action by the all D-amino acid peptide (D-peptide) literature inhibitor D-TLKIVW. We provide below a detailed introduction to this inhibitor as well as sum up the experimental insights on the R3 model, discussed in Chapter 2, which are most relevant for the modelling process.

4.2.1 Inhibition of tau by D-TLKIVW

The inhibitor used in this study is the first D-peptide developed to inhibit tau aggregation, the hexapeptide D-TLKIVW developed by Sievers *et al.* [134]. In this pioneering study, a series of peptides was computationally designed based on the X-ray crystal structure (obtained from microcrystals) of the R3 hexapeptide VQIVYK [66], one of tau's two aggregation hotspots [72]. VQIVYK packs in what the authors have defined as a dry steric zipper, in which a pair of β-sheets are interdigitated via the interface-facing side chains (see Figure 4.1 A). Sawaya *et al.* crystallised fragments of a variety of aggregating proteins, which presented a variety of packing arrangements; tau's VQIVYK in particular packed in individual β sheets made of parallel in-register strands (see Figure 4.1 A i and ii), an interface consisting of the same residues from each sheet (Val306, Ile308 and Tyr310), with both sheets parallel as well in the longitudinal axis [66]. D-TLKIVW was docked on the VQIVYK structure and packed on top of either β-sheet, forming backbone hydrogen bonds with a few residues as well as a bond with the side chain of Gln307. From the peptides selected by the computational design approach, D-TLKIVW was the strongest inhibitor of tau aggregation, delaying the onset of aggregation of the hexapeptide VQIVYK and the 3R construct K12 (see Figure 4.1C) in a concentration dependent manner [134].

4.2.2 Summary of qualitative findings on R3 aggregation from the previous chapter

From the previous chapter, we know that there are two visible species in the reaction, which correspond to monomer and dimer. Also, the formation of dimer under oxidative conditions is necessary for the primary (but not seeded) aggregation of R3, suggesting it is an on-pathway species. These observations were used here to inform the design of the pool of steps to be tested following the systematic strategy and optimisation settings described in Chapter 3. Finally, the inhibitor D-TLKIVW delays aggregation of R3 as shown by monitoring with ThT fluorescence, which makes it a suitable candidate for mechanism modelling using our method.

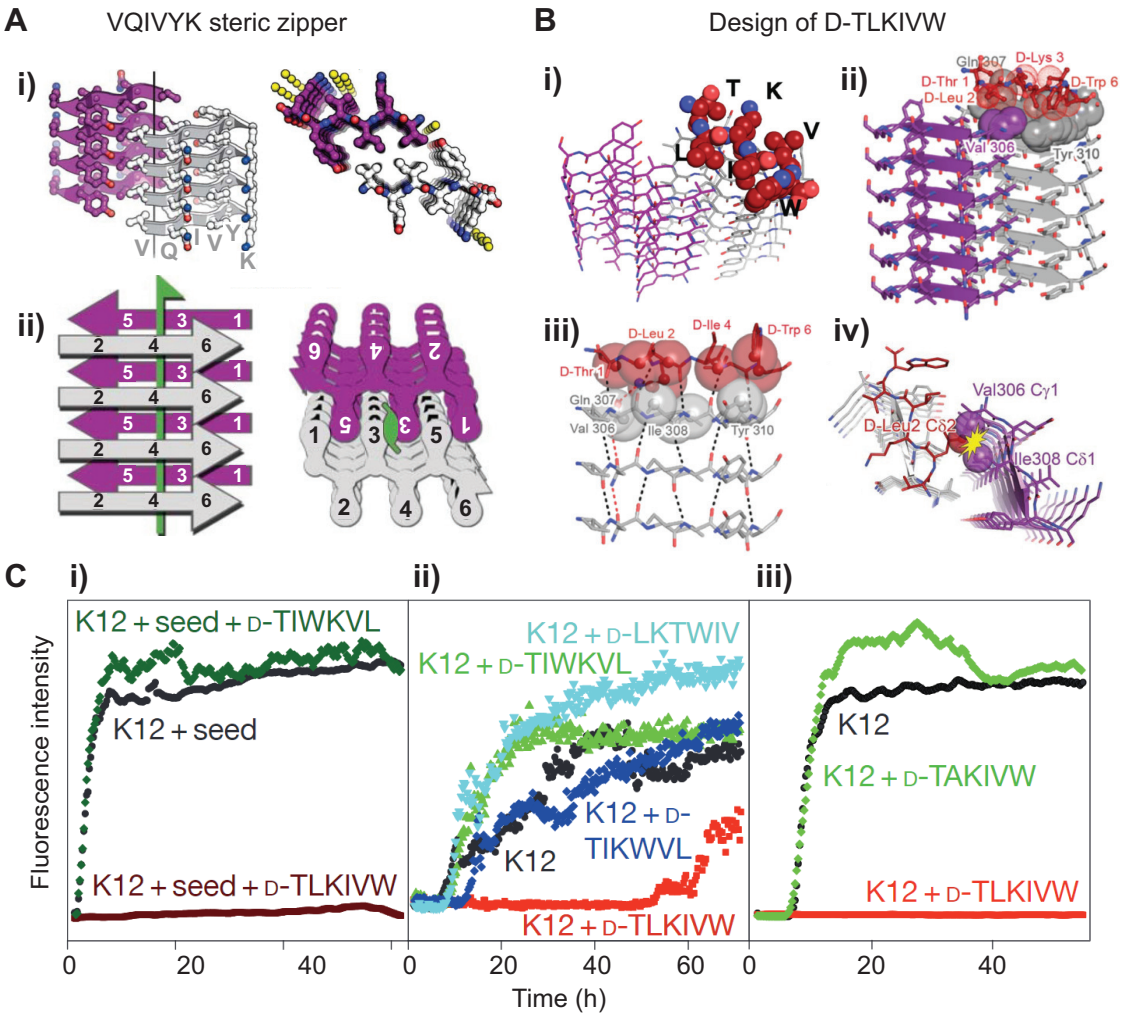


Figure 4.1: D-TLKIVW, a computationally designed inhibitor of tau aggregation. A. Atomic resolution structure of VQIVYK aggregates (adapted from [66]). (i) The interdigitated interface between the two β-sheets (purple and grey, shown in stick and ribbon representation) is shown from a side-view (left) and a top view (right). (ii) Schematic representation of the class 1 steric zipper formed by VQIVYK, from a side-view (left) and a top view (right), with the green arrow indicating the direction of growth of the fibrils. Each sheet is in register and arranged in a parallel configuration, also parallel with respect to each other along the growth axis. B. D-TLKIVW represented at the docked binding site (adapted from [134]. (i) Sphere representation of D-TLKIVW on top of one of the strands of the steric zipper. (ii) Side view showing the packing of D-TLKIVW on top of the grey strand and interacting with Val306 on the opposite strand. (iii) Close side-view on the packing of D-TLKIVW on top of the β sheet strand (the purple strand has been removed). Hydrogen bonds in black correspond to side chain bonds along the β sheet. Several bonds are predicted between the inhibitor and the strand residues. (iv) Steric clash between the Leu2 of D-TLKIVW docked on the grey strand and a VQIVYK molecule at the same z position in the fibril axis, but on the purple strand, suggesting elongation would be blocked at both strands. C. Inhibition of aggregation by D-TLKIVW (adapted from [134]). (i) Inhibition of seeded aggregation of tau K12 (using 50 µM K12, 0.25 % K12 seeds v/v and 500 µM of either peptide). (ii) D-TLKIVW, but not a scrambled version, inhibits primary aggregation of tau K12. (iii) The inhibition of primary aggregation is lost if Leu2 is replaced by alanine, in agreement with the steric clash represented in B iv. In ii) and iii), equimolar 50 µM concentrations of K12 and peptides were used.

In this chapter, we apply the algorithm presented in chapter 3 to the model system peptide R3. We fit the data to an extensive set of models, from which a most likely mechanism of R3 aggregation is found. Next we apply the same method to probe a variety of inhibitory actions of the peptide D-TLKIVW and arrive at a best model of inhibition.

4.3 Methods

4.3.1 Experimental

Kinetic traces were obtained following the method described in the previous chapter, with the only difference that the second dataset of R3-only experiments was carried out in 90% D_2O-based PBS. The datasets contained the following traces:

- R3-Only Dataset 1: 50 µM, 100 µM (2 replicates) and 400 µM.
- R3-Only Dataset 2: 47 µM, 94 µM (2 replicates, one of them used only for cross-validation), 188 µM (2 replicates) and 376 µM.
- R3+D-TLKIVW: 100 µM R3 + 121 µM D-TLKIVW, 100 µM R3 + 48 µM D-TLKIVW and 50 µM R3 + 134 µM D-TLKIVW.

The R3 concentrations refer to the equivalent total monomeric concentration.

4.3.2 Computational

Initial rates fitting

We initially aimed to probe the early stage reactions involving monomer by using the initial rates method, a type of differential method that can be applied to experimentally determine the rate law of a reaction. If we consider a reaction involving m monomers (P_1) and n dimers (P_2), the rate equation can be written as: $v = k{P_1}^m{P_2}^n$ where k is the rate constant and m and n represent the order of the reaction with respect to monomer and dimer, respectively. If we perform experiments with variable monomer concentration and constant dimer concentration, measure the monomer initial rates of decay as $v_0 = \frac{dP_1}{dt}$, and fit a straight line to a

logarithmic plot of rate versus monomer concentration, the slope of the line would indicate the order of the reaction with respect to monomer. In the case of a reaction involving only monomer, the requirement of having constant dimer concentration is removed, although if the dimer is the product of the reaction it should ideally not be present in a significant amount. Even though in our case there was dimer signal present at the early timepoints (especially for the first dataset), we applied the approach described above as an approximation to gain insight on the mechanism. By taking all the datasets and considering the monomer signal as a direct proxy for concentration, we fitted the initial rates of reaction (in units of monomer signal $\cdot h^{-1}$) as a straight line of the form $P_1signal(t) = v_0 t + P_1signal_0$, using the datapoints for the first 2 hours (first dataset) or the first 4 hours (second dataset). Then, independently for each dataset, we fitted $ln(v_0) = ln(k) + mln(P_1signal_0)$, yielding m (the reaction order for monomer) as the slope. The fits were performed using the fit function of Gnuplot, which uses an implementation of non-linear least-squares fitting, with equal weighting of each datapoint.

Overview of model exploration strategy

The modelling process followed the algorithm described and discussed in Chapter 3. The analysis was performed in two stages: we first used the data of aggregation of R3-only (1st dataset) to obtain a model we call "best protein-only model". Second, we tested models of inhibition using the same principle, and starting with a model that included all the chemical steps of the "best protein-only model". The rates of those steps were kept fixed at the values of the best protein only fit, and only the rates of the inhibitor related steps were allowed to vary during the optimisation. In parallel, another dataset of R3-only was acquired, in D_2O based buffer, and was modelled independently using the same strategy as the first R3-only dataset.

As described in Chapter 3, we calculated an average per point $\overline{\chi^2}_{RSS}$ by dividing the residual sum of squares by the number of datapoints. For the rest of this Chapter, we refer to this $\overline{\chi^2}_{RSS}$ as χ^2. In all cases, the criterion for accepting a model of higher complexity, i.e. with more fitted parameters, (or for reducing

complexity in pruning stages) was an improvement to the second significant figure (0.1) with respect of the simpler one. If the models had equal complexity, the lowest χ^2 model was always accepted.

Broadly speaking, there were three types of model testing generations:

- **Model building.** A parent model was the basis for generating more complex children models, each of which shared all the chemical steps of the parent and included an additional chemical step from the pool.
- **Pruning.**. In order to prevent overfitting, the potential overall winner model was pruned by generating a set of models that each lacked a specific chemical step of the original model.
- **Specific hypothesis probing.** Specific steps were tested against alternatives (i.e. nucleation tests for protein-only first dataset, inhibition mechanisms for the inhibitor dataset). This represents a deviation from the algorithm outlined in Chapter 3 and is discussed in the Results section.

R3-only 1st dataset

We started with a model of linear polymerisation, designed a pool of steps to test and tested them in the systematic way described in the previous chapter, in which each model tested included a new step and all the rates of the steps in the model were optimised. This was done for a total of four generations until the pool of steps was exhausted. Polishing (explained below) was applied from generation 4. The winner of generation 4 was pruned (1^{st} pruning) and in parallel became the basis for nucleation tests (in which the nucleation step was switched for other steps from the original pool, with all other steps remaining the same). Extra models were retrospectively added to generation 2 in parallel to the nucleation tests from the winner of generation 4. There was a final 2^{nd} pruning stage from the winner of the 1^{st} pruning stage.

R3-only 2nd dataset We included extra steps into the pool (explained below) and followed the algorithm in a similar fashion, involving 4 model building generations and two pruning generations.

Inhibitor dataset

There were 11 generations of models, in which polishing was turned on from generation 7. Two generations of pruning were performed from the winner of generation 10. In parallel to the 2nd pruning stage (i.e. with the same parent, the winner of the 1st pruning stage), some models were tested to probe specific hypotheses.

Model Optimisation

Implementation of model Each model consists of a set of chemical steps. In terms of implementation, there are two main kinds of chemical steps, those that involve a single equilibrium (such as dimerisation from two monomers, or inhibitor binding to the nucleus), and those which occur iteratively, such as the elongation of fibrils, or the binding of inhibitor to fibrils. The latter have been implemented using for loops. Each model is implemented on a per chemical equilibria basis, since that facilitates keeping track of species involved in many different steps. As an example, this is the chemical equilibrium, differential equations and code implementation for the dimerisation / nucleation step in which two monomers reversibly form a dimer:

Chemical equilibria

$$\mathrm{P}_1 + \mathrm{P}_1 \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{P}_2 \tag{4.1}$$

$$\frac{\mathrm{d}[\mathrm{P}_1]}{\mathrm{dt}} = -2\mathrm{k}_+[\mathrm{P}_1]^2 + 2\mathrm{k}_-[\mathrm{P}_2] \tag{4.2}$$

$$\frac{\mathrm{d}[\mathrm{P}_2]}{\mathrm{dt}} = \quad \mathrm{k}_+[\mathrm{P}_1]^2 - \quad \mathrm{k}_-[\mathrm{P}_2] \tag{4.3}$$

Rate constants

//rates

kp[0] = fabs(*(params->m["kp1"]));

```
km[0] = fabs(*(params->m["km1"]));
```

where kp1 and km1 correspond to k_+ and k_- as defined in the above equilibria. The arrays kp and km get multiplied by dt in a different function, prior to the calculation of the derivative and update of the system at time $t + dt$.

Calculation of concentrations at time $t + dt$

```
//forward
double kpcurr1 = kp[0]*xpt[0]*xpt[0];
xp[0] += -kpcurr1;
xp[0] += -kpcurr1;
xp[1] += kpcurr1;

//backward
double kmcurr1 = km[0]*xpt[1];
xp[1] += -kmcurr1;
xp[0] += kmcurr1;
xp[0] += kmcurr1;
```

where the arrays xp and xpt are identical before the execution of the code above, tracking the concentration of monomer (index 0) and dimer (index 1) at time t, and xp gets updated with the concentrations at time $t + dt$.

The complete equations and implementations of the best D-TLKIVW inhibition model, which includes the best R3-only model (first dataset) are included in the Appendix.

Parameters The parameters can be grouped in three categories:

- Initial conditions parameters.
- Rate of reaction parameters.

- Others.

Initial conditions parameters. The parameter "frac" corresponds to the fraction of total equivalent monomeric concentration that is in monomeric state at time 0. The parameter "srac" is used to relate the signal intensity to equivalent monomeric concentration (in µM). They are expressed as follows, where P_1 refers to monomer, and P_2 to dimer:

$$frac = \frac{[P_1]_0}{[P_1]_0 + 2[P_2]_0} \tag{4.4}$$

$$srac = \frac{P_1 signal}{[P_1]} = \frac{P_2 signal}{2[P_2]} \tag{4.5}$$

The concentrations of monomer and dimer are initialised as:

$$[P_1]_0 = frac[P_{Tot}] \tag{4.6}$$

$$[P_2]_0 = \frac{[P_{Tot}](1 - frac)}{2} \tag{4.7}$$

where P_{Tot} is the total equivalent monomeric concentration. For the calculation of χ^2 between the fit and the data, the concentrations of monomer and dimer are converted to signal as follows:

$$P_1 signal = srac[P_1] \tag{4.8}$$

$$P_2 signal = 2srac[P_2] \tag{4.9}$$

The optimiser can either fit a global value for these parameters (i.e. the same for all samples) or can fit individual srac and frac values to account for sample variability and/or pipetting errors (called "polishing"). For the protein-only modelling process, initially all models were fit with a global frac and srac, and later the process of "polishing" allows both to be fitted individually. In particular, it's important to note that allowing srac to vary results in effectively changing the concentration ratios between samples, and should be done only once it has been established that that particular model can fit the data well with a single global srac. In the second protein-only dataset, frac was kept fixed at a value of 0.9995 throughout

the modelling process and only allowed to vary in the polishing stage. For the modelling of inhibitor action, initially frac and srac were not allowed to vary; they were only allowed to vary, and in a sample-specific way, at the polishing stage.

Rate of reaction parameters. Those are the parameters that correspond to the rates of chemical steps, such as in this example:

$$\mathrm{P}_1 + \mathrm{P}_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P}_2 \tag{4.10}$$

Within the code implementation of the system of ODEs, these are multiplied by dt. All reaction rates have units of h^{-1} (unimolecular, such as k_1) or $\mu\mathrm{M}^{-1}$ h^{-1} (bimolecular, such as k_{-1}), with the exception of nucleation steady-state approximations which are discussed later in more detail.

Other parameters These are parameters that don't correspond to rates but are part of the definition of a chemical step, such as the "n" in the steady state approximation of nucleation described below. These don't get multiplied by dt.

Optimisation settings The overall optimisation process started with the calculation of the χ^2 using the input parameters (initial simulated χ^2). Then there was an initial gradient-descent step, which yielded the initial reference χ^2. This χ^2 (and the corresponding modified parameters) was accepted if it was lower than the initial simulated χ^2, and stored as the first reference χ^2. Otherwise, the input parameters and the initial simulated χ^2 were taken as the reference. Then, a series of iterations of a perturbation-minimisation process started: in each iteration, the parameters would be perturbed, and gradient-descent would be applied to find a new minimum. The χ^2 after gradient-descent would then be compared to the reference best χ^2: the new parameters would be accepted (and the best χ^2 updated), if the χ^2 had been reduced. In the process of calculating the χ^2, penalties would be applied. The optimisation was set to stop if 20 iterations of perturbation-minimisation had been performed without an improvement of the reference best χ^2. Settings were defined for each of the procedures (integration, penalty calculation, gradient-descent and basin-hopping) as follows:

- **Integration.** δt for calculation was set by dividing the total time by the parameter "rune", which was set as $5 \cdot 10^4$. δt ranged between 1.4 s and 8.4 s depending on the duration of the experiments (20-120 hours). The maximum length of fibrils was set at 800 (this corresponds to N/2 in the implementation in the Appendix).

- χ^2 **penalties.** Maximum thresholds were set for the parameters srac and frac. For frac, the threshold of 1 corresponds to a physical limit, since above that there would need to be a negative dimer concentration. For srac, the threshold was set to 5. The penalties were applied as:

$$Penalty = f * (deviation)^2 \tag{4.11}$$

  where f is a multiplication factor set to 10^5 and deviation is the squared difference between the value and the threshold (to ensure it's a positive number).

- **Gradient-descent.** We used the gsl_multifit_nlinear library (GSL 2.7), with gsl_multifit_nlinear_trust as the type of nonlinear least squares minimization method, and using Levenberg-Marquadt (default) as the solver. The method of integration was forward Euler, as discussed in chapter 3. Most parameters were set at their default values using gsl_multifit_nlinear_default_parameters(), unless mentioned explicitly. The increment step for the derivative calculation was set to 10^{-15} h. The descent was set to converge with gtol and xtol of 10^{-4}, ftol=0 and maximum iterations = 5.

- **Basin-Hopping.** Each parameter was perturbed by adjusting it according to a lognormal distribution. The setting "heat" corresponded to the σ of the distribution and was varied between 0.01 and 0.2 in separate optimisation runs. For each parameter initialisation mode (see below), an optimisation was run with four heat values. The parameter-polishing runs (explained below) were run three times with heat 0.01,0.05 and 0.1.

Parameter initial values There were two main modes of parameter initialisation, called "naive guess" and "inherited guess".

Naive guess mode. For protein only fitting, initial values of global frac and srac were estimated as an average between the calculated frac and srac for each sample. The following expression relates frac to the initial monomeric ($P_1signal_0$) and dimeric signal ($P_2signal_0$) as:

$$frac = \frac{\frac{P_1signal_0}{srac}}{\frac{P_1signal_0}{srac} + \frac{P_2signal_0}{2srac}} \tag{4.12}$$

$$= \frac{2P_1signal_0}{2P_1signal_0 + P_2signal_0} \tag{4.13}$$

$$= \frac{2}{2+\alpha} \tag{4.14}$$

where

$$\alpha = \frac{P_2signal}{P_1signal} \tag{4.15}$$

From frac we can calculate srac as:

$$srac = \frac{P_1signal_0}{[P_{Tot}]frac} \tag{4.16}$$

where P_{Tot} corresponds to the total monomeric equivalent concentration (in µM).

For the modelling of inhibitor actions, individual frac and srac were initialised (in a fixed way, not fitted) as the fitted parameter values from the best protein only model of the sample with the same concentration (of protein); the values from one of the two protein-only 100 µM replicates were used since this sample had the closest initial monomer and dimer signal to both inhibitor samples with 100 µM protein.

The remaining parameters corresponding to rates were initialised at 10^{-9}, except for the initialisation of rates for linear polymerisation (the first protein only model) which were initialised as 10^{-4} (k_+) and 10^{-6} (k_-). For the dimer+n monomers steady-state nucleation step (described below), n was initialised either at 10^{-9}, 1, 2, 3 or 4.

Inherited guess mode. For the chemical steps that were common between a model and its parent, the final parameter values of the parent model were passed

to the child model as initial guesses. This applied to all parameter types, including frac, srac, rates and any other non rate parameters.

For the main R3-only dataset, the modelling included tests of different nucleation mechanisms at a late stage, i.e. stemming from and comparing to the winner of the 1st pruning generation. For these tests, in addition to the "naive guess" and "inherited guess" parameter initialisation modes, the following variations of the inherited modes were used:

- **Nucleation-naive inherited guess mode.** All rates were inherited except the forward and backward rates of the nucleation reaction (which is different to the nucleation reaction in the parent model), which were initialised at 10^{-9}.
- **Nucleation-slowed inherited guess mode.** All rates were inherited, and the forward and backward rates of the nucleation reaction were divided by a factor of 10.

Polishing runs Regardless of the inheritance mode, after the main optimisation run, polishing (fitting sample-specific frac and srac parameters) was run if the main optimisation had been approved (according to the metrics described in the next section). For polishing, the fitted values were given as initial guesses for the polishing run, in which all parameters and sample-specific srac and frac were allowed to vary. Three polishing runs were done for each main run, using "heat" settings of 0.01, 0.05 and 0.1, and the best result was taken as the polished result.

Direct-polishing runs Specifically for the nucleation tests in the first protein-only dataset, all the four inheritance modes were also run, in parallel to the usual two-step optimisation, directly with polishing (sample-specific frac and srac) turned on.

Validation of individual optimisation results A series of validation steps were implemented to ensure the quality of the optimisation results. These metrics, described below, were calculated for each run (main and polishing), independently for each sample and the largest value amongst samples was compared to the threshold of acceptance.

- **Constant Mass.** We assessed any deviations of the calculated total monomeric equivalent protein concentration and the sample concentration, as follows:

$$\delta[P] = \frac{\mid [\overline{P}]_{calc} - [P] \mid}{[P]} \cdot 100 \tag{4.17}$$

where $[\overline{P}]_{calc}$ was the average calculated mass (over time). We also assessed that the mass fluctuation over time would not be too large, by calculating:

$$\sigma_{P\,fractional} = \frac{\sigma_P}{[P]} \cdot 100 \tag{4.18}$$

This was calculated for each sample, the test was passed if the largest concentration deviation and the largest fractional standard deviation (within the dataset) were both below 0.5%, and failed otherwise.

- **Absence of negative mass.**

The fractional negative oligomeric concentration was calculating by summing over the negative and positive oligomeric masses, over all time points, as follows:

$$[Oligomer]_{negative} = \frac{\sum[P_n]_{negative}}{\sum[P_n]_{negative} + \sum[P_n]_{positive}} \cdot 100 \tag{4.19}$$

The results were approved if the largest fractional negative oligomeric concentration was below 0.5%.

- **Numerical integration accuracy.**. The ${\chi^2}_{true}$ was calculated using a non-linear ODE solver (gsl_odeiv2_step_msadams). Then the χ^2 calculated with the Euler method (the outcome of the optimisation) was compared to the true χ^2:

$${\chi^2}_{error} = \frac{\mid {\chi^2}_{Euler} - {\chi^2}_{true} \mid}{{\chi^2}_{true}} 100 \tag{4.20}$$

The accuracy test was passed if ${\chi^2}_{error} < 0.5\%$, and failed otherwise.

4.3.3 Composition of the chemical steps pool

Protein-only pool

Figure 4.2 shows the steps that were included in the pool for modelling the protein-only data. Steps in panel A(i) were used for both datasets, while steps in panel A(ii) were only used for the first dataset, and steps B were only used in the second dataset pool.

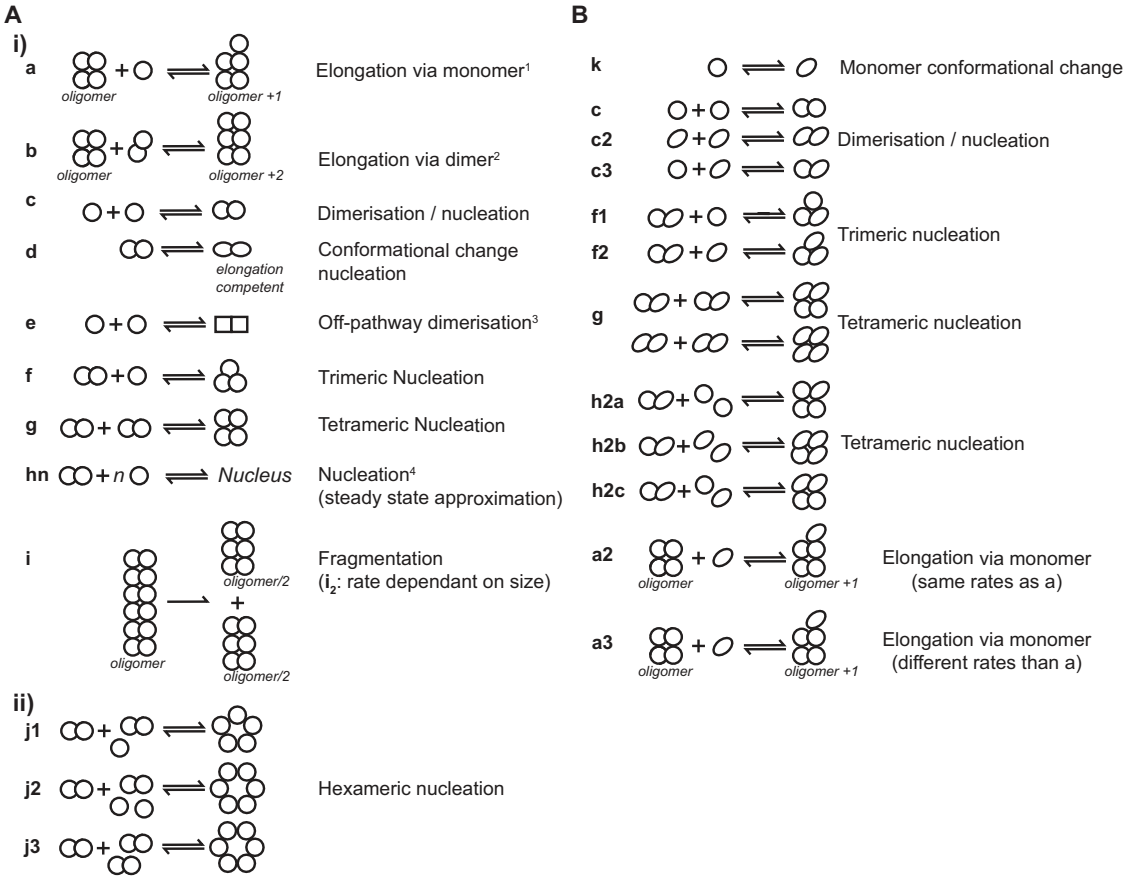


Figure 4.2: Pool of steps proposed for the aggregation of R3. A. (i) Steps included in both datasets. (ii) Steps included only in the first datset, in the nucleation tests. B. Steps included only for the second dataset. Notes: 1: If a is the only reaction, the mechanism of the model corresponds to linear polymerisation. 2: In the second dataset, the step b (elongation by dimer; shown in A) involves either of the dimers that can be formed in steps c1, c2 or c3, depending on which one is part of the model. 3: In this reaction, the resulting dimer doesn't participate in any further reactions.4: n can either be free to vary (h_n) or fixed (e.g.h_2). If n=1, this step is equivalent to f.

Inhibitor pool

For the inhibitor modelling, the steps shown in Figure 4.3 were included. It is important to note that steps i', j' and k', corresponding to elongation and fragmentation reactions of inhibitor-bound oligomers, were only added to the pool from generation 4, once inhibitor binding to the oligomers had been selected as part of the generation winning model (in generation 3).

Regarding the inhibitor actions that affect the nucleus, binding and disruption were tested both independently (c'1 and f'1) and as a single composite mechanism (c'1 and f'2) of the following form:

$$\mathrm{P}_4 + \mathrm{I} \underset{k_{-b}}{\overset{k_b}{\rightleftharpoons}} \mathrm{P}_4\mathrm{I} \tag{4.21}$$

$$\mathrm{P}_4\mathrm{I} \xrightarrow{k_d} \mathrm{P}_2 + 2\,\mathrm{P}_1 + \mathrm{I} \tag{4.22}$$

4.3.4 Mathematical approximations

Single-step steady-state nucleation reaction

Step h of the protein-only pool, with all its variations, corresponds to a multistep nucleation mechanism involving a dimer and 2 or a variable "n" number of monomers, which has been modelled as a single-step steady state approximation.

Single-step tetrameric nucleation The full derivation of this approximation starts with a single-step tetrameric nucleation in which a dimer (P_2) initially binds a monomer (P_1) to form a trimer (P_3), which then binds another monomer to form the tetrameric nucleus (P_4). Each reaction is associated with forward and backward rates as follows:

$$\mathrm{P}_2 + \mathrm{P}_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P}_3 \tag{4.23}$$

$$\mathrm{P}_3 + \mathrm{P}_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{P}_4 \tag{4.24}$$

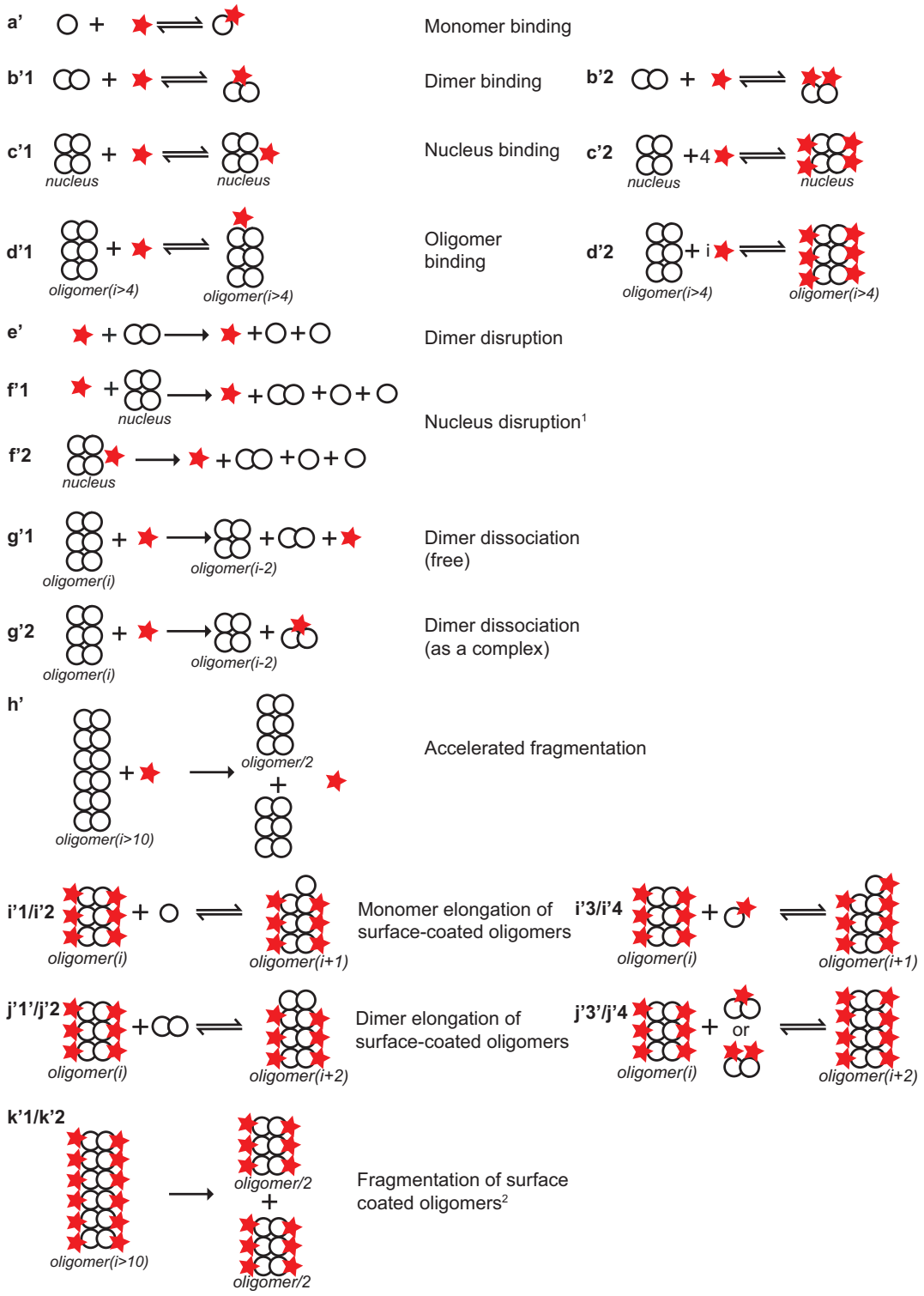


Figure 4.3: Pool of steps proposed for the inhibition of R3 aggregation by D-TLKIVW. In steps i', j' and k', the following details apply: 1 and 3 indicate the rates of the reaction are fixed at the same values the protein-only best model reactions, 2 and 4 they are allowed to vary. Note that in several steps (i'1,i'2, j'1, j'2, j'3 and j'4 elongating by dimer-single inhibitor complex) the current depiction would lead to a mass imbalance; this imbalance and the necessary correction applied to compensate for it are discussed in detail in the next section. Steps i' to k' were added to the pool in generation 4 once step d'2 had been included in the parent model. For steps i and j, if used in a model including nucleus binding (c'1/c'2), two variants were tested: one in which the inhibitor-nucleus complex can also participate in those reactions (b) and one in which it doesn't (a).

Thus, the system of ODEs that describes the evolution in the concentration of all species is the following:

$$\frac{\mathrm{d[P_1]}}{\mathrm{dt}} = -\mathrm{k_1[P_1][P_2]} + \mathrm{k_{-1}[P_3]} - \mathrm{k_2[P_1][P_3]} + \mathrm{k_{-2}[P_4]} \tag{4.25}$$

$$\frac{\mathrm{d[P_2]}}{\mathrm{dt}} = -\mathrm{k_1[P_1][P_2]} + \mathrm{k_{-1}[P_3]} \tag{4.26}$$

$$\frac{\mathrm{d[P_3]}}{\mathrm{dt}} = +\mathrm{k_1[P_1][P_2]} - \mathrm{k_{-1}[P_3]} - \mathrm{k_2[P_1][P_3]} + \mathrm{k_{-2}[P_4]} \tag{4.27}$$

$$\frac{\mathrm{d[P_4]}}{\mathrm{dt}} = +\mathrm{k_2[P_1][P_3]} - \mathrm{k_{-2}[P_4]} \tag{4.28}$$

If we assume that the species P_3 is unstable and therefore never builds up, we can consider it in a steady state ($\frac{d[P3]}{dt} = 0$) and we can isolate $[P_3]$ as a function of $[P_1]$, $[P_2]$ and $[P_4]$:

$$\mathrm{[P_3]} = \frac{\mathrm{k_1[P_1][P_2]} + \mathrm{k_{-2}[P_4]}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}$$

Substituting this into the original ODE system yields:

$$\frac{\mathrm{d[P_1]}}{\mathrm{dt}} = -2\frac{\mathrm{k_1k_2}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_1]^2[P_2]} + 2\frac{\mathrm{k_{-1}k_{-2}}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_4]} \tag{4.29}$$

$$\frac{\mathrm{d[P_2]}}{\mathrm{dt}} = -\frac{\mathrm{k_1k_2}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_1]^2[P_2]} + \frac{\mathrm{k_{-1}k_{-2}}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_4]} \tag{4.30}$$

$$\frac{\mathrm{d[P_4]}}{\mathrm{dt}} = +\frac{\mathrm{k_1k_2}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_1]^2[P_2]} - \frac{\mathrm{k_{-1}k_{-2}}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}}\mathrm{[P_4]} \tag{4.31}$$

The rates in these equations can be redefined as single composite forward and backward rates, $\mathrm{k_s}$ and $\mathrm{k_{-s}}$ (the subindex s stands for "steady-state approximation"):

$$\mathrm{k_s} = \frac{\mathrm{k_1k_2}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}} \tag{4.32}$$

$$\mathrm{k_{-s}} = \frac{\mathrm{k_{-1}k_{-2}}}{\mathrm{k_{-1}} + \mathrm{k_2[P_1]}} \tag{4.33}$$

However, in these expressions, the composite rates have a dependency on the concentration of monomer ($[P_1]$), which is not constant. We can simplify the expressions and remove the dependency, in order to be able to fit these composite rates as a single rate, by making the assumption that $\mathrm{k_{-1}} \ggg \mathrm{k_2}[P_1]$. This can be interpreted as the energy barrier to the final product P_4 being very high (see

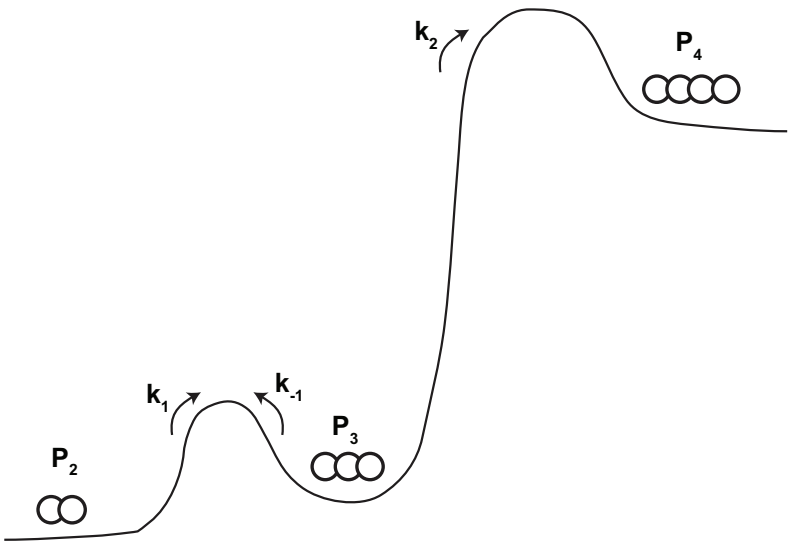


Figure 4.4: Free energy landscape for the two-step tetrameric nucleation mechanism. There is a pre-equilibrium between the dimer P_2 and the trimer P_3, and the formation of tetramer or P_4 is the rate limiting step, with a high energy barrier between P_3 and P_4, which corresponds to a small k_2 rate (relative to k_{-1}).

Figure 4.4). This further supports the earlier assumption of a steady-state of P_3, since if the forward rate to P_4 is very slow compared to the breakage of P_3, P_2 and P_3 will be in a fast equilibrium. This simplifies the rates as:

$$\mathrm{k_s} = \frac{\mathrm{k_1 k_2}}{\mathrm{k_{-1} + k_2[P_1]}} \approx \frac{\mathrm{k_1 k_2}}{\mathrm{k_{-1}}} \tag{4.34}$$

$$\mathrm{k_{-s}} = \frac{\mathrm{k_{-1} k_{-2}}}{\mathrm{k_{-1} + k_2[P_1]}} \approx \frac{\mathrm{k_{-1} k_{-2}}}{\mathrm{k_{-1}}} = \mathrm{k_{-2}} \tag{4.35}$$

Finally, this can be represented as a single step reaction, with the approximated composite rates:

$$\mathrm{P_2 + 2\,P_1} \underset{k_{-s}}{\overset{k_s}{\rightleftharpoons}} \mathrm{P_4} \tag{4.36}$$

Figure 4.4 shows a potential energy landscape for this multistep reaction, consistent with the approximations made, which corresponds to step h_2 in Figure 4.2.

Single-step dimer + n monomers nucleation

Building on the previous approximation, we can further generalise this single step nucleation reaction as a reaction between a dimer and a number n of monomers (in intermediate steps), corresponding to step h_n in Figure 4.2, which opens the

possibility to allow n to vary during the optimisation. The multiple step chemical reaction can be represented as:

$$\mathrm{P}_2 + \mathrm{P}_1 \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P}_3 \tag{4.37}$$

$$\mathrm{P}_3 + \mathrm{P}_1 \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{P}_4 \tag{4.38}$$

$$\mathrm{P}_4 + \mathrm{P}_1 \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} \mathrm{P}_5 \tag{4.39}$$

$$\vdots \tag{4.40}$$

$$\mathrm{P}_i + \mathrm{P}_1 \underset{k_{-(i-1)}}{\overset{k_{(i-1)}}{\rightleftharpoons}} \mathrm{P}_{(i+1)} \tag{4.41}$$

$$\vdots \tag{4.42}$$

$$\mathrm{P}_{(n+1)} + \mathrm{P}_1 \underset{k_{-n}}{\overset{k_n}{\rightleftharpoons}} \mathrm{P}_{(n+2)} \tag{4.43}$$

where i = 2,3,...,n+1

We have already shown that we can approximate the two step nucleation to a tetramer to a single step with composite rates. Now, we want to prove by induction that the same is true for the n^{th} approximation, such that:

$$\mathrm{P}_2 + \mathrm{iP}_1 \underset{k_{-s}}{\overset{k_s}{\rightleftharpoons}} \mathrm{P}_{(i+2)} \tag{4.44}$$

$$\mathrm{P}_{(i+2)} + \mathrm{P}_1 \underset{k_{-(i+1)}}{\overset{k_{(i+1)}}{\rightleftharpoons}} \mathrm{P}_{(i+3)} \tag{4.45}$$

can be approximated to:

$$\mathrm{P}_2 + (\mathrm{i}+1)\,\mathrm{P}_1 \underset{\tilde{k}_{-s}}{\overset{\tilde{k}_s}{\rightleftharpoons}} \mathrm{P}_{(i+3)} \tag{4.46}$$

Similarly as before, the equations that represent these two steps are:

$$\frac{\mathrm{d}[\mathrm{P}_1]}{\mathrm{dt}} = -\mathrm{ik_s}[\mathrm{P}_2][\mathrm{P}_1]^{\mathrm{i}} + \mathrm{ik_{-s}}[\mathrm{P}_{(\mathrm{i}+2)}] - \mathrm{k_{i+1}}[\mathrm{P}_1][\mathrm{P}_{(\mathrm{i}+2)}] + \mathrm{k_{-(i+1)}}[\mathrm{P}_{(\mathrm{i}+3)}] \tag{4.47}$$

$$\frac{\mathrm{d}[\mathrm{P}_2]}{\mathrm{dt}} = -\mathrm{k_s}[\mathrm{P}_2][\mathrm{P}_1]^{\mathrm{i}} + \mathrm{k_{-s}}[\mathrm{P}_{(\mathrm{i}+2)}] \tag{4.48}$$

$$\frac{\mathrm{d}[\mathrm{P}_{(\mathrm{i}+2)}]}{\mathrm{dt}} = +\mathrm{k_s}[\mathrm{P}_2][\mathrm{P}_1]^{\mathrm{i}} - \mathrm{k_{-s}}[\mathrm{P}_{(\mathrm{i}+2)}] - \mathrm{k_{(i+1)}}[\mathrm{P}_1][\mathrm{P}_{(\mathrm{i}+2)}] + \mathrm{k_{-(i+1)}}[\mathrm{P}_{(\mathrm{i}+3)}] \tag{4.49}$$

$$\frac{\mathrm{d}[\mathrm{P}_{(\mathrm{i}+3)}]}{\mathrm{dt}} = +\mathrm{k_{(i+1)}}[\mathrm{P}_1][\mathrm{P}_{(\mathrm{i}+2)}] - \mathrm{k_{-(i+1)}}[\mathrm{P}_{(\mathrm{i}+3)}] \tag{4.50}$$

Similarly as before, we first consider the intermediate $P_{(i+2)}$ in a steady state where $\frac{d[P_{(i+2)}]}{dt} = 0$, and isolate the concentration of this intermediate with respect to the other species involved as:

$$[\mathrm{P_{(i+2)}}] = \frac{\mathrm{k_s}[\mathrm{P_2}][\mathrm{P_1}]^{\mathrm{i}} + \mathrm{k_{-(i+1)}}[\mathrm{P_{(i+3)}}]}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]} \tag{4.51}$$

Then we substitute $[P_{(i+2)}]$ in the equations of the system and obtain:

$$\frac{\mathrm{d[P_1]}}{\mathrm{dt}} = -(i+1)\frac{\mathrm{k_s k_{(i+1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_2}][\mathrm{P_1}]^{(\mathrm{i+1})} + (i+1)\frac{\mathrm{k_{-s}k_{-(i-1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_{(i+3)}}] \tag{4.52}$$

$$\frac{\mathrm{d[P_2]}}{\mathrm{dt}} = -\quad \frac{\mathrm{k_s k_{(i+1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_2}][\mathrm{P_1}]^{(\mathrm{i+1})} + \quad \frac{\mathrm{k_{-s}k_{-(i-1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_{(i+3)}}] \tag{4.53}$$

$$\frac{\mathrm{d[P_4]}}{\mathrm{dt}} = +\quad \frac{\mathrm{k_s k_{(i+1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_2}][\mathrm{P_1}]^{(\mathrm{i+1})} - \quad \frac{\mathrm{k_{-s}k_{-(i-1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]}[\mathrm{P_{(i+3)}}] \tag{4.54}$$

The dependency on monomer concentration can be removed from the composite rates, as before, by assuming that $\mathrm{k_{-s}} \ggg \mathrm{k_{(i+1)}}[P_1]$, which yields the following rates and "single-step" equilibrium:

$$\tilde{\mathrm{k}}_{\mathrm{s}} = \frac{\mathrm{k_s k_{(i+1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]} \approx \frac{\mathrm{k_s k_{(i+1)}}}{\mathrm{k_{-s}}} \tag{4.55}$$

$$\tilde{\mathrm{k}}_{\mathrm{-s}} = \frac{\mathrm{k_{-s}k_{-(i+1)}}}{\mathrm{k_{-s}} + \mathrm{k_{(i+1)}}[\mathrm{P_1}]} \approx \frac{\mathrm{k_{-s}k_{-(i+1)}}}{\mathrm{k_{-s}}} = \mathrm{k_{-(i+1)}} \tag{4.56}$$

This means that the complete system of successive monomer additions that yields $[P_{(n+2)}]$ can be modelled as a single step of the form:

$$\mathrm{P_2} + \mathrm{nP_1} \underset{\tilde{k}_{-s}}{\overset{\tilde{k}_s}{\rightleftharpoons}} \mathrm{P_{(n+2)}} \tag{4.57}$$

When n=1, the composite rates become single rates since there is only one chemical step, and when n=2, this reduces to the tetrameric nucleation case we proved previously.

Extensions of this approximation to pentameric and hexameric nucleation

The same steady-state approximation framework was applied to derive single-step reactions for the formation of pentamers and hexamers in the explicit formulations of hn, h3 and h4, as well as through the steps j1, j2, j3 (Figure 4.2).

Mass balance considerations for fragmentation reactions

Protein-only models We model fragmentation as a unimolecular irreversible reaction that breaks an oligomer or fibril in half:

$$\mathrm{P}_i \xrightarrow{k_f} \mathrm{P}_{(i/2)} + \mathrm{P}_{i-(i/2)} \tag{4.58}$$

where the subindex of P represents how many monomers it is made of (for example, P_i is an oligomer or fibril made of i monomers), accounting for the fact that i can be even or odd. The change of concentration of each species is defined as:

$$\frac{\mathrm{d[P_i]}}{\mathrm{dt}} = -\mathrm{k_f[P_i]} \tag{4.59}$$

$$\frac{\mathrm{d[P_{(i/2)}]}}{\mathrm{dt}} = +\mathrm{k_f[P_i]} \tag{4.60}$$

$$\frac{\mathrm{d[P_{i-(i/2)}]}}{\mathrm{dt}} = +\mathrm{k_f[P_i]} \tag{4.61}$$

In practice, since i needs to be an integer, the implementation of this reaction in the code is as follows:

$$\Delta xp[i] = -\mathrm{k_f}\, xp[i] \tag{4.62}$$

$$\Delta xp[frag] = +\mathrm{k_f}\, xp[i] \tag{4.63}$$

$$\Delta xp[i - frag + Corr] = +\mathrm{k_f}\, xp[i] \tag{4.64}$$

where $frag = i/2$(floored division), and the concentration of P_i is stored in location i of array xp (xp stores protein-only species), and so on. The index location of the second half fibril is i-frag+Corr, where Corr is an index correction necessary to keep mass balance constant for specific model implementations in which the size of a fibril monitored at xp[i] is not i monomers, which will be described in more detail below.

Justification of index correction Corr In the general case, let's assume that from the nucleus onwards, oligomers or fibrils of length (i+j) monomers are stored at index i of the array xp. For example, in a model with a dimeric nucleus stored at position xp[4], j=-2, and if a pentameric nucleus is stored at position xp[4],

j=1. For mass balance to be maintained, loss of substrate(s) needs to equal the gain in product(s). In this case in particular, we need to consider the equivalent monomeric mass/concentration:

$$Loss\,of\,Substrate = Gain\,of\,Products \tag{4.65}$$

$$\mathrm{k_f}\; xp[i]\;(i+j) = \mathrm{k_f}\; xp[i]\;(frag+j) + \mathrm{k_f}\; xp[i]\;(i-frag+Corr+j) \tag{4.66}$$

$$(i+j) = (frag+j) + (i-frag+Corr+j) \tag{4.67}$$

$$Corr = -j \tag{4.68}$$

We can see that $Corr = 0$ when $j = 0$, that is, when the location in the array xp corresponds to the number of monomers that make up that species, as expected. If that's not the case, modifying the index by Corr keeps mass balance, as desired.

Finally, there are instances of the implementation of the steady state dimer + n monomers approximation described above in which n is allowed to take non integer values during optimisation, which corresponds to allowing a mixture of nucleus sizes. These require a further concentration correction to keep mass balance in the fragmentation step.

Justification of concentration correction for models including steady-state dimer + n monomers nucleation, where n can be decimal

In the implementation of this nucleation mechanism, we arbitrarily define xp[4] as the location to store the nucleus, which contains 2+n monomers:

$$species : size$$

$$xp[4] : 2+n \tag{4.69}$$

$$xp[5] : 2+n+1 \tag{4.70}$$

$$... \tag{4.71}$$

$$xp[i] : 2+n+(i-4) = i+n-2 \tag{4.72}$$

From the nucleus onwards, the size of oligomers/fibrils is thus $(i+n-2)$. Therefore, if n is an integer, $(n-2)$ would correspond to the variable j described

above, and $Corr = -(n-2) = 2-n$. We define nucCorr in an analogous way, but corresponding to the closest integer j, for cases when n is not an integer, as $nucCorr = 2 - round(n)$, using the C function round(n) that returns the integer closest to n. Thus the implementation contains an index correction for the second half fibril as previously:

$$\Delta xp[i] = -\mathrm{k_f}\, xp[i] \tag{4.73}$$

$$\Delta xp[frag] = -\mathrm{k_f}\, xp[i] \tag{4.74}$$

$$\Delta xp[i - frag + nucCorr] = -\mathrm{k_f}\, xp[i] \tag{4.75}$$

We then consider the mass balance of this fragmentation implementation by looking at the monomeric equivalent loss and gain:

$$Loss\, of\, Substrate = Gain\, of\, Products \tag{4.76}$$

$$\mathrm{k_f}\, xp[i]\, (i+n-2) = \mathrm{k_f}\, xp[i]\, (frag+n-2) + \mathrm{k_f}\, xp[i]\, (i - frag + nucCorr + n - 2) \tag{4.77}$$

$$i + n - 2 = i + nucCorr = i + 2 - round(n) \tag{4.78}$$

$$n = round(n) \tag{4.79}$$

We can see that the mass is only balanced when n=round(n), i.e. when n is an integer, as expected. Therefore we introduce a concentration correction factor CF:

$$\Delta xp[i] = -\mathrm{k_f}\, xp[i] \tag{4.80}$$

$$\Delta xp[frag] = -\mathrm{k_f}\, xp[i] \tag{4.81}$$

$$\Delta xp[i - frag + nucCorr] = -\mathrm{k_f}\, xp[i]\, CF \tag{4.82}$$

By considering the monomeric equivalent mass balance again, we can isolate this correction factor:

$$Loss\,of\,Substrate = Gain\,of\,Products \tag{4.83}$$

$$\mathrm{k_f}\; xp[i]\;(i+n-2) = \mathrm{k_f}\; xp[i]\;(frag+n-2) + \mathrm{k_f}\; xp[i]\; CF\;(i-frag+nucCorr+n-2) \tag{4.84}$$

$$i+n-2 = (frag+n-2) + CF\;(i-frag+nucCorr+n-2) \tag{4.85}$$

$$CF = \frac{i+n-2-(frag+n-2}{i-frag+2-round(n)+n-2} = \frac{i-frag}{i-frag+n-round(n)} \tag{4.86}$$

Finally, we define nucRem as $nucRem = n - round(n)$ and implement the reaction as:

$$\Delta xp[i] = -\mathrm{k_f}\; xp[i] \tag{4.87}$$

$$\Delta xp[frag] = -\mathrm{k_f}\; xp[i] \tag{4.88}$$

$$\Delta xp[i-frag+nucCorr] = -\mathrm{k_f}\; xp[i]\;\frac{i-frag}{i-frag+nucRem} \tag{4.89}$$

In summary, in this implementation, the index correction nucCorr places the second half fibril in the array location corresponding to the closest integer size value, and the correction factor adjusts the oligomeric concentration to compensate for the difference between the rounded value of n and the actual decimal value of n, in a way that the overall monomeric concentration (mass balance) is kept constant.

Inhibitor models - fragmentation of inhibitor-bound species In the end-binding model tested in the last generation (model ID 142), k'4 was implemented as depicted in Figure 4.3, generating a protein-only fibril and an inhibitor-bound fibril, since this is the easiest way to keep mass balance without the need of any approximations.

Surface binding of inhibitor

We wanted to compare a surface binding model of inhibitor with an end binding model, for which we developed an approximation described below (represented as step d'2) in Figure 4.3).

If we propose a single binding site per protein unit, this can be modelled as a series of successive binding steps of the form:

$$\mathrm{P}_i + \mathrm{I} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{P}_i\mathrm{I}_1 \tag{4.90}$$

$$\mathrm{P}_i\mathrm{I}_1 + \mathrm{I} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{P}_i\mathrm{I}_2 \tag{4.91}$$

$$\vdots \tag{4.92}$$

$$\mathrm{P}_i\mathrm{I}_{n-1} + \mathrm{I} \underset{k_{-n}}{\overset{k_n}{\rightleftharpoons}} \mathrm{P}_i\mathrm{I}_n \tag{4.93}$$

Focusing on the forward binding reactions, If we consider a Langmuir-like process with the elementary rate depending on the number of free sites, we can express the rate (r_f) as:

$$r_f = k_+ \sum_{n=0}^{i} (i-n)[P_iI_n][I] \tag{4.94}$$

where i represents the length of the fibril, n the number of sites occupied by inhibitor and the (i-n) inside the sum turns the concentration of fibril-inhibitor species P_iI_n into a concentration of free sites, which gives the dependency of the rate on free sites.

At this point, we need to consider the computational aspects of implementing this mechanism as is. For each oligomer/fibril of length i, we would need to keep track of the concentration of each P_iI_n over time. This would require an additional $i(i-2)$ calculations and memory locations with respect to the inhibitor end binding model (where two species of P_i interconvert, free and bound to one inhibitor molecule). The value of this fully explicit calculation would lie in obtaining distributions of free and bound states, therefore allowing to verify the Langmuir model assumptions and propose alternatives. However, large species are invisible in NMR measurements; in our case, we are observing the monomer and dimer species. Therefore, given that it is not possible to compare the distributions of free and bound fibrils with data, we decided to reduce the computational complexity and consider two species:

free fibril ($P_i free$) and inhibitor bound ($P_i bound$). Rearranging the previous dI/dt equation, and with the approximation of constant $P_i I_n$, we obtain the forward rate:

$$r_f = k_+[P_i free]\frac{1}{[P_i free]}\sum_{n=0}^{i-1}(i-n)[P_i bound][I] = \tag{4.95}$$

$$= k_+[P_i free]\overline{i-n}[I] \tag{4.96}$$

where $\overline{i-n}$ is the average of free sites in a fibril with length i.

In terms of $\overline{i-n}$, this would take a value between 0 and 1. We could define it as a free parameter to optimise for each value of i, or as a single parameter that is function of i, if we consider it could increase with i. Alternatively, we could approximate it to 1 if we assume all sites are always saturated, for example if there is cooperativity and the microscopic rate limiting step is the first binding event, followed by much faster events. We could also incorporate it into a rate k'_+, if we assume only a weak dependency on i.

Again, as outlined above, the consequences of these assumptions would lie in different distributions of free and bound fibrils, which we can't probe experimentally. For simplicity, then, we have approximated it to 1, compatible with saturation of all the sites, which translates to a rate forward of:

$$r_f = k_+[P_i free][I] \tag{4.97}$$

and a rate of change of inhibitor, in the forward binding reaction, as:

$$\frac{dI}{dt} = -ik_+[P_i free][I] \tag{4.98}$$

This approximation is similar to the assumption in the fibril fragmentation reaction that all species break in half (or as close to as possible, if the length is odd). In both cases, certain aspects of the physical complexity that we cannot experimentally probe against have been removed from the mathematical treatment accordingly.

In summary, the approximation implemented for surface binding of inhibitor is as follows, with a single chemical equilibria for each P_i species:

$$\mathrm{P}_i\mathrm{free} + \mathrm{iI} \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{P}_i\mathrm{bound} \tag{4.99}$$

The rates of change of each species are as follows:

$$\frac{\mathrm{d[P_i free]}}{\mathrm{dt}} = -\mathrm{k_+[P_i free][I]} + \mathrm{k_-[P_i bound]} \tag{4.100}$$

$$\frac{\mathrm{d[P_i bound]}}{\mathrm{dt}} = +\mathrm{k_+[P_i free][I]} - \mathrm{k_-[P_i bound]} \tag{4.101}$$

$$\frac{\mathrm{d[I]}}{\mathrm{dt}} = -\mathrm{ik_+[P_i free][I]} + \mathrm{ik_-[P_i bound]} \tag{4.102}$$

Most importantly, the approximation implemented captures the higher oscillations in free inhibitor concentration that will result from a surface binding model with respect to an end binding one (in the i factor in $\frac{dI}{dt}$). These will affect the kinetics in a measurable way, given that the inhibitor participates in other interactions with protein species whose concentrations we can monitor.

Finally, the same approximation was applied for the surface binding of inhibitor to either dimer (step b'2) or nucleus (step c'2).

Mass balance corrections for elongation reactions in models including surface binding of inhibitor

These are small mass balance corrections that are required in order to fit some combinations of chemical steps as part of the same model and keep mass balance constant. This is due to implementation decisions to simplify the calculations computationally, which are described above, as well as the adjustments required.

If i is the number of monomeric subunits in an oligomer or fibril, and N is the maximum length of fibrils included in the system, we are keeping track of the concentration of species in two arrays of size N, "xp" for protein-only species and "xpC" for protein-inhibitor complexes. For oligomers (species larger than nucleus), the location xp[i] contains a protein oligomeric species of size i. The content of xpC is also a protein oligomeric species of size i, but the inhibitor content depends on the mode of inhibitor binding. For end binding, xpC[i] contains a fibril of

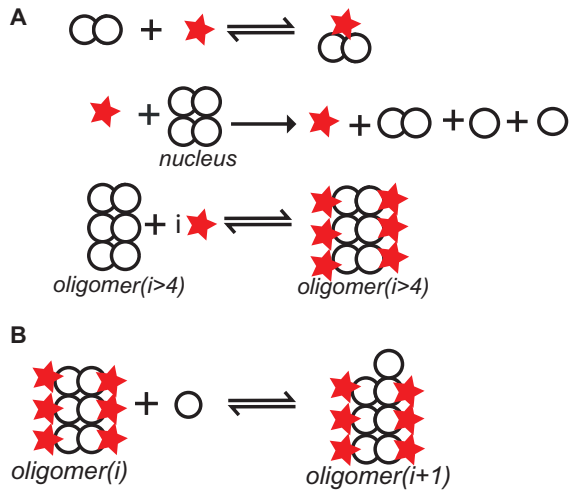


Figure 4.5: Example model requiring mass balance corrections.A. Model b'1 d'2 f'1, including surface binding of inhibitor to oligomers (P_iI_i). B. Elongation by monomer step that generates $P_iI_{(i-1)}$.

length i and two inhibitor molecules (one at each end), so P_iI_2. In models with surface inhibitor binding, xpC[i] keeps track of the concentration of a fibrils of length i with i molecules of inhibitor bound, which we can refer to as P_iN_i, as was described in the previous section.

Therefore, mass imbalances can be generated when reactions that involve protein-inhibitor complexes generate products that don't have the same protein-inhibitor composition. In particular, this was encountered for models including surface binding of inhibitor, since that step was selected in the modelling process (if end-binding had been selected, different corrections would have been needed, following the same logic). This is illustrated with an example by looking at the model I24 with inhibitor binding to dimer, nucleus disruption and inhibitor surface-binding (Figure 4.5). In this model, some protein-inhibitor complexes (dimer+1 inhibitor, and surface-bound oligomers and fibrils) are created. Reactions i' through to k' (Figure 4.3), which have surface-bound oligomers as reactants, can be added to this model and sometimes cause mass imbalances, as in the case considered below in detail where elongation by monomer (Figure 4.5B) is added.

This corresponds to this chemical equilibria, represented also as array locations

(note that monomer is stored in xp[0]):

$$\mathrm{P_8I_8} + \mathrm{P_1} \quad \underset{k_{-el}}{\overset{k_{el}}{\rightleftharpoons}} \quad \mathrm{P_9I_8} \tag{4.103}$$

$$\mathrm{xpC[8\,]} + \mathrm{xp[0\,]} \underset{k_{-el}}{\overset{k_{el}}{\rightleftharpoons}} \mathrm{xpC[9\,]} \tag{4.104}$$

The product of the reaction is a species of the form $P_iI_{(i-1)}$; however, when it participates in the surface-binding reaction between oligomers and inhibitor, it will be interpreted as P_iI_i and this will result in the release of i (9 in this case) molecules of inhibitor in the dissociation step.

The imbalance can be seen when considering the change of concentration of the species in the forward reaction:

$$\Delta\,[P_8I_8] = \Delta\,[P_1] = -\Delta\,[P_9I_8] = -\mathrm{k_{el}}\,[P_8I_8][P_1] + \mathrm{k_{-el}}\,[P_9I_8] \tag{4.105}$$

The mass balance is correct for protein subunits:

$$Loss\,of\;Substrate = Gain\,of\;Products \tag{4.106}$$

$$\Delta\,Conc \cdot i + \Delta\,Conc\;\cdot 1 = \Delta\,Conc \cdot (i+1) \tag{4.107}$$

$$(i+1) = (i+1) \tag{4.108}$$

But incorrect for inhibitor molecules:

$$Loss\,of\;Substrate = Gain\,of\;Products \tag{4.109}$$

$$\Delta\,Conc \cdot i + \Delta\,Conc\;\cdot 0 \neq \Delta\,Conc \cdot (i+1) \tag{4.110}$$

$$(i+0) \neq (i+1) \tag{4.111}$$

The exact solution to this problem would be to implement several protein-inhibitor arrays, for example in this case an array with i-1 inhibitor molecules. However we can implement a simple adjustment that allows to keep the computational complexity to a minimum, with only the xp and xpC arrays. The solution, defined generally for a case in which a species P_pI_i is generated in an elongation reaction

(where $p \neq i$), is to remove (in the forward reaction) and add (in the backward reaction) (p-i) equivalents of free inhibitor concentration. For the example above with elongation by free monomer, the system becomes:

$$\frac{\mathrm{d[P_pI_p]}}{\mathrm{dt}} = -\mathrm{k_{el}[P_pI_p][P_1]} \quad +\mathrm{k_{-el}[P_{p+1}I_{p+1}]} \tag{4.112}$$

$$\frac{\mathrm{d[P_i]}}{\mathrm{dt}} = -\mathrm{k_{el}[P_pI_p][P_1]} \quad +\mathrm{k_{-el}[P_{p+1}I_{p+1}]} \tag{4.113}$$

$$\frac{\mathrm{d[P_{p+1}I_{p+1}]}}{\mathrm{dt}} = +\mathrm{k_{el}[P_pI_p][P_1]} \quad -\mathrm{k_{-el}[P_{p+1}I_{p+1}]} \tag{4.114}$$

$$\frac{\mathrm{d[I]}}{\mathrm{dt}} = -\mathrm{ak_{el}[P_pI_p][P_1]}+\mathrm{ak_{-el}[P_{p+1}I_{p+1}]} \tag{4.115}$$

with the last equation corresponds to the adjustment by a equivalents of inhibitor concentration, 1 in this case. Note that this adjustment does not alter the rate expressions, and the reaction remains order zero with respect to free inhibitor.

This logic was followed to check and maintain mass balance for all cases of elongation reactions involving surface-bound oligomers, and Table 4.1 summarises the corrections implemented. Note that in cases where the nucleus is allowed to elongate, it requires a specific correction if it is bound to only one molecule of inhibitor. Also, in some cases there are no corrections applied, when all species involved are of the form P_iI_i. Whenever there was a need for such an adjustment, if the alternative exact model was possible (for example, elongation by surface bound dimer, or the surface-bound nucleus also elongating), it was explicitly tested, which meant deviating slightly from the original step selection process in which mutually exclusive (or redundant) steps, such as the binding of dimer to one (b'1) or two (b'2 - surface binding) inhibitor molecules, would not be tested in parallel (within "sibling" models) again.

A similar logic was used to keep mass balance in cases with end-binding of inhibitor to fibrils and elongation of bound-fibrils by bound monomer or dimer. For simplicity we have assumed a single binding event at one of the fibril ends. This would cause mass balance issues if the single-bound fibril elongates by a monomer-inhibitor or dimer-inhibitor complex, since there would be a second molecule of

inhibitor bound to the fibril which would not be accounted for in the array. This has been solved in an analogous way to the elongation of surface-bound fibrils, by releasing one equivalent of inhibitor concentration in the forward reaction and removing it in the backwards reaction - again, the rate of the backwards reaction is order zero with respect to free inhibitor.

Table 4.1: Mass balance corrections for elongating steps involving inhibitor surface-bound oligomers. The units refer to equivalents of inhibitor concentration that are removed and added to the pool of free inhibitor in the forward and backward reaction, respectively.

Elongation by	Oligomeric correction	Nucleus correction	
	1	3 if 	0 if 
	0		
	2		
	1		
	0		

4.4 Results and Discussion

4.4.1 Mechanism of R3 aggregation

Concentration dependency of the monomer initial decay rates

Figure 4.6 shows the kinetic traces for two protein-only datasets, referred to as first and second dataset, and an analysis of the order of reactions involving monomer at the beginning of the aggregation process. In the first dataset, there was a substantial amount of dimer present at t=0h in all concentrations, building-up and decaying in the 100 and 50 µM samples, while already decaying in the 400 µM sample. In contrast, in the second dataset the initial concentration of dimer was small throughout the concentration range.

We expect many elementary reactions to occur during the full timecourse of the aggregation reaction and we are not expecting a simple overall mechanism or rate law.

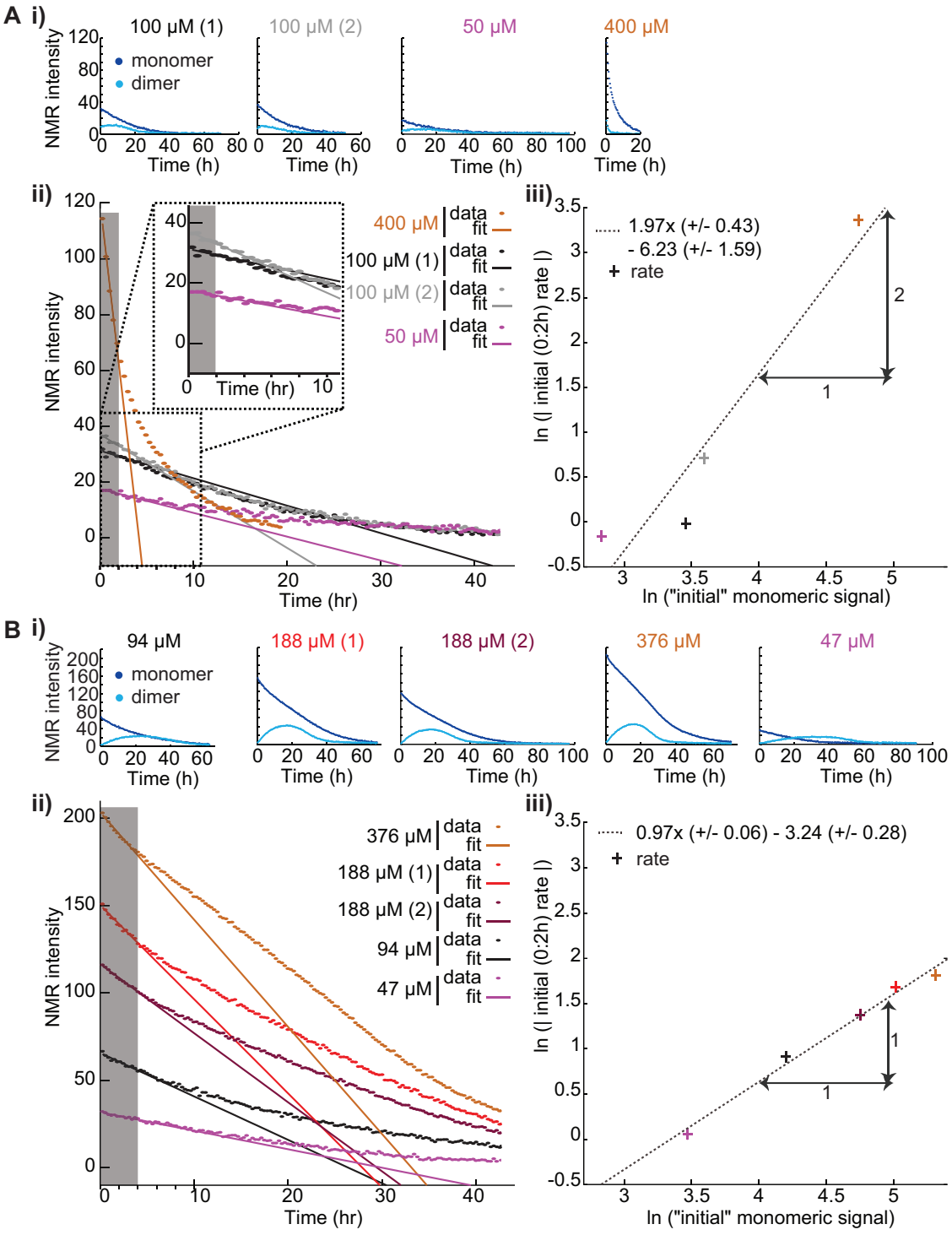


Figure 4.6: Initial rates analysis of protein-only data. Analysis of the order of reactions involving monomer at initial times for the first (A) and second (B) protein only datasets. (i) Complete dataset showing monomer (dark blue) and dimer (light blue) traces. (ii) Initial rate fits, using the first 2 hours and 4 hours of data for the first dataset and the second dataset, respectively, as indicated by the shaded area. (iii) Plot of ln(rate) versus ln(signal), in which the fitted slope value indicates the order of the reaction.

However, we thought we could obtain insights on the early events in the aggregation pathway by using the initial rates method. This would not be possible with other methods such as half lives or integral methods that would require looking at a longer time period. The initial rates of monomer decay in the first dataset followed a second order dependency with respect to monomer signal (taken as a direct proxy for concentration). This indicates that at early stages there is a reaction involving two monomers. In contrast, the initial rates in the second dataset show a first order dependency on monomer signal, indicating a unimolecular reaction of the monomer. This had implications for the design of the pool of steps inputted into the modelling algorithm, as seen below in the respective sections.

Modelling results

First dataset: The aggregation of R3 involves dimerisation, nucleation, elongation and fragmentation events The application of the modelling algorithm to the first dataset of R3-only data yielded the results schematically shown in Figure 4.7 and reported fully in Table 4.2 (with model IDs used in text). Briefly, it can be seen how the systematic search explored a number of models of complexity ranging between 4 parameters (linear polymerisation, χ^2=48.6, ID 1 in Table 4.2, Figure 4.7 D i) and 19 (ID 24 Table 4.2) and arrived at a final best model with 17 parameters (9 rate constants and 4 pairs of sample-specific *srac* and *frac* (as part of the polishing procedure - see Methods), χ^2=1.22, ID 25 in Table 4.2) which maximised the goodness of fit. Figure 4.7 C shows that in this process, there were four generations of model building (with polishing turned on at generation 4), two pruning stages and specific nucleation tests in parallel to the first pruning (not depicted in full - discussed below in a separate section). The final best model, the winner of the 1st pruning stage (Figure 4.7 D ii) can be described as a tetrameric nucleation model with fragmentation as a secondary process. The formation of the nucleus involves a dimerisation reaction and a nucleation reaction involving one dimer and two monomers, fitted as a single-step steady-state approximation (see Methods).

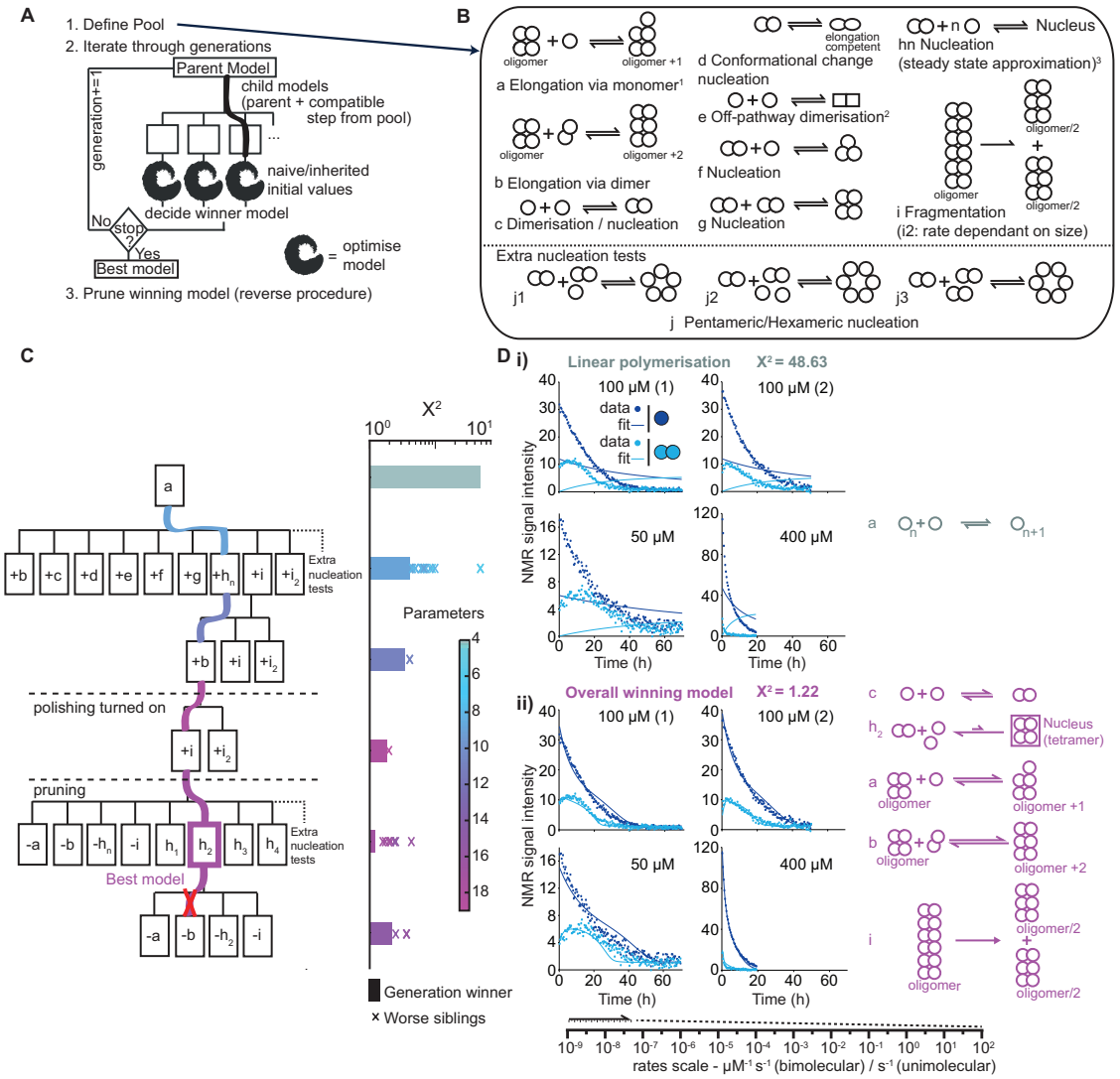


Figure 4.7: Mechanistic model for the aggregation of R3. A. Overview of the algorithm employed to search the model space (Chapter 3). B. Pool of steps (see Figure 4.2). Notes: 1: If a is the only reaction, this corresponds to linear polymerisation. 2: The resulting dimer doesn't participate in any further reactions. 3: n can either be free to optimise (hn) or fixed (e.g. h2). If n=1, this step is equivalent to f. C. Resulting tree representation of the modelling process, showing the number of models tested per generation and the evolution of χ^2 and number of parameters. The reaction codes indicate the steps added to each model on top of the contents from the parent model and correspond to the schematics in B. D. Fits and schematic representation for the initial linear polymerisation model (i) and the final best model (ii), with ID 1 and 25 in Table 4.2. Dark blue: monomer, cyan: dimer. The length of the chemical equilibria arrows corresponds to the order of magnitude of the rate constant (discretised to the closest 0.5). Note that the rate constants of step h2 are composite (see the steady-state approximation in Methods) and therefore are neither unimolecular or bimolecular. The monomer fraction (frac), and concentration/signal ratio (*srac*) parameters and the equivalent concentration ($C_{eq} = \frac{srac}{srac_{ref}} C$), based on the reference srac ($srac_{ref}$) and sample concentration (C), taking the 100 µM (1) sample as the reference, are: 100 µM (1): frac = 0.82, srac = 0.43, C_{eq} = 100 µM; 100 µM (2): frac = 0.91, srac = 0.44, C_{eq} = 101.5 µM; 400 µM: frac = 0.93, srac = 0.35, C_{eq} = 322 µM, 50 µM: frac = 0.78, srac = 0.39, C_{eq} = 45 µM.

Table 4.2: Modelling results for the R3-only first dataset. Best result for each model tested in the systematic modelling process of the first dataset of R3. The winning models of each stage are shown in bold and the overall best model is shown in bold italicised. 2-ext: extra nucleation tests added at the generation 2 level. P1: First pruning. P2: Second pruning. NT: Nucleation tests run in parallel to the first pruning. p: parameters, σ: perturbation in basin-hopping like optimisation. Asterisks refer to parameter initialisation modes in which polishing was already turned on, as explained in the Methods section. The initial (unpolished optimisation, if polishing is turned on) and fitted (for the final model result reported in the table) n values are indicated when the nucleation step is hn.

ID	Gen.	Model	p	χ^2	Parent?	Polish?	Guess	σ
1	**1**	**a**	**4**	**48.63**	**Yes**	**No**	**naive**	**0.01**
2	2	a,b	6	6.79	No	No	inherited	0.05
3	2	a,c	6	6.49	No	No	inherited	0.1
4	2	a,c,hn	9	4.31	No	No	inherited, n=1e-09(->n=1.4)	0.01
5	**2**	**a,c,hn**	**9**	**4.09**	**Yes**	**No**	**inherited, n=1(->n=1.6)**	**0.05**
6	2	a,c,hn	9	4.10	No	No	inherited, n=2(->n=1.7)	0.05
7	2	a,c,hn	9	6.69	No	No	inherited, n=3(->n=2.4)	0.1
8	2	a,i	5	48.63	No	No	inherited	0.01
9	2	a,e	6	9.03	No	No	naive	0.01
10	2	a,c,g	8	4.94	No	No	inherited	0.05
11	2	a,c,h2	8	4.99	No	No	naive	0.2
12	2	a,c,f	8	6.11	No	No	naive	0.01
13	2	a,c,d	8	7.22	No	No	naive	0.1
14	2	a,i2	6	48.63	No	No	naive	0.01
15	2- ext	a,c,h3	8	8.38	No	No	naive	0.2
16	2- ext	a,c,j1	8	4.49	No	No	inherited	0.1
17	2- ext	a,c,h4	8	9.91	No	No	inherited	0.2
18	2- ext	a,c,j2	8	6.13	No	No	naive	0.1
19	2- ext	a,c,j3	8	4.46	No	No	naive	0.1
20	3	a,c,hn,i	10	4.08	No	No	inherited (n=1.6->n=1.6)	0.01
21	**3**	**a,b,c,hn**	**11**	**3.47**	**Yes**	**No**	**naive, n=1(->n=1.7)**	**0.05**
22	3	a,c,hn,i2	11	4.07	No	No	inherited (n=1.6->n=1.6)	0.01
23	**4**	**a,b,c,hn,i**	**18**	**1.62**	**Yes**	**Yes**	**naive, n=2 (->n=3.4)**	**0.1**
24	4	a,b,c,hn,i2	19	1.95	No	Yes	naive, n=1 (->n=1.5)	0.1
25	***P 1***	***a,b,c,h2,i***	***17***	***1.22***	***Yes***	***Yes***	***inherited***	***0.2***
26	P 1	b,c,hn,i	16	4.31	No	Yes	naive, n=3 (->n=2.8)	0.1
27	P 1	a,c,hn,i	16	2.22	No	Yes	naive, n=2 (->n=1.7)	0.01
28	P 1	a,b,c,hn	17	2.06	No	Yes	naive, n=1 (->n=1.5)	0.01
29	P 1	a,b,c,i	15	2.49	No	Yes	naive	0.05
30	NT	a,b,c,f,i	17	2.33	No	Yes	inherited	0.2
31	NT	a,b,c,d,i	17	2.29	No	Yes	naive*	0.05
32	NT	a,b,c,g,i	17	1.68	No	Yes	naive*	0.05
33	NT	a,b,c,h3,i	17	1.56	No	Yes	inherited* nuc*	0.05
34	NT	a,b,c,h4,i	17	1.73	No	Yes	inherited*	0.05
35	NT	a,b,c,i,j1	17	1.80	No	Yes	inherited nuc*	0.05
36	NT	a,b,c,i,j3	17	2.00	No	Yes	inherited* nuc	0.1
37	NT	a,b,c,i,j2	17	1.32	No	Yes	inherited nuc*	0.1
38	P 2	b,c,h2,i	15	3.66	No	Yes	inherited	0.1
39	**P 2**	**a,c,h2,i**	**15**	**2.12**	**Yes**	**Yes**	**inherited**	**0.1**
40	P 2	a,b,c,h2	16	3.62	No	Yes	naive	0.05
41	P 2	a,b,c,i	15	2.49	No	Yes	naive	0.05

The nucleus and oligomers/fibrils elongate by addition of monomer or dimer at distinct rates that are independent of their size. Finally, oligomers/fibrils (species made of more than 10 subunits) fragment irreversibly.

Evolution of the goodness of fit through the stages leading to the final best model Figure 4.8 shows in more detail how the application of the algorithm progressed through stages towards the final result, with the goodness of fit improving at each stage. In the first three generations, there were systematic residual trends for both the monomer and the dimer traces. There were very prominent systematic deviations in generation 1, which were reduced and changed in generation 2, when the nucleation reaction (dimer+n monomers, n=1.5) was incorporated into the best model. The same trend was maintained in generation 3, but with smaller dimer deviations - this is consistent with the new step incorporated involving the dimer species (elongation). In generation 4, the monomer residuals evened out but the dimer deviations stayed similar, and finally in the winning model, resulting from the first pruning generation, the dimer residuals also evened out. The steps were incorporated in the following order: elongation via monomer (equivalent to linear polymerisation) in generation 1, single-step dimer+n monomers nucleation (n=1.5) in generation 2, elongation via dimer in generation 3 and fragmentation in generation 4. In the 1st pruning stage, the winning model had one less parameter than its parent, since the parameter n of the steady-state nucleation approximation was fixed to 2. In contrast, its parent model, the winner of generation 4, had been optimised to a final n of 3.3 (ID 23 in Table 4.2). This reduction of complexity was accepted since the χ^2 was reduced, and as a consequence the nucleation step was explicitly defined as the formation of a tetramer.

Justification of the complexity of the best model based on the results of the second pruning stage Figure 4.9 shows the worsening of the fits and reappearance of systematic residual deviations in all the models tested in the 2nd pruning stage, compared to the best model (winner of the 1st pruning stage). The χ^2s of all pruned models were higher than that of the winner (Figure 4.9 i) by

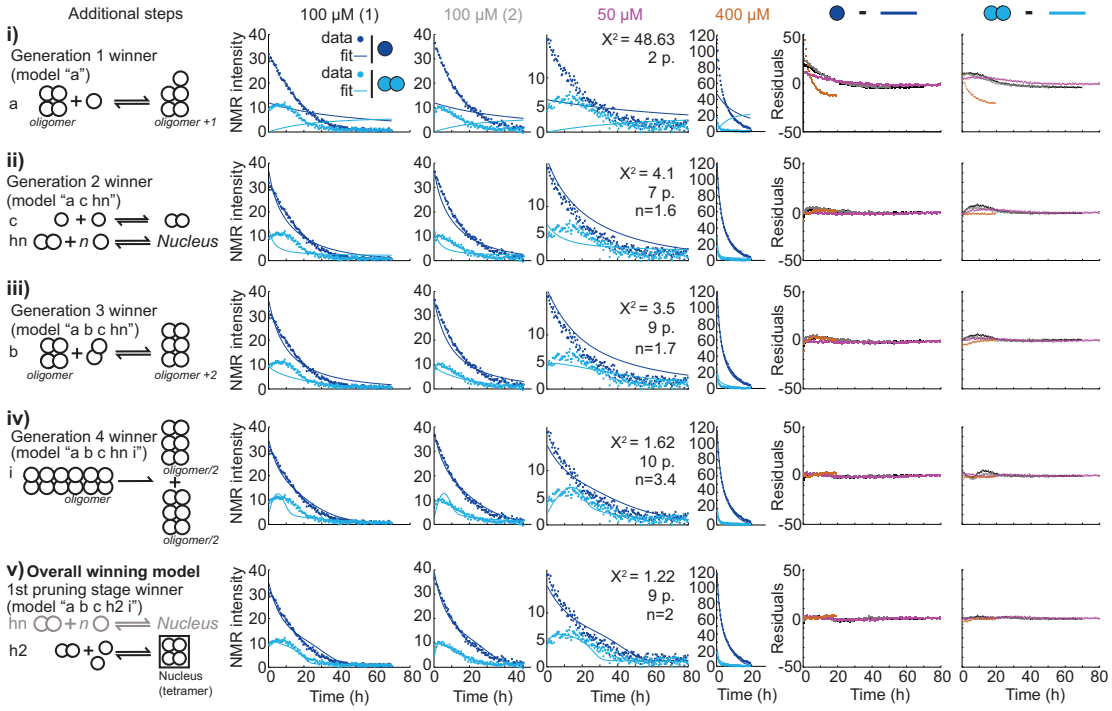


Figure 4.8: Evolution of goodness of fit through the modelling process. Each row shows the winner of the generation indicated in the left, in terms of its global goodness of fit (χ^2) and residuals (data - fit) for the monomer and dimer traces separately. The number of parameters reflects only those that correspond to a chemical reaction, excluding the initial monomeric fraction and signal to concentration ratio. The steps that have been added to the model at each generation are schematically shown on the left. The results shown in each row correspond to ID 1, 5, 21, 23 and 25, respectively, in Table 4.2.

a larger amount than the 0.1 cut-off defined in the Methods. Also, systematic residuals reappeared for both the monomer and dimer traces in most of the pruned models (with the exception of removing monomer elongation, which mainly affected the monomer fits). None of these models were accepted and therefore the winning model of the first pruning stage, model "a b c h2 i" (ID 25 in Table 4.2), became the final best model.

It seems at first sight that there is little correlation between the importance of each step in terms of how early it was added in the evolution process (meaning it had been the best in the generation compared to other steps in sibling models) and how much impact on the χ^2 its removal in the pruning stage had. The goodness of fit as expressed in χ^2 (the quantitative measure used in the modelling process) was more resilient to the removal of the elongation by dimer ($\chi^2 = 2.13$, Figure 4.9 ii)

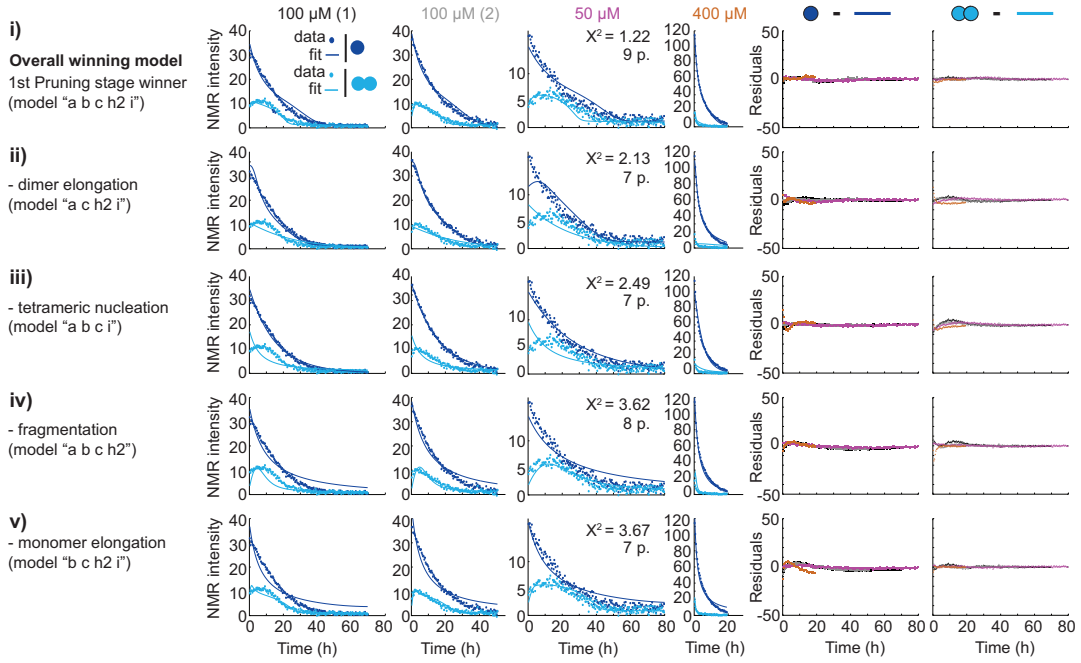


Figure 4.9: Results of the second pruning stage of the first R3-only dataset. Top row: winning model in the first pruning (reference model for comparison). From the second row, each row shows one of the pruned models and which step was removed from the reference model. Each row shows the global goodness of fit (χ^2), the fits for each sample and the residuals (data - fit) for the monomer and dimer traces separately. The number of parameters reflects only those that correspond to a chemical reaction, excluding the initial monomeric fraction and signal to concentration ratio. The results shown in each row correspond to ID 25, 39, 41, 40 and 38, respectively, in Table 4.2.

and tetrameric nucleation ($\chi^2 = 2.49$, Figure 4.9 iii) steps than of the other steps. However, the features of build-up and decline of the dimer traces were completely lost in all concentrations in the case of removing tetrameric nucleation and in the 50 µM sample if removing elongation by dimer. Meanwhile, larger increases in χ^2 in the models without fragmentation ($\chi^2 = 3.62$, Figure 4.9 iv) or monomer elongation ($\chi^2 = 3.67$, 4.9) resulted from systematic deviations mainly in the monomer traces, yet those models recapitulated the features of the dimer traces. Overall, and most importantly for the purpose of avoiding overfitting, it was clear that there were significant losses of goodness of fit in all pruned models. This supported fitting the data with a final best model with 9 chemical rate constants (and 8 auxiliary parameters that define initial monomer fraction and signal to concentration ratio).

Challenges in the optimisation of the nucleus size The results of the 1st pruning stage represent a (rare) instance in which the optimiser has not found the global minimum for a model, since nucleation step h2, with a tetrameric nucleus (in the final best model, see Figure 4.8 v) is an explicit implementation (n=2) of the more general hn (dimer and n monomers make nucleus) steady-state approximation nucleation step that was part of the winner model of generation 4 (see Figure 4.8 v). Therefore the global minimum with n=2 and χ^2 of 1.22 should have been found for the winner model of generation 4. Instead, the best fit at that generation had n=3.4 and a χ^2 of 1.62. We attribute this specific instance of suboptimal optimiser performance to the specific nature of the n parameter, which introduces higher powers in the concentration dependencies of the derivatives, compared with other chemical steps, which leads to numerical instabilities.

Due to this finding, we wanted to verify that the h2 nucleation step was the best fit for the data, given the importance of the nucleation reaction within the aggregation mechanism and the discrepancies in the literature. For this we performed additional nucleation tests (grid search), presented below, with which we can confidently conclude that a tetrameric nucleus is essential for the best fitting model presented above and remains the best fitting nucleation mechanism.

Comparison of nucleation hypotheses at two stages of the modelling

We performed an exhaustive quantitative and qualitative comparison of a wide range of nucleation reactions. This was done at two levels of model complexity: relatively simple models, equivalent in complexity to generation 2 (with linear polymerisation as the parent model); and more complex models, equivalent to the 1st pruning stage (with the winner of generation 4 as the parent model). Given that generation 2 had already tested many nucleation mechanisms as part of the original routine, the extra tests involved defining explicit implementations of the hn general case as well as adding larger nucleus sizes. In the more complex stage, we took the best model of generation 4 and systematically switched the h2 nucleation step for all other available steps, including as well explicit formulations of hn up to n=4 and

larger nucleus sizes (j1,j2,j3). Two additional parameter initialisation inherited mode variants were used, which specifically cleared the information regarding nucleation rates (see Methods); in some cases, these setting variants led to the best results. The results of these tests are recorded in Table 4.2 (2/2-ext and NT categories, respectively, for lower and higher complexity of models compared).

Figure 4.10 shows the goodness of fit depending on the nucleation reaction at the two complexity levels. In the generation 2 equivalent stage, the model of reference was the winner of that generation (ID 5 in Table 4.2), while at the later stage, the reference models were the winner of generation 4 with step hn (ID 23, $\chi^2 = 1.62$) and the final best model (ID25, $\chi^2 = 1.22$). Importantly, the model with h2, the best model found with the original algorithm, remained the overall best model. There is little correlation between the ranking at the two stages of the modelling, however the best step (h2) in the later generation corresponds to an explicit implementation of the more general hn step that was selected in the winner model of generation 2. In generation 2, the parameter n was fitted to 1.5, indicating a mixture of trimer and tetramer nuclei, while at generation 4 it was fitted to 3.4. Based on their quantitative goodness of fit at the later generation, and the qualitative visual inspection of systematic residuals (for the dimer trace in particular), we can group the nucleation steps in three categories: good, close and poorly fitting (Figure 4.10 B).

Overall classification of nucleation steps based on goodness of fit and qualitative assesment The good steps include the tetrameric nucleation step of the best model (h2), the dimer+n monomers hn step that was part of all models since being incorporated in generation 2 (optimised to n=3.3 in generation 4), as well as similar mechanisms (see Figure 4.10 B i). All "good" steps involve a dimer and more than one monomer (h3, with three monomers, and h4, with four, are explicit formulations of the general dimer+n monomers steady-state approximation), except for step j2 which involves two dimers and two monomers. Models in the "close" category (Figure 4.10 B ii) show an initial build-up of dimer,

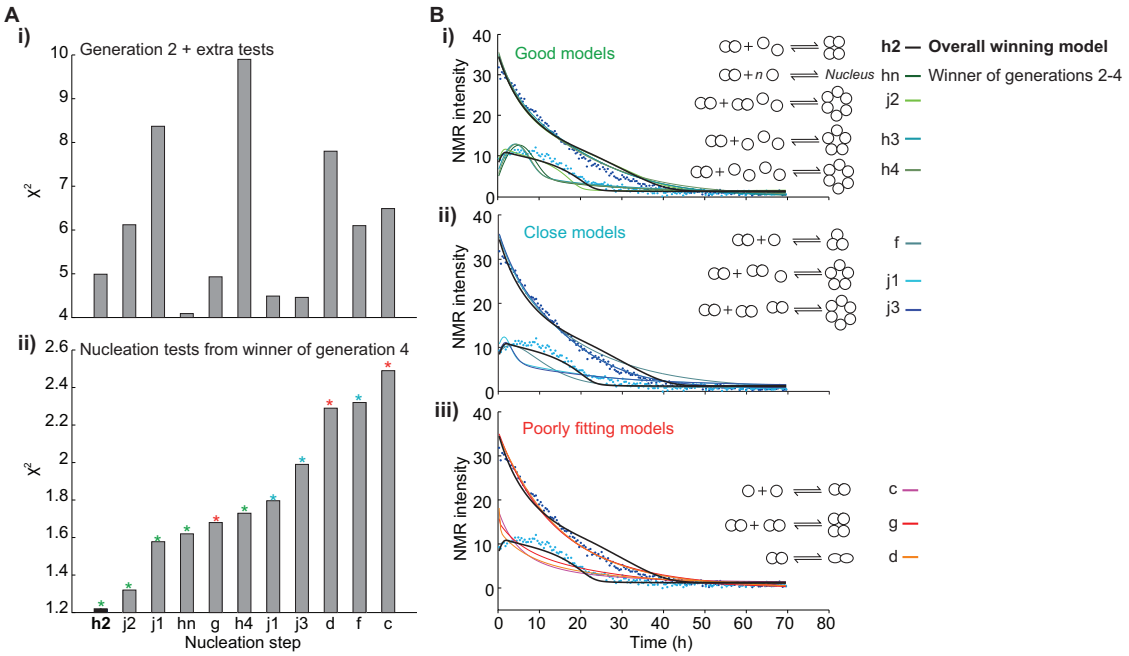


Figure 4.10: Comparison of nucleation steps at an early and late stage of the modelling process. A. χ^2 vs nucleation step type. All other steps in the mechanism are common, even if their rates have been optimised to different values. (i) Generation 2 and extended nucleation tests. The parameter n of step *hn* was fitted to 1.5). (ii) Nucleation tests stage equivalent to the 1st pruning stage. The parameter n of step *hn* was fitted to 3.4. The nucleation steps are ordered from left to right based on χ^2 (lowest to highest) in the more complex stage (ii). B. The types of nucleation steps are grouped into good (i), close (ii) and poorly fitting (iii) based on a combination of goodness of fit at the nucleation tests stage (A ii) and visual inspection, paying particular attention to how closely they match the build-up and decline of the dimer trace (i.e. whether there are large systematic residuals). The fits shown correspond to the 100 µM (1) sample in Figure 4.8 and others. The final best model (h2) is shown in all categories for reference. The model IDs (Table 4.2) of the models shown are: h2 - 25, hn - 23, j2 - 37, h3 - 33, h4 - 34, f - 30, j1 - 35, j3 - 36, c - 29, g - 32, d - 31.

but not sustained enough; these are mechanisms with a trimeric nucleus (f) or a nucleation reaction with at least two dimers, and either a third dimer (j3) or a monomer (j1). The steps j1, j2 and j3 are multistep mechanisms that were fitted to steady-state approximations with single composite rates, in an analogous way to the hn family of steps. Finally, the poorly fitting steps (see Figure 4.10 C) are those with either a dimeric nucleus, with (d) or without (c) a conformational change that makes it elongation-competent, or a tetrameric nucleus formed by two dimers (g). It is worth noting that the model with tetrameric nucleation by

assembly of two dimers (g) is in the poor fitting category despite having a χ^2 of 1.68, in the range of the good models (1.22-1.73); however, it did not recapitulate the build-up and decline of the dimer in either the 50 µM (not shown) or 100 µM traces. Finally, as expected due to the abundance of dimer in the traces, the model with a dimeric nucleus performed poorly even in the context of a relatively complex mechanism containing all other important steps.

Dimer residuals allow further discrimination between well-fitting models Within the mechanisms of nucleation that gave a good fit, two patterns can be distinguished, as shown in Figure 4.11. Crucially, only the final best model (h2, $\chi^2 = 1.22$) and the second best (j2, $\chi^2 = 1.32$) have small systematic residuals for both monomer and dimer. In contrast, the explicit versions of hn, h3 and h4, unsurprisingly, have a similar pattern to the winner of generation 4 (hn, fitted to n=3.3). The second best model (j2, ID 37 in Table 4.2) has a higher χ^2, above the 0.1 threshold, and worse systematic residuals for the dimer trace in particular. This model is also more complex than the h2 step even if it has the same amount of fitted rate constants, given that one more (three versus two) elementary reaction has been included in the steady-state approximation (see Methods). With the same logic, the remaining models in this well-fitting category also are more complex than the one with h2, on top of having worse fits. The conclusion from these extended tests regarding the nucleation step was that the model with the h2 step (tetrameric nucleation from a dimer and two monomers), which was arrived at following the original algorithm, stands as the overall best model, given it had the lowest χ^2 and minimal complexity amongst the closest fitting alternatives.

Discussion of the first dataset winning model

Following the application of the modelling algorithm to the first dataset, we selected a minimally complex best model. The rates shown in Figure 4.7 indicate that the formation of the tetrameric species is the rate limiting step, with a very slow forward assembly rate and very fast dissociation rate. The nucleus of a reaction is typically defined as the first species in the pathway for which the subsequent

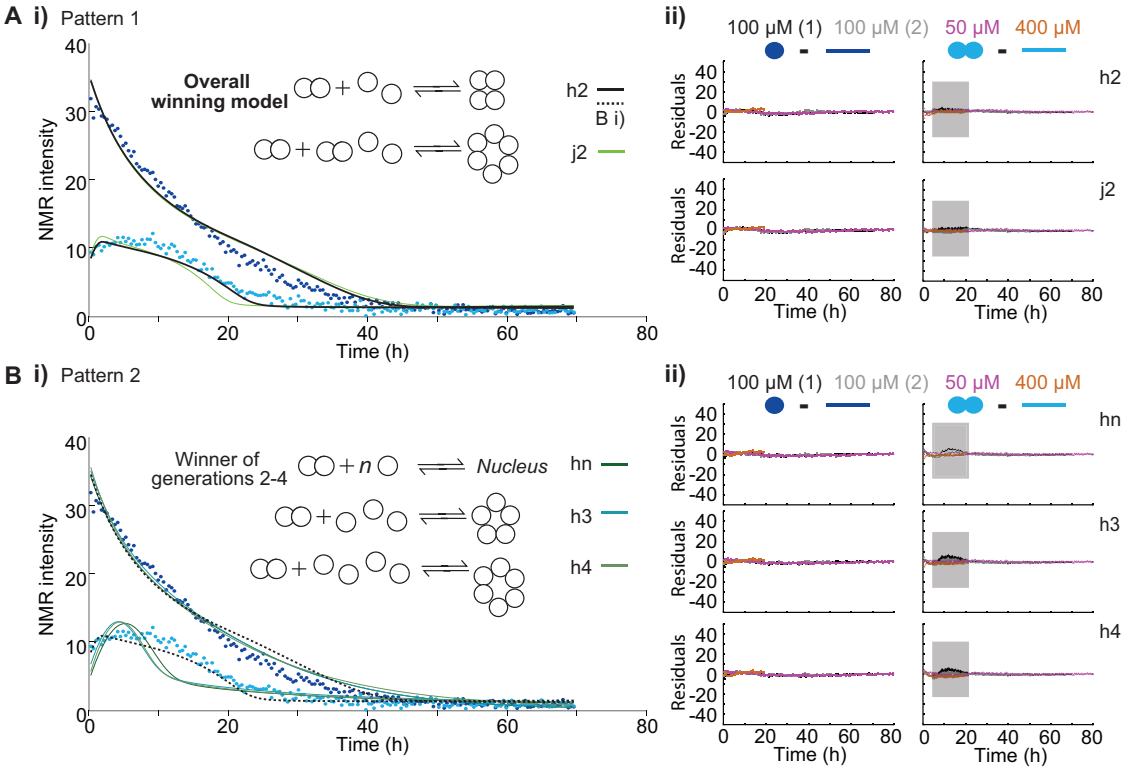


Figure 4.11: Comparison of good nucleation mechanisms. A. The winning best model (h2, ID 25) and the model with mechanism j2, which had the second best fit both in terms of χ^2 and features. (i) Fits of models with h2 and j2 nucleation types. (ii) Monomer and dimer residuals (data - fit) for each model in the first pattern. B. The winner model of generation 4 (hn; n fitted to 3.4) shares similar features with models with explicit implementations (h3 and h4). (i) Fits of the models; the absolute best model (h2) is also shown in dashes for reference. (ii) Monomer and dimer residuals (data - fit) for each model in the second pattern. Model IDs in Table 4.2 are the same as in Figure 4.10

.

reaction proceeds faster than the breakage of the nucleus [38]. In this case, two reactions involve the nucleus in the forward direction, elongation by monomer and by dimer, and the magnitude of the elongation by dimer forward rate is higher than the backwards dissociation rate of the nucleation step. These aspects of the final model are consistent with the definition of the tetramer as the nucleus.

Furthermore, the unfavourable nature of the nucleus assembly and its fast consumption in elongation reactions is reflected, as expected, in the tetramer not being an abundant species in the course of the reaction - the predicted signal for the nucleus does not exceed 0.45 signal units within all the samples, in comparison with 15-110 for monomer and 5-20 for dimer. Given the scarcity of this species, if there

were a unique chemical shift signature for the histidine residues corresponding to the nucleus, it would likely be below detection; in addition, the signal calculated does not consider the additional relaxation of a larger species or potential broadening due to exchange. It is also expected that the diffusion experiment performed in non-aggregating conditions, in which the nature of the dimer was established, would not have reflected any significant amount of nucleus, since that would have led to aggregation. However, it is worth noting that for the 3R construct K12, not only dimers but also traces of higher aggregates (trimers and tetramers) had been detected by native PAGE, in oxidising but not reducing conditions [119]. Therefore it would be interesting to examine in more detail if the R3 nucleus can be detected by NMR, using multidimensional experiments and/or exchange experiments, or other biophysical techniques.

Comparison of full mechanism with initial rates insights In light of this detailed complete mechanism, we can better interpret the early insights from the analysis of the initial rates of monomer decay (Figure 4.6). The only two steps in the mechanism that are second order with respect to monomer are dimerisation and nucleation. These two reactions might contribute differently to the rate dependencies analysed in each sample, since the different samples were "caught" at different stages of the process. For the 400 µM sample we are likely observing the nucleation step since there is already a substantial dimer build up prior to the recording of the first datapoint, and at t=0h it is already declining, getting used up in the nucleation (and elongation) reactions. For the remaining samples, we are more likely observing the dimerisation step, since it is faster than nucleation, and the dimer curve is still building up, or a combination of the two.

The analysis of the initial rates was limited in that the monomer concentration is not properly isolated, since the concentration of dimer, both a reactant and a product of reactions involving monomer, is neither constant nor zero. A reflection of the limitations is the fact that the datapoint further away from the fit in Figure 4.6 A iii is that of the 100 µM (1) sample, which has a higher amount of dimer at t=0h

relative to the other replicate. In the case of 400 µM, the good agreement with an initial second order dependency of monomer consumption further supports that the reaction reflected is the nucleation reaction, therefore it doesn't matter as much that the dimer was present in a high amount at the start of aggregation. The agreement of the early insights and the detailed chemical kinetics mechanism highlights the usefulness of applying this simpler approach, even given the limitations discussed.

The effect of the polishing process on the final fit concentrations is modest. The largest deviation affects the 400 µM sample, which is treated as having an effective concentration of 322 µM (see Figure 4.7). This would imply a 20% experimental error in that sample with respect to the others. Since that is the highest concentration sample, a slightly longer (with respect to the other samples) dead-time before the first datapoint (approximately 15 minutes for all samples) would have the largest effect on the concentration dependency, leading to the sample appearing to have a lower concentration.

Finally, it is worth noting that it was very difficult to fit the dimer traces well, as that was the last issue to be solved in the modelling process. The biggest improvement in goodness of fit came from defining the nucleus size in generation 2, with a dimerisation and nucleation involving a dimer and n monomers (hn; n=1.5). Yet, even with a nucleation reaction that was very close to the final best (dimer and two monomers; hn), there were systematic deviations in the dimer traces. A more complex model was required for those to flatten, with the main contribution coming from adding the elongation by dimer step. Further, the extensive nucleation tests showed that a good fit is very sensitive to the nucleation step; if the nucleation reaction was not that of the winning model or a very similar one, there were prominent systematic deviations in the dimer traces. In conclusion, the correct nucleation step was not enough to remove the systematic deviations in the dimer traces if other important steps were missing, but it is essential for the overall best model fit, as can be seen by the comparison of nucleation reactions in more complex models. We think this has strong implications when considering the choice of model systems and aggregation conditions. As discussed in the

introduction, one of the areas of controversy in the modelling of tau has been the influence of redox conditions. Many literature model systems contain cysteine mutations (even in constructs containing only one cysteine, which is our case [79] [78]), which prevents probing potential dimerisation via disulfide linkage. Our results suggest that not only is the dimer state of the R3 model system crucial for its aggregation, but that the availability of distinct NMR chemical shift signatures of the monomer and dimer species allowed us to discriminate very finely between similar models of primary nucleation.

Analysis of the second dataset reveals a monomer conformational change prior to dimerisation

Figure 4.12 shows the result of independently modelling the second dataset of R3-only traces (see Methods for experimental differences). The pool of steps was very similar to the one for the first dataset, with the addition of a unimolecular monomer conformational change step (k) and the variations in nucleation and elongation steps that directly result from the existence of two distinct monomeric species (e.g. c2 and c3 for the dimerisation/nucleation step). The final best model, discussed in detail below, had a global fit χ^2 of 3.1, and was arrived at by using the data from five samples; a sixth sample (a repeat of 94 µM) not used for modelling is shown as crossvalidation. The global fit of the six samples was 3.7 when the crossvalidation sample's individual frac and srac parameters were given the values fitted for 376 µM, with which the crossvalidation traces had the best agreement in terms of proportional initial monomer signal. The global fit including the extra sample improved to a $\chi^2 = 2.8$ after allowing only srac and frac of the crossvalidation sample to be optimised.

The process taken to model the second dataset is summarised in Figure 4.13 and recorded fully in Table 4.3. There were six generations of model building, where polishing was turned on in generation 6. The monomer conformational change step (k) and heterogeneous dimerisation (from an original monomer and a changed monomer; c3) were incorporated in generation 2, which improved the fit substantially and flattened the monomer and dimer residuals. This was followed by tetrameric

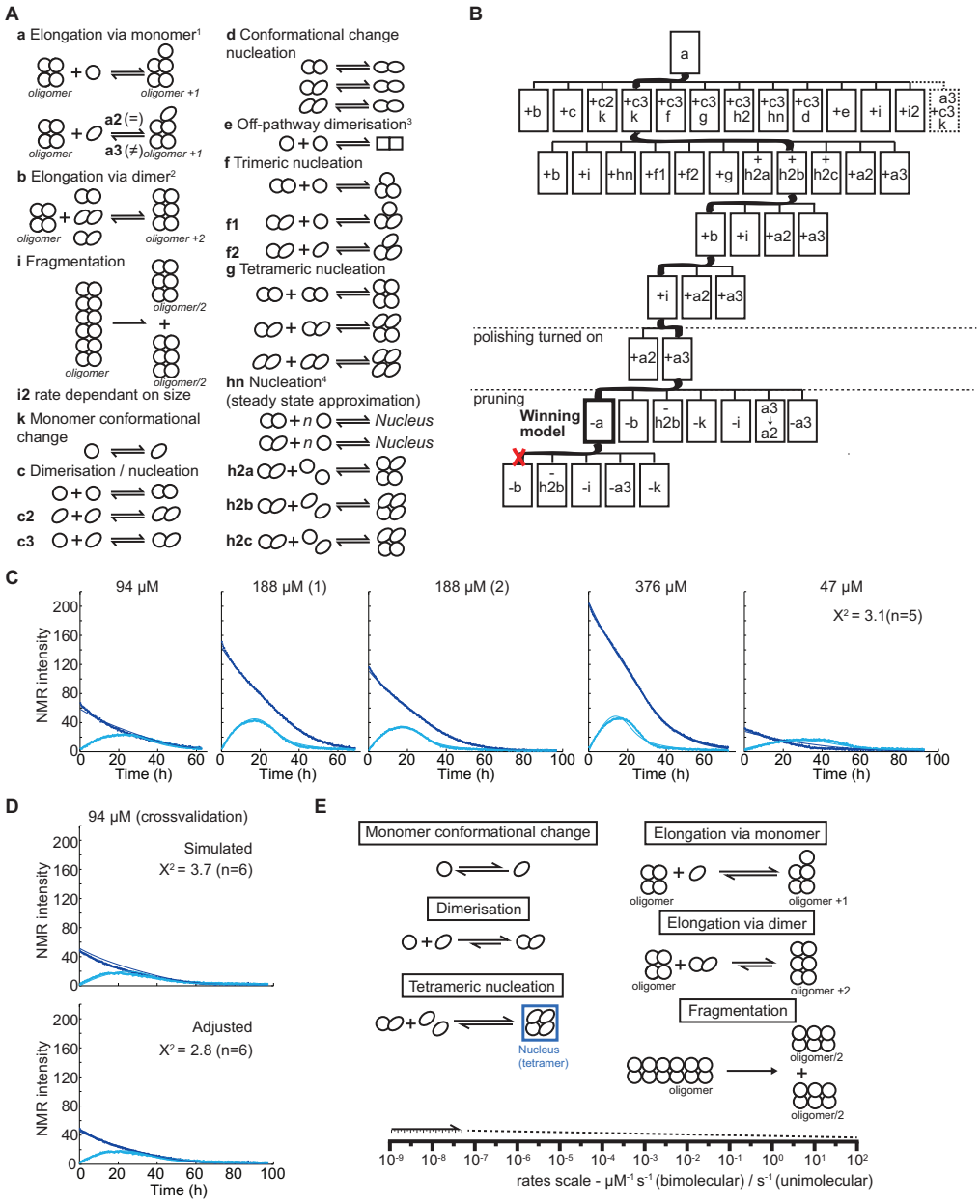


Figure 4.12: Mechanistic analysis of the second dataset of R3-only data. A. Pool of steps. Notes: 1: If a is the only reaction, this corresponds to linear polymerisation. step b involves either of the dimers from steps c, c2 or c3, depending on which one is part of the model. 3: The resulting dimer doesn't participate in any further reactions.4: n can either be free to optimise (hn) or fixed (h2). If n=1, this step is equivalent to f. B. Trajectory of the systematic search. C. Fits corresponding to the winning model. The monomer fraction (frac), and concentration/signal ratio ($srac$) parameters and the equivalent concentration ($C_{eq} = \frac{srac}{srac_{ref}} C$), based on the reference srac ($srac_{ref}$) and sample concentration (C), taking the 94 µM sample as the reference, are: 94 µM (1): frac = 0.93, srac = 0.67, C_{eq} = 94 µM; 188 µM (1): frac = 0.96, srac = 0.80, C_{eq} = 223.5 µM; 188 µM (2): frac = 0.97, srac = 0.61, C_{eq} = 172 µM, 376 µM: frac = 1.00, srac = 0.55, C_{eq} = 307.5 µM, 47 µM: frac = 0.88, srac = 0.68, C_{eq} = 48 µM, 94 µM (crossvalidation): frac = ,0.93, srac = 0.51, C_{eq} = 72 µM. D. Crossvalidation of the model with another 94 µM sample. χ^2 are global (n=6). (i) Simulation of the crossvalidation sample with parameter values from the best fitted model (frac and srac from the 376 µM sample). (ii) Fit of the crossvalidation sample allowing only its specific frac and srac parameters to be optimised. E. Mechanism of the winning model (ID 41 in Table 4.3). The length of the chemical equilibria arrows corresponds to the order of magnitude of the rate (discretised to the closest 0.5).

nucleation from dimer and two changed monomers in generation 3, at which point the dimer fits started to recapitulate the characteristic build-up and decline of the data, and elongation by dimer in generation 4, which further improved the dimer fits in particular. The addition of fragmentation led to small further improvements in χ^2 (30.8 vs 33.8) as did the elongation by changed monomer (30.6). Polishing was turned on at this point, in generation 6, due to the observed convergence of the modelling process, and had a large effect (see Figure 4.13 vii vs vi) on the χ^2 (larger than for the first dataset). Before turning on polishing, and since generation 3, the fits had systematically underestimated the signal of both monomer and dimer for the first 188 µM (1) sample (in red) while overestimating them for the 376 µM sample. Since the polishing effectively allows changing the concentration of each sample (see legend of Figure 4.12), which is why it is not carried out until the late stages of the modelling, it's important to inspect this process in some detail. In this case, and with respect to the 94 µM sample, the polishing led to the 188 µM (1) sample being readjusted to 223 µM (a 20 % increase), the 188 µM (2) sample to 172 µM (a 10% decrease), and the 376 µM to 307 µM (a 20 % decrease), while the 47 µM remained unadjusted. This can be better rationalised by taking into account experimental details for this dataset. The first two samples, 94 µM and 188 µM (two first from the left in the figures), were acquired from a common high concentration stock, while the remaining three (and the crossvalidation sample) were part of a second set of experiments from a different stock (both stocks made from dissolving the same batch of peptide). Either experimental variability in handling the reagents, or different gain settings in the NMR spectrometer, or a combination of both, resulted in a consistently lower signal for three out of four samples of the second set of experiments (other than the 47 µM). This is best seen by comparing the initial monomeric signal, e.g. between the 188 µM (1) and (2) replicates (black and red traces), the 94 µM first sample and crossvalidation sample, and the roughly only 1.5 times (versus 2-fold expected) larger signal of the 376 µM sample compared to the 188 µM (1) (red) sample. These differences explain why in the crossvalidation fits the best unoptimised fit was obtained when the second

sample of 94 µM was given the srac and frac values of a sample of the second set of experiments (376 µM, see Figure 4.12 D i), rather than those of its corresponding replicate. This variability would be best reduced for future work, for example by acquiring the entire dataset from a single higher concentration initial stock and in one set of experiments, close to each other. However, the polishing feature of our modelling approach, applied only to fine tune well fitting models, can compensate for it within a global fitting procedure. In this case, it also indicates that there was not only systematic error, given the lack of adjustment of the 47 µM sample effective concentration. Most importantly, the final best model (Figure 4.13 viii), as well as the late generation winners (Figure 4.13 iv and v) recapitulated the features of the monomer and dimer traces well *without* the fine-tuning allowed by the polishing of the srac and frac parameters, confirming the data has been well fit in a global way.

Table 4.3: Modelling results for the R3-only second dataset. Best result for each model tested in the systematic modelling process of the second dataset of R3. The winning models of each stage are shown in bold and the overall best model is shown in bold italicised. 2-ext: extra model run at the level of complexity equivalent to generation 2. p: parameters, σ: perturbation in basin-hopping like optimisation. The asterisks in the Parent? category indicate that the model with ID 5 was the winner of generation 2 used to continue with the process, while the model with ID 17, tested later, has a lower χ^2 - this is discussed in the text. P1: First pruning. P2: Second pruning.

ID	Gen.	Model	p.	χ^2	Parent?	Polish?	Guess	σ
1	**1**	**a**	**3**	**571.17**	**Yes**	**No**	**naive**	**0.01**
2	2	a,b	5	164.58	No	No	inherited	0.05
3	2	a,c	5	280.75	No	No	naive	0.05
4	2	a,c2,k	7	212.82	No	No	naive	0.05
5	**2**	**a,c3,k**	**7**	**104.83**	**Yes***	**No**	**naive**	**0.2**
6	2	a,c3,f	7	171.28	No	No	inherited	0.2
7	2	a,c3,g	7	190.61	No	No	inherited	0.01
8	2	a,c3,h2	7	280.82	No	No	naive	0.1
9	2	a,c3,hn	8	142.33	No	No	inherited, n=1e-09(->n=6e-03)	0.01
10	2	a,c3,hn	8	142.29	No	No	inherited, n=1(->n=3e-03)	0.2
11	2	a,c3,hn	8	281.25	No	No	naïve, n=2 (->0.5)	0.2
12	2	a,c3,hn	8	136.44	No	No	inherited, n=3 (->0.4)	0.01
13	2	a,c3,d	7	142.51	No	No	inherited	0.2
14	2	a,e	5	198.85	No	No	inherited	0.2
15	2	a,i	4	571.17	No	No	inherited	0.1
16	2	a,i2	5	571.17	No	No	inherited	0.05
17	2 - ext	a3,c3,k	7	90.15	No*	No	naive	0.1
18	3	a,b,c3,k	9	54.55	No	No	inherited	0.01
19	3	a,c3,i,k	8	103.55	No	No	inherited	0.01
20	3	a,c3,hn,k	10	63.76	No	No	naïve, n=1e-09 (->2e-02)	0.2
21	3	a,c3,hn,k	10	63.75	No	No	naïve, n=1 (->2e-02)	0.1
22	3	a,c3,hn,k	10	115.49	No	No	inherited, n=2 (->5e-04)	0.05
23	3	a,c3,hn,k	10	105.03	No	No	inherited, n=3 (->8e-04)	0.05
24	3	a,c3,f1,k	9	100.70	No	No	naive	0.01
25	3	a,c3,f2,k	9	53.38	No	No	inherited	0.2
26	3	a,c3,g,k	9	59.85	No	No	naive	0.05
27	3	a,c3,h2a,k	9	150.49	No	No	inherited	0.01
28	**3**	**a,c3,h2b,k**	**9**	**51.00**	**Yes**	**No**	**naive**	**0.05**

Table 4.3 continued from previous page

ID	Gen.	Model	p.	Chi2	Parent?	Polish?	Guess	σ
29	3	a,c3,h2c,k	9	61.52	No	No	inherited	0.1
30	3	a,a2,c3,k	7	103.62	No	No	naive	0.2
31	3	a,a3,c3,k	9	64.25	No	No	naive	0.05
32	**4**	**a,b,c3,h2b,k**	**11**	**33.35**	**Yes**	**No**	**naive**	**0.2**
33	4	a,c3,h2b,i,k	10	50.96	No	No	inherited	0.01
34	4	a,a2,c3,h2b,k	9	50.89	No	No	inherited	0.01
35	4	a,a3,c3,h2b,k	11	50.10	No	No	inherited	0.05
36	**5**	**a,b,c3,h2b,i,k**	**12**	**30.76**	**Yes**	**No**	**inherited**	**0.01**
37	5	a,a2,b,c3,h2b,k	11	32.49	No	No	inherited	0.05
38	5	a,a3,b,c3,h2b,k	13	31.53	No	No	inherited	0.1
39	6	a,a2,b,c3,h2b,i,k	21	5.09	No	Yes	inherited	0.05
40	**6**	**a,a3,b,c3,h2b,i,k**	**23**	**3.27**	**Yes**	**Yes**	**inherited**	**0.01**
41	***P1***	***a3,b,c3,h2b,i,k***	***21***	***3.13***	***Yes***	***Yes***	***naive***	***0.2***
42	P1	a,a3,c3,h2b,i,k	21	27.46	No	Yes	naive	0.05
43	P1	a,a3,b,c3,i,k	21	20.62	No	Yes	naive	0.2
44	P1	a,b,c3,h2b,i	19	131.82	No	Yes	naive	0.01
45	P1	a,a3,b,c3,h2b,k	22	3.61	No	Yes	inherited	0.1
46	P1	a,a2,b,c3,h2b,i,k	21	4.95	No	Yes	inherited	0.05
47	P1	a,b,c3,h2b,i,k	21	5.15	No	Yes	naive	0.2
48	P2	a3,c3,h2b,i,k	19	46.35	No	Yes	inherited	0.01
49	P2	a3,b,c3,i,k	19	20.57	No	Yes	inherited	0.2
50	**P2**	**a3,b,c3,h2b,k**	**20**	**3.37**	**Yes**	**Yes**	**inherited**	**0.2**
51	P2	b,c3,h2b,i,k	19	12.07	No	Yes	inherited	0.05
52	P2	a,b,c3,h2b,i	19	131.82	No	Yes	naive	0.01

The winning model is the result of the 1st pruning stage, which is shown in detail in Figure 4.14. The removal of the step corresponding to elongation by original (unchanged) monomer (which was incorporated in generation 1 by default in the modelling strategy, since it is equivalent to linear polymerisation) leads to a small gain in goodness of fit (3.1 vs 3.3, see ii) vs i)) with a reduction of complexity by 2 parameters, which is therefore accepted by default. The removal of the fragmentation step has a modest effect on the fit, while there are larger effects when removing the elongation by changed monomer or constraining the rates of that step to those of the original monomer, but in all these cases the main features of the traces are recapitulated. In contrast, removing any of the other steps (see vi, vii, viii) leads to the reappearance of clear systematic residuals, in the dimer traces in particular.

The elongation step using conformationally changed monomer fits the data better also at an earlier stage In line of the result of the pruned model without step *a*, we performed a test at an earlier stage equivalent to generation 2 by replacing this step, in the winner model of generation 2, with the changed monomer counterpart (a3), and allowing it to inherit the rates fitted for the *a* step in the parent model (from generation 1) in the inherited initial guesses mode.

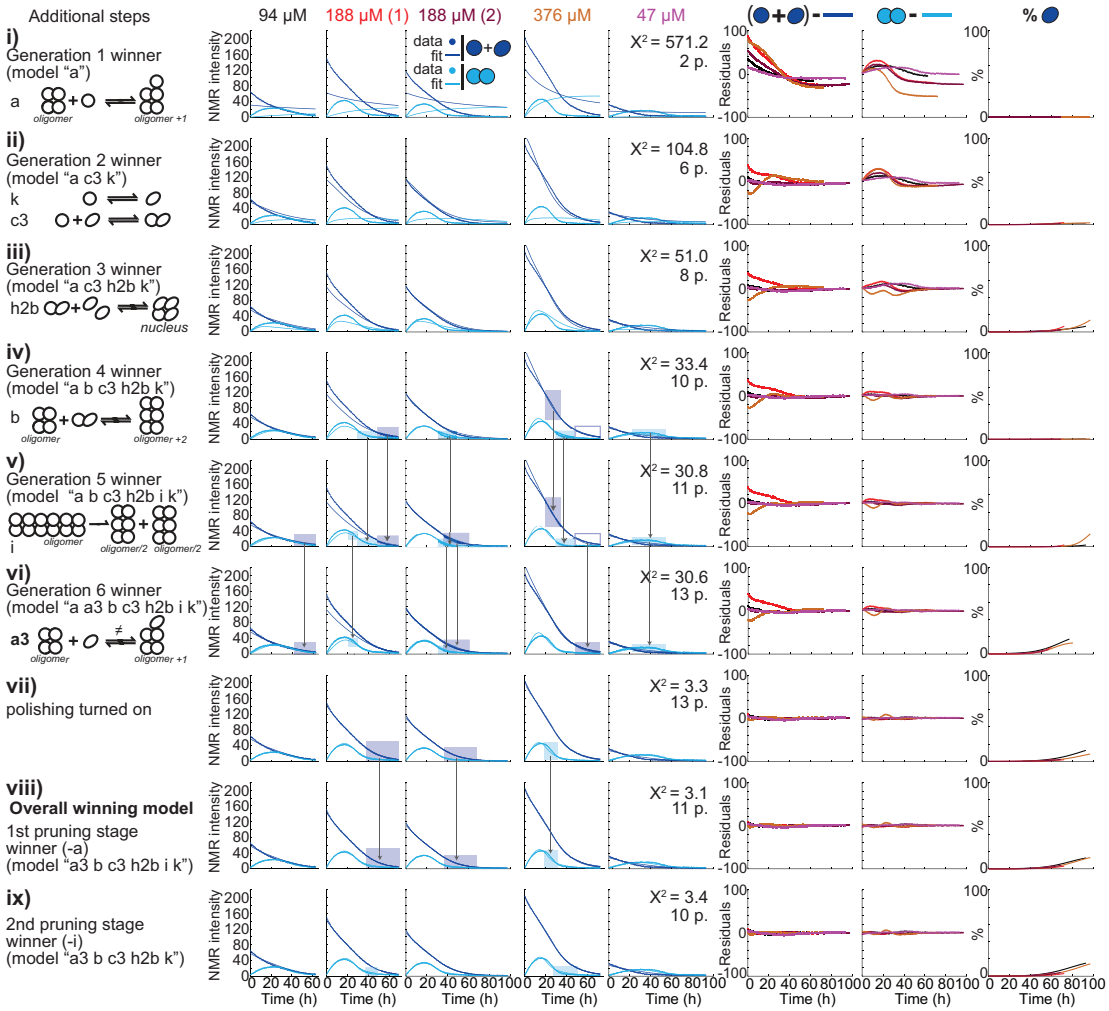


Figure 4.13: Modelling path for the R3-only second dataset. Each row shows the winner of the generation indicated in the left, in terms of its global goodness of fit (χ^2) and residuals (data - fit) for the monomer and dimer traces separately. The number of parameters reflects only those that correspond to a chemical reaction, excluding the initial monomeric fraction and signal to concentration ratio. The steps that have been added to the model at each generation are shown on the left. The results shown in each row correspond to ID 1, 5, 28, 32, 36, 40 (vi; unpolished and vii; polished), 41 and 50, respectively, in Table 4.3.

This resulted in a better fit with respect to the generation winner (90.4 vs 104.8, see Figure 4.14 B), which means this alternative model would have won that generation. However, it's clear that both models still have substantial residuals, both in the monomer and especially the dimer traces. The alternative path that would ensue from taking that new winner model through the full modelling process was not further explored. In the future, when multiple species of monomer are present, it seems sensible to consider these two types of elongation reactions earlier in the modelling process, in an analogous way to what was done for the dimerisation and nucleation steps.

Further, the lower χ^2 of a pruned model with fewer parameters indicates another instance of the optimiser struggling to leave a local minimum, since the winner model of generation 6 could have achieved the same goodness of fit by setting the rates of the *a* step to zero. However, as in the previous case (the nucleation step in the modelling of the first dataset), the full course of the modelling process did arrive to the minimally complex best solution.

In the 2$^{\text{nd}}$ pruning stage, similar to the first, the closest fit (i.e. least loss of goodness of fit) corresponds to removing fragmentation, and the largest to removing the conformational change of monomer (see Figure 4.15). In this case, the removal of any of the other steps leads to large systematic deviations, mainly for the dimer traces, including that of *a3* (iii). In the 1$^{\text{st}}$ pruning stage, the removal of either elongation by monomer step had minor effects, indicating a certain degree of redundancy between them, which is consistent with the removal of one of them in the 1$^{\text{st}}$ stage. In contrast, the loss of goodness of fit when removing *a3* indicates there are no further redundant steps.

The amount of conformationally changed monomer is negligible at early stages of the aggregation reaction Finally, it's interesting to discuss the relative abundance of the two monomer species in the final model and in the course of the modelling process. In the path through generation/stage winners (Figure 4.13), the conformationally changed species is consistently not abundant at early

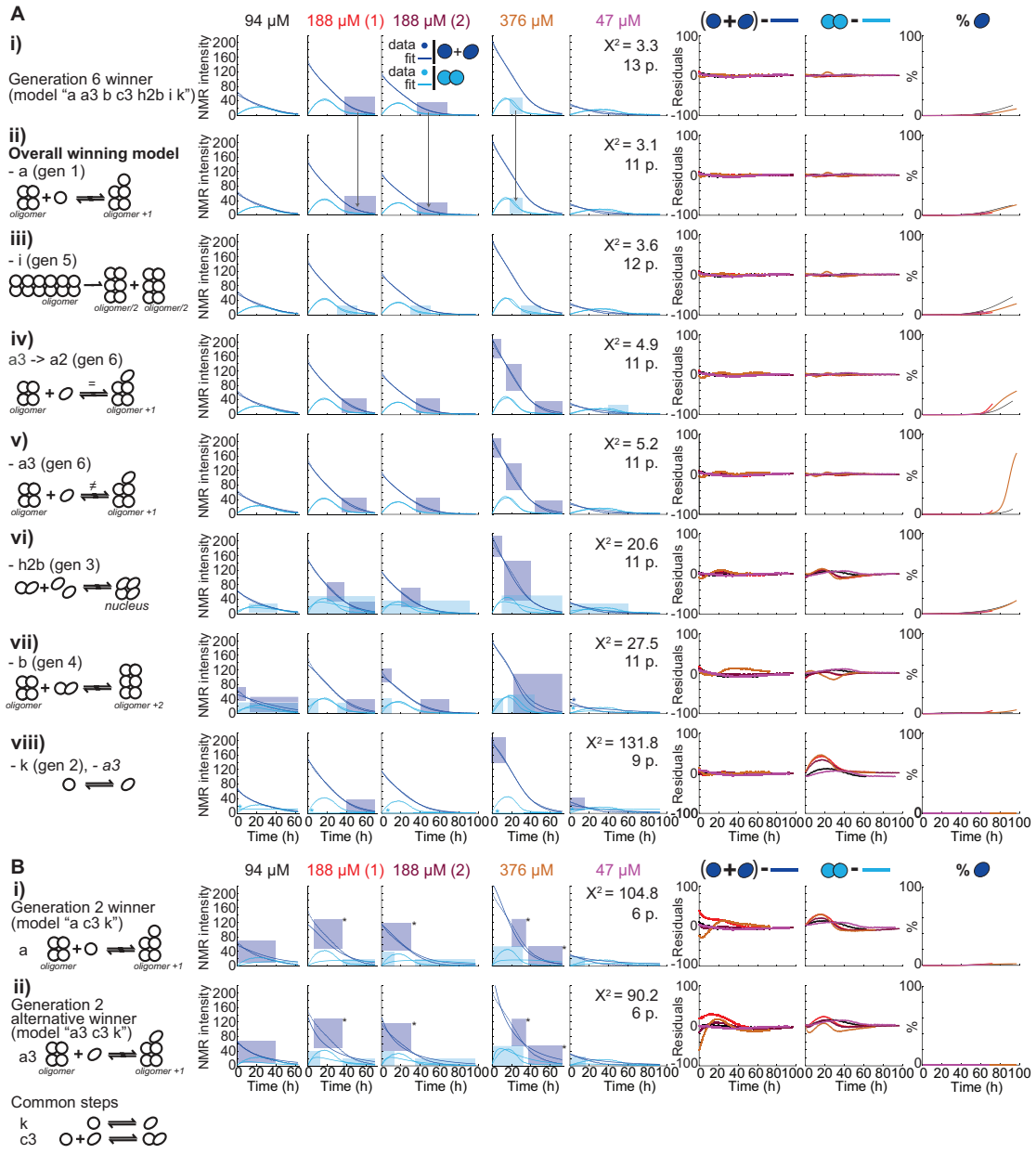


Figure 4.14: First pruning stage for the R3-only second dataset. A. Each row shows one of the pruned models, in terms of its global goodness of fit (χ^2) and residuals (data - fit) for the monomer and dimer traces separately. The number of parameters reflects only those that correspond to a chemical reaction, excluding the initial monomeric fraction and signal to concentration ratio. The steps that have been added to the model at each generation are shown on the left. The results shown in each row correspond to ID 40, 41, 45, 46, 47, 43, 42 and 44 in Table 4.3. B. Result of changing the elongation by monomer step in generation 2. The models depicted include the winner of the generation ($\chi^2 = 104.8$) and the newly tested model ($\chi^2 = 90.2$), and correspond to ID 5 and 17 respectively, in Table 4.3. The shaded areas show improvements of systematic deviations in the 90.2 fit model with respect to the 104.8 model, while the asterisks indicate better fit in the 104.8 model.

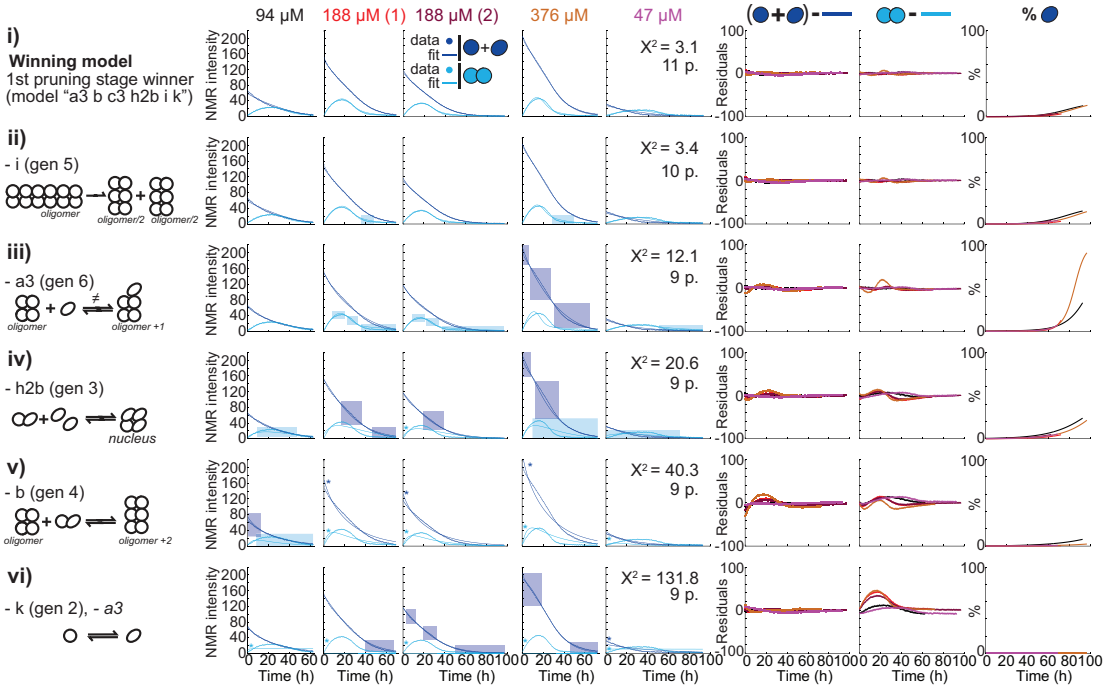


Figure 4.15: Second pruning stage for the R3-only second dataset. Each row shows one of the pruned models, in terms of its global goodness of fit (χ^2) and residuals (data - fit) for the monomer and dimer traces separately. The number of parameters reflects only those that correspond to a chemical reaction, excluding the initial monomeric fraction and signal to concentration ratio. The steps that have been added to the model at each generation are shown on the left. The results shown in each row correspond to ID 41, 50, 51, 49, 48 and 52 in 4.3.

stages (0-40h) and generally builds up to around 20 % only in the late second half of the reactions. The exception is the generation 4 winner, when elongation by dimer is incorporated in the model, in which there is no substantial build-up. In the pruning stages, there are exceptions to the build-up to 20 % as well, in particular in the models in which *a3* has been removed or replaced by *a2*. This can be understood since the ability of the changed monomer to elongate reduces its concentration, and this is completely abolished if *a3* is removed, and partially if it is happening at the same rates as the original monomer (step *a*) and those are slower. That is the case; in orders of magnitude, the generation 6 winner has forward and backward rates of 10^{-4} and 10^{-3} for *a* (original monomer), while for *a3* (conformationally changed monomer) they are 10^{0} and 10^{-1}, respectively. Another noticeable aspect of the build-up of the changed monomeric species is that it coincides with the decline in

dimer concentration, which is sensible since the changed monomer is a reactant in the dimerisation reaction and also in the nucleation reaction involving the dimer.

Discussion of the second dataset model

For this dataset, the rate limiting step is the conformational change of monomer. Even if the forward rate constant of nucleus assembly is larger than the forward rate for the monomer conformational change, given that there is a second order dependency on the changed monomer and that this species is very scarce (negligible in the first 40 hours), the rate of nucleus assembly will still be slow. A similar logic applies to the formation of the heterogeneous dimer, since that also has a dependency (first order) on the concentration of the scarce changed monomer.

This agrees with the findings from the initial rates method, for which this dataset is better suited, since the concentration of dimer at t=0h is much lower with respect to that of monomer than in the first dataset. The unimolecular early event revealed by this analysis would correspond to the conformational change step (k) incorporated into the pool for the modelling of this dataset. In this case neither the dimerisation or nucleation events, which are second order for monomer, even if also essential steps in the mechanism, dominate the rate law at the early stages (t < 4h).

Comparison of the first and second dataset models

There are a few differences between the best final models for the two datasets, and in the path taken by the algorithm, which are interesting to comment on, including some that once rationalised appear much more modest.

The main difference between the models obtained from the first and second protein only datasets concerns the monomer conformational change step that is required in order to fit the second dataset well, and which is reflected in the linear dependency of the initial rate of monomer decay. In contrast, the first dataset showed a second order dependency for initial monomer consumption and was well described without including any monomer unimolecular step. The rest of the steps included in the final models are the same, most importantly regarding the identity of the nucleus and the species involved in nucleation. There are some differences in the fitted rate

constants, of which the biggest is the faster forward nucleation rate constant for nucleation in the second dataset. However, as discussed above, the effective rate of nucleus formation will be reduced by the scarcity of the changed monomer species.

In order to integrate the two models, an early hypothesis was that the second dataset had been caught at an earlier time of the aggregation process when the unimolecular conformational change was rate-limiting, while the first one had already pre-equilibrated into a majoritary population of changed monomer to start with. However, the results disagree with this hypothesis, since the changed monomer does not build-up in the first 40 hours of the reaction, and only to a maximum of 20 %. This suggests a small amount of the changed monomer is enough to drive the reaction forward, and that might already have built-up by the start of the first dataset reactions. Alternatively, the initial conformational change might be slower in a 90 % D_2O buffer compared to a 90 % H_2O buffer.

The common steps were also included in the final model in the same order in the modelling process (see Figure 4.8 and 4.13). The goodness of fit of the final fits is also comparable; since the average per point χ^2 reported (see Chapter 3) are not divided by the error of each point, they are proportional to the signal magnitude. That is a bit below two-times higher (1.5-1.75 taking an average of the srac values for each dataset) for the second dataset compared to the first one. Therefore the final best fit of $\chi^2 = 3.1$ units would be equivalent to a 1.77-2.06 fit for the first dataset, and that is closer to the 1.22 fit that was initially apparent.

Another apparent difference, which in reality is a similarity, is the difference in the order of impact of each step in the pruning stage (see Figure 4.9 vs 4.15). In the second dataset case, there seems to be a much smaller impact of removing the fragmentation step on the goodness of fit (3.4 vs 3.1, compared to 3.62 vs 1.22 in the first dataset). However, as explained above, the features of the traces (in particular the dimer, which is the hardest to fit well) are maintained in the model without fragmentation both for the first and second datasets. Similarly, the removal of elongation by monomer has the largest χ^2 effect for the first dataset, while it is the second in the second dataset, but in both cases the build-up and decline of the

dimer traces are maintained. Finally, and similarly for both datasets, the removal of either tetrameric nucleation or elongation by dimer leads to the reappearance of significant dimer residuals, as well as worse χ^2 values.

4.4.2 Mechanism of action of D-TLKIVW on R3 aggregation

General features of R3 aggregation in the presence of D-TLKIVW

Figure 4.16 shows a comparison of the amide and methyl regions of the 1H NMR spectra in aggregation reactions of R3-only (first dataset, 100 µM) and R3 with the inhibitor D-TLKVIW (100 µM + 48 µM), at the start of the reaction, a middle point and an end point for each sample. Several features can be noted: 1) there is a slower decay of monomer signal and a larger build-up and delayed decline of the dimer signal in the presence of inhibitor (compare the t=23h timepoint of the R3-only sample with the t=34h timepoint of the sample with inhibitor); 2) there are no chemical shift perturbations (CSPs) larger than the linewidth in the resolved histidine resonances or the inhibitor peaks (see Figure 4.16 C); 3) there are well-resolved D-TLKIVW peaks (shown are the D-Tryptophan aromatic peaks and a methyl region doublet for D-Isoleucine), which decay approximately by 20-30% over the timecourse of the experiment. Importantly, the fact that the D-tryptophan aromatic peaks can be identified and don't overlap with the histidine resonances of the R3 monomer or dimer justifies that the concentrations of R3 species can be reliably monitored following the same peaks as in the R3-only experiments.

In the study where the inhibitor D-TLKIVW was developed, the authors reported a lack of CSPs of the inhibitor indole peak in conditions of great excess of monomeric VQIVYK (the starting hexapeptide of the R3 repeat, and one of the hotspots for aggregations, which was used to structurally design this inhibitor) [134]. This is in agreement with our observations, however without a full titration in our case. Similarly, the decline in inhibitor signal is consistent with Sievers *et al.* observing concentration-dependent decays upon titration with either pre-formed aggregates of the construct K19 or of the hexapeptide VQIVYK.

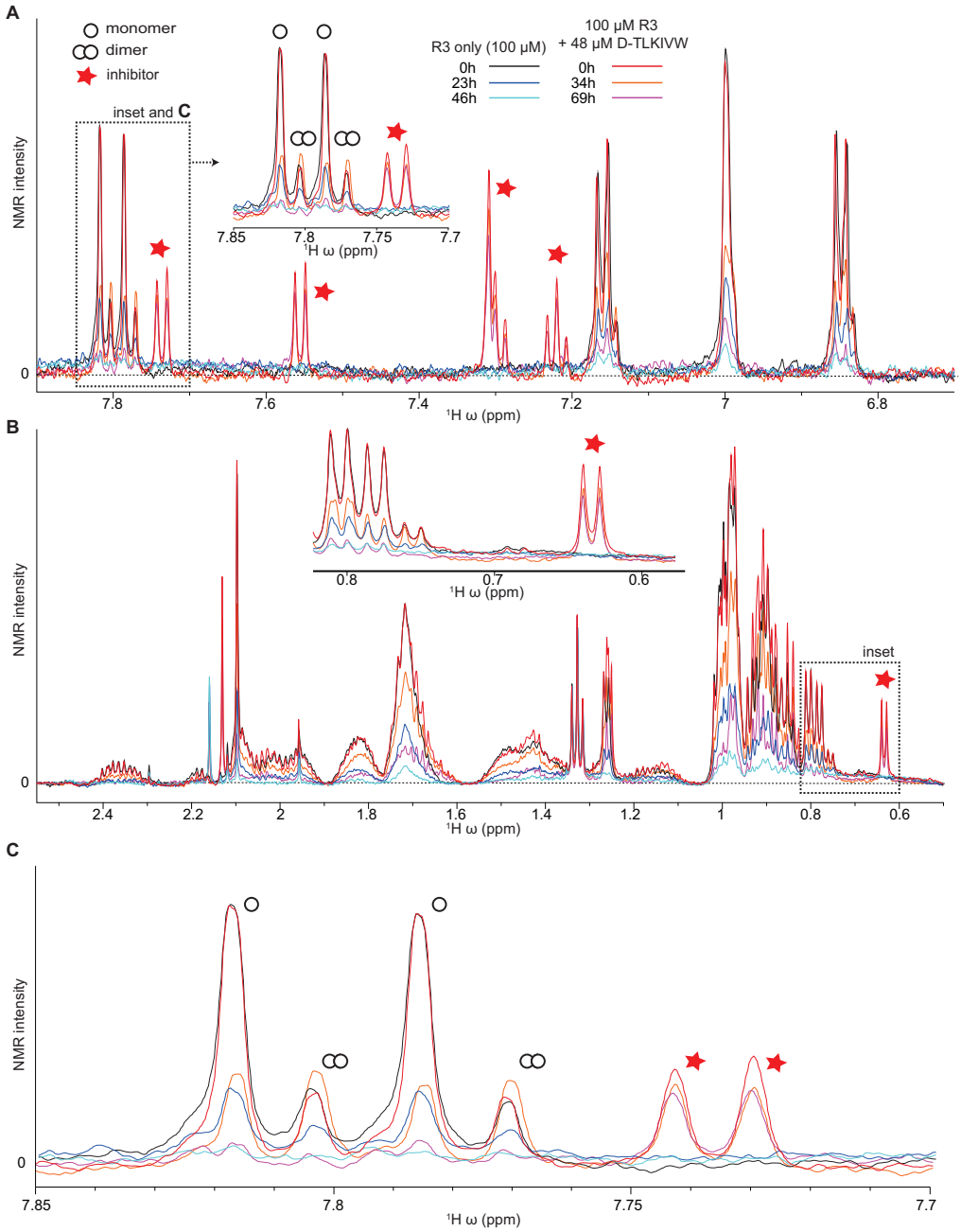


Figure 4.16: Overlay of ^{1}H amide and methyl regions in the presence or absence of inhibitor. The non inhibitor data corresponds to 100 µM R3, and the inhibitor data to 100 µM R3 with 48 µM D-TLKIVW. Signal intensity for the no-inhibitor sample was normalised by a factor of 1.1 to have an equivalent signal at the first time point, and shifted upfield by 0.005 ppm (A) or 0.0038 (B). A. Amide region. The inset shows the monomer and dimer peaks corresponding to histidines (around 7.8 ppm) that were used to quantify aggregation (see also C). Several peaks (red stars) appear between 7.2 and 7.7 ppm which correspond to the terminal tryptophan residue in D-TLKIVW (by comparison to reference spectra in *http* : *//www.hmdb.ca/spectra/nmr_one_d/*1608). B. Methyl region. The inhibitor doublet (red stars) corresponds to the isoleucine residue. C. Zoomed in region around the R3 histidine peaks and aromatic resonances from D-TLKIVW.

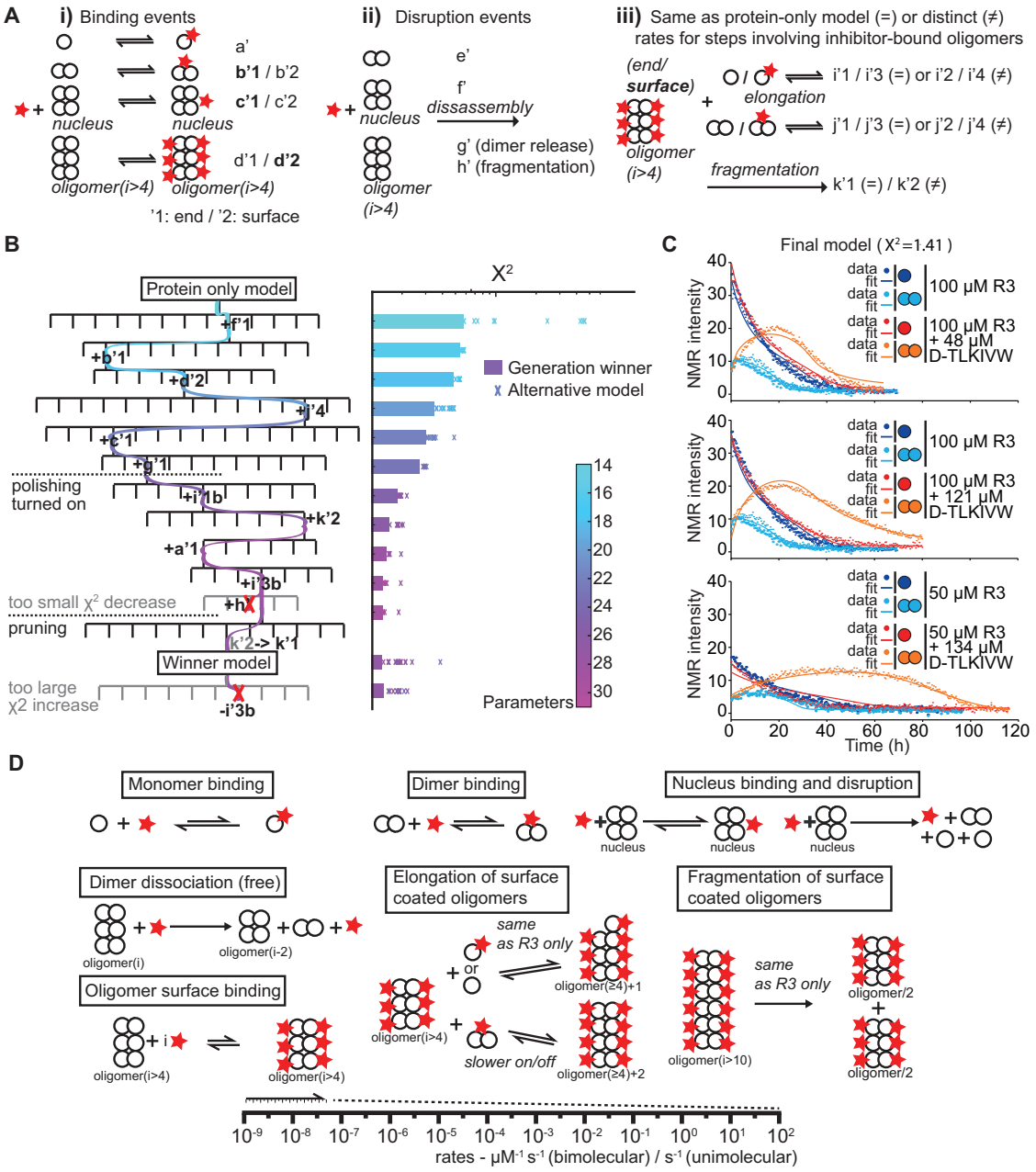


Figure 4.17: Overview of modelling results for the action of the D-TLKIVW peptide on R3. A. Summary of pool of steps tested, including binding events, disruption events and elongation/fragmentation of inhibitor-bound species; where multiple similar steps exist, the variants present in the final model are shown (see Figure 4.3 for more details). B. Tree-like description of the modelling path, including generations of increasing complexity and two pruning stages. The letter codes correspond to the steps in A / Figure 4.3. The number of parameters is the sum of protein-only model parameters and inhibitor model parameters. C. Data and fits for the final winning model (ID 125 in Table 4.4). The data for the sample 100 µM R3 + 48 µM D-TLKIVW was truncated at 64 hours for kinetic modelling. The monomer fraction (frac), and concentration/signal ratio ($srac$) parameters and the equivalent concentration ($C_{eq} = \frac{srac}{srac_{ref}} C$), based on the reference srac ($srac_{ref}$) and sample concentration (C), taking the 100 µM R3 + 121 µM D-TLKIVW sample as the reference, were: 100 µM R3 + 121 µM: frac = 0.91, srac = 0.4, $[R3_{eq}] = 100$ µM; 100 µM R3 + 48 µM: frac = 0.87, srac = 0.46, $[R3_{eq}] = 115$ µM; 50 µM R3 + 134 µM: frac = 0.70, srac = 0.36, $[R3_{eq}] = 45$ µM. D. Mechanism of the final winning model. The length of the chemical equilibria arrows corresponds to the order of magnitude of the rate (discretised to the closest 0.5).

The mechanism of inhibition of D-TLKIVW is multifaceted and affects multiple species in the pathway to aggregates

The inhibition by the peptide D-TLKIVW is summarised in 4.17 and the complete results are recorded in Table 4.4. The inhibition mechanisms proposed in the pool, summarised in A (see Figure 4.3 for a more detailed depiction), correspond to interactions with all important species identified in the mechanism obtained from modelling the first dataset of R3 data, namely monomer, dimer, tetrameric nucleus and larger oligomers/fibrils (any species larger than nucleus, treated collectively). They include end or surface binding to each species independently, disruption events and distinct hypotheses on whether inhibitor-bound species participate in elongation or fragmentation reactions that occur in the R3-only mechanism, and with the same or different rates. The algorithm progressed through eleven generations of model building, in which the winner model of the last one was not accepted since the χ^2 improvement was below the threshold, and two stages of pruning, with the overall best model resulting from the first pruning stage.

The evolution of the modelling process is shown in Figures 4.18 and 4.19. The top row shows the R3-only best model fits of the equivalent R3 concentration (from the first dataset) and their poor agreement with the data from the samples with inhibitor, in line with the qualitative features shown in Figure 4.16 and discussed above. For all the samples, the main systematic deviations occur in the dimer traces, whose build-up is delayed yet reaches higher values in the presence of inhibitor. As expected, the effects are concentration dependent with respect to inhibitor (e.g. compare 121 µM vs 48 µM inhibitor concentration, with constant R3 concentration).

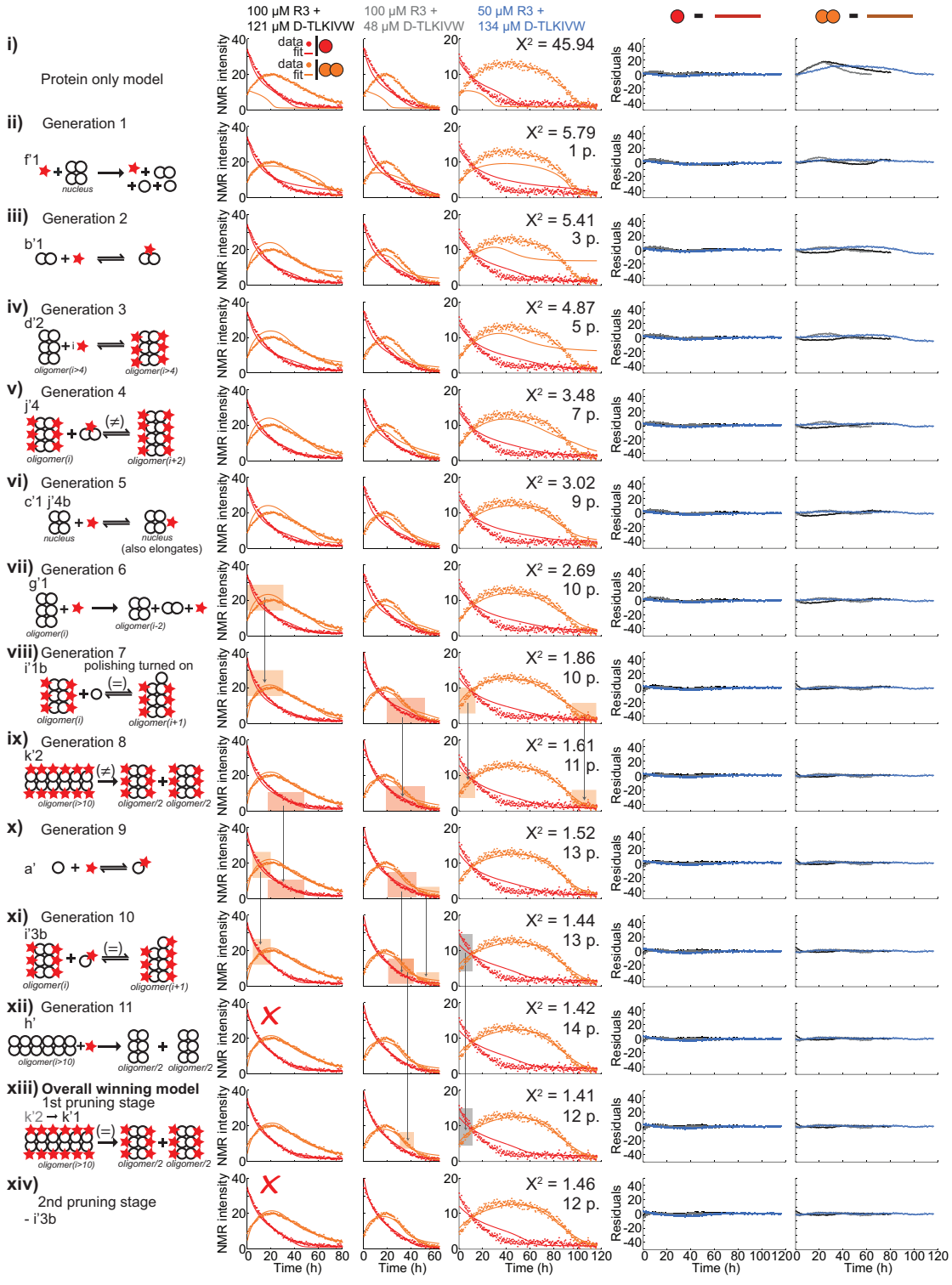


Figure 4.18: Evolution of modelling process for D-TLKIVW inhibition mechanism - Fits and residuals. Top row: best protein-only model (ID 25 in Table 4.2) shown against the kinetic traces of the inhibitor dataset. From 2nd row to the last: winner of each generation. Polishing was turned on in generation 7. The number of parameters refers only to rate parameters, excluding frac and srac. The inhibitor models displayed correspond to entries with IDs 9, 14, 29, 49, 54, 72, 84, 95, 98, 109,113, 125 and 135 in Table 4.4.

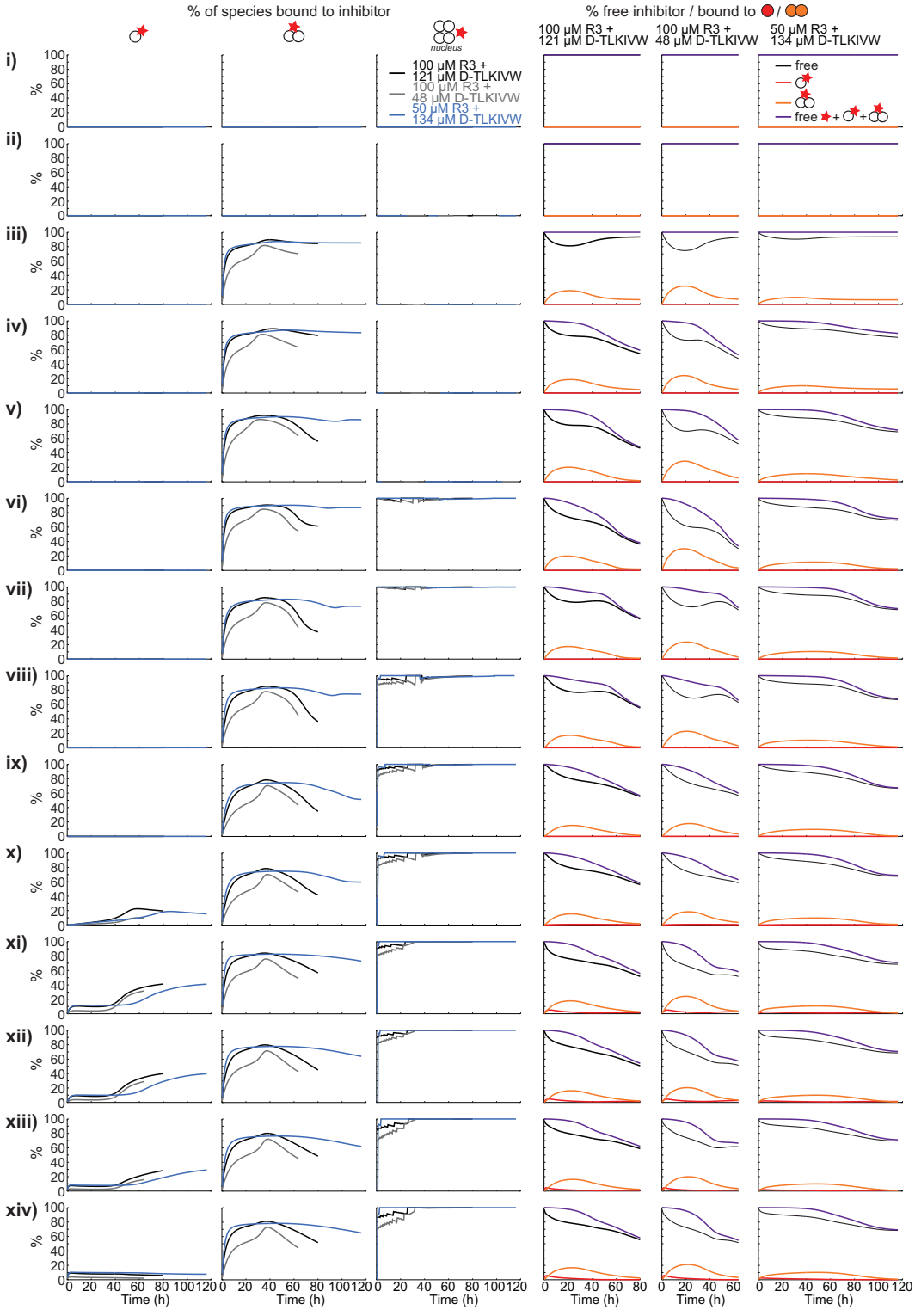


Figure 4.19: Evolution of modelling process for D-TLKIVW inhibition mechanism - Analysis of binding between small species and inhibitor. The models correspond to the same as in Figure 4.18. Left three panels. For each model, left to right: percentage of monomer bound to inhibitor, percentage of dimer bound to inhibitor, percentage of nucleus bound to inhibitor. Right three panels. For each dataset, distribution of inhibitor concentration amongst bound to monomer, bound to dimer, free, free+bound to either monomer or dimer.

The disruption of the nucleus was the first step to be incorporated to the model, which led to a big improvement in the dimer residuals yet also to larger deviations of the fits from the monomer decay (which are less affected by the presence of inhibitor, see Figure 4.18 i). Following this, binding to dimer and surface binding to oligomers were incorporated. Surface-binding to oligomers was better than end binding but not by a lot (χ^2 of 4.87 vs 5.00). In generation 4, elongation of the surface-bound oligomers by dimer-inhibitor complexes was incorporated, with different rates compared to the elongation of free oligomers by free dimer. The binding to nucleus was incorporated in generation 5 and in generation 6 a step that releases free dimers from oligomers was incorporated, with a small but noticeable improvement in the fits (and difference in χ^2 clearly above the 0.1 cut-off; see Figure 4.18 vii vs vi). Polishing was turned on in generation 7, after observing the improvements in χ^2 were getting smaller as the features of all traces were mostly recapitulated by the fitted curves. From generation 7 (other than the monomer binding step in generation 9), all the steps that were incorporated (or modified, in the pruning) belong to the category of elongation and fragmentation steps involve inhibitor-bound species. Finally, the winner of generation 11 was not accepted because of a too small reduction in χ^2. The winner of the first pruning stage was, since it reduced complexity by 1 parameter while having a marginally smaller χ^2, but not that of the second pruning stage since there was a loss of χ^2 without any reduction in complexity. The best models fit have the relatively larger deviations in the fits to concentration 50 +134 and in particular the monomer trace, in line with what was observed in the other two datasets, since that has the lowest signal of all the traces.

Analysis and comments on the amounts of inhibitor-bound populations

From Figure 4.19, we can also comment on the binding of inhibitor to monomer, dimer and nucleus as percentage of the respective species that is in complex form with respect to total, and similarly the percentage of inhibitor that is either free (black), bound to monomer (red) or dimer (orange), or the sum of free and bound to small species (purple, including both monomer and dimer). The difference between

the purple trace and 100 % is then the amount of inhibitor bound to larger species starting from nucleus. The step of inhibitor binding to dimer is incorporated the first in the modelling process, in generation 3, and from that stage and consistently onwards most of the dimer is in the complex form, with no evident change when the complex is allowed to elongate with different rates in generation 4. This is reflected in the inhibitor as well, and we can see that the percentage of inhibitor bound to dimer (orange traces) reaches 10 to 30 % depending on the sample, and that it follows a similar build-up and decline pattern to the dimer traces in the dataset. Therefore it seems that the existence of a dimer-inhibitor complex is one of the ways in which the inhibitor model manages to recapitulate the larger and more sustained build-up of dimer, by sequestering the dimer in a state where it cannot participate in some of the protein-only mechanism reactions, such as nucleation or elongation steps. In agreement with this hypothesis, the elongation on and off rates for the complex are slower compared to those of the protein-only by free dimer, and the models with elongation by bound-dimer at the same rates as protein-only (step j'1) fit quite poorly, sometimes worse than the parent model (this is further discussed below).

In contrast, the percentage of monomer that is in complex with the inhibitor is smaller (maximum of 20-40 %) and only builds up after 40 hours, once the monomer concentration has already declined substantially, when this step is incorporated in generation 9. It does increase when this complex is allowed to elongate, in generation 10, and this is reflected in a consistently very low percentage of inhibitor that is in a complex with monomer (red traces). In the case of the nucleus, it is almost 100 % bound to the inhibitor in all generations after, and including, generation 5 when it is included in the model.

In the traces of inhibitor, we can see that the gap between the percentage of free inhibitor (black), or the sum of free and bound to either monomer or dimer (purple), is largely due to the percentage bound to dimer (orange)s, which as commented above is following the shape of the dimer traces in the data the generation 3 model winner when the binding was incorporated. The amount of free inhibitor at the late stages is almost the same as at the start in that generation. From generation 4,

there starts to be a substantial decay of the free or bound to small species traces at the end of the experiment, which is larger in generation 4 and 5 and then is reduced again and stays very similar throughout the remaining stages. In generation 4 and 5, the larger decay is rationalised by the steps added, since the bound-oligomers are able to elongate, including the nucleus-inhibitor complex, which leads to more inhibitor-bound oligomers and therefore less free inhibitor. The rationale behind calculating the free plus bound to monomer and dimer percentage is that the inhibitor in those complexes should be visible in the 1D spectra, since the monomer and dimer are visible and the complex only has a molecular weight six amino acids larger than these free R3 species. This is discussed again later when contrasting the kinetic and spectral evidence for the binding of inhibitor to small species.

The decline of the free inhibitor traces, in particular for the 100 µM R3 samples, becomes smoother as the generations progress. Remarkably, in the final best model (xiii), there is around 70 % of visible inhibitor left at t=70h for the 100 µM R3 + 121 µM D-TLKIVW, which is in very close agreement with the experimental observations in Figure 4.16 even though the concentration of visible inhibitor was not used as a constraint in the modelling.

Results of extensions (and rationale for those) of the modelling strategy applied to the inhibitor data In the course of the generation building process, there were three instances of minor deviations from the general model building procedure defined as part of the algorithm, which are discussed here. The first one concerns the acceptance of the winner model of generation 9, with χ^2 of 1.52 (see Figure 4.18 x), which is within the 0.1 cut-off with respect to its parent (χ^2 = 1.61, see Figure 4.18 ix). When the binding of monomer was incorporated in the model, it opened up the question on whether this monomer-inhibitor complex would be able to elongate, and whether it would be with the same rate constants of the elongation of free oligomers or different ones. At that stage the model already included the steps of inhibitor-bound oligomers elongating by free monomer and by bound dimer. This was tested in two steps between generation 9 (binding step

incorporated, no involvement in elongation) and generation 10 (elongation occurs, either with same rates (step i'1b) or different rates (i'3b)). In generation 10 the model including elongation with the same rates had a χ^2 of 1.44, without any additional parameters, so it was accepted by default in comparison to generation 9. It can be considered that globally from generation 8 to generation 10 the χ^2 dropped from 1.61 to 1.44, which is within the 0.1 cut-off. The χ^2 with different rates was 1.45 (ID110) therefore clearly not justifying the additional complexity with respect to the sibling model with 1.44. This successive addition of the binding step, then the elongation step was done in the same way as for the dimer binding step (incorporated in generation 2, its involvement in elongation steps tested from generation 3), with the rationale being to avoid adding too many new parameters to optimise at the same time. It's worth noting that the binding of the nucleus and its ability to elongate or not were tested simultaneously in generation 5, yet the step of elongation of bound-oligomers by dimer had already been incorporated in generation 4, meaning the models of generation 5 only had two new parameters to optimise.

The second deviation from the general model building procedure concerns the two interactions with the nucleus identified in the modelling process: direct disruption and the binding into a complex. Since it seemed these two steps could be redundant and/or part of a single two step mechanism, this possibility was tested (c'1,f'2) in generations 4 and 5; in both cases, the models with the separate interactions had a better fit than the binding equilibrium - disruption model: in generation 4, 4.45 (ID 35) vs 4.91 (ID 37), in generation 5, the generation winner 3.02 (54) vs 3.11 (ID 58). This will be further discussed in the pruning stage below. It's interesting to compare the results of both generations, since the binding to the nucleus led to only a modest improvement in fit in generation 4 (4.45 vs the 4.87 of the parent), while in generation 5 it improved the χ^2 from 3.48 (v, ID 49) to 3.02 (vi, ID 54) but specifically in the case that it was elongating by dimer-inhibitor complex as well (the model without this elongation competency was not better than the parent, with χ^2 of 3.47 (ID 55). This indicates the binding to nucleus itself is not beneficial for the fitting unless it is coupled to the ability of the nucleus-inhibitor complex to elongate.

Finally, and thirdly, there was a model tested in order to avoid the usage of the mass corrections described in 4.3.4 for the cases when surface-bound oligomers can elongate by addition of species that don't contain an equal number of inhibitor molecules per protein subunit (i.e. the inhibitor-dimer complex, free monomer and dimer, and single-bound nucleus elongating into a surface-bound oligomer). This affected generation 4; the step of inhibitor binding to dimer *b'1* was replaced by surface-binding (*b'2*) in the model ID51, which removed the need for corrections, but had a worse fit and therefore the step *b'1* was maintained in the generation winner. In relation to that, it's worth noting that when the binding to nucleus (which elongates) was incorporated in the model in generation 5, a full set of exhaustive combinations (single, surface-binding or equilibrium disruption; with or without the ability to elongate) was tested, of which the winning model (ID 54), with a single inhibitor molecule binding to nucleus, was the best in goodness of fit. In the later generations, there were no more alternative models tested and instead the mass corrections were applied when required.

The first pruning reveals another instance of optimisation converging to a local minimum The improvement of the fit in the winning model of the first pruning stage shows another case of the optimiser not finding the global minimum, even if by a very small margin, since in generation 10 the winning model could have been optimised to the χ^2 of 1.41 by setting the rate of fragmentation to the value of the protein-only model step. This is probably achieved, in the pruning stage, by lowering the amount of monomer-inhibitor complex in the pruned model (see Figure 4.19 xiv vs x). Still, this instance of optimiser limitations is more similar to the minor improvement in the first pruning stage of the second protein-only dataset (when elongation by original monomer was removed) than the equivalent stage of the first protein-only dataset (when the nucleation step *hn* had been modified to *h2*, with a constrained n=2; see Figure 4.9). Furthermore, looking into the generation 8 results in more detail, we can see that the related model with equal fragmentation rate (*k'1* had been fitted to 1.61 - ID 94; vs 1.62 of the winning

model), therefore the result of the pruning stage is not unexpected. It seems thus that for the specific case of inhibitor modelling, in which some steps can be included but with equal or different rates to the protein-only counterpart step (therefore not always adding more optimisable parameters), it might be useful to modify the modelling procedure to include a 0.1 cut-off across the generation, to evaluate the winner of the generation against its related siblings of lesser complexity.

Table 4.4: Modelling results for the R3 + D-TLKIVW dataset. Best result for each model tested in the systematic modelling process of the inhibitor dataset. The winning models of each stage are shown in bold and the overall best model is shown in bold italicised. P1: First pruning. P2: Second pruning. P2 - ext: second pruning, testing of additional hypotheses. p: parameters, σ: perturbation in basin-hopping like optimisation. The number of parameters is reported as the sum of the 9 chemical rate constants of the R3-only mechanism (even if they are kept constant in the optimisation of inhibitor models), the chemical rate constants of steps involving inhibitor and the global or sample-specific frac/srac parameters.

ID	Gen.	Model	p.	χ^2	Parent?	Polish?	Guess	σ
1	1	a’	15	44.49	No	No	n	0.1
2	1	b’1	15	9.65	No	No	n	0.01
3	1	b’2	15	9.85	No	No	n	0.05
4	1	c’1	15	5.80	No	No	n	0.01
5	1	c’2	15	5.80	No	No	n	0.01
6	1	d’1	15	6.89	No	No	n	0.01
7	1	d’2	15	7.32	No	No	n	0.01
8	1	e’	14	45.94	No	No	n	0.01
9	**1**	**f’1**	**14**	**5.79**	**Yes**	**No**	**n**	**0.01**
10	1	g’2	14	42.58	No	No	n	0.01
11	1	g’1	14	24.25	No	No	n	0.1
12	1	h’	14	45.94	No	No	n	0.01
13	2	a’,f’1	16	5.79	No	No	i	0.01
14	**2**	**b’1,f’1**	**16**	**5.41**	**Yes**	**No**	**n**	**0.05**
15	2	b’2,f’1	16	5.53	No	No	i	0.01
16	2	c’1,f’1	16	5.79	No	No	i	0.2
17	2	c’2,f’1	16	5.79	No	No	n	0.05
18	2	d’1,f’1	16	5.65	No	No	i	0.01
19	2	d’2,f’1	16	5.64	No	No	i	0.01
20	2	e’,f’1	15	5.79	No	No	i	0.01
21	2	f’1,g’2	15	5.72	No	No	n	0.01
22	2	f’1,g’1	15	5.79	No	No	i	0.01
23	2	f’1,h’	15	5.79	No	No	i	0.01
24	3	a’,b’1,f’1	18	5.37	No	No	i	0.2
25	3	b’1,c’1,f’1	18	5.39	No	No	n	0.01
26	3	b’1,c’2,f’1	18	5.41	No	No	i	0.01
27	3	b’1,c’1,f’2	18	5.41	No	No	n	0.01
28	3	b’1,d’1,f’1	18	5.00	No	No	i	0.01
29	**3**	**b’1,d’2,f’1**	**18**	**4.87**	**Yes**	**No**	**i**	**0.01**
30	3	b’1,e’,f’1	17	5.41	No	No	i	0.01
31	3	b’1,f’1,g’2	17	5.41	No	No	i	0.01
32	3	b’1,f’1,g’1	17	5.41	No	No	i	0.01
33	3	b’1,f’1,h’	17	5.27	No	No	i	0.01
34	4	a’,b’1,d’2,f’2	20	4.87	No	No	i	0.2
35	4	b’1,c’1,d’2,f’2	20	4.45	No	No	n	0.01
36	4	b’2,c’2,d’2,f’1	20	4.87	No	No	i	0.01
37	4	b’1,c’1,d’2,f’2	20	4.91	No	No	i	0.05
38	4	b’1,d’2,e’,f’2	19	4.87	No	No	i	0.01
39	4	b’1,d’2,f’2,g’2	19	4.87	No	No	i	0.01
40	4	b’1,d’2,f’2,g’1	19	4.87	No	No	i	0.01
41	4	b’1,d’2,f’2,h’	19	3.89	No	No	n	0.01

Table 4.4 continued from previous page

ID	Gen.	Model	p.	Chi2	Parent?	Polish?	Guess	σ
42	4	b'1,d'2,f'2,i'1	18	4.76	No	No	i	0.01
43	4	b'1,d'2,f'2,i'2	20	4.62	No	No	n	0.01
44	4	b'1,d'2,f'2,j'1	18	4.82	No	No	i	0.01
45	4	b'1,d'2,f'2,j'2	20	3.63	No	No	i	0.05
46	4	b'1,d'2,f'2,k'1	18	4.12	No	No	i	0.01
47	4	b'1,d'2,f'2,k'2	19	4.05	No	No	n	0.1
48	4	b'1,d'2,f'1,j'3a	18	4.53	No	No	i	0.01
49	**4**	**b'1,d'2,f'1,j'4a**	**20**	**3.48**	**Yes**	**No**	**i**	**0.01**
50	4	b'2,d'2,f'1,j'3a	18	3.90	No	No	i	0.01
51	4	b'2,d'2,f'1,j'4a	20	3.61	No	No	i	0.05
52	5	a',b'1,d'2,f'1,j'4a	22	3.44	No	No	i	0.2
53	5	b'1,c'1,d'2,f'1,j'4a	22	3.47	No	No	i	0.01
54	**5**	**b'1,c'1,d'2,f'1,j'4b**	**22**	**3.02**	**Yes**	**No**	**n**	**0.1**
55	5	b'1,c'2,d'2,f'1,j'4a	22	3.47	No	No	i	0.01
56	5	b'1,c'2,d'2,f'1,j'4b	22	3.08	No	No	n	0.05
57	5	b'1,c'1,d'2,f'2,j'4a	22	4.89	No	No	n	0.05
58	5	b'1,c'1,d'2,f'2,j'4b	22	3.11	No	No	i	0.2
59	5	b'1,d'2,e',f'1,j'4a	21	3.47	No	No	i	0.01
60	5	b'1,d'2,f'1,g'2,j'4a	21	3.47	No	No	i	0.01
61	5	b'1,d'2,f'1,g'1,j'4a	21	3.43	No	No	i	0.2
62	5	b'1,d'2,f'1,h',j'4a	21	3.32	No	No	i	0.1
63	5	b'1,d'2,f'1,i'1,j'4a	22	3.46	No	No	i	0.01
64	5	b'1,d'2,f'1,i'2,j'4a	22	3.46	No	No	i	0.01
65	5	b'1,d'2,f'1,j'4a,k'1	20	3.24	No	No	i	0.01
66	5	b'1,d'2,f'1,j'4a,k'2	21	3.21	No	No	i	0.01
67	5	b'1,d'2,f'1,j'4a,j'1a	20	3.58	No	No	i	0.01
68	5	b'1,d'2,f'1,j'4a,j'2a	22	3.44	No	No	i	0.05
69	6	a',b'1,c'1,d'2,f'1,j'4b	24	2.98	No	No	i	0.05
70	6	b'1,c'1,d'2,e',f'1,j'4b	23	2.99	No	No	i	0.01
71	6	b'1,c'1,d'2,f'1,g'2,j'4b	23	2.96	No	No	i	0.01
72	**6**	**b'1,c'1,d'2,f'1,g'1,j'4b**	**23**	**2.69**	**Yes**	**No**	**n**	**0.05**
73	6	b'1,c'1,d'2,f'1,h',j'4b	23	2.97	No	No	i	0.01
74	6	b'1,c'1,d'2,f'1,i'1b,j'4b	22	2.87	No	No	i	0.1
75	6	b'1,c'1,d'2,f'1,i'2b,j'4b	24	2.87	No	No	i	0.01
76	6	b'1,c'1,d'2,f'1,j'4b,k'1	22	3.05	No	No	i	0.1
77	6	b'1,c'1,d'2,f'1,j'4b,k'2	23	2.98	No	No	i	0.2
78	6	b'1,c'1,d'2,f'1,j'4b,j'1b	22	2.98	No	No	n	0.01
79	6	b'1,c'1,d'2,f'1,j'4b,j'2b	24	2.96	No	No	i	0.1
80	7	a',b'1,c'1,d'2,f'1,g'1,j'4b	27	1.94	No	Yes	i	0.1
81	7	b'1,c'1,d'2,e',f'1,g'1,j'4b	26	1.98	No	Yes	i	0.01
82	7	b'1,c'1,d'2,f'1,g'2,g'1,j'4b	26	1.90	No	Yes	n	0.2
83	7	b'1,c'1,d'2,f'1,g'1,h',j'4b	26	1.92	No	Yes	n	0.01
84	**7**	**b'1,c'1,d'2,f'1,g'1,i'1b,j'4b**	**25**	**1.86**	**Yes**	**Yes**	**n**	**0.1**
85	7	b'1,c'1,d'2,f'1,g'1,i'2b,j'4b	27	1.85	No	Yes	i	0.05
86	7	b'1,c'1,d'2,f'1,g'1,j'4b,k'1	25	1.95	No	Yes	i	0.2
87	7	b'1,c'1,d'2,f'1,g'1,j'4b,k'2	26	1.91	No	Yes	i	0.2
88	7	b'1,c'1,d'2,f'1,g'1,j'4b,j'1b	25	2.17	No	Yes	n	0.1
89	7	b'1,c'1,d'2,f'1,g'1,j'4b,j'2b	27	1.91	No	Yes	i	0.01
90	8	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b	27	1.83	No	Yes	i	0.05
91	8	b'1,c'1,d'2,e',f'1,g'1,i'1b,j'4b	28	1.89	No	Yes	i	0.05
92	8	b'1,c'1,d'2,f'1,g'2,g'1,i'1b,j'4b	26	1.88	No	Yes	i	0.01
93	8	b'1,c'1,d'2,f'1,g'1,h',i'1b,j'4b	26	1.86	No	Yes	i	0.1
94	8	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'1	25	1.62	No	Yes	i	0.01
95	**8**	**b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'2**	**26**	**1.61**	**Yes**	**Yes**	**n**	**0.05**
96	8	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'1b	25	2.03	No	Yes	n	0.1
97	8	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'2b	27	1.84	No	Yes	n	0.01
98	**9**	**a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'2**	**28**	**1.52**	**Yes**	**Yes**	**i**	**0.1**
99	9	b'1,c'1,d'2,e',f'1,g'1,i'1b,j'4b,k'2	27	1.59	No	Yes	i	0.05
100	9	b'1,c'1,d'2,f'1,g'2,g'1,i'1b,j'4b,k'2	27	1.63	No	Yes	i	0.01
101	9	b'1,c'1,d'2,f'1,g'1,h',i'1b,j'4b,k'2	27	1.60	No	Yes	i	0.1
102	9	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'1b,k'2	26	1.85	No	Yes	i	0.05
103	9	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'2b,k'2	28	1.60	No	Yes	i	0.05
104	10	a',b'1,c'1,d'2,e',f'1,g'1,i'1b,j'4b,k'2	29	1.51	No	Yes	i	0.05
105	10	a',b'1,c'1,d'2,f'1,g'2,g'1,i'1b,j'4b,k'2	29	1.47	No	Yes	i	0.01
106	10	a',b'1,c'1,d'2,f'1,g'1,h',i'1b,j'4b,k'2	29	1.49	No	Yes	i	0.2
107	10	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'1b,k'2	28	1.89	No	Yes	i	0.05
108	10	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,j'2b,k'2	30	1.47	No	Yes	i	0.01
109	**10**	**a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,k'2**	**28**	**1.44**	**Yes**	**Yes**	**i**	**0.05**

Table 4.4 continued from previous page

ID	Gen.	Model	p.	Chi2	Parent?	Polish?	Guess	σ
110	10	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'4,k'2	30	1.45	No	Yes	i	0.05
111	11	a',b'1,c'1,d'2,e',f'1,g'1,i'1b,j'4b,i'3b,k'2	29	1.45	No	Yes	i	0.01
112	11	a',b'1,c'1,d'2,f'1,g'2,g'1,i'1b,j'4b,i'3b,k'2	29	1.49	No	Yes	i	0.01
113	**11**	**a',b'1,c'1,d'2,f'1,g'1,h',i'1b,j'4b,i'3b,k'2**	**29**	**1.42**	**Yes**	**Yes**	**i**	**0.05**
114	11	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,j'1b,k'2	28	1.89	No	Yes	i	0.2
115	11	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,j'2b,k'2	30	1.44	No	Yes	i	0.01
116	P1	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'2	26	1.58	No	Yes	i	0.05
117	P1	a',c'1,d'2,f'1,g'1,i'1b,i'3b,k'2	24	2.13	No	Yes	i	0.05
118	P1	a',b'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,k'2	26	1.58	No	Yes	i	0.01
119	P1	a',b'1,c'1,d'2,g'1,i'1b,j'4b,i'3b,k'2	27	1.93	No	Yes	n	0.1
120	P1	a',b'1,c'1,d'2,f'1,i'1b,j'4b,i'3b,k'2	27	1.59	No	Yes	i	0.2
121	P1	a',b'1,c'1,d'2,f'1,g'1,j'4b,i'3b,k'2	28	1.95	No	Yes	n	0.2
122	P1	a',b'1,c'1,d'2,f'1,g'1,i'1b,i'3b,k'2	26	1.91	No	Yes	i	0.2
123	P1	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'2	28	1.45	No	Yes	i	0.2
124	P1	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b	27	1.88	No	Yes	i	0.1
125	***P1***	***a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,k'1***	***27***	***1.41***	***Yes***	***Yes***	***i***	***0.2***
126	P1	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'3b,i'3b,k'2	26	1.86	No	Yes	n	0.01
127	P2	b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'1	25	1.62	No	Yes	i	0.05
128	P2	a',c'1,d'2,f'1,g'1,i'1b,i'3b,k'1	23	2.07	No	Yes	i	0.05
129	P2	a',b'1,d'2,f'1,g'1,i'1b,j'4b,i'3b,k'1	25	1.83	No	Yes	i	0.2
130	P2	a',b'1,c'1,f'1,g'1	23	4.24	No	Yes	i	0.2
131	P2	a',b'1,c'1,d'2,g'1,i'1b,j'4b,i'3b,k'1	26	1.82	No	Yes	i	0.1
132	P2	a',b'1,c'1,d'2,f'1,i'1b,j'4b,i'3b,k'1	26	1.58	No	Yes	i	0.05
133	P2	a',b'1,c'1,d'2,f'1,g'1,j'4b,i'3b,k'1	27	1.90	No	Yes	n	0.05
134	P2	a',b'1,c'1,d'2,f'1,g'1,i'1b,i'3b,k'1	25	1.90	No	Yes	i	0.2
135	**P2**	**a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,k'1**	**27**	**1.46**	**Yes**	**Yes**	**i**	**0.1**
136	P2	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'4b,i'3b	27	1.86	No	Yes	i	0.2
137	P2	a',b'1,c'1,d'2,f'1,g'1,i'1b,j'3b,i'3b,k'1	25	1.91	No	Yes	i	0.05
138	P2 - ext	a',b'1,c'1,d'2,f'2,g'1,i'1b,j'4b,i'3b,k'1	27	1.72	No	Yes	i	0.2
139	P2 - ext	c'1,d'2,f'1,g'1,i'1b,k'1	21	2.21	No	Yes	n	0.2
140	P2 - ext	a',b'1,c'1,d'2,f'1,g'1,i'1b,k'1	25	2.07	No	Yes	i	0.05
141	P2 - ext	a',b'1,c'1,d'2,f'1,g'1,k'1	25	2.99	No	Yes	i	0.1
142	P2 - ext	a',b'1,c'1,d'1,f'1,g'1,i'1b,j'4b,i'3b,k'1	27	2.19	No	Yes	i	0.05
143	P2 - ext	a',b'1,c'1,f'2,g'1	23	4.23	No	Yes	n	0.1
144	P2 - ext	a',b'1,d'2,g'1,i'1b,j'4b,i'3b,k'1	24	2.27	No	Yes	i	0.05
145	P2 - ext	a',b'1,c'1,d'2,f'1,g'1,i'1a,j'4a,i'3a,k'1	27	1.68	No	Yes	i	0.01

Finally, another feature of this modelling process that is a consequence of the steps that have rate constants equal to protein-only steps is that not all models across a generation need to have a lower χ^2 than their parent; this is true for many of the models with such steps and is reasonable since the parent model is no longer a subset of the steps of the children models, in terms of optimisable rate constants. For example, step *j'1* leads to consistently worse fits than the parent model in the models with IDs 67,96, 102, 107 and 114.

Discussion of the importance of each step and impact on the fit based on the results of the second pruning stage We can analyse the relative importance of each inhibitory action by looking into more detail at the second and final pruning stage of the modelling process, comparing the effects in χ^2 and systematic deviations that result from the removal of specific steps. This is shown in

Figure 4.20 and 4.21, with the steps ordered based on the impact of their removal on the fit, from smaller to larger. The removal of elongation by bound monomer leads to a marginally worse fit ($\chi^2 = 1.46$), with the same amount of parameters, which was not accepted. The fit had a reduced percentage of monomer that is bound to inhibitor to a maximum of 10 % which doesn't vary with time (see ii) vs i) of Figure 4.21). This suggests the effect of binding monomer is partially independent from its ability to elongate (e.g. has to do with preventing it from participating in reactions that involve free monomer, such as dimerisation and nucleations), which is expected since in terms of the elongation reaction it makes no difference whether it is free or bound, as it was happening with the same rates.

Similarly to what was discussed above for the first pruning stage winner, this fit of 1.46 could have been achieved earlier by the generation 9 winner (fitted to 1.52), since that had the fragmentation step with a free optimisable rate and therefore was a superset of steps. This would have required both optimising the fragmentation rate to the same value as the protein-only model, and reducing the amount of monomer-inhibitor complex, as seen with the 1.46 result.

The next closest model in χ^2 is the one without dimer dissociation from oligomers/fibrils (1.58), a step which was incorporated in generation 6 and led to minor improvements (2.69 vs 3.02), for which reason the polishing was turned on from generation 7. The removal of the binding to monomer (and elongation by bound monomer) has a similar small effect (1.62) and an equal fit to the same model that was the second best in generation 8, as has been discussed before (ID 94).

Removing either the direct disruption of nucleus or the binding to nucleus has similar effects in χ^2 with minor differences in the residuals that appear. This similarity in the fit agrees with the result of the first generation, in which the goodness of fit was very similar between the models including these two steps (5.79 for disruption, ID 9, 5.80 for binding, IDs 4/5, the latter having one more parameter). When the direct disruption is removed, even more of the nucleus is bound by the inhibitor (see Figure 4.21 v), another indication of the redundancy of these steps, discussed below in the section of custom tests.

Next comes a set of models with similar χ^2 worsening, all of which involve removing fragmentation or elongation steps of the inhibitor-coated fibrils, and most of which don't reduce complexity (other than the elongation by bound dimer with different rates to the protein-only fibrils, shown in ix)). Removing the elongation by bound dimer or forcing it to have the same rates as the protein-only model step has very similar effects in χ^2, yet different residual patterns, and different effects on the abundance of free/bound protein species (discussed below). Most of the losses in fit (but not all) affect the dimer traces; this suggests the contribution to a good fit from this step relies mainly on the slower elongation rates, which contribute to the larger and longer build-up of dimer.

Looking into them in more detail, removing fragmentation of the inhibitor-bound oligomers (vii) leads to a reduced amount of percentage of monomer bound to inhibitor (see Figure 4.21 vii). This is in agreement with the χ^2 of 1.86, the same as the winner of generation 7, which had two fewer parameters since it didn't include the binding of monomer step. At generation 8 the binding of monomer only led to a minor improvement to 1.83 (ID 90), so overall this seems to indicate a direct link between more elongation ends of bound oligomers being available and a greater amount (and contribution to the fit) of inhibitor-bound monomer. If elongation by free monomer is removed (viii), we might expect a higher proportion of bound monomer, which can elongate at the ends of inhibitor-bound oligomers. However we observe the opposite, together with a reduction of inhibitor-bound nucleus. It seems that in order for more free monomer to be consumed in elongation reactions by protein-only oligomers, the model has increased the amount of free nuclei, and the proportional decline in bound nuclei (and elongation ends of inhibitor-bound oligomers) is coupled to the lower abundance of inhibitor bound monomers. Also, we can see that removing elongation by free monomer has more impact than the binding to monomer or elongation by bound monomer, which might be due to the fact that in the final best model bound monomer is a minor species in comparison. The removal of elongation by bound dimer, or constraining its rates to those of the protein-only oligomers, has similar effects on χ^2 but leads to opposite readjustments

in the percentages of small species bound to inhibitor, which can be rationalised since the rates of elongation of protein-only oligomers are faster, meaning the pruning is having opposite effects on the fate of inhibitor-dimer complexes. If elongation is removed, there is a decline in inhibitor-dimer complexes which is needed to adjust the traces of this species, given it cannot be used up in elongation reactions. If the elongation of this complex is faster than in the best model, the model readjusts by reducing the amount of bound nucleus and thus available ends of elongation of bound-oligomers, which in turn affects elongation by bound monomer and leads to a marked reduced in the amount of the inhibitor-monomer complex.

Removing the binding to the dimer step has quite noticeable effects on the fits (mainly on the dimer traces) and also on the decays of free inhibitor; it seems to lead to a much stronger binding of the inhibitor to the large oligomeric/fibril species, especially in the samples with 100 µM R3. Theses large effects are in agreement with this step having been incorporated quite early in the modelling process (in generation 3). Finally, though, the biggest loss in goodness of fit comes with the removal of binding of inhibitor to oligomers/fibrils, which leads to similar patterns of systematic deviations to that earlier simple model with only nucleus disruption and dimer binding (see Figure 4.18 iii). It also has the consequence of keeping the inhibitor mostly in a free or bound to dimer state throughout the reaction.

Finally, we extended the second pruning stage with a few extra models in order to probe some open questions from the final winning model, which concern the interactions of inhibitor with the visible monomer and dimer species, the nucleus and the oligomers (see Figure 4.22 and 4.23).

There is some redundancy between the binding to and disruption of the nucleus by D-TLKIVW The models in Figure 4.22 B probe the interactions with the nucleus. The potential redundancy of the binding and direct disruption effects was confirmed by the model shown in i), which had a much larger impact on the goodness of fit. When both steps were removed, there were large systematic deviations from the dimer traces in all three samples, and some of which had not

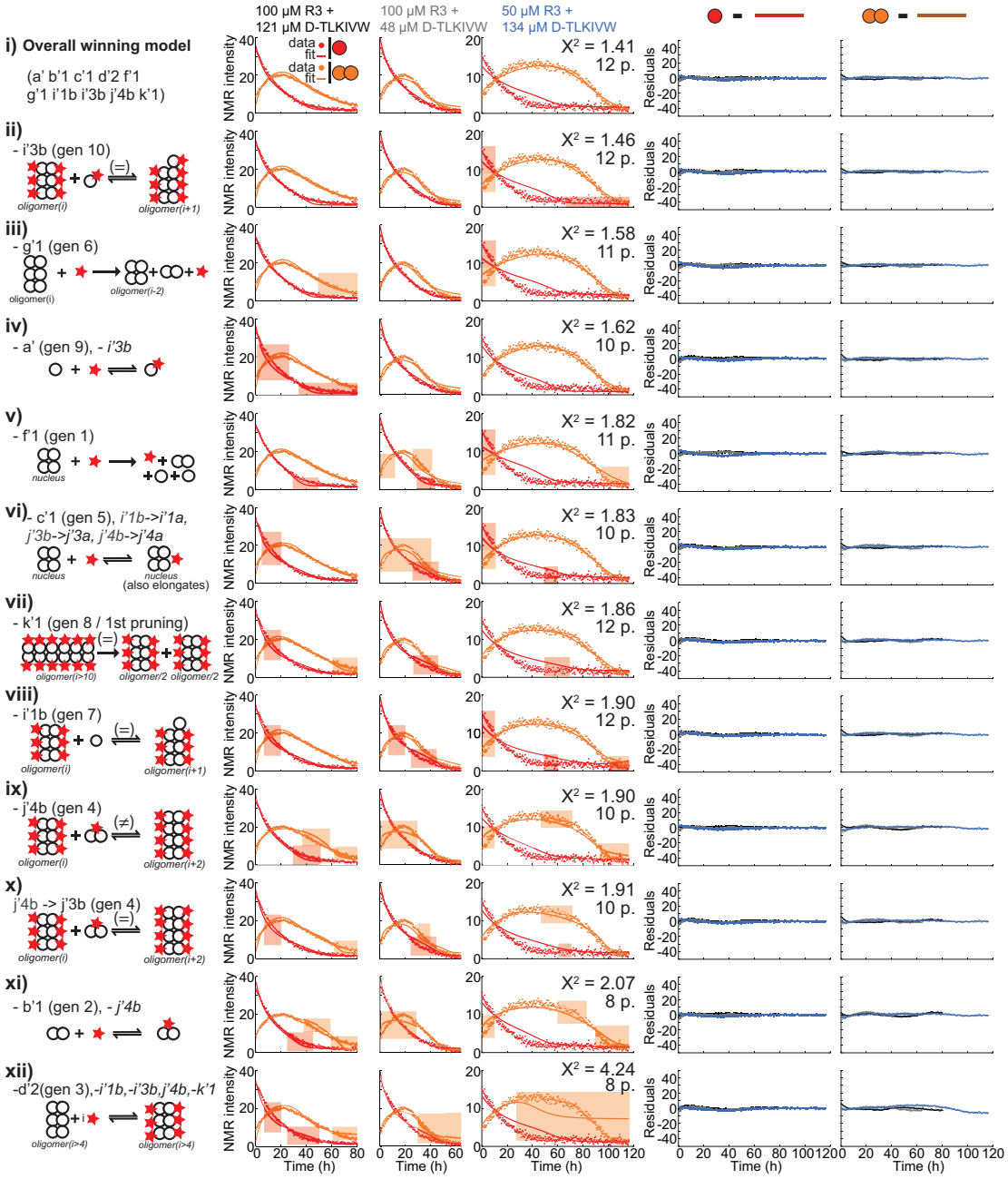


Figure 4.20: Final individual-step pruning stage of the modelling of inhibition by D-TLKIVW - Fits and residuals. Top row: winning model of the modelling process. From 2nd row to the last: models involving the removal of one step (depicted) from the winning model. Note that in some cases additional steps (indicated in italics) had to be removed, since they involved a species that was not present in the system anymore due to the removal of the specific step. The number of parameters refers only to rate parameters, excluding frac and srac. The models displayed correspond to entries with IDs 125, 135,132,127,131,129, 136,133, 134, 137, 128 and 130 in Table 4.4.

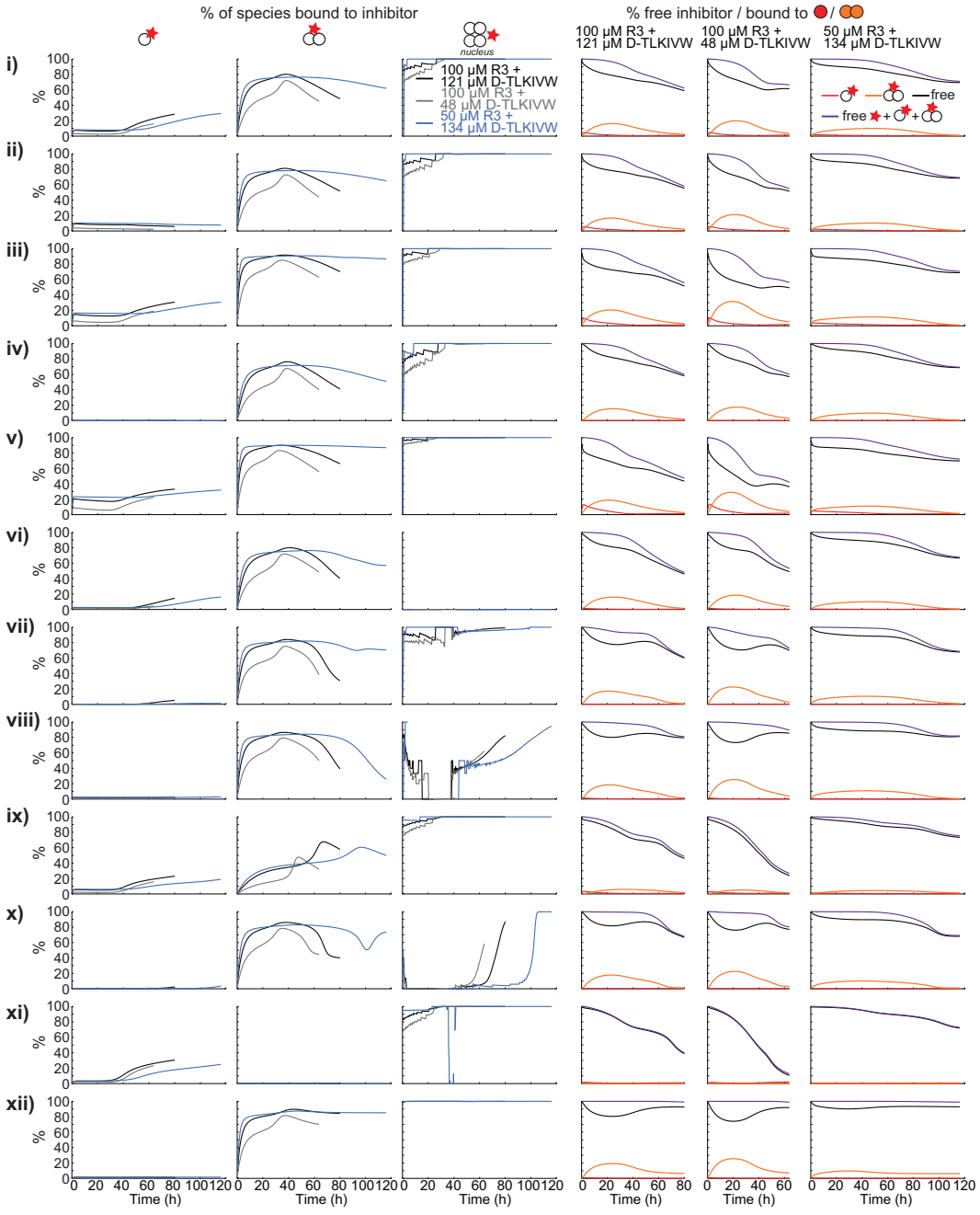


Figure 4.21: Final individual-step pruning stage of the modelling of inhibition by D-TLKIVW - Analysis of binding between small species and inhibitor. The models correspond to the same as in Figure 4.20. Left three panels. For each model, left to right: percentage of monomer bound to inhibitor, percentage of dimer bound to inhibitor, percentage of nucleus bound to inhibitor. Right three panels. For each dataset, distribution of inhibitor concentration amongst bound to monomer, bound to dimer, free, free+bound to either monomer or dimer.

appeared either of the single pruned models ($\chi^2 = 2.27$, see Figure 4.20 v and vi). There were also deviations in the monomer which were similar to the model without the disruption step. Both single pruned models had large changes in the amount of free inhibitor, by increasing its decay for both the 100 µM R3 samples, and in the combined pruning model this effect was even larger, in the case of the 100 µM R3 + 48 µM D-TLKIVW bringing the free inhibitor percentage to almost zero for the second half of the reaction. The percentage of monomer in a complex with inhibitor was also increased in the double pruned model, and similar to the model without direct disruption. Overall this confirms the partial redundancy of the steps, which was suggested by the previous results in which goodness of fit was lost with the removal of either, given the greater loss with the removal of both. It also indicates that when considered as the sum of two effects, the interaction of the inhibitor with the nucleus acquires higher importance as a mechanism of action, only surpassed by the the binding to oligomers.

When the bound nucleus is not able to elongate, there is a smaller loss of goodness of fit compared to the removal of the binding to nucleus (1.68 vs 1.83). Therefore the binding itself has a contribution to the fit, even if in the earlier generation 5, when that step was incorporated, it required the ability to elongate by bound dimer (the only elongation step for inhibitor-bound oligomers in that model) in order to improve the goodness of fit with respect to the parent. Unlike the case of binding to monomer, even if the bound nucleus is not able to elongate, most of the nucleus is in a complex with inhibitor (see Figure 4.23, compare the winner model in i) and the 1.68 fit in iii)). Bound nucleus elongates by free monomer or bound monomer at the same rates as free nucleus, therefore the effect might be either a stabilisation effect of the nucleus (can't dissociate while it's bound) or an effect due to the slower aggregation by dimer. The lack of elongation of the bound nucleus led to a reduction in the percentage of dimer bound to inhibitor, yet at the same time there is more binding of the inhibitor to the oligomers, as indicated by much reduced free inhibitor. These common readjustments amongst the models in Figure 4.22 i) and ii) can be understood as the model needing to create more elongation ends of

bound oligomers in order to support the consumption of the dimer as dimer-inhibitor complex, which elongates more slowly than the free counterpart on the inhibitor-free oligomers. Yet this is not achieved completely (even if in the case of 100 µM R3 + 48 µM D-TLKIVW, the free inhibitor drops to zero, similar to the previous model), and therefore leads to a reduction on the amount of dimer-inhibitor complex.

The model with an equilibrium two step mechanism of nucleus disruption (iii) has the same amount of parameters as the final best model, but a worse fit with increased systematic residuals mainly in the dimer traces. Also, the replacement of the independent steps by this equilibrium step in a model analogous to that shown in Figure 4.20 xii, in which binding of inhibitor to oligomers was removed, resulted in an identical χ^2 and fit in terms of systematic residuals and abundance of the inhibitor-protein complexes. In conclusion, it seems that the inclusion of both the nucleus direct disruption and binding steps, and as independent events affecting the nucleus, is justified by the results at multiple stages of the modelling, even if they do display a certain degree of redundancy in the final model.

At the late stage, surface-binding of inhibitor gives a much better fit compared to end-binding Another relevant test was between the surface and end binding models of elongation, given the rationale behind the design of this inhibitor as well as the EM data showing nanogold-labelled inhibitor bound to ends of fibrils[134]. Figure 4.22 and 4.23 C show that the end binding model has a clearly worse fit with more systematic residuals mainly in the dimer traces, and almost all the inhibitor would be free at the end of the reaction, in disagreement with the data in 4.16. It is important to note that the modelling did not incorporate constraints on the concentration of free inhibitor, which makes this agreement between observed decline (20-30% for the 100 µM R3 + 48 µM D-TLKIVW in Figure 4.16) and predicted decline remarkable. Overall, the evidence in our kinetic analysis supports a model with surface binding better than one with end-binding.

Custom pruning confirms the relatively larger importance of inhibitor binding to dimer versus monomer Finally, another set of tests was performed to probe in more detail the importance of the binding of the inhibitor to the small visible species, monomer and dimer, given the lack of clear evidence of binding both in our data (see 4.16) and in the study where this inhibitor was developed. In the single pruning stage, we had already seen, for the case of the monomer, that its removal had a moderate impact on the goodness fit (the third smallest impact), to a χ^2 of 1.61 (see Figure 4.20 and 4.21, (iv)), and that there was an even smaller loss of goodness of fit when removing the ability of the complex to elongate ($\chi^2 = 1.46$, see Figure 4.20 and 4.21, (ii)). In comparison, the removal of dimer related steps had a larger impact; the removal of binding of dimer had the second largest effect on goodness of fit in the pruning stage, only after the removal of binding to oligomers, with a $\chi^2 = 2.07$ and again, similarly to monomer, some contribution coming from the ability of the complex to elongate (at a different rate than the inhibitor-free oligomers, in this case, $\chi^2 = 1.90$, which is worse than removing binding to the monomer), as indicated by the loss of fit in Figures 4.20 and 4.21 ix and x, and some that is independent of the elongation of the complex. The relative larger importance of binding of dimer is also reflected in the population of dimer bound to inhibitor being majoritary and building up quickly from the onset of aggregation reactions, while the monomer remained mainly mostly free until at later times (and at which point its total concentration is greatly reduced).

We tested the effects of removing binding to both monomer and dimer, which yielded a substantially worse fit, with $\chi^2 = 2.21$ (see Figures 4.22 and 4.23 D ii), in particular with much larger systematic deviations observed in the dimer traces. This together with the previous results indicates that it can be interpreted as most of the effect coming from the absence of binding to the dimer, given the previous results above described. Interestingly, when the binding to both monomer and dimer is allowed but not the elongation by either complex, the systematic deviations are very similar and the χ^2 is only marginally better (2.07, see D iii); meaning that when evaluated jointly, most of the effects of the binding of monomer

and dimer are dependant on the complexes being able to elongate. In line with this, we probed the hypothesis that all inhibitor-bound species are not able to elongate and found that the fit is significantly affected ($\chi^2 = 2.99$, see D i)), yet the residuals that appear differ from the subset model without elongation by monomer or dimer inhibitor complexes. The ability of bound-oligomers to elongate by free monomer is an important feature of the best final model, confirmed by this custom test and already indicated earlier by the single removal of that step (viii). This also indicates that the binding of oligomers does have an important contribution to the good fit of the best model, but that a large part of that effect lies in the ability of the bound oligomers to still elongate. Overall, from the extended set of tests, we can deduce that it is not possible to obtain a well fitting model without binding to oligomers, binding to dimer, and the ability of bound oligomers and dimer-inhibitor complex to participate in elongation, while the binding of monomer (and the elongation of that complex) is more dispensable.

Kinetic versus spectral evidence for inhibitor binding to monomer and dimer As discussed above, the step of D-TLKIVW binding to the dimer (and to a lesser extent, to the monomer) contributes to the goodness of fit of the final model (Figures 4.20 and 4.22), and in particular the fits predict the dimer to be in a predominantly inhibitor-bound state (see Figure 4.19 xiii). In the earlier qualitative overview of spectral changes in the presence of the inhibitor, we did not observe CSPs for either the R3 species (histidine peaks) or the D-tryptophan and D-leucine peaks from D-TLKIVW, which was consistent with the literature [134].

However, the lack of CSPs does not necessarily imply a lack of binding. The R3 resonances correspond to the histidines, which are located in the C terminus, while the predicted binding site is at the N-terminal starting sequence VQIVYK. In addition to that, chemical exchange between multiple populations can occur in a intermediate regime in which either or both states are broadened out beyond detection. We therefore did not biase the modelling process into excluding these potential interactions, for which upon analysis of the results there is supportive

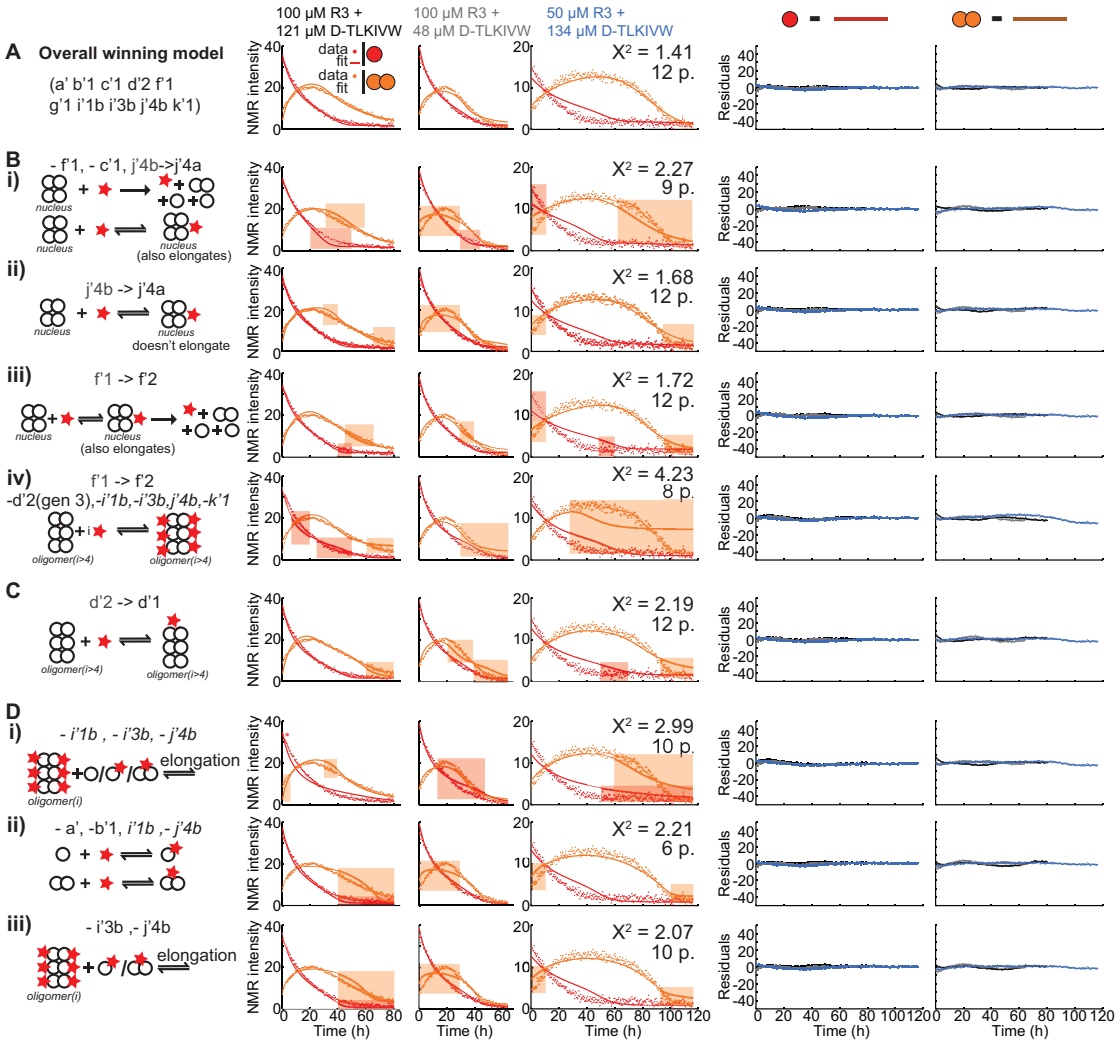


Figure 4.22: Custom testing of mechanistic hypothesis of inhibition by D-TLKIVW - Fits and residuals. Top row: winning model of the modelling process. From 2nd row to the last: models involving the removal of multiple steps or changes to specific steps. The number of parameters refers only to rate parameters, excluding frac and srac. The models displayed correspond to entries with IDs 125, 144,145,138,143,142,141,139 and 140 in Table 4.4.

kinetic evidence. Further work is needed to probe these predicted bound states and potentially obtain constraints for the modelling.

Discussion of the inhibition mechanism

D-TLKIVW is one of four inhibitors designed computationally on the basis of the crystal structures of the hexapeptide VQIVYK (the beginning of the R3 repeat and of our R3 fragment), in order to block the binding of other hexapeptides in both

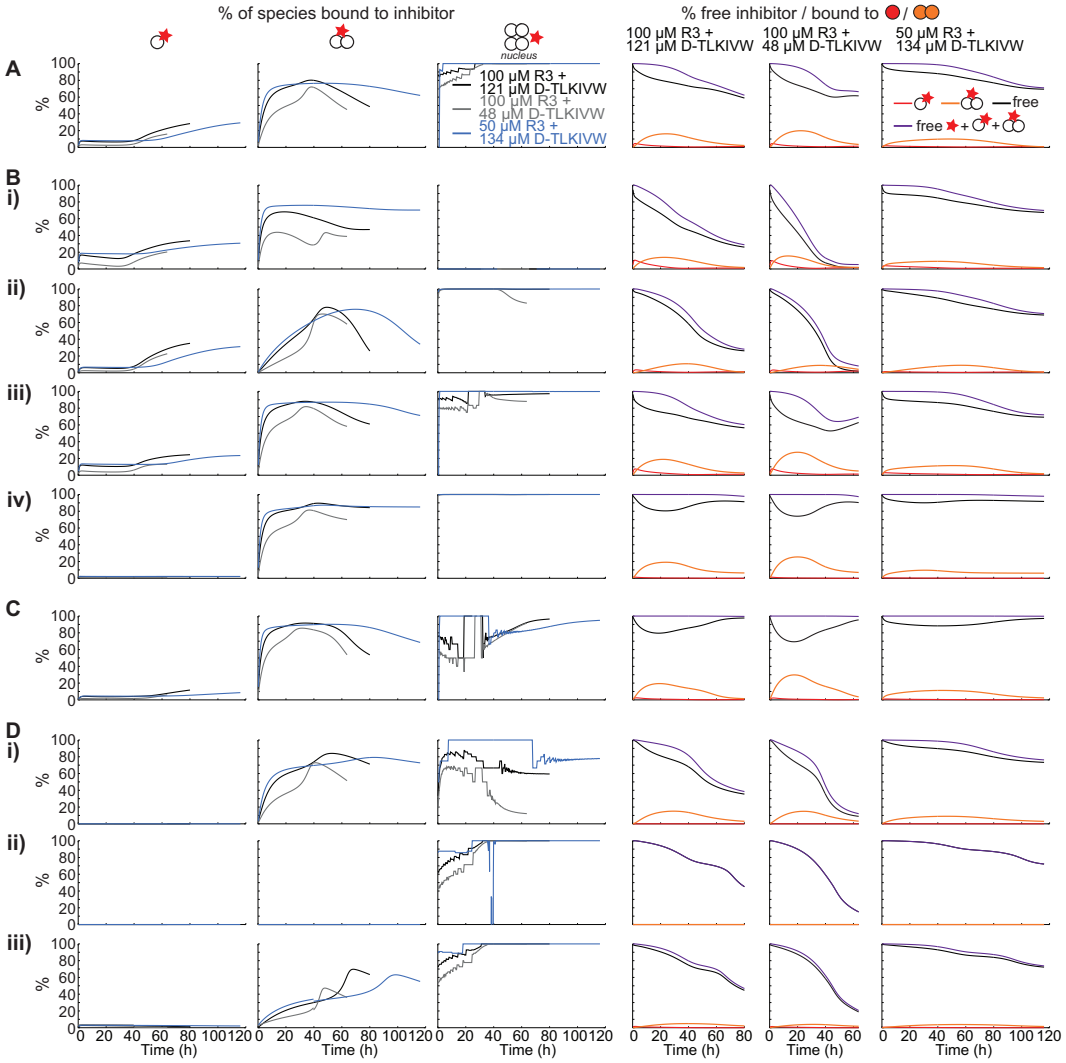


Figure 4.23: Custom testing of mechanistic hypothesis of inhibition by D-TLKIVW - Analysis of binding between small species and inhibitor. The models correspond to the same as in Figure 4.22. Left three panels. For each model, left to right: percentage of monomer bound to inhibitor, percentage of dimer bound to inhibitor, percentage of nucleus bound to inhibitor. Right three panels. For each dataset, distribution of inhibitor concentration amongst bound to monomer, bound to dimer, free, free+bound to either monomer or dimer.

the longitudinal (elongation) and horizontal (accross the steric zipper) axis, and acts on VQIVYK as well as three-repeat constructs K12 and K19 [134], blocking seeded and primary elongation. Their experiments were performed in reducing conditions (under which R3 doesn't aggregate on a timescale of days) and initiated by heparin. D-TLKIVW was found to bind fibrils with µM affinity and proposed mainly as an elongation blocker, even though a delay in the onset of primary aggregation was also observed. The results here offer a more complete picture of its action, allowing us to also assess the relative importance of each step.

4.4.3 Integration of the kinetic model with structural information

The tetrameric nucleus can be structurally understood in the context of the recent atomic resolution structures of Alzheimer's disease fibrils, which adopt mainly PHF but also SF morphologies[65] (see Figure 4.24A). The SF longitudinal interface is formed entirely of residues in the R3 repeat, which would be accessible to the R3 fragment. In this two-by-two model, the interface between monomers along the growth axis would likely generate a stronger interaction (perhaps aided by disulfide bridges, even if the distances between rows are too large for the bond to be present in the fibrils) compared to the interface across; oncse a dimer was formed, the less favourable binding of two monomers along that interface would constitute the rate-limiting nucleation step. The existence of polymorphisms could be attributed to nucleation happening along distinct interfaces for SF and PHFs; however this is subject to reorganisation as it has been shown that SFs convert to PHFs both *in vitro*[70] and in Alzheimer's disease patients [253]. Within this model, given enough structural resemblance between small species and fibrils (perhaps via partial sampling of more ordered conformations), it would be reasonable that D-TLKIVW could bind to either.

4.4.4 Discussion of the performance of the modelling algorithm

In chapter 3, we designed an algorithm to explore model space in a systematic way, with a pool of chemical steps, a building stage of generations of models of

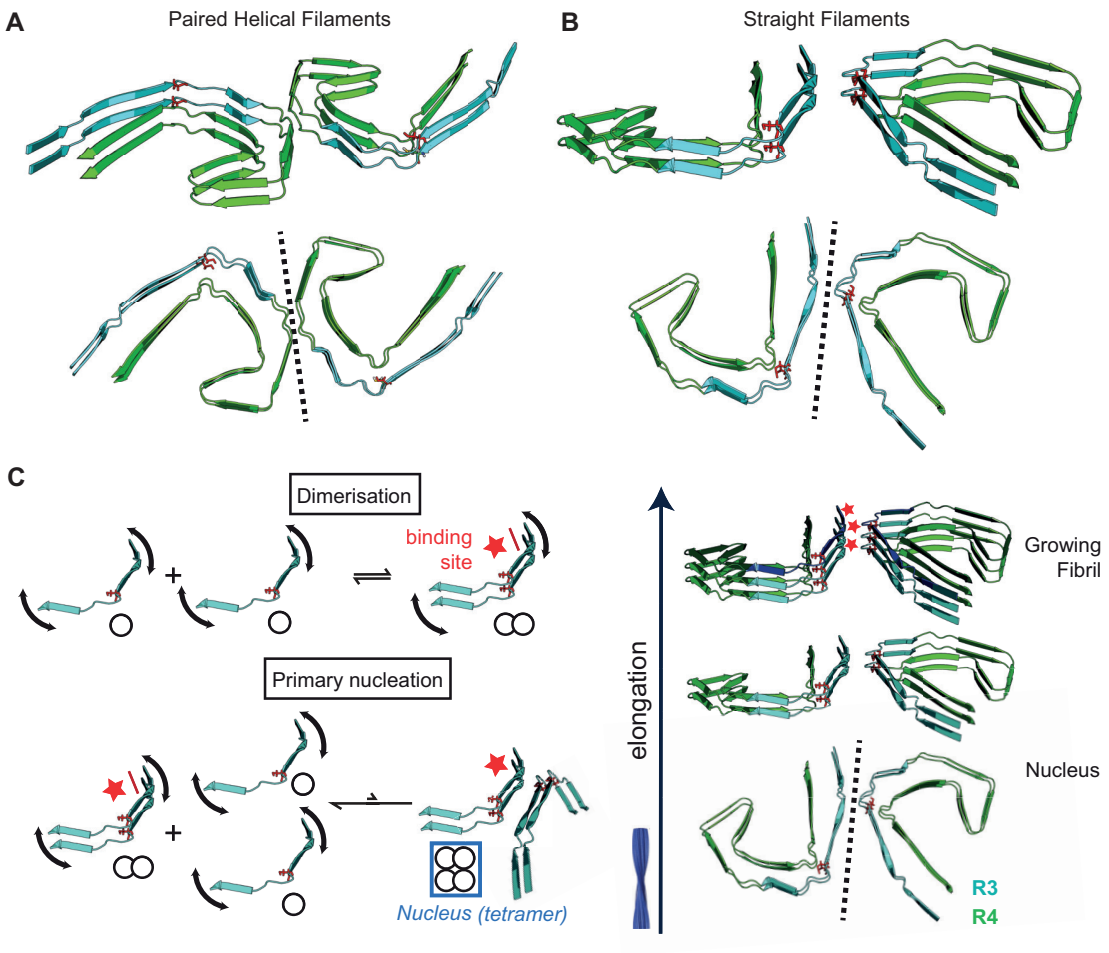


Figure 4.24: Structural model for the aggregation of R3. A. Paired Helical Filaments have an interface formed by residues in the R3 and R4 repeats. B. Straight filaments have an interface made entirely of R3 residues. C. Left: Illustration of the early steps in the aggregation mechanism. Arrows indicate that the monomeric and dimeric species are dynamic disordered species. A potential binding site for D-TLKIVW on the dimer and nucleus is indicated. Right: Illustration of the growing axis of the fibrils. The data represented corresponds to PDB accession numbers 5O3L (paired helical filaments) and 5O3T (straight filaments) and was displayed using PyMOL under an educational license.

increasing complexity and a final pruning stage to avoid overfitting. The application of this abstract framework allowed us to find the best fitting models over a very large set for both protein aggregation and the explicit inhibition by the inhibitor. There were a few instances, discussed above, where the optimiser (using basin-hopping like global search and local gradient descent) did not find the true global minimum. However, we believe that a certain degree of redundancy in our approach (i.e. some models being optimised twice, in the building-up stage and the pruning stage, as well as running parallel runs for the same model with multiple settings for some of the

optimiser) has been beneficial in dealing with this challenging optimisation problem.

An aspect worth discussing, in relationship to the results of chapter 3, is that of the accuracy of the numerical integration method used for LMA. Forward Euler is accurate when the time interval taken, dt, is small enough for the rates of evolution of the system. Since most of the optimisation runs were validated by the more accurate variable step multistep method, it follows that our heuristic of setting dt to $5 \cdot 10^{-}4$ was suitable for this system.

4.5 Conclusions

This chapter represents the culmination of the work described in the previous three chapters. We aimed to obtain a detailed step by step mechanistic understanding of the aggregation of a biologically relevant fragment of tau, as well as explicitly modelling the inhibitory actions of the literature pioneering D-peptide D-TLKIVW. With this aim, we applied the two-layer optimisation computational strategy described in chapter 3 (building on previous work[37] to the selected construct of tau, R3, characterised in chapter 2. We can see the overall mechanism of aggregation and inhibition by D-TLKIVW obtained from the first dataset in Figure 4.25.

Regarding the modelling process, we observed a few instances in which the lowest-layer optimisation protocol struggled to find the *true* global minimum, even when provided with a combination of optimisation settings regarding the parameter initialisation and perturbation stage of basin-hopping. Such cases involved the nucleation step for the protein-only mechanism (fitted best by the steady-state approximation of dimer and two monomers, but the best fit not being achieved in a superset of possibilities if the amount of monomers in that reaction is allowed to vary) and the fragmentation step of the inhibitor-coated oligomers.

However, in both cases the use of a top-layer optimisation algorithm that includes a certain level of redundancy (e.g. the same model with the same chemical reactions has been tested at multiple stages of the modelling process) enabled us to bypass these limitations and achieve a true global fit. In particular, the pruning stage (introduced to prevent overfitting) was instrumental in this sense.

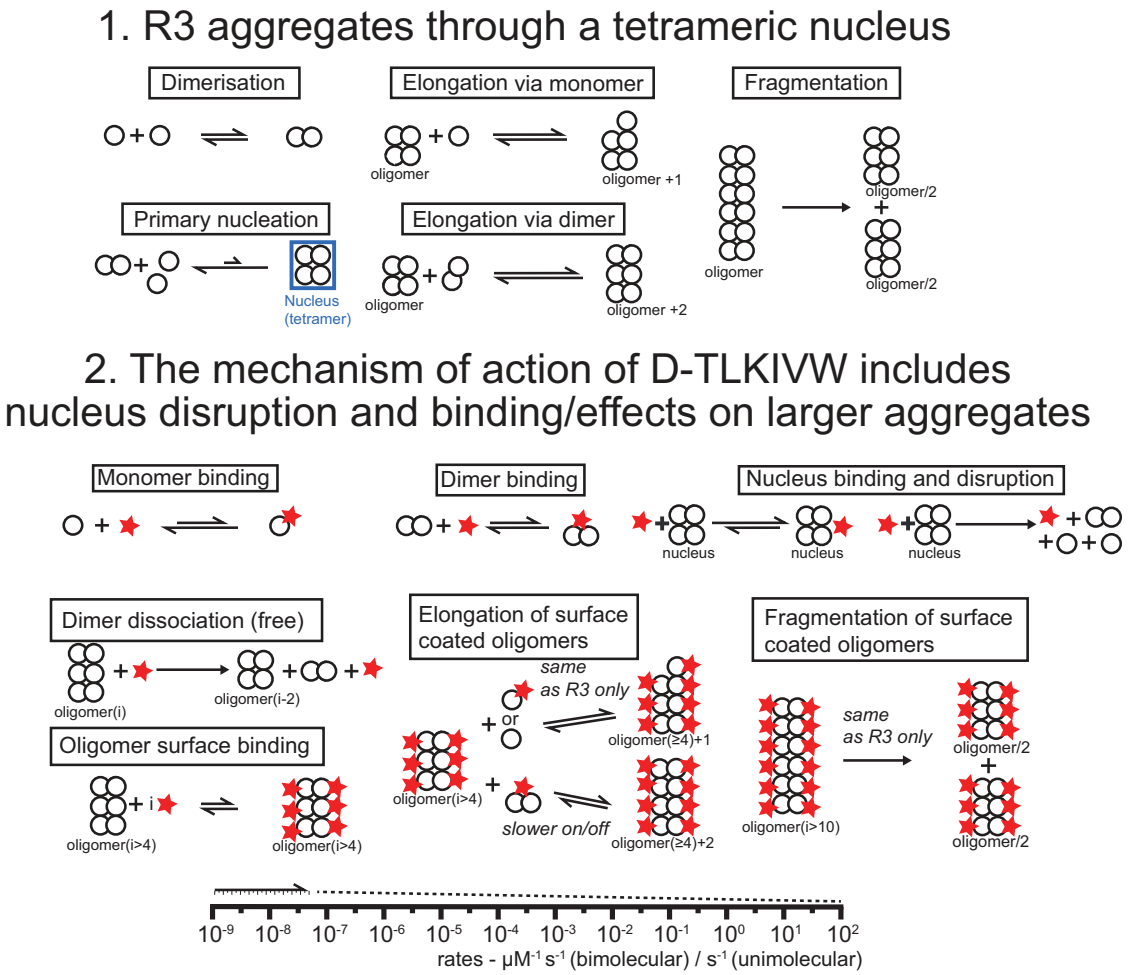


Figure 4.25: Overall mechanism of aggregation of R3 and inhibition by D-TLKIVW. Chemical steps of the final best model for R3 aggregation (1st dataset) and for the best inhibition model. These correspond to ID 25 in Table 4.2 and ID 125 in Table 4.4. The length of the arrows indicates the order of magnitude of the rate, discretised to the closest 0.5.

Additionally, in the case of the protein-only nucleation step, testing an exhaustive set of potential nucleation reactions after other steps had been optimised to good values allowed us to verify that the model obtained following the original strategy was indeed the best fitting of all.

A key feature observed during the modelling process was how much harder it was to fit well the dimer traces, in comparison to those of monomer. This was most striking in the comparison of nucleation mechanisms for the first protein-only dataset (see Figures 4.10 and 4.11), for which the dimer fits were particular informative

to distinguish the best fitting mechanism with confidence. This highlights the value of having found conditions in which both monomer and dimer species can be monitored independently and simultaneously, in particular given the trends in the literature of performing reactions in reduced conditions and/or modifying the cysteine residues for fluorophore labelling.

The best model for R3 aggregation included dimerisation, tetrameric nucleation by addition of a dimer and two monomers, elongation by monomer and dimer, and fragmentation. A small amount of a different conformation of monomer was essential for the fitting of the second dataset (in D_2O conditions) but was not detected kinetically in the first dataset, while all other steps in the model were the same. It is worth mentioning that in comparison with literature models [80][78], the nucleation reaction has been more precisely modelled, as attested by the sensitivity of the fits (especially in the dimer traces) to relatively small variations in that step. Additionally, the tetrameric nucleus identified from its kinetic fingerprint can be easily rationalised structurally from the recent cryoEM structures (i.e. for Alzheimer's, [65]). Finally, we have provided an explicit mechanism of action of the literature inhibitor D-TLKIVW which extends beyond its proposed role in elongation blocking and includes an essential step of nucleus disruption as well as more subtle effects on elongation of the bound fibrils and interactions with small species.

The models discussed above are of higher complexity than most literature models; however, the implementation of several criteria to prevent overfitting allows us to conclude they are minimally complex.

4.6 Further work

The main focus of this section of the thesis was to investigate the kinetics of tau aggregation. In the process, though, the results have revealed insights on other aspects of this system of study, as well as on the methodology used. I present below a few directions of interest for further research building on the current results.

Structural features of the monomer and dimer We have already gained structural insights, in the process of modelling the kinetics of R3, such as the observation that the histidine residues have a proton chemical shift unique to the monomer and dimer environments. This is interesting given that they are located in the C terminal end of the peptide, while the cysteine is at the middle of the sequence, which suggest the histidines (and other residues) could be involved in the dimer interface. We could probe transient contacts at the dimer interface using two-dimensional NMR methods on isotopically labelled R3. This seems feasible given the long timescales of aggregation, specifically in terms of the time during which there are substantial amounts of monomer and dimer.

Further to this, literature studies in a wide range of systems (from full length tau to the smallest hexapeptide models) and techniques (ensemble versus single-molecule) have shown evidence of aggregation prone transient conformations in the monomer. In our case, in the aggregation in D_2O, a unimolecular event was needed to fit the kinetic data well. The best fitting model included a unimolecular conversion that only generates a very small population of conformationally changed monomer for most of the nucleation stage, building to around 20 % at the end. If we could detect this distinct conformation using multidimensional NMR and/or exchange studies [231], we would both gain information on the structure and dynamics of this conformational change for the R3 model, as well as potential constraints to include in the kinetic modelling process.

Interactions of D-TLKIVW with monomer and dimer Our kinetic model of inhibition included binding of D-TLKIVW to the monomer and dimer species, which was important for the goodness of fit yet did not manifest in unambiguous signs of chemical exchange for these visible species. Higher-dimensional experiments and exchange experiments could help validate these interactions, as well as introduce fitting constraints, as in the case of the monomer conformational change.

Modelling strategy: optimising the process Finally, it proved very challenging to develop a modelling strategy to find a global minimum both in model space and parameter space, as discussed in Chapter 3 and throughout the description of the results in this chapter. The final implementation included parallel runs of optimisation for each model, with variable initialisation modes and perturbation levels: these were recorded not only in terms of final χ^2, but also other metrics such as the time taken for the optimisation, the number of iterations of the basin-hopping like routine and the evolution of the overall global minimum χ^2 at each iteration. While there were no obvious trends in the performance of distinct conditions, a more exhaustive analysis could help understand the performance of the modelling approach developed and potentially allow to *optimise* the optimisation process itself. Given the generalisable nature of our approach, this could be beneficial for the study of other tau model systems and other proteins within the protein aggregation field.

5

Partitioning of small molecules in biomolecular condensates

Contents

5.1 Abstract 188
5.2 Introduction 189
5.2.1 FUS 189
5.2.2 DDX4 193
5.2.3 Motivation for the partitioning experiments into DDX4 *in vitro* condensates 196
5.2.4 Nucleotide effects on phase separation 196
5.2.5 Literature evidence of chemical features influencing nucleotide-condensate interactions 197
5.3 Contributions 203
5.4 Methods 204
5.4.1 Protein expression 204
5.4.2 Partitioning of small molecules in condensates 205
5.4.3 Computational analysis 210
5.4.4 Comparison of methods to determine pellet volume . . . 212
5.4.5 Measurements of absolute partitioning of ATP, cAMP, GTP, cGMP and deoxyribonucleosides 213
5.5 Results and Discussion 216
5.5.1 Lipoamide partitions into *in vitro* FUS condensates . . 216
5.5.2 Development of a semi automated pipeline for analysis of partitioning: test case results 217
5.5.3 Partitioning into DDX4^{N1}NoCys is affected by the nucleobase (G vs A) and the triphosphate vs cyclic moiety of a set of nucleotides 218
5.5.4 Development of a PCV tube based method for determining partition coefficient 222

5.5.5 Comparison of partitioning trends of thiamine, GTP, cGMP, ATP and cAMP in DDX4 depending on construct and pH 225
5.5.6 Purine deoxyribonucleosides partition more than pyrimidines into DDX4^{N1} WT 227
5.5.7 Rationalisation of the partitioning trends as a reflection of protein-nucleotide interactions in the condensate phase 229
5.6 Conclusions 237
5.7 Further work 240

5.1 Abstract

In this chapter, I present results concerning the partitioning of small molecules into two proteins, FUS and DDX4. These are well studied model systems for liquid-liquid phase separation, with orthogonal properties in terms of salt dependency and interactions driving condensation. In the case of FUS, the finding that the molecule lipoamide (which disrupts stress granules in cells and affects FUS condensates *in vitro*) is ten-fold enriched into *in vitro* FUS-GFP condensates suggests a mode of action based on stronger interactions with FUS in the condensate phase. In the case of DDX4, we investigated the contributions of chemical features of nucleotides to the interactions with the condensed phase that drive their partitioning into *in vitro* condensates of the N-terminal domain DDX4^{N1}. The triphosphate moiety of nucleotides contributed positively to the partitioning, as seen by comparing triphosphate and cyclic versions of adenosine and guanosine nucleotides. The interactions were also influenced by the nature of the nucleobase, with partitioning trends decreasing in the order G > A > C ∼ U. This is discussed in the context of the main interactions that can be established between nucleotides and protein residues as well as findings from heterotypic condensates formed by RNA homopolymers and model peptides. The increasing evidence of molecular selectivity by the environment of condensates supports the hypothesis that phase separation leads to a distinct chemical solvent environment.

5.2 Introduction

Membraneless organelles (MLOs) carry out a range of essential functions in cells, in particular concerning nucleic acids. MLOs are permeable to small molecules in the cytoplasm [146] and several studies have suggested their environment constitutes a distinct chemical solvent to the surrounding bulk [254] [152]. Understanding the rules that determine partitioning or exclusion of small molecules into MLOs, both regarding small molecules and the protein/nucleic acid constituents of the MLO, would be beneficial in developing therapeutic strategies to modulate phase separation.

In this chapter, we have developed NMR-based methods for measuring partitioning coefficient of small molecules into *in vitro* condensates of two orthogonal well studied models of phase separation, the stress granule protein FUS and the nuage protein DDX4. The introduction below focuses on the known characteristics of the phase separation of these model systems. Further, given the frequent presence of nucleic acids in MLOs, and their modulatory role of phase separation, we have explored the influence of some features of nucleotides on partitioning. To aid the discussion of those results, I present a review of the literature concerning the effects of nucleotides on phase separation.

5.2.1 FUS

Overview

The FUS ("FUsed in Sarcoma") protein was identified as part of the oncogenic fusion protein FUS-CHOP (also called TLS-CHOP) in myxoid liposarcoma, and found to bind RNA *in vitro* [255]. FUS is a 526 amino acid protein encoded by 15 exons [256]. FUS contains a low complexity Prion-like domain (PLD), three RGG rich domains (one N-terminal, two C-terminal), an RNA binding region (RRM), a zinc finger motif, a nuclear export signal and an atypical PY nuclear localisation signal, which can be seen in Figure 5.1.

FUS is located both in the nucleus and the cytoplasm in most cell types (more in the nucleus in neurons, and exclusively in glia [257]), binds RNA *in vivo* in both

compartments and shuttles in between [258]. In the nucleus, it plays a role in the regulation of RNA transcription and splicing, as well as genome maintenance; a cytoplasmic function in mRNA transport has also been identified in neurons [259]. Mutations in the nuclear localisation sequence, located in the C terminus (last 18 residues), lead to impaired nuclear import[260] and excessive localisation into the cytoplasm, linked to incorporation into cytoplasmic stress granules [261] and which can be rescued by adding the wild-type C-terminus to the mutant protein [262].

After initial emphasis on its role in cancer, FUS has been linked to neurodegenerative disease, in particular amyotrophic lateral scleroris (ALS) [263][264] and frontotemporal dementia (FTD) [265],which is the second most common form of dementia after AD[266]. FUS mutations cause 3 % of familial ALS and 1 % of sporadic ALS; most affect the nuclear localisation sequence [267]. In contrast, autopsy-confirmed FTD cases with FUS pathology are not accompanied by mutations in FUS [268][269][270]. The disruption of RNA homeostasis has been proposed as a common pathogenic mechanism for both ALS and FTD [271].

Phase separation of FUS

FUS was shown to phase separate into liquid like droplets in cells and *in vitro* (see Fig 5.1 B) that show the characteristics of these compartments: rapid internal rearrangement, shown by recovery after photobleaching (Fig 5.1 Ci), a roughly spherical shape, shown with experiments with shear flow (Fig 5.1 Cii), and coalescence (Fig 5.1 Ciii) [147]. It has been shown that *in vitro* FUS droplets can "harden" and fibrillise over time and this is accelerated by the ALS-associated mutation G156E [147].

The phase separation of full length FUS is much more favourable than that of the PLD domain, with a saturation concentration two orders of magnitude smaller for full length[198]. The N-terminal PLD domain (1-165) is rich in tyrosines (27 occurrences) but contains no charged residues, and remains disordered, lacking stable secondary structure, in the phase separated state [157][159]. Its phase

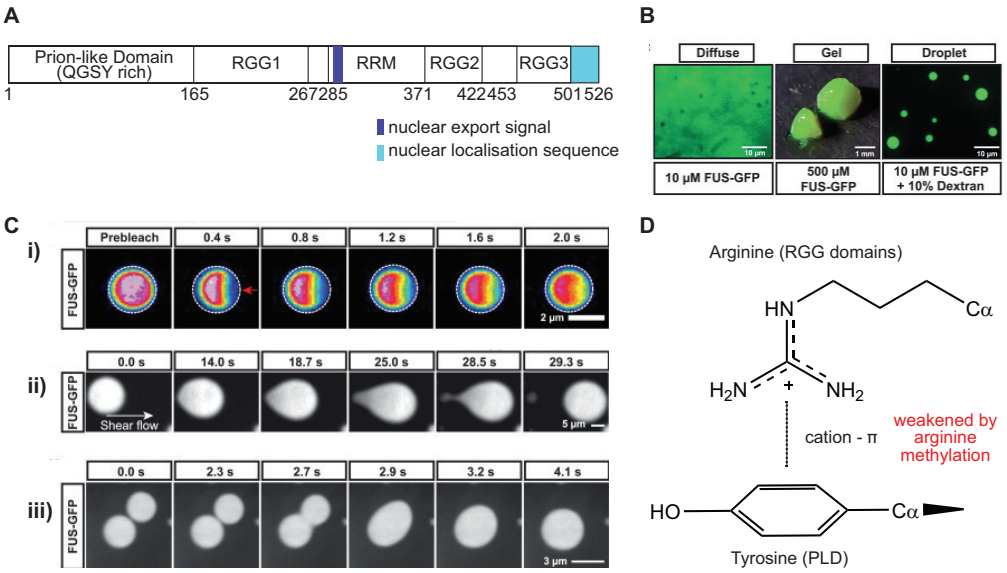


Figure 5.1: Overview of FUS phase separation. A. Domain organisation of FUS. PLD: prion-like domain (low complexity), RGG: arginine-glycine-glycine repeat domain, RRM: RNA recognition motif, NES: nuclear export signal, NLS: nuclear localisation sequence, ZnF: zinc finger domain. Adapted from [259]. B. Full length FUS can exist in diffuse, phase separated (droplet) and hydrogel states. Taken from [147]. C. Liquid-like behaviour of FUS: rapid internal rearrangement (i), spherical shape (ii) and fusion (iii). Taken from [147]. D. Cation-π interactions between N-terminal tyrosine residues and arginines drive the phase separation of FUS, which is downregulated by methylation.

separation is likely driven by tyrosine-tyrosine π-π interactions; in the case of full-length, the presence of both the tyrosine-rich PLD (there are no tryptophan or arginine residues in the PLD) and the arginine rich disordered RGG domains in the C-terminal domain (165-526) allows for cation-π interactions to be stronger drivers of phase separation [198]. This was supported by evidence from mutagenesis studies and by the finding that in the wider family of FUS proteins, the product of the number of tyrosine and arginine residues negatively correlated with the saturation concentration [198]. These cation-π interactions are fined-tuned by the methylation state of the arginine residues[194] (Fig 5.1 D). Electrostatic interactions also contribute to the phase separation of full length FUS, enhancing the aromatic interactions, and other residues, mainly glutamine, serine and glycine, modulate the dynamics in the condensate phase[198]. In addition, the phase separated and hydrogel states of FUS are stabilised by intermolecular β-sheet hydrogen bonds in the PLD domain[272]. Overall, the sticker-spacer model proposed for the

family of FUS proteins proposes that stickers (containing tyrosines and arginines) drive phase separation, while spacers (containing glycine, serine and glutamine) modulate condensate dynamics[198].

FUS phase separation is also modulated by salt concentration, showing phase separation at low ionic strength [147] with a reentrant phase at high ionic strength [201], temperature [273], protein concentration[194] and specific interactions with the nuclear import receptor, which was proposed to act as a chaperone for non-nuclear FUS [273][212][194]. Phase separation and aggregation of FUS are disrupted by phosphorylation of the PLD domain[274].

Modulation of stress granules and FUS condensates by lipoamide

The work presented in this chapter relates to a study aiming to test the druggability of MLOs by focusing in particular on stress granules [275]. Stress granules are cytoplasmic ribonucleoprotein (RNP) granules formed by phase separation, consisting of protein and RNA, which assemble upon the inhibition of translation initiation by small molecules or stress stimuli (reviewed in [276]). They contain a variety of proteins, in particular translation factors, protein chaperones and RNA-binding proteins, the latter including ALS and FTLD-associated proteins TDP-43 and FUS [277]. For several components of stress granules, disease associated mutations have been shown to affect *in vitro* phase separation of the isolated proteins [147] or *in vivo / ex vivo* of the entire stress granule compartment [278] [279], which is manifested as hardening, and in some cases fibrillar aggregation.

In the study this work relates to [275], lipoamide and the related lipoic acid were selected from a library of small molecules in a two step screening process; first, in cell systems they reduced the formation of stress granules and the localisation of FUS into these upon arsenate stress; second, they affected the phase separation behaviour of FUS *in vitro*. Importantly for its therapeutic potential, lipoamide was shown to act on stress granules in a specific way, not affecting other MLOs such as P bodies, Cajal bodies and the nucleolus. Lipoamide rescued the phenotype of ALS-associated FUS mutants *in vivo* in different systems (*Caenorhabditis elegans*,

Drosophila melanogaster and patient derived induced pluripotent stem cells). In terms of the molecular mechanism of action of lipoamide, in-cell NMR studies using ^{15}N-labelled lipoamide showed that lipoamide enters cells regardless of stress conditions, and is not metabolised. However, using standard protein-ligand probing techniques, including NMR and ITC, an interaction between FUS and lipoamide was not observed. In this context, we aimed to contribute to a more detailed mechanistic understanding of the action of lipoamide on FUS and stress granules, by taking the approach of probing interactions in the condensed phase through the measurement of its partition coefficient for FUS in *in vitro* condensates.

5.2.2 DDX4

DDX4 (sometimes referred to as VASA homolog) is a dominant component of the mammalian nuage[280], a conserved membraneless organelle that has an epigenetic role in male germ cells, functionally related to somatic cell P-bodies [281]. DDX4 contains a central DEAD-box RNA helicase domain[282], not essential for its phase separation either *in vitro* or in cells, surrounded by disordered N- and C- terminal regions [148]. Figure 5.2 shows the domain organisation of DDX4 and the features most relevant for its phase separation.

Features of DDX4 phase separation

The N-terminal domain construct (1-236; DDX4^{N1}) can phase separate on its own, which is driven by electrostatic interactions and cation-π interactions [148]. Charge-charge interactions occur between alternating charged blocks of 8-10 amino acids (see Figure 5.2 A-B) and cation-π interactions involve phenylalanine and arginine residues in FG and RG groupings, which have conserved spacings amongst DDX4 orthologues (Figure 5.2 D). The phase diagram of DDX4 is influenced by salt concentration (attributed to charge screening effects - see Figure 5.2 C), temperature (UCST behaviour) and arginine methylation (inhibitory). DDX4^{N1} is disordered in both the dilute and condensate phase and reaches concentrations of 300-400 mg/ml in the latter[156]. The translational diffusion of DDX4^{N1} is a hundred fold slower in

the condensate phase with respect to the dilute phase, yet local dynamics remain fast as seen by relaxation measurements of the amide backbone, and there are abundant interactions leading to Nuclear Overhauser Effects, mainly involving phenylalanine, arginine, lysine, leucine, threonine and valine. The diffusion of small molecule probes can be modelled with an excluded volume effect, while DDX4^{N1} is further slowed down consistently with the transient intermolecular interactions observed.

Chemical environment of DDX4

The existence of a chemical solvent environment in phase separated condensates, distinct to the surrounding bulk, is a feature hypothesized to emerge from the phenomenon of phase separation [148] [16]. In fact, this has been proposed as the biological criterion to distinguish condensates from protein dynamic hubs, with which they share multivalent fast exchanging weak interactions [152]. Effects on the rates and/or specificity of biochemical reactions have been observed in condensates of a range of systems [16], but there isn't yet a consensus on whether their environment constitutes a *true* chemical solvent.

The dielectric constant and molecular selectivity are two of the properties of a solvent that have been explored in the model system DDX4^{N1}. In terms of selectivity, DNA and RNA oligonucleotides differentially partition into DDX4^{N1} *in vitro* condensates, with their enrichment following an inverse correlation with stability [254]. Single stranded and compact structures are enriched in the condensates, while double stranded are excluded. In particular, double stranded DNA melts inside the condensates, with a denaturation effect comparable to 4 M guanidinium hydrochloride. In terms of the dielectric constant, the environment inside DDX4^{N1} condensates has been compared to polar solvents like DMSO and acetonitrile, given their values of 47 and 38, respectively, and values predicted to be 45 $\pm$ 13 [148] or even lower (20 [155]) for the DDX4^{N1} condensate phase. For reference, the dielectric constant of bulk water is 80, which means that there is a notable difference with the DDX4^{N1} condensate environment, even if its water content is estimated to be 75 % [156].

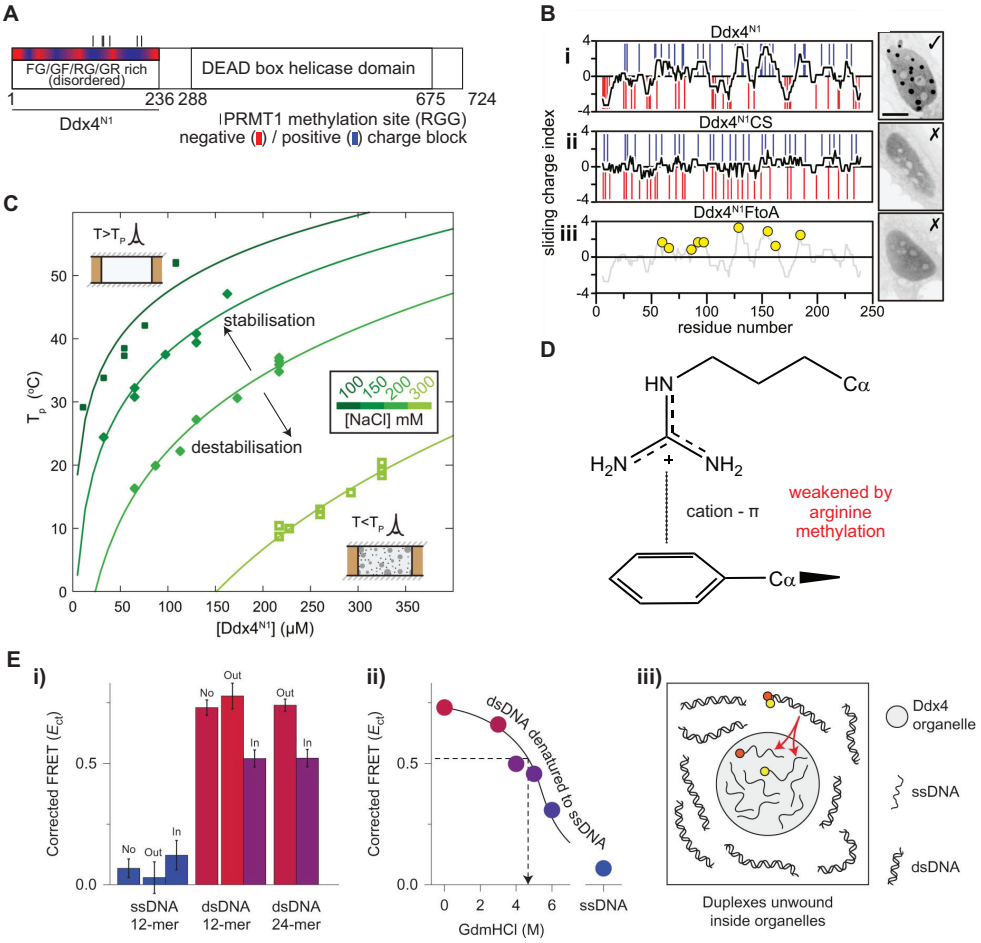


Figure 5.2: Overview of DDX4 phase separation. A. Domain organisation of DDX4. B. Charge-charge interactions between opposite charge blocks and cation-π interactions are required for phase separation of $DDX4^{N1}$, as indicated by the lack of condensate formation when charge is scrambled or when phenylalanine residues are mutated to alanine. C. The phase diagram of $DDX4^{N1}$ is affected by salt concentration. D. Cation-π interactions occur between phenylalanine residues and arginine residues. E. The condensate environment of $DDX4^{N1}$ melts double stranded DNA. B and C were taken from [148], E from [254].

5.2.3 Motivation for the partitioning experiments into DDX4 *in vitro* condensates

In order to further explore the hypothesis that the condensate phase constitutes a solvent, we studied the partitioning of a set of nucleotides into $DDX4^{N1}$, one of the best understood model systems for phase separation. This was designed within the framework of not only showing selectivity of DDX4 for these molecules, but also including a set of systematic tests to probe the chemical space around single nucleotides, given the importance of protein-nucleic acid interactions in phase separation. Through these, we aimed to find correlations between the features and the partitioning into condensates enrichment, potentially generalisable to chemical rules.

The small molecule set consisted of triphosphate and cyclic nucleotides (ATP, cAMP, GTP and cGMP) and the deoxyribonucleosides dA, dC, dG and dU. ATP, cAMP, GTP and cGMP are important metabolites in cells involved in cellular signalling, one of the functions related to phase separation: ATP is the cell's most important energy source and a phosphoryl group donor, cyclic AMP and cyclic GMP are second messengers [283] [284] and GTP is part of the signalling cascade of nitric oxide/cGMP. Furthermore, ATP has been increasingly studied as a regulator of protein stability in general as well as a LLPS modulator and protein stabiliser in a hydrolysis independent matter [285] [286], which is described in more detail in the next section. Finally, the set of deoxyribonucleosides was chosen to probe the influence of the nucleobase moiety on the interactions of nucleotides with the condensate phase of DDX4.

5.2.4 Nucleotide effects on phase separation

ATP was originally shown to inhibit the phase separation of full-length FUS, at mid millimolar concentration, as well as inhibit protein aggregation of amyloid forming proteins, leading to the hypothesis that a protein homeostasis action in this concentration regime could explain that cells maintain 2-8 mM concentrations of this metabolite, well above the necessary concentration for most ATP-utilising

enzymes [285] [287]. Further studies have shown a non monotonic effect of ATP on LLPS, as an enhancer at low concentrations and inhibitor at high concentrations, for a range of proteins including FUS [288] [289], the PLD of TDP-43 [290] and the SARS-CoV-2 nucleocapsid protein [291]. Similar biphasic action has been observed by RNA [292] and ssDNA [293]. Finally, the ALS-linked D169G mutation, in the RRM1 domain of TDP-43, previously poorly understood in terms of pathogenesis, has been shown to allosterically disrupt the binding of ATP to RRM1 [294].

ATP has also been shown to stabilise around 25 % of the insoluble proteome [286], inhibit fibrillisation of aggregation-prone proteins such as amyloid β [285] [295] and dissolve preformed aggregates of the ALS-linked FUS G156E mutant [147] [285]. ATP inhibits fibrillation of the TDP-43 RRM domain, with an affinity in the millimolar range for the RRM1 and RRM2 domains [296].

5.2.5 Literature evidence of chemical features influencing nucleotide-condensate interactions

Interactions between ATP and phase separating proteins

The induction of the phase separation of FUS-NTD by ATP was proposed to require cation-π interactions with arginine/lysine residues of the RGG1 domain [289]. The rationale for this was that the PLD, rich in tyrosines but devoid of charged residues, was not induced to phase separate by either ATP or ssDNA at much higher concentrations than required for the enhancement of the entire NTD domain (PLD+RGG1) or the CTD (see Figure 5.1 for domain organisation), suggesting that π-π interactions between the PLD tyrosines and the nucleobases in ATP/ssDNA were too weak in those conditions. Further to this, NMR titrations showed mainly chemical shift perturbations (CSPs) in arginine peaks and some lysines.

NMR titration studies of interactions of ATP with the prion-like domain of TDP43 showed that arginines were not required to be in an arginine rich domain for ATP binding; even though TDP-43 PLD is not RGG rich, its phase separation was modulated by ATP and analysis of CSPs showed specific binding of ATP to three N-terminal arginine residues (out of a total of five arginines), and subsequent

perturbations in neighbouring residues that were dependant upon the presence of the arginines [290]. Further, when the arginine residues were mutated to lysine, the CSPs of HSQC peaks remained unsaturated, indicating a lower affinity of ATP for lysine versus arginine.

In terms of general nonspecific protein binding, it has been shown that ATP could bind weakly to two folded proteins (mainly in the flexible loops) and α-synuclein as representative of a disordered protein; in the case of ubiquitin, an RGR sequence in an exposed loop was affected, suggesting coordination of the triphosphate group; in the case of synuclein, different types of residues were bound. [297]. It is worth noting that there are no arginines in the sequence of α-synuclein, and only one lysine out of 15 was affected, indicating a relative binding preference for arginine as in the study with TDP-43 [290].

Arginines have a prominent role in our understanding of phase separation, since many phase separating domains are RGG rich. In fact, an analysis of 22 phase separating proteins of the FUS family showed that the saturation concentration could be predicted from the product of the number of arginine and tyrosine residues, in an inversely proportional relationship [198]. Arginine to lysine mutations have been shown to reduce phase separation propensity in several systems [212] [298] [159].

Importance of the triphosphate moiety of ATP

For full length FUS, ATP inhibited phase separation at mid millimolar concentration, while ADP and AMP had lower inhibitory effects by a factor of 5 [285]. Similarly, for the N-terminal domain (PLD+RGG1), ATP had stronger effects in disrupting phase separation than AMP, while adenosine didn't show detectable effects at the solubility limit; their binding to FUS as shown by NMR titrations followed the same relative trends [289].

In the case of the PLD domain of TDP-43, which is the C-terminal low complexity domain of TDP-43, and is not RGG/RG rich, ATP vs ADP and AMP showed different effects on phase behaviour as well as binding/dynamics assessed by NMR titrations [290]. ATP biphasically modulated phase separation and induced chemical

shift differences (largely to arginines) that saturated at a 1:100 protein:ATP ratio. ADP showed similar CSPs as ATP, while AMP did not induce CSPs even at a 1:500 ratio. However, ADP did not show a biphasic modulatory effect, instead causing aggregation and precipitation in a monotonic way. At high excess of either ATP or AMP, HSQC signals were either broadened (ATP) or non-detectable (AMP), suggesting oligomerisation mediated by the nucleotides. The fact that the triphosphate group is important for modulating phase separation has been linked to its hydration properties: ATP (and ADP/AMP to a lower extent) have two types of water; constrained water, which moves slower than bulk water, and hypermobile water around the phosphates [299]; in a study of non-specific binding of ATP to three proteins, hydration of the proteins (in particular of α synuclein as a model for IDPs) was affected by ATP, as indicated by reduced protein-solvent exchange [297].

Finally, the triphosphate group was responsible for many CSPs measured for α-synuclein, as a model of an IDP, which was shown by ATP and adenosine affecting the same aromatic / hydrophobic residues (suggesting π-π interactions involving the base), with additional CSPs caused only by ATP (and similarly the other triphosphates tested), in aromatic, hydrophobic and polar residues [297].

Importance of the nucleobase moiety of ATP

After the discovery of ATP's modulatory action on phase separation, relatively little attention has been devoted to the action of the remaining nucleotide triphosphates. GTP was found to have similar or stronger inhibitory action on full length FUS, yet this was not considered very relevant due to the much lower concentration of free GTP in cells [285]. In the case of the ParABS partition system, the phase separation of the ParB component (which has CTPase activity) is positively modulated by CTP but destabilised by CDP, ATP, GTP and UTP, which requires the integrity of the CTP-binding site and has been explained as CTP-induced oligomerisation that favours phase separation [300].

In terms of non specific protein binding of nucleotides to α-synuclein as a model IDP protein, the four studied triphosphates (ATP, GTP, CTP and UTP) had similar binding, in terms of direction and magnitude of the CSPs in the HSQC [297].

Neither study is informative on nucleobase trends for general non-specific interactions with residues relevant for phase separation proteins; in the former, a highly specific binding-site mediated effect was reported, while in the latter the model protein chosen did not contain any arginines and the data was not analysed in terms of distinct amino acid types, such as aromatic residues capable of π-π interactions.

However, there are studies of RNA homopolymer-peptide/protein coacervates which reveal trends that could be relevant for the single nucleotides, and which further highlight the differences between arginine and lysine residues in terms of phase separation. These are discussed in the next section.

Insights from polynucleotide - protein LLPS The study of reentrant phase separation behaviour of coacervates made of polyadenine(polyA)/polyuridine(polyU) and R-rich/K-rich peptides, in which the polynucleotides induced phase separation at intermediate concentrations, inhibiting it at the high concentration regime, provide insights on the relative strength of interactions on both the nucleotide and protein sides[298]. The decondensation boundary for both R-rich and K-rich peptides was higher (i.e. the condensates stayed stable until higher polynucleotide concentrations) for mixtures with polyA with respect to polyU, and the effect was stronger for R-rich peptides, which overall showed wider regions of demixed state in the phase diagrams compared to the K-rich peptides. Also, the viscosity of the condensates depended on the combination of R/K rich peptide and polyA/polyU nucleic acid, with polyA-R-rich condensates being the most viscous. Additionally, a construct of FUS encompassing its R-rich domains had slower dynamics in mixtures with polyA compared to polyU. All these effects suggested that the strongest interaction was between the purine base polymer polyA and arginines, in either the R-rich peptide or the FUS construct.

In another study, PR-repeat (PR) peptides had greater affinity for, and higher partitioning in, polyA versus polyC and polyU constructs (polyC ≈ polyU), which was attributed to the increased aromatic interaction strength of the double ring versus the single ring bases [301]. Interestingly, PR peptides had the lowest affinity for polyG RNA, which was attributed to the formation of very stable G-quadruplexes. polyA, polyC and polyU all could induce the formation of PR-homopolymeric RNA coacervates, while polyG yielded non spherical morphologies. Similar to the results of Alshareedah *et al.*, the polyA-PR condensates had higher viscosity and slower dynamics. However, unlike the previous study, differential effects of the RNAs were not observed for PK-repeat peptides, which was attributed to the additional aromatic interactions enabled by the additional dipole and quadrupole moments of the guanidinium groups of arginine being most important. In particular, π-π interactions between the nucleobases and arginine were proposed to be the source of the ability of distinct RNA homopolymers to tune the dynamics of the PR-droplets but not of PK-droplets. This was further supported by the finding that the FUS LCD, rich in tyrosines but without cationic residues, was preferentially partitioned in polyA versus polyC condensate subcompartments. However, hnRNPA2 (less rich in tyrosines, and which contains arginines) did not show a preference, which in that case indicated a degree of complexity in the rules governing these interactions higher than for the "simple" R-rich or K-rich model systems.

In conclusion, the current evidence from two homopolymeric polynucleotide-model peptide coacervate studies indicates that purine nucleobases can form stronger interactions with cationic residues than pyrimidine bases, and in particular with arginine [298][301]. Due to the formation of G-quadruplexes from polyG, the only comparison between guanine and adenine bases is in the context of ATP vs GTP and indicated similar interactions with full length FUS, even though it was not probed extensively [285].

Requirement of the nucleobase of ATP for its effects on proteins

There have been several studies aiming to deconvolve the contributions of the adenine and the triphosphate (TPP) moieties to the diverse actions of ATP on protein stability, phase separation and aggregation.

ATP was proposed as a hydrotope [285], a type of molecule that can solubilize hydrophobic substances in water and is routinely used as such in the chemical industry [302]. However this has been questioned; a thorough analysis of hydrotope behaviour led to the conclusion that two different activities coexist in ATP: a kosmotropic Hofmeister salt (the triphosphate group) and a fibrillisation inhibitor (the adenosine group) [295]. In the former case, the salting-out action and inhibition of the aggregation of thermally denatured proteins effects were sensitive to the presence of triphosphate, but not the base, and proposed to stem from the reduction of availability of the exposed hydrophobic residues, which is in agreement with NMR studies showing reduced protein-solvent exchange for flexible loops of folded proteins and widely for disordered α-synuclein [297]. In the latter case, concerning effects on fibrillisation, ATP and adenosine had more similar effects than ATP and sodium triphosphate, but a role for the triphosphate could not be ruled out [295].

The description of ATP as the combination of two groups is helpful in reconciling some of the contrasting observations as to whether the nucleobase was required for the observed effect, such as the lower inhibitory effect of TPP versus ATP for phase separation of FUS, and lack of inhibition by TPP of fibrillisation of amyloid β 42 [285] [295], yet a similar action in stabilising folded proteins [295] [303].

However, some features remain unexplained, such as the requirement for the nucleobase for arginine-rich peptides to have a higher decondensation boundary than lysine-rich [298], since in that study the boundary effect was not present when using polyphosphate, in contrast with the finding that the phase diagram of a series of mutants of IDP histatin with TPP was sensitive to the arginine/lysine content [304].

What is the mechanism of action of nucleic acids on phase separation?

Taken together, it is clear that nucleic acids interact with and modulate a range of phase separating systems, which lead to contrasting behaviours both *in vitro* and *in vivo*. The mechanism by which this is accomplished is unclear, and we propose to address it at the level of the fundamental building blocks: nucleotides. There are some literature studies on the relative contribution of the nucleobase and the triphosphate moieties of ATP to its various effects on phase-separating proteins, as well as studies of heterotypic polyRNA-peptide systems, which I have summarised above, yet no systematic comparison of binding or partitioning trends of single nucleotides with different nucleobases. This is one of the focuses of this chapter, within the broader aim of exploring the distinct solvent environment of condensates *via* their selectivity for small molecules. The finding that purine bases are enriched with respect to pyrimidines is discussed in the context of trends in nucleic acid-protein non covalent interactions.

This research required the development of novel NMR-based methods tailored to the specific challenges of studying phase separation systems *in vitro*, and particular attention has been given to methods to measure the volume of the condensate phase, which is required to calculate absolute partition coefficients. Their development is described in detail in the following Methods section and throughout the results obtained using them.

5.3 Contributions

Some of the work presented in this chapter was performed in collaboration with or by others than the author of this thesis; these contributions are acknowledged below. ^{15}N lipoamide was synthesized and provided by Satoshi Kishigami, who had previously optimised the conditions for NMR experiments. Expression of DDX4^{N1}NoCys and the design and setup of the relative partitioning experiments with ATP, cAMP, GTP, cGMP and thiamine were performed by Spencer A. Cahill; the NMR data acquisition for that dataset was done by Charlie Buchanan. The

development of condensate volume measurement methods was carried out in collaboration with Richard Wheeler (microcentrifuge tube method) and James Eaton (PCV-based method). The design and experimental setup of the partitioning experiments of DDX4^{N1} deoxyribonucleosides were performed together with James Eaton, and the NMR data was acquired by James Eaton.

5.4 Methods

5.4.1 Protein expression

FUS-GFP was obtained as part of a collaboration with the laboratory of Anthony Hyman (Dresden, Germany). The N-terminal construct DDX4^{N1} (residues 1-236) wildtype (WT) and the cysteine mutant DDX4^{N1}NoCys (containing four Cys->Ser mutations) were expressed as fusion proteins with an N-terminal TEV cleavable SUMO-His tag (pET-SUMO vector). Expression was done in BL21-Codon Plus (DE3)-RIL competent cells (Agilent), which contain extra copies of tRNA-encoding genes in order to compensate for the different distribution of tRNAs between *Escherichia coli* and human cells, requiring antibiotic selection with chloramphenicol (for the tRNAs plasmid) as well as kanamycin (for the DDX4 plasmid). Standard protein expression procedures were followed, with expression induced by 1 mM IPTG for 18-24h at 22 ºC, with agitation. Cells were spun down (soft spin) at 4000 g and 4 ºC for 15 minutes and the pellets resuspended in pre-cooled lysis buffer (20 mM sodium phosphate, 6 M guanidinium chloride, 10 mM imidazole, pH 8). Lysis was done using a microfluidiser with three passes. The solute fraction (with DDX4) was separated from all other fractions from cell lysis via centrifugation at 16500 rpm and 4 ºC for 1h. Purification was done with a combination of nickel-based affinity purification, on-column TEV cleavage, inverse nickel-based purification (since the cleaved DDX4 protein does not have a His tag anymore) and size exclusion chromatography. Size exclusion chromatography was performed to final buffer containing 20 mM sodium phosphate and 300 mM NaCl (to keep the protein at dilute phase), at pH=8, with additional 5 mM TCEP only for the wild type protein.

to prevent aggregation. After lysis, all purification, cleavage and concentration steps were performed at room temperature to prevent phase separation.

5.4.2 Partitioning of small molecules in condensates

FUS-Lipoamide

Phase separation set-up In order to measure the partitioning of a small molecule in biomolecular condensates, we designed the workflow depicted in Figure 5.3. Briefly, phase separation was performed in a 1.5 ml microcentrifuge tube with known bottom internal radius.

Phase separation of FUS-GFP, at room temperature, was achieved by diluting 12.5 µl of protein stock at 170 µM concentration (in salty HEPES buffer – 50 mM HEPES, 500 mM KCl and 5% glycerol, at pH 7.25, DTT 1 mM) with 247 µl of salt-free buffer containing ^{15}N-(±)-lipoamide (50 mM HEPES, 5% D_2O, 105 µM [^{15}N]-(±)-lipoamide, 1.05% DMSO, at pH 8), resulting in samples of 260 µl with 8 µM FUS-GFP, 100 µM [^{15}N]-(±)-lipoamide, 1 % DMSO and 25 mM KCl.

The sample was then centrifuged using a bench centrifuge at a low speed (4000g) for 10 minutes at room temperature. Once there was a visibly separate condensate phase, referred to as the pellet, photographs were taken with a smartphone for volume measurement (explained in more detail in the next section). Then the supernatant was carefully removed from the tube and placed in another tube to make up the dilute phase NMR sample. This was done in multiple steps, removing first most of the volume and later only a few µl which were discarded from the subsequent sample (the incomplete recovery of supernatant was accounted for in the calculations later). Then the pellet was resuspended in the same buffer corresponding to the conditions of phase separation in order to make up the NMR sample, for comparison of the signal of supernatant and pellet by NMR, as explained below.

FUS-GFP (+ lipoamide) volume fraction measurements After phase separation of the sample and centrifugation, perpendicular view pictures of the condensate phase were taken using any regular smartphone; the wall width was

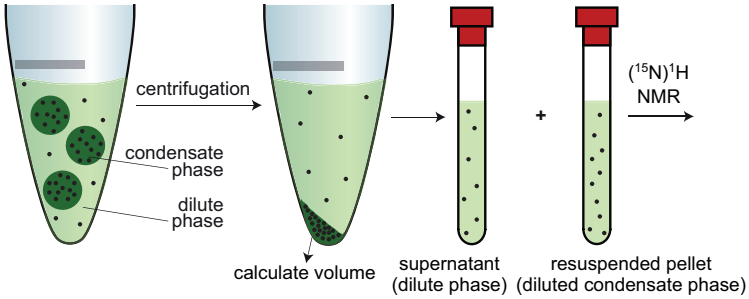


Figure 5.3: Overview of the experimental protocol for partition coefficient determination. The protein is depicted in dark and light green for the condensate and soluble phase, respectively.

not visible, which implies that the view of the pellet was refracted by the walls, projecting the inner diameter onto the outer (see Fig 5.4). The volume of condensate was calculated as the volume of a spherical cap, based on measurements of the pellet circular radius (a) and previous knowledge of the inner radius of the semi-spherical bottom of a 1.5 ml the microcentrifuge tube (r) (which gives the scaling for the pellet radius a), by applying the following equations:

$$V = \frac{\pi}{3} r^3 (2 + cos\theta)(1 - cos\theta)^2 \tag{5.1}$$

$$\theta = sin^{-1} \frac{a}{r} \tag{5.2}$$

where a = pellet circular radius, r = microcentrifuge tube inner radius and θ = subtended angle.

NMR intensity measurements The resuspended condensate or supernatant were loaded in D_2O-matched 5 mm Shigemi tubes and analysed by ^1H detected ^{15}N edited ^1H sensitivity enhanced HSQC ($(^{15}N)^1H$) at 10 ^oC. Spectra were acquired on a 14.1 T Varian Inova spectrometer equipped with a 5 mm z-axis gradient triple resonance room temperature probe. Previously, the conditions of NMR measurement had been optimised (by Satoshi Kishigami, a collaborator in this project). At 10 ^oC and pH below 8.6, the signal of the trans-amide proton showed minimal exchange and was stable over 10 hours [275].

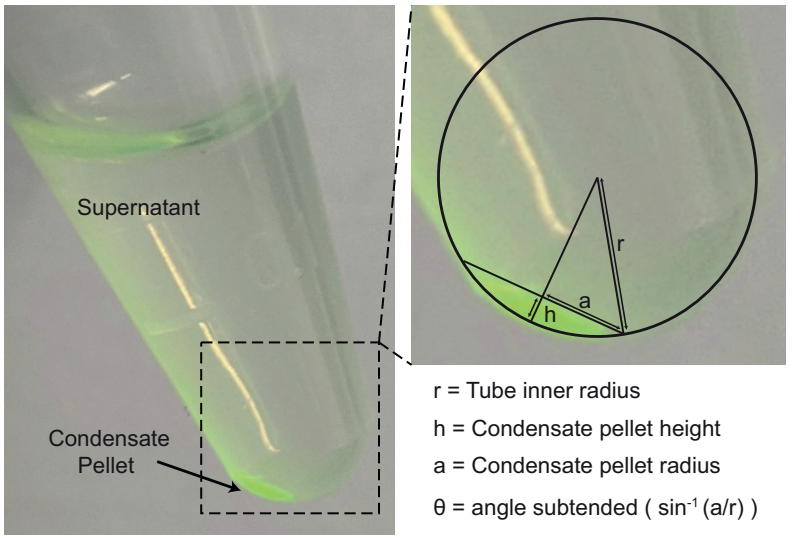


Figure 5.4: Volume fraction measurement. The microcentrifuge tube inner diameter was projected onto the outer in this view; therefore the known r value was taken as the inner diameter, and the measurement on the image gave the scaling needed to relate measured a to real units a.

The free induction decay was recorded for an acquisition time of 0.0624 s and a sweep width of 8 kHz, recorded over 1000 points and with a recovery delay of 1 s. To achieve adequate signal to noise, the resuspended condensate was scanned for 20 h (using 5 four-hour experiments, during which the signal did not decay), while the supernatant was scanned for 4 h.

Partitioning calculations We define the partition coefficient (PC) as the ratio between the concentration of small molecule (SM) in the condensate ($[SM]_{Cond}$) and the dilute ($[SM]_{Cond}$) phases:

$$PC = \frac{[SM]_{Cond}}{[SM]_{Dil}} \tag{5.3}$$

We know that the total amount of small molecule distributes between the phases as follows:

$$[SM]_{Tot} * V_{Tot} = [SM]_{Cond} * V_{Cond} + [SM]_{Dil} * V_{Dil} \tag{5.4}$$

where $[SM]_{Tot}$ is the total concentration, V_{Tot} the total volume, and V_{Cond} and V_{Dil} the volumes of the condensate and dilute phase, respectively.

The intensity ratio between the two phases, or signal factor, can be described as:

$$SF = \frac{S_{Dil}}{S_{RCond}} = \frac{[SM]_{Dil}}{[SM]_{RCond}} \tag{5.5}$$

where $[SM]_{RCond}$ is the concentration of small molecule in the condensate once resuspended with a volume V_{Res}:

$$[SM]_{RCond} = \frac{V_{Cond} * [SM]_{Cond}}{V_{Cond} + V_{Res}} \tag{5.6}$$

Combining the previous two equations, we can define $[SM]_{Dil}$ as:

$$[SM]_{Dil} = SF * \frac{V_{Cond} * [SM]_{Cond}}{V_{Cond} + V_{Res}} \tag{5.7}$$

And finally we can substitute $[SM]_{Dil}$ in the initial partitioning equation to obtain:

$$PC = \frac{[SM]_{Cond}}{\frac{SF * V_{Cond} * [SM]_{Cond}}{V_{Cond} + V_{Res}}} = \frac{V_{Cond} + V_{Res}}{SF * V_{Cond}} \tag{5.8}$$

This final equation allows us to calculate the partitioning coefficient from the measured volume of condensate V_{Cond}, the experimentally defined resuspension volume V_{Res} and the intensity factor SF measured by NMR as explained above.

The partition coefficient was calculated per each replicate, and the result is reported as average ± standard deviation.

Additionally, we can rearrange equation 5.4.2 and substitute $[SM]_{Dil}$ in 5.4.2, also considering that $V_{Dil} = V_{Tot} - V_{Cond}$, as follows:

$$[SM]_{Cond} = \frac{[SM]_{Tot} * V_{Tot}}{V_{Cond} + \frac{V_{Tot} - V_{Cond}}{PC}} \tag{5.9}$$

Finally, from this last expression and using the obtained partition coefficient, we can easily use equation 5.4.2 to calculate the concentration of small molecule

in the dilute phase. Finally, the fraction of signal from the condensate signal is calculated as follows:

$$S_{CondFraction} = \frac{S_{RCond}}{S_{RCond} + S_{Dil}} = \frac{S_{RCond}}{S_{RCond} + SFS_{RCond}} = \frac{1}{1 + SF} \tag{5.10}$$

Relative partitioning measurements of ATP, cAMP, GTP and cGMP in DDX4^{N1}NoCys

Phase separation set-up Initially DDX4^{N1}NoCys was buffer exchanged from a H_2O based sodium phosphate buffer into a D_2O based sodium phosphate buffer using phase separation. We have observed that at the conditions of 100-150 uM protein and 100-150 mM NaCl approximately half the protein amount is in the condensate phase, which makes this viable (the dilute phase can be recovered and used later).

250 µl of 374 µM DDX4^{N1} NoCys (in H_2O 20 mM sodium phosphate, 300 mM NaCl, pH 8) were mixed with the same volume of D_2O 20 mM sodium phosphate pH 6.2. This led to phase separation of the protein. The sample was centrifuged and after removal of the supernatant the pellet was resuspended in 300 µl of D_2O 20 mM sodium phosphate, 150 mM NaCl, pH 7.2 (the equivalent conditions of the phase separated sample). This sample was still phase separated and used as the protein stock for three samples; for each sample, 100 µl of DDX4^{N1}NoCys suspension was mixed with small molecule stocks at 10 mM (for final 1mM concentration of each small molecule) and buffer up to a volume of 200 µl. After incubation for 10 minutes at 4 ^{o}C, the samples were centrifuged and separated into pellet and supernatant samples. The pellet sample was resuspended in 200 µl of buffer and the supernatant sample was topped up to 200 µl with buffer (D_2O 20 mM sodium phosphate, 150 mM NaCl, pH 7.2). In order to ensure the samples would be in the dilute phase for NMR analysis, the salt concentration was raised to 500 mM NaCl by adding 20 µl of a 4 M NaCl stock.

The dataset consisted of six samples (n=1 for each) that came from two rounds of the procedure described above:

- Thiamine + ATP

- Thiamine + cAMP
- Thiamine + ATP + cAMP
- Thiamine + GTP
- Thiamine + cGMP
- Thiamine + GTP + cGMP

One stock of D_2O exchanged DDX4^{N1}NoCys was used for the three samples with thiamine and adenine nucleotides, and the other stock for the samples with thiamine and guanine nucleotides.

NMR measurements ^{1}H water suppressed spectra were acquired using a 22.3 T (950 MHz) Bruker spectrometer equipped with a 5 mm TCI cryoprobe (Department of Biochemistry, University of Oxford).

Relative partitioning calculations This was based on the NMR measurement of signal factor, which is the quantity used to obtain the partition coefficient. For this set of experiments, since volume fractions were not available, quantification of relative partitioning was based on the relative SF versus a reference molecule SF, which is present in all the samples, which corresponded to thiamine.

5.4.3 Computational analysis

We designed a pipeline for automated processing of the data using the python module nmrglue. Briefly, there are 3 stages to the analysis:

1. **Analysis of the small molecule control samples and identification of peaks** This was done applying a hat function (1/0) to the processed spectra (5-9 ppm), in which spectral points with values above a user-defined threshold would be converted to 1, and otherwise replaced by 0s. Afterwards the new binary array was subtracted from itself with a 1 point displacement in order to identify the edges of the peaks where 0 becomes 1 and viceversa. These

locations were then saved in arrays of left edges and right edges, and the center was calculated as the mean point and reported in ppm as the reference value for each peak.

2. **Processing of pellet and supernatant spectra for the extraction of signal factors** The pellet and supernatant spectra were processed using the Python package nmrglue. Then, for each molecule region found in step **1**, the correlation of the pellet and supernatant peaks was maximised (using the function correlate from scipy) and the pellet signal shifted appropriately. This was done to facilitate visual comparison of the peak shapes to detect signs of exchange. If baseline correction was required, it was then applied next. Finally, the signal factor was calculated as the ratio between the signal per region in supernatant / pellet, taking the maximum value within the region defined (original region widened by the default or customised "width" parameter - see a discussion in the Results section). A series of Figures representing the entire processed regions for pellet, supernatant and control molecule spectra, as well as the individual regions, were saved.

3. **Calculation of statistics** Based on the signal factors, relative signal factors were calculated (versus thiamine) as a ratio of the average signal factor (across multiple peaks of the same molecule included in the quantification - referred to as internal error later in the Results section), propagating the error accordingly. If the pellet volume was known (as an input), the absolute partition coefficient, both for each peak individually and per molecule as average ± standard deviation, would also be calculated. A series of plots was automatically produced of all the per peak and per molecule statistics calculated.

The analysis pipeline took a few seconds to complete. In general, for each sample it was typical to run the code 3 or 4 times. The first time, we used a default width (3.0) and no baseline correction for all peaks. In some samples that was already appropriate and the analysis was complete, if none of the peaks had

different baselines and the initial region window didn't miss any peaks. However in most samples baseline correction would be needed for some of the peaks, in which case the edges of the region had to be flat, which sometimes requires adjusting the width parameter. Also, in case of exchange in any peaks, we would update the settings input file to exclude them from quantification. A few examples of manual adjustments required are discussed in the Results section.

5.4.4 Comparison of methods to determine pellet volume

Two different methods were compared using samples made from the same protein stock (400 µM DDX4^{N1} WT in 20 mM sodium phosphate, 300 mM NaCl, 5 mM TCEP, pH 8) and phase separated to the same conditions using a no salt buffer (20 mM sodium phosphate, 5 mM TCEP, pH 8). The final conditions were 133 µM protein, 100 mM NaCl, 5 mM TCEP, pH 8. Each sample consisted of 59 uL of protein stock and 118 no salt buffer. The samples were incubated for one hour at 5 ºC (cold room) and centrifuged for 10 minutes at 2500 g (at 10 ºC). In the microcentrifuge method, the samples were made in a microcentrifuge 1.5 ml tube, imaged after centrifugation and the volume calculated as described above for the FUS-lipoamide project. In the PCV tube method, the samples were made directly in the PCV tube

The initial method described above in section 5.4.2 was compared with a new method based on using PCV tubes (Helena Biosciences) designed for cellular fraction measurement. The two methods had differences not only in the measuring of the volume but also in how the resuspended pellet sample was recovered from the tube, which are shown in the Figure 5.8. The differences with the original method consist of initiating the phase separation reaction in the PCV tube instead of a microcentrifuge tube, and after centrifugation measuring the volume of pellet using the Volume Reading device that is calibrated for the PCV tubes. Another new step in the protocol consisted of measuring the volume of leftover supernatant: after removal of the supernatant, without disturbing the pellet, the PCV tube was centrifuged again to bring down any remaining supernatant, and the total volume remaining measured using the same device, from which the leftover supernatant volume was extracted.

5.4.5 Measurements of absolute partitioning of ATP, cAMP, GTP, cGMP and deoxyribonucleosides

Phase separation set-up Samples were buffer exchanged by phase separation at room temperature with a D_2O-based buffer: 1 part of the $DDX4^{N1}$ (WT or NoCys) stock solution (400 µM protein, 20 mM sodium phosphate H_2O-buffer, 300 mM NaCl, pH=8 and 5 mM TCEP) was mixed with 2 parts of D_2O-based buffer (20 mM sodium phosphate, pH 8.0, 5 mM TCEP). The phase separated mixture was centrifuged at 1000g for 10 minutes and the supernatant was removed and stored away. The pellet was resuspended in high-salt D_2O-based buffer (20 mM sodium phosphate, 300 mM NaCl, pH 8.0, 5 mM TCEP) to a final concentration of 250 µM (at which conditions it was still phase separated). The resuspension was mixed by pipetting before taking out 59 µl volumes into PCV tubes. Compounds were prepared individually in 5-10 mM stock solutions in NaCl-free D_2O-based buffer (20 mM sodium phosphate, 300 mM, pH 8.0, 5 mM TCEP) and pH was adjusted to 8. Then mixtures of small molecules were made by combining the stocks, for a concentration of 750 µM of each small molecule in the molecule mix (and a final 500 µM concentration after mixing with the protein). Partitioning experiments were carried out by mixing 118 µl of small molecule mix (750 µM of each small molecule in 20 mM sodium phosphate, pH 8.0, 5 mM TCEP) into the 59 µl of protein already present in the PCV tube.

Pellet volume measurements Pellet and leftover supernatant volumes were measured using the PCV tube based method described in detail above.

NMR Measurements ^1H water suppressed spectra were acquired using a 22.3 T (950 MHz) Bruker spectrometer equipped with a 5 mm TCI cryoprobe (Department of Biochemistry, University of Oxford). Typical acquisition times were 5 min for small molecule only control and supernatant samples and 1h 50 min for pellet samples.

Quantification of partition coefficients Volumes relating to experimental handling of the samples were added to consideration in the PC calculation, due to the fact that in the lipoamide sample we used Shigemi tubes (for minimal dilution of the pellet sample), therefore we were measuring all the signal available in the sample of pellet and dilute phase. Here we used 3 mm tubes, therefore the dilutions performed are relevant. We chose the resuspension and additional volumes needed to make up the samples to be very consistent across pellet and supernatant, which means that those extra terms have very little impact in this setting, but the expressions developed are the most exact and should be used if any of the volumes is substantially different for any reason.

To study the impact of knowing the volume of leftover supernatant on the absolute partition coefficient, we calculated the partition coefficient in two ways: without and with accounting for the signal of molecule that comes from the leftover supernatant.

Without considering the contribution of leftover supernatant

Briefly, considering the sample handling volumes results in the following relations between the concentration of molecule in each supernatant and pellet NMR samples, and the concentration of the molecule in the dilute/condensate phase, respectively:

$$[Dil]_{Sample} = \frac{[Dil]V_{Dil}}{(V_{Dil} + Vextra_{Dil})} \tag{5.11}$$

$$[Cond]_{Sample} = \frac{[Cond]V_{Cond}}{(V_{Cond} + V_{Res} + Vextra_{RC})} \tag{5.12}$$

The signal factor is then:

$$SF = [Dil]_{Sample}/[RCond]_{Sample} \tag{5.13}$$

Rearranging gives:

$$[Dil] = SF * \frac{[RCond] * V_{Cond}}{V_{Cond} + V_{Res} + Vextra_{RC}} * \frac{(V_Dil + Vextra_S)}{V_{Dil}} \tag{5.14}$$

Note that this expression shows how the concentration in the dilute phase sample is smaller than the concentration in the dilute phase by a factor that directly comes from the added volume Vextra(S).

Finally, incorporating this expression into the expression for partition coefficient yields:

$$PC = \frac{V_{Dil} * (V_{Cond} + V_{Res} + Vextra_{RC})}{SF * V_{Cond} * (V_{Dil} + Vextra_S)} \tag{5.15}$$

where V_{Dil} is the volume of dilute sample taken for analysis, V_{Res} is the resuspension volume added to the pellet, $Vextra_{RC}$ is the extra volume added to the pellet and $Vextra_S$ is the extra volume added to the supernatant. Those were added to increase the salt concentration to 300 mM and to make up enough sample for the 3mm tubes.

If we account for leftover supernatant (V_{LOS}), then the concentration of molecule in the resuspended condensate needs to be modified as:

$$[RCond]_{Sample} = \frac{[Cond]V_{Cond} + [Dil]V_{LOS}}{V_{Cond} + V_{Res} + Vextra_{RC}} \tag{5.16}$$

This introduces a term depending on the concentration in the dilute phase ([Dil]), and after applying the same reorganisation than previously, the partition coefficient is:

$$PC = \frac{1 - \alpha SFV_{LOS}}{\alpha SFV_{LOS}} \tag{5.17}$$

where α includes terms related to the experimental handling volumes mentioned above as follows:

$$\alpha = \frac{V_{Dil} + Vextra_S}{V_{Dil}(V_{Cond} + V_{LOS} + V_{Res} + Vextra_{RC})} \tag{5.18}$$

The concentrations of molecule in each phase are defined by the same equations derived before for lipoamide.

5.5 Results and Discussion

5.5.1 Lipoamide partitions into *in vitro* FUS condensates

Figure 5.5 shows the application of the designed method for the quantification of partitioning of lipoamide into FUS *in vitro* condensates. This method involved the physical separation of the condensate and dilute phases into two NMR samples (Figure 5.5 A). As described in the Methods, by measuring the NMR signal ratio (Signal Factor) between the samples (an example of the overlay of spectra is shown in Figure 5.5 B) and the volume of condensate phase as a centrifuged pellet (an example of the measurement is shown in Figure 5.5 C), we could quantify the partition coefficient of lipoamide into the droplets.

The volume fraction of the condensate phase of FUS condensates was determined to be on average 0.15 ± 0.02 % (Figure 5.5 D i) and the signal factor as measured by NMR was 65.45 ± 12.62. From these measurements, the calculated partition coefficient (PC) was 10.34 ± 1.74 (Figure 5.5 D ii), which indicates that lipoamide is enriched in the condensate phase of FUS condensates. This would be equivalent to $-\Delta G = 5.5 \pm 0.29$ kJ mol^{-1} (or 1.34 ± 0.07 kcal mol^{-1}) , which is energetically comparable to the contribution of a hydrogen bond in protein-ligand interactions (0.5-1.8 kcal mol^{-1}[305]), or its stabilisation effect for a protein in an aqueous environment (average at 3.2 kJ mol^{-1}[306]). The tenfold difference in concentration between the two phases is apparent in the comparison of the NMR signal fraction corresponding to the condensate phase and the volume fraction of the condensate phase (Figure 5.5 D i). Given a total average lipoamide concentration of 100 µM in the phase separated mixture, the effective calculated concentration in the condensate phase is $\sim$ 1 mM; in contrast, the concentration in the dilute phase is only minimally reduced with respect to the average. Lipoamide disrupts phase separation of stress granules (of which FUS is a component) in cells, and affects phase separation of FUS *in vitro*, yet it does not show binding to FUS using gold standard interaction-probing techniques[275]. These results suggest a potential mechanism of action of lipoamide on phase separation based on its preferential localisation in the condensate phase.

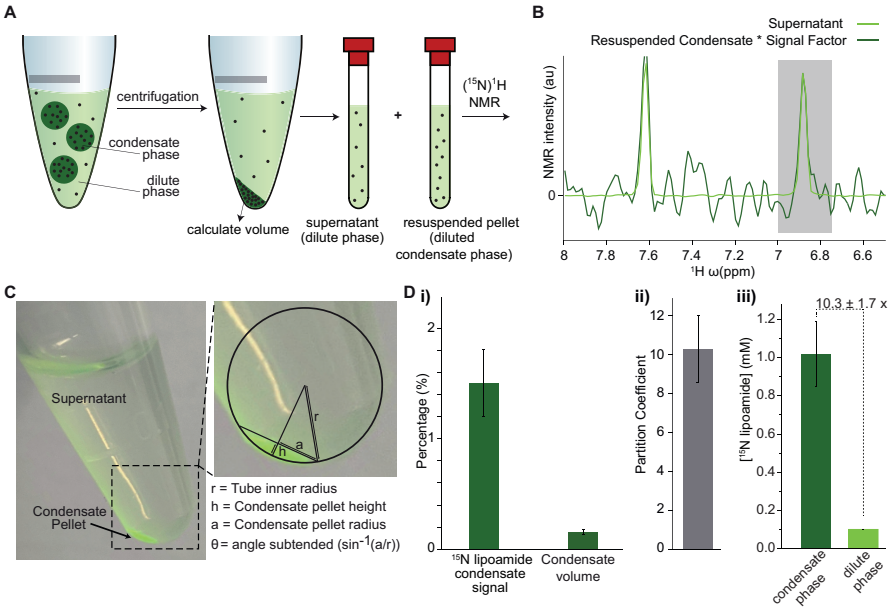


Figure 5.5: Measurement of the lipoamide partition coefficient into FUS condensates by NMR. A-C, Methodology for determination of partition of ^{15}N lipoamide into FUS condensates *in vitro*. A. Schematic showing the sample preparation process. B.Example of ^{15}N edited ^{1}H NMR signal around the ^{15}N cis and trans amide protons of lipoamide in representative spectra of the dilute phase and condensate phase of the same sample. The signal is previously normalised to account for different numbers of scans, since more scans were required for the pellet sample to achieve adequate signal to noise. The resuspended pellet (condensate phase) spectrum is shown scaled by an experimentally determined signal factor (for the trans-amide proton), used in the calculation of the partition coefficient and signal fraction (see Methods). C.Measurements for calculation of the condensate phase (pellet) volume from macro photographs of the sample within a microcentrifuge tube. D. Results of the partitioning experiments, shown as mean ± standard deviation; uncertainties derive from independent repeated measurements (n = 4). (i) Measured ^{15}N edited ^{1}H NMR signal fraction and condensate volume fraction. (ii) Calculated partition coefficient. (iii) Concentration of racemic ^{15}N-lipoamide in the condensate and the surrounding dilute phase.

5.5.2 Development of a semi automated pipeline for analysis of partitioning: test case results

Figure 5.6 shows some of the outputs produced by the analysis code that systematise and quicken the analysis of the data, as well as examples of small manual adjustments required. The main output for verification of analysis being done correctly is the multiplot figure in which the spectra of supernatant (dilute phase) and resuspended pellet (condensate phase), scaled by the automatically calculated signal factor (based on maximum intensity in the region - see Methods)

are plotted for each peak region identified for each molecule, aligned with spectra of each molecule on its own (see Figure 5.6 B ii). This helps verify that the processing has been done well, and aids in making two small manual adjustments: the modification of the width of the region (with respect to the center of the peak) and the application of baseline correction, in this order. The width of the region taken for analysis needs to be adjusted in three situations: 1) the location of the peak center is different enough to that in the molecule-only control sample that the default window misses the peak in the phase separated protein-molecule samples, 2) the location of a more intense peak of another molecule of the mixture is close enough that the intensity of the other peak is mistakenly used for the quantification (e.g. Figure 5.6 A) and 3) if baseline correction is required, in order to make sure the signal at the edges is suitable for the linear subtraction applied (for example, not including the start of another peak, as in Figure 5.6 B). In the case of the protein regions used to quantify the enrichment (increased density) in the condensate phase, taking the maximum intensity automatically to calculate the signal factor worked better for the amide region than the methyl region if there were sharp peaks from the small molecules in the latter; in such cases, we applied a manual adjustment to the signal factor using the signal factor based on integral calculation as the basis. The processing of the data produces outputs including the signal factor, and if pellet volume is known (included together with experimental handling volumes in a sample information input file), the partition coefficient, both for each peak individually and per molecule as average ± standard deviation, as a well as a set of plots representing these quantities. Regarding the averaging, by default all the peaks of a molecule would be included, but we often had to manually exclude peaks due to exchange or the overlap of peaks of different molecules (see Figure 5.6 B).

5.5.3 Partitioning into DDX4^{N1}NoCys is affected by the nucleobase (G vs A) and the triphosphate vs cyclic moiety of a set of nucleotides

Figure 5.7 shows the results of applying the semi automated processing method described above to the dataset of samples containing thiamine and one or two

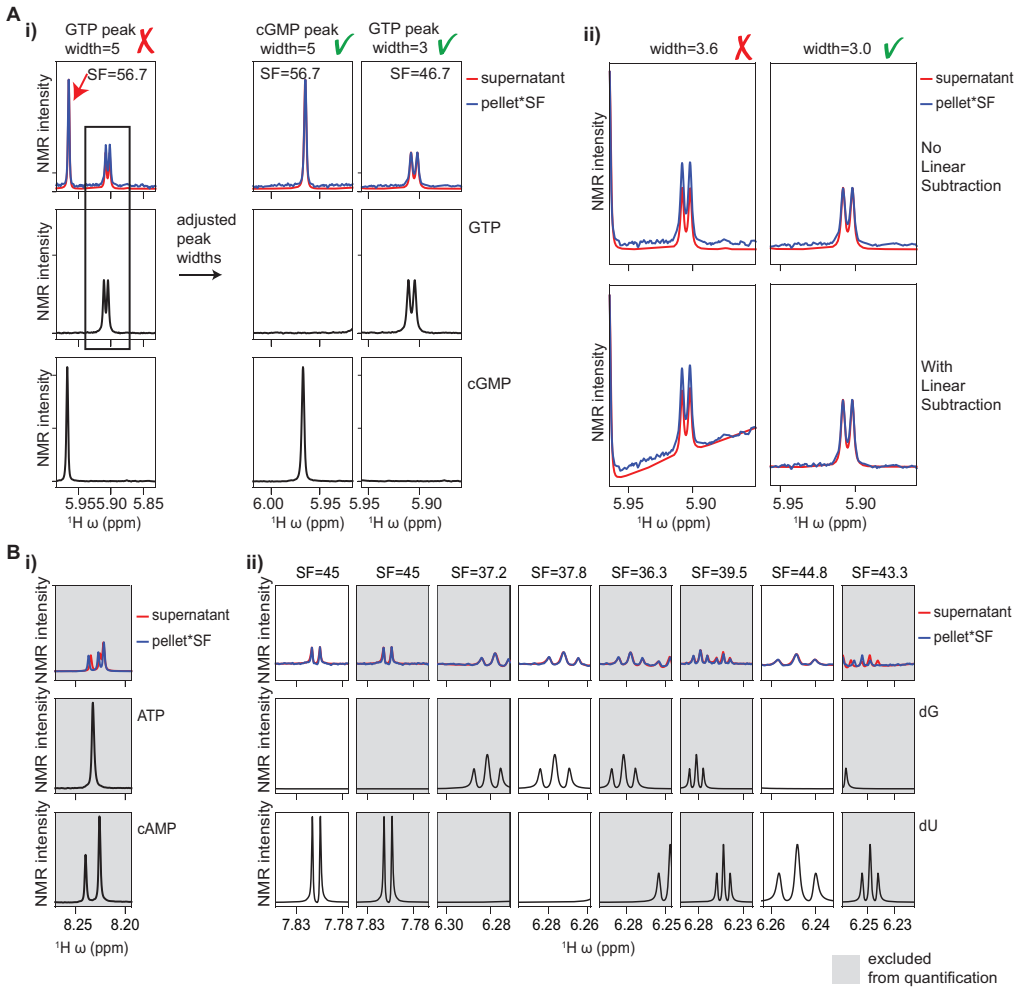


Figure 5.6: Adjustments in the semiautomated NMR processing pipeline. A. Width of the region needs to be adjusted to avoid nearby stronger intensity peaks (i) and/or for correct baseline subtraction (ii).B. Exclusion of peaks due to overlap between molecules (i) or in the case of multiplets to avoid underestimating the error between peaks corresponding to individual atoms (ii). (ii) also serves as a representative example of outputs used to verify the processing workflow.

molecules from the sets of ATP,cAMP, GTP and cGMP (Figure 5.7 A). There were three aims of this experiment: 1) to test a semiautomated processing pipeline, 2) to determine if thiamine could be used as a reference standard molecule across distinct combinations of small molecules, and thus gain insight on their relative partitioning (see table of sample composition and chemical structures in Figure 5.7 A), and 3) to probe the effects of backbone charge (triphosphate vs cyclic) and aromatic ring substituents (guanine vs adenine bases) on partitioning into DDX4^{N1}NoCys, taken as a model system.

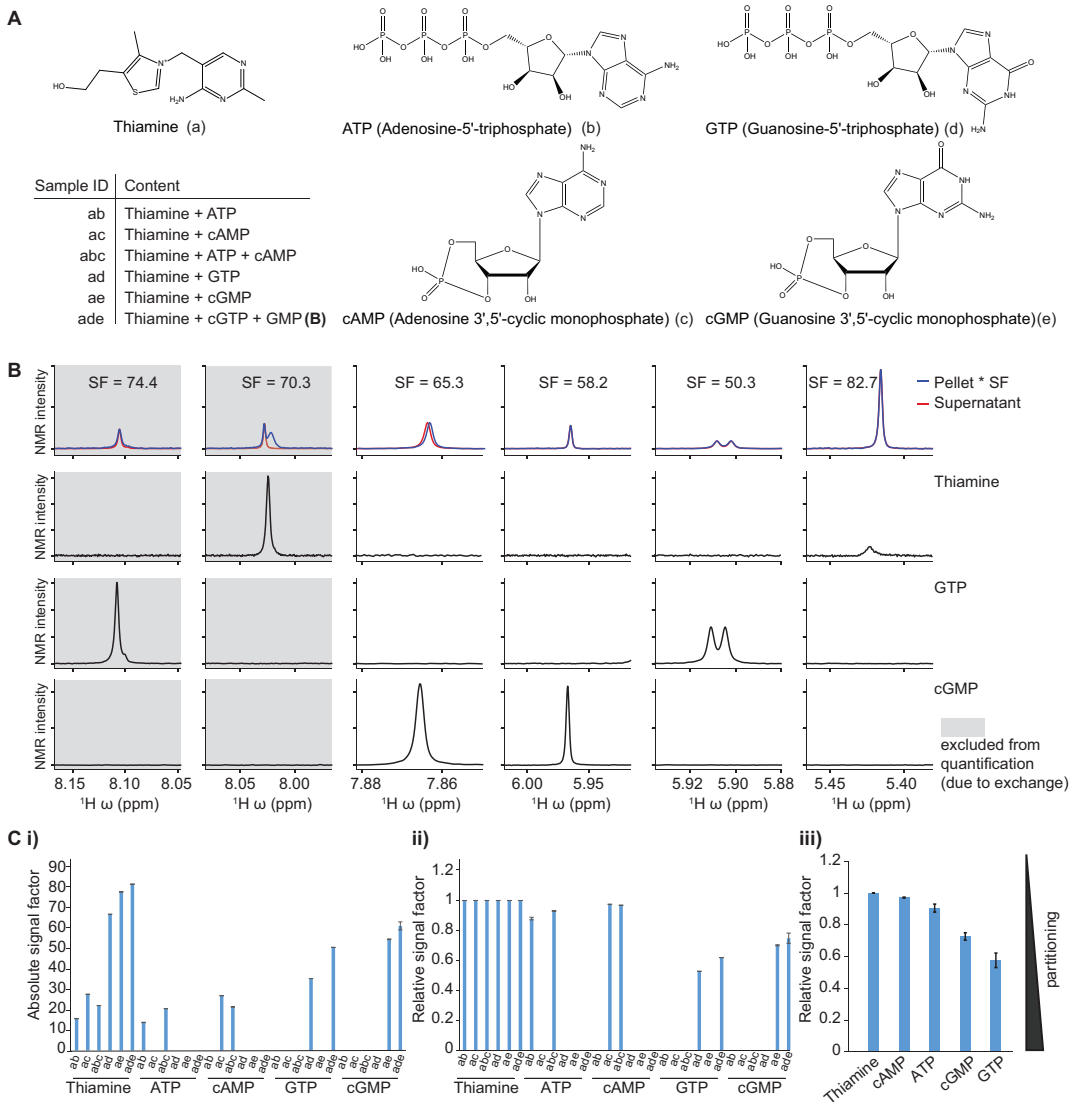


Figure 5.7: Relative partitioning of guanine and adenine nucleotides into DDX4^{N1}NoCys *in vitro* condensates. A. Chemical structures of ATP, cAMP, GTP, cGMP and Thiamine and composition of the samples used. B. Example of semi automated processing for the sample containing thiamine, ATP and cAMP. The peaks not included for quantification are shaded in grey. C. i) Measured NMR signal factor (signal supernatant / signal pellet, after scan normalisation) for each molecule (i), and relative signal factor (ii) (signal factor / signal factor of thiamine - taken as the average of two peaks) per individual sample containing the molecule. Absolute Signal factor Results are shown as a single measurement (if only one peak was available in the aromatic region) or mean ± standard deviation (across multiple peaks), with error propagation applied included in the error for the relative signal factor quantity. (iii) Average relative signal factor vs thiamine for each molecule (mean ± standard deviation from two samples, see table in A). The experiments reported here were carried out by Spencer A. Cahill (protein expression and set-up of partitioning samples) and Charlie Buchanan (NMR spectra acquisition).

First, Figure 5.7 B shows an example of the application of the processing pipeline to the sample thiamine + GTP + cGMP. The 8.10 ppm peak of GTP and the 8.0 ppm peak of thiamine were excluded from quantification since they showed signs of exchange.

Second, Figure 5.7 C i and ii show the value of assessing relative signal factor ratios of each molecule with respect to the chosen standard, thiamine. In i) we can see that the absolute NMR-measured signal factors of each molecule are not the same across samples containing this molecule. This is expected since the NMR intensities of supernatant and pellet samples (and their ratio, signal factor (SF)) depend not only on the molecule's intrinsic property of partitioning but also on the fraction of condensate phase (which can be affected by the presence of the molecules in different ways across the sample) and on the amount of leftover supernatant, both of which are subject to experimental variability. In this dataset, both volume fraction and volume of leftover supernatant were not known, which precludes calculations of partition coefficients and further analysis of the variations in absolute signal factor. For this reason, we focused on the comparison of relative signal factor measurements, taking thiamine as the reference. In iii), we can see that the relative signal factor of all molecules with respect to that of thiamine was consistent across samples containing the same molecule, within error (for example, the relative signal factor of ATP in samples ab and abc), even if the absolute signal factors were different. This confirms that the behaviour of each molecule was independent of the presence of others and the relative trends are maintained.

Third, Figure 5.7 D shows the average of relative signal factor for each molecule across the two samples that contained it. The four nucleotides partitioned more than thiamine into DDX4^{N1}NoCys droplets. The main trend shown by the results is that guanine-containing molecules (GTP/cGMP) had a smaller relative signal factor, and therefore a larger partition coefficient, than adenine-containing molecules (ATP/cAMP). This can thus be attributed to the distinct substituents present on the indole group. In addition, the triphosphate versions were slightly more enriched

than the cyclic versions, which was more pronounced in the GTP/cGMP pair. Both findings are discussed below in conjunction with other results.

In summary, this experiment allowed us to validate the approach of studying chemical trends via relative partitioning experiments, and test the benefits of using a semiautomated approach for the analysis of the NMR data. We observed that the partitioning of this set of four nucleotides is sensitive to the substituents of the indole ring, given the relative enrichment of guanine vs adenine-containing nucleotides, as well as the presence of the triphosphate moiety (being more enriched than the cyclic counterpart, for the same base).

5.5.4 Development of a PCV tube based method for determining partition coefficient

There were two main limitations of the method developed for the study of lipoamide partitioning into FUS droplets, and which were addressed by further method development: 1) when observed from a top view, the pellets formed at the bottom of the microcentrifuge tube had a concave appearance, meaning that the real pellet volume was smaller than the semi-spherical cap delimited by the top of the pellet meniscus observed in the perpendicular views and used in the calculations, and 2) the fact that we could not determine the amount of leftover supernatant that remains in the tube after removing the volume of supernatant required for dilute phase NMR sample and as much as possible without disturbing the pellet.

We addressed this by developing the pipeline further by using packed cell volume tubes, designed to measure the cell volume fraction of cell samples, instead of microcentrifuge tubes, for the phase separation reaction, and measuring the volume of pellet using the companion calibrated measuring device. A key aspect of these devices is that they are designed for very viscous samples (we had previously found that FUS stuck to the glass of various capillaries tested) and that they have a thin graded cylinder for measurements with a precision of 0.025 µl.

Figure 5.8 A shows a direct comparison of the two methods in terms of the measured pellet volume, using samples of the same concentration of $DDX4^{N1}$ WT.

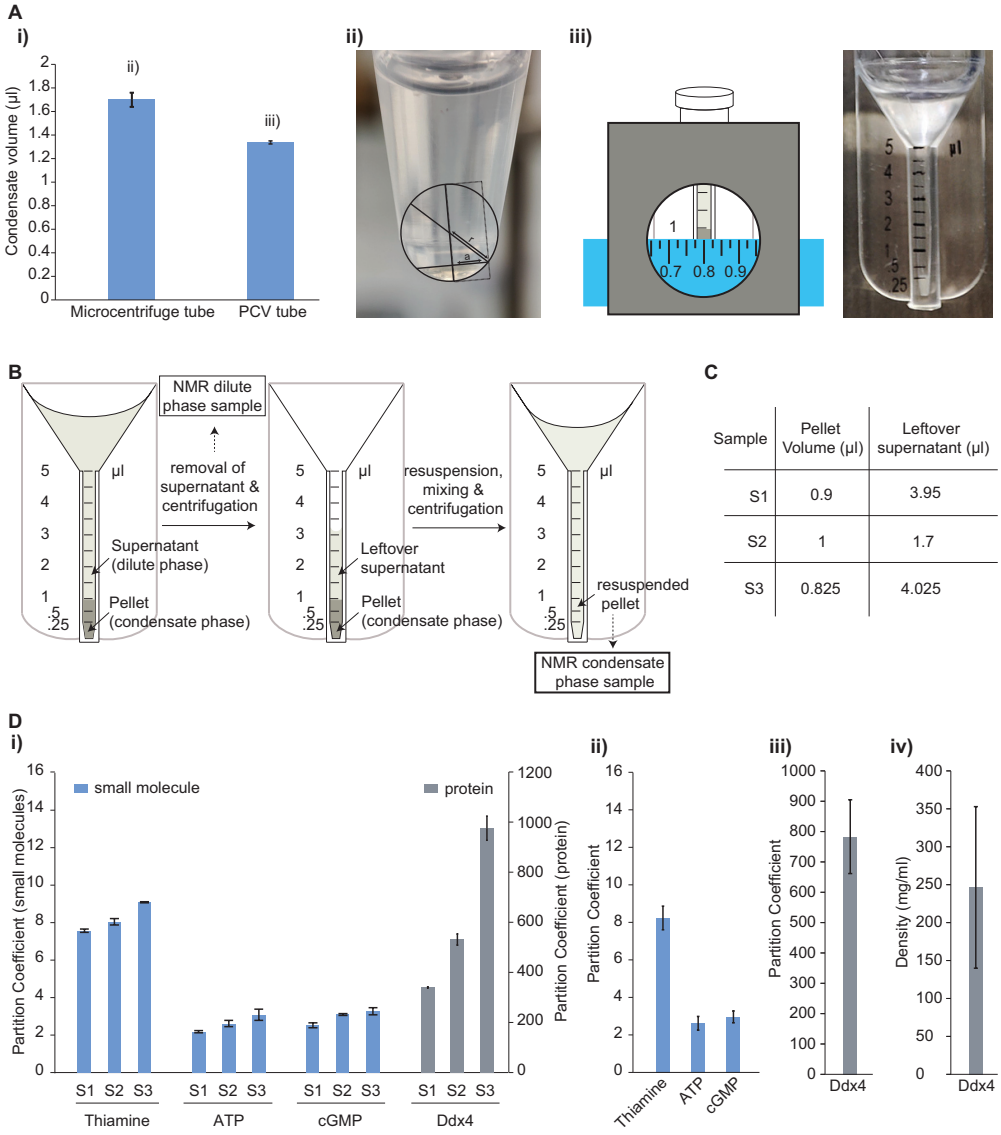
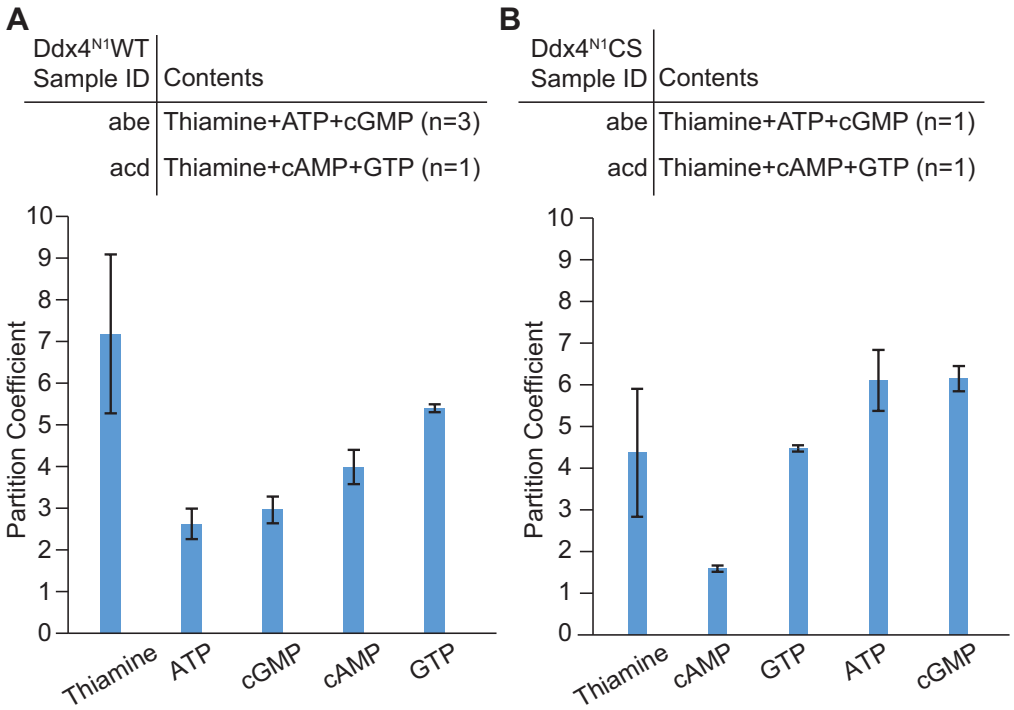


Figure 5.8: Experimental method developments. A. Volume of condensate phase (pellet formed after centrifugation) measured using two different methods (see Methods for complete description) using the same conditions (buffer conditions, protein concentration and sample volume). Results shown are the average ± standard deviation for three replicates (microcentrifuge tube method - ii)) and two replicates (PCV tube - iii)). Note that the image of the PCV tube in (iii) does not correspond to any of the replicates that yielded the measurements shown in the barchart in (i), but to another sample. B. Overview of the protocol used to measure the amount of leftover supernatant after removal of the dilute phase sample. C. Pellet volume and leftover supernatant for individual triplicate samples of $DDX4^{N1}$ with thiamine, ATP and cGMP, measured using the PCV tube based method. D. (i) Determined partition coefficients for small molecules and protein for individual triplicate samples of a mixture of thiamine, ATP and cGMP, using the PCV tube volume measurement method and accounting for the leftover supernatant. Values are reported as average ± standard deviation from internal error (using different peaks of the molecule). (ii) Average ± partition coefficient of the individual replicates shown in D i) (n=3). (iii) and (iv) Average protein partition coefficient (iii) and density (iv) in the samples shown in D i. The experiments reported here were designed and performed together with James Eaton.

The two methods showed good reproducibility but the volume measured with the cell counting device is smaller (by around 25 %). This is in line with the expectations outlined above. To further validate the new PCV based method for partition experiments, we used three repeats of phase separated mixtures containing WT DDX4^{N1}, thiamine, ATP and cGMP. The results are shown in Figure 5.8 B-D. The correction for leftover supernatant volume (see schematic) is important for the reproducibility of the method; for thiamine, for example, without the correction $PC = 11.7 \pm 1.66$, and with the correction $PC = 8.3 \pm 0.6$. This is expected since the leftover supernatant volume depends on pipette handling and is different amongst samples, as can be seen in the table in Figure 5.8 B. Also, as expected, the calculated partition coefficient is lower when accounting for the leftover supernatant; the leftover supernatant contributes to the signal in the NMR pellet sample, which biases the result to a higher apparent concentration and thus partition coefficient in the condensate phase. We concluded that even though there was quite good agreement in terms of measured volume fraction between the two methods, the PCV method (using the leftover supernatant correction) was able to reduce a source of experimental variability, and for this reason established it as the reference method for determining absolute partition coefficients.

From the comparison of Figure 5.8 D i and D ii, it seems that for the small molecules the internal error in PC from a single sample (coming from using different peaks for the same molecule) is only marginally smaller than the error across the set of triplicates. The results revealed a larger variation in protein partitioning (also with respect to the internal error of each sample), compared to the small molecules. Interestingly, the fluctuations around the mean of the measured partition coefficient for the small molecules showed a strong correlation with the protein PC / density in the condensate phase of the corresponding individual sample (Pearson coefficients of 0.98-0.99). For DDX4^{N1}, the density at the same salt conditions (100 mM NaCl) was found to be around 400 mg/mg (for a single sample [156]), which is in the range of the values obtained here.



C

Protein construct	pH	Partitioning trend
Ddx4^{N1}WT	8	Thiamine > GTP > cAMP > cGMP ~ ATP
Ddx4^{N1}CS	8	cGMP ~ ATP > GTP ~ Thiamine > cAMP
Ddx4^{N1}CS	7	GTP > cGMP > ATP > cAMP > Thiamine

Figure 5.9: The partitioning trend of a set of aromatic molecules varies with protein construct and pH. Partitioning of thiamine, ATP, cGMP, cAMP and GTP at pH 8 in DDX4^{N1}WT (A) and Ddx4^{N1}NoCys (B), assessed using the PCV method and semi automated data processing. The value for thiamine in A (n=4) and B (n=2), and ATP (n=3) and cAGMP (n=3) in A are reported as mean ± standard deviation of independent replicates; all other values are reported as mean ± standard deviation from technical error (different peaks in the same sample); see tables for details of samples. C. Partitioning trends for the same set of molecules in three distinct conditions; the pH 8 trends correspond to the data shown in A and B, the pH 7 trend corresponds to figure 5.7. The experiments reported here were designed and performed together with James Eaton.

5.5.5 Comparison of partitioning trends of thiamine, GTP, cGMP, ATP and cAMP in DDX4 depending on construct and pH

Figure 5.9 shows the comparison of the partitioning of the previous set of thiamine and nucleotides at distinct conditions. All the molecules tested were enriched in the condensate phase, at pH 8, for both the WT and NoCys constructs. The trends in partitioning were different at pH 8 for the NoCys mutant, and also

different to WT at pH 8. Neither of these samples shows the trend of enrichment of triphosphate vs cyclic, and of guanine vs adenine bases observed in the previous dataset (see table in Figure 5.9 C and Figure 5.7 C iii).

Figure 5.8 shows that the replicability of the results was very good for the $DDX4^{N1}$ WT + thiamine+ATP+cGMP sample; however, comparing the two WT protein samples there was a larger error for the thiamine PC (see 5.9 A) which is coming from values of $\sim$ 7.5,8 and 9 from the triplicates shown in 5.8 and a value $\sim$ 4 for the cAMP and GTP sample. Similarly, in the case of NoCys protein, thiamine had a PC of $\sim$ 6 in the ATP+cAMP sample, versus $\sim$ 3 in the cAMP+GTP sample.

We speculate that the variations in thiamine partition coefficient might be due to small experimental differences in pH and/or percentage of D_2O in the samples due to the buffer exchange procedure prior to the partitioning experiments, since the samples were prepared on different days from a different batch of buffer exchanged protein (from the same batch of recombinant $DDX4^{N1}$ in H_2O sodium phosphate buffer - see Methods) and buffers, and we observed strong effects of D_2O on the phase diagram of $DDX4^{N1}$ WT and NoCys. In the buffer exchange (done by phase separation), we resuspended the pellet in 300 mM NaCl D_2O based buffer to a concentration equivalent to the original conditions in H_2O buffer (around 400 µM), in which we were expecting it to become dilute again. However, after equilibration at room temperature and extensive mixing, the solution stayed phase separated. We did not characterise the phase diagram fully, yet we diluted the concentration up to 250 µM and still observed phase separation, which is equivalent to at least the effect of lowering the concentration of NaCl from 300 mM to 200 mM in H_2O buffer (see Figure 5.2 C). This qualitative observation of a strong D_2O effect is in line with literature observations on other proteins. The phase separated regime of bovine serum albumin, in the presence of trivalent salts, was broadened by D_2O, and a 30 % content of D_2O lowered the transition temperature by 13 ^{o}C [203]. Similarly, strong effects of D_2O on the phase diagram were observed for a fragment of the androgen receptor, with the transition temperature dropping by 25 ^{o}C at 50 % D_2O [307]. Given the strong impact of D_2O on protein-protein interactions driving

LLPS, it seems plausible that small differences in D_2O content (from adjusting the pH of buffers, for example, since we used H_2O based acid and base solutions) could also influence the interactions between the protein and small molecules in the condensate phase, which drive partitioning.

For these reasons, I believe the buffer exchange into D_2O buffer, prior to the partitioning experiments, is one of the steps of the method which should be investigated and more carefully controlled. Overall, further work is needed to elucidate the source of the variability observed between experiments performed on different days, and make modifications to the protocol accordingly.

Finally, the differences seen at pH 8 for the WT versus NoCys constructs could be due to the effects of the four cysteine to serine mutations, or the presence of 5 mM TCEP as a reducing agent in the case of the WT protein (to prevent aggregation), which could interact differently with the small molecules. We did not probe this further, and instead focused on the WT construct for the remaining experiments presented in this chapter.

5.5.6 Purine deoxyribonucleosides partition more than pyrimidines into DDX4^{N1} WT

We next probed the influence of the nucleobase on the partitioning into DDX4^{N1} by using a set of four deoxyribonucleosides (see Figure 5.10). In both samples, the purine nucleoside showed larger partitioning than the pyrimidine. The values for thiamine partitioning were very consistent across the two samples, as can be seen in the very small error bar in Figure 5.10 B i (coming from the average of the two samples, and not the internal between peaks error that is reported for the bases). Therefore we could compare the absolute partition coefficients of all the deoxyribonucleosides. dG was clearly more enriched than the rest, dA being the closest, marginally within the experimental error with respect to dC $\approx$ dU. The preferential partitioning of dG versus dA reproduced the trend seen for ATP/cAMP vs GTP/cGMP in the previous section (see Figure 5.7).

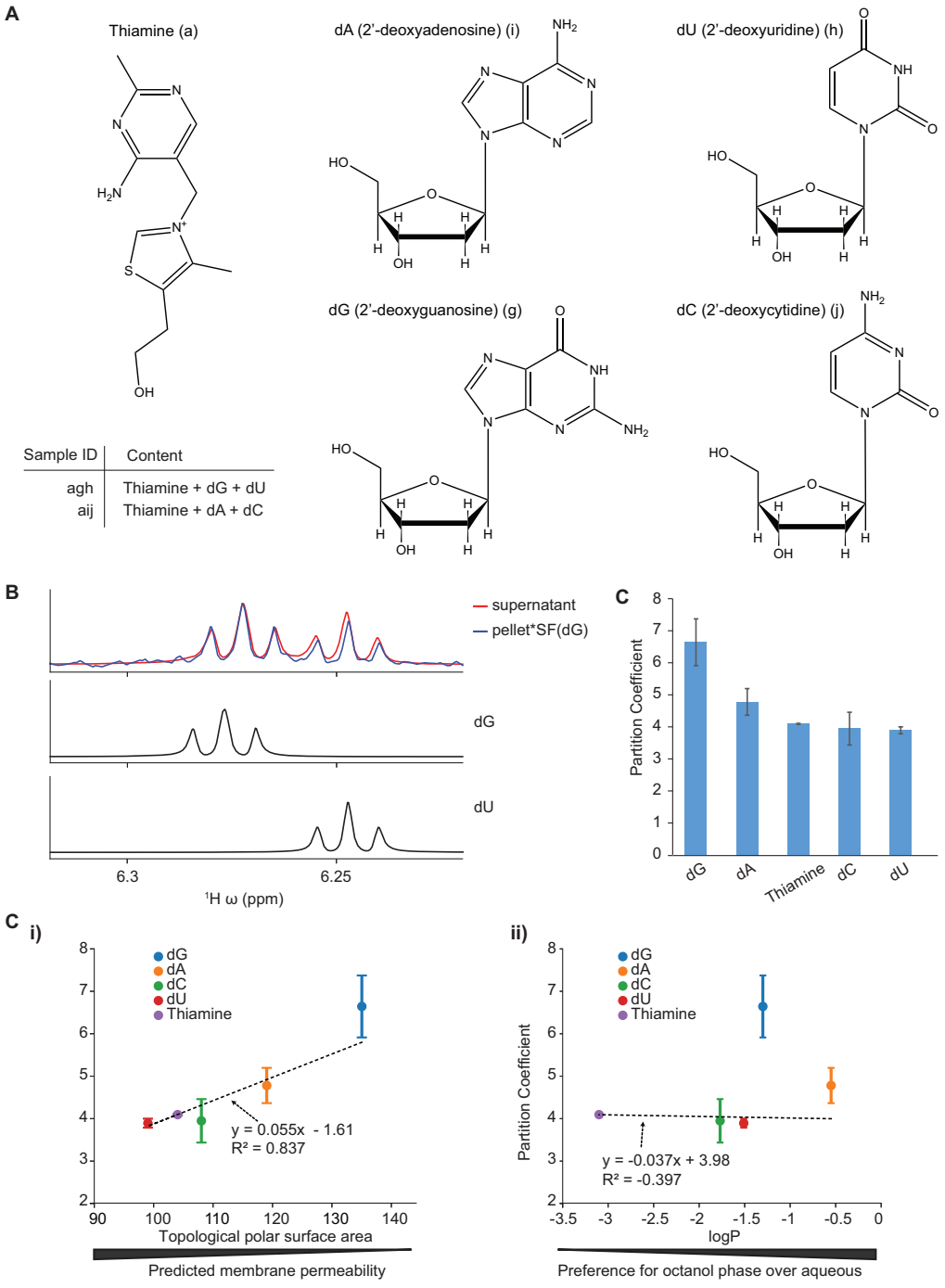


Figure 5.10: Partitioning of deoxyribonucleosides into DDX4^{N1} WT. A. Chemical structures of deoxyribonucleosides and thiamine. B. Representative NMR spectra of the dilute (supernatant) and condensate (pellet) phases of the agh sample, with the pellet intensity normalised by the signal factor of dG. dG is enriched versus dU, as shown by dU requiring a larger signal factor to match its supernatant intensity. C.Partition coefficient of deoxyribonucleosides and thiamine. For the nucleosides, the result is shown as average ± standard deviation across multiple aromatic peaks in one sample,and for thiamine the result is shown as average ± standard deviation across two independent samples (see Methods for composition of this dataset). D. Correlation between the partition coefficient and topological surface area (i) (values from PubChem) or logP (ii) (values from PubChem or DrugBank), calculated as a linear regression, weighted by the standard deviation of each point.The experiments reported here were designed and performed together with James Eaton.

The partition coefficient trends of the deoxyribonucleosides do not correlate significantly with logP, which is a measure of partitioning into an organic phase (octanol) relative to water, and one of the five properties related to the Lipinski rules, used for the prediction of absorption in drug discovery[308]. There is though a strong positive correlation with the topological polar surface area[309]. This descriptor is a two-dimensional version, faster to compute, of the three-dimensional polar surface area, which has been found to inversely correlate with intestinal absorption and blood-brain barrier permeability [310][311]. One of the features that distinguishes membraneless organelles from membrane-bound organelles is the greater permeability of membraneless organelles [145][146]. The fact that membrane permeability and droplet partitioning are inversely correlated emphasizes the distinct environment that is created by phase separation, both with respect to the bulk cytoplasm and the membrane boundaries.

Also, the correlation of the partitioning with a descriptor related to polar interactions (TPSA), but not with the solubility in the nonpolar solvent octanol, is in agreement with the finding that the permittivity of the condensed phase of $DDX4^{N1}$ resembles that of polar solvents DMSO and acetonitrile[148], even if it contains 75% water[156].

5.5.7 Rationalisation of the partitioning trends as a reflection of protein-nucleotide interactions in the condensate phase

We have explored the interactions in the condensed phase of $DDX4^{N1}$ through a series of experiments to determine relative and absolute partitioning coefficients of a set of test molecules. These revealed a trend of preferential enrichment for triphosphate nucleotides versus their cyclic counterpart (for $DDX4^{N1}$NoCys at pH = 7), and higher partitioning of guanine-containing nucleotides versus adenine counterparts, both in the NoCys mutant at pH 7 and in the WT at pH 8. We then chose the WT construct and pH 8 conditions to further probe purine and pyrimidine deoxyribonucleosides, which yielded the trend dG $>$ dA $\geq$ dC $\sim$ dU.

The finding that the triphosphate group (versus cyclic) has a positive effect on partitioning (5.7) follows a similar trend with literature results of ATP vs ADP/AMP comparisons regarding the disruption of phase separation, for FUS-GFP [285] and for the N-terminal domain of FUS[289], also reflected in binding trends in the latter case. Given that triphosphate nucleotides are more negatively charged than the cyclic counterparts (-3 vs -1), it is interesting to note that these preferential interactions of the triphosphate nucleotide happen across a range of protein net charges; FUS-GFP and FUS NTD are positively charged at pH 7 (+5 and +2, respectively), while DDX4^{N1}NoCys is negatively charged (-4). This lack of correlation with the net charge is consistent with the DDX4 literature, since both DNA and RNA oligonucleotides, also negatively charged, partition into DDX4^{N1} *in vitro* condensates[254], as does messenger RNA (Hickling *et al.* , in preparation). Regarding protein partitioning, both a more positive and a more negative mutant form of GFP (in terms of surface charge) were more enriched than the WT GFP in DDX4^{N1} droplets[254]. Overall, the trends of partitioning of a range of proteins could not be explained by simple metrics like the pI or molecular weight; instead, there was a correlation with the amino acid composition (in particular, favouring proteins with a higher content of proline, arginine and tyrosine), which pointed at aromatic interactions being important for partitioning, as they are for the formation of the phase separated state.

Stronger effects/interactions of purine (polyA) versus pyrimidine (polyU and/or polyC) RNA homopolymers have been described for arginine-rich model peptides as well as full length FUS (which contains three RGG domains - see Figure 5.1)[298][301]. We have recapitulated this effect from single nucleosides, for the RGG domain of DDX4 (see Figure 5.10). Further, we have also been able to probe trends within the purine nucleobases (see Figure 5.7 and 5.10), which had not been possible in the context of RNA homopolymers given that polyG forms G-quadruplexes [301].

In both studies with cationic peptides, π-π or cation-π interactions between the nucleobases and cationic residues, arginines in particular, had been proposed to explain the observations. The phase separation of DDX4^{N1} is driven by cation-π

interactions between phenylalanine and arginine residues, very abundant in the sequence (6% and 10%, respectively), as shown by the destabilisation of the mutant containing DL-3-fluorophenylalanine instead of phenylalanine [148]. However, NMR studies of the condensed phase have revealed abundant intermolecular interactions involving most amino acids [156]. It is worth considering the wider range of interactions available to the nucleobases and residues in $DDX4^{N1}$, in order to interpret the observed trends.

Protein-nucleotide interactions have been an object of study for the last fifty years, with two main types of studies: database searches of PDB crystal structures and quantum chemical energy calculations, very often in combination, with the distance-based identified interactions being analysed computationally. What follows below is a review of the main trends identified and a discussion on how these can be applied to rationalise the partitioning trends observed, focusing on the chemical signatures of the bases and the sequence characteristics of $DDX4^{N1}$ already outlined in the introduction.

Cation-π interactions

Cation-π interactions are abundant in protein, protein-ligand and protein-nucleic acid structures[193]. Pioneer studies by Dougherty-Gallivan of a set of 600 protein structures combined a distance-based search criteria with *ab initio* / force-field energetic calculations, and found abundant cation-π interactions as well as showed biases in favour of tryptophan, as the aromatic residue, and arginine over lysine[189]. Cation-π interactions have received increasing attention in the phase separation field, given their importance for phase transitions and their regulation (by arginine methylation) as discussed in the introduction for both DDX4 and FUS [148][194].

Studies of DNA-protein and RNA-protein complexes have showed very abundant cation-π interactions, mostly involving arginine side chains, and with a higher frequency of purine versus pyrimidine bases[191][192]. The population trend for the DNA bases decreases according to G > A > C > T [191][312], while for RNA it follows G > A > U > C [192]. In these database mining studies, accompanying

energy calculations showed the same trends in the energetics of the interaction, highlighting in particular arginine-guanine as the strongest interaction, and showed that pyrimidine interactions can often be energetically unfavourable, in particular for thymine[191][192][312].

Regarding the cation side of the interaction, a quantum chemistry comparative study of the relative strength of arginine versus lysine for cation-π interactions has rationalised the higher prevalence of arginine in these interactions, even if lysine shows a stronger interaction in the gas-phase[313]. In the case of arginine, but not lysine, the geometries obtained by gas-phase had good agreement with those experimentally observed in a dataset of protein structures. The authors proposed that the arginine-arene interaction remains strong in a dielectric medium since it involves both electrostatic and dispersion attraction, in contrast to the electrostatic dominated lysine-arene interaction.

It seems that the trend we have observed of dG partitioning more than dA, and both more than the pyrimidines, agrees with the population trends of cation-π in crystal structures as well as the results of energetics calculations.

π-π interactions

Analysis of π-π stacking in DNA-protein X-ray crystal structures has shown that pyrimidines are involved in more interactions than the purines (with frequency decreasing as T > C > A > G), even if the strongest calculated energies correspond to interactions with C or G[196]. However, the self-stacking (π-π) tendency of nucleobases decreases in the order adenine > guanine > cytosine ≈ uracil; the fact that the purine bases self-associate more than the pyrimidine bases has been attributed to their larger aromatic systems [314].

Furthermore, gas-phase calculations of all pairwise DNA/RNA nucleobase-aromatic amino acid π-π stacking interactions showed that on average the interaction energy decreases in the order G > A > T ∼ U ∼ C (except for the specific tryptophan-nucleobase case where G > T > A ∼ U ∼ C), which suggests geometric constraints might apply in the context of structured complexes, especially including

double stranded DNA[315]. For RNA-protein complexes, the overall trend for nucleobase-amino acid interactions (grouping π-π and cation-π) is more sensitive to the amino acid (with Y,F and R overrepresented) than to the nature of the nucleobase (A>U G>C), perhaps reflecting the greater degree of conformational flexibility of RNA [197].

There is little agreement between the trends observed and the population frequencies for π-π interactions in either DNA-protein or RNA-protein complexes, yet better agreement regarding stronger self-stacking of purine versus pyrimidine nucleotides (yet with adenine having stronger self-association than guanine) and even better with gas phase calculated interaction strengths. Given there are no structural constraints in single nucleotides interacting in the context of a disordered protein, the latter would be a reasonable point of reference.

Hydrogen bonds

Hydrogen bonding is fundamental for the recognition of nucleic acids by proteins [305][316][317]. Non-specific bonding plays an important role in stability; two thirds of DNA-protein bonds involve the DNA backbone[316], and for RNA 20 % of bonds involve the 2'OH on the ribose. The original predictions for specific recognition of base-pairs in the DNA major and minor grooves were that bidentate hydrogen bonds would be able to distinguish between nucleobases better than single bonds[318]. The proposed interactions (compatible with the base-pairing bonds, see Figure 5.11 A and B) involved arginine (guanidinium group, with two donors) which would select for guanine, and asparagine or glutamine (carboxamide group, with one acceptor and one donor) for adenine, for the major groove (also known as the Hogsteen edge), and asparagine or glutamine would bind guanine in the minor groove (also known as the sugar edge). These predicted interactions have indeed been the most observed in database analysis of both DNA-protein and RNA-protein complexes, with the exception of the major groove Asn/Gln-Adenine interaction for RNA-protein complexes[316] [317]. Guanine in particular had the largest observed versus expected ratio of hydrogen bonds, which was attributed to it containing

the greatest number of potential hydrogen bonding atoms that are available for interaction at the edges of the G-C base pair[316]. The extraction of overall trends was complicated by the coexistence of several types of interactions in the dataset, including single interactions, bidentate interactions with a base or base-pair, and more complex interactions along the DNA double strand (the latter of which are beyond the scope of this discussion since our focus is on single nucleotides). Finally, an analysis of nucleobase - amino acid pseudo pairs in protein-nucleotide complexes (heavily biased in favour of complexes with adenine nucleotides) showed further trends of interactions also at the Watson-Crick edge, available in single nucleotides [319]. The authors highlighted that asparagine, glutamine, aspartic acid, glutamic acid and arginine can orient coplanarly with the bases and make two bonds at the Watson-crick edge; in addition, serine, threonine and also aspartic acid could make a double bond using both the side chain and backbone. Guanine interactions with Asp and Glu were seen at the Watson-Crick edge , the interaction with Asp (see Figure 5.11 C) being most represented, as well as with arginine in the Hogsteen edge as in the previous studies. For adenine, interactions with Asn,Gln, Asp and Glu were found both at the Watson-Crick (mainly Asn and Asp - see Figure 5.11 D) and Hogsteen edges, even if in the latter case the majority of interactions were with the protein backbone. The latter study is most relevant for the current discussion, since single nucleotides are not base-paired and therefore can bond with protein residues at any interaction edge, as well as being much smaller so less constrained geometrically. Overall, the fact that guanine was the most frequent nucleobase engaging in hydrogen bonds in protein-nucleic acid complexes [316] and frequently interacts with arginine and aspartic acid in single nucleotide complexes, both of which are abundant in DDX4, highlights the likely contribution of hydrogen bonding to the higher partitioning of dG vs dA. Relating to the correlation of partitioning with TPSA, guanine has one more hydrogen bond donor atom than adenine, which would allow it to satisfy four (vs three) hydrogen bonds.

Finally, and returning to the previous discussion of triphosphate nucleotides having higher partitioning than the cyclic counterparts, findings from protein-DNA

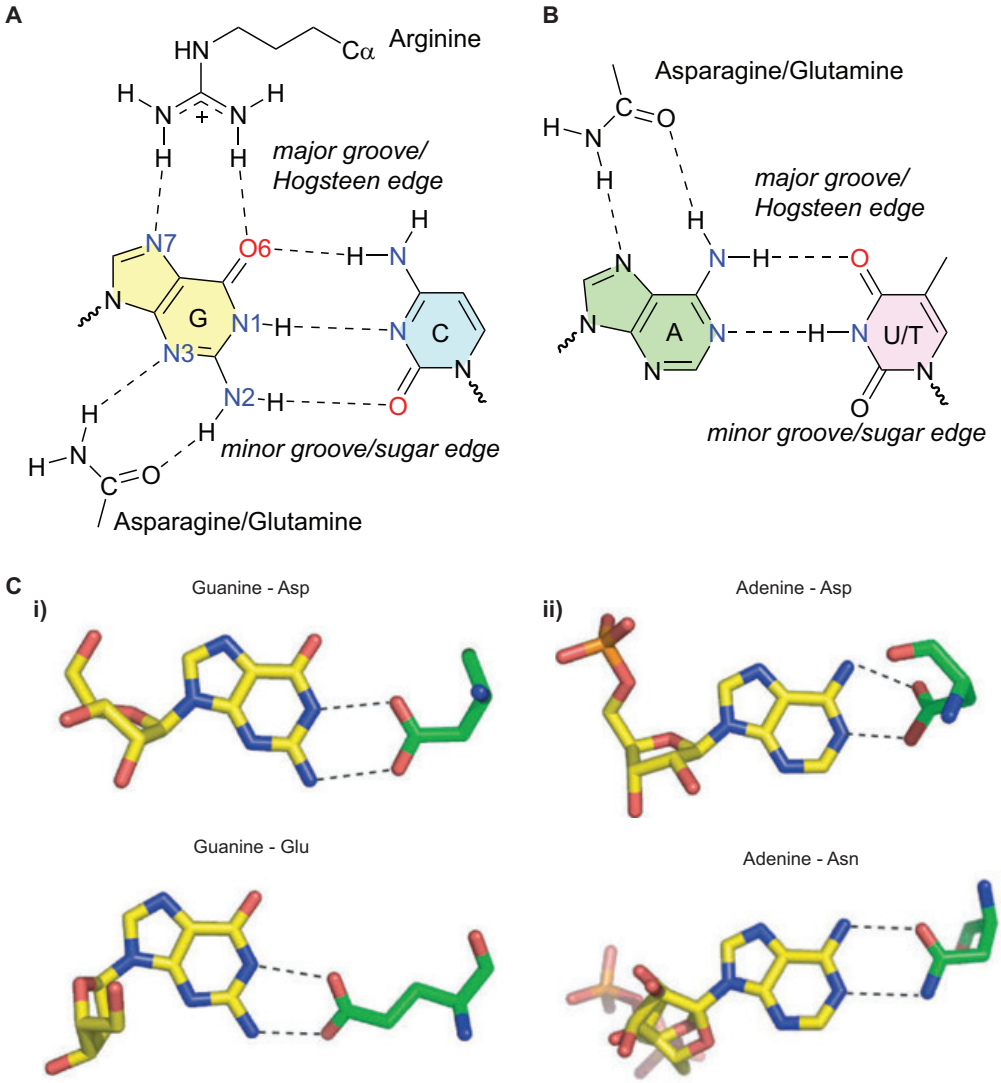


Figure 5.11: Hydrogen bonding of nucleobases. A. Guanine-Cytosine. B. Adenine-Thymine/uracil. A and B show the hydrogen bonds that form the Watson-Crick base-pairs of DNA/RNA nucleobases, as well as the bidentate bonding patterns of specific amino acids to the major and minor grooves of the bases, which have been proposed to contribute to selective recognition of the base pairs. C. Main pseudo-pairs observed between purine bases (guanine-asp/glu in i), adenine-Asp/Asn in ii)) and single nucleotides, with hydrogen bonds (sometimes water mediated, especially for adenine) at the Watson-crick edge. Taken from [319].

data mining studies also suggest that hydrogen bonding could explain the additional interactions of the triphosphate moiety (other than charge-charge interactions with cationic residues). Two thirds of DNA-protein H-bonds (from 129 crystal structures) involve the DNA backbone phosphate groups[316], and in particular the side chains of arginine, lysine, serine, theronine, asparagine and glutamine (in decreasing order of frequency). A more recent comparison using a larger structure database (over 3000 structures) showed similar trends with arginine being the highest contributor with two-fold as many hydrogen bonds as the next most prevalent amino acids (lysine, threonine and serine)[320]. Given that arginine (20 instances) and serine (33 instances) are the second and third most frequent amino acids in the 236 residue chain of $DDX4^{N1}$ (only after glycine, with 44 instances), it seems plausible that the potential for many such hydrogen bonds to stabilise backbone phosphate groups can explain the relatively higher partitioning of the triphosphate versus the cyclic nucleotides. Furthermore, it has been shown for ATP that it binds weakly and in a non specific way to proteins, especially disordered or in flexible regions, replacing protein-solvent hydrogen bonds (seen by NMR as reduced protein-solvent hydrogen exchange [297]).

Multiple interactions

Cation-π interactions have been found to frequently be accompanied by a hydrogen bond, for example for arginine binding to a DNA strand, which has been called the cation-π H-bond stair motif[321], and in cation-π interactions at protein-protein interfaces, which are very abundant and also mainly involve arginine [322].

Considerations on the correlations between partitioning trends and interaction trends

Overall, there is an good level of agreement between experimentally observed (and/or calculated) non-covalent interactions and the partitioning trends observed, which is better for cation-π and hydrogen bonding than for π-π.

If we consider the solvent properties of the condensate phase, we might place higher weight on the hydrogen bond capabilities as determining factors for the

partitioning. It has been shown using MD simulations that for guanine and cytosine in particular, the solvation environment changes the relative strength of aromatic versus hydrogen bond interactions.[323]. MD simulation extracted binding energies between all pairs of nucleobases and aromatic or charged amino acids were compared in two environments, water and methanol. The energies calculated for guanine and cytosine interactions showed the greatest variation with the nature of the solvent; guanine in particular had the strongest interactions in either solvent, but those were of π-π nature (with aromatic amino acids) in water, while in methanol the strongest interactions were hydrogen bonds with acidic residues. Further to that, energy calculations in a range of solvent mixtures (0-1 methanol fraction) for all bases with aspartate (representing hydrogen bonding) and tryptophan (π-π) showed a linear correlation between the strength of guanine/cytosine aspartate hydrogen bonding and the methanol fraction [324].

There is a large amount of acidic residues as part of the alternating charge blocks of DDX4 [148], which could engage in hydrogen bonding. An important contribution of the hydrogen bonding interactions to the partitioning would also be in agreement with the strong correlation between the TPSA metric and the partitioning trends. Finally, the fact that arginine has been seen to simultaneously engage in a cation-π interaction and a hydrogen-bond, and that it is very enriched in DDX4, could point to abundant combined multivalent interactions.

5.6 Conclusions

The study of partitioning of small molecules into protein phase separated condensates motivated the development of an NMR-based method and a semiautomated processing pipeline. NMR-based partitioning studies have the advantage of being label-free with respect to fluorescence approaches, as well as being informative on any chemical processes occurring in either dilute or condensate phase; however, it is necessary to physically separate and recover the two phases and measure the volume of condensate fraction. An initial protocol was designed for the study with lipoamide and FUS; as a follow up, another method for measuring the fraction

of condensate volume was developed. The two methods compared quite well in terms of pellet fraction, yet we chose to take forward the PCV tube based method given it would allow us to reduce the potential bias / variability in partitioning coming from the volume of supernatant fraction that remains with the pellet separated fraction. Finally, the design of relative partitioning experiments using mixtures of small molecules prompted and benefited from the development of data processing code, which used with some manual adjustments greatly sped up and systematised the analysis.

Regarding the implications of the results presented, the study of lipoamide partitioning into FUS condensates provides a potential mechanism of action for the regulation of phase separation by a small molecule, which is based on its enrichment into the condensate phase. This highlights the need (and opportunity) for new approaches to drug discovery for these systems [223].

Through a series of systematic comparisons of nucleotides partitioning into $DDX4^{N1}$ constructs (see Figure 5.12 A), we have established that all the molecules partitioned into the condensate phase, and that the condensate phase selectively enriched for some more than others. The finding of selective enrichment provides further evidence to the hypothesis that the condensate phase acts as a distinct solvent to the surrounding bulk, in agreement with selectivity being one of the properties of solvents.

Further, we investigated a few chemical features and observed two trends (see Figure 5.12 B). First, there was a positive contribution of the triphosphate group to partitioning, in line with some literature results but in this case also for a protein with a negative net charge. Second, the partitioning depended on the individual nucleobase in the order (higher PC to lower PC) $dG > dA \geq dC \sim dU$, which reproduces literature trends of purine vs pyrimidine RNA homopolymer - R-rich peptide coacervates but extends the chemical space probed by including guanine nucleotides, which were not amenable to RNA homopolymers studies due to the formation of G-quadruplexes. The partitioning trend within nucleobases correlated with the descriptor topological surface area (a count of heteroatoms) but

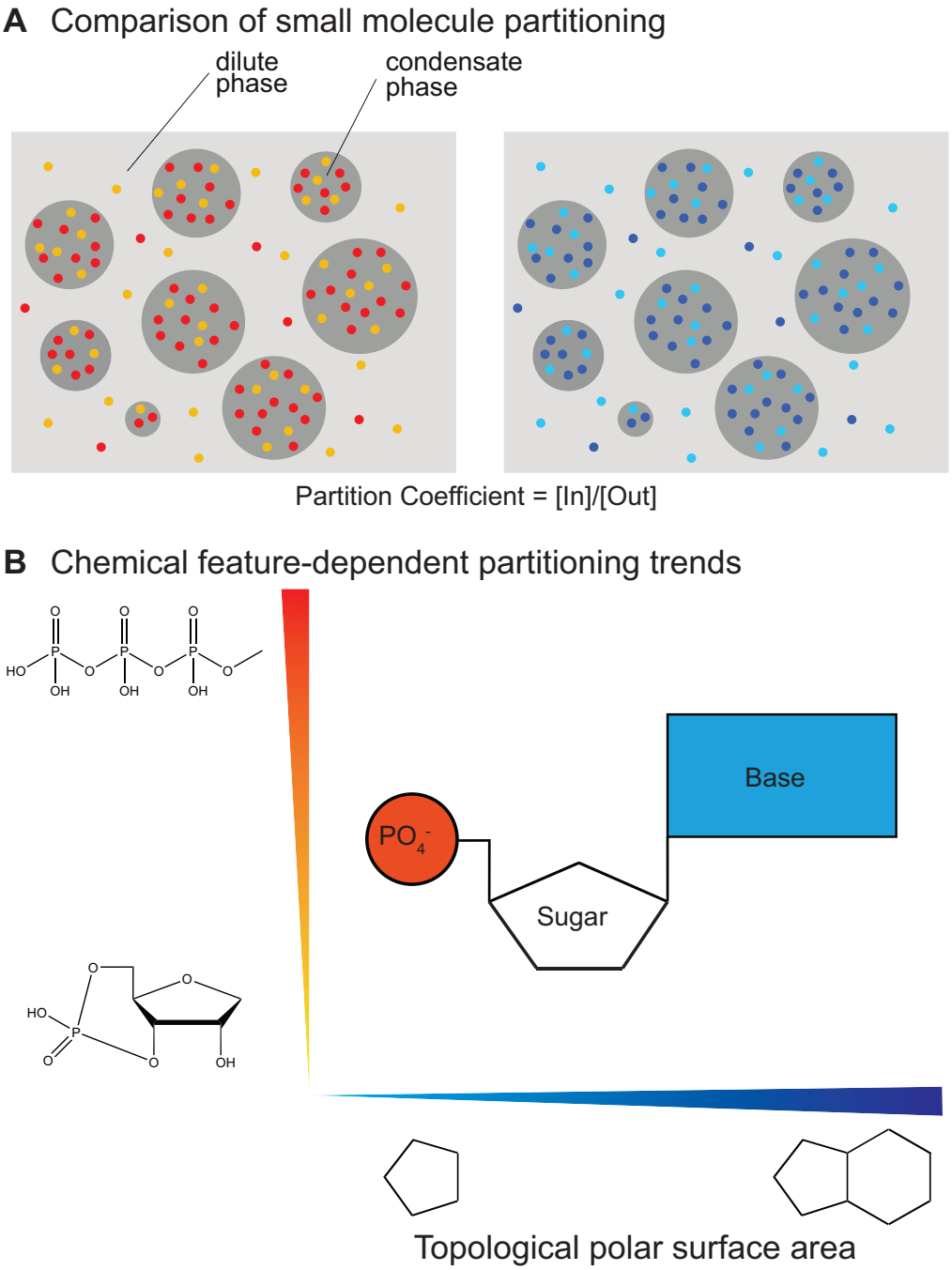


Figure 5.12: Schematic summary of the investigation of the influence of chemical features on the partitioning of nucleotides into DDX4 condensates. Nucleotides were selected as scaffolds for systematic analysis of chemical features along two dimensions of chemical space and their effects on relative partitioning into model system DDX4^{N1}. The top diagram shows a schematic of the results obtained, in which all nucleotides were enriched in the condensate phase, but their relative enrichment depended on charged phosphates (red vs orange) and the nucleobase (blue vs cyan). The colours correspond to the axis of partitioning represented in the bottom. A triphosphate moiety drives partitioning into the condensed phase, relative to the cyclic phosphate moiety. The nucleobase also influenced partitioning, with purines favoured over pyrimidines, with partition coefficients correlating with the topological polar surface area of the deoxynucleosides compared.

not with logP (a measure of solubility in octanol), which highlights the differences between incorporation into these compartments versus crossing of lipid bilayers. Considering the literature on non-covalent interactions between proteins and nucleic acids / nucleotides, we proposed that the partitioning trend could be rationalised via differences in strength of cation-π or hydrogen bonds; cation-π interactions have been shown to drive phase separation of DDX4 (amongst other proteins), while hydrogen bonding is a ubiquitous interaction within protein structures and in particular with nucleic acids, and is directly related to the descriptor TPSA, which is correlated with the trend. Further work remains to elucidate the relative contribution of these interactions, as well as probe a wider chemical space in terms of both phase separating protein and small molecules.

In conclusion, the study of small molecule interactions with the condensed phase of phase separating proteins can provide insights into the distinctive features of these environments, as well as inform strategies for the modulation of this fundamental cellular compartmentalisation mechanism.

5.7 Further work

Method development

As discussed regarding the data presented in Figure 5.9, there was a larger variation in the PC of thiamine in several samples. In contrast, the two samples of thiamine and deoxynucleosides (Figure 5.10) showed variability for thiamine PC in line with the interreplicate variability observed for the $DDX4^{N1}$ sample in Figure 5.8). Phase separation is very sensitive to variables such as temperature, pH and salt, yet those were controlled in our experimental design. However given the unexplained variability, more experiments are required to uncover any hidden variables and improve the consistency of the method. In particular, the buffer exchange step prior to partitioning experiments should be investigated as a source of variability, given the strong effect of D_2O on the phase diagram of $DDX4^{N1}$, consistent with the literature, and which suggests that small variations in D_2O could also have effects on the partitioning of small molecules.

Partitioning trends

The short term extension of the work presented in this chapter would involve completing the exploration of the chemical space of nucleotides in relation to their interactions with DDX4 in the condensate phase, as well as orthogonal studies in proteins with different sequence composition features, such as the FUS N-terminal PLD domain (rich in tyrosines, devoid of cationic residues). In particular, we could probe: 1)the influence of cyclisation (AMP vs cAMP), 2) the phosphate contributions sequentially (ATP vs ADP vs AMP), 3) the contribution of the adenine base (ATP vs tripolyphosphate (TPP), 4) the effect of the 2'OH group on the ribose base (e.g. A vs dA) and 5) the relative contribution of cation-π and hydrogen bonds to the partitioning (e.g. introducing a fluorine into the nucleobase, vs unmodified);the latter three follow-up experiments are rationalised in more detail below.

ATP vs TPP As discussed in the introduction, there is some contradictory evidence on whether the TPP moiety contributes to the phase separation effect of nucleotides, in particular ATP. In the study that showed higher decondensation boundary for R-rich peptides than K-rich peptides, in mixtures with homopolymeric RNA (pA or pU), polyphosphate did not show differential boundaries for the two peptides [298]. Similarly, TPP had lower effects than ATP on phase separation of FUS [285]. However, in another study the phase diagram of an IDP with TPP was sensitive to the arginine/lysine content of a series of mutants [304]. Therefore it will be interesting to compare the partitioning of TPP with ATP / GTP, in order to deconvolve the contributions from the sugar-base and from the triphosphate. We might expect lower partitioning of TPP based on the loss of probable aromatic interactions of the base; however, triphosphate would also be a smaller molecule, which might facilitate more non-specific interactions.

RNA vs DNA bases Regarding the 2'OH characteristic of RNA), this group is involved in 20% of hydrogen bonds[325] and 14% of sugar-pi interactions in RNA-

protein complexes [197]. Therefore we would hypothesize that the corresponding ribonucleosides would partition more into DDX4 than the deoxyribonucleosides.

Fluorination of nucleobases The introduction of a fluorine atom as an aromatic substituent in the ring of a nucleobase would have a large impact on the charge density distribution, which is crucial for cation-π interactions; in contrast, it would add another heteroatom able to accept a hydrogen bond. This could help us uncover the relative contribution of aromatic vs hydrogen bonding to the partitioning of these molecules into this particular phase separation model system.

In the longer term, we would aim to expand the chemical space probed in our systematic tests, as well as use NMR structural and exchange methods to investigate the relative strengths and effects of the interactions proposed to drive the partitioning trends observed.

6 Conclusions

In this thesis, I have explored questions regarding protein assembly, both physiological and pathological, from a biophysical point of view and using mainly NMR, a technique particularly suited to these systems. The results are relevant at both extremes of the continuum between basic and translational research, given that a more nuanced understanding of fundamental phenomena is intrinsically linked with opportunities to modulate them in novel ways. This is particularly relevant in these areas of research, since protein aggregation and phase separation are associated with severe human diseases that have few therapeutic options and a high rate of clinical trials failure. It is a belief shared by many researchers, which I subscribe to, that the lack of a detailed mechanistic understanding of these processes is a major roadblock in the drug discovery pipeline. I present below a few concluding remarks, first regarding the tau project presented in chapters 2, 3 and 4, and finally the phase separation work in chapter 5.

6.1 Mechanism of aggregation and inhibition of tau

There is an extensive rich body of literature on the aggregation of tau, yet there are still many open questions which have challenged the excellent researchers

in the field for decades. The main question that I believe this work contributes to answer is the nature of the nucleus of the aggregation pathway of tau. This has remained ill-defined throughout mechanistic studies of several model systems. We have obtained a model for aggregation that includes disulfide-bond mediated dimerisation and the formation of a tetrameric nucleus from the assembly of a dimer and two monomers, as well as elongation and fragmentation. This has been extensively tested against other nucleation mechanisms with nuclei in the range of dimer through to hexamer, which I believe is the most exhaustive set of comparisons performed for a tau construct.

Several challenges, some general to protein aggregation studies, and others specific to the case of tau, had to be overcome in order to obtain this result. A particular complication in the study of tau, in comparison with amyloid β (the other protein implicated in Alzheimer's disease), is its high solubility and low propensity to aggregate in *in vitro* settings. This has traditionally required the use of inducers of aggregation, shorter more aggregation prone constructs, or both, although recently there has been an increasing effort towards inducer-free systems. This switch is related to the long ongoing controversy in the field regarding the biological relevance of studies using inducers, and to which the recent developments of cryoEM have added another dimension, with some studies providing evidence of differences of structure between *in vitro* fibrils formed with inducers and fibrils from patient tissues, yet others showing endogenous cofactors similar to the inducers commonly used present in *in vivo* fibrils.

We wanted to work with an inducer-free system, given that the modelling methods developed previously in our lab allow us to model the intrinsic aggregation pathway of the protein independently from the actions of other biomolecules, such as inducers or inhibitors, the actions of which are probed in an explicit manner. Therefore, elucidating a mechanism of a "purely tau" system is only the first step towards modelling the influence of other molecules of biological or therapeutic interest. In this sense, another contribution of this work is the

detailed characterisation of the kinetic effects of a literature inhibitor of tau, D-TLKIVW, developed in a pioneering computational study, and which we show has a multifaceted mechanism of action.

The model system we used, "R3", a 26mer that constitutes most of the R3 repeat of tau, located in its microtubule binding repeat domain, and which begins with the hexapeptide VQIVYK sequence, one of the two hotspots of aggregation. R3 aggregated in a reproducible manner, in the absence of inducers or agitation, in a suitable scale for NMR-based monitoring, and the aggregates formed had the ability to template the aggregation of full length endogenous cellular tau, in agreement with other studies that defined R3 as the minimal seeding-competent construct. Further, two distinct chemical signatures were resolved in 1D ^{1}H spectra for the two C terminal histidine residues, and which corresponded to a monomer and a dimer. The monitoring of the dimer trace, in particular, was instrumental in discriminating between nucleation mechanisms.

Finally, the optimisation procedure leading to the mechanistic descriptions above-mentioned is the subject of a substantial part of the thesis, in line with the difficulty of modelling such processes as systems of nonlinear ODEs. This work extended the previous method, which combined a local search algorithm with a global search basin-hopping like routine, by adding a top layer algorithm at the level of the model space. This algorithm combined elements of local search (incorporating distinct steps in parallel, and choosing the path with the best goodness of fit) and global perturbation, by adding new chemical steps and combining different modes of initialisation of the parameters. In combination with developments in other layers of the optimiser, we achieved a strategy that allowed us to find a best model for each dataset which represents a global minimum both in model and parameter space.

6.2 Probing the chemical environment of *in vitro* biomolecular condensates

Phase separation is emerging as a key mechanism by which eukaryotic organisms (and probably prokaryotic too) organise processes in cells in a complementary way

to the traditional membrane-based compartmentalisation. From a physicochemical point of view, an exciting area of research is the hypothesis that the inside of the condensates (also known as membraneless organelles) constitutes a *true* chemical solvent, for which there is increasing evidence at multiple levels. One aspect of condensates that permeates the literature is their complex composition, yet also selectivity, in terms of proteins, nucleic acids and small molecules, which is increasingly linked to functions such as the regulation of enzymatic processes. In the absence of delimiting membranes, such selectivity seems analogous to that of a solvent, and it follows to consider the partitioning trends of biomolecules in condensates. Again, this has implications for the basic understanding of phase separation, as well as huge translational potential given that the proteins typically found, and driving, the formation of condensates are rich in disordered regions, which turns them into classically *undruggable* targets and asks for novel strategies for their modulation based on understanding the chemical rules in play.

Early work in this thesis addressed a specific case of phase separation modulation, where lipoamide was found to alter phase separation of FUS, a protein linked to neurodegenerative diseases. The results of this collaboration contributed a piece of evidence concerning the mechanism of action of lipoamide, which we showed was enriched by ten-fold in the condensate phase. This project required the development of a method to study partitioning with NMR, including the measurement of the condensate volume fraction. Next, while further developing this method, we addressed the question of selectivity by studying partitioning into another phase separating system, DDX4, a component of the nuage MLO in male germ cells. We explored chemical trends in partitioning by comparing the partitioning of nucleotides, which have a well defined scaffold yet naturally rich in chemical variation, and which are the building blocks of nucleic acids, essential components and regulators of phase separation *in vivo*. We observed preferential partitioning of purine versus pyrimidine nucleobase nucleosides, as well as enrichment of triphosphate nucleotides versus cyclic counterparts. Our results are in agreement with the "solvent hypothesis", and the methods developed are label-free

and thus generalisable to studies of a wide range of chemical features in orthogonal protein systems, which I hope will be useful in improving our understanding of this aspect of phase separation.

Overall, this thesis highlights the value of detailed mechanistic studies of protein aggregation and phase separation, from a fundamental and applied point of view; I think the work presented makes modest but valuable contributions to these exciting fields of research, in terms of insights on specific systems as well as generally applicable methods developed to meet (some of) the abundant challenges.

Appendices

A
Appendix

Contents

A.1 Full implementation of best R3 + D-TLKIVW mechanistic model 249
A.2 Fitted parameters for best models 260

A.1 Full implementation of best R3 + D-TLKIVW mechanistic model

Implementation

//nucleation_2

//forward

double kpcurr1 = kp[0]*xpt[0]*xpt[0];

xp[0] += -kpcurr1;

xp[0] += -kpcurr1;

xp[1] += kpcurr1;

//backward

double kmcurr1 = km[0]*xpt[1];

xp[1] += -kmcurr1;

Chemical equilibria

Dimerisation

$$\mathrm{P}_1 + \mathrm{P}_1 \overset{kp[0}{\underset{}{\rightleftharpoons}}][\mathrm{km}[0\,]]\mathrm{P}_2 \qquad \text{(A.1)}$$

```
xp[0] += kmcurr1;
xp[0] += kmcurr1;
```

```
//nucleation_n
//forward
double kpcurr2 = kp[1]*xpt[1]*pow(xpt[0],
kp[2]);
xp[0] += -kpcurr2*kp[2];
xp[1] += -kpcurr2;
xp[4] += kpcurr2;
//backward
double kmcurr2 = km[1]*xpt[4];
xp[4] += -kmcurr2;
xp[0] += kmcurr2*kp[2];
xp[1] += kmcurr2;
```

Nucleation (steady-state approximation)

$$\mathrm{P_2} + 2\,\mathrm{P_1} \underset{km[1]}{\overset{kp[1]}{\rightleftharpoons}} \mathrm{P_4} \qquad \text{(A.2)}$$

```
for(int i=4; i<N/2-5;++i)
{
//elongation_1
//forward
double kpcurr3 = kp[i]*xpt[i]*xpt[0];
xp[i] += -kpcurr3;
xp[0] += -kpcurr3;
xp[i+1] += kpcurr3;
//backward
double kmcurr3 = km[i]*xpt[i+1];
xp[i+1] += -kmcurr3;
xp[i] += kmcurr3;
xp[0] += kmcurr3;
```

Monomer elongation

// free fibrils

$$\mathrm{Pi} + \mathrm{P_1} \underset{km[i]}{\overset{kp[i]}{\rightleftharpoons}} \mathrm{P_{i+1}} \qquad \text{(A.3)}$$

```
//elongation_2
//forward
double kpcurr4 = kp[N/2+i]*xpt[i]*xpt[1];
xp[i] += -kpcurr4;
xp[1] += -kpcurr4;
xp[i+2] += kpcurr4;
//backward
double kmcurr4 = km[N/2+i]*xpt[i+2];
xp[i+2] += -kmcurr4;
xp[i] += kmcurr4;
xp[1] += kmcurr4;
```

Dimer elongation

// free fibrils

$$\mathrm{P_i} + \mathrm{P_2} \underset{km[N/2+i]}{\overset{kp[N/2+i]}{\rightleftharpoons}} \mathrm{P_{i+2}} \tag{A.4}$$

```
//fragmentation
//fragmentation corrections
int nucCorr= +2 - round(kp[2]);
double nucRem = kp[2]-round(kp[2]);
//forward
if (i > 10)
{
int frag = i/2;
double kpdcurr5 = kpd[0]*xpt[i];
xp[i] += -kpdcurr5;
xp[frag] += kpdcurr5;
xp[i-frag +nucCorr] += kpdcurr5*(i-frag)/(i-
frag+nucRem);
```

Fragmentation of oligomers / fibrils

// free fibrils

$$\mathrm{P}_i \xrightarrow{kpd[0]} \mathrm{P}_{(i/2)} + \mathrm{P}_{(i/2)} \tag{A.5}$$

```
//inhibitor-bound oligomers fragment (same
rate)
int frag_j = i/2;
double kpdfrag = kpd[0]*xptC[i];
```

// inhibitor-bound fibrils

$$\mathrm{P}_i, bound \xrightarrow{kpd[0]} \mathrm{P}_{(i/2,bound)} + \mathrm{P}_{(i/2,bound)} \tag{A.6}$$

```
xpC[i] += -kpdfrag;
xpC[frag_j] += kpdfrag;
xpC[i-frag_j +nucCorr] += kpdfrag*(i-
frag_j)/(i-frag_j+nucRem);
}
```

```
//oligomers + drug elongate by monomer
(same rate)
//forward
double kpcurrelm = kp[i]*xptC[i]*xpt[0];
xpC[i] += -kpcurrelm;
xp[0] += -kpcurrelm;
xpC[i+1] += kpcurrelm;
if (i==4)
{
xpC[1] += -kpcurrelm*4; //for mass balance
if elongating nucleus
}
else
{
xpC[1] += -kpcurrelm;
}
//backward
double kmcurrelm = km[i]*xptC[i+1];
xpC[i+1] += -kmcurrelm;
xpC[i] += kmcurrelm;
xp[0] += kmcurrelm;
if (i==4)
{
xpC[1] += +kmcurrelm*4; //for mass
```

Monomer elongation (of inhibitor-bound oligomers/fibrils)

```
// by free monomer
// inhibitor-bound fibrils
```

$$\mathrm{P_{i,bound}} + \mathrm{P_1} \underset{km[i]}{\overset{kp[i]}{\rightleftharpoons}} \mathrm{P_{i+1,bound}} \qquad \text{(A.7)}$$

```
balance if elongating nucleus
}
else
{
xpC[1] += +kmcurrelm;
}
                                        // by inhibitor-monomer complex
//inhibitor-bound oligomers elongate by
inhibitor-monomer complex (same rate)
//forward
double kpcurrelbm = kp[i]*xptC[i]*xptC[2];
xpC[i] += -kpcurrelbm;
xpC[2] += -kpcurrelbm;
xpC[i+1] += kpcurrelbm;
if (i==4)
{
xpC[1] += -kpcurrelbm*3; //for mass
balance if elongating nucleus
}
//backward
double kmcurrelbm = km[i]*xptC[i+1];
xpC[i+1] += -kmcurrelbm;
xpC[i] += kmcurrelbm;
xpC[2] += kmcurrelbm;
if (i==4)
{
xpC[1] += +kmcurrelbm*3; //for mass
balance if elongating nucleus
}
```

$$\mathrm{P_{i,bound}} + \mathrm{P_1I} \underset{km[i]}{\overset{kp[i]}{\rightleftharpoons}} \mathrm{P_{i+1,bound}} \quad \text{(A.8)}$$

```
//inhibitor-bound oligomers elongate by dimer (different rate)
//forward
double kpcurreldd = kpd[3]*xptC[i]*xptC[3];// by dimer-inhibitor complex
xpC[i] += -kpcurreldd;
xpC[3] += -kpcurreldd;
xpC[i+2] += kpcurreldd;
if (i==4)
{
xpC[1] += -kpcurreldd*4; //for mass balance if elongating nucleus
}
else
{
xpC[1] += -kpcurreldd; //for mass balance if elongating species larger than nucleus
}
//backward
double kmcurreldd = kmd[3]*xptC[i+2];
xpC[i+2] += -kmcurreldd;
xpC[i] += kmcurreldd;
xpC[3] += kmcurreldd;
if (i==4)
{
xpC[1] += kmcurreldd*4; //for mass balance if elongating nucleus
}
else
{
xpC[1] += kmcurreldd; //for mass balance
```

Dimer elongation (of inhibitor-bound oligomers/fibrils)

// inhibitor-bound fibrils

$$\mathrm{P_{i,bound}} + \mathrm{P_2} \underset{kmd[3]}{\overset{kpd[3]}{\rightleftharpoons}} \mathrm{P_{i+2,bound}} \quad \text{(A.9)}$$

if elongating species larger than nucleus

}

if (i>4)

{ **Inhibitor surface-binding to oligomers/fibrils**

//Inhibitor binds oligomers (surface binding,

1 per protein molecule)

$$\mathrm{P}_i\mathrm{free} + \mathrm{iI} \underset{kmd[5]}{\overset{kpd[5]}{\rightleftharpoons}} \mathrm{P}_i\mathrm{bound}(\mathrm{I_i}) \quad \text{(A.10)}$$

//forward reaction

double kpcurrolis = kpd[5]*xpt[i]*xptC[1];

xp[i] -= kpcurrolis; // oligomeric protein

xpC[1] -= kpcurrolis*i; // free inhibitor.

1 molecule per protein (surface)

xpC[i] += kpcurrolis; // protein + inhibitor

(bound)

// backward reaction

double kmcurrolis = kmd[5]*xptC[i];

xp[i] += kmcurrolis; // oligomeric protein

xpC[1] += kmcurrolis*i; // free inhibitor.

1 molecule per protein (surface)

xpC[i] -= kmcurrolis; // protein + inhibitor

(bound)

}

//Inhibitor dissociates dimer from oligomer **Inhibitor dissociates dimer from fibrils**

//forward reaction

if (i>5)

$$\mathrm{P}_i \xrightarrow{kpd[6]} \mathrm{P}_{i-2} + \mathrm{P}_2 \quad \text{(A.11)}$$

{

//forward reaction

double kpdcurrdiss= kpd[6]*xpt[i]*xptC[1];

xp[i] -= kpdcurrdiss; // oligomer

```
xp[i-2] += kpdcurrdiss; // shortened oligomer
released
xp[1] += kpdcurrdiss; // free dimer
} // end of if loop for dimer dissociation

} //end of big for loop

//additional inhibitor steps
```

Inhibitor binds monomer

```
//Inhibitor binds monomer
//forward reaction
double kpdcurrmon = kpd[8]*xpt[0]*xptC[1];
xp[0] -= kpdcurrmon; // monomeric protein
xpC[1] -= kpdcurrmon; // free inhibitor
xpC[2] += kpdcurrmon; // protein +
inhibitor (bound)
//backward reaction
double kmdcurrmon = kmd[8]*xptC[2];

xp[0] += kmdcurrmon; // monomeric
protein
xpC[1] += kmdcurrmon; // free inhibitor
xpC[2] -= kmdcurrmon; // protein +
inhibitor (bound)
```

$$P_1 + I \underset{kmd[8]}{\overset{kpd[8]}{\rightleftharpoons}} P_1I_1 \qquad \text{(A.12)}$$

Inhibitor binds dimer

```
//Inhibitor binds dimer
//forward reaction
double kpdcurrdim = kpd[1]*xpt[1]*xptC[1];
xp[1] -= kpdcurrdim; // dimeric protein
xpC[1] -= kpdcurrdim; // free inhibitor
```

$$P_2 + I \underset{kmd[1]}{\overset{kpd[1]}{\rightleftharpoons}} P_2I \qquad \text{(A.13)}$$

```
xpC[3] += kpdcurrdim; // dimeric protein
+ inhibitor (bound)
//backward reaction
double kmdcurrdim = kmd[1]*xptC[3];

xp[1] += kmdcurrdim; // dimeric protein
xpC[1] += kmdcurrdim; // free inhibitor
xpC[3] -= kmdcurrdim; // dimeric protein
+ inhibitor (bound)
```

Inhibitor binds nucleus

```
//Inhibitor binds nucleus
//forward reaction
double kpdcurrnuc = kpd[4]*xpt[4]*xptC[1];
xp[4] -= kpdcurrnuc; // dimeric protein
xpC[1] -= kpdcurrnuc; // free inhibitor
xpC[4] += kpdcurrnuc; // dimeric protein
+ inhibitor (bound)
//backward reaction
double kmdcurrnuc = kmd[4]*xptC[4];

xp[4] += kmdcurrnuc; // dimeric protein
xpC[1] += kmdcurrnuc; // free inhibitor
xpC[4] -= kmdcurrnuc; // dimeric protein
+ inhibitor (bound)
```

$$\mathrm{P}_4 + \mathrm{I} \underset{kmd[4]}{\overset{kpd[4]}{\rightleftharpoons}} \mathrm{P}_4\mathrm{I} \qquad \text{(A.14)}$$

Inhibitor disrupts nucleus

```
// Inhibitor disrupts nucleus
// forward reaction
double kpdcurrdisn = kpd[2]*xpt[4]*xptC[1];
xp[4] -= kpdcurrdisn; // nucleus (xp[4])
xp[1] += kpdcurrdisn; // release dimer
```

$$\mathrm{P}_4 + \mathrm{I} \xrightarrow{kpd[2]} \mathrm{P}_2 + 2\,\mathrm{P}_1 + \mathrm{I} \qquad \text{(A.15)}$$

```
xp[0] += kpdcurrdisn*kp[2]; // nucleus
(xp[4]) + inhibitor (bound)
```

Mass balance calculation

```
////calculate aggregation concentration
double model::CalcConcX()
{
double fConc=0.0;
fConc+= xp[0]*(+1);
fConc+= xp[1]*(+2);
fConc+= xp[4]*(+*(params->m["kp13"])+2);
for(int i=0;i<N-(5);++i)
{
fConc+= xp[5+i]*(+*(params->m["kp13"])+i+3);
}
for(int i=0;i<N-(5);++i)
{
fConc+= xpC[5+i]*(+*(params->m["kp13"])+i+3);
}
fConc+= xpC[2]*(+1);
fConc+= xpC[3]*(+2);
fConc+= xpC[4]*(+*(params->m["kp13"])+2);
return fConc;
}
////calculate chaperone concentration
double model::CalcConcC()
{
double cConc=0.0;
```

```
cConc+= xpC[1]*(+1);
cConc+= xpC[2]*(+1);
cConc+= xpC[3]*(+1);
cConc+= xpC[4]*(+1);
for(int i=0;i<N-(5);++i)
{
cConc+= xpC[5+i]*(+i+5);
}
return cConc;
}
```

The winning model of the R3-only second dataset had one additional chemical step, the monomer conformational change (see Figure **4.2**, step k):

$$\mathrm{P}_1 \underset{k_-}{\overset{k_+}{\rightleftharpoons}} \mathrm{P}_{1,\mathrm{changed}} \tag{A.1}$$

Additionally, the dimerisation and nucleation reactions included both P_1 and $P_{1,changed}$, and elongation used up $P_{1,changed}$ (see Figure **4.2** B). Array locations of each species (within xp) were adjusted as required, following the same principles as the above implementation.

A.2 Fitted parameters for best models

Table A.1 relates the rate constant array positions from the implementation above with the code of the chemical equilibria in Figure 4.2 (protein-only steps) or 4.3 (inhibitor steps), the reaction they correspond to and their fitted values in the best models for the R3-only 1st dataset winning model (9 rates) and the R3 + D-TLKIVW best model (remaining rates).

Table A.1: Rate constants fitted for the R3-only best model (first dataset) and for the R3 + D-TLKIVW best model. The reaction codes correspond to the depictions in Figure **4.2** (protein-only steps) and Figure **4.3** (inhibitor steps). Reaction rates have units of h^{-1} (unimolecular, such as km[0]) or μM^{-1} h^{-1} (bimolecular, such as kp[0]), with the exception of those of the nucleation steady-state approximation (step h2).

rate (array)	Name in Figure 4.2 / 4.3	Meaning	Fitted value
kp[0]	c	Dimerisation (forward)	$5.7 \cdot 10^{-4}$
km[0]	c	Dimerisation (backward)	$3.3 \cdot 10^{-4}$
kp[1]	h2	Nucleation (forward)	$9.0 \cdot 10^{-8}$
km[1]	h2	Nucleation (backward)	1.3
kp[4...N/2]	a / i'1b / i'3b	Monomer elongation (forward)	$2.7 \cdot 10^{-1}$
km[4...N/2]	a / i'1b / i'3b	Monomer elongation (backward)	$7.3 \cdot 10^{-1}$
kp[N/2+0...N]	b	Dimer elongation (forward)	20
km[N/2...N]	b	Dimer elongation (backward)	29
kpd[0]	i	Fragmentation (forward)	$1.3 \cdot 10^{-1}$
kpd[8]	a'	Inhibitor - monomer binding (forward)	$9.7 \cdot 10^{-4}$
kmd[8]	a'	Inhibitor - monomer binding (backward)	1.4
kpd[1]	b'1	Inhibitor - dimer binding (forward)	$2.1 \cdot 10^{-3}$
kmd[1]	b'1	Inhibitor - dimer binding (backward)	$7.1 \cdot 10^{-2}$
kpd[4]	c'1	Inhibitor - nucleus binding (forward)	1.1
kmd[4]	c'1	Inhibitor - nucleus binding (backward)	$4.3 \cdot 10^{-3}$
kpd[5]	d'2	Inhibitor - fibril surface binding (forward)	$1.4 \cdot 10^{-5}$
kmd[5]	d'2	Inhibitor - fibril surface binding (backward)	$3.2 \cdot 10^{-5}$
kpd[2]	f'1	Inhibitor - nucleus disruption (forward)	4.2
kdp[6]	g'1	Inhibitor - dissociates free dimer (forward)	$4.3 \cdot 10^{-1}$
kpd[3]	j'4b	Inhibitor-bound fibrils elongate by dimer-inhibitor complex (forward)	$2.1 \cdot 10^{-2}$
kmd[3]	j'4b	Inhibitor-bound fibrils elongate by dimer-inhibitor complex (backward)	$1.4 \cdot 10^{-3}$

Table A.2 details the rate constant names, meaning and fitted values for the best model fitted to the R3-only second dataset.

Table A.2: Rate constants fitted for the R3-only second dataset best model. The reaction codes correspond to the depictions in Figure **4.2**.Reaction rates have units of h^{-1} (unimolecular, such as km[0]) or μM^{-1} h^{-1} (bimolecular, such as kp[0]), with the exception of those of the nucleation steady-state approximation (step h2b).

Code in Figure 4.2	Meaning	
k	Monomer conformational change (forward)	$1.5 \cdot 10^{-2}$
k	Monomer conformational change (forward)	$3.7 \cdot 10^{-1}$
c3	Dimerisation (forward)	$2.7 \cdot 10^{-1}$
c3	Dimerisation (backward)	$1.9 \cdot 10^{-4}$
h2b	Nucleation (forward)	5.1
h2b	Nucleation (backward)	$6.5 \cdot 10^{-1}$
a3	Changed monomer elongation (forward)	2.6
a3	Changed monomer elongation (backward)	$6.4 \cdot 10^{-1}$
b	Dimer elongation (forward)	$2.2 \cdot 10^{-2}$
b	Dimer elongation (backward)	$7.2 \cdot 10^{-4}$
i	Fragmentation (forward)	$7.9 \cdot 10^{-2}$

References

(1) Chiti, F.; Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annual Review of Biochemistry* **2017**, *86*, 27–68.

(2) Hartl, F. U.; Bracher, A.; Hayer-Hartl, M. Molecular chaperones in protein folding and proteostasis. *Nature* **2011**, *475*, 324–332.

(3) Dobson, C. M. The Amyloid Phenomenon and Its Links with Human Disease. *Cold Spring Harbor Perspectives in Biology* **2017**, a023648.

(4) Marsh, J. A.; Teichmann, S. A. Structure, Dynamics, Assembly, and Evolution of Protein Complexes. *Annual Review of Biochemistry* **2015**, *84*, 551–575.

(5) Dutta, S.; Berman, H. M. Large Macromolecular Complexes in the Protein Data Bank: A Status Report. *Structure* **2005**, *13*, 381–388.

(6) Hohmann, T.; Dehghani, F. The Cytoskeleton-A Complex Interacting Meshwork. *Cells* **2019**, *8*, doi: 10.3390/cells8040362, 362.

(7) Evans, G. Protecting us... even our tears... 2021, `https://www.ebi.ac.uk/pdbe/news/protecting-us-even-our-tears`.

(8) Goodsell, D. Molecule of the Month: Ribosomal Subunits, 2000, `https://pdb101.rcsb.org/motm/10`.

(9) Fletcher, D. A.; Mullins, R. D. Cell mechanics and the cytoskeleton. *Nature* **2010**, *463*, 485–492.

(10) Wright, P. E.; Dyson, H. J. Intrinsically Disordered Proteins in Cellular Signaling and Regulation. *Nature Reviews Molecular Cell Biology* **2015**, *16*, 18–29.

(11) Lafontaine, D. L.; Riback, J. A.; Bascetin, R.; Brangwynne, C. P. The nucleolus as a multiphase liquid condensate. *Nature Reviews Molecular Cell Biology* **2021**, *22*, 165–182.

(12) Wright, P. E.; Dyson, H. J. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology* **1999**, *293*, 321–331.

(13) Ward, J. J.; Sodhi, J. S.; McGuffin, L. J.; Buxton, B. F.; Jones, D. T. Prediction and Functional Analysis of Native Disorder in Proteins from the Three Kingdoms of Life. *Journal of Molecular Biology* **2004**, *337*, 635–645.

(14) Dunker, A. K.; Silman, I.; Uversky, V. N.; Sussman, J. L. Function and structure of inherently disordered proteins. *Current Opinion in Structural Biology* **2008**, *18*, 756–764.

(15) Hyman, A. A.; Weber, C. A.; Jülicher, F. Liquid-Liquid Phase Separation in Biology. *Annual Review of Cell and Developmental Biology* **2014**, *30*, 39–58.

(16) Banani, S. F.; Lee, H. O.; Hyman, A. A.; Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nature Reviews Molecular Cell Biology* **2017**, *18*, 285–298.

(17) Ross, C. A.; Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nature Medicine* **2004**, *10*, S10–S17.

(18) Greenwald, J.; Riek, R. Biology of Amyloid: Structure, Function, and Regulation. *Structure* **2010**, *18*, 1244–1260.

(19) Iadanza, M. G.; Jackson, M. P.; Hewitt, E. W.; Ranson, N. A.; Radford, S. E. A new era for understanding amyloid structures and disease. *Nature Reviews Molecular Cell Biology* **2018**, *19*, 755–773.

(20) Fowler, D. M.; Koulov, A. V.; Alory-Jost, C.; Marks, M. S.; Balch, W. E.; Kelly, J. W. Functional Amyloid Formation within Mammalian Tissue. *PLoS Biology* **2005**, *4*, e6.

(21) Maji, S. K.; Perrin, M. H.; Sawaya, M. R.; Jessberger, S.; Vadodaria, K.; Rissman, R. A.; Singru, P. S.; Nilsson, K. P. R.; Simon, R.; Schubert, D.; Eisenberg, D.; Rivier, J.; Sawchenko, P.; Vale, W.; Riek, R. Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* **2009**, *325*, 328–332.

(22) Saad, S.; Cereghetti, G.; Feng, Y.; Picotti, P.; Peter, M.; Dechant, R. Reversible protein aggregation is a protective mechanism to ensure cell cycle restart after stress. *Nature Cell Biology* **2017**, *19*, 1202–1213.

(23) Tycko, R. Amyloid Polymorphism: Structural Basis and Neurobiological Relevance. *Neuron* **2015**, *86*, 632–645.

(24) Housmans, J. A. J.; Wu, G.; Schymkowitz, J.; Rousseau, F. A guide to studying protein aggregation. *The FEBS Journal* **2023**, *290*, 554–583.

(25) Baldwin, A. J.; Knowles, T. P.; Tartaglia, G. G.; Fitzpatrick, A. W.; Devlin, G. L.; Shammas, S. L.; Waudby, C. A.; Mossuto, M. F.; Meehan, S.; Gras, S. L.; Christodoulou, J.; Anthony-Cahill, S. J.; Barker, P. D.; Vendruscolo, M.; Dobson, C. M. Metastability of native proteins and the phenomenon of amyloid formation. *Journal of the American Chemical Society* **2011**, *133*, 14160–14163.

(26) Gao, X.; Carroni, M.; Nussbaum-Krammer, C.; Mogk, A.; Nillegoda, N. B.; Szlachcic, A.; Guilbride, D. L.; Saibil, H. R.; Mayer, M. P.; Bukau, B. Human Hsp70 Disaggregase Reverses Parkinson's-Linked α-Synuclein Amyloid Fibrils. *Molecular Cell* **2015**, *59*, 781–793.

(27) Scior, A.; Buntru, A.; Arnsburg, K.; Ast, A.; Iburg, M.; Juenemann, K.; Pigazzini, M. L.; Mlody, B.; Puchkov, D.; Priller, J.; Wanker, E. E.; Prigione, A.; Kirstein, J. Complete suppression of Htt fibrilization and disaggregation of Htt fibrils by a trimeric chaperone complex. *The EMBO Journal* **2018**, *37*, 282–299.

(28) Nachman, E.; Wentink, A. S.; Madiona, K.; Bousset, L.; Katsinelos, T.; Allinson, K.; Kampinga, H.; McEwan, W. A.; Jahn, T. R.; Melki, R.; Mogk, A.; Bukau, B.; Nussbaum-Krammer, C. Disassembly of Tau fibrils by the human Hsp70 disaggregation machinery generates small seeding-competent species. *Journal of Biological Chemistry* **2020**, *295*, 9676–9690.

(29) Frieden, C. Protein aggregation processes: In search of the mechanism. *Protein Science* **2007**, *16*, 2334–2344.

(30) Morris, A. M.; Watzky, M. A.; Finke, R. G. Protein aggregation kinetics, mechanism, and curve-fitting: A review of the literature. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2009**, *1794*, 375–397.

(31) Jucker, M.; Walker, L. C. Self-propagation of pathogenic protein aggregates in neurodegenerative diseases. *Nature* **2013**, *501*, 45–51.

(32) Walker, L. C.; Diamond, M. I.; Duff, K. E.; Hyman, B. T. Mechanisms of Protein Seeding in Neurodegenerative Diseases. *JAMA Neurology* **2013**, *70*, 304–310.

(33) Mahler, H.-C.; Friess, W.; Grauschopf, U.; Kiese, S. Protein aggregation: Pathways, induction factors and analysis. *Journal of Pharmaceutical Sciences* **2009**, *98*, 2909–2934.

(34) Hawe, A.; Sutter, M.; Jiskoot, W. Extrinsic fluorescent dyes as tools for protein characterization. *Pharmaceutical Research* **2008**, *25*, 1487–1499.

(35) Biancalana, M.; Koide, S. Molecular mechanism of Thioflavin-T binding to amyloid fibrils. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2010**, *1804*, 1405–1412.

(36) Xue, W. F.; Homans, S. W.; Radford, S. E. Systematic analysis of nucleation-dependent polymerization reveals new insights into the mechanism of amyloid self-assembly. *Proceedings of the National Academy of Sciences of the United States of America* **2008**, *105*, 8926–8931.

(37) Tkachenko, O.; Benesch, J. L. P.; Baldwin, A. J. αB-crystallin inhibits amyloidogenesis by disassembling aggregation nuclei. *bioRxiv* **2018**, doi: 10.1101/300541.

(38) Bishop, M. F.; Ferrone, F. A. Kinetics of nucleation-controlled polymerization. A perturbation treatment for use with a secondary pathway. *Biophysical Journal* **1984**, *46*, 631–644.

(39) Meisl, G.; Kirkegaard, J. B.; Arosio, P.; Michaels, T. C. T.; Vendruscolo, M.; Dobson, C. M.; Linse, S.; Knowles, T. P. J. Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nature Protocols* **2016**, *11*, 252–272.

(40) Linse, S. et al. Kinetic fingerprints differentiate the mechanisms of action of anti-Aβ antibodies. *Nature Structural and Molecular Biology* **2020**, *27*, 1125–1133.

(41) Neve, R. L.; Harris, P.; Kosik, K. S.; Kurnit, D. M.; Donlon, T. A. Identification of cDNA clones for the human microtubule-associated protein tau and chromosomal localization of the genes for tau and microtubule-associated protein 2. *Molecular Brain Research* **1986**, *1*, 271–280.

(42) Weingarten, M. D.; Lockwood, A. H.; Hwo, S. Y.; Kirschner, M. W. A protein factor essential for microtubule assembly. *Proceedings of the National Academy of Sciences of the United States of America* **1975**, *72*, 1858–1862.

(43) Goedert, M.; Eisenberg, D. S.; Crowther, R. A. Propagation of Tau Aggregates and Neurodegeneration. *Annual Review of Neuroscience* **2017**, *40*, 189–210.

(44) Goedert, M.; Spillantini, M. G.; Jakes, R.; Rutherford, D.; Crowther, R. A. Multiple isoforms of human microtubule-associated protein tau: sequences and localization in neurofibrillary tangles of Alzheimer's disease. *Neuron* **1989**, *3*, 519–526.

(45) Oblinger, M. M.; Argasinski, A.; Wong, J.; Kosik, K. S. Tau gene expression in rat sensory neurons during development and regeneration. *Journal of Neuroscience* **1991**, *11*, 2453–2459.

(46) Goedert, M.; Spillantini, M. G.; Crowther, R. A. Cloning of a big tau microtubule-associated protein characteristic of the peripheral nervous system. *Proceedings of the National Academy of Sciences* **1992**, *89*, 1983–1987.

(47) Fischer, I.; Baas, P. W. Resurrecting the Mysteries of Big Tau. *Trends in Neurosciences* **2020**, *43*, 493–504.

(48) Cleveland, D. W.; Hwo, S. Y.; Kirschner, M. W. Physical and chemical properties of purified tau factor and the role of tau in microtubule assembly. *Journal of Molecular Biology* **1977**, *116*, 227–247.

(49) Schweers, O.; Schönbrunn-Hanebeck, E.; Marx, A.; Mandelkow, E. Structural studies of tau protein and Alzheimer paired helical filaments show no evidence for β-structure. *Journal of Biological Chemistry* **1994**, *269*, 24290–24297.

(50) Mylonas, E.; Hascher, A.; Bernadó, P.; Blackledge, M.; Mandelkow, E.; Svergun, D. I. Domain conformation of tau protein studied by solution small-angle X-ray scattering. *Biochemistry* **2008**, *47*, 10345–10353.

(51) Jeganathan, S.; von Bergen, M.; Brutlach, H.; Steinhoff, H. J.; Mandelkow, E. Global hairpin folding of tau in solution. *Biochemistry* **2006**, *45*, 2283–2293.

(52) Elbaum-Garfinkle, S.; Rhoades, E. Identification of an aggregation-prone structure of tau. *Journal of the American Chemical Society* **2012**, *134*, 16607–16613.

(53) Morris, M.; Knudsen, G. M.; Maeda, S.; Trinidad, J. C.; Ioanoviciu, A.; Burlingame, A. L.; Mucke, L. Tau post-translational modifications in wild-type and human amyloid precursor protein transgenic mice. *Nature Neuroscience* **2015**, *18*, 1183–1189.

(54) Alquezar, C.; Arya, S.; Kao, A. W. Tau Post-translational Modifications: Dynamic Transformers of Tau Function, Degradation, and Aggregation. *Frontiers in Neurology* **2021**, *11*, 595532.

(55) Noble, W.; Hanger, D. P.; Miller, C. C.; Lovestone, S. The Importance of Tau Phosphorylation for Neurodegenerative Diseases. *Frontiers in Neurology* **2013**, *4*, 83.

(56) Wegmann, S.; Biernat, J.; Mandelkow, E. A current view on Tau protein phosphorylation in Alzheimer's disease. *Current Opinion in Neurobiology* **2021**, *69*, 131–138.

(57) Drubin, D. G.; Kirschner, M. W. Tau protein function in living cells. *Journal of Cell Biology* **1986**, *103*, 2739–2746.

(58) Barbier, P.; Zejneli, O.; Martinho, M.; Lasorsa, A.; Belle, V.; Smet-Nocca, C.; Tsvetkov, P. O.; Devred, F.; Landrieu, I. Role of tau as a microtubule-associated protein: Structural and functional aspects. *Frontiers in Aging Neuroscience* **2019**, *11*, 204.

(59) Brandt, R.; Léger, J.; Lee, G. Interaction of tau with the neural plasma membrane mediated by tau's amino-terminal projection domain. *Journal of Cell Biology* **1995**, *131*, 1327–1340.

(60) Morris, M.; Maeda, S.; Vossel, K.; Mucke, L. The Many Faces of Tau. *Neuron* **2011**, *70*, 410–426.

(61) Lee, V. M.; Goedert, M.; Trojanowski, J. Q. Neurodegenerative Tauopathies. *Annual Review of Neuroscience* **2003**, *24*, 1121–1159.

(62) Scheltens, P.; Strooper, B. D.; Kivipelto, M.; Holstege, H.; Chételat, G.; Teunissen, C. E.; Cummings, J.; van der Flier, W. M. Alzheimer's disease. *The Lancet* **2021**, *397*, 1577–1590.

(63) Braak, H.; Braak, E. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* **1991**, *82*, 239–259.

(64) Goedert, M.; Clavaguera, F.; Tolnay, M. The propagation of prion-like protein inclusions in neurodegenerative diseases. *Trends in Neurosciences* **2010**, *33*, 317–325.

(65) Fitzpatrick, A. W.; Falcon, B.; He, S.; Murzin, A. G.; Murshudov, G.; Garringer, H. J.; Crowther, R. A.; Ghetti, B.; Goedert, M.; Scheres, S. H. Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* **2017**, *547*, 185–190.

(66) Sawaya, M. R.; Sambashivan, S.; Nelson, R.; Ivanova, M. I.; Sievers, S. A.; Apostol, M. I.; Thompson, M. J.; Balbirnie, M.; Wiltzius, J. J.; McFarlane, H. T.; Madsen, A.; Riekel, C.; Eisenberg, D. Atomic structures of amyloid cross-β spines reveal varied steric zippers. *Nature* **2007**, *447*, 453–457.

(67) Wille, H.; Drewes, G.; Biernat, J.; Mandelkow, E.-M.; Mandelkow, E. Alzheimer-like Paired Helical Filaments and Antiparallel Dimers Formed from Microtubule-associated Protein Tau In Vitro. *Journal of Cell Biology* **1992**, *118*, 573–584.

(68) Goedert, M.; Jakes, R.; Spillantini, M. G.; Hasegawa, M.; Smith, M. J.; Crowther, R. A. Assembly of microtubule-associated protein tau into Alzheimer-like filaments induced by sulphated glycosaminoglycans. *Nature* **1996**, *383*, 550–553.

(69) Friedhoff, P.; Schneider, A.; Mandelkow, E. M.; Mandelkow, E. Rapid assembly of Alzheimer-like paired helical filaments from microtubule-associated protein tau monitored by fluorescence in solution. *Biochemistry* **1998**, *37*, 10223–10230.

(70) King, M. E.; Ahuja, V.; Binder, L. I.; Kuret, J. Ligand-dependent tau filament formation: Implications for Alzheimer's disease progression. *Biochemistry* **1999**, *38*, 14851–14859.

(71) Kampers, T.; Friedhoff, P.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. RNA stimulates aggregation of microtubule-associated protein tau into Alzheimer-like paired helical filaments. *FEBS Letters* **1996**, *399*, 344–349.

(72) Von Bergen, M.; Friedhoff, P.; Biernat, J.; Heberle, J.; Mandelkow, E. M.; Mandelkow, E. Assembly of tau protein into Alzheimer paired helical filaments depends on a local sequence motif ((306)VQIVYK(311)) forming β structure. *Proceedings of the National Academy of Sciences of the United States of America* **2000**, *97*, 5129–34.

(73) Von Bergen, M.; Barghorn, S.; Li, L.; Marx, A.; Biernat, J.; Mandelkow, E. M.; Mandelkow, E. Mutations of tau protein in frontotemporal dementia promote aggregation of paired helical filaments by enhancing local β-structure. *Journal of biological chemistry* **2001**, *276*, 48165–74.

(74) Falcon, B.; Zhang, W.; Murzin, A. G.; Murshudov, G.; Garringer, H. J.; Vidal, R.; Crowther, R. A.; Ghetti, B.; Scheres, S. H.; Goedert, M. Structures of filaments from Pick's disease reveal a novel tau protein fold. *Nature* **2018**, *561*, 137–140.

(75) Fichou, Y.; Lin, Y.; Rauch, J. N.; Vigers, M.; Zeng, Z.; Srivastava, M.; Keller, T. J.; Freed, J. H.; Kosik, K. S.; Han, S. Cofactors are essential constituents of stable and seeding-active tau fibrils. *Proceedings of the National Academy of Sciences of the United States of America* **2018**, *115*, 13234–13239.

(76) Zhang, W.; Falcon, B.; Murzin, A. G.; Fan, J.; Crowther, R. A.; Goedert, M.; Scheres, S. H. Heparin-induced tau filaments are polymorphic and differ from those in alzheimer's and pick's diseases. *eLife* **2019**, *8*, e43584.

(77) Lövestam, S.; Koh, F. A.; van Knippenberg, B.; Kotecha, A.; Murzin, A. G.; Goedert, M.; Scheres, S. H. Assembly of recombinant tau into filaments identical to those of Alzheimer's disease and chronic traumatic encephalopathy. *eLife* **2022**, *11*, e76494.

(78) Rodriguez-Camargo, D. C.; Sileikis, E.; Chia, S.; Axell, E.; Bernfur, K.; Cataldi, R. L.; Cohen, S. I.; Meisl, G.; Habchi, J.; Knowles, T. P.; Vendruscolo, M.; Linse, S. Proliferation of Tau 304-380 Fragment Aggregates through Autocatalytic Secondary Nucleation. *ACS Chemical Neuroscience* **2021**, *12*, 4406–4415.

(79) Al-Hilaly, Y. K.; Pollack, S. J.; Vadukul, D. M.; Citossi, F.; Rickard, J. E.; Simpson, M.; Storey, J. M.; Harrington, C. R.; Wischik, C. M.; Serpell, L. C. Alzheimer's Disease-like Paired Helical Filament Assembly from Truncated Tau Protein Is Independent of Disulfide Crosslinking. *Journal of Molecular Biology* **2017**, *429*, 3650–3665.

(80) Friedhoff, P.; von Bergen, M.; Mandelkow, E. M.; Davies, P.; Mandelkow, E. A nucleated assembly mechanism of Alzheimer paired helical filaments. *Proceedings of the National Academy of Sciences of the United States of America* **1998**, *95*, 15712–15717.

(81) Barghorn, S.; Mandelkow, E. Toward a Unified Scheme for the Aggregation of Tau into Alzheimer Paired Helical Filaments. *Biochemistry* **2002**, *41*, 14885–14896.

(82) Foote, A. K.; Manger, L. H.; Holden, M. R.; Margittai, M.; Goldsmith, R. H. Time-resolved multirotational dynamics of single solution-phase tau proteins reveals details of conformational variation. *Physical Chemistry Chemical Physics* **2019**, *21*, 1863–1871.

(83) Chirita, C. N.; Congdon, E. E.; Yin, H.; Kuret, J. Triggers of Full-Length Tau Aggregation: A Role for Partially Folded Intermediates. *Biochemistry* **2005**, *44*, 5862–5872.

(84) Von Bergen, M.; Barghorn, S.; Biernat, J.; Mandelkow, E.-M.; Mandelkow, E. Tau aggregation is driven by a transition from random coil to β sheet structure. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease* **2005**, *1739*, 158–166.

(85) Mukrasch, M. D.; Bibow, S.; Korukottu, J.; Jeganathan, S.; Biernat, J.; Griesinger, C.; Mandelkow, E.; Zweckstetter, M. Structural Polymorphism of 441-Residue Tau at Single Residue Resolution. *PLoS Biology* **2009**, *7*, e1000034.

(86) Eschmann, N. A.; Georgieva, E. R.; Ganguly, P.; Borbat, P. P.; Rappaport, M. D.; Akdogan, Y.; Freed, J. H.; Shea, J. E.; Han, S. Signature of an aggregation-prone conformation of tau. *Scientific Reports* **2017**, *7*, 44739.

(87) Carmel, G.; Mager, E. M.; Binder, L. I.; Kuret, J. The structural basis of monoclonal antibody Alz50's selectivity for Alzheimer's disease pathology. *Journal of Biological Chemistry* **1996**, *271*, 32789–32795.

(88) Manger, L. H.; Foote, A. K.; Wood, S. L.; Holden, M. R.; Heylman, K. D.; Margittai, M.; Goldsmith, R. H. Revealing Conformational Variants of Solution-Phase Intrinsically Disordered Tau Protein at the Single-Molecule Level. *Angewandte Chemie International Edition* **2017**, *56*, 15584–15588.

(89) Shammas, S. L.; Garcia, G. A.; Kumar, S.; Kjaergaard, M.; Horrocks, M. H.; Shivji, N.; Mandelkow, E.; Knowles, T. P.; Mandelkow, E.; Klenerman, D. A mechanistic model of tau amyloid aggregation based on direct observation of oligomers. *Nature Communications* **2015**, *6*, 7025.

(90) Cremades, N.; Cohen, S. I.; Deas, E.; Abramov, A. Y.; Chen, A. Y.; Orte, A.; Sandal, M.; Clarke, R. W.; Dunne, P.; Aprile, F. A.; Bertoncini, C. W.; Wood, N. W.; Knowles, T. P.; Dobson, C. M.; Klenerman, D. Direct Observation of the Interconversion of Normal and Toxic Forms of α-Synuclein. *Cell* **2012**, *149*, 1048–1059.

(91) Hardy, J. A.; Higgins, G. A. Alzheimer's disease: the amyloid cascade hypothesis. *Science* **1992**, *256*, 184–185.

(92) Gómez-Isla, T.; Hollister, R.; West, H.; Mui, S.; Growdon, J. H.; Petersen, R. C.; Parisi, J. E.; Hyman, B. T. Neuronal loss correlates with but exceeds neurofibrillary tangles in Alzheimer's disease. *Annals of Neurology* **1997**, *41*, 17–24.

(93) Brier, M. R. et al. Tau and Aβ imaging, CSF measures, and cognition in Alzheimer's disease. *Science translational medicine* **2016**, *8*, 338ra66.

(94) Li, C.; Götz, J. Tau-based therapies in neurodegeneration: opportunities and challenges. *Nature Reviews Drug Discovery* **2017**, *16*, 863–883.

(95) Congdon, E. E.; Sigurdsson, E. M. Tau-targeting therapies for Alzheimer disease. *Nature Reviews Neurology* **2018**, *14*, 399–415.

(96) Wang, L.; Bharti; Kumar, R.; Pavlov, P. F.; Winblad, B. Small molecule therapeutics for tauopathy in Alzheimer's disease: Walking on the path of most resistance. *European Journal of Medicinal Chemistry* **2021**, *209*, 112915.

(97) Novak, P.; Kontsekova, E.; Zilka, N.; Novak, M. Ten years of tau-targeted immunotherapy: The path walked and the roads ahead. *Frontiers in Neuroscience* **2018**, *12*, 798.

(98) Ji, C.; Sigurdsson, E. M. Current Status of Clinical Trials on Tau Immunotherapies. *Drugs* **2021**, *81*, 1135–1152.

(99) DeVos, S. L. et al. Tau reduction prevents neuronal loss and reverses pathological tau deposition and seeding in mice with tauopathy. *Science Translational Medicine* **2017**, *9*, eaag0481.

(100) Alzforum Antisense Therapy Stifles CSF Tau in Mild Alzheimer's Disease, 2022, `https://www.alzforum.org/news/conference-coverage/antisense-therapy-stifles-csf-tau-mild-alzheimers-disease-0`.

(101) Aillaud, I.; Funke, S. A. Tau Aggregation Inhibiting Peptides as Potential Therapeutics for Alzheimer Disease. *Cellular and Molecular Neurobiology* **2022**, *43*, 951–961.

(102) Limorenko, G.; Lashuel, H. A. Revisiting the grammar of Tau aggregation and pathology formation: how new insights from brain pathology are shaping how we study and target Tauopathies. *Chemical Society Reviews* **2022**, *51*, 513–565.

(103) Kontsekova, E.; Zilka, N.; Kovacech, B.; Skrabana, R.; Novak, M. Identification of structural determinants on tau protein essential for its pathological function: Novel therapeutic target for tau immunotherapy in Alzheimer's disease. *Alzheimer's Research and Therapy* **2014**, *6*, 45.

(104) Kontsekova, E.; Zilka, N.; Kovacech, B.; Novak, P.; Novak, M. First-in-man tau vaccine targeting structural determinants essential for pathological tau-tau interaction reduces tau oligomerisation and neurofibrillary degeneration in an Alzheimer's disease model. *Alzheimer's Research and Therapy* **2014**, *6*, 44.

(105) Novak, P. et al. ADAMANT: a placebo-controlled randomized phase 2 study of AADvac1, an active immunotherapy against pathological tau in Alzheimer's disease. *Nature Aging* **2021**, *1*, 521–534.

(106) Theunis, C.; Crespo-Biel, N.; Gafner, V.; Pihlgren, M.; López-Deber, M. P.; Reis, P.; Hickman, D. T.; Adolfsson, O.; Chuard, N.; Ndao, D. M.; Borghgraef, P.; Devijver, H.; Leuven, F. V.; Pfeifer, A.; Muhs, A. Efficacy and Safety of A Liposome-Based Vaccine against Protein Tau, Assessed in Tau.P301L Mice That Model Tauopathy. *PLoS ONE* **2013**, *8*, e72301.

(107) ClinicalTrials.gov A Study to Evaluate the Safety, Tolerability and Immunogenicity of Tau Targeted Vaccines in Participants With Early Alzheimer's Disease, (date accessed: 07/06/2023), 2022, `https://clinicaltrials.gov/ct2/show/NCT04445831`.

(108) Streffer, J.; Mermoud, J.; Sol, O.; Vukicevic, M.; Fiorini, E.; Gollwitzer, E.; Hliva, V.; Hickman, D.; Gray, J.; Donati, P.; Deber, M. L.; Rongere, J.; Pfeifer, A.; Kosco-Vilbois, M.; Scheltens, P. ACI-35.030 and JACI-35.064, two novel anti-phospho-tau vaccines for the treatment of Alzheimer's disease: interim phase 1B/2A data on safety, tolerability and immunogenicity (Oral Communicated, presented at CTAD Conference; Nov-Dec 2022). *The Journal of Prevention of Alzheimer's Disease* **2022**, *9*, S8–S50.

(109) Dhillon, S. Aducanumab: First Approval. *Drugs* **2021**, *81*, 1437–1443.

(110) Walsh, S.; Merrick, R.; Milne, R.; Brayne, C. Aducanumab for Alzheimer's disease? *BMJ* **2021**, *374*, n1682.

(111) Kiss, R.; Csizmadia, G.; Solti, K.; Keresztes, A.; Zhu, M.; Pickhardt, M.; Mandelkow, E.; Tóth, G. Structural Basis of Small Molecule Targetability of Monomeric Tau Protein. *ACS Chemical Neuroscience* **2018**, *9*, 2997–3006.

(112) Wischik, C. M.; Edwards, P. C.; Lai, R. Y.; Roth, M.; Harrington, C. R. Selective inhibition of Alzheimer disease-like tau aggregation by phenothiazines. *Proceedings of the National Academy of Sciences of the United States of America* **1996**, *93*, 11213–11218.

(113) Schirmer, R. H.; Adler, H.; Pickhardt, M.; Mandelkow, E. Lest we forget you - methylene blue... *Neurobiology of Aging* **2011**, *32*, 2325.e7–2325.e16.

(114) Crowe, A.; James, M. J.; Virginia, M. Y.; Smith, A. B.; Trojanowski, J. Q.; Ballatore, C.; Brunden, K. R. Aminothienopyridazines and Methylene Blue Affect Tau Fibrillization via Cysteine Oxidation. *Journal of Biological Chemistry* **2013**, *288*, 11024–11037.

(115) Congdon, E. E.; Wu, J. W.; Myeku, N.; Figueroa, Y. H.; Herman, M.; Marinec, P. S.; Gestwicki, J. E.; Dickey, C. A.; Yu, W. H.; Duff, K. E. Methylthioninium chloride (methylene blue) induces autophagy and attenuates tauopathy in vitro and in vivo. *Autophagy* **2012**, *8*, 609–622.

(116) Wischik, C. M.; Staff, R. T.; Wischik, D. J.; Bentham, P.; Murray, A. D.; Storey, J. M.; Kook, K. A.; Harrington, C. R. Tau Aggregation Inhibitor Therapy: An Exploratory Phase 2 Study in Mild or Moderate Alzheimer's Disease. *Journal of Alzheimer's Disease* **2015**, *44*, 705–720.

(117) Baddeley, T. C.; McCaffrey, J.; Storey, J. M.; Cheung, J. K.; Melis, V.; Horsley, D.; Harrington, C. R.; Wischik, C. M. Complex Disposition of Methylthioninium Redox Forms Determines Efficacy in Tau Aggregation Inhibitor Therapy for Alzheimer's Disease. *Journal of Pharmacology and Experimental Therapeutics* **2015**, *352*, 110–118.

(118) TauRx.com TauRx Announces Results from Phase 3 Alzheimer's Disease Study, LUCIDITY, Assuring Path for Regulatory Submissions, (date accessed: 06/05/2023), 2022, `https://taurx.com/news-insights/taurx-announces-results-from-phase-3-alzheimers-disease-study-lucidity-assuring-path-for-regulatory-submissions`.

(119) Schweers, O.; Mandelkow, E. M.; Biernat, J.; Mandelkow, E. Oxidation of cysteine-322 in the repeat domain of microtubule-associated protein tau controls the in vitro assembly of paired helical filaments. *Proceedings of the National Academy of Sciences of the United States of America* **1995**, *92*, 8463–8467.

(120) Akoury, E.; Pickhardt, M.; Gajda, M.; Biernat, J.; Mandelkow, E.; Zweckstetter, M. Mechanistic basis of phenothiazine-driven inhibition of Tau aggregation. *Angewandte Chemie International Edition* **2013**, *52*, 3511–3515.

(121) Al-Hilaly, Y. K.; Pollack, S. J.; Rickard, J. E.; Simpson, M.; Raulin, A. C.; Baddeley, T.; Schellenberger, P.; Storey, J. M.; Harrington, C. R.; Wischik, C. M.; Serpell, L. C. Cysteine-Independent Inhibition of Alzheimer's Disease-like Paired Helical Filament Assembly by Leuco-Methylthioninium (LMT). *Journal of Molecular Biology* **2018**, *430*, 4119–4131.

(122) Lee, A. C. L.; Harris, J. L.; Khanna, K. K.; Hong, J. H. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. *International Journal of Molecular Sciences* **2019**, *20*, 2383.

(123) Lien, S.; Lowman, H. B. Therapeutic peptides. *Trends in Biotechnology* **2003**, *21*, 556–562.

(124) Liu, M.; Li, X.; Xie, Z.; Xie, C.; Zhan, C.; Hu, X.; Shen, Q.; Wei, X.; Su, B.; Wang, J.; Lu, W. D-Peptides as Recognition Molecules and Therapeutic Agents. *The Chemical Record* **2016**, *16*, 1772–1786.

(125) Fosgerau, K.; Hoffmann, T. Peptide therapeutics: current status and future directions. *Drug Discovery Today* **2015**, *20*, 122–128.

(126) Goodman, M.; Chorev, M. On the Concept of Linear Modified Retro-Peptide Structures. *Accounts of Chemical Research* **1979**, *12*, 1–7.

(127) Doti, N.; Mardirossian, M.; Sandomenico, A.; Ruvo, M.; Caporale, A. Recent Applications of Retro-Inverso Peptides. *International Journal of Molecular Sciences* **2021**, *22*, 8677.

(128) Smith, G. P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **1985**, *228*, 1315–1317.

(129) Smith, G. P. Phage Display Simple Evolution in a Petri Dish (Nobel Lecture). *Angewandte Chemie International Edition* **2019**, *58*, 14428–14437.

(130) Wu, C. H.; Liu, I. J.; Lu, R. M.; Wu, H. C. Advancement and applications of peptide phage display technology in biomedical science. *Journal of Biomedical Science* **2016**, *23*, 8.

(131) Schumacher, T. N.; Mayr, L. M.; Minor, D. L.; Milhollen, M. A.; Burgess, M. W.; Kim, P. S. Identification of d-Peptide Ligands Through Mirror-Image Phage Display. *Science* **1996**, *271*, 1854–1857.

(132) Dawson, P. E.; Muir, T. W.; Clark-Lewis, I.; Kent, S. B. Synthesis of Proteins by Native Chemical Ligation. *Science* **1994**, *266*, 776–779.

(133) Ciemny, M.; Kurcinski, M.; Kamel, K.; Kolinski, A.; Alam, N.; Schueler-Furman, O.; Kmiecik, S. Protein–peptide docking: opportunities and challenges. *Drug Discovery Today* **2018**, *23*, 1530–1537.

(134) Sievers, S. A.; Karanicolas, J.; Chang, H. W.; Zhao, A.; Jiang, L.; Zirafi, O.; Stevens, J. T.; Münch, J.; Baker, D.; Eisenberg, D. Structure-based design of non-natural amino-acid inhibitors of amyloid fibril formation. *Nature* **2011**, *475*, 96–100.

(135) Seidler, P. M.; Boyer, D. R.; Rodriguez, J. A.; Sawaya, M. R.; Cascio, D.; Murray, K.; Gonen, T.; Eisenberg, D. S. Structure-based inhibitors of tau aggregation. *Nature Chemistry* **2018**, *10*, 170–176.

(136) Griner, S. L.; Seidler, P.; Bowler, J.; Murray, K. A.; Yang, T. P.; Sahay, S.; Sawaya, M. R.; Cascio, D.; Rodriguez, J. A.; Philipp, S.; Sosna, J.; Glabe, C. G.; Gonen, T.; Eisenberg, D. S. Structure based inhibitors of amyloid β core suggest a common interface with Tau. *eLife* **2019**, *8*, e46924.

(137) Dammers, C.; Yolcu, D.; Kukuk, L.; Willbold, D.; Pickhardt, M.; Mandelkow, E.; Horn, A. H.; Sticht, H.; Malhis, M. N.; Will, N.; Schuster, J.; Funke, S. A. Selection and Characterization of Tau Binding D-Enantiomeric Peptides with Potential for Therapy of Alzheimer Disease. *PLoS ONE* **2016**, *11*, e0167432.

(138) Zhang, X.; Zhang, X.; Zhong, M.; Zhao, P.; Guo, C.; Li, Y.; Wang, T.; Gao, H. Selection of a d -Enantiomeric Peptide Specifically Binding to PHF6 for Inhibiting Tau Aggregation in Transgenic Mice. *ACS Chemical Neuroscience* **2020**, *11*, 4240–4253.

(139) Malhis, M.; Kaniyappan, S.; Aillaud, I.; Chandupatla, R. R.; Ramirez, L. M.; Zweckstetter, M.; Horn, A. H.; Mandelkow, E.; Sticht, H.; Funke, S. A. Potent Tau Aggregation Inhibitor D-Peptides Selected against Tau-Repeat 2 Using Mirror Image Phage Display. *ChemBioChem* **2021**, *22*, 3049–3059.

(140) Arosio, P.; Vendruscolo, M.; Dobson, C. M.; Knowles, T. P. Chemical kinetics for drug discovery to combat protein aggregation diseases. *Trends in Pharmacological Sciences* **2014**, *35*, 127–135.

(141) Selkoe, D. J.; Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Molecular Medicine* **2016**, *8*, 595–608.

(142) Brangwynne, C. P.; Eckmann, C. R.; Courson, D. S.; Rybarska, A.; Hoege, C.; Gharakhani, J.; Jülicher, F.; Hyman, A. A. Germline P granules are liquid droplets that localize by controlled dissolution/condensation. *Science* **2009**, *324*, 1729–1732.

(143) Azaldegui, C. A.; Vecchiarelli, A. G.; Biteen, J. S. The emergence of phase separation as an organizing principle in bacteria. *Biophysical Journal* **2021**, *120*, 1123–1138.

(144) Dignon, G. L.; Best, R. B.; Mittal, J. Biomolecular Phase Separation: From Molecular Driving Forces to Macroscopic Properties. *Annual Review of Physical Chemistry* **2020**, *71*, 53–75.

(145) Wei, M. T.; Elbaum-Garfinkle, S.; Holehouse, A. S.; Chen, C. C. H.; Feric, M.; Arnold, C. B.; Priestley, R. D.; Pappu, R. V.; Brangwynne, C. P. Phase behaviour of disordered proteins underlying low density and high permeability of liquid organelles. *Nature Chemistry* **2017**, *9*, 1118–1125.

(146) Schuster, B. S.; Reed, E. H.; Parthasarathy, R.; Jahnke, C. N.; Caldwell, R. M.; Bermudez, J. G.; Ramage, H.; Good, M. C.; Hammer, D. A. Controllable protein phase separation and modular recruitment to form responsive membraneless organelles. *Nature Communications* **2018**, *9*, 1–12.

(147) Patel, A. et al. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation. *Cell* **2015**, *162*, 1066–1077.

(148) Nott, T. J.; Petsalaki, E.; Farber, P.; Jervis, D.; Fussner, E.; Plochowietz, A.; Craggs, T. D.; Bazett-Jones, D. P.; Pawson, T.; Forman-Kay, J. D.; Baldwin, A. J. Phase Transition of a Disordered Nuage Protein Generates Environmentally Responsive Membraneless Organelles. *Molecular Cell* **2015**, *57*, 936–947.

(149) Strom, A. R.; Emelyanov, A. V.; Mir, M.; Fyodorov, D. V.; Darzacq, X.; Karpen, G. H. Phase separation drives heterochromatin domain formation. *Nature* **2017**, *547*, 241–245.

(150) Brangwynne, C. P. Phase transitions and size scaling of membrane-less organelles. *Journal of Cell Biology* **2013**, *203*, 875–881.

(151) Peng, A.; Weber, S. C. Evidence for and against Liquid-Liquid Phase Separation in the Nucleus. *Non-coding RNA* **2019**, *5*, 50.

(152) Forman-Kay, J. D.; Ditlev, J. A.; Nosella, M. L.; Lee, H. O. What are the distinguishing features and size requirements of biomolecular condensates and their implications for RNA-containing condensates? *RNA* **2022**, *28*, 36–47.

(153) Darling, A. L.; Liu, Y.; Oldfield, C. J.; Uversky, V. N. Intrinsically Disordered Proteome of Human Membrane-Less Organelles. *Proteomics* **2018**, *18*, 1700193.

(154) Uversky, V. N. Intrinsically disordered proteins in overcrowded milieu: Membrane-less organelles, phase separation, and intrinsic disorder. *Current Opinion in Structural Biology* **2017**, *44*, 18–30.

(155) Lin, Y. H.; Song, J.; Forman-Kay, J. D.; Chan, H. S. Random-phase-approximation theory for sequence-dependent, biologically functional liquid-liquid phase separation of intrinsically disordered proteins. *Journal of Molecular Liquids* **2017**, *228*, 176–193.

(156) Brady, J. P.; Farber, P. J.; Sekhar, A.; Lin, Y. H.; Huang, R.; Bah, A.; Nott, T. J.; Chan, H. S.; Baldwin, A. J.; Forman-Kay, J. D.; Kay, L. E. Structural and hydrodynamic properties of an intrinsically disordered region of a germ cell-specific protein on phase separation. *Proceedings of the National Academy of Sciences of the United States of America* **2017**, *114*, E8194–E8203.

(157) Burke, K. A.; Janke, A. M.; Rhine, C. L.; Fawzi, N. L. Residue-by-Residue View of In Vitro FUS Granules that Bind the C-Terminal Domain of RNA Polymerase II. *Molecular Cell* **2015**, *60*, 231–241.

(158) Ryan, V. H.; Dignon, G. L.; Zerze, G. H.; Chabata, C. V.; Silva, R.; Conicella, A. E.; Amaya, J.; Burke, K. A.; Mittal, J.; Fawzi, N. L. Mechanistic view of hnRNPA2 low complexity domain structure, interactions, and phase separation altered by disease mutation and arginine methylation. *Molecular Cell* **2018**, *69*, 465–479.

(159) Murthy, A. C.; Dignon, G. L.; Kan, Y.; Zerze, G. H.; Parekh, S. H.; Mittal, J.; Fawzi, N. L. Molecular interactions underlying liquid-liquid phase separation of the FUS low-complexity domain. *Nature Structural and Molecular Biology* **2019**, *26*, 637–648.

(160) Brangwynne, C. P.; Mitchison, T. J.; Hyman, A. A. Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes. *Proceedings of the National Academy of Sciences of the United States of America* **2011**, *108*, 4334–4339.

(161) Platani, M.; Goldberg, I.; Swedlow, J. R.; Lamond, A. I. In vivo analysis of Cajal body movement, separation, and joining in live human cells. *Journal of Cell Biology* **2000**, *151*, 1561–1574.

(162) Shin, Y.; Brangwynne, C. P. Liquid phase condensation in cell physiology and disease. *Science* **2017**, *357*, eaaf4382.

(163) Luo, Y.; Na, Z.; Slavoff, S. A. P-Bodies: Composition, Properties, and Functions. *Biochemistry* **2018**, *57*, 2424–2431.

(164) Banani, S. F.; Rice, A. M.; Peeples, W. B.; Lin, Y.; Jain, S.; Parker, R.; Rosen, M. K. Compositional Control of Phase-Separated Cellular Bodies. *Cell* **2016**, *166*, 651–663.

(165) Feric, M.; Vaidya, N.; Harmon, T. S.; Mitrea, D. M.; Zhu, L.; Richardson, T. M.; Kriwacki, R. W.; Pappu, R. V.; Brangwynne, C. P. Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell* **2016**, *165*, 1686–1697.

(166) Gibson, B. A.; Doolittle, L. K.; Schneider, M. W.; Jensen, L. E.; Gamarra, N.; Henry, L.; Gerlich, D. W.; Redding, S.; Rosen, M. K. Organization of Chromatin by Intrinsic and Regulated Phase Separation. *Cell* **2019**, *179*, 470–484.

(167) Pessina, F.; Gioia, U.; Brandi, O.; Farina, S.; Ceccon, M.; Francia, S.; d'Adda di Fagagna, F. DNA Damage Triggers a New Phase in Neurodegeneration. *Trends in Genetics* **2021**, *37*, 337–354.

(168) Cho, W. K.; Spille, J. H.; Hecht, M.; Lee, C.; Li, C.; Grube, V.; Cisse, I. I. Mediator and RNA polymerase II clusters associate in transcription-dependent condensates. *Science* **2018**, *361*, 412–415.

(169) Sabari, B. R. et al. Coactivator condensation at super-enhancers links phase separation and gene control. *Science* **2018**, *361*, eaar3958.

(170) Guo, Y. E. et al. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature* **2019**, *572*, 543–548.

(171) Yasuda, S.; Tsuchiya, H.; Kaiho, A.; Guo, Q.; Ikeuchi, K.; Endo, A.; Arai, N.; Ohtake, F.; Murata, S.; Inada, T.; Baumeister, W.; Fernández-Busnadiego, R.; Tanaka, K.; Saeki, Y. Stress- and ubiquitylation-dependent phase separation of the proteasome. *Nature* **2020**, *578*, 296–300.

(172) Sun, D.; Wu, R.; Zheng, J.; Li, P.; Yu, L. Polyubiquitin chain-induced p62 phase separation drives autophagic cargo segregation. *Cell Research* **2018**, *28*, 405–415.

(173) Chong, P. A.; Forman-Kay, J. D. Liquid–liquid phase separation in cellular signaling systems. *Current Opinion in Structural Biology* **2016**, *41*, 180–186.

(174) Abbondanzieri, E. A.; Meyer, A. S. More than just a phase: the search for membraneless organelles in the bacterial cytoplasm. *Current Genetics* **2019**, *65*, 691–694.

(175) Muthunayake, N. S.; Tomares, D. T.; Childers, W. S.; Schrader, J. M. Phase-separated bacterial ribonucleoprotein bodies organize mRNA decay. *Wiley Interdisciplinary Reviews: RNA* **2020**, *11*, e1599.

(176) Ladouceur, A. M.; Parmar, B. S.; Biedzinski, S.; Wall, J.; Tope, S. G.; Cohn, D.; Kim, A.; Soubry, N.; Reyes-Lamothe, R.; Weber, S. C. Clusters of bacterial RNA polymerase are biomolecular condensates that assemble through liquid-liquid phase separation. *Proceedings of the National Academy of Sciences of the United States of America* **2020**, *117*, 18540–18549.

(177) Janissen, R.; Arens, M. M.; Vtyurina, N. N.; Rivai, Z.; Sunday, N. D.; Eslami-Mossallam, B.; Gritsenko, A. A.; Laan, L.; de Ridder, D.; Artsimovitch, I.; Dekker, N. H.; Abbondanzieri, E. A.; Meyer, A. S. Global DNA Compaction in Stationary-Phase Bacteria Does Not Affect Transcription. *Cell* **2018**, *174*, 1188–1199.

(178) Saurabh, S.; Chong, T. N.; Bayas, C.; Dahlberg, P. D.; Cartwright, H. N.; Moerner, W. E.; Shapiro, L. ATP-responsive biomolecular condensates tune bacterial kinase signaling. *Science Advances* **2022**, *8*, eabm6570.

(179) Guilhas, B.; Walter, J. C.; Rech, J.; David, G.; Walliser, N. O.; Palmeri, J.; Mathieu-Demaziere, C.; Parmeggiani, A.; Bouet, J. Y.; Gall, A. L.; Nollmann, M. ATP-Driven Separation of Liquid Phase Condensates in Bacteria. *Molecular Cell* **2020**, *79*, 293–303.

(180) O'Flynn, B. G.; Mittag, T. The role of liquid–liquid phase separation in regulating enzyme activity. *Current Opinion in Cell Biology* **2021**, *69*, 70–79.

(181) Peeples, W.; Rosen, M. K. Mechanistic dissection of increased enzymatic rate in a phase-separated compartment. *Nature Chemical Biology 2021 17:6* **2021**, *17*, 693–702.

(182) Klosin, A.; Oltsch, F.; Harmon, T.; Honigmann, A.; Jülicher, F.; Hyman, A. A.; Zechner, C. Phase separation provides a mechanism to reduce noise in cells. *Science* **2020**, *367*, 464–468.

(183) Riback, J. A.; Brangwynne, C. P. Can phase separation buffer cellular noise? *Science* **2020**, *367*, 364–365.

(184) Yoo, H.; Triandafillou, C.; Drummond, D. A. Cellular sensing by phase separation: Using the process, not just the products. *Journal of Biological Chemistry* **2019**, *294*, 7151–7159.

(185) Alberti, S.; Gladfelter, A.; Mittag, T. Considerations and Challenges in Studying Liquid-Liquid Phase Separation and Biomolecular Condensates. *Cell* **2019**, *176*, 419–434.

(186) Brangwynne, C. P.; Tompa, P.; Pappu, R. V. Polymer physics of intracellular phase transitions. *Nature Physics* **2015**, *11*, 899–904.

(187) Zhang, X.; Lin, Y.; Eschmann, N. A.; Zhou, H.; Rauch, J. N.; Hernandez, I.; Guzman, E.; Kosik, K. S.; Han, S. RNA stores tau reversibly in complex coacervates. *PLoS Biology* **2017**, *15*, e2002183.

(188) Das, S.; Amin, A. N.; Lin, Y. H.; Chan, H. S. Coarse-grained residue-based models of disordered protein condensates: utility and limitations of simple charge pattern parameters. *Physical Chemistry Chemical Physics* **2018**, *20*, 28558–28574.

(189) Gallivan, J. P.; Dougherty, D. A. Cation-π interactions in structural biology. *Proceedings of the National Academy of Sciences of the United States of America* **1999**, *96*, 9459–9464.

(190) Xiu, X.; Puskar, N. L.; Shanata, J. A.; Lester, H. A.; Dougherty, D. A. Nicotine binding to brain receptors requires a strong cation-π interaction. *Nature* **2009**, *458*, 534–537.

(191) Wintjens, R.; Lievin, J.; Rooman, M.; Buisine, E. Contribution of Cation-π Interactions to the Stability of Protein-DNA Complexes. *Journal of Molecular Biology* **2000**, *302*, 395–410.

(192) Zhang, H.; Li, C.; Yang, F.; Su, J.; Tan, J.; Zhang, X.; Wang, C. Cation-π interactions at non-redundant protein-RNA interfaces. *Biochemistry (Moscow)* **2014**, *79*, 643–652.

(193) Dougherty, D. A. The cation-π interaction. *Accounts of Chemical Research* **2013**, *46*, 885–893.

(194) Qamar, S. et al. FUS Phase Separation Is Modulated by a Molecular Chaperone and Methylation of Arginine Cation-π Interactions. *Cell* **2018**, *173*, 720–734.

(195) Yakovchuk, P.; Protozanova, E.; Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **2006**, *34*, 564–574.

(196) Wilson, K. A.; Kellie, J. L.; Wetmore, S. D. DNA–protein π-interactions in nature: abundance, structure, composition and strength of contacts between aromatic amino acids and DNA nucleobases or deoxyribose sugar. *Nucleic Acids Research* **2014**, *42*, 6726–6741.

(197) Wilson, K. A.; Kung, R. W.; D'Souza, S.; Wetmore, S. D. Anatomy of noncovalent interactions between the nucleobases or ribose and π-containing amino acids in RNA-protein complexes. *Nucleic Acids Research* **2021**, *49*, 2213–2225.

(198) Wang, J.; Choi, J. M.; Holehouse, A. S.; Lee, H. O.; Zhang, X.; Jahnel, M.; Maharana, S.; Lemaitre, R.; Pozniakovsky, A.; Drechsel, D.; Poser, I.; Pappu, R. V.; Alberti, S.; Hyman, A. A. A Molecular Grammar Governing the Driving Forces for Phase Separation of Prion-like RNA Binding Proteins. *Cell* **2018**, *174*, 688–699.

(199) Fromm, S. A.; Kamenz, J.; Ncldeke, E. R.; Neu, A.; Zocher, G.; Sprangers, R. In Vitro Reconstitution of a Cellular Phase-Transition Process that Involves the mRNA Decapping Machinery. *Angewandte Chemie International Edition* **2014**, *53*, 7354–7359.

(200) Conicella, A. E.; Zerze, G. H.; Mittal, J.; Fawzi, N. L. ALS Mutations Disrupt Phase Separation Mediated by α-Helical Structure in the TDP-43 Low-Complexity C-Terminal Domain. *Structure* **2016**, *24*, 1537–1549.

(201) Krainer, G. et al. Reentrant liquid condensate phase of proteins is stabilized by hydrophobic and non-ionic interactions. *Nature Communications* **2021**, *12*, 1085.

(202) Elbaum-Garfinkle, S.; Kim, Y.; Szczepaniak, K.; Chen, C. C. H.; Eckmann, C. R.; Myong, S.; Brangwynne, C. P. The disordered P granule protein LAF-1 drives phase separation into droplets with tunable viscosity and dynamics. *Proceedings of the National Academy of Sciences of the United States of America* **2015**, *112*, 7189–7194.

(203) Braun, M. K.; Wolf, M.; Matsarskaia, O.; Vela, S. D.; Roosen-Runge, F.; Sztucki, M.; Roth, R.; Zhang, F.; Schreiber, F. Strong Isotope Effects on Effective Interactions and Phase Behavior in Protein Solutions in the Presence of Multivalent Ions. *Journal of Physical Chemistry B* **2017**, *121*, 1731–1739.

(204) Onuchic, P. L.; Milin, A. N.; Alshareedah, I.; Deniz, A. A.; Banerjee, P. R. Divalent cations can control a switch-like behavior in heterotypic and homotypic RNA coacervates. *Scientific Reports* **2019**, *9*, 12161.

(205) Crabtree, M. D.; Holland, J.; Kompella, P.; Babl, L.; Turner, N.; Baldwin, A. J.; Nott, T. J. Repulsive electrostatic interactions modulate dense and dilute phase properties of biomolecular condensates. *bioRxiv* **2021**, doi: 10.1101/2020.10.29.357863.

(206) Ruff, K. M.; Roberts, S.; Chilkoti, A.; Pappu, R. V. Advances in Understanding Stimulus-Responsive Phase Behavior of Intrinsically Disordered Protein Polymers. *Journal of Molecular Biology* **2018**, *430*, 4619–4635.

(207) Garcia-Quiroz, F.; Chilkoti, A. Sequence heuristics to encode phase behaviour in intrinsically disordered protein polymers. *Nature Materials* **2015**, *14*, 1164–1171.

(208) Adame-Arana, O.; Weber, C. A.; Zaburdaev, V.; Prost, J.; Jülicher, F. Liquid Phase Separation Controlled by pH. *Biophysical Journal* **2020**, *119*, 1590–1605.

(209) Kroschwald, S.; Munder, M. C.; Ruer, M.; Hyman, A. A.; Alberti, S. Different Material States of Pub1 Condensates Define Distinct Modes of Stress Adaptation and Recovery. *Cell Reports* **2018**, *23*, 3327–3339.

(210) Parchure, A.; Tian, M.; Stalder, D.; Boyer, C. K.; Bearrows, S. C.; Rohli, K. E.; Zhang, J.; Rivera-Molina, F.; Ramazanov, B. R.; Mahata, S. K.; Wang, Y.; Stephens, S. B.; Gershlick, D. C.; von Blume, J. Liquid–liquid phase separation facilitates the biogenesis of secretory storage granules. *Journal of Cell Biology* **2022**, *221*, e202206132.

(211) Garcia-Quiroz, F.; Fiore, V. F.; Levorse, J.; Polak, L.; Wong, E.; Pasolli, H. A.; Fuchs, E. Liquid-liquid phase separation drives skin barrier formation. *Science* **2020**, *367*, eaax9554.

(212) Hofweber, M.; Hutten, S.; Bourgeois, B.; Spreitzer, E.; Niedner-Boblenz, A.; Schifferer, M.; Ruepp, M. D.; Simons, M.; Niessing, D.; Madl, T.; Dormann, D. Phase Separation of FUS Is Suppressed by Its Nuclear Import Receptor and Arginine Methylation. *Cell* **2018**, *173*, 706–719.

(213) Tsai, W. C.; Gayatri, S.; Reineke, L. C.; Sbardella, G.; Bedford, M. T.; Lloyd, R. E. Arginine Demethylation of G3BP1 Promotes Stress Granule Assembly. *Journal of Biological Chemistry* **2016**, *291*, 22671–22685.

(214) Zhang, H.; Elbaum-Garfinkle, S.; Langdon, E. M.; Taylor, N.; Occhipinti, P.; Bridges, A. A.; Brangwynne, C. P.; Gladfelter, A. S. RNA Controls PolyQ Protein Phase Transitions. *Molecular Cell* **2015**, *60*, 220–230.

(215) Molliex, A.; Temirov, J.; Lee, J.; Coughlin, M.; Kanagaraj, A. P.; Kim, H. J.; Mittag, T.; Taylor, J. P. Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization. *Cell* **2015**, *163*, 123–133.

(216) Banerjee, P. R.; Milin, A. N.; Moosa, M. M.; Onuchic, P. L.; Deniz, A. A. Reentrant Phase Transition Drives Dynamic Substructure Formation in Ribonucleoprotein Droplets. *Angewandte Chemie International Edition* **2017**, *56*, 11354–11359.

(217) Mehta, S.; Zhang, J. Liquid–liquid phase separation drives cellular function and dysfunction in cancer. *Nature Reviews Cancer* **2022**, *22*, 239–252.

(218) Alberti, S.; Dormann, D. Liquid–Liquid Phase Separation in Disease. *Annual Review of Genetics* **2019**, *53*, 171–194.

(219) Wegmann, S. et al. Tau protein liquid–liquid phase separation can initiate tau aggregation. *The EMBO Journal* **2018**, *37*, e98049.

(220) Ray, S. et al. α-Synuclein aggregation nucleates through liquid–liquid phase separation. *Nature Chemistry* **2020**, *12*, 705–716.

(221) Carey, J. L.; Guo, L. Liquid-Liquid Phase Separation of TDP-43 and FUS in Physiology and Pathology of Neurodegenerative Diseases. *Frontiers in Molecular Biosciences* **2022**, *9*, 826719.

(222) Ambadipudi, S.; Reddy, J. G.; Biernat, J.; Mandelkow, E.; Zweckstetter, M. Residue-specific identification of phase separation hot spots of Alzheimer's-related protein tau. *Chemical Science* **2019**, *10*, 6503–6507.

(223) Wheeler, R. J. Therapeutics-how to treat phase separation-associated diseases. *Emerging Topics in Life Sciences* **2020**, *4*, 331–342.

(224) Biesaga, M.; Frigolé-Vivas, M.; Salvatella, X. Intrinsically disordered proteins and biomolecular condensates as drug targets. *Current Opinion in Chemical Biology* **2021**, *62*, 90–100.

(225) Kroschwald, S.; Maharana, S.; Alberti, S. Hexanediol: a chemical probe to investigate the material properties of membrane-less compartments. *Matters* **2017**, *3*, e201702000010.

(226) Schanda, P. Development and application of fast NMR methods for the study of protein structure and dynamics. **2007**, PhD thesis. Université Joseph-Fourier - Grenoble I.

(227) Fernández, C.; Wider, G. TROSY in NMR studies of the structure and function of large biological macromolecules. *Current Opinion in Structural Biology* **2003**, *13*, 570–580.

(228) Wishart, D. S. Interpreting protein chemical shift data. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2011**, *58*, 62–87.

(229) Johnson, C. S. Diffusion ordered nuclear magnetic resonance spectroscopy: principles and applications. *Progress in Nuclear Magnetic Resonance Spectroscopy* **1999**, *34*, 203–256.

(230) Karunanithy, G. Development of NMR methodology for the study of complex molecular and biomolecular systems. **2019**, PhD thesis. University of Oxford.

(231) Baldwin, A. J.; Kay, L. E. NMR spectroscopy brings invisible protein states into focus. *Nature Chemical Biology* **2009**, *5*, 808–814.

(232) Casablancas-Antràs, V. Asymmetry and allostery in IgE-Fc. **2016**, MRes thesis. King's College London.

(233) Karamanos, T. K.; Kalverda, A. P.; Thompson, G. S.; Radford, S. E. Mechanisms of amyloid formation revealed by solution NMR. *Progress in Nuclear Magnetic Resonance Spectroscopy* **2015**, *88-89*, 86–104.

(234) Schlepckow, K.; Schwalbe, H.; Schlepckow, [K.; Schwalbe, H. Molecular Mechanism of Prion Protein Oligomerization at Atomic Resolution. *Angewandte Chemie International Edition* **2013**, *52*, 10002–10005.

(235) Bellomo, G.; Bologna, S.; Gonnelli, L.; Ravera, E.; Fragai, M.; Lelli, M.; Luchinat, C. Aggregation kinetics of the Aβ1-40 peptide monitored by NMR. *Chem. Commun* **2018**, *54*, 7601.

(236) Hutchison, M. T.; Bellomo, G.; Cherepanov, A.; Stirnal, E.; Fürtig, B.; Richter, C.; Linhard, V.; Gurewitsch, E.; Lelli, M.; Morgner, N.; Schrader, T.; Schwalbe, H. Modulation of Aβ42 Aggregation Kinetics and Pathway by Low-Molecular-Weight Inhibitors. *ChemBioChem* **2023**, *24*, e202200760.

(237) Hoffmann, A. R. F.; Caillon, L.; Shadai, L.; Vazquez, S.; Spath, P.-A.; Carlier, L.; Khemté, L.; Lequin, O. Time dependence of NMR observables reveals salient differences in the accumulation of early aggregated species between human islet amyloid polypeptide and amyloid-β. *Physical Chemistry Chemical Physics* **2018**, *20*, 9561.

(238) Ceccon, A.; Tugarinov, V.; Torricella, F.; Clore, G. M. Quantitative NMR analysis of the kinetics of prenucleation oligomerization and aggregation of pathogenic huntingtin exon-1 protein. *Proceedings of the National Academy of Sciences of the United States of America* **2022**, *119*, e2207690119.

(239) Murthy, A. C.; Fawzi, N. L. The (un)structural biology of biomolecular liquid-liquid phase separation using NMR spectroscopy. *Journal of Biological Chemistry* **2020**, *295*, 2375–2384.

(240) Gamblin, T. C.; Berry, R. W.; Binder, L. I. Current Topics Modeling Tau Polymerization in Vitro: A Review and Synthesis. *Biochemistry* **2003**, *42*, 15009–15017.

(241) Adamcik, J.; Sánchez-Ferrer, A.; Ait-Bouziad, N.; Reynolds, N. P.; Lashuel, H. A.; Mezzenga, R. Microtubule-Binding R3 Fragment from Tau Self-Assembles into Giant Multistranded Amyloid Ribbons. *Angewandte Chemie* **2016**, *128*, 628–632.

(242) Wilkins, D. K.; Grimshaw, S. B.; Receveur, V.; Dobson, C. M.; Jones, J. A.; Smith, L. J. Hydrodynamic radii of native and denatured proteins measured by pulse field gradient NMR techniques. *Biochemistry* **1999**, *38*, 16424–31.

(243) Marsh, J. A.; Forman-Kay, J. D. Sequence Determinants of Compaction in Intrinsically Disordered Proteins. *Biophysical Journal* **2010**, *98*, 2383–2390.

(244) Falcon, B.; Cavallini, A.; Angers, R.; Glover, S.; Murray, T. K.; Barnham, L.; Jackson, S.; O'Neill, M. J.; Isaacs, A. M.; Hutton, M. L.; Szekeres, P. G.; Goedert, M.; Bose, S. Conformation determines the seeding potencies of native and recombinant Tau aggregates. *Journal of Biological Chemistry* **2015**, *290*, 1049–65.

(245) Carlomagno, Y. et al. The AD tau core spontaneously self-assembles and recruits full-length tau to filaments. *Cell Reports* **2021**, *34*, 108843.

(246) Stöhr, J. et al. A 31-residue peptide induces aggregation of tau's microtubule-binding region in cells. *Nature Chemistry* **2017**, *9*, 874–881.

(247) Annadurai, N.; Malina, L.; Malohlava, J.; Hajduch, M.; Das, V. Tau R2 and R3 are essential regions for tau aggregation, seeding and propagation. *Biochimie* **2022**, *200*, 79–86.

(248) Novak, M.; Kabat, J.; Wischik, C. M. Molecular characterization of the minimal protease resistant tau unit of the Alzheimer's disease paired helical filament. *The EMBO Journal* **1993**, *12*, 365–370.

(249) Oosawa, F.; Kasai, M. A theory of linear and helical aggregations of macromolecules. *Journal of Molecular Biology* **1962**, *4*, 10–21.

(250) Wales, D. J.; Doye, J. P. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. *Journal of Physical Chemistry A* **1997**, *101*, 5111–5116.

(251) Olson, B.; Hashmi, I.; Molloy, K.; Shehu, A. Basin Hopping as a General and Versatile Optimization Framework for the Characterization of Biological Macromolecules. *Advances in Artificial Intelligence* **2012**, *2012*, 674832.

(252) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P., *Numerical Recipes in C. The Art of Scientific Computing*, 2nd; Cambridge University Press: 1992, pp 1–1018.

(253) Perry, G.; Kawai, M.; Tabaton, M.; Onorato, M.; Mulvihill, P.; Richey, P.; Morandi, A.; Connolly, J. A.; Gambetti, P. Neuropil threads of Alzheimer's disease show a marked alteration of the normal cytoskeleton. *Journal of Neuroscience* **1991**, *11*, 1748–1755.

(254) Nott, T. J.; Craggs, T. D.; Baldwin, A. J. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nature Chemistry* **2016**, *8*, 569–575.

(255) Crozat, A.; Åman, P.; Mandahl, N.; Ron, D. Fusion of CHOP to a novel RNA-binding protein in human myxoid liposarcoma. *Nature* **1993**, *363*, 640–644.

(256) Åman, P.; Panagopoulos, I.; Lassen, C.; Fioretos, T.; Mencinger, M.; Toresson, H.; Höglund, M.; Forster, A.; Rabbitts, T. H.; Ron, D.; Mandahl, N.; Mitelman, F. Expression Patterns of the Human Sarcoma-Associated Genes FUS and EWS and the Genomic Structure of FUS. *Genomics* **1996**, *37*, 1–8.

(257) Andersson, M. K.; Ståhlberg, A.; Arvidsson, Y.; Olofsson, A.; Semb, H.; Stenman, G.; Nilsson, O.; Åman, P. The multifunctional FUS, EWS and TAF15 proto-oncoproteins show cell type-specific expression patterns and involvement in cell spreading and stress response. *BMC Cell Biology* **2008**, *9*, 37.

(258) Zinszner, H.; Sok, J.; Immanuel, D.; Yin, Y.; Ron, D. TLS (FUS) binds RNA in vivo and engages in nucleo-cytoplasmic shuttling. *Journal of Cell Science* **1997**, *110*, 1741–1750.

(259) Deng, H.; Gao, K.; Jankovic, J. The role of FUS gene variants in neurodegenerative diseases. *Nature Reviews Neurology* **2014**, *10*, 337–348.

(260) Dormann, D.; Rodde, R.; Edbauer, D.; Bentmann, E.; Fischer, I.; Hruscha, A.; Than, M. E.; MacKenzie, I. R.; Capell, A.; Schmid, B.; Neumann, M.; Haass, C. ALS-associated fused in sarcoma (FUS) mutations disrupt transportin-mediated nuclear import. *EMBO Journal* **2010**, *29*, 2841–2857.

(261) Bosco, D. A.; Lemay, N.; Ko, H. K.; Zhou, H.; Burke, C.; Kwiatkowski, T. J.; Sapp, P.; Mckenna-Yasek, D.; Brown, R. H.; Hayward, L. J. Mutant FUS proteins that cause amyotrophic lateral sclerosis incorporate into stress granules. *Human Molecular Genetics* **2010**, *19*, 4160–4175.

(262) Vance, C.; Scotter, E. L.; Nishimura, A. L.; Troakes, C.; Mitchell, J. C.; Kathe, C.; Urwin, H.; Manser, C.; Miller, C. C.; Hortobágyi, T.; Dragunow, M.; Rogelj, B.; Shaw, C. E. ALS mutant FUS disrupts nuclear localization and sequesters wild-type FUS within cytoplasmic stress granules. *Human Molecular Genetics* **2013**, *22*, 2676–2688.

(263) Vance, C. et al. Mutations in FUS, an RNA processing protein, cause familial amyotrophic lateral sclerosis type 6. *Science* **2009**, *323*, 1208–1211.

(264) Kwiatkowski, T. J. et al. Mutations in the FUS/TLS gene on chromosome 16 cause familial amyotrophic lateral sclerosis. *Science* **2009**, *323*, 1205–1208.

(265) Neumann, M.; Rademakers, R.; Roeber, S.; Baker, M.; Kretzschmar, H. A.; MacKenzie, I. R. A new subtype of frontotemporal lobar degeneration with FUS pathology. *Brain* **2009**, *132*, 2922–2931.

(266) Langenhove, T. V.; Zee, J. V. D.; Broeckhoven, C. V. The molecular basis of the frontotemporal lobar degeneration–amyotrophic lateral sclerosis spectrum. *Annals of Medicine* **2012**, *44*, 817–828.

(267) Mackenzie, I. R.; Rademakers, R.; Neumann, M. TDP-43 and FUS in amyotrophic lateral sclerosis and frontotemporal dementia. *The Lancet Neurology* **2010**, *9*, 995–1007.

(268) Urwin, H. et al. FUS pathology defines the majority of tau-and TDP-43-negative frontotemporal lobar degeneration. *Acta Neuropathologica* **2010**, *120*, 33–41.

(269) Snowden, J. S.; Hu, Q.; Rollinson, S.; Halliwell, N.; Robinson, A.; Davidson, Y. S.; Momeni, P.; Baborie, A.; Griffiths, T. D.; Jaros, E.; Perry, R. H.; Richardson, A.; Pickering-Brown, S. M.; Neary, D.; Mann, D. M. The most common type of FTLD-FUS (aFTLD-U) is associated with a distinct clinical form of frontotemporal dementia but is not related to mutations in the FUS gene. *Acta Neuropathologica* **2011**, *122*, 99–110.

(270) Mackenzie, I. R.; Neumann, M. Fused in Sarcoma Neuropathology in Neurodegenerative Disease. *Cold Spring Harbor Perspectives in Medicine* **2017**, *7*, a024299.

(271) Ling, S.-C.; Polymenidou, M.; Cleveland, D. W. Review Converging Mechanisms in ALS and FTD: Disrupted RNA and Protein Homeostasis. *Neuron* **2013**, *79*, 416–438.

(272) Murray, D. T.; Kato, M.; Lin, Y.; Thurber, K. R.; Hung, I.; Mcknight, S. L.; Correspondence, R. T. Structure of FUS Protein Fibrils and Its Relevance to Self-Assembly and Phase Separation of Low-Complexity Domains. *Cell* **2017**, *171*, 615–627.

(273) Yoshizawa, T. et al. Nuclear Import Receptor Inhibits Phase Separation of FUS through Binding to Multiple Sites. *Cell* **2018**, *173*, 693–705.

(274) Monahan, Z.; Ryan, V. H.; Janke, A. M.; Burke, K. A.; Rhoads, S. N.; Zerze, G. H.; O'Meally, R.; Dignon, G. L.; Conicella, A. E.; Zheng, W.; Best, R. B.; Cole, R. N.; Mittal, J.; Shewmaker, F.; Fawzi, N. L. Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity. *The EMBO Journal* **2017**, *36*, 2951–2967.

(275) Wheeler, R. J. et al. Small molecules for modulating protein driven liquid-liquid phase separation in treating neurodegenerative disease. *bioRxiv* **2019**, doi: 10.1101/721001.

(276) Protter, D. S. W.; Parker, R. Principles and Properties of Stress Granules. *Trends in Cell Biology* **2016**, *26*, 668–679.

(277) Jain, S.; Wheeler, J. R.; Walters, R. W.; Agrawal, A.; Barsic, A.; Parker, R. ATPase-Modulated Stress Granules Contain a Diverse Proteome and Substructure. *Cell* **2016**, *164*, 487–498.

(278) Mackenzie, I. R. et al. TIA1 Mutations in Amyotrophic Lateral Sclerosis and Frontotemporal Dementia Promote Phase Separation and Alter Stress Granule Dynamics. *Neuron* **2017**, *95*, 808–816.

(279) Murakami, T.; Qamar, S.; Lin, J. Q.; Vendruscolo, M.; Kaminski, C. F.; St, P.; Correspondence, G.-H. ALS/FTD Mutation-Induced Phase Transition of FUS Liquid Droplets and Reversible Hydrogels into Irreversible Hydrogels Impairs RNP Granule Function. *Neuron* **2015**, *88*, 678–690.

(280) Kotaja, N.; Bhattacharyya, S. N.; Jaskiewicz, L.; Kimmins, S.; Parvinen, M.; Filipowicz, W.; Sassone-Corsi, P. The chromatoid body of male germ cells: Similarity with processing bodies and presence of Dicer and microRNA pathway components. *Proceedings of the National Academy of Sciences of the United States of America* **2006**, *103*, 2647–2652.

(281) Kotaja, N.; Sassone-Corsi, P. The chromatoid body: a germ-cell-specific RNA-processing centre. *Nature Reviews Molecular Cell Biology* **2007**, *8*, 85–90.

(282) Cordin, O.; Banroques, J.; Tanner, N. K.; Linder, P. The DEAD-box protein family of RNA helicases. *Gene* **2006**, *367*, 17–37.

(283) Sassone-Corsi, P. The cyclic AMP pathway. *Cold Spring Harbor Perspectives in Biology* **2012**, *4*, a011148.

(284) Denninger, J. W.; Marletta, M. A. Guanylate cyclase and the NO/cGMP signaling pathway. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* **1999**, *1411*, 334–350.

(285) Patel, A.; Malinovska, L.; Saha, S.; Wang, J.; Alberti, S.; Krishnan, Y.; Hyman, A. A. ATP as a biological hydrotrope. *Science* **2017**, *356*, 753–756.

(286) Sridharan, S.; Kurzawa, N.; Werner, T.; Günthner, I.; Helm, D.; Huber, W.; Bantscheff, M.; Savitski, M. M. Proteome-wide solubility and thermal stability profiling reveals distinct regulatory roles for ATP. *Nature Communications* **2019**, *10*, 1155.

(287) Rice, A. M.; Rosen, M. K. Cell biology: ATP controls the crowd. *Science* **2017**, *356*, 701–702.

(288) Kang, J.; Lim, L.; Song, J. ATP enhances at low concentrations but dissolves at high concentrations liquid-liquid phase separation (LLPS) of ALS/FTD-causing FUS. *Biochemical and Biophysical Research Communications* **2018**, *504*, 545–551.

(289) Kang, J.; Lim, L.; Lu, Y.; Song, J. A unified mechanism for LLPS of ALS/FTLD-causing FUS as well as its modulation by ATP and oligonucleic acids. *PLoS Biology* **2019**, *17*, e3000327.

(290) Dang, M.; Lim, L.; Kang, J.; Song, J. ATP biphasically modulates LLPS of TDP-43 PLD by specifically binding arginine residues. *Communications Biology* **2021**, *4*, 714.

(291) Dang, M.; Li, Y.; Song, J. ATP biphasically modulates LLPS of SARS-CoV-2 nucleocapsid protein and specifically binds its RNA-binding domain. *Biochemical and Biophysical Research Communications* **2021**, *541*, 50–55.

(292) Maharana, S. et al. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science* **2018**, *360*, 918–921.

(293) Wang, L.; Kang, J.; Lim, L.; Wei, Y.; Song, J. TDP-43 NTD can be induced while CTD is significantly enhanced by ssDNA to undergo liquid-liquid phase separation. *Biochemical and Biophysical Research Communications* **2018**, *499*, 189–195.

(294) Dang, M.; Song, J. ALS-causing D169G mutation disrupts the ATP-binding capacity of TDP-43 RRM1 domain. *Biochemical and Biophysical Research Communications* **2020**, *524*, 459–464.

(295) Mehringer, J.; Do, T. M.; Touraud, D.; Hohenschutz, M.; Khoshsima, A.; Horinek, D.; Kunz, W. Hofmeister versus Neuberg: is ATP really a biological hydrotrope? *Cell Reports Physical Science* **2021**, *2*, 100343.

(296) Dang, M.; Kang, J.; Lim, L.; Li, Y.; Wang, L.; Song, J. ATP is a cryptic binder of TDP-43 RRM domains to enhance stability and inhibit ALS/AD-associated fibrillation. *Biochemical and Biophysical Research Communications* **2020**, *522*, 247–253.

(297) Nishizawa, M.; Walinda, E.; Morimoto, D.; Kohn, B.; Scheler, U.; Shirakawa, M.; Sugase, K. Effects of Weak Nonspecific Interactions with ATP on Proteins. *Journal of the American Chemical Society* **2021**, *143*, 11982–11993.

(298) Alshareedah, I.; Kaur, T.; Ngo, J.; Seppala, H.; Kounatse, L. A. D.; Wang, W.; Moosa, M. M.; Banerjee, P. R. Interplay between Short-Range Attraction and Long-Range Repulsion Controls Reentrant Liquid Condensation of Ribonucleoprotein-RNA Complexes. *Journal of the American Chemical Society* **2019**, *141*, 14593–14602.

(299) Mogami, G.; Wazawa, T.; Morimoto, N.; Kodama, T.; Suzuki, M. Hydration properties of adenosine phosphate series as studied by microwave dielectric spectroscopy. *Biophysical Chemistry* **2011**, *154*, 1–7.

(300) Babl, L.; Giacomelli, G.; Ramm, B.; Gelmroth, A. K.; Bramkamp, M.; Schwille, P. CTP-controlled liquid–liquid phase separation of ParB. *Journal of Molecular Biology* **2022**, *434*, 167401.

(301) Boeynaems, S.; Holehouse, A. S.; Weinhardt, V.; Kovacs, D.; Lindt, J. V.; Larabell, C.; Bosch, L. V. D.; Das, R.; Tompa, P. S.; Pappu, R. V.; Gitler, A. D. Spontaneous driving forces give rise to protein-RNA condensates with coexisting phases and complex material properties. *Proceedings of the National Academy of Sciences of the United States of America* **2019**, *116*, 7889–7898.

(302) Kunz, W.; Holmberg, K.; Zemb, T. Hydrotropes. *Current Opinion in Colloid and Interface Science* **2016**, *22*, 99–107.

(303) Bye, J.; Murray, K.; Curtis, R. ATP and tri-polyphosphate (TPP) suppress protein aggregate growth by a supercharging mechanism. *Biomedicines* **2021**, *9*, 1646.

(304) Lenton, S.; Hervø-Hansen, S.; Popov, A. M.; Tully, M. D.; Lund, M.; Skepö, M. Impact of Arginine-Phosphate Interactions on the Reentrant Condensation of Disordered Proteins. *Biomacromolecules* **2021**, *22*, 1532–1544.

(305) Fersht, A. R. The hydrogen bond in molecular recognition. *Trends in Biochemical Sciences* **1987**, *12*, 301–304.

(306) Baldwin, R. L. Energetics of Protein Folding. *Journal of Molecular Biology* **2007**, *371*, 283–301.

(307) Bielskutė, S.; Garcia-Cabau, C.; Frigolé-Vivas, M.; Szulc, E.; Mol, E. D.; Pesarrodona, M.; García, J.; Salvatella, X. Low amounts of heavy water increase the phase separation propensity of a fragment of the androgen receptor activation domain. *Protein Science* **2021**, *30*, 1427–1437.

(308) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **2001**, *46*, 3–26.

(309) Ertl, P.; Rohde, B.; Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *Journal of Medicinal Chemistry* **2000**, *43*, 3714–3717.

(310) Palm, K.; Stenberg, P.; Luthman, K.; Artursson, P. Polar molecular surface properties predict the intestinal absorption of drugs in humans. *Pharmaceutical Research* **1997**, *14*, 568–571.

(311) Kelder, J.; Grootenhuis, P. D.; Bayada, D. M.; Delbressine, L. P.; Ploemen, J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs. *Pharmaceutical Research* **1999**, *16*, 1514–1519.

(312) Gromiha, M. M.; Santhosh, C.; Suwa, M. Influence of cation-π interactions in protein-DNA complexes. *Polymer* **2004**, *45*, 633–639.

(313) Kumar, K.; Woo, S. M.; Siu, T.; Cortopassi, W. A.; Duarte, F.; Paton, R. S. Cation–π interactions in protein–ligand binding: theory and data-mining reveal different roles for lysine and arginine. *Chemical Science* **2018**, *9*, 2655–2665.

(314) Scheller, K. H.; Hofstetter, F.; Mitchell, P. R.; Prijs, B.; Sigel, H. Macrochelate Formation in Monomeric Metal Ion Complexes of Nucleoside 5'-Triphosphates and the Promotion of Stacking by Metal Ions. Comparison of the Self-Association of Purine and Pyrimidine 5'-Triphosphates Using Proton Nuclear Magnetic Resonance. *Journal of the American Chemical Society* **1981**, *103*, 247–260.

(315) Rutledge, L. R.; Campbell-Verduyn, L. S.; Wetmore, S. D. Characterization of the stacking interactions between DNA or RNA nucleobases and the aromatic amino acids. *Chemical Physics Letters* **2007**, *444*, 167–175.

(316) Luscombe, N. M.; Laskowski, R. A.; Thornton, J. M. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. *Nucleic Acids Research* **2001**, *29*, 2860–2874.

(317) Cheng, A. C.; Chen, W. W.; Fuhrmann, C. N.; Frankel, A. D. Recognition of Nucleic Acid Bases and Base-pairs by Hydrogen Bonding to Amino Acid Side-chains. *Journal of Molecular Biology* **2003**, *327*, 781–796.

(318) Seeman, N. C.; Rosenberg, J. M.; Rich, A. Sequence-specific recognition of double helical nucleic acids by proteins. *Proceedings of the National Academy of Sciences of the United States of America* **1976**, *73*, 804–808.

(319) Kondo, J.; Westhof, E. Classification of pseudo pairs between nucleotide bases and amino acids by analysis of nucleotide–protein complexes. *Nucleic Acids Research* **2011**, *39*, 8628–8637.

(320) Yu, B.; Pletka, C. C.; Iwahara, J. NMR Observation of Intermolecular Hydrogen Bonds between Protein Tyrosine Side-Chain OH and DNA Phosphate Groups. *Journal of Physical Chemistry B* **2020**, *124*, 1065–1070.

(321) Rooman, M.; Vin, J. L.; Buisine, E.; Wintjens, R. Cation-π/H-bond Stair Motifs at Protein-DNA Interfaces. *Journal of Molecular Biology* **2002**, *319*, 67–76.

(322) Crowley, P. B.; Golovin, A. Cation-π Interactions in Protein-Protein Interfaces. *Proteins: Structure, Function and Bioinformatics* **2005**, *59*, 231–239.

(323) Ruiter, A. D.; Zagrovic, B. Absolute binding-free energies between standard RNA/DNA nucleobases and amino-acid sidechain analogs in different environments. *Nucleic Acids Research* **2015**, *43*, 708–718.

(324) Ruiter, A. D.; Polyansky, A. A.; Zagrovic, B. Dependence of Binding Free Energies between RNA Nucleobases and Protein Side Chains on Local Dielectric Properties. *Journal of Chemical Theory and Computation* **2017**, *13*, 4504–4513.

(325) Treger, M.; Westhof, E. Statistical analysis of atomic contacts at RNA–protein interfaces. *Journal of Molecular Recognition* **2001**, *14*, 199–214.