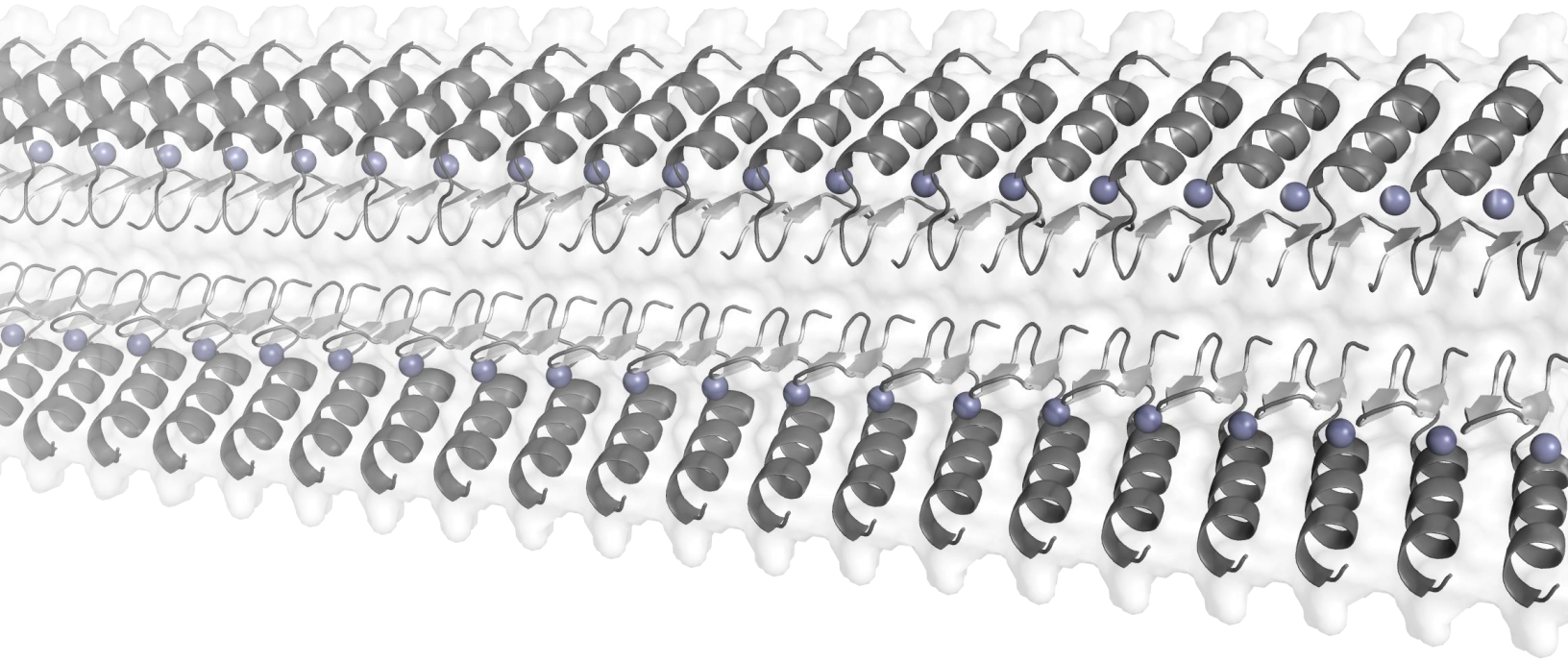
Controlling the self-assembly of *de novo* designed amyloid structures



Mads Jeppesen

Master thesis in nanoscience

Supervisor: Ebbe Sloth Andersen

Co-supervisor: Daniel Otzen

Collaborators: Per Greisen and Ingemar André

Aarhus University

June 2018

“What I cannot create, I do not understand.”

- Richard Feynman

Abstract

De novo protein design is an emerging methodology that can be used to design entirely new proteins not seen in nature. It promises to completely revolutionize protein engineering, and already a range of *de novo* designed proteins have been created such as novel enzymes or artificial viruses. Although great strides have been made in *de novo* protein design, the design of conformational switches and protein assemblies remain challenging. A type of protein assembly that has for many years garnered significant attention are amyloids. These have important roles in many neurological diseases but have recently also been used in a range of technological applications. The project explained in this thesis attempts to *de novo* design an amyloid conformational switch that can be induced to fold into an amyloid in the presence of zinc and unfold when zinc is removed. A procedure was developed to *de novo* design these structures. To develop it, an up to date analysis of zinc sites deposited in the Protein Data Bank were carried out, and a new algorithm was created. Initial test showed that the amyloid structure could be created using this procedure with a zinc site inside and between the individual monomers making up the amyloid. The test also showed that the generated structures needed to be optimized, and in the future large-scale sampling and improvements to the computational procedure must be carried out. When satisfiable computational models have been created these models have to be experimentally characterized. The successful design of de novo designed amyloid conformational switches would have many applications both for fundamental research, technological applications, and for the advancement of the de novo protein design.

Acknowledgements

First and foremost, I would like to thank all the people in the ESA lab. To my supervisor, Ebbe Sloth Andersen for being open minded and crazy enough to allow me to start a project involving protein design and for supporting me all the way through to the end in both the project and my career. Thank you to Andreas for the support you have given me in any way possible, for the discussions we had and for the many coffee expeditions to the iNANO foyer.

Thanks goes to Ingemar André for allowing me to visit his group in Lund. You showed a genuine interest in the project and in helping me develop it and your inputs were highly valuable. Thank you also to Christoffer Norn for your support.

Thank you to Per Greisen for your inputs, helping me understand Rosetta and for troubleshooting all the terrible mistakes I made. It was a rough beginning with not a lot of progress. I hope we can work together in the future as well.

Thank you to Daniel Otzen for inputs in the beginning and for being ready for me to join the lab and do experiments.

Thank you to Jesper Karlsen for allowing us to use Rosetta on the computer cluster and for helping to set us up.

Thank you to Novo Nordisk for funding me personally and for inviting me to Bagsværd and all the exciting events.

Finally, I would like to thank all my friends and family who were very supportive and understanding throughout. Thank you to my parents for picking me up late in the evening at the office and for always welcoming me home when I needed it.

List of abbreviations

BB:	Backbone
BM:	Binding mode
CCD:	Cyclic Coordinate Descent
CN:	Coordination number
CPD:	Computational protein design
DOF:	Degrees of freedom
GM:	Geometric center
HV:	Helix vector
MC:	Monte Carlo
MDPA:	Metal Directed Protein Assembly
MSD:	Multi state design
PDB:	Protein Data Bank
PPI:	Protein-Protein Interaction.
SA:	Simulated annealing
SC:	Side chain
SS:	Secondary structure
SSD:	Single state design
SZ:	Steric zipper

Table of Figures

Appendix A

Appendix B

Appendix C

Table of Tables

Table of Equations

Table of Contents

1 Introduction

1.1 Protein structure

Proteins are the molecules of life. They exist in all living entities and carry out a range of function such as catalyzing reactions [1], transducing signals [2], transportation [3], mechanical movement [4], and as structural building blocks [5]. Because of their fundamental role in biology, much attention has been invested in understanding their role and mechanisms, but equally so in finding ways to use them for new purposes that benefit humanity.

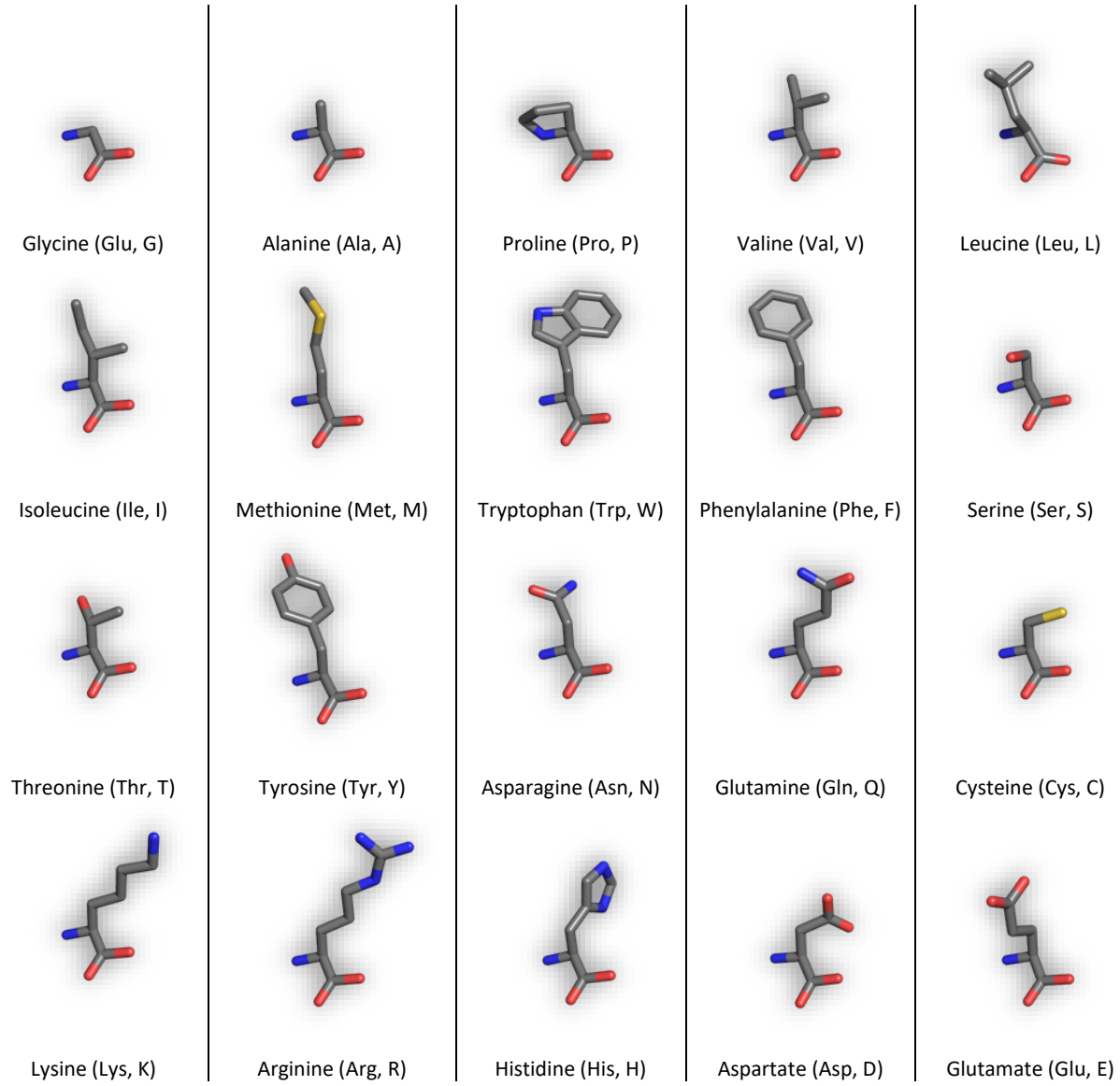
It is the structure of proteins that allow them to carry out these diverse functions, and to understand proteins, and why they are important, one needs to understand their structure. In this subchapter the principal features of protein structure and some definitions will be explained.

For an elaboration on some of the concepts, the reader is referred to reference [6].

1.1.1 Primary structure

Proteins consist of one or more unbranched linear heteropolymers called peptides (if the polymer is short) or polypeptides (if the polymer is long). The monomers of proteins are amino acids and genes in the genome of organisms encode the instruction for building proteins, that is, the sequence of amino acids in the polymer. The sequence constitutes the primary structure of a protein. Consecutive triplets of bases in genes, called codons, determines the amino acid sequence, which without modification can consist of up to 20 different canonical amino acids. All 20 canonical amino acids are shown in **Figure 1-1.a**. Amino acids have an identical part which will constitute part of the backbone (BB), and a unique part, called the side chain (SC), which is specific to each amino acid. The amino acids are connected through peptide bonds and the start of the of the protein chain is designated the N-terminal and starts with the amino group. Similarly, the end of the protein is designated as the C-terminal and ends with the carboxyl group. The structure of a hypothetical short protein is shown in **Figure 1-1.b** along with the definitions mentioned.

a



b

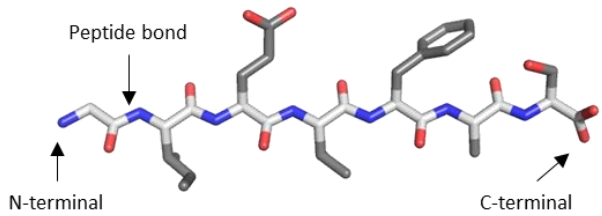


Figure 1-1: Names and definitions of the 20 canonical amino acids and the protein chain. a) All 20 canonical amino acids are shown along with their long name and two shorthand descriptions. Grey is carbon, blue is nitrogen, red is oxygen and yellow is sulfur. Hydrogens are not shown. **b)** A hypothetical protein is shown along with the definitions described in the main text (white is the BB, grey is the SCs).

1.1.2 Degrees of freedom

Proteins have certain degrees of freedom (DOF) and **Figure 1-2.a** highlights the most important ones along with their definitions. The most important atom names of amino acids are also shown. In this thesis, capital letters are used instead of the greek letters to describe atom names so α=A, β=B, γ=G, δ=D, ε=E. The torsional DOF of the BB is psi (ψ), phi (ϕ) and omega (ω). The ψ and ϕ torsion angles are quite restricted, which is evident from the Ramachandran plot [7] (**Figure 1-2.b**). Due to the partially double bond character of the peptide bond, the ω torsion angle is even more restricted and it only takes on values that keep the peptide bond planar. Because of steric repulsion the trans conformation (ω ≈ 180 degrees) is almost exclusively preferred over the cis conformation (ω ≈ 0 degrees), although in proline, the cis conformation is more pronounced. Since the ω angle is very restricted, the direction of the protein chain is determined mostly by the values of ψ and ϕ. The torsional DOF of a SC is given by a set of chi angles (χ) and again because of steric clashes, these angles are similarly quite constrained. The orientation of amino acids is therefore determined by the combinations of its χ torsion angles. These combinations can be classified into rotamers, which is discussed further in Chapter 3 along with rotamer libraries. Other degrees of freedom such as bond lengths and bond angles does not significantly change as much as the DOFs mentioned, and they are therefore relatively fixed.

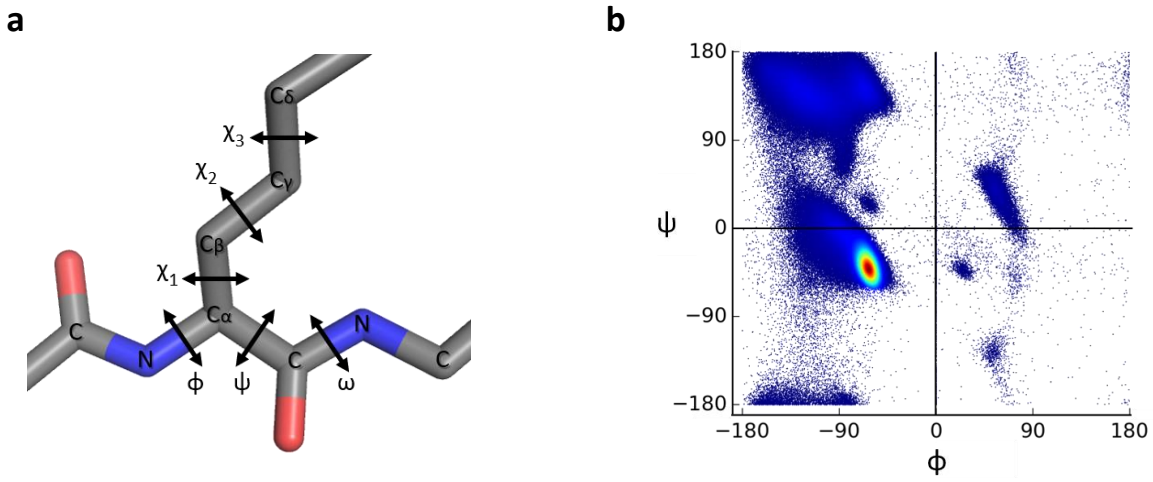


Figure 1-2: Torsion angle DOFs and the Ramachandran plot. a) The definitions of the torsion angle DOF for a polypeptide chain. Torsion angles are defined by 4 atoms starting from the N-terminal. ϕ: C-N-Cα-C, ψ: N-Cα-C-N, ω: Cα-C-N-C, χ_1: N-Cα-Cβ-Cγ etc. **b)** A Ramachandran plot showing ψ and ϕ angles (300.000 datapoints) for a subset of proteins from the PDB. The map is plotted as a heatmap showing the density of ψ and ϕ (from dark blue (low) to red (high)). It's clear ψ and ϕ angles only take certain values.

1.1.3 Beyond primary structure

As seen in **Figure 1-1.a,** each amino acid is unique, and chemically they have different features. It is the individual features of amino acids in the sequence of a protein, that ultimately determines its structure, and it is the structure that determines a protein's function. There are several designations associated with protein structure, and the lowest level; the primary structure, has already been described. The second lowest is the secondary structure (SS), which refers to the 3-dimensional structures of segments in the proteins, formed by hydrogen bonds between the BB C=O and N-H groups. Two important SS elements are worth mentioning. The right handed alpha helix [8] is made by having ϕ and ψ torsion angles around -60° and -50°, respectively. The β-sheet [9] which can be parallel or antiparallel is made by having ϕ and ψ torsion angles around -140 ˚ and 135 ˚ for the parallel orientation and -120 ˚ and 120 ˚ for the antiparallel orientation. In the Ramachandran plot in **Figure 1-2.b** these structures correspond to the high-density regions (light blue and red). The next two levels of structural classification are the tertiary, which refers to the overall 3-dimensional structure of a single polypeptide chain, and the quaternary structure which refers to the structure of several interacting polypeptide chains.

In the following two subchapters the structure and some properties of two classes of proteins are discussed: zinc proteins and amyloids. These will be important for the discussion on the design in Chapter 2.

1.2 The structure of zinc proteins

Proteins are known in many cases to associate with other molecules or metal ions for functional or structural purposes. In particular abundance are metal ions with one estimate claiming that they are present in half of all proteins [10]. Looking at all crystalized structures and statistics from the MetalPDB (metalweb) [11], it is evident that a range of different metals are associated with proteins, with the 10 most prominent being zinc (Zn), sodium (Na), potassium (K), magnesium (Mg), calcium (Ca) and the 1 row transition metals: manganese (Mn), iron (Fe), cobalt (Co), nickel (Ni), and cobber (Cu). Zinc is the second most abundant metal in proteins. The relative stability of the latter class and zinc, is empirically given by the Irving–Williams series as Mn(II) < Fe(II) < Co(II), Ni(II) < Cu(II) > Zn(II) where the position of zinc is variable [12].

1.2.1 Zinc and its coordination sphere in proteins

Zinc is rather different from the other transition metals mentioned above. Zn(II) is redox inert with filled d-orbitals and has the electron configuration [Ar]$3d^{10}$. As mentioned, zinc is one of most abundant metals in proteins, and some estimates that up to 10% of all proteins in the human proteome incorporates zinc [13]. As so, zinc plays important roles in biology, most of which are structural and catalytic [14]. The most prominent amino acid ligands to zinc are cysteine (Cys), histidine (His), aspartate (Asp) and glutamate (Glu) (**Figure 1-3.a**), with Cys and His being about twice as prominent as Asp and Glu [15]. Zinc is also known to bind to other ligands in and around proteins, such as water and the BB N-H and C=O groups [15, 16]. Cys coordinates to zinc though its sulfur atom (SG), histidine through either of its nitrogen groups (NE2 and ND1) and aspartate and glutamate either monodentate through either of its oxygen groups (OD1, OD2 or OE1, OE2, respectively) or through both of its oxygen groups (OD1|OD2 and OE1|OE2 respectively). In this thesis "|" designates bidentate binding, and all of the binding modes (BM) are designated with ZN:ATOM, where ATOM is replaced by the coordinating atom or atoms (Figure **1-3.a**).

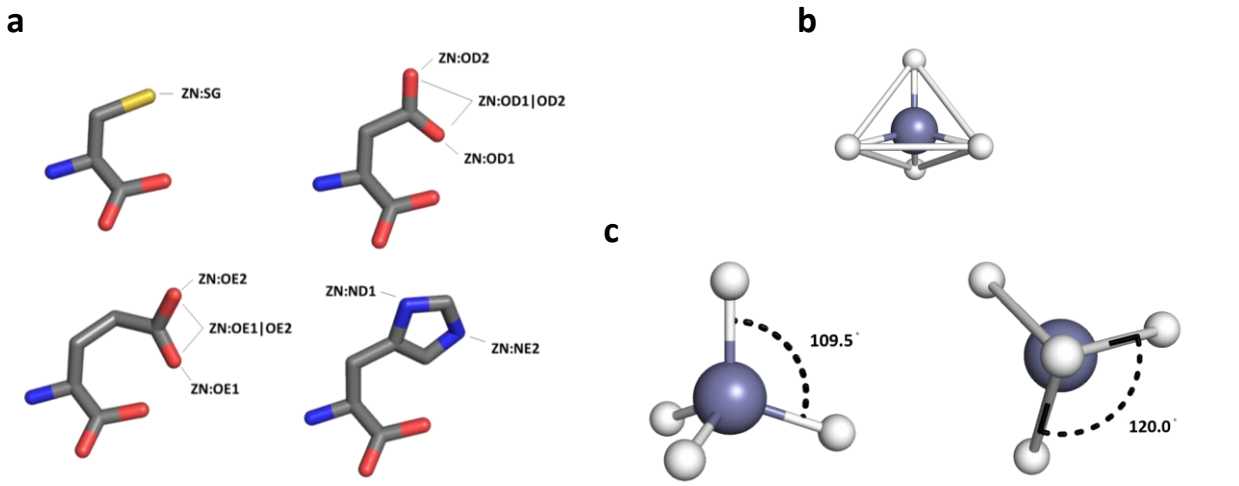


Figure 1-3: Common ligand residues and atoms to zinc and the tetrahedral coordination sphere. **a)** The amino acid structure of the most commons groups found bound to zinc along with designation of the binding modes. All residues can bind monodentate, but the carboxyl group of Asp and Glu can bind bidentate to zinc. **b)** The ideal tetrahedral coordination sphere of zinc. Zinc is shown in grey and ligands in white. A line shown between ligands illustrate the tetrahedral sphere which resembles a pyramid. **c)** The ideal tetrahedral coordination sphere has the following angle as shown. Left: the angle between 2 ligands in a plane spanned by the 2 ligands is 109.5°. Right: the angle between 2 ligands spanned by 3 ligands in a plane is 120.0°.

The coordination number (CN) of zinc proteins has been reported to range between 2-8 binding ligands, with the most common CN and geometry being 4 ligands in a tetrahedral arrangement and with most sites seeming to be mononuclear [15, 17, 18]. The ideal tetrahedral site is shown in **Figure 1-3b**. The angle between all ligands in an ideal tetrahedral arrangement is 109.5˚, with the separation between 3 ligands in a plane being 120˚(**Figure 1-3 c**).

1.2.2 Secondary structure of zinc proteins

One study [18] divided zinc proteins into specific classes depending on the structural protein environment around the zinc atom in a radius of 5 Å [18, 19]. Using this classification system, 75% of the original nonredundant zinc protein dataset used in that study could be divided into 10 distinct classes, while 16% could be divided into pseudo classes, that is, broadly similar classes, while the rest could not. **Figure 1-4** shows representatives of the 10 classes.

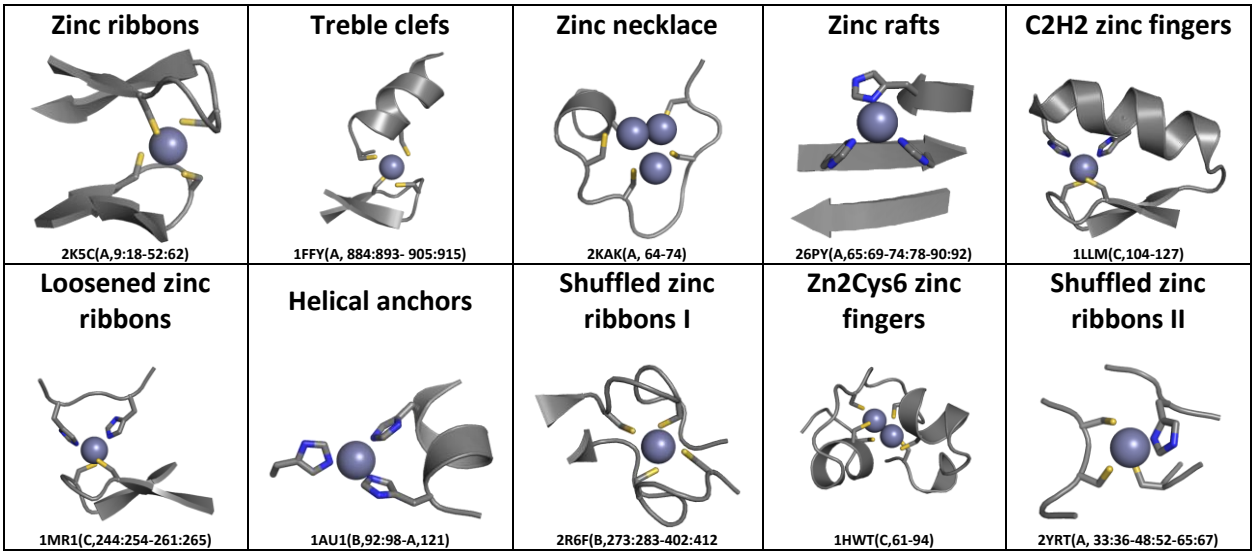


Figure 1-4: The 10 structural classes of zinc sites. The images show representatives for each class along with its PDB ID. The following is a rough description, see the reference [18] for more information. Zinc ribbons are composed of two β-hairpins each providing two ligands. Treble clefs consist of N/C-terminal β-hairpin and a C/N-terminal α-helix each providing 2 ligands. Zinc-necklaces has multiple zinc ligands oriented in a loop structure resembling a necklace. Zinc rafts consist of 3 adjacent β-strands, harboring 2-3 ligands. C2H2 zinc fingers consist of a β-hairpin and an α-helix each providing 2 ligands. Loosened zinc ribbons consist of a β-hairpin and a coil harboring 2 ligands each. Helical anchors consist of an α-helix providing 2 ligands, and another variable segment proving 1-2 ligands. Shuffled zinc ribbons I and II consist of zinc-binding-loops connecting 2 β-strands in different sheets. Zn2Cys6 zinc fingers consist of an α-helix and an extended coil.

1.2.3 Stability of small zinc proteins

As mentioned, zincs role in proteins is mainly structural and catalytic. But zinc has been known for a long time to also play a role in the folding and stability of proteins [20-24]. Small proteins which do not have a large hydrophobic interface, which is a major driving force for folding and stability of larger proteins [25, 26], are relying on extra binding forces such as metal coordination to fold and stabilize themselves [27, 28]. A class of proteins that have been significantly studied in this area is the zinc finger proteins, which are a small class of zinc binding proteins that have diverse functions, but primarily function as transcription factors [29]. One of the major classes of zinc fingers is the C2H2 zinc fingers which have a structure as identified above in **Figure 1-4** [29, 30]. Metals such as zinc has also been shown to affect folding in the class of proteins called amyloids, which will be discussed below and returned to in Chapter 2.

1.3 The Structure of Amyloids

Amyloids refer to a diverse set of proteins or peptides that form elongated and unbranched fibers with certain structural characteristics [31]. Amyloids are known to form a characteristic cross-β pattern, which consists of two or more β sheets stacking against each other (**Figure 1-5.a**). The distance between strands in the cross-β structure is 4.8 Å and the spacing between sheets can range from 5 to 16 Å. The β sheets run perpendicular to the fibrillar axis, and the residues of the involved sheets intertwine with each other to create a tight interface called a steric zipper (SZ) (**Figure 1-5.b**). The cross-β interface makes up the protofilament of amyloids, and two or more of such filaments can interact to create larger amyloid fibrils. **Figure 1-5.c** shows an example of a protofilament. The width of these amyloid fibers can vary from approximately 5-25 nm, with a typical size of around 10 nm, and with a length of up to 10 μm [32].

The first atomic structure of the cross-β spine structure solved by crystallography was of the 7-residue yeast peptide Sup35 [33]. Since then many more atomic structures of the cross-β spine has been deposited in the Protein Data Bank (PDB) [34]. Different from β sheets, as found in other proteins, the sheets in amyloid proteins consist of repetitive copies of the protein segments. From observations gained from many structural analyses such as crystallography, a classification system based on symmetry has emerged [35].

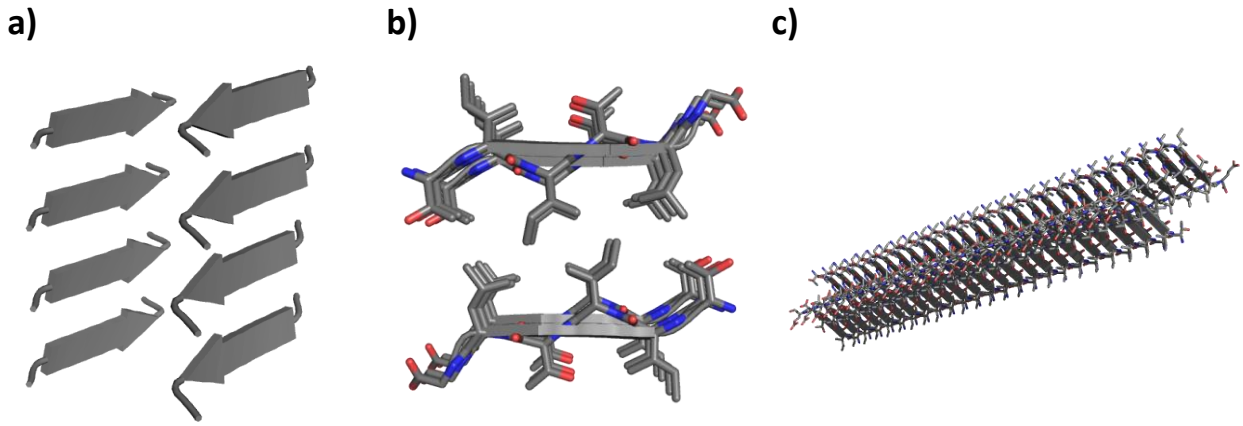
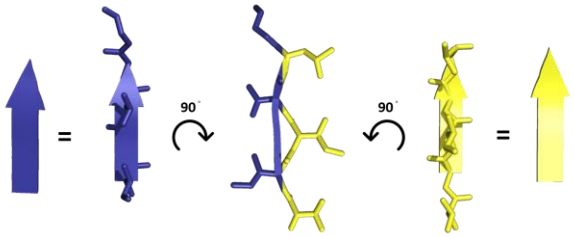


Figure 1-5: The general structure of amyloids. **a)** The cross-β structure typical of amyloid structures, where 2 or more sheets stack against each other. See **Figure 1-6** for the different possible symmetries the cross-β structure can make. **b)** The cross-β structure forms a tight interface known as a steric zipper. The interdigitating SC's from each strand looks like the teeth of a zipper and hence the name. **c)** Part of a crystal structure of a protofilament. This protofilament can act together with other protofilaments to create higher order structures called fibrils. The PDB ID of the structure shown is 4RP7.

Several layers of symmetry can be enumerated, but for this thesis, only the 10 elementary symmetry classes defined at the cross-β spine level are relevant (**Figure 1-6**). Each strand in the sheet has two nonidentical sides and if the two sides of the strands are colored blue and yellow it becomes easier to visualize the classes (**Figure 1-6.a**). 4 top level classes arise from the way the β-strands associate with each other inside 1 sheet (**Figure 1-6.b** top). The strands can either be parallel, or antiparallel, and the residue identities coming out from 1 side of the strand, can all point the same way or change direction for every successive strand along the fibril axis (change from blue to yellow or vice versa). If the strands successively change direction, it is called equifacial, and if not, it is called antifacial (**Figure 1-6.b** top). The rest of the symmetry classes arises from the way a pair of β sheets are associated with one another, and what the 4 top level symmetries are within 1 sheet (**Figure 1-6.b** bottom). Only 7 of those 10 classes have been experimentally identified so far (class 1, 2, 4, 5, 6, 7 and 8) and many structures for each class have been deposited in the PDB [34].

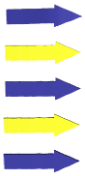
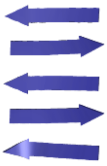
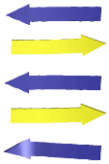
Having discussed the structures of 2 types of proteins, it is now time to discuss what proteins can do for us, and how a new method for protein engineering called *de novo* protein design has come of age.

a)



b)

Symmetry within one sheet

Parallel β sheet		Antiparallel β sheet	
Antifacial	Equifacial	Antifacial	Equifacial
			

Symmetry between sheets

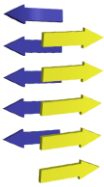
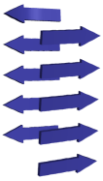
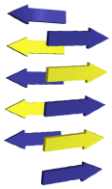
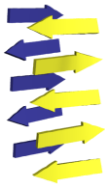
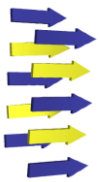
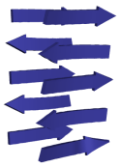
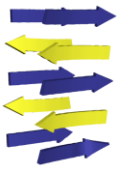
Class 1	Class 4	Class 10	Class 5	Class 8
				
Class 2	Class 3	Class 9	Class 6	Class 7
				

Figure 1-6: The 10 symmetry classes of the cross-β spine. a) The two nonidentical sides of a β strand can be colored either blue or yellow and simplified by the ribbons diagram as an arrow. **b)** Top: 4 classes can be classified based on the symmetry within 1 sheet. The sheet can be parallel or antiparallel, and the faces of the two sides of a strand can change between facing inward and outward (equifacial) or only facing inward or outward (antifacial), for every consecutive strand in the sheet. 10 symmetry classes can be classified based on the symmetry between sheets. Only class 1, 2, 4, 5, 6, 7 and 8 have been experimentally verified.

1.4 *De novo* Protein design

1.4.1 Engineered proteins are shaping the world

Humans have for many years used proteins for their own custom purposes. Even before we knew what proteins were, we were actively enjoying the fruits of what evolution had provided for us in terms of natural proteins, for instance in brewing beer and making cheese.

Beginning in the 1980's, with the advent of techniques such as site directed mutagenesis [36, 37], scientist began exploring the modification of proteins by changing their amino acid sequence [38]. At the time, scientist knew that proteins had huge potential for use in the industrial sector, in particular enzymes, but in many cases natural proteins were not suitable in industrial environments or the catalytic efficiency of the enzymes were not high enough. With these methods scientist could improve for instance the stability or reactivity of enzymes, and the field of protein engineering (or protein design) was born [38]. Two complementary approaches, called irrational and rational methods, are now the cornerstones of protein engineering. Irrational methods rely on random or "irrational" mutations made in a protein sequence followed by rounds of screening that selects for wanted characteristics. These methods are also called directed evolution [39]. Rational methods rely on rational choices on which residues should be mutated in a protein sequence. This is often accompanied by computational approaches. When rational designs are aided by computers it is called computational protein design (CPD) [40]. Over the last 4 decades, protein engineering have become a huge scientific field, and today we are using engineered proteins in a variety of areas such as in laundry detergents [41, 42], in foods [43], in the textile industry [44] and in medicine [45] among other things. Engineered proteins have truly shaped the world we live in.

1.4.2 The dark matter of proteins

Most efforts in protein engineering has been on natural proteins, where a small number of mutations have been carried out, to improve their intrinsic properties for an application. Using only existing proteins as a template for design is inherently limited because applications can, for the most part, only extend the functions provided to us by evolution. Furthermore, the available

folds used by natural proteins is limited (folds are defined as the placement of SSs in space, and estimates suggest that 1.000-10.000 folds exist [46-48]), so to create entirely new functions not seen in nature - "re"-engineering existing proteins is not always enough. The protein space (comprising the *sequence space*: the space of all potential sequences. The *structure space*: the space of all potential structures, and the *function space*: the space of all potential functions [48]) that we have not explored has been termed the "dark matter" of proteins [49], or another favorite of mine: the "never born proteins" [50]. Computer simulations that have probed this space indicate that it is larger than the known space of proteins [49, 51], and theoretically it should be many orders of magnitude larger [52]. Protein engineering efforts that explore this unknown space is termed *de novo* protein design; which is the design of new proteins from scratch, with sequences unseen in nature, having new folds or functions. Discovering new functional proteins in the "dark matter universe" promises to completely revolutionize technology [53, 54] from medicine, to computing, to industrial production and beyond.

The following subchapter will touch upon what has already been accomplished in d*e novo* protein design. Before moving on, note that the term "*de novo*" is used in connection with different methodologies, and with varying degree of novelty in the designed sequence, structure and function in the literature. This thesis will focus on *de novo* designs were CPD play a major role in the design process, and designs with a significant degree of novelty in sequence and structure, and function. Methods like directed evolution are still import for *de novo* design, in particular for downstream design and optimization, and novel proteins can even be derived from directed evolution [55, 56], but this will not be the focus of this thesis.

1.4.3 State of the art in *de novo* protein design

In recent years, with the advancement of computational power, computational algorithms, and cost and efficiency of DNA synthesis, the field of *de novo* protein design has seen great advances in the level of control that has been obtained on protein structure. This has meant a vast amount of new protein structures with novel sequences, folds and functions have been created. It would be impossible to enumerate all the incredible designs made in the past decade, but the following highlights important milestones**.**

The sequences and structures of small αβ-proteins [57-62], repeat proteins [63-69] and helical bundles [70, 71] (**Figure 1-7.a-c**) can now be completely designed from scratch, that is, every single amino acid in the protein. Already, small αβ-proteins have been tested as potential therapeutics for the treatment of influenza [72] and through massively parallel design, they have been used to understand protein folding [61]. Helical bundles has recently been used to create membrane proteins [73], membrane transporters [74], composite materials [75], ligand binding proteins [76], enzymes [77] and in protein assembly design such as in fibers [78, 79] and containers [80].

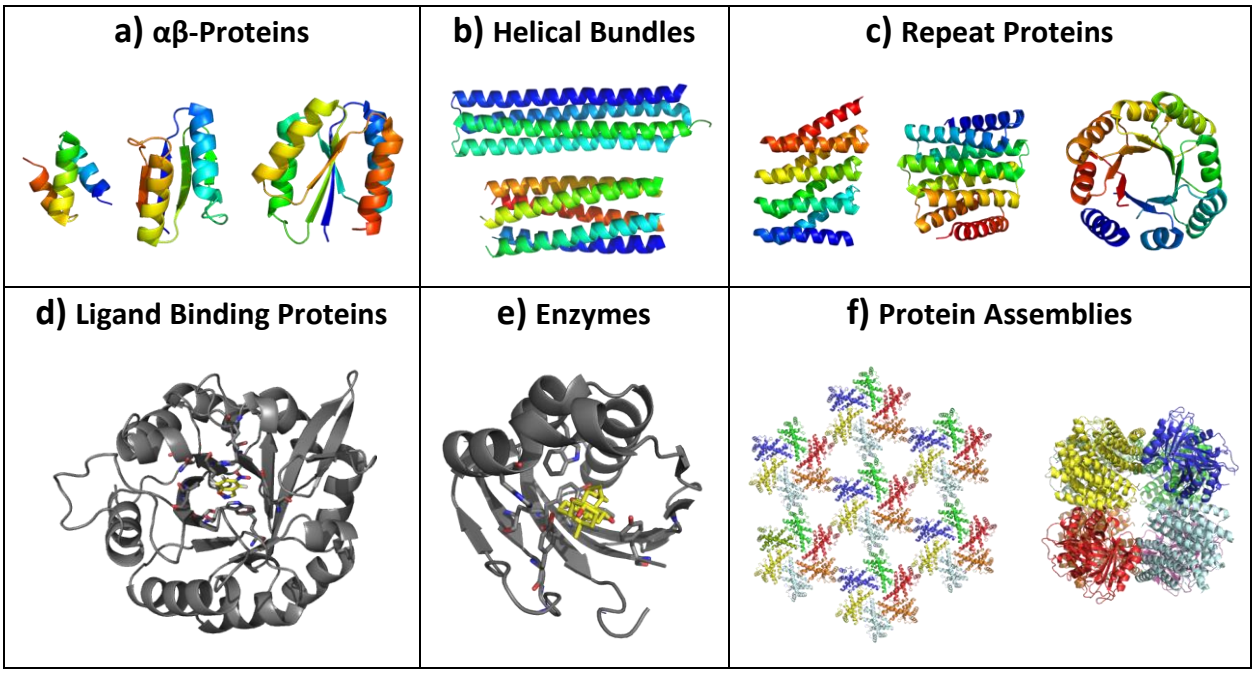


Figure 1-7: State of the art in *de novo* protein design. a) 3 examples of αβ-proteins (ref (left to right): [61], [59], [58]). **b)** Top: 4-helix-bundle. Bottom: 5-helix-bundle (ref: [70]). **c)** 3 examples of repeat proteins. 2 with linear symmetry (left and middle) (Ref [64]) and 1 with internal symmetry (Ref [68]). **d)** An example of a *de novo* designed ligand binding protein. The protein has been designed to bind the steroid digoxigenin (DIG) shown in yellow (ref: [81]). **e)** An example of an enzyme with the reaction intermediate state shown in yellow. The enzyme has been designed to catalyze the Kemp elimination reaction (ref: [82]). **f)** Examples of two types of protein assemblies. Left: 2D layers (ref: [83]). Right: 3D cage (ref: [84]). A single interface has been designed between the monomers to create these assemblies.

De novo ligand binding proteins [81, 85] and enzymes [82, 86-89] can also be created starting form natural scaffolds (**Figure 1-7.d-e**). *De novo* enzymes have been developed to potentially treat disease, such as celiac disease [90], or to decrease greenhouse gas emission [91], and *de novo* ligand binding proteins have been used as potential biosensors, such as in the detection of

fentanyl [92]. Finally, protein assemblies (**Figure 1-7.f**) designed through protein-protein interactions (PPIs), such as the development of hollow cages [84, 93], 1D fibers [94, 95], 2D layers [83] and 3D crystals [96], have been reported. The designed cages have already been used to design an artificial virus [97], which has potential uses in drug delivery.

De novo protein design has come a long way, but there are still many challenges to be overcome. 2 important challenges is the design of conformational switches [54] and designing PPIs to create proteins assemblies [98, 99]. These are discussed in more detail later.

The rest of this chapter will first be devoted to the theory behind protein folding and *de novo* protein design. This will lay the groundwork for understanding the status of *de novo* protein switch design and *de novo* protein assembly design, which will be elaborated upon at the end of the chapter. The chapter will then conclude with discussing a kind of protein assembly that for a long time has garnered significant attention, namely amyloids, and why they are important targets for *de novo* design.

1.4.4 Protein Folding and Energy Landscapes

A complete understanding of protein folding has not been accomplished, and over more than 50 years of work, it has come to be called the "the protein folding problem" [100]. One of the problems that baffled scientist in the beginning was how a protein could find its native state (the structural state that the protein is normally found to be in to carry out its function), out of all its non-native states. As an example, imaging a protein consisting of 100 residues, and assume that only the ϕ and ψ angles of that protein can move, and together ϕ and ψ can take only 10 discrete states (extreme simplification). Such a protein would be able to have about 10^{100} different structural states. This is many times larger than the number of electrons in the universe, and if a protein were to randomly sample all its states, it would take that protein much longer time to fold, than what experimentally was known to be true for proteins at the time. (This paradox was later called Levinthal's paradox [101]). But advances have been made over the years, and two important theories are behind our current understanding of protein folding [102]. The first theory (later to be called the Anfinsen dogma, or the thermodynamic hypothesis [103]), for which the groundwork was laid by Anfinsen and others back in the 1960's [104], is that proteins fold to the

state with lowest free energy. This means that there is a reason to way to identify the native state out of all those astronomically many states. But the Levinthal's paradox still stands - how can the protein find its lowest free energy state? The solution came later in the 1980's with the introduction of energy landscapes [105, 106]. A simplified 2-dimensional energy landscape of a typical protein is shown in **Figure 1-8.a.**

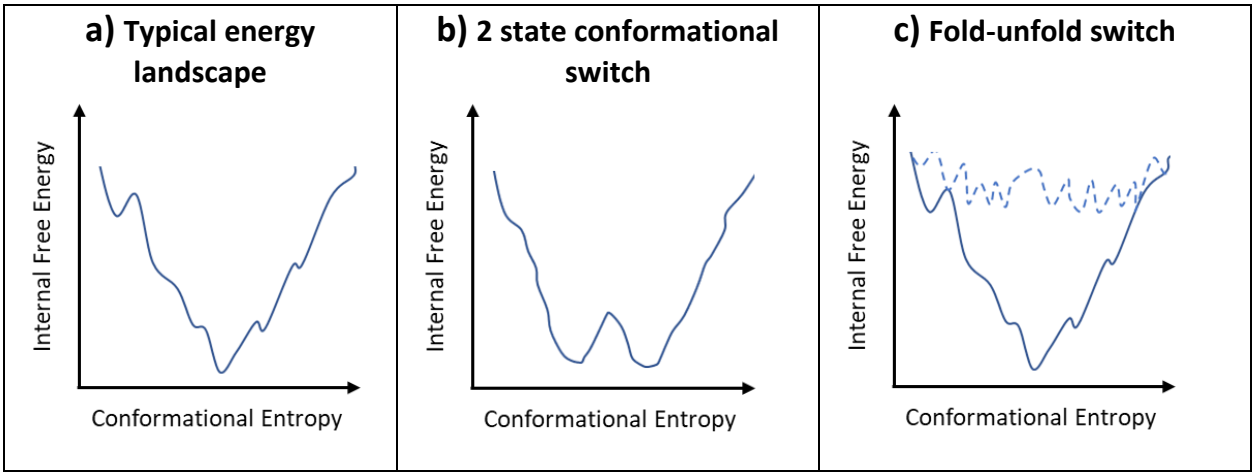


Figure 1-8: Three types of energy landscapes. a) A hypothetical energy landscape of a protein having a smooth funnel shape without too many local minima. A protein having this energy landscape has a single native state at the lowest internal free energy state at the bottom of the funnel. **b)** A hypothetical energy state of a protein acting as a 2-state conformational switch. The protein can switch between the 2 states, corresponding to the 2 low internal free energy states at the bottom of the energy landscape. The two low energy minima are not necessarily present at the same time as shown here. Another low energy minimum can for instance be induced by a ligand or another protein, and in such an interaction the protein will switch to the new local energy minima made by the ligand or protein. **c)** 2 overlaid hypothetical energy landscapes are shown: An energy landscape leading the protein to its native stave (normal line), and an energy landscape without a specific low energy state (dashed line). Many proteins can be induced to switch between the two energy landscapes (and therefore other conformational states), such as by ligand binding or thermal heating.

The vertical axis represents the internal free energy, which is the sum of all free energy terms (ion bonds, torsional energies, hydrogen bonds etc.), except conformational entropy which is represented on the horizontal axis. It is important to note that all the landscapes represented in **Figure 1-8** is a projection of a multidimensional energy surface, since the conformational entropy is a function of all DOF (ϕ, ψ, χ and so on for every residue in the protein).

Folding happens as a diffusion-like process where the conformational entropy is changed by random Brownian forces, with an overall trajectory towards the lowest internal free energy

state (or closely energetically equivalent sub-states [107]) at the bottom of the energy landscape [102]. Along the protein's journey around the energy landscape, it will move around in all directions, but since the probability of existing at a point with a lower internal free energy is larger, than one that is higher, the overall trajectory of the protein will be towards the lowest internal free energy at the bottom of the energy landscape. This solves Levinthal's paradox, because the protein does not need to sample all states, but just a certain number of states, that in total moves it towards its native state.

It is important to note that the energy landscape can be different for different proteins, such as having different levels of roughness (see reference [102] for a visualization of different types of energy landscapes). Roughness is marked by many bumps on the energy landscape surface, and a protein with a rough energy landscape is said to be frustrated [108]. Through evolution, proteins have evolved to smoothen out their energy landscapes to create a funneled energy landscape with minimal frustration, with many conformational states having high internal free energy, and few with low free energy (see **Figure 1-8.a**) [106, 109, 110]. But frustration is also necessary for protein function such as seen in **Figure 1-8.b**. Most functional proteins do not exist in a single state as is the case for proteins in **Figure 1-8.a**, but exerts their function through structural conformational changes between two or more low energy states (**Figure 1-8.b**) [107, 111]. Proteins that change between distinct states, for instance in the presence of a ligand, are called conformational switches, and proteins with such mechanisms are fundamental to many processes in biology including catalysis, signal transduction and PPIs [107, 111, 112].

The energy landscape of a simpler switch, which instead of changing between distinct conformational states, can switch between folding and unfolding is shown in **Figure 1-8.c.** In theory, all proteins are folding-unfolding switches because proteins continually sample all their states (from the Boltzmann distribution), but such proteins are not really considered switches, because the time they spend in the unfolded state, after they have folded, is very minimal. Instead, for a protein to really act as a folding-unfolding switch, the energy landscape has to be significantly changed, from an energy landscape that leads to a single folded state (**Figure 1-8.c** normal line) to an energy landscape that leads to a random unfolded state (**Figure 1-8.c** dashed line). A protein that acts like such a switch is for instance a small zinc protein (see section 1.2.3),

such as the C2H2 zinc finger described earlier. This is a very important point, because it will be one of the main mechanisms of the switch discussed in Chapter 2.

1.4.5 Design of *de novo* proteins: SSD and MSD

Having discussed protein folding it is now time to discuss *de novo* protein design. Before delving into this, it's important to make a distinction between what is called Single State Design (SSD) and Multistate State Design (MSD) [113]. For the SSD methodology, a single sequence is designed for a single target structure, while for the MSD methodology, a single sequence is designed for multiple target structures. The problem of designing proteins, or "the inverse folding problem" [114], as it is sometimes called, can be then be stated as follows: *starting from a single target structure or multiple target structures, how can a sequence be designed that would allow the protein to fold into that structure or structures*. Computationally this is a difficult problem to solve. Again, imagine a protein with 100 residues. The sequence space of such a protein is enormous, because it would have 20^{100} different potential combinations. Like protein folding, computers cannot sample all these states, so efficient methods have been developed to do so. The way to tackle the above search problem, is to guide the search through sequence space in such a way that efficiently leads to the protein having the lowest free energy structure or structures for that sequence. The sequence space is sampled using optimization algorithms (see also Chapter 3), and in SSD this happens under the constraint of an objective function that serves to optimize a sequence on a single structure, while in MSD a fitness function is evaluated that serves to optimize one sequence for an ensemble of structures [113, 115-117]. The objective or fitness functions are based on free energy models, and while most design efforts focus on finding the lowest free energy state or states (Anfinsen's dogma), the energy landscape (such as funnel shapes, or two or more low energy states) is rarely designed explicitly. Although folding simulations closely follows finished designs to evaluate that the structures do indeed fold into their designed state.

Proteins are not solid objects, but dynamic entities better represented by an ensemble of structures with many low minima sub-states [107, 118]. SSD approaches are therefore inherently incorrect, but has circumvented the "incorrectness", largely by incorporating SC minimization

[119] and BB flexibility [120] into the design process. MSD methods can more directly approach the description of many protein sub-states, by incorporating each sub-state into the design. But even though it can outperform SSD in some cases [121], SSD has proven to be the most successful method so far. For instance, all the designs mentioned in section 1.4.3 have used SSD. Challenges of MSD will be mentioned shortly in relation to protein switches which will be discussed next.

1.4.6 Design of *de novo* conformational switches

The design a protein conformational switches are impeded by 3 main obstacles: **1)** The creation of multiple low energy states. **2)** Control of the kinetic barrier between the states. **3)** Incorporation of a switch that can perturb the equilibrium between the states. Furthermore, a protein can be designed to switch between specific states such as in **Figure 1-8.b** or between a folded and unfolded state such as in **Figure 1-8.c**. Protein conformational switches using both approaches have been designed, but these have been mostly done using existing proteins combined with for instance fusion, chemical modification or small scale mutation schemes [122-124], and can therefore not really, according to the perspective of this thesis, be considered *de novo* designs. The design of truly *de novo* conformational switches, particularly switches between distinct states, is a huge challenge in the field and is still largely unsolved [54]. A method that has shown some progress is MSD, and it is suited for this purpose because it can simultaneously design for both states of the switch. Two successful MSD designs of *de novo* conformational switches includes a zinc-finger to coiled coil transition design [125] and a membrane protein transporter that could transport zinc across a membrane [74]. Although successes such as these show that the *de novo* design of conformational switches between many states is possible, successful attempts have been limited. MSD is a young methodology, that has not had the same maturation time as SSD, with inherent energy evaluation challenges [113], increased computational complexity by incorporating more states in the design process, and for conformational switches, difficulties in controlling the energy landscape [126]. On the other hand, folding-unfolding switches are easier targets because first: one can rely on SSD to just stabilize the folded state, and this approach does not have the same challenges as MSD, such as energy evaluation and increased computational complexity. And secondly, because proteins are

naturally in an equilibrium between folded unfolded states, one does not necessarily have to design for the unfolded state, but just a single folded state, while only making sure the folded state is stabilized or destabilized with the presence of a component that can perturb the equilibrium to either the folded or unfolded state (from now on I will refer to such a component, as an inducer). Using small proteins such as zinc proteins discussed previously, the control of stability without a zinc ion present is easier, because intrinsically they do not have the high hydrophobic surface area necessary for stability and are therefore expected to unfold when the zinc ion is not present. However, the difficulty of designing folding-and-unfolding switches is to control the unwanted aggregation that might arise from interaction between many unfolded proteins. These problems are discussed further in chapter 6, when the results of this thesis have been presented.

1.4.7 Design of *de novo* protein assemblies

While the design of conformational switches is still largely an uncharted field, the design of protein assemblies has seen more progress, and a few of the designs have already been mentioned in section 1.4.3. However the major challenge in *de novo* protein assembly design is the design of PPIs [98, 99], and to the authors knowledge, no more than 1 unique interface between subunits has been reported to be *de novo* designed. The higher order assemblies such as the cages, 2D layers, or even the repeat proteins as seen in **Figure 1-7**, which seems to have more than 1 unique interface are designed assuming symmetry (see section 3.2.6). This means that all interfaces are the same and not unique. The success of protein design has been largely helped by applying symmetry, along with intrinsic properties such as avidity [127]. Other strategies that can help in designing PPIs is to use metals [128, 129]. Metal Directed Protein Assembly (MDPA) design has been used in many cases to direct self-assembly such as in simple homodimer design [130] to 1D, 2D and 3D assemblies [131]. Protein assemblies, in particular symmetric assemblies are ubiquitous in nature [132], and there is a need to further understand their properties for basic research, but also to use them for technological applications. One type of protein assembly that has attracted significant attention in both these areas are amyloids, which will be discussed next.

1.4.8 Amyloids - the good and the bad

Misfolding and aggregation of proteins are major factors in many diseases [133]. Amyloids are aggregates of misfolded proteins, and has for many decades attracted significant attention, because of its association with a number of diseases such Alzheimer's, Parkinson and type II diabetes just to mention a few [134]. Through decades of intensive research we are still lacking in our knowledge of how exactly they contribute to disease, and in particular, which of the intermediate states between monomer and full fibril are toxic [135]. Consequently, this means that therapeutic strategies have been limited, and the struggle to find treatments are therefore still ongoing [135].

Besides their pathological role, amyloids have in recent time also been found to have functional roles in bacteria as well as in humans. In bacteria for instance, amyloids have been shown to function in biofilm formation [136], and in humans, amyloids have been found to be associated with memory generation [137] and in melanin production [138]. Furthermore, amyloids have now also become attractive materials in a range of technological application (see below). There are several attributes that makes amyloids attractive. The previously described cross-β structure creates incredibly strong non-covalent interfaces, owing to many factors such as the extensive hydrogen bonding within sheets, and the van der Waals interactions of the SZ between sheets [31, 139]. The many strong interaction interfaces endows the fibers with excellent mechanical properties, reaching for instance the mechanical strength comparable to that of steel, and the mechanical stiffness of spider silk [140]. Furthermore, the fibers also exhibit high thermal and chemical stability [141, 142]. These factors, along with their rapid self-assembly, fibrillar structure, ease of production and low cost, makes them attractive as materials from the nano scale [143-145] to the macro scale [146], and in both living and nonliving settings [147]. Amyloids have for instance been used as scaffolds, and functionalized with simple protein engineering and chemical techniques to create underwater glue [148], nanowires [149] and biocatalyst [150-152]. Amyloids have also been used for stem cell differentiation control, tissue-engineering [153, 154], drug delivery [155], and as biosensors [156] among others.

Although *de novo* protein design is an emerging methodology with a huge potential, and amyloids are attractive building materials for a range of applications, few *de novo* designs of

amyloids have been carried out. The most prominent *de novo* design of an amyloid is a design from the André lab [94]. His team used the crystal structure of an amyloid structure (Sup35 as described previously) and designed a β-α-β protein monomer, for which the 2 β-strands were kept “as is” from the crystal structure. The monomer was shown to assemble into an amyloid under the right experimental conditions.

In the next chapter, I will present the design of a *de novo* amyloid conformational switch, which is the main design goal of this thesis and project.

2 Design description: Aim and Architecture

2.1 Aim

The aim of this thesis is to computationally *de novo* design a reversible protein conformational switch that can be induced to be in an unfolded state without the presence of zinc, and controllably fold into a well folded amyloid structure state in the presence of zinc (**Figure 2-1**).

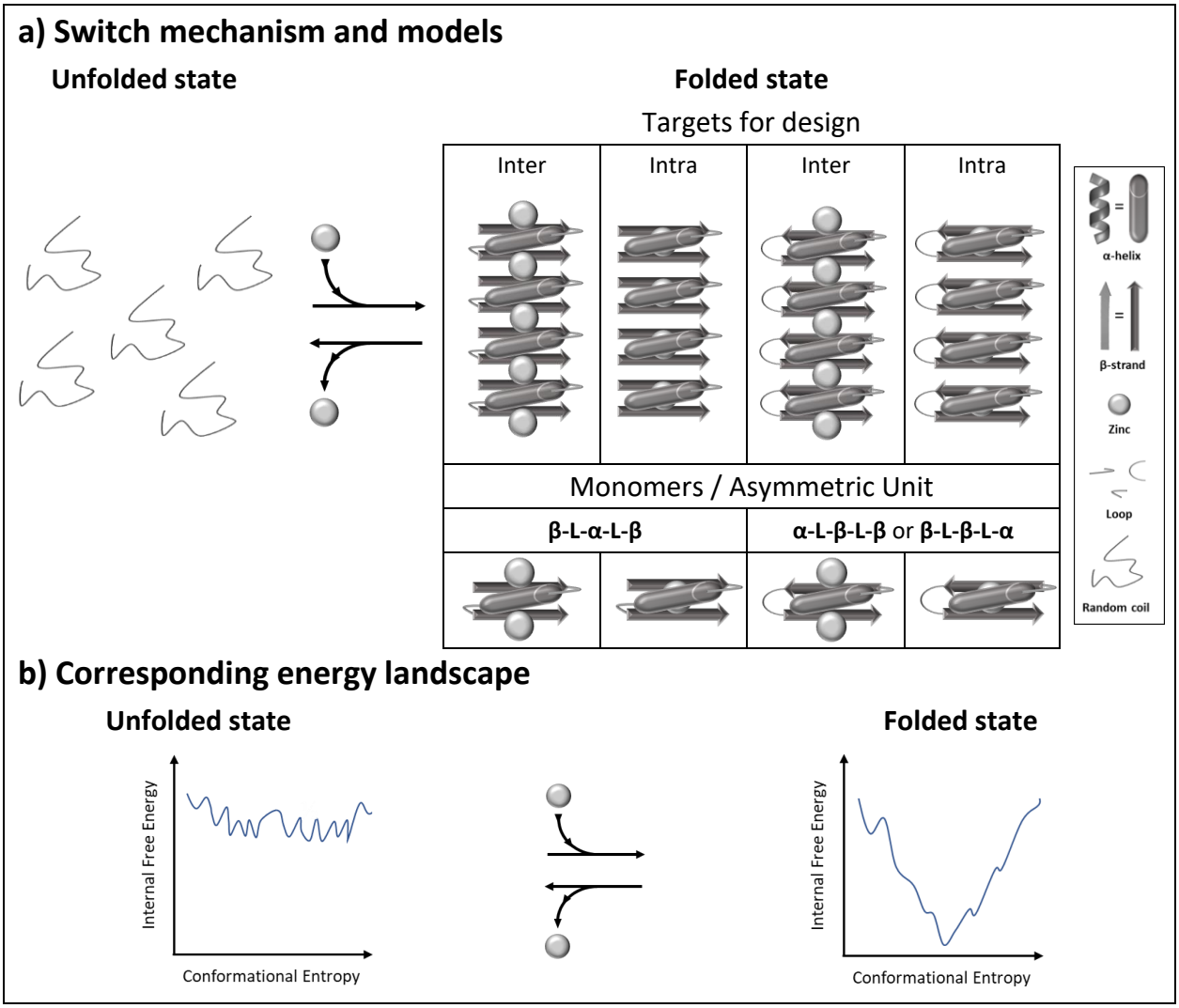


Figure 2-1: Design architecture of the protein switch. a) Right: 6 monomers (β-L-α-L-β, α-L-β-L-β and β-L-β-L-α where β: β-strand, α: α-helix, L: loop), are designed to assemble into an amyloid structure (drawing shows only 1 side), having zinc either between the monomers (inter) or within each monomer (intra). The design is expected to disassemble when zinc is removed (left). The β-L-α-L-β monomers have parallel β-strands while the others have anti-parallel β-strands. **b)** Left: the energy landscape without zinc is expected to not have a stable thermodynamic state, while the energy landscape of the amyloid structure has (right).

2.2 Design Architecture

This is the basic outline of the design architecture. Chapter 6 and 7 will discuss further design, optimization and challenges of the described approach. The Rosetta software suit will be used to design the proteins (see Materials 4.1).

2.2.1 Design approach

The design approach I will pursue uses an SSD approach where only the stable amyloid structure (**Figure 2-1.a** right) along with its interactions with zinc ions will be designed. The unfolded state (**Figure 2-1.a** left) will not be specifically designed, but the protein is expected to transition to this state using strategies as described below (see section 2.2.4). The amyloid structure will consist of multiple copies of a single monomer (**Figure 2-1.a** right), and the other identical monomers in the amyloid structure will be related to each other by symmetry operations (see section 3.2.6). Thus, the design target is only a single monomer in the amyloid structure. Because of the challenges of designing PPIs, it was decided that the design would start from crystal structures deposited in the PDB, similar to an approach used in the André lab [94]. This decreases the number of interfaces to explicitly design because the SZ interface is kept "as is" from the crystal structure. Specifically, the BB and SCs of the SZ interface was decided to be kept as is from the crystal structure. This means that 2 β-strands in one of the sheets, comprising the SZ, is used in the design, but is not redesigned. The only parts to be designed are the 2 loops and the single helix, along with a zinc site.

2.2.2 Choice of secondary structure

The SS of the monomers consist of either β-L-α-L-β, α-L-β-L-β or β-L-β-L-α (β: β-strand, α: α-helix, L: loop), where the involved SCs (protruding into the SZ) and the BB of the β-strands are not explicitly designed. Inspiration for the SS is derived from natural zinc proteins (**Figure 1-4**) and amyloid proteins (**Figure 1-6**). Natural proteins are always a good starting point for designs, because they are proof of principles of folds that work. Because of size limitation of the coordination sphere of zinc, potential designs must include at least 2 β-strands that forms part of the SZ, and the rest of the SS of the monomer should help position zinc with residue ligands in

a tetrahedral arrangement. Only designs with small loops, with known geometries, have been successfully build *de novo* [58, 59], and design of larger loops remain challenging [157]. Furthermore, to the author's knowledge, *de novo* design of loops having binding capabilities to ligands has not been reported. Based on this, it was decided not to have coordinating residue ligands to zinc emanating from loop structures, but instead from either β-sheets or α-helices. Looking at the natural classes of zinc sites (**Figure 1-4**), only the zinc rafts and helical anchors are in that category while the others most often have ligands in loops structures. It was reasoned that the combination of the two could be combined with either 2-3 ligands from a β-sheet and then 1-2 ligands from the helix, and then combine the two SS with loops that would give an overall structure like the C2H2 zinc fingers (**Figure 1-4**). The SS of the monomer would then be α-L-β-L-β or β-L-β-L-α containing 1 zinc site between helix and sheet (monomer to C2H2 zinc fingers). To further probe the "dark universe of proteins", and test other switch mechanisms, it was decided to close the loop of the α-helix yielding a β-L-α-L-β structure (similar to the design from André's lab), and to put the zinc ion inside (intra) one monomer, as before, but also between (inter) two monomers (**Figure 2-1.a** right). The choice of loop length between the different SSs is derived from previously small *de novo* designs [58, 59]. Small monomer targets were used because it would ease design but also because small zinc proteins do not have a stable thermodynamic structure without zinc present as described earlier (see 1.2.3).

2.2.3 Choice of inducer

A ligand was chosen as the inducer, because for application purposes, the switch should be individually addressed (compared to heat or pH switches that affect its surroundings). It should also be feasible, and in that regard, ligand binding designs has already been accomplished [76, 81]. Zinc was chosen as the ligand because of previously success of designs [128, 130, 131], its redox stability, its simple tetrahedral coordination sphere (**Figure 1-3**), which also seems to be heavily preferred, and its abundance in nature and the PDB [34], which makes it possible to gather knowledge based statistical metrics usable in the design process. Furthermore, zinc is also chemically labile and therefore folding can happen under thermodynamic control. Lastly, a metal is also favorable for designs because their bonds are even stronger than hydrogen bonds, and

this might rescue non-optimal designs. To ease design, only mononuclear zinc is used in the design.

2.2.4 Switch mechanism

The amyloid is expected to switch between the two states using the following 2 strategies:

1) **Intermonomer assembly:** A zinc site connecting all the monomers is designed (**Figure 2-1.a** (the two inter designs)). The hypothesis is that the association between the monomers will be driven by zinc. Zinc will further drive the formation of local noncovalent interactions (within and between monomers) to obtain the correct folding and self-assembly of the monomers into the amyloid structure as designed. When zinc is removed (see below), the interaction between monomers will decrease, and so will the forces zinc have had in stabilizing the monomers. If this effect is significant enough, the monomers are expected to disengage from the amyloid structure and become random coils. This strategy is similar to previously MDPA designs (see section 1.4.7).

2) **Intramonomer assembly:** A zinc site within each monomer will be designed (**Figure 2-1.a** right (intra)). The mechanism is expected to be slightly different from the mechanism described above. The hypothesis is that the stabilization of the individual monomer, mainly by zinc binding, will position the BB and SCs of the monomer, to facilitate the association of the monomers into the amyloid structure. When zinc is removed (see below) the stability of the individual monomer along with nonlocal interactions between monomers will decrease, and if significant enough the monomers are expected to disengage from the amyloid structure and become random coils.

In both cases the equilibrium will be shifted to the folded state by adding zinc, and to the unfolded state by chelators such as EDTA, that removes zinc from the structures (**Figure 2-1.a**). The energy landscapes of the structures are expected to look like the ones in **Figure 2-1.b**. For the unfolded state, the energy landscape should not contain any stable thermodynamic states with low conformation entropy. The folded state on the other hand should contain a stable thermodynamic state, with a corresponding low conformational entropy. The marginal stability

of small zinc proteins without zinc, and zincs role in folding (see section 1.2.3) is as indicated, thought to play huge role in the switch mechanism.

3 Computational Techniques

3.1 Score functions

The total energy ($\Delta \mathrm{E}_{total}$) of the score function used in Rosetta is a weighted linear combination of individual terms as given in **(eq 3-1)**:

$$\Delta \mathrm{E}_{total} = \sum_i w_i E_i(\Theta_i, aa_i) \qquad \textbf{(eq 3-1)},$$

where w_i is the weight for the i'th term, and E_i the energy of individual terms as a function of geometric degrees of freedom (Θ) and amino acid identities (aa). The terms are either knowledge based or physical based. Knowledge based score terms are based on statistics from the PDB. Such score terms could for instance be terms that involves optimal rotamers (see section 3.2.3). Physical based score terms are based on physical models. The electrostatic or van der Waals forces are often implemented in this way.

The score function in Rosetta is flexible meaning that terms can be added, or reweighted, and different terms are often used in different circumstances. Rosetta's score function classes can be divided into two general classes. The low-resolution class uses a reduced representation of SCs called centroids that replaces everything beyond the CB atoms into a single super atom located at the residue center of mass (from CB and beyond). Furthermore, the terms in this class are very simple. For more information on the low resolution class terms and centroids see reference [158]. The high-resolution class use all atoms of the protein in the calculations, and the terms can be more complicated. For more information on all the most up to date terms see reference [159].

3.2 Protein modelling

3.2.1 Monte Carlo with simulated annealing

One approach to solving an optimization problem is using Monte Carlo (MC) methods. Broadly speaking, MC methods rely on stochastic or random moves to optimize a given problem. For each random move, the move is either accepted or rejected, and the algorithm stops at some point

when the problem has been sufficiently optimized. To accept or reject moves, such as the insertion of an amino acid identity or changing a rotamer (see section 3.2.2), Rosetta uses the Metropolis Criterion [160]. If a move leads to a lower total energy (see section 3.1), the move is always accepted, and if it does not the probability (p) of accepting the move is given by **(eq 3-2):**

$$\mathrm{p} = e^{-\frac{\Delta \mathrm{E}}{k_B T}} \qquad \textbf{(eq 3-2)},$$

Where $\Delta \mathrm{E}$ is the difference in energy between the previous move and the current move, $k_B$is the Boltzmann factor and T the temperature. In Rosetta, MC is often accompanied by a simulated annealing (SA) approach. In MC with SA the temperature factor (T) as above is high in the beginning of the search, while it is lowered as the search goes on. Several ramping steps can be applied if needed.

3.2.2 Gradient descent minimization

Another approach to solving an optimization problem is using gradient descent minimization. Given a function, such as an energy function, gradient descent minimization algorithms use the gradient at the current point on the function to find the minimum of the function. Using an energy function (which is the case in Rosetta), the gradient descent methods can be described with the following equation:

$$x_{n+1} = x_n - \varepsilon \nabla E \qquad \textbf{(eq 3-3)}$$

Where x_n is the current position on the energy surface and x_{n+1} is the new position. To go to the x_{n+1} position, a step is taken ($\varepsilon \nabla E$) along the negative slope of the gradient (∇E) with a constant factor ε. ∇E is a vector of partial derivates of all the energetic terms and the energetic terms are a function of the movable DOF (see also section 3.1). This could for instance be the torsion angles of the BB. Rosetta uses different flavors of the general approach outlined above such as the Broyden–Fletcher–Goldfarb–Shanno method.

Minimization combined with MC sampling is a very effective strategy for sampling proteins structures [161] (for both design and structure prediction), and these are the main sampling methods used in Rosetta and this thesis.

3.2.3 Rotamers and rotamer libraries

Rotamers are conformational isomers of rotations around sp3-sp3 hybridized bonds. Amino acids have certain rotamers that differs by a specific set of optimal χ torsion angles. In the case of Met for instance (which only has sp3-sp3 hybridized bonds) 3 specific set of optimal values of its χ torsion angles, or rotamers, can be seen in the χ probability distributions. Non-rotameric χ-values (χ-values not having close to discrete states) exist in sp3-sp2 hybridized bonds. These can for instance be seen for Asn or Glu. Rotamer libraries consist of specific sets of optimal χ-values for all amino acids (can include rotamers and non-rotamer DOFs). This library can then be used in protein modelling to only search for optimal χ-values and can be quickly inserted in a protein by for instance MC methods (see 3.2.1). The optimal χ-values are dependent on ϕ and ψ values of the given residue, and rotamer libraries called BB dependent rotamer libraries include these as another variable. Rosetta uses the Dunbrack BB dependent rotamer library [162].

3.2.4 Cyclic Coordinate Descent (CCD)

Cyclic Coordinate Descent (CCD) is a loop modelling algorithm used to close loops in protein structures. It is originally described in reference [163]. Starting from a loop break, 3x2 superimposable BB atoms (N, C and CA) can be defined from both end of the loops for which one wishes to close (the last and first residue in the break). One end is kept fixed and the fixed 3 atoms are defined as F1, F2 and F3. The other end is movable, and the 3 movable atoms are defined as M1, M2, M3. To close the loop, the 3x2 atoms must be superimposed. The sum of the squared distance (S) between F1 and M1 ($F1M1$), F2 and M2 ($F2M2$) and F3 and M3 ($F1M1$) can be described as in **(eq 3-4)**:

$$S = |F1M1|^2 + |F2M2|^2 + |F3M3|^2 \qquad \textbf{(eq 3-4)},$$

The BB torsion angles (ϕ and ψ) of any residue preceding the movable atoms that minimizes S in **(eq 3-4)** can be analytically found. The idea of the CCD algorithm is then to change each BB torsion angle one at a time iteratively until S is minimized as much as possible.

Rosetta's implementation of the traditional CCD algorithm involves several modifications. The protocol starts with a random break somewhere in the loop. This is then follow by an iterative

approach of: 1) Either MC fragment insertion or small MC BB ϕ and ψ changes. Fragment insertion are insertion of a set of ϕ and ψ values that are derived from real proteins structures (such as from the PDB) [164, 165]. 2) CCD as originally implemented. 3) Gradient descent minimization (see section 3.2.2). See reference [166] for more information on Rosetta's implementation of the CCD algorithm.

3.2.5 Rosetta Match

The Rosetta Match algorithm was originally conceived for the use in enzyme design as is described in reference [167] . Its modifications in Rosetta 3.x can be found in reference [168].

Given a protein scaffold, a ligand one wishes to design into that scaffold, and 6 coordination parameters (CPs) that holds information on how specific amino acids bind to the ligand, the Rosetta Match builds that ligand into the scaffold with the residues coordinating as described by the 6 CPs. The way the algorithm does this is by first placing each defined residue and the ligand together according to the 6 CPs, at defined positions in the scaffold. Different rotamers of the residue along with sampling around the specified 6 CPs is carried out at each position. Each clash-free placement of the ligand and residue is called a hit. For each hit the 6 CPs are binned. If hits for all the different residues, defined to bind to the ligand, places the ligand in the same bin (i.e. with the same coordinates) it is called a match. The algorithm will output all matches found, that is all the residues that are a part of the match, along with the ligand and the scaffold.

3.2.6 Symmetry

The description of symmetry as applied in Rosetta3.x can be found in reference [169]. Symmetry is used to ease design of larger symmetrical protein systems. The way this is done is by applying sampling and scoring for only one of the asymmetric units of the symmetrical system. This subunit is called the master. When sampling is applied to the master subunit, such as MC moves or minimization, the changes are replicated to all other of the identical subunits (called slaves) of the symmetric system. Specifically, this is done by defining reference frames for each asymmetric unit, through coordinate systems such as seen in **Figure 3-1.a**, where the changes to the master in one reference frame is replicated to the other reference frames. In the setup shown in **Figure**

3-1.a there are 13 coordinates systems, were 6 of them defines the references frames for each of the 6 subunits (A-F), and the 7 others define reference frames for those coordinate systems. This is also the setup used in high resolution optimization of the structures (see section 5.5). Scoring is simplified because it is only done within the master subunit and for its interactions with the other slaves. For instance, consider the system shown in **Figure 3-1.b** (same as in **Figure 3-1.a**). There are 4 unique interfaces with respect to the master subunit and these can describe the whole system. The energy calculation for this system is then simplified to:

$$\Delta \mathrm{E}_{total} = 6\mathrm{E}_{C} + 4\mathrm{E}_{C:A} + 3\mathrm{E}_{C:D} + 2\mathrm{E}_{C:B} + 2\mathrm{E}_{C:E} \qquad \textbf{(eq 3-5)},$$

where E_{C} is the master subunit and ":" refers to an interface. For instance, the interface $C:A$ is the interface between the C and A subunit.

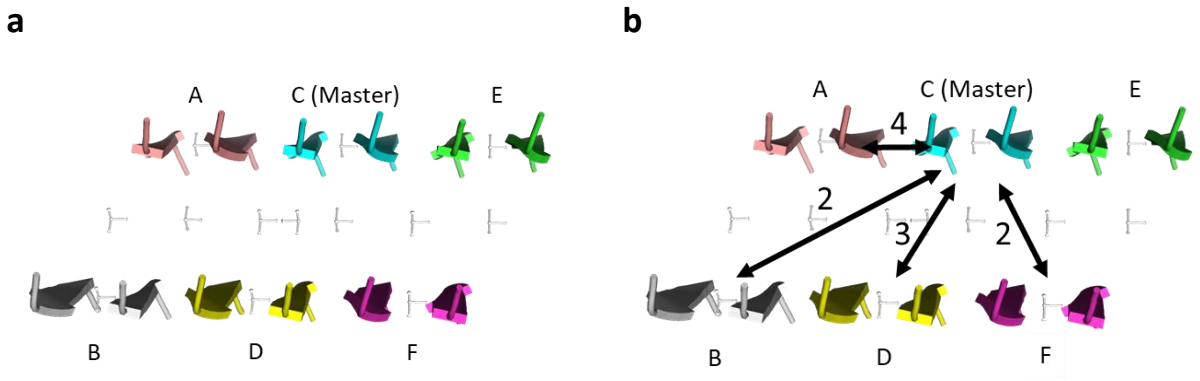


Figure 3-1: How symmetry works. a) An example of a symmetric setup (also the symmetric setup used in this project). 6 identical monomers describe the whole system (shown in different colors) and are named A-F. C (cyan) is the master subunit and the others are the slaves. Each subunit is placed in a coordinate system (used for sampling and scoring) and other coordinate systems are used to place these coordinate systems. All sampling done to the master is replicated to the other subunits. **b)** Scoring is simplified with symmetry. There are only 4 unique interfaces in the system, because many of the interfaces are symmetrically identical. For instance, the interface between C and A is the same interface as between C and E. Each unique interface of the master subunit (shown with a black arrow) is multiplied with the number next to it, which corresponds to the amount of symmetrically identical interfaces of the system.

4 Materials and Methods

This section includes the Materials and Methods used to produce the results shown in Chapter 5. The reader should read Chapter 5 before reading this chapter. The methods as described here assumes the results have been presented.

4.1 Materials

Rosetta [170], PyRosetta [171] and RosettaScripts [172] were used in this project. Two version of Rosetta were used. A local (non-mpi) Rosetta version (17.36.59679) was installed on a computer and an mpi version (17.52.59948) was installed on a computer cluster. Scripts seen in Appendix C using the static version of Rosetta refers to the first version, while scripts using the mpi version refers to the latter. The computer cluster consist of several supercomputers with the following specs: CPU: Intel Xeon E5-2699 v3, Processors: 72, clk (turbo): 2.30 GHz (3.60 GHz) and RAM 384 Gb. The local computer has the following processor: CPU: Intel Core i3-6100 CPU @ 2.30 GHZ and RAM 8.00 GB.

PyRosetta4 for python 2.7 was installed (release 154), and RosettaScripts are implemented in the both of the Rosetta versions. The following python 2.7 modules were also used: NymPy (version 1.13.3), SciPy (version 0.17.0) and Matplotlib (version 2.1.1). For generating most of the figures in this thesis, and for carrying out other tasks, open source PyMOL version 1.8.4.0 was used.

4.2 Methods

4.2.1 Amyloid data mining

The PDB (www.rcsb.org) was searched for PDB entries containing the words “Amyloids”, “Steric Zipper” and “Protein Fibril” (Dec. 2017). For each of the searched words, the PDB entries were manually inspected, either from the website display image or by looking at the PDB file with PyMOL, to identify a SZ structure. Those structures that were identified to contain a SZ, and having a resolution of higher than 2.5 Å, were chosen as potential candidates for design.

Resolution was checked on either the PDB database, the EMDataBank (www.emdatabank.org) or in the original article of the corresponding PDB entry. The resolution for the amyloid structures ranged between 0.85 Å (PDB ID: 2OL9) to 2.3 Å (PDB ID: 2Y29), and the resolution cut-off was not used in the end to filter structures out. The list does not include all deposited amyloid structures at the time (for instance 5K2G were missed on later inspection). Some of the PDB entries is the same structure but with higher resolution obtained from new experiments (for instance PDB entries: 1YJO, 1YJP, 2OMM and the respective PDB entries: 5K2E, 5K2F, 5K2H)[173]. Only structures based on X-ray or electron crystallography were considered and not NMR structures (see for instance 2KIB). This was done to avoid dealing with ensemble structures. A full list of PDB entries used in this study along with their symmetry classification (see Methods 4.2.2) are given in Appendix D.1.

4.2.2 Symmetry classification

The PDB entries selected from Methods 4.2.1 were manually inspected using PyMOL to classify the amyloid structures according to the classification system outlined in the Introduction (see section 1.3. It is based on the classification system by Eisenberg et al. [35]). In PyMOL, symmetry mates were generated, and using either the cartoon representation or the N- and C- terminal as a guide, two interacting sheets connected with a visible SZ were used as the basis of the classification. Some of the structures were already classified by Eisenberg et al. (from reference [31]), and the previously classified structures were not classified anew.

4.2.3 Scaffold selection criteria

Table 4-1 list the 10 criteria that were used to select the 2 β-strand scaffolds, and they are explained in further depth below.

Criteria 1) and 2) are explained in section 5.2. Criterion 3) and 4) stems from the fact that an amyloid fiber is designed were only 2 β-sheets should associate with each other through the SZ. No adjacent sheets or ligands are modelled in the design, so removing them might render the SZ interface, that is kept fixed from the crystal structure, non-ideal to design from. This is because the environment most likely influences the association of the SZ and the positions of the different

atoms involved in the SZ. Being able to remove ligands and adjacent sheets also renders the modelling procedure easier because these factors are then not considered.

Table 4-1: Criteria used to select appropriate 2 β-strands scaffolds from crystal structures.

1	β-L-α-L-β monomers should be derived from parallel β-strands, and α-L-β-L-β or β-L-β-L monomers from antiparallel β-strands.	**2**	The 2 β-sheets of the SZ interface should be related to each other by a 2-fold rotation around the fibril axis.
3	The SZ interface should not significantly span more than and 2 sheets.	**4**	No ligands such as metals or water should be present in SZ interface or be integral to the association of the SZ.
5	No hydrogen bonds should be unsatisfied between β-strands.	**5**	No glycine should be present in the β-strands.
7	No net charges should exist in the SZ interface	**7**	The SZ interface should not be V-shaped
9	At least 5 residues should be present in the SZ interface.	**10**	Every residue and rotamer should be uniquely assigned in the SZ interface

Criterion 5-9 are based on energy considerations. β-sheets with unfulfilled hydrogen bonds are less stable than β-sheets having all their hydrogen bonds fulfilled (considering that they are of the same length). The same is the case for β-sheets with proline or glycine. These residues have been shown to be destabilizing in β-sheets [174, 175]. Net charges in the SZ interface can potentially inhibit amyloid formation and were therefore removed [176, 177]. V-shaped SZ were also discarded because these did not have enough interaction between the sheets of the SZ. V shaped SZ refers to the direction of the two sheets of the SZ, where one end of the β-sheets is significantly closer to each other than the other end.

10) For ease of design it was chosen not to consider structures with many non-uniquely defined rotamers in the SZ.

To have an even amount of α-L-β-L-β or β-L-β-L-α and β-L-α-L-β monomers, the latter monomers were chosen based on self-association of the protofilaments. Self-association is based on how much the two sheets of the protofilament interacts alone, that is, the most self-associated sheets are the ones with not too many peptides or ligands around them, with the largest SZ interface.

4.2.4 Cutting out scaffolds from crystal structures

PDB entries were opened with PyMOL and symmetry mates were generated. A SZ interface was then identified and 2 β-strands were selected and saved as a separate file. For generating the symmetry files described later a larger part of the steric zipper interface was saved. 6 strands on each side of the SZ (12 in total) was saved.

4.2.5 Zinc site data mining

The full dataset consisted of 2945 protein crystal structures containing 1 or more zinc sites. The PDB files for each crystal structure were obtained by a query search in the PDB (www.rcsb.org), specifying that the PDB entries should contain zinc and have a resolution of higher than 2.5 Å. The subset of the full dataset containing only mononuclear tetrahedral sites was generated by cross refencing the PDB entries of the full dataset with PDB IDs designated as mononuclear tetrahedral sites by the MetalPDB (metalweb). PDB entries containing mononuclear tetrahedral sites were identified by using the advanced search option and specifying mononuclear site features and tetrahedral geometry. The subset contained 412 crystal structures. The full list of PDB files for both datasets is given in Appendix D.2.

4.2.6 Zinc site analysis

For each zinc site in either the full dataset or the subset, only coordinating atoms in a distance of 3.0 Å to zinc were used in the analysis. Furthermore, the binding mode (BM) of Asp or Glu were determined as bidentate if the distance of the two oxygen groups did not differ by more than 0.5 Å. If the difference was larger, the closest coordinating atom was accepted as the coordinating atom. Similarly, for the other BMs the closest atom to zinc was accepted as the coordinating atom. The anti-BM of Asp and Glu (the BM of which were not used in the analysis) were identified by measuring the torsion angle between either ZN-OD1-CG-CB or ZN-OD2-CG-CB (for Asp) or ZN-OE1-CD-CG or ZN-OE2-CD-CG (for Glu). If these values were equal to or between -90° and 90° they were identified as an anti-BM and were therefore not used in the analysis.

For the full dataset, all residues at a distance of 3.0 Å were considered but only data for His, Cys, Asp and Glu were used. For the subset containing only mononuclear tetrahedral sites,

only designated coordinating atoms from the MetalPDB (metalweb) were considered.

Two scripts were developed to carry out the analysis. The uses of them are described in Appendix C.1. The scripts carry out the data acquisition as well as the generation of the density plots and fitting, which is discussed in the next section below.

4.2.7 Generation of the density plots and fitting

The data acquired from carrying out the steps in Methods 4.2.5 and 4.2.6 were analyzed. First, the absolute values of negative bond angles and torsion angles were calculated and then added by 180°, so that all angle data points were measured from 0-360°. The coordination parameters (CP) were then divided into 50 bins. The height (h_j) of each bin (j) was set and normalized according to **(eq 4-1)**.

$$h_j = \frac{y_j}{n \cdot b} \qquad \textbf{(eq 4-1)},$$

where n is the sum of all data points for that CP, y_j is the amount of data points in the j bin, and b is the bin width which were set by the histogram method of the NumPy module (see Materials 4.1).

All 81 parametric probability distributions (pdf) available in the SciPy module (see Materials 4.1) were fitted to each generated density histogram using each pdf's fit method. The fit method uses maximum likelihood estimates (MLE) to fit the data.

Each fit was evaluated based on the least sum of squared errors (SSE). The SSE was calculated as in **(eq 4-2)**.

$$\mathrm{SSE} = \sum_{j}^{50} (h_j - pdf_j)^2 \qquad \textbf{(eq 4-2)},$$

where h_j is the same as before, and pdf_j is the height of that pdf at the center of the bin. The best pdfs were the ones having the lowest SSE scores. Several different pdfs could fit the data well. Normal pdfs seemed to fit all data well, except for the D CP of His, Glu and Asp. Normal distributions **(eq 4-4)** were then fitted to those and the Johnson SU pdf **(eq 4-5)** were fitted to

the D CP of the His, Asp, and Glu. To $scale$ and move (loc) the 2 pdfs, the input variable x is modified to $\boldsymbol{x}$ as in **(eq 4-3).**

$$\boldsymbol{x} = \frac{x - loc}{scale} \qquad \textbf{(eq 4-3)}.$$

The 2 pdfs can then be given as the following:

$$\text{Normal pdf} = \frac{\exp(-\boldsymbol{x}^2 / 2)}{scale \cdot \sqrt{2\pi}} \qquad \textbf{(eq 4-4)},$$

$$\text{Johnson SU pdf} = \frac{\text{b}}{scale \cdot \sqrt{\boldsymbol{x}^2 - 1}} \Phi(a + b \cdot \log(x + \sqrt{\boldsymbol{x}^2 + 1}) \qquad \textbf{(eq 4-5)}.$$

The fitted data yielded for the Normal pdf the $scale$ and loc parameters, while for the Johnson SU pdf this yielded, $scale$, loc, a and b.

The mean (μ) of the CP distributions were calculated as in **(eq 4-6).**

$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \textbf{(eq 4-6)},$$

where x_i is the i'th value of the measured CP and n is the total amount of measurements of that CP. The standard deviation is then calculated as in **(eq 4-7).**

$$\sigma = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n - 1} \qquad \textbf{(eq 4-7)}.$$

As the pdfs were numerically expressed in a list of 100.000 datapoints, the maximum or the optimal value (Max) of the CP was found by finding the maximum value in that list.

4.2.8 Sampling of all residues, BMs, CPs and rotamers around a tetrahedral zinc atom

Each residue was first positioned at the two vertices of the tetrahedral coordination sphere according to the optimal values of the corresponding CP and BM. 3 values for each of the non-free CPs were then chosen to be sampled. The values were the optimal value ± 1 standard deviation. This yielded 3^4 discrete states for the ZN:SG BM which has 2 free CPs and 3^5 states for the other BMs. The 2 free CPs for the ZN:SG BM were then sampled at 0°, 90°, 180° and 270°,

yielding $3^4 \cdot 4^2$ states. The 1 free CP for the other BMs were sampled at 0°, 60°,120°, 180°,240° and 300° yielding $3^5 \cdot 4^2$ states. For each state all the rotamers of the corresponding residues were build. For all states and rotamers the CB-CB distances were measured. Appendix C.2 shows the scripts and their uses for carrying out this analysis as well as how many CB atoms were defined for each BM.

4.2.9 Step 1: Generation of BB structures

Starting from a 2 β-strand scaffold. A script is used to mutate the non-SZ interface and build a loop and an α-helix. The script to do this using the 2α3β2β-2y29 model, and all the following scripts described in this section and their uses can be found in Appendix C.4. The loop and α-helix are built without optimization (because it is just a simple extension) and the α-helix torsion angles are set to ϕ=-57° and ψ=-47°. Another script was used to generate the loops with the CCD algorithm.

Now that the entire structure has been generated, an algorithm was developed to sample it. See section 3.2.6 and Appendix C.7 for how the symmetry was set up. The loop connecting the helix to the 2 β-strand scaffold is sampled as described. Then the BB structures are subjugated to filters which are described below.

Filter 1 calculates the steric repulsion between atom pairs using Rosetta's van der Waals (vdw) score term. It is non-zero only in the case when 1) the interatomic distance between the pairs are less than the sum of the atoms van der Waals radii, and when 2) the interatomic distance does not depend on the torsion angles of a single residue (see also reference [158, 178]). In the case were the vdw terms is non-zero, the structures were filtered out. Importantly, alanine centroids were used, except in the SZ, where the centroids of the SZ residues were used.

To apply the 3 other filters a plane, geometric center and a helix vector must be defined. The geometric center ($\overrightarrow{GM}$) is calculated using the coordinates of atoms $\overrightarrow{a_i}$ of interest, and the total number of atoms n, as seen in **(eq 4-8)**:

$$\overrightarrow{GM} = \frac{\sum_{i=1}^{n} \overrightarrow{a_i}}{n} \qquad \textbf{(eq 4-8)}.$$

For the α-helix all the CA atoms of the α-helix are used.

To create the helix vector (for only 1 of the helices), inspiration was taken from Kahn et al. 1998 [179]. Because the helix is ideal, the calculation of the helix vector is easier. The positions of 4 CA atoms from the first 4 residues in the helix are calculated (CA1, CA2, CA3, CA4). Then A1, B1, A2, B2 can be defined as in **(eqs 4-9)**:

$$\overrightarrow{A1} = \overrightarrow{CA1} - \overrightarrow{CA2}$$
$$\overrightarrow{B1} = \overrightarrow{CA3} - \overrightarrow{CA2}$$
$$\overrightarrow{A2} = \overrightarrow{CA2} - \overrightarrow{CA3} \qquad \textbf{(eqs 4-9)},$$
$$\overrightarrow{B2} = \overrightarrow{CA4} - \overrightarrow{CA3}$$

Then 2 normalized vectors ($\overrightarrow{V1}$ and $\overrightarrow{V2}$) can be defined as in **(eqs 4-10)**:

$$V1 = \frac{\overrightarrow{A1} + \overrightarrow{B1}}{|\overrightarrow{A1} + \overrightarrow{B1}|}$$
$$\overrightarrow{V2} = \frac{\overrightarrow{A2} + \overrightarrow{B2}}{|\overrightarrow{A2} + \overrightarrow{B2}|} \qquad \textbf{(eqs 4-10)},$$

Finally, the helix vector ($\overrightarrow{HV}$) can be defined as in **(eq 4-11)**:

$$\overrightarrow{HV} = \frac{\overrightarrow{V1} \times \overrightarrow{V2}}{|\overrightarrow{V1} \times \overrightarrow{V1}|} \qquad \textbf{(eq 4-11)},$$

For the plane, 3 points ($\vec{a}$, $\vec{b}$ and $\vec{c}$) on the amyloid β-sheet are defined. This is done using the CB atoms. In the case of the models used in the results, these would be the CB atom of residue 26 for point a, CB atom of residue 24 for point b and residue 19 (of its symmetry mate) for point c. A coordinate system (with an x, y, and z component) with respect to this plane is then defined as:

$$\vec{x}_{plane} = \frac{\vec{b} - \vec{a}}{|\vec{b} - \vec{a}|} \qquad \textbf{(eqs 4-12)},$$

$$\vec{y}_{plane} = \frac{\vec{c} - \vec{a}}{|\vec{b} - \vec{a}|}$$

$$\vec{z}_{plane} = \frac{\vec{x}_{plane} \text{ x } \vec{y}_{plane}}{|\vec{x}_{plane} \text{ x } \vec{y}_{plane}|}$$

Filter 2 then filters based on the angle (θ_{F2}) between the $\overrightarrow{HV}$ and the $\vec{z}_{plane}$ vector. θ_{F2} is given as in **(eq 4-13)**:

$$\theta_{F2} = 90 - \cos^{-1}\left(\frac{\vec{z}_{plane} \cdot \overrightarrow{HV}}{\left|\vec{z}_{plane}\right|\left|\overrightarrow{HV}\right|}\right) \quad \textbf{(eq 4-13)},$$

If θ_{F2} is not between -30 to 30 degrees, the BB structures are filtered out.

Filter 3 filters based on the angle θ_{F3} between the $\overrightarrow{HV}$ vector and $\vec{y}_{plane}$ vector.

$$\theta_{F3} = 90 - \cos^{-1}\left(\frac{\vec{y}_{plane} \cdot \overrightarrow{HV}}{\left|\vec{y}_{plane}\right|\left|\overrightarrow{HV}\right|}\right) \quad \textbf{(eq 4-14)},$$

If θ_{F3} is not between -30 to 30 degrees, the BB structures are filtered out.

Filter 4 filters based upon the distance between the geometric center of the α-helix and the closest point on the amyloid plane. This distance (D_{F4}) is the same as a projection of a vector ($\overrightarrow{V_J}$) from a point in the plane to $\overrightarrow{GM}$ onto the $\vec{z}_{plane}$vector. Point a is chosen as this point in the plane and $\overrightarrow{V_J}$ is then given as in **(eq 4-15)**:

$$\overrightarrow{V_J} = \vec{a} - \overrightarrow{GM} \quad \textbf{(eq 4-15)},$$

The projection (P_j) is then given as in **(eq 4-16)**:

$$\overrightarrow{P_J} = \frac{\overrightarrow{V_J} \cdot \vec{z}_{plane}}{|\vec{z}_{plane}|^2} \vec{z}_{plane} \quad \textbf{(eq 4-16)},$$

$D_{F4}$is then given as in **(eq 4-17)**:

$$D_{F4} = |\overrightarrow{P_J}| \quad \textbf{(eq 4-17)},$$

If this distance is not between 3.15-11 Å the BB structures are rejected.

To only search certain positions in Rosetta Match, a position file is outputted containing

all residues of the non-SZ interface, non-loop residues and residues of the α-helices having their CB atom below the plane of the two helices. The helix plane of the two helices are defined with the x-component ($\vec{x}_{helix}$) of the $\overrightarrow{HV}$ vector, and the y-component ($\vec{y}_{helix}$) by the difference between the 2 $\overrightarrow{GM}$ vectors $\overrightarrow{GM}(\text{helix1})$ and $\overrightarrow{GM}(\text{helix2})$ (remember there are 2 helices in the symmetric system for this algorithm). Importantly, the $\vec{z}_{helix}$ must be defined so that it points away from the amyloid plane. The equations are defined in **(eqs 4-18):**

$$\vec{x}_{helix} = \overrightarrow{HV}$$

$$\vec{y}_{helix} = \frac{\overrightarrow{GM}(\text{helix2}) - \overrightarrow{GM}(\text{helix1})}{|\overrightarrow{GM}(\text{helix2}) - \overrightarrow{GM}(\text{helix1})|} \qquad \textbf{(eqs 4-18)},$$

$$\vec{z}_{helix} = \frac{\vec{x}_{helix} \text{ x } \vec{y}_{helix}}{|\vec{x}_{helix} \text{ x } \vec{y}_{helix}|}$$

Then for each residues CB atom in the 2 helices a vector (V_{CB}) between CB ($\overrightarrow{CB}$) and a point in the helix plane, chosen as one of the helices $\overrightarrow{GM}$, is calculated. This is shown in **(eq 4-19)**:

$$\overrightarrow{V_{CB}} = \overrightarrow{CB} - \overrightarrow{GM} \qquad \textbf{(eq 4-19)},$$

Only residues with a θ_{pos} angle (as defined in **(eq 4-20)**) less than 0 degrees are outputted in the position file. This is why the direction of the $\vec{z}_{helix}$ is important.

$$\theta_{pos} = 90 - \cos^{-1}\left(\frac{\vec{z}_{helix} \cdot \overrightarrow{V_{CB}}}{|\vec{z}_{helix}||\overrightarrow{V_{CB}}|}\right) \qquad \textbf{(eq 4-20)}.$$

A local computer (see Materials 4.1) were used to generate around 400 structures for the 2α3β2β-2y29 model. This took about 2 minutes

4.2.10 Generation of the Ramachandran plots and selected grid points

The 70% homology (standard) Top8000 database was downloaded from the website: kinemage.biochem.duke.edu. This dataset represents about 8000 high quality filtered protein chains used in the MolProbity software [180]. Every ψ and ϕ torsion angle was then calculated

from the dataset, except in the following cases: 1) the first and last residue of the chains was not counted. 2) pre-Pro residues (residues preceding Pro), Gly, Pro, Ile, Val and residues designated as UNK were not counted either. This yielded $>10^6$ datapoints. 5x5 grids were made starting from (ψ, ϕ) = (-180, -180), (-180, -175), (-175, -180), (-175, -175). Midpoints of the grids were, in the case of the first grid (ψ ϕ) = (-177.5, -177.5). The amount of data points in each grid were then calculated and the grids with highest number of data points were selected. 30 grids were selected for the A and B ABEGO bins, while 10 were selected for the E and G bins. See also **Figure 5-8.a**. Appendix C.3 shows how the scripts developed for this purpose works.

4.2.11 Symmetry setups

A script was developed to make a symmetry file for each amyloid structure. The script and its use are shown in Appendix C.7. Two different kinds of symmetry files were used. For the generation of BB structures in step 1, 5 coordinate systems and 2 monomers were used. For step 3, 13 coordinate systems and 13 monomers were used. A visual example of the coordinate systems used in step 3 can be seen in Appendix C.7. See also section 3.2.6.

4.2.12 Step 2: Design of a zinc site with Rosetta Match

The parametrization of zinc is specified by a params file and is shown Appendix C.8. Commands and files used to generate matches in the 2α3β2β-2y29 model by the Rosetta Match algorithm are shown in Appendix C.5.

Each BM was sampled in the same way as the analysis done earlier to calculate distances between CB atoms (see Methods 4.2.8 and section 5.4), but the rotamer sampling was handled internally by the Rosetta Match algorithm. The construction of the T_A, T_B and T_{AB} CPs replaces the L_1 and L_2 atoms with the virtual atoms of V1, V2, V3 and V4. The combination of virtual atoms was chosen wisely because the construction as outlined in section 5.5.2 will place the coordinating residues to zinc at each vertex of the tetrahedral coordination sphere.

To generate matches for the 2α3β2β-2y29 model, the scripts shown in Appendix C.5. were run on 1 of the supercomputers (see Materials 4.1) using 70 cores. This took about 2 days. The Rosetta Match algorithm was aborted before too many structures were outputted. A subset of the outputted files was then manually inspected, and a single structure was chosen for the

downstream design of the intermonomer. Similarly, one structure was chosen as the intramonomer.

4.2.13 Step 3: Zinc site optimization and sequence design

The two structures outputted from Rosetta Match (input was the 2α3β2β-2y29 model) were turned into the representatives of the asymmetric unit. The intramonomer was identified by having all its ligand residues to zinc coming from within 1 monomer, while the intermonomer was defined by the residue ligands coming from two monomers. To make an asymmetric unit, the intramonomer of the Rosetta match output was saved in its own file without modification. Similarly, for the intermonomer, this was also done, but the His ligand was manually moved to its symmetric position before saving.

The symmetry setup used in step 3 is described in 4.2.11. The total score is calculated using the master subunit. The total score is 6x the energy of the master subunit + the interaction with other monomers for which there are 4x, 3x, 2x and 2x interfaces. See also the symmetry files and the visual example in Appendix C.7.

The penalty that is applied to the score function to optimize the zinc site is given as a sum of harmonic constraints as below, involving only the CPs labelled as free are constrained. For all BMs except for Cys the following penalty will be applied as in **(eq 4-21):**

$$\text{Penalty} = \sum_{i=1}^{4}\left(\frac{(\text{D}-\text{opt})^2}{std}+\frac{(\text{A}_B-\text{opt})^2}{std}+\frac{(T_B-\text{opt})^2}{std}+\frac{(A_A-\text{opt})^2}{std}+\frac{(T_A-\text{opt})^2}{std}\right) \quad \textbf{(eq 4-21)},$$

Where opt is the optimal values of the corresponding CPs as given in **Table 5-1**, and the sum is over all 4 coordinating residues. For Cys it would be given by **(eq 4-22):**

$$\text{Penalty} = \sum_{i=1}^{4}\left(\frac{(\text{D}-\text{opt})^2}{std}+\frac{(\text{A}_B-\text{opt})^2}{std}+\frac{(A_A-\text{opt})^2}{std}+\frac{(T_A-\text{opt})^2}{std}\right) \qquad \textbf{(eq 4-22)},$$

The optimization and sequence design step carried out in step 3 iterates 3 times over two sub-steps: 1) SC identity and rotamer optimization (using the Rosetta Packer module) and 2) rotamer and BB minimization (using the Rosetta minimization module). Subset 1 involves random insertion of SC identities and rotamers for those SC identities using a MC approach with simulated annealing. Substeb 2 involves gradient descent minimization, where the torsion angle space of

both the BB and the rotamers are optimized. At the start of the trajectory the fa_rep term is very low but is slowly increased over the trajectory. The amino acid identities are not changed in step 1. The SZ residues are not allowed to change in step 1 or 2. The ligand residues to zinc are allowed to be optimized, but not designed (insertion of new amino acids). The penalty given in **(eq 4-21)** and **(eq 4-22)** is applied to Rosetta's score function throughout the trajectory. The score function is high resolution and uses the newest version of the score function (ref2015 [159]). Weights for the penalties were set to 1.0.

The intermonomer and intramonomer structures chosen from the Rosetta match output were run individually on 1 of the supercomputers. Each model was run using 70 cores, for about 3 hours. This produced 140 structures for each model. These were then manually inspected, and the best structures (based on a qualitative inspection of the optimal coordination-sphere, hydrogen bonds, clashes and packing) were selected. Commands and files that were used in step 3 involving design and optimization are given in Appendix C.6.

“Success is not final, failure is not fatal: it is the courage to continue that counts.”

- Winston Churchill

5 Results

5.1 Initial design efforts

Developing the protocol for designing the structures as described in Chapter 2 proved to be a significant challenge. The Initial design efforts of the project were focused on designing the BB structures of the amyloid structures, similar to what has been done for *de novo* designed αβ-proteins (**Figure 1-7.a**) (such as in references [58, 59]), and a zinc site similarly to what has been done in the *de novo* design of enzymes or ligand proteins (**Figure 1-7.d-e**), using only Cys and His as residue ligands to zinc (such as in reference [130]). This approach, and all approaches described below, including the current algorithm described in section 5.5, use the Rosetta software suit (see Materials 4.1).

In the first approach, as mentioned above, the entire BB structure of an αβ-protein was designed followed by the design of a zinc site into the generated BB structures. Appendix A.1 has a more in-depth discussion of this approach, but it was eventually abandoned, and new approaches were sought. Initially two separate algorithms were developed. The first algorithm focused on first placing zinc on the 2 β-strands making up the monomer of the amyloid fiber with some of the residue ligands originating from the β-strands, while the rest were free floating in space. An α-helix was then "docked" onto those free-floating residue ligands and filtered based on clashes and optimal zinc site geometry. Appendix A.2 elaborates further on this method, but this approach was also abandoned. The second algorithm that was developed focused on sampling an α-helix in different positions above the 2 β-strands, and then designing the entire zinc site at once when the proper α-helix position had been found. Appendix A.3 has an in-depth discussion on this approach, but it was also abandoned in favor of the algorithm that is currently used to design the amyloid structures (elaborated upon in section 5.5). Several things were learned from these initial design efforts:

1) To increase chances of successful designs, designs must start from many different starting structures. Therefore, the 2 β-strands that are part of the non-designed steric zipper, must be selected from a large set of crystal structures. Furthermore, these crystal structures must be carefully selected based on designability. The result of this selection is shown in section 5.2

2) Designing a zinc site with only His and Cys is limited. Only 1 Cys in the tetrahedral coordination sphere must preferably be designed, because oxidation of the sulfur groups of 2 cysteines in close proximity can lead to disulfide bonds which can interfere with the integrity of the designed structures. This means, that in principle the tetrahedral coordination sphere of zinc could only consist of 3-4 His and 0-1 Cys. Asp and Glu were added in the design to: a) increase the diversity of interaction to zinc, b) increase the chance of getting favorable interaction between the α-helices and the β-strands, c) neutralize the positive charges on zinc by the negative charge of the carboxylate groups, which is important for stability, and finally d) ease downstream design because the oxygen groups do not need a hydrogen donor in the second coordination sphere, when they are coordinated to zinc. Lastly, to properly design a zinc site, the geometry between the zinc and the residue ligands (His, Cys, Glu and Asp) must be properly parametrized. The parametrization of the residue ligands is shown in section 5.3.

3) The space between the β-strands and α-helices, that would allow a zinc to fit in between them must be considered when the BB is initially sampled. This demands that the volume of the tetrahedral coordination sphere of zinc is considered explicitly in the design. Section 5.4 shows the results of an analysis of the volume of zinc's tetrahedral coordination sphere.

The 2 β-strand selection, the volume of zinc, and the parametrization of zinc and the residue ligands are used in the final design protocol which is described in section 5.5. Finally, in section 5.6 the structures generated using this protocol are shown and evaluated.

5.2 Scaffold selection of 2 β-strands

2 β-strands that should constitute part of the monomer in the amyloid structure were selected and cut out from amyloid crystals structures derived from the PDB (**Figure 5-1**). These 2 β-strands are referred to as a "scaffold", because the BB and SC of the steric zipper interface is not redesigned.

To obtain as many 2 β-strand scaffolds as possible, the entire PDB was manually searched for amyloid crystal structures (see Methods 4.2.1). This yielded 84 amyloid crystal structures which were then manually assigned to their symmetry group (see Methods 4.2.2 and section 1.3 for information on the symmetry groups). To select 2 β-strands from crystal structures that would be useful as a scaffold in the design, a list of criteria was used in the selection process. These criteria were based on amyloid symmetry classes, ease of modelling, energetics and self-association. All criteria are listed in **Table 4-1** in Methods 4.2.3. The class and symmetry criterion are important and elaborated upon here, while Methods 4.2.3 elaborates further on the other criteria.

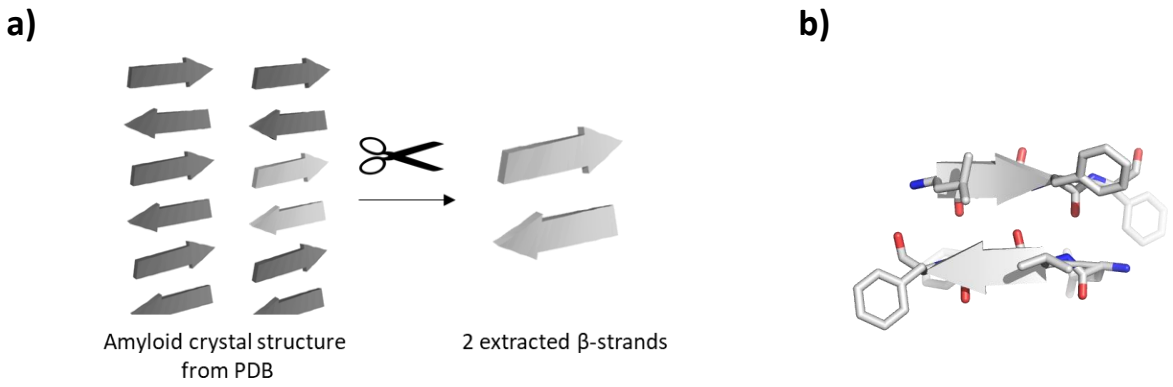


Figure 5-1: The process of cutting out 2 β-strands from the PDB. a) An amyloid crystal structure is shown to the left. 2 β-strands were identified (white) in the crystal structure and cut out (right). **b)** An example of the structure of a 2 β-strand scaffold that is used in downstream design. PDB ID: 2Y29.

To keep loop structures short, monomers with the β-L-α-L-β SS should be derived from parallel β-strands (amyloid symmetry classes 1, 2 and 4), while monomers with the α-L-β-L-β or β-L-β-L-α SS should be derived from antiparallel β-strands (amyloid symmetry classes 5, 6, 7 and 8).

The choice of **1)** not redesigning the BB or the SC of the steric zipper interface, and **2)** that only 1 monomer should reproduce the entire amyloid structure through symmetry operations,

posed some restrictions on what crystal structures could be used for design. For the α-helix not to be designed in the steric zipper interface, a 2-fold symmetric rotation must be possible for the monomer while keeping the steric zipper interface intact. Amyloid symmetry classes, where the 2 β-sheets are only related by translation (amyloid symmetry classes 2, 4 and 6) could not satisfy this criterion and were therefore discarded. Furthermore, the identity of the asymmetric unit also played a role in what scaffolds could be used and is discussed further in Appendix A.4.

Following the criteria as in **Table 4-1** yielded only 4 potential 2 β-strand scaffolds for the α-L-β-L-β or β-L-β-L-α monomers. More scaffolds were available for the β-L-α-L-β SS monomers, but to keep it even, only 4 structures were selected for that type of monomer based upon how the 2 β-sheets self-associated (see Methods 4.2.3).

In total 8 scaffolds were selected for downstream design: 4 for β-L-α-L-β monomer (PDB ID 1YJP, 4NIO, 4NIP and 4RP7), and 4 for either the α-L-β-L-β or β-L-β-L-α monomers (PDB ID: 2Y2A, 2Y29, 3FR1 and 3OW9). Each scaffold consists of two β-strands and a steric zipper interface that were cut out from the crystal structures (see Methods 4.2.4) and **Figure 5-1.b**.

5.3 Parametrization of ligand geometries

6 coordination parameters (CP): D, A_B, T_B, A_A, T_A and T_{AB} were defined for each of the BMs (see section 1.2.1 for a description of the BMs) of His (ZN:NE2 and ZN:ND1), Cys (ZN:SG), Glu (ZN:OE1, ZN:OE2, ZN:OE1|OE2) and Asp (ZN:OD1, ZN:OD2, ZN:OD1|OD2). The 6 CPs were analyzed using data obtained from the PDB. The definition of the 6 CPs are shown in **Figure 5-2** for the ZN-NE2 BM and **Table 5-1** (page 52) list the definitions of all atoms involved for each BM and CP. The definitions of the CPs were carefully defined so that the bond distance (D), bond angles (A_B), tetrahedral coordination sphere of zinc (A_A and T_A) and the imidazole ring and carboxylate group's preference for orienting planarly towards the zinc site [181] could be properly modelled.

2945 high resolution crystal structures containing 1 or more zinc sites were selected from the PDB (see Methods 4.2.5). The full dataset (all 2945 crystal structures) was used to obtain the values for the 3 CP: D, A_B and T_B, while a subset of the full dataset containing only mononuclear tetrahedral sites was used to obtain values for the other 3 CP: A_A, T_A, T_{AB}.

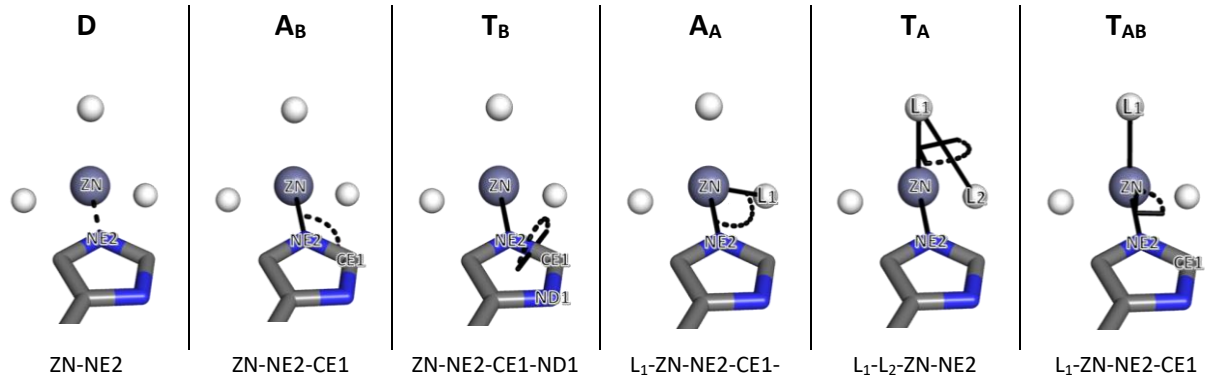
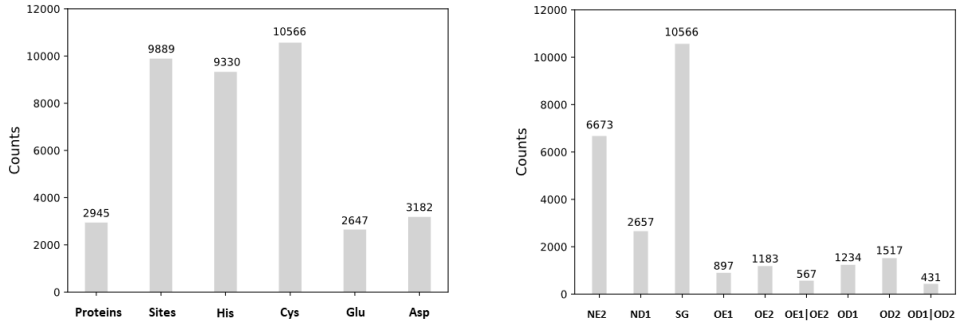


Figure 5-2: The 6 coordination parameters for the ZN-NE2 binding mode. The 6 CPs (D, A_B, T_B, A_A, T_A and T_{AB}) are shown at the top, and the atoms involved in the corresponding CP are indicated in the image. The bottom of the image indicates how the CP is defined. The CPs are defined by the atoms read from left to right. For instance, the T_B CP is defined as the torsion angle from ZN to NE2 to CE1 to ND1. Histidine is shown below with carbon as grey and nitrogen as blue. Zinc is shown in light grey, and 3 other ligand residues (in a tetrahedral coordination sphere) are shown in white. L1 and L2 refers to two of those other residues.

a) Full dataset



b) Subset

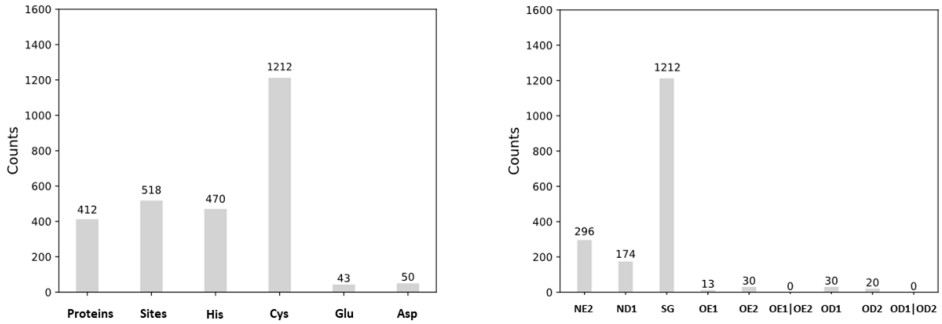


Figure 5-3: Proteins, sites, residues and binding modes analyzed in the two datasets. **a)** Information obtained from the full dataset. **b)** Information obtained from subset containing only mononuclear tetrahedral sites. Left: The amount of proteins (unique crystal structures), sites, and residues investigated are shown, and the exact number is shown in top of the bars. Right: The number of unique BMs investigated are shown.

Coordinating atoms were accepted in the analysis only if they fulfilled certain criteria (see Methods 4.2.5).

Because the anti-coordination mode of Asp and Glu have significantly lower energy than the syn-BM [182], these were identified and not used in the final datasets (see Methods 4.2.6). **Figure 5-3** shows the amount of proteins, zinc sites, residues, and BMs obtained from both datasets.

For each zinc site in either the full dataset or the subset, values were obtained for each CP and BM present in that site, and these were used to create a density histogram for each BM and CP such as those shown in **Figure 5-4** for the ZN-NE2 BM (see Methods 4.2.7). The plots for the other 8 BMs can be found in Appendix B.1.

Not all CPs have optimal values. For instance, the T_{AB} CP of the ZN-NE2 BM does not have an optimal value (**Figure 5-4**) but spans a range from 0-360 degrees more or less evenly. CPs that seeming to have optimal values such as the T_{AB} CP of the ZN-ND1 BM (see Appendix B.1 and **Figure 5-5**) were identified to come from the steric hindrance between the BB of the residue and zinc, and these should not be explicitly modelled, since sterically hindrance is considered explicitly in Rosetta's score function. The CPs that do not have optimal values, or pseudo optimal values coming from steric hindrance, are designated as "free" (see also **Table 5-1**) and are not explicitly modelled (see later).

For the CPs having optimal values not coming from steric hindrance (referred to as "non-free" CPs), different parametric probability distributions functions (pdfs) were fitted to the density plots using maximum likelihood estimates (MLE), and the best fits were evaluated based upon the least sum of squared errors (SSE) (see Methods 4.2.7). Two probability distributions were selected. Normal pdfs were fitted to all of the non-free CPs for each BM, except for the D CP for His, Glu and Asp. The distributions of the D CPs for these BMs are right skewed, so the Johnson SU pdf were used instead (see Methods 4.2.7). The fits are shown in **Figure 5-4** and in Appendix B.1 along with the density histograms. The optimal values for each of the fits were numerically found as the maximum of the corresponding pdfs (indicated as Max) along with the sample standard deviation (σ) which were also calculated (see Methods 4.2.7).

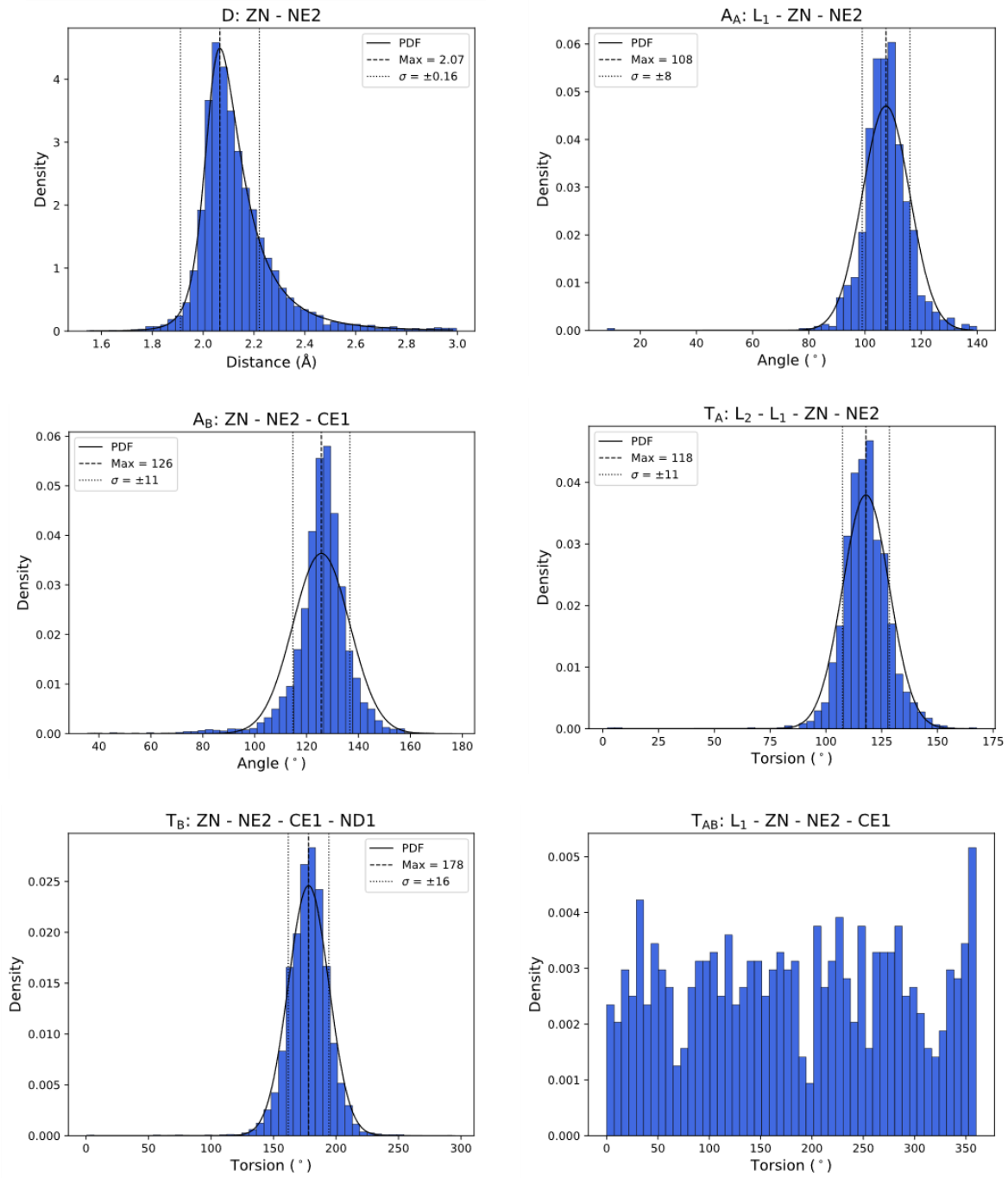


Figure 5-4: Density histogram and pdf fits for the CPs of the ZN:NE2 binding mode. Each individual figure is labelled with the corresponding CP on top, along with the involved atoms defining that CP. See also **Figure 5-2**. The figures show a density histogram for each CP, and if not labelled as "free" (see main text and **Table 5-1**), such as the T_{AB} CP, the pdf that was fitted to the density histogram shown (labelled PDF). The maximum value or optimal value (Max) is also shown along with the standard deviation (σ), and plotted as Max - σ and Max + σ.

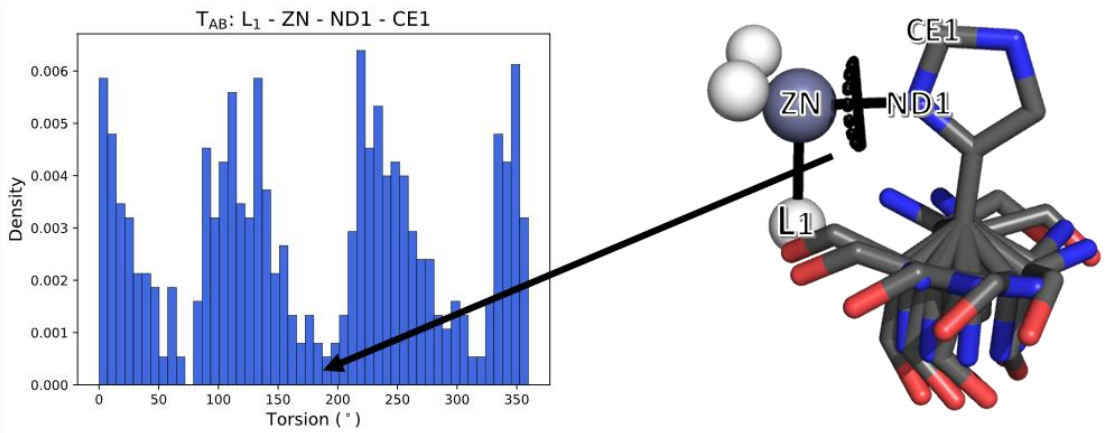


Figure 5-5: Sterical hindrance modulates the CP distribution. Left: the density histogram of the T_{AB} CP for the ZN:ND1 BM is shown. It looks like there are 3 optimal values at 0°, 120°, 240°, and 3 non-optimal values at 60°, 180°and 300°. Right: The rotamers of His are shown along with zinc and the positions of the other 3 residues (shown as a white sphere). The atoms involved in the T_{AB} CP are also marked. The sterical hindrance of the BB of His in the ZN:ND1 BM modulates the density histogram. The angle (shown as the black maker to the right) of the T_{AB} is approximately 180 degrees and an arrow shows the corresponding position in the density histogram. The L_1 coordination residue clashes with the BB at this angle, and therefore the probability of seeing this angle for the ZN:ND1 BM is low. This happens at every 120° interval which corresponds to the minima at 60°, 180° and 300°.

Table 5-1: Optimal values and standard deviation for each CP and BM. The columns shows the BMs and the optimal values and standard deviations that are used to model them (see section 5.4 and section 5.5 for their uses). The rows show the 6 CPs along with atoms that defines the corresponding CP of that BM. The definition is read from left to right, so for the ZN:NE2 BM, the T_A CP describes the torsion angle from L_2 to L_1 to ZN and then to NE2. See also Figure 5-2 for the definition of L_2 and L_1. The A_A and T_A CPs were set to ideal values (marked as idealized) for a tetrahedral coordination sphere. Some values are marked as "free" which means they are not modelled. The standard deviation of A_A and T_A for the ZN:OE1|OE2 and ZN:OD1|OD2 BM were set to ± 15, and are not derived from the analysis described in the text, as the others are.

	D	A_B	T_B	A_A	T_A	T_{AB}
ZN:NE2	**ZN-NE2**	**ZN-NE2-CE1**	**ZN-NE2-CE1-ND1**	**L_1-ZN-NE2-**	**L_2-L_1-ZN-NE2**	**L_1-ZN-NE2-CE1**
	2.07 ± 0.16	126 ± 11	178 ± 16	109.5 ± 8 (idealized)	120 ± 11 (idealized)	**Free**
ZN:ND1	**ZN-ND1**	**ZN-ND1-CE1**	**ZN-ND1-CE1-NE2**	**L_1-ZN-ND1-**	**L_2-L_1-ZN-ND1**	**L_1-ZN-ND1-CE1**
	2.07 ± 0.16	120 ± 8	181 ± 17	109.5 ± 8 (idealized)	120 ± 9 (idealized)	**Free**
ZN:SG	**ZN-SG**	**ZN-SG-CB**	**ZN-SG-CB-CA**	**L_1-ZN-SG-**	**L_2-L_1-ZN-SG**	**L_1-ZN-SG-CB**
	2.34 ± 0.11	105± 8	**Free**	109.5 ± 7 (idealized)	120 ± 8 (idealized)	**Free**
ZN:OE1	**ZN-OE1**	**ZN-OE1-CD**	**ZN-OE1-CD-CG**	**L_1-ZN-OE1-CD-**	**L_2-L_1-ZN-OE1**	**L_1-ZN-OE1-CD**
	2.03 ± 0.20	123 ± 13	181 ± 24	109.5 ± 10 (idealized)	120 ± 14 (idealized)	**Free**
ZN:OE2	**ZN-OE2**	**ZN-OE2-CD**	**ZN-OE2-CD-CG**	**L_1-ZN-OE2-CD-**	**L_2-L_1-ZN-OE2**	**L_1-ZN-OE2-CD**
	2.02 ± 0.19	122 ± 13	178 ± 25	109.5 ± 11 (idealized)	120 ± 12 (idealized)	**Free**
ZN:OE1|OE2	**ZN-CD**	**ZN-CD-OE1**	**ZN-CD-OE1-OE2**	**L_1-ZN-CD**	**L_2-L_1-ZN-CD**	**L_1-ZN-CD-OE1**
	2.63 ± 0.23	63 ± 7	-2 ± 17	109.5 ± 15 (idealized)	120 ± 15 (idealized)	**Free**
ZN:OD1	**ZN-OD1**	**ZN-OD1-CG**	**ZN-OD1-CG-CB**	**L_1-ZN-OD1-CG-**	**L_2-L_1-ZN-OD1**	**L_1-ZN-OD1-CG**
	2.01 ± 0.19	124 ± 12	173 ± 26	109.5 ± 13 (idealized)	120 ± 14 (idealized)	**Free**
ZN:OD2	**ZN-OD1**	**ZN-OD1-CG**	**ZN-OD1-CG-CB**	**L_1-ZN-OD1-CG-**	**L_2-L_1-ZN-OD1**	**L_1-ZN-OD1-CG**
	2.01 ± 0.17	124 ± 11	183 ± 26	109.5 ± 10 (idealized)	120 ± 11 (idealized)	**Free**
ZN:OD1|OD2	**ZN-CG**	**ZN-CG-OD1**	**ZN-CG-OD1-OD2**	**L_1-ZN-CG**	**L_2-L_1-ZN-CG**	**L_1-ZN-CG-OD1**
	2.66 ± 0.22	62 ± 7	0 ± 16	109.5 ± 15 (idealized)	120 ± 15 (idealized)	**Free**

To see if there was any difference between the full dataset and the subset containing mononuclear tetrahedral sites, the D, A_B and T_B CPS were also analyzed for the subset, as previously described. The results are shown in Appendix B.2. Comparison between the ZN-OD1|OD2 and the ZN-OD1|OD2 BMs were not possible because these were not observed in the subset (see **Figure 5-3.b**), but for the other BM the three parameters were very similar and well within in the standard distribution. This analysis showed that using the CP obtained from a general analysis of zinc sites, and not only mononuclear tetrahedral zinc sites, seemed to be valid, but further analysis must be carried out to determine if a difference really exist (see Future work).

To correctly model a tetrahedral geometry, the optimal value found for the A_A and T_A CPs were not used in the design, but instead set to the ideal values for a tetrahedral geometry (A_A = 109.5° and T_A = 120.0°). The standard deviations found for these CPs were still used (which models BM flexibility – see later). Furthermore, since no data was obtained for the ZN-OD1|OD2 and the ZN-OD1|OD2 BMs for the subset, the standard deviation for the A_A and T_A CPs, were set to ± 15°, while the T_{AB} is modelled as free. **Table 5-1** shows the non-free optimal values and standard deviation that was either obtained from this analysis or set to ideal values (indicated as ideal) or modelled as free (indicated as free).

The optimal value and the standard deviation are used in the protein of the amyloid structures in 3 ways: **1)** to model the volume of the zinc coordination sphere, which is used to properly sample a BB structure, **2)** to put zinc and 4 coordinating residues into an appropriately sampled BB structure (using Rosetta Match), and **3)** to optimize the coordination geometry during sequence design optimization.

Section 5.4 shows the results of an analysis that was used to model the volume of zinc (**1**), while section 5.5 elaborates on how the results of section 5.4, and how (**2**) and (**3**) are used in the design.

5.4 Distances between zinc ligands

To explicitly model the volume of a mononuclear tetrahedral zinc site, 2 pairs of each of the amino acids: His, Cys, Glu and Asp, and their corresponding BMs were sampled around a zinc

atom, and the distances between all the involved CB atoms (CB-CB distances) were calculated (**Figure 5-6**).

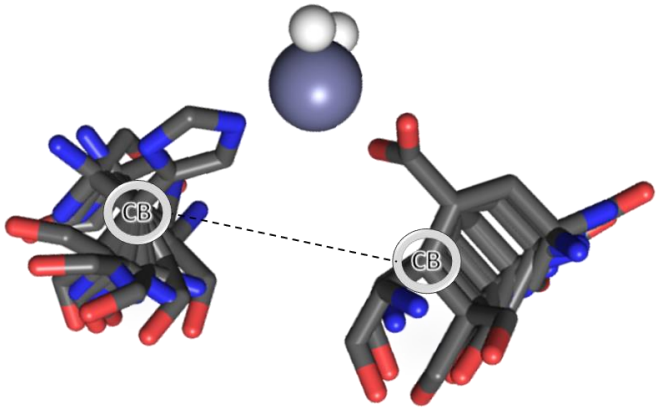


Figure 5-6: Sampling of residues, BMs, CPs and their rotamers around a zinc site. 2 amino acids (exemplified by the His and its ZN:NE2 BM and Glu and its ZN:OE1 BM) were positioned at the two vertices of the tetrahedral coordination sphere of zinc according to their optimal CPs. These were then sampled around 1 standard deviation for each CP along with their rotamers, and all coordinates of the CB atoms were recorded. CB-CB distances were then calculated for all residues and BMs. A discrete sample point with all of the respective rotamers of the involved amino acids are shown, and the distance (dashed line) are shown between the 2 involved CB atoms.

His, Cys, Asp and Glu were positioned at two of the vertices of the tetrahedral coordination sphere of zinc (see **Figure 5-6**) according to their ideal values of their BMs and CPs (see **Table 5-1**). Then each BM and CP was sampled with 1 standard deviation away from the optimal value as is given in **Table 5-1** (although with non-rounded values - see Appendix A.7) along with the rotamers possible for that amino acid (see Methods 4.2.8). The CPs indicated as free were sampled as well. For each discrete sample point the position of the CB atom for all rotamers was recorded and the minimum and maximum distance between two CB atoms of the two residues sampled at the two vertices were calculated. **Table 5-2** lists the minimum and maximum CB-CB distance for each BM combination, along with the largest (green) and smallest (red) distances measured across all BM combinations.

The largest distance possible is 10.9 Å between the ZN:NE2 and ZN:NE2 BMs, and the shortest possible is 2.1 Å between the ZN:ND1 and ZN:ND1 BM. The minimum and maximum distances are used explicitly in the design process which is described next.

Table 5-2: Minimum and maximum CB-CB distances for all combinations of BMs. The minimum and maximum CB-CB distances calculated by the sampling described in the main text is shown. Green highlights the largest distances observed across all BM, and red highlights the shortest.

	ZN:NE2	ZN:ND1	ZN:SG	ZN:OE1	ZN:OE2	ZN:OE1|OE2	ZN:OD1	ZN:OD2	ZN:OD1|OD2
ZN:NE2	6.5-10.9	4.2-9.3	4.7-8.9	5.9-10.2	5.8-10.7	5.9-10.3	5.7-9.6	5.8-9.6	5.8-9.5
ZN:ND1		2.1-7.3	2.2-7.1	3.5-8.5	3.7-9.0	3.6-8.6	3.2-7.9	3.3-7.9	3.2-7.9
ZN:SG			2.7-6.8	4.0-8.1	4.1-8.6	4.1-8.3	3.7-7.6	3.8-7.5	3.8-7.5
ZN:OE1				5.3-9.4	5.3-9.9	5.3-9.6	5.0-8.9	5.1-8.8	5.1-8.8
ZN:OE2					5.3-10.4	5.3-10.1	5.0-9.4	5.1-9.4	5.1-9.3
ZN:OE1|OE2						5.3-9.7	5.1-9.0	5.1-9.0	5.1-8.9
ZN:OD1							4.7-8.3	4.8-8.3	4.9-8.2
ZN:OD2								4.9-8.3	5.0-8.2
ZN:OD1|OD2									5.0-8.1

5.5 Computational design

The protocol developed to design the structures as outlined in Chapter 2 is described in this section. The protocol happens in 3 steps. In step 1 the BB of the target structures are sampled. In step 2 a zinc site is designed into the sampled BB structures, and in the last step, step 3, the zinc site and BB is optimized, and a sequence is designed. Each individual step is described next.

5.5.1 Step 1: Generation of BB structures

To fit a zinc site into BB structures, and for the BB structures to be favorable for downstream design, the sampling of the BB structures, in the first step, must be guided appropriately. The described procedure will use one of the designed α-L-β-L-β monomers, having the 2 β-strands scaffolds from the PDB ID 2Y29, and loop lengths of 3 and 2 (referred to as the α3β2β-2y29 model) as an example. The approach is general, but a slight modification in loop building must occur for the β-L-α-L-β monomers, which is addressed later.

Because the SZ is kept fixed, only loops and an α-helix are designed. When sampling the BB of the two SSs there are several metrics that can be used to guide the search. For the α-helix, it is important that N-H and C=O groups of the BB makes favorable hydrogen bonds with each other. The α-helix must also have favorable helix-helix-packing, and be in contact with the β-

sheet, although it cannot clash with its adjacent helices or the β-sheet. Furthermore, the length of the α-helix must be appropriately chosen [58, 59]. Lastly the α-helix should also be able to appropriately fit a zinc site between it and the β-sheet. For loops, depending on the SS structure preceding and following them, the loop should be between 2-5 residues in length, and they should have their ϕ and ψ angles in specific torsion bins. The reader is referred to the following references [58, 59] for a more in depth description of how SS relates to loop length and torsion bins. The specific bins that the residues must lie in follows the ABEGO numbering and is explained in **Figure 5-8**. Lastly, the loops should also cap the α-helix with hydrogen bond interactions [183].

The following procedure captures all of this in the first part of the design involving BB sampling (see Methods 4.2.9 for details). The procedure starts from the 2 β-strand scaffolds which in the case of the α3β2β-2γ29 model would be cut out from the 2Y29 crystal structure (**Figure 5-1)**. The interface not facing the steric zipper is then mutated to alanine and a loop and an ideal helix, having ϕ=-57° and ψ=-47°, are built from one of the ends of the 2 β-strand scaffold, and a loop connecting the two β-strands is build using Rosetta's CCD algorithm (**Figure 5-7**).

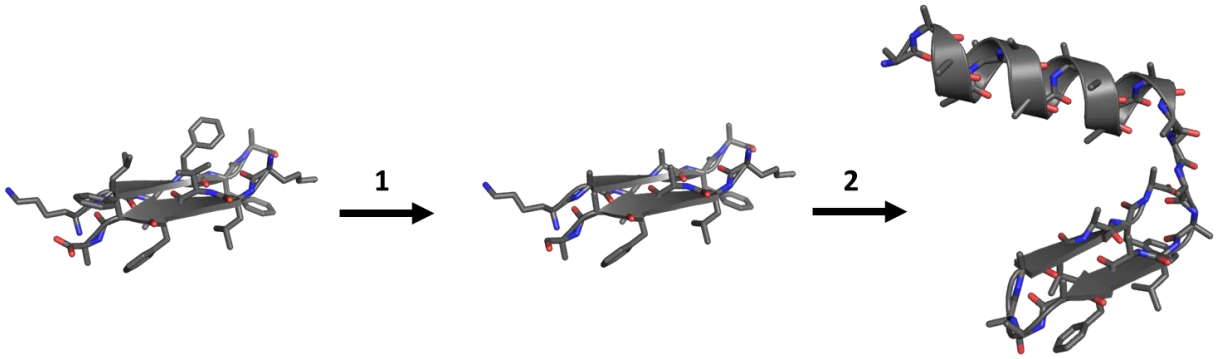


Figure 5-7: Building the starting structure from 2 β-strand scaffolds. The α3β2β-2γ29 model is shown as an example. In step 1 the non-SZ part of the 2 β-strand scaffolds extracted from the PDB is initially mutated to alanine. The residues not mutated are converted into loops in step 2. In step 2, a loop connecting to a α-helix is built. The loop between the 2 β-strands are built using Rosetta's CCD algorithm.

The α-helices built for the α3β2β-2γ29 model is 14 residues long. This is approximately the length of the 2 β-strands. Loops connecting the β-strand of the α3β2β-2γ29 with an alpha helix should have a loop length of 2-4 residues with the ABEGO assignment of GB, GBA, and BAAB, respectively. A loop length of 2, in some cases could not pack the α-helix appropriately together with the β-sheet, owing to the fact the SZ of the β-sheet is fixed, limiting the loops movability.

For the α3β2β-2y29 model a loop length of 3 is built following the GBA bin assignment (**Figure 5-8.b**). The GBA loop is also favorable because it can cap the α-helix. The loop connecting the 2 β-strands should have a length of 2 and a torsion bin assignment of either EA or GG. A loop length of 5 is also appropriate but for the structures build here, which have individual loops closely together, short loops seemed to be preferred. A loop length of 2 is built for the α3β2β-2y29 model (**Figure 5-7**). ABEGO bins assignment for this loop is not carried out since it will change during optimization in step 3 (see section 5.5.3).

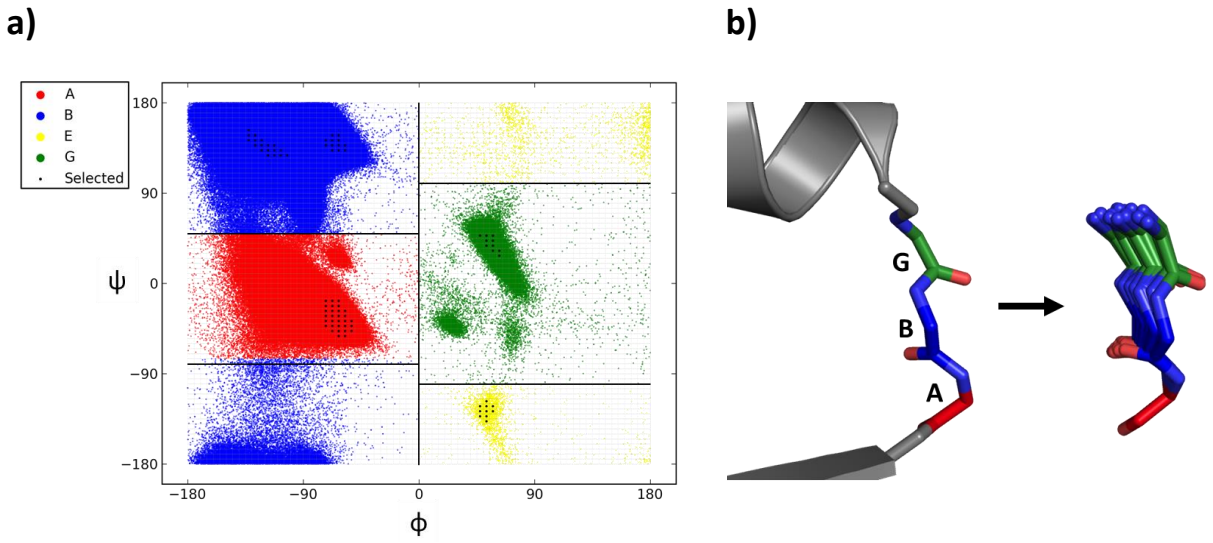


Figure 5-8: Sampling of loops in specific grids of ABEGO bins. a) Ramachandran plot generated from the analysis described in the main text and Methods 4.2.10. The Ramachandran plot is colored accordingly to ABEGO bins (A (red): $-180 \leq \phi < 0$ and $-75 \leq \psi < 50$). B (blue): $-180 \leq \phi < 0$ and $50 \leq \psi < 180$ or $-180 \leq \psi < -75$. E (yellow): $0 \leq \phi < 180$ and $100 \leq \psi < 180$ or $-180 \leq \psi < -100$. G (green): $0 \leq \phi < 180$ and $-100 \leq \psi < 100$. ABEG have $\omega \approx 180$ (trans); O has $\omega \approx 0$ (cis). The selected grids are shown as black dots. **b)** Left: the loop connecting to the α-helix of the α3β2β-2y29 model is assigned to GBA bins. The first residue preceding the α-helix is assigned to G, the next is B and the latter is A. Right: the GBA loop is sampled around the GBA bins, within the selected grid points

The rest of this section of step 1 is carried out using a new algorithm that was developed specifically for the design of the amyloid structures (see Methods 4.2.9 for details). The motivation and innerworkings of the algorithm is explained next.

To only favor highly ideal BB structures, used in later downstream design, the α-helix is kept as is (that is ideal) and is sampled as a rigid body by changing the torsion angles in the loop connecting it to the β-strand. A dataset of close to 8000 high resolution crystal structures were used to generate a Ramachandran plot as shown in **Figure 5-8.a** (see Methods 4.2.10). The

Ramachandran plot was then divided into ABEGO bins and 5x5 degrees grids. 30 ϕ and ψ combinations, from the midpoints in A and the B ABEGO bin grids, were chosen based on the grids having the most data points. Similarly, 10 ϕ and ψ combinations were chosen for the E and G bins. The selected ϕ and ψ combinations are shown in **Figure 5-8.a**. The loops are then sampled around those grids points, according to the favorable ABEGO assignment, and the α-helix will move accordingly (**Figure 5-8.b**).

Since the monomers should be designed in an amyloid environment, a symmetric system consisting of 2 adjacent monomers are used in the first step sampling process (**Figure 5-9.a**) (see Methods 4.2.11 for how the symmetry was set up). This system can also handle both the design of an intermonomer and intramonomer zinc site (see Chapter 2). At this stage the SCs are represented as centroids (**Figure 5-9.a**). Sampling the loops as described (see also **Figure 5-8.b**) produces an ensemble of starting structures as seen in **Figure 5-9.b**. Only those structures that pass certain filters are outputted, which is described next.

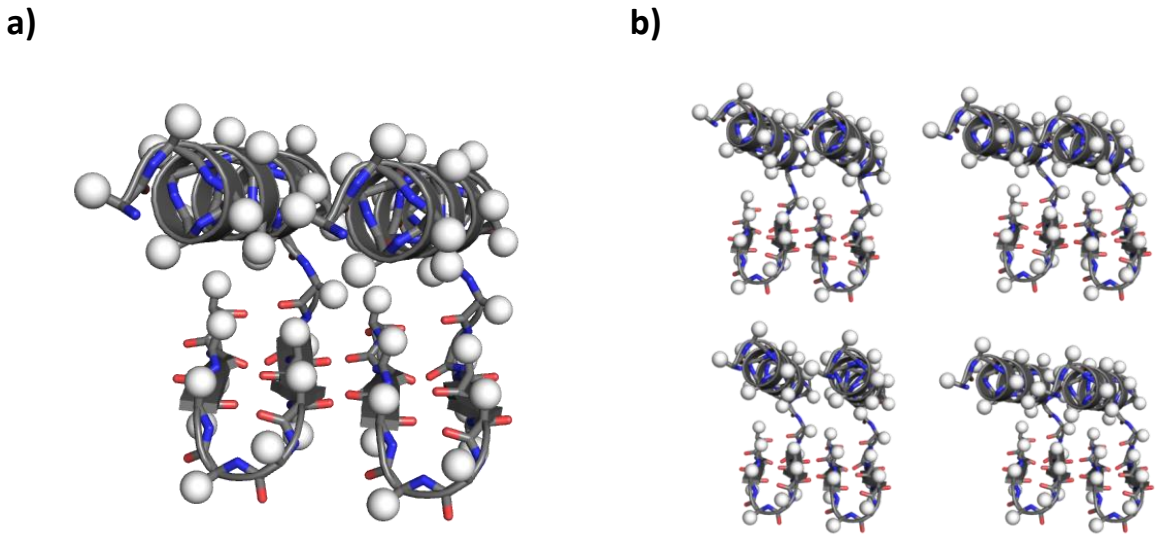


Figure 5-9: Symmetric setup and sampling of BB structures. a) The symmetric setup is shown. 2 identical monomers are positioned perpendicular to each other, in such a way that the 2 β-strands scaffolds are superimposable with the amyloid structure from where they came. The 2 monomers move in unison. The SCs are represented as centroids shown in white. **b)** 4 examples of the structures generated from the algorithm.

Filter 1 (see **Figure 5-10.a**) takes care of internal clashes by evaluating the steric repulsion between atoms. Other filters take care of appropriate positioning of the α-helices. For this

purpose, it was found that describing the β-sheet of the SZ as a plane, and one of the α-helices having a vector along its axis and a geometric center was suitable (see **Figure 5-10.b**). 3 filters removed structures with bad α-helix positioning. First, Filter 2 (see **Figure 5-10.a**) filters based on the minimum and maximum allowed angles between the α-helix vector and a vector normal to the β-sheet plane. Only sampled protein structures having this value between -30 to 30 degrees were accepted. This was done to allow favorable interaction between the β-sheet and the α-helix. Secondly, Filter 3 (see **Figure 5-10.a**) filters based on the minimum and maximum allowed angles between the α-helix vector and a vector in the β-sheet parallel to the fibril. Only sampled protein structures having this value between -30 to 30 degrees were accepted This was done to acquire favorable helix-helix-packing. The last filter, Filter 4 (see **Figure 5-10.a**) filtered structures based on the shortest distance between the geometric center of the α-helices and the β-sheet plane. This distance is based on the minimum and maximum CB-CB analysis shown in **Table 5-2** and will be explained below.

To allow all BMs to fit between the α-helix and β-sheet plane, a distance of minimum 2.1 Å and maximumly 10.9 Å can exist between the CB atoms of making up a favorable zinc site, as calculated in **Table 5-2**. Since the distance calculated here is between the β-sheet plane and geometric center of the α-helices, the minimum allowed distance is set to 3.15 Å (distance between the geometric center to the CB atoms of an ideal helix) and the maximum distance is set to 11 Å. The maximum distance of 11 Å is lower than what is maximally allowed from the CB-CB distances measured in **Table 5-2** (page 55) (would allow a distance 14.05 Å (10.9 Å+3.15 Å) between β-strands (see also below)), but such large distances between β-sheets and α-helices are rarely seen in nature [184]. The cut-off is set according to what is observed in nature.

Filtering and sampling the loops with the ideal α-helices as a rigid body, in a symmetric system containing 2 copies of identical monomers, yields good starting structures for step 2 which is the design of the zinc site. The algorithm developed in step 1, returns all amino acids to the full atom representation, and outputs certain positions that the Rosetta Match algorithm needs to search. Only BB structures that are below a plane, defined by the helix vector and a vector between the two helices of the symmetric system, are searched by the Rosetta Match algorithm (see Methods 4.2.9).

Other efforts that tried to improve the algorithm for step 1 were pursued, but these were not implemented. Appendix A.5 elaborates on this.

a)

Filter 1: Steric repulsion
Filter 2: α-helix - β-sheet interaction
Filter 3: Helix-helix-packing
Filter 4: Allow room for a zinc site

b)

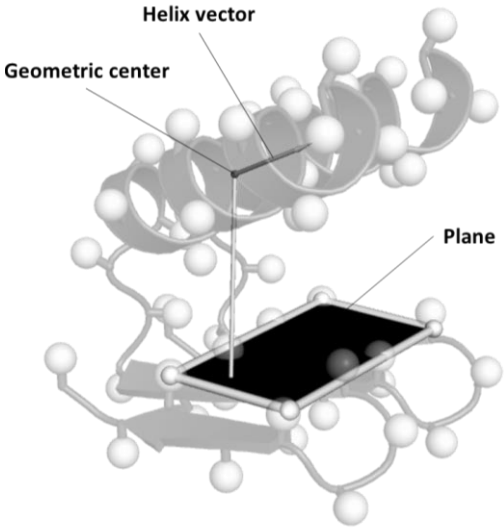


Figure 5-10: Filtering of sampled BB structures. **a)** The 4 filters that BB structures must pass. The filters are explained further in the main text. **b)** The setup used to describe the system (used by Filter 2-4). A geometric center, a helix vector and a plane are used to filter the structures in Filter 2-4.

5.5.2 Step 2: Design of a zinc site

For each BB structure generated in step 1, the Rosetta Match algorithm is used to build a zinc site (see Methods 4.2.12 for details). For this purpose, zinc is represented as ligand with 5 atoms, 4 of which are virtual atoms (V1, V2, V3 and V4) and 1 which is the real zinc atom (**Figure 5-11.a**). The 4 virtual atoms are positioned at a distance of 2.20 Å from the zinc atom, and at the vertices of an ideal tetrahedron (see section 1.2.1 and **Figure 1-3.b-c**). The Rosetta Match algorithm samples all BMs and their corresponding rotamers at each allowed position set from the algorithm in step 1 (see **Figure 5-11.b**) with the zinc ligand sampled between the non-free CPs at optimal values and standard deviation as given in **Table 5-1**, and the free CPs at specific intervals (see Methods 4.2.12). All BMs are sampled 4 times at each allowed position, but with different connections to the virtual atoms. The D, A_B and T_B CPs are always sampled similarly since these do not involve the virtual atoms. The T_A, T_B and T_{AB} CPs on the other hand are sampled with first (respectively) L1 and L2 changed to V4 and V2, then V4 and V3, then V4 and V1 and finally V2 and

V1. Each accepted position (see section 3.2.5) of a BM and its positioning of the zinc ligand is called a hit. 4 hits found for 4 BMs with 4 different connections to the virtual atoms are shown in **Figure 5-11.c**. 4 of such hits that place the zinc ligand with close coordinates in space (see section 3.2.5) are called a match. Matches are outputted by the Rosetta Match algorithm. Because the BB structures outputted from step 1 consist of 2 monomers, matches for both intermonomers and intramonomers can be found. Examples of matches for the α3β2β-2y29 model are shown for both cases in **Figure 5-12.a-b**.

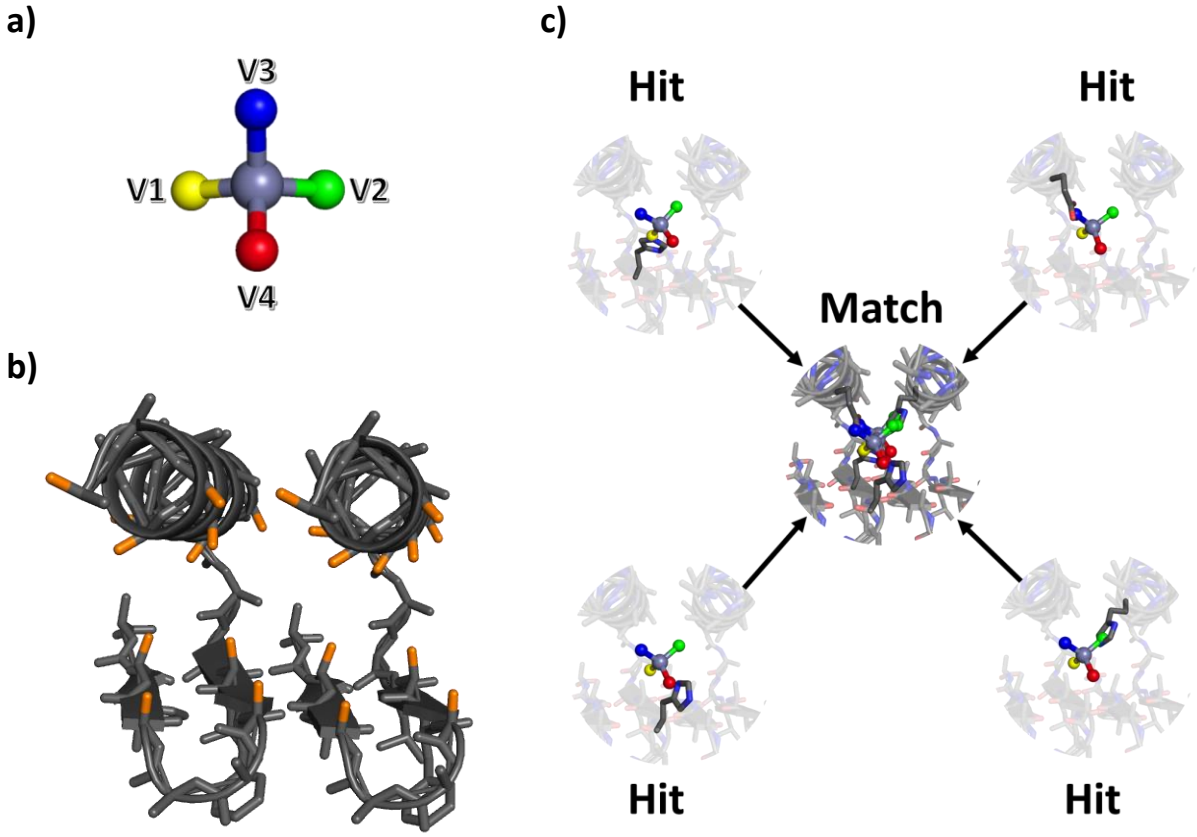


Figure 5-11: Design of a zinc site. **a)** The parametrization of the zinc ligand. It consist of a zinc atom (shown in grey) and then 4 non-identical virtual atoms called V1, V2, V3 and V4 shown in different colors. The virtual atoms were placed at the vertices of a ideal tetrahedron. **b)** shows one of the outputted structures from step 1 for the α3β2β-2y29 model. The positions the Rosetta Match algorithm can search through are highlighted with orange. These positions were determined by the algorithm developed for step 1. **c)** The inner workings of the Rosetta Match algorithm, using the parametrization of zinc as described (see **a** and main text) and His and Asp as an example. For His and Asp each rotamer is sampled at each allowed positions set from step 1 and constrained under the CPs calculated in the results. 4 hits (accepted positions of BMs and the zinc ligand) are shown for each connection to the virtual atoms. 4 hits that put the zinc ligand in close proximity in space are called a match and is outputted by the Rosetta Match algorithm.

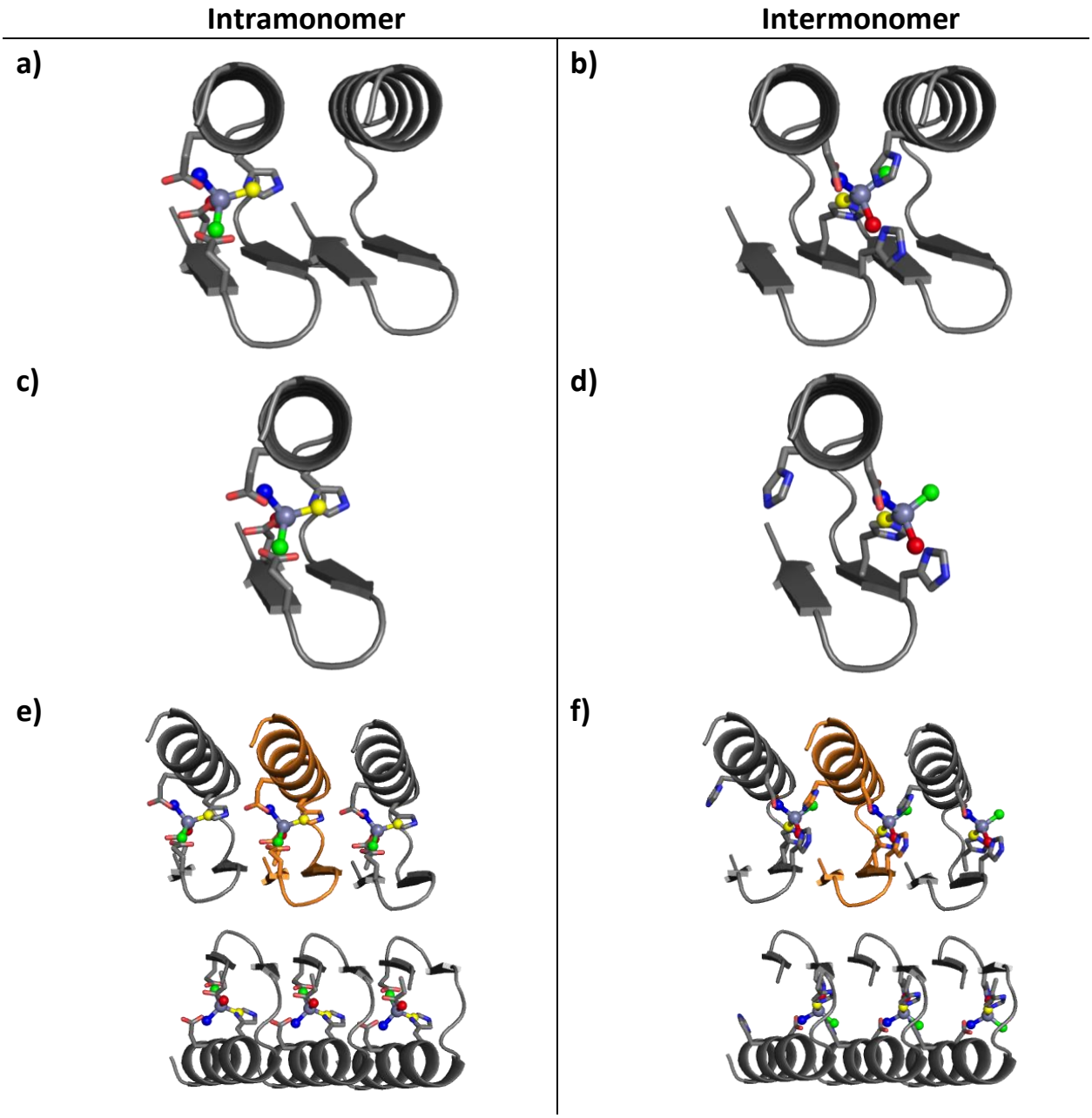


Figure 5-12: Matches for the intra and intermoner designs and the symmetric setup. Left (**a**, **c** and **e**) shows matches and the symmetric setup based on the intermonomer interaction, and right (**b**, **d** and **f**) shows the same for the intermonomer interactions. **a-b)** The output from Rosetta Match. The residue ligands from the intramonomer zinc site only originates from within 1 monomer, while the residue ligands in the intermonomer originates from both monomers (1 Asp and 2 His from the left one, and 1 His from the right one). **c-d)** The asymmetric unit must be defined differently for the two types. For the intramonomer the asymmetric unit can be immediately extracted from the Rosetta Match output while for the intermonomer the residue ligands have to be transferred to a single monomer (His in the case shown). **e-f)** The symmetric system for the optimization step consist of 6 monomers. The monomer shown in orange is the master subunit. All energy contributions are calculated with respect to it. The contributions come from within the master subunit itself and its interactions with the 5 other subunits.

5.5.3 Step 3: Zinc site optimization and sequence design

When a proper zinc site has been built into the BB structures, and both intramonomer and intermonomer zinc sites have been identified (**Figure 5-12.a-b**) from the Rosetta Match output, the structures have to be optimized (see Methods 4.2.13 for details).

First, the asymmetric unit of the two types of monomers have to be defined (**Figure 5-12 .c-d**). For the intramonomer, all residue ligands to zinc must originate from within 1 monomer. Because this is the case, a single monomer from the Rosetta Match output can immediately be represented as the asymmetric unit (**Figure 5-12.d-c**). For the intermonomer this is not the case. The residue ligands must originate from more than one monomer and a modification must occur to the output. In the case of the α3β2β-2y29 intermonomer model shown in **Figure 5-12.b**, the His from one of the monomers must be transferred to a single monomer such as it is shown in **Figure 5-12.d**

During optimization and sequence design, the asymmetric unit must be represented as it would look like in the full amyloid fibril. The contribution to Rosetta's energy score between atoms spaced 10 Å apart is negligible. The monomers in this design is spaced approximately 10 Å apart so 6 monomers are enough to describe the full amyloid system (**Figure 5-12.e-f**) (less might be more appropriate and more efficient - see Appendix A.6). The master subunit is shown in orange (**Figure 5-12.e-f**), and only the energy within the master subunit as well as the interactions it has with the atoms in its vicinity (the other 5 monomers) are calculated (see also Methods 4.2.13).

Constraints are applied to residue ligands to zinc according to their BMs and CPs. As for the Rosetta Match output, the virtual atoms are used for the T_A, T_B and T_{AB} CPs. Harmonic constraints are applied to all CPs of the respective BMs found using Rosetta Match, and these constraints use the optimal value and standard deviation obtained from **Table 5-1** (see Methods 4.2.13).

The sequence design and optimization steps are carried out using Rosetta's full atom scoring function, using all 20 canonical amino acids, and expressing the system as 6 identical monomers positioned with the 2 β-strand scaffolds as they would be from their respective crystal structure. An iterative approach between MC insertion of amino acids and rotamer sampling is

followed by BB and rotamer minimization (see Methods 4.2.13). The residue ligands to zinc are not designed, but their rotamers are sampled and minimized, while the harmonic constraints are applied. The SZ zipper is neither sampled or minimized during design.

5.6 Structure evaluation

The previous section outlined the development and procedure of a protocol that can generate the structures as described in Chapter 2. The protocol is a very recent development, and therefore large-scale sampling has not been carried out, and not all the models as described in Chapter 2 has been designed. A small-scale test of the whole protocol (step 1-3) was carried out with the α3β2β-2y29 model. It is designed as outlined in the protocol described in section 5.5. The entire procedure for each step (step 1-3) can be repeated by following the steps given in Appendix C.4 (step 1), C.5 (step 2) and C.6 (step 3). Methods section 4.2.9 (step 1), 4.2.12 (step 2) and 4.2.13 (step 3) also explains how to produce the results shown in this section. The procedure is also explained here briefly. Following step 1, the 2 β-strand scaffold from the PDB ID 2Y29 was cut out. Then a loop and a helix were designed as described in section 5.5 (the α3β2β-2y29 model at this stage is shown in **Figure 5-7**). This model was then sampled as described in step 1. This produced around 400 structures. One of those structures was then inputted to the Rosetta Match algorithm as in step 2 (the model at this stage is shown in **Figure 5-11** and **Figure 5-12**). This produced many structures, and a subset of the outputs were investigated. 1 intermonomer and 1 intramonomer model was selected (see Methods 4.2.12). The Rosetta Match algorithm was run slightly differently as described in section 5.5 (see Appendix A.7). Optimization and sequence design were then carried out as described in step 3 for the intermonomer and intramonomer and 140 structures were generated for each. One structure was then selected for the intermonomer and intramonomer model (see Methods 4.2.13). The selected models are shown in **Figure 5-13**.

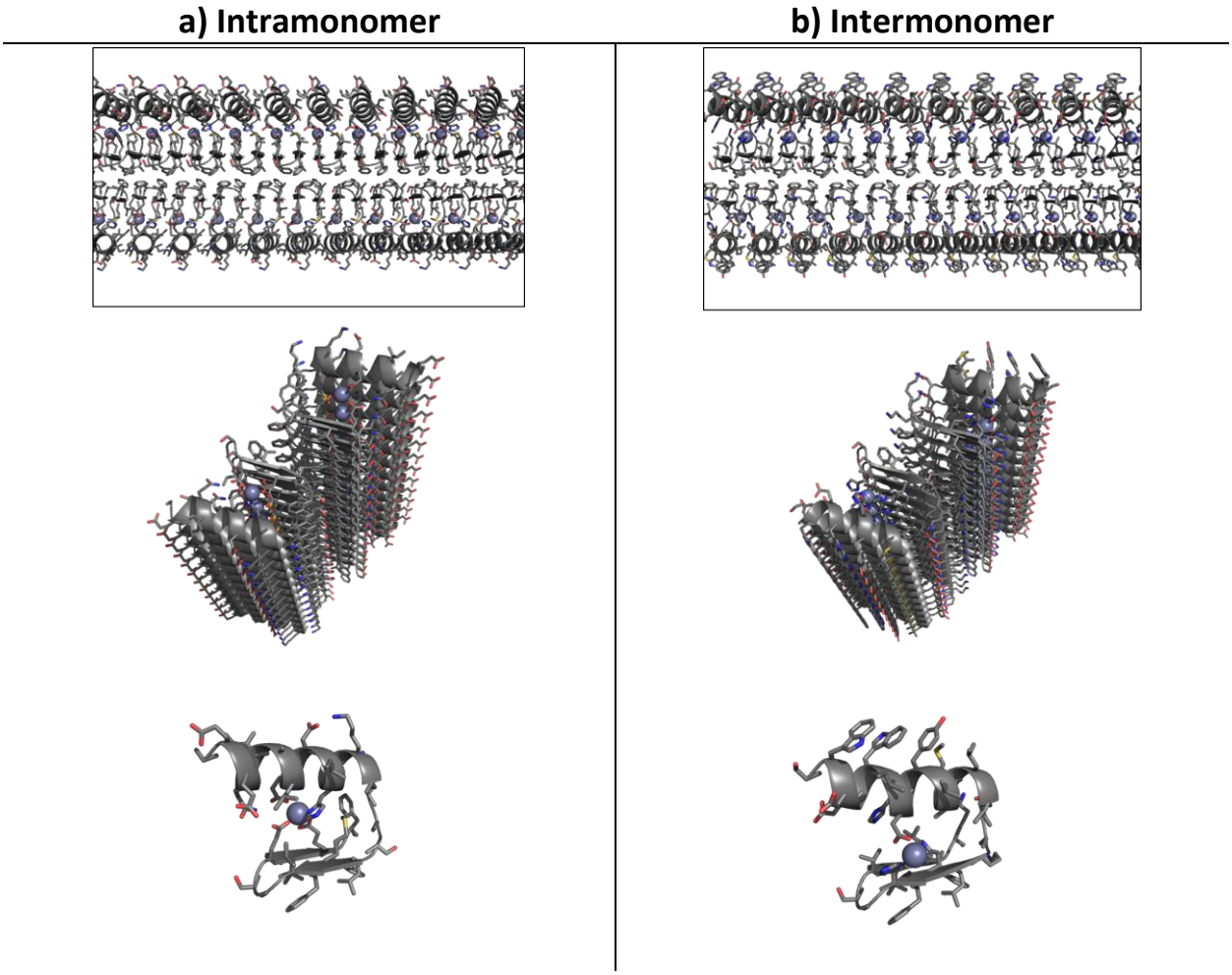


Figure 5-13: The designed intramonomer and intermonomer models. a) The Intramonomer model. **b)** The intermonomer model. Top and middle figure show two different views of the models in the full fibril. Bottom image shows the monomer for each type.

A few metrics were then evaluated for these structures. **Table 5-3** shows the score of all the score terms for the models before and after the optimization carried out in step 3. The optimized models have low energy sequences given by the total score at the bottom of the table. These energy contributions stem from reduction of the attractive and repulsive van der Waals terms (fa_atr and fa_rep), and electrostatic terms (fa_elec) and the ref energy term which described the energy difference between the folded and unfolded state. Looking at individual terms for each amino acid (data not shown), most of the residues contributes to the fa_atr and ref terms, while proper position of zinc and the removal of clashes (such as between A3 and P27) had major contributions to the fa_elec and fa_rep terms, respectively.

Table 5-3: All weighted score terms and their values recorded for the models. Starting from the outer most left column. Score terms shows the 20 terms used to score the models (terms are from the ref2015 score function). The next column shows the weight associated with each term, and the latter 4 columns shows weighted score for both the original optimized intramonomer and intermonomer models.

		Intramonomer		Intermonomer	
Score terms	**Weights**	**Original**	**Optimized**	**Original**	**Optimized**
fa_atr	1	-692.2	-935.2	-682.3	-995.6
fa_rep	0.55	732.8	183.3	1477.3	256.3
fa_sol	1	492.6	651.6	436.1	638.4
fa_intra_rep	0.005	0.8	2.01	0.907	2.3
fa_intra_sol_xover4	1	18.6	40.8	15.4	43.2
lk_ball_wtd	1	-10.4	-10.0	-8.1	-4.92
fa_elec	1	-495.0	-673.9	-211.3	-460.4
pro_close	1.25	0	0	0	0
hbond_sr_bb	1	-100.4	-88.5	-100.4	-80.5
hbond_lr_bb	1	-24.2	-26.9	-24.2	-30.0
hbond_bb_sc	1	0	-10.9	0	-21.4
hbond_sc	1	0	-15.2	0	-9.2
dslf_fa13	1.25	0	0	0	0
omega	0.4	15.6	22.8	15.6	24.4
fa_dun	0.7	130.6	259.7	122.3	243.3
p_aa_pp	0.6	10.1	-14.9	0.73	-32.9
yhh_planarity	0.625	0	0	0	0
ref	1	152.0	36.1	181.1	108.6
rama_prepro	0.45	86.4	28.6	76.0	24.2
Total		317.2	-550.5	1299.1	-294.3

Certain terms also go up in energy, these are the fa_dun term, which describes the favorability of rotamers, and the fa_sol term which describes the desolvation energy. Analyzing the score for each residue again, the energetic contributions were confirmed to come from most of the residues and not only from a few as was the case for the fa_atr term.

Table 5-4 shows the value of the CPs before and after the optimization step for all the involved BMs. The penalty score (P) is shown in the last column of the table. The penalty score highlights that in case of the intramonomer, the coordination sphere is less optimal, while for the intermonomer it is optimized. The penalties still show high values in both cases and looking at the individual CP values of the BMs, it can be seen that these in many cases are not very optimal for both the original and the optimized structures (optimal values are shown in **Table 5-1**).

Manual inspection of all the generated structures clearly indicated that these were not ideal (data not shown), so it seems to be a general trend.

Another problem that was observed was that the structures had cavities as seen in **Figure 5-14.** This was especially the case of the intermonomers as seen in **Figure 5-14.b**. Natural proteins are often well packed, and the cavities observed could pose a problem for stable folding.

Table 5-4: CP and penalty values for the original and optimized BMs. The BMs designed for the models are shown in the first row, and these are followed by the values of the individual CPs for that BM as given in the outer most left column. Both the intramonomer and intermonoer are shown as well as the original and optimized models. The total penalty score (P) applied to all the BMs are given in 1 model are shown in the last column

	Intramonomer								**Intermonomer**							
	Original				**Optimized**				**Original**				**Optimized**			
BM	**ND1**	**OE2**	**OE1**	**OE1|OE2**	**ND1**	**OE2**	**OE1**	**OE1|OE2**	**NE2**	**NE2**	**OD1|OD2**	**ND1**	**NE2**	**NE2**	**OD1|OD2**	**ND1**
D	1.91	2.71	2.98	2.84	2.19	2.00	2.00	2.68	2.23	1.40	3.33	3.44	2.32	2.24	2.67	2.34
A_B	120.2	87.5	137.7	106.7	127.6	159.6	147.7	87.5	114.7	159.9	51.9	143.5	114.6	112.7	43.8	138.4
T_B	163.9	138.6	234.5	12.5	180.0	47.4	161.3	17.3	178.1	212.5	19.5	184.7	180.0	215.5	5.9	165.8
A_A	101.4	88.2	139.7	93.1	99.6	142.4	130.2	84.0	100.0	92.2	108.3	100.5	120.4	114.3	111.0	100.2
T_A	129.4	110.9	145.3	133.7	142.2	170.2	139.0	106.6	97.3	92.5	128.1	170.6	150.1	104.3	127.0	161.9
P	832.8				1464				779.4				554.0			

Another finding that is worth mentioning is that at least in the case of intramonomer, many of the coordinating residues to zinc are shown to have hydrogen bonds with neighboring residues. Having hydrogen bonds in the second coordination sphere is very important for stability. Furthermore, since Asp is binding bidentate to zinc (according to the Rosetta Match placement) all hydrogen bonds are satisfied for the coordinating residues.

a) Intramonomer **b) Intermonomer**

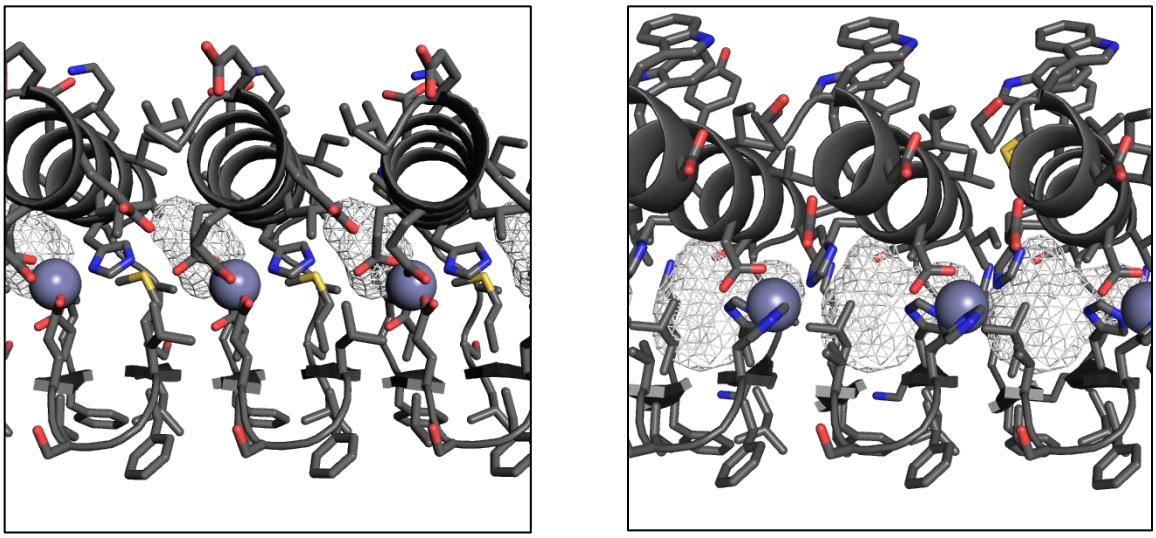


Figure 5-14 Cavities exists in the models. a) and **b)** shows cavities (seen as a white mesh) inside (mostly in the case of b) or in the interface (mostly in the case of a). The cavities are smaller for the intramonomer models, and larger for the intermonomer models.

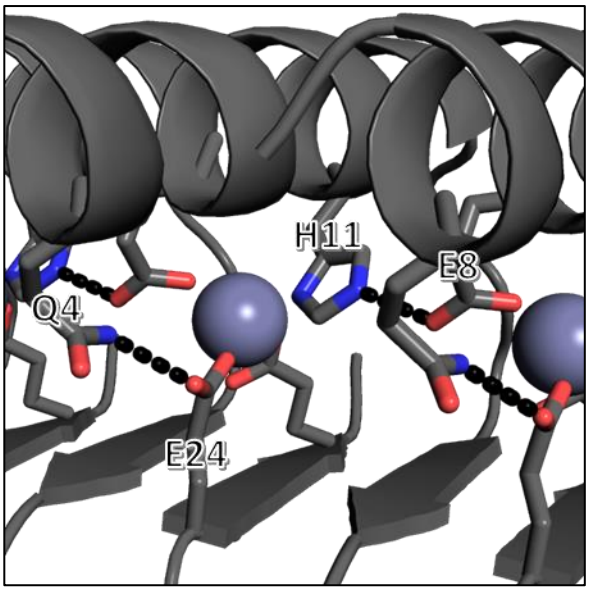


Figure 5-15: The intramonomer model has extensive hydrogenbonds. Hydrogen bonds are shown as dashed black lines. Hydrogen bonds can be seen between H11 and E9 and E24 and Q4. Asp (D26) (seen behind zinc in grey) is binding bidentate to zinc (according to the Rosetta Match output). The whole coordination sphere (E24, H11, E8 and D26) has satisfied all their hydrogen bonds.

6 Discussion

6.1 The design of a protocol for making amyloid switches.

In Chapter 2 the conceptualization of the amyloid switches was outlined, and in Chapter 5, a protocol for designing them was described and evaluated for a single model.

The protocol starts with a library of 2 β-strand scaffolds that are cut out of a set of selected amyloid crystal structures. Two loops and an α-helix are then build on the 2 β-strand scaffolds to create starting α-β-structures. An algorithm was then developed to specifically sample two symmetric copies of the starting structures that yielded good BB structures with room between zinc and the α-helix and β-strands. To design this algorithm, close to 3000 individual crystal structures containing zinc were analyzed to obtain optimal CPs for the BMs between zinc, His, Cys, Glu and Asp. In the designed algorithm, this was used to place the helix in a proper orientation to the β-strands that would allow a zinc site to be build. The generated structures were then inputted to the Rosetta Match algorithm, which used the optimal CPs to design a zinc site into the sampled two symmetric copies of the α-β-structures. When a zinc site is designed for both the intramonomer and intermonomer models, the zinc sites and BB structures are optimized, and a sequence is designed.

A small-scale run was used to test the above protocol. It was found that low energy sequences could be found and that at least in the case of the intramonomer, all coordinating residues to zinc had their hydrogen bonds completely satisfied.

Four problems were also found for the generated structures: 1) The rotamers were not ideal which was indicated by the high fa_dun score. 2) The structures had high solvation free energy which was indicated by fa_sol term. 3) Cavities could be seen in the structures. 4) The CP values found were far from ideal.

The structures need to be further optimized and in particular the four problems should be taken into account in future works (See Chapter 7). Two concrete steps need to be taken. The first is to carry out large-scale sampling to see if the problems indicated can be improved. The problems could stem from not having sampled the structure space or sequence space sufficiently or potentially the structures cannot be optimized sufficiently. This could be an inherent problem

with the protocol or the structures cannot be created because they are inherently unstable.

But the *de novo* design of an αβ-monomer has been shown to fold into an amyloid structure [94], and metals such as zinc have shown to function in amyloid formation [177, 185], which indicates that a stable fold might exist for the targets designed for here.

6.2 CP analysis

To the best of the authors knowledge, the CP analysis carried out in this project represents the most up to date statistical analysis of the zinc coordination sphere in the PDB with respect to the 4 residues and 9 BMs. To the authors knowledge, all of the 6 CPs for the 9 BMs have not been investigated in the PDB before. Most efforts only focus on very few parameters such as bond lengths (see for instance reference [15]), and because only few parameters have been the focus of such investigations it is hard to compare them to the literature. Although the parameters that are heavily investigated such as bond lengths (D) are very similar to previous investigations [15, 186].

The difference in BMs within the same residue and similar atoms (the ZN:ND1 and ZN:NE2, ZN:OD1 and ZN:OD2 and ZN:OE1 and ZN:OE2) is not very pronounced, which means that they could potentially be searched simultaneously in the Rosetta Match algorithm in step 2 to minimize the search time.

An analysis was done on mononuclear tetrahedral sites but no significant difference between the full dataset and the subset dataset was found. To see if it is worth modelling the CPs using only mononuclear tetrahedral sites, two different datasets, one consisting only of mononuclear tetrahedral sites and one containing all the other sites (potentially split up into different datasets according to the structural site) must be compared. A difference might be observed. For instance, binuclear sites or 6 coordination residues could have steric hindrance effects.

The analysis of zinc sites as described here used very simple metrics based on distances and angles to evaluate what atoms where considered coordinating and to remove anti-BMs for Asp and Glu. More care or metrics could be applied to evaluate the coordination sphere. Manual investigation of zinc coordination spheres has revealed many flaws associated with using such

simple approaches as done here. Zinc sites could for instance be misinterpreted based on missing electron densities or symmetry related molecules [15]. Other approaches could also be used such as Density Functional Theory (DFT) calculations, or the CP parameters could be set to ideal values such as was done for the A_A and T_A CPs. For instance, the TB parameter of His, Glu and Asp could be set 180° as is done in other studies [167, 181].

The way that the CPs are implemented in the protocol is a simplification which might pose a problem in modelling. The standard deviation is used to model the flexibility evenly around the optimal value. In step 1-2 it is used to sample the coordinating residues to zinc around their optimal values and in step 3 it is used to penalize coordination with harmonic constraints. This is a problem for modelling, in particular, the D CP of His, Glu and Asp, because the probability distributions are not evenly distributed around the optimal value. The fitted pdfs or similar pdfs could be used directly to better model the coordination sphere.

The analysis of the volume that the coordination sphere of zinc has was simply based on the distances between the CB atoms of the rotamers of the coordinating residues. To better model the zinc coordination sphere, the 3-dimensional structure of the zinc coordination sphere has to be taken into account.

6.3 Future Challenges and applications

There are many challenges in the design of the amyloid switch as described here. The high thermodynamic stability of amyloids, which is even more stable than the native state of most natural proteins [187], could potentially decrease the reversibility of the designed switch. Previous protein engineering efforts [94, 188] have had success with reversing the state under different conditions, which indicates that the switch might work.

It is also hard to control the random coil state without zinc present, for instance the protein in this state might aggregate uncontrollably. The zinc ion might also make unwanted interaction with the BB, or the proximity of all the charges from zinc might inhibit amyloid folding.

De novo design of proteins is the ultimate frontier for testing our understanding how proteins work and fold. It's a discipline that applies this understanding directly in the design process, and any *de novo* designs serves as an important feedback that can aid in improving our theories.

Furthermore, probing the "dark matter of the protein universe", gives us insight into what protein folds are possible, which in turn can contribute to our understanding of protein structure, stability, dynamics and kinetics. Even simple designs can be used to investigate these metrics [189], but in particular, large-scale screening of protein designs, such as has been done in the Baker lab [61], is a great example of how truly powerful this approach is. Such an approach could be used to understand the folding and stability of the structures described here. To this end, it is of importance to understand further how protein conformational switches work, not only because they are important in many biological roles, including disease, but also because they have important implications for technological applications, including the development of biosensors and smart materials. The work presented in this thesis might open new routes to the rational design of *de novo* switches and in particular to create switches between protein assemblies.

Amyloids are important agents in several diseases. The switch designs might serve as a model that can be used to understand amyloid folding and the mechanisms in disease. Especially the intermediate states between monomer and full fibril, are of particular importance (see section 1.4.8) but are hard to study because of the fast fibrillation of amyloids. The switch designs might be able to slow down the fibrillation process so that it can be better studied. Such an approach has also already been done with success [190]. As mentioned in section 1.4.8, there is no cure for many of the diseases associated with amyloids, and these designs might contribute to the knowledge of the role in diseases related to amyloids, and therefore to the development of better therapeutics. Finally, as mentioned, metals play a role in amyloid formation, and the designs may be able to elucidate more about the interaction between the two, in particular the kinetics.

As mentioned in section 1.4.8, amyloids have attracted significant attention as materials. To date, few *de novo* designs of amyloids have been carried out. *De novo* protein design is not limited to what nature has to offer, and efforts in expanding the potential use of amyloids using *de novo* design should have many applications. If the amyloid switches designed in this thesis work, they could be used for controlled cell compartmentalization and triggered drug delivery. Many metals could be substituted for zinc for new purposes such as in metal sensors, or for

cobber, which upon protein binding leads to blue color excitation, could be used to create new biosensors, or cobber's conducting properties could be used to build self-assembling nanoconductors. Finally, the designs could also be used for reversible super-glue, in self-healing materials such as in body armor, or increase the control of cell growth for use in tissue engineering, and many other purposes.

7 Future work

The future work that must be done is explained in 2 sections. First, the optimization and expansion of the computational design approach are explained in section 7.1. The future experimental work that must be done for the amyloid switches are explained in section 7.2.

7.1 Computational optimization and expansion

The improvements of the computational approach for designing the amyloid switches, along with the computational predictions that can be used to evaluate the designs are shown in **Figure 7-1** (step 1-4 and step Y). Approaches that adds to the current protocol are shown in green and blue.

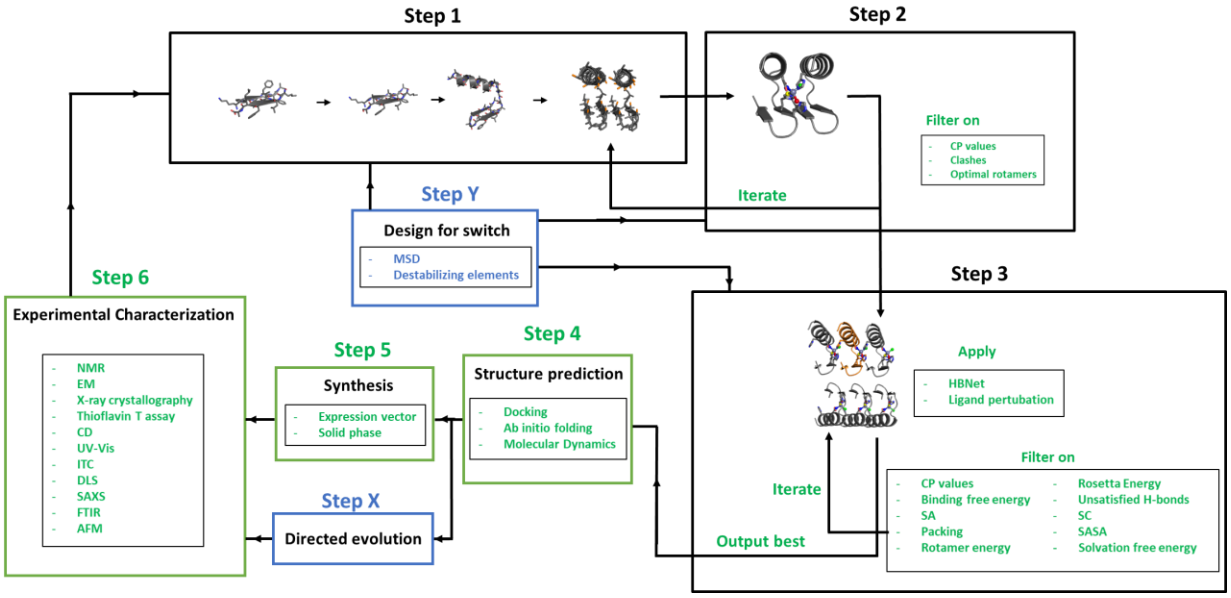


Figure 7-1: Optimization of the protocol. Step 1-3 (black boxes and text) are the steps already implemented, while step 4-6 are new steps (green boxes and text) that should be implemented in future works. Step Y and X (blue boxes and texts) are steps that can be added in the future after the green steps has been carried out (can also potentially be added together with the green steps). The design procedure starts at step 1 and then ends at step 1 again connected through step 6. See main text for a further explanation of each step.

7.1.1 Improvements to the main algorithm

The improvements of the main computational approach as outlined in the section 5.5 (step 1-3) are explained below and can be seen in **Figure 7-1**.

Step 2 of the protocol is the design of the zinc site using Rosetta Match. To avoid sampling to many non-ideal structures the outputs of the Rosetta Match algorithm must be filtered. This can be done using the CPs, such that only matches with optimal CPs within a cutoff value (for instance the standard deviation) are outputted. Clashes, for instance between the zinc and the coordinating residues as well as optimal rotamers (which was a problem identified in section 5.6) can be considered. Furthermore, an iteration between the last step in step 1 (sampling of two symmetric α-β-proteins) and step 2 can be iterated between. For instance, Rosetta Match might not find many matches, or there might be many clashes, or the CP values or rotamers might not be ideal. If this is the case a new BB should quickly be resampled from step 1. If needed the loop sampling can also be extended by searching more grid points in the Ramachandran plot (**Figure 5-8**).

Step 3 of the protocol involves zinc site optimization, BB minimization and sequence design in a symmetric environment that could describe the entire amyloid structure (to see how the symmetric system could be optimized see Appendix A.6). The algorithm in step 3 can be improved in two main ways: by optimizing the search and by applying filters to filter out "bad" structures.

The improvement of the search can happen in two ways: 1) An algorithm could be applied to specifically search for hydrogen bond networks. In section 5.6 it was shown that hydrogen bond networks could be designed for the intramonomer model, and this is a very important metric that must be considered. Furthermore, in a symmetric system, such as the amyloids designed here, hydrogen bond networks could be extended along the entire fibril. Hydrogen bond networks called "ladders" are already a strong driving force for fibrillation of amyloids [31, 191], and furthermore the destruction and reconstruction of hydrogen bond networks by zinc binding could be a switch mechanism that could perturb the two states. An algorithm has recently been developed (and implemented in Rosetta) to search for hydrogen bond networks. It is called HBNet [192], and it could be implemented in into the main algorithm. Furthermore, the zinc

ligand could be perturbed at different points throughout the trajectory to avoid local energy minima trapping. Many filters can be applied to quickly filter bad trajectories in the protocol as well as evaluating the best structures. Metrics that can be applied to both optimize the zinc geometry, binding forces and interface energy are the following: The solvent-accessible surface area (SASA), the binding free energy, surface area (SA) and surface complementarity (SC) between the individual monomers of the amyloid, unsatisfied hydrogen bonds, packing, rotamer energy and solvation free energy (the latter 3 were also problems identified in section 5.6).

Lastly, it is important to note that the entire protocol (step 1-3) should be automated and more sampling needs to be done to increase the chances of finding low energy structures. More sampling was one of the main points discussed in section 6.1. Furthermore, there should be a connection between structural characterization and design. Such that the structures that are experimentally characterized can be used to optimize the design (shown as the arrow leading to step 1 in **Figure 7-1**).

7.1.2 Structure validation

To validate the generated structures, 3 computer simulations can be carried out (**Figure 7-1**, step 4). First docking, such as RosettaLigand docking [193], could be used to see if zinc does bind to the generated structures. Secondly, structure prediction methods could be used to evaluate that the amyloid structure do indeed fold into the structures designed. This could be very hard to implement. The amyloid system described here is very complex. It consists of many individual monomers, and each monomer binds to a zinc atom. Normally, without relying on information such as co-evolution-based constraints, only small monomeric proteins can be predicted accurately using *ab initio* structure prediction methods [54, 100, 194]. However, methods for simultaneous folding and docking of symmetric proteins [195] and metal proteins [181] have been carried out, and these might be methods that could be combined to predict folding of the structures described here.

Lastly, molecular dynamics (MD) methods could be used to investigate the stability of the amyloid structures, or to get information on other metrics.

7.1.3 Designing the β-L-β-L-α and β-L-α-L-β monomers and extending the loop

The design of the β-L-β-L-α monomers is the same as for the α-L-β-L-β monomers described earlier, although the α-helix should be extended from the C-terminal instead. The design of the β-L-α-L-β monomers are slightly different. Loop building could happen in the first step as for the α-L-β-L-β described, but it is probably more ideal to build it after the zinc site has been inserted because the loops are extended from the helix, and if build before it might inhibit the downstream design of the zinc site (although that could be implemented as constraints in the loop building algorithm). The ideal loops build in section 5.6 might not be applicable to these structures either. New loop building approaches might have to be sought out such as Kinematic loop closure [196]. Except for loop building, the other steps for designing the β-L-α-L-β monomers should be similar to the α-L-β-L-β monomers.

7.1.4 Expansion of the approach

There are many ways the design can be expanded. The interfaces of the SZ could be redesigned instead of being kept as is from the crystal structure. An MSD approach could be implemented to design switches between distinct states, or to make sure that certain structures are not designed for (negative design) (**Figure 7-1** step Y). Other metals could be implemented or even other ligands using the same strategy described here. Functional components could also potentially be designed into the structures deliberately. To further probe "the dark matter universe" other folds could also be tried out. If the amyloid structures are too stabile, destabilizing elements can also be designed into the amyloid structures for it to easily convert to the unfolded state (**Figure 7-1** step Y).

Rosetta Match is one of many approaches to computationally design ligand binding proteins [197]. Other approaches could be used as well. Ligand docking could for instance be combined with BB sampling and sequence design.

Directed evolution approaches (**Figure 7-1** step X) could also be used to optimize the designs. Designs could be expressed in yeast (for instance in conjunction with yeast surface display or other methods such as phage display could be used) and the thioflavin T assay (or other assays) could be combined with FACS to screen for designs able to form amyloid structures

(fluorescent signal). EDTA could then be applied and designs able to switch to the random state from the amyloid state could be screened for (loss of fluorescent signal).

7.2 Structural characterization

The designed structures can either be made by solid-phase synthesis or using expression vectors (**Figure 7-1,** step 5). For designing many different structures, expression vectors are usually preferred. Several structural analyses can be carried out (**Figure 7-1**, step 6). Initially in the design process, experimental methods that can quickly search for correctly designed structures or switch mechanisms would be ideal. A very fast assay that could be used to investigate which structures respond to zinc, or unfold when zinc is removed, with for instance EDTA, is the thioflavin T fluorescence assay [198]. The structures that respond to zinc can be used for further optimization. Only correctly folded structures could be further investigated (and the switch method could be computationally designed by incorporating destabilizing elements into the structures later (**Figure 7-1** step Y)) or structures that also unfold when zinc is removed could be investigated. The Thioflavin T assay does not directly verify zinc binding, and other experiments must be carried out to determine that zinc binds. Isothermal titration calorimetry (ITC) or ultraviolet–visible spectroscopy (UV-vis) methods could be applied to detect zinc binding. Structural experiments could then be carried out. Methods such as Fourier transform infrared spectroscopy (FTIR) or circular dichroism (CD) could be used to detect that the secondary structures form correctly, and fiber diffraction methods can be used to detect that the SZ and the general amyloid structure forms. Transmission electron microscopy (TEM), small angle x-ray scattering (SAXS), dynamic light scattering (DLS) and atomic force microscopy (AFM) can be used to detect overall morphologically and size. Lastly nuclear magnetic resonance (NMR), cryo-EM and X-ray crystallography can be used to determine the atomic structure of the designed amyloids. All the amyloid structures used in this project have been determined with X-ray crystallography, but solid-state NMR (ssNMR) [199] and cryo-EM [200] has also been used to obtain high resolution structures of amyloids.

8 Conclusion

A procedure for generating de novo designed amyloid structures consisting of symmetrical αβ-monomers and containing zinc sites was developed. To create this procedure, an up to date analysis of zinc sites deposited the Protein Data Bank were carried out to obtain probability distributions of 6 different coordination parameters for 9 different binding modes of histidine, cysteine, aspartate and glutamate to zinc. A new algorithm was also developed to specifically design the amyloid structures. The procedure was tested using a single starting structure and small-scale sampling. The test showed that a zinc site could be designed into the αβ-monomers while these were assembled into an amyloid structure. Two different types of amyloid structures were generated in the test. One amyloid structure consisted of an αβ-monomer where the zinc site was placed in between the monomers, while the other amyloid structure had the zinc site within the monomers. It was found that for both models, low energy sequences could be designed, and in one of the models all coordinating ligands to zinc had completely satisfied their hydrogen bonds. Challenges were also identified in the test. The structures had non-optimal rotamers, solvation free energy and coordination to zinc. Furthermore, large cavities were also observed in the models. These findings indicated that the structure needed to be optimized. The first step in designing de novo designed switches, the main goal of this thesis, has been taken, although the computational procedure must be expanded, and large-scale sampling must be carried out to further optimize the structures. When satisfiable structures have been designed these must be experimentally characterized. *De novo* designed amyloid switches are attractive for both fundamental research and technological applications. They could potentially be used to understand amyloid diseases, for which there are no cure, or be used as materials in a range of different application from the macro scale to the nano scale.

References

[1] J.G. Zalatan, D. Herschlag, The far reaches of enzymology, Nature Chemical Biology 5(8) (2009) 516-520.
[2] T. Pawson, Protein Modules and Signaling Networks, Nature 373(6515) (1995) 573-580.
[3] C.R.D. Lancaster, Structural biology - Ion pump in the movies, Nature 432(7015) (2004) 286-287.
[4] R.D. Vale, R.A. Milligan, The way things move: Looking under the hood of molecular motor proteins, Science 288(5463) (2000) 88-95.
[5] T. Vignaud, L. Blanchoin, M. Thery, Directed cytoskeleton self-organization, Trends in Cell Biology 22(12) (2012) 671-682.
[6] J.S. Richardson, The Anatomy and Taxonomy of Protein Structure, Advances in Protein Chemistry Volume 341981, pp. 167-339.
[7] G.N. Ramachandran, C. Ramakrishnan, V. Sasisekharan, Stereochemistry of polypeptide chain configurations, J Mol Biol 7 (1963) 95-9.
[8] L. Pauling, R.B. Corey, H.R. Branson, The Structure of Proteins - 2 Hydrogen-Bonded Helical Configurations of the Polypeptide Chain, P Natl Acad Sci USA 37(4) (1951) 205-211.
[9] L. Pauling, R.B. Corey, Configurations of Polypeptide Chains with Favored Orientations around Single Bonds - 2 New Pleated Sheets, P Natl Acad Sci USA 37(11) (1951) 729-740.
[10] A.J. Thomson, H.B. Gray, Bio-inorganic chemistry, Current Opinion in Chemical Biology 2(2) (1998) 155-158.
[11] V. Putignano, A. Rosato, L. Banci, C. Andreini, MetalPDB in 2018: a database of metal sites in biological macromolecular structures, Nucleic Acids Res 46(D1) (2018) D459-D464.
[12] H. Irving, R.J.P. Williams, The Stability of Transition-Metal Complexes, J Chem Soc (Oct) (1953) 3192-3210.
[13] C. Andreini, L. Banci, I. Bertini, A. Rosato, Counting the zinc-proteins encoded in the human genome, J Proteome Res 5(1) (2006) 196-201.
[14] D.S. Auld, Zinc coordination sphere in biochemical zinc sites, Biometals 14(3-4) (2001) 271-313.
[15] M. Laitaoja, J. Valjakka, J. Janis, Zinc coordination spheres in protein structures, Inorg Chem 52(19) (2013) 10983-91.
[16] M.M. Harding, M.W. Nowicki, M.D. Walkinshaw, Metals in protein structures: a review of their principal features, Crystallography Reviews 16(4) (2010) 247-302.
[17] S.F. Sousa, A.B. Lopes, P.A. Fernandes, M.J. Ramos, The Zinc proteome: a tale of stability and functionality, Dalton Trans (38) (2009) 7946-56.
[18] C. Andreini, I. Bertini, G. Cavallaro, Minimal functional sites allow a classification of zinc sites in proteins, PLoS One 6(10) (2011) e26325.
[19] C. Andreini, I. Bertini, G. Cavallaro, R.J. Najmanovich, J.M. Thornton, Structural analysis of metal sites in proteins: non-heme iron sites as a case study, J Mol Biol 388(2) (2009) 356-80.
[20] E.H. Cox, G.L. McLendon, Zinc-dependent protein folding, Curr Opin Chem Biol 4(2) (2000) 162-5.
[21] G. Parraga, S. Horvath, L. Hood, E.T. Young, R.E. Klevit, Spectroscopic studies of wild-type and mutant "zinc finger" peptides: determinants of domain folding and structure, Proc Natl Acad Sci U S A 87(1) (1990) 137-41.
[22] G. Parraga, S.J. Horvath, A. Eisen, W.E. Taylor, L. Hood, E.T. Young, R.E. Klevit, Zinc-dependent structure of a single-finger domain of yeast ADR1, Science 241(4872) (1988) 1489-92.
[23] L. Regan, Protein Design - Novel Metal-Binding Sites, Trends Biochem Sci 20(7) (1995) 280-285.
[24] W. Li, J. Zhang, J. Wang, W. Wang, Metal-coupled folding of Cys2His2 zinc-finger, J Am Chem Soc 130(3) (2008) 892-900.

[25] C. Nick Pace, J.M. Scholtz, G.R. Grimsley, Forces stabilizing proteins, FEBS Lett 588(14) (2014) 2177-84.
[26] K.A. Dill, Dominant forces in protein folding, Biochemistry 29(31) (1990) 7133-55.
[27] F. Polticelli, G. Raybaudi-Massilia, P. Ascenzi, Structural determinants of mini-protein stability, Biochem Mol Biol Edu 29(1) (2001) 16-20.
[28] E.G. Baker, G.J. Bartlett, K.L. Porter Goff, D.N. Woolfson, Miniprotein Design: Past, Present, and Prospects, Acc Chem Res 50(9) (2017) 2085-2092.
[29] S.S. Krishna, I. Majumdar, N.V. Grishin, Structural classification of zinc fingers: survey and summary, Nucleic Acids Res 31(2) (2003) 532-50.
[30] J. Miller, A.D. McLachlan, A. Klug, Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes, EMBO J 4(6) (1985) 1609-14.
[31] D.S. Eisenberg, M.R. Sawaya, Structural Studies of Amyloid Proteins at the Molecular Level, Annu Rev Biochem 86 (2017) 69-95.
[32] M. Schleeger, C.C. vandenAkker, T. Deckert-Gaudig, V. Deckert, K.P. Velikov, G. Koenderink, M. Bonn, Amyloids: From molecular structure to mechanical properties, Polymer 54(10) (2013) 2473-2488.
[33] R. Nelson, M.R. Sawaya, M. Balbirnie, A.O. Madsen, C. Riekel, R. Grothe, D. Eisenberg, Structure of the cross-beta spine of amyloid-like fibrils, Nature 435(7043) (2005) 773-8.
[34] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucleic Acids Res 28(1) (2000) 235-42.
[35] M.R. Sawaya, S. Sambashivan, R. Nelson, M.I. Ivanova, S.A. Sievers, M.I. Apostol, M.J. Thompson, M. Balbirnie, J.J. Wiltzius, H.T. McFarlane, A.O. Madsen, C. Riekel, D. Eisenberg, Atomic structures of amyloid cross-beta spines reveal varied steric zippers, Nature 447(7143) (2007) 453-7.
[36] P. Carter, Site-Directed Mutagenesis, Biochem J 237(1) (1986) 1-7.
[37] C.A. Hutchison, S. Phillips, M.H. Edgell, S. Gillam, P. Jahnke, M. Smith, Mutagenesis at a Specific Position in a DNA-Sequence, Journal of Biological Chemistry 253(18) (1978) 6551-6560.
[38] J.A. Brannigan, A.J. Wilkinson, Protein engineering 20 years on, Nat Rev Mol Cell Biol 3(12) (2002) 964-70.
[39] M.S. Packer, D.R. Liu, Methods for the directed evolution of proteins, Nat Rev Genet 16(7) (2015) 379-94.
[40] I. Samish, The Framework of Computational Protein Design, Methods Mol Biol 1529 (2017) 3-19.
[41] D.A. Estell, T.P. Graycar, J.A. Wells, Engineering an Enzyme by Site-Directed Mutagenesis to Be Resistant to Chemical Oxidation, Journal of Biological Chemistry 260(11) (1985) 6518-6521.
[42] L. Vojcic, C. Pitzler, G. Korfer, F. Jakob, R. Martinez, K.H. Maurer, U. Schwaneberg, Advances in protease engineering for laundry detergents, New Biotechnol 32(6) (2015) 629-634.
[43] V.J. Jensen, S. Rugh, Industrial-Scale Production and Application of Immobilized Glucose-Isomerase, Method Enzymol 136 (1987) 356-370.
[44] A. Madhu, J.N. Chakraborty, Developments in application of enzymes for textile processing, J Clean Prod 145 (2017) 114-133.
[45] P.H. Tobin, D.H. Richards, R.A. Callender, C.J. Wilson, Protein Engineering: A New Frontier for Biological Therapeutics, Curr Drug Metab 15(7) (2014) 743-756.
[46] C. Chothia, Proteins - 1000 Families for the Molecular Biologist, Nature 357(6379) (1992) 543-544.
[47] A.F.W. Coulson, J. Moult, A unifold, mesofold, and superfold model of protein fold use, Proteins-Structure Function and Genetics 46(1) (2002) 61-71.
[48] R. Kolodny, L. Pereyaslavets, A.O. Samson, M. Levitt, On the Universe of Protein Folds, Annual Review of Biophysics, Vol 42 42 (2013) 559-582.
[49] W.R. Taylor, V. Chelliah, S.M. Hollup, J.T. MacDonald, I. Jonassen, Probing the "dark matter" of protein fold space, Structure 17(9) (2009) 1244-52.

[50] C. Chiarabelli, D. De Lucrezia, P. Stano, P.L. Luisi, The World of the "Never Born Proteins", Origins Life Evol B 39(3-4) (2009) 308-309.
[51] P. Cossio, A. Trovato, F. Pietrucci, F. Seno, A. Maritan, A. Laio, Exploring the universe of protein structures beyond the Protein Data Bank, PLoS Comput Biol 6(11) (2010) e1000957.
[52] D.N. Woolfson, G.J. Bartlett, A.J. Burton, J.W. Heal, A. Niitsu, A.R. Thomson, C.W. Wood, De novo protein design: how do we expand into the universe of possible protein structures?, Curr Opin Struct Biol 33 (2015) 16-26.
[53] O. Alvizo, B.D. Allen, S.L. Mayo, Computational protein design promises to revolutionize protein engineering, Biotechniques 42(1) (2007) 31-+.
[54] P.S. Huang, S.E. Boyken, D. Baker, The coming of age of de novo protein design, Nature 537(7620) (2016) 320-7.
[55] A.D. Keefe, J.W. Szostak, Functional proteins from a random-sequence library, Nature 410(6829) (2001) 715-718.
[56] M.A. Fisher, K.L. McKinley, L.H. Bradley, S.R. Viola, M.H. Hecht, De Novo Designed Proteins from a Library of Artificial Sequences Function in Escherichia Coli and Enable Cell Growth, Plos One 6(1) (2011).
[57] B. Kuhlman, G. Dantas, G.C. Ireton, G. Varani, B.L. Stoddard, D. Baker, Design of a novel globular protein fold with atomic-level accuracy, Science 302(5649) (2003) 1364-8.
[58] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T.B. Acton, G.T. Montelione, D. Baker, Principles for designing ideal protein structures, Nature 491(7423) (2012) 222-7.
[59] Y.R. Lin, N. Koga, R. Tatsumi-Koga, G. Liu, A.F. Clouser, G.T. Montelione, D. Baker, Control over overall shape and size in de novo designed proteins, Proc Natl Acad Sci U S A 112(40) (2015) E5478-85.
[60] G. Bhardwaj, V.K. Mulligan, C.D. Bahl, J.M. Gilmore, P.J. Harvey, O. Cheneval, G.W. Buchko, S.V. Pulavarti, Q. Kaas, A. Eletsky, P.S. Huang, W.A. Johnsen, P.J. Greisen, G.J. Rocklin, Y. Song, T.W. Linsky, A. Watkins, S.A. Rettie, X. Xu, L.P. Carter, R. Bonneau, J.M. Olson, E. Coutsias, C.E. Correnti, T. Szyperski, D.J. Craik, D. Baker, Accurate de novo design of hyperstable constrained peptides, Nature 538(7625) (2016) 329-335.
[61] G.J. Rocklin, T.M. Chidyausiku, I. Goreshnik, A. Ford, S. Houliston, A. Lemak, L. Carter, R. Ravichandran, V.K. Mulligan, A. Chevalier, C.H. Arrowsmith, D. Baker, Global analysis of protein folding using massively parallel design, synthesis, and testing, Science 357(6347) (2017) 168-175.
[62] E. Marcos, B. Basanta, T.M. Chidyausiku, Y. Tang, G. Oberdorfer, G. Liu, G.V. Swapna, R. Guan, D.A. Silva, J. Dou, J.H. Pereira, R. Xiao, B. Sankaran, P.H. Zwart, G.T. Montelione, D. Baker, Principles for designing proteins with cavities formed by curved beta sheets, Science 355(6321) (2017) 201-206.
[63] S. Ramisch, U. Weininger, J. Martinsson, M. Akke, I. Andre, Computational design of a leucine-rich repeat protein with a predefined geometry, Proc Natl Acad Sci U S A 111(50) (2014) 17875-80.
[64] T.J. Brunette, F. Parmeggiani, P.S. Huang, G. Bhabha, D.C. Ekiert, S.E. Tsutakawa, G.L. Hura, J.A. Tainer, D. Baker, Exploring the repeat protein universe through computational protein design, Nature 528(7583) (2015) 580-4.
[65] L. Doyle, J. Hallinan, J. Bolduc, F. Parmeggiani, D. Baker, B.L. Stoddard, P. Bradley, Rational design of alpha-helical tandem repeat proteins with closed architectures, Nature 528(7583) (2015) 585-8.
[66] K. Park, B.W. Shen, F. Parmeggiani, P.S. Huang, B.L. Stoddard, D. Baker, Control of repeat-protein curvature by computational protein design, Nat Struct Mol Biol 22(2) (2015) 167-74.
[67] F. Parmeggiani, P.S. Huang, S. Vorobiev, R. Xiao, K. Park, S. Caprari, M. Su, J. Seetharaman, L. Mao, H. Janjua, G.T. Montelione, J. Hunt, D. Baker, A general computational approach for repeat protein design, J Mol Biol 427(2) (2015) 563-75.
[68] P.S. Huang, K. Feldmeier, F. Parmeggiani, D.A.F. Velasco, B. Hocker, D. Baker, De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy, Nat Chem Biol 12(1) (2016) 29-34.
[69] D. Nagarajan, G. Deka, M. Rao, Design of symmetric TIM barrel proteins from first principles, BMC Biochem 16 (2015) 18.

[70] P.S. Huang, G. Oberdorfer, C. Xu, X.Y. Pei, B.L. Nannenga, J.M. Rogers, F. DiMaio, T. Gonen, B. Luisi, D. Baker, High thermodynamic stability of parametrically designed helical bundles, Science 346(6208) (2014) 481-485.
[71] A.R. Thomson, C.W. Wood, A.J. Burton, G.J. Bartlett, R.B. Sessions, R.L. Brady, D.N. Woolfson, Computational design of water-soluble alpha-helical barrels, Science 346(6208) (2014) 485-8.
[72] A. Chevalier, D.A. Silva, G.J. Rocklin, D.R. Hicks, R. Vergara, P. Murapa, S.M. Bernard, L. Zhang, K.H. Lam, G. Yao, C.D. Bahl, S.I. Miyashita, I. Goreshnik, J.T. Fuller, M.T. Koday, C.M. Jenkins, T. Colvin, L. Carter, A. Bohn, C.M. Bryan, D.A. Fernandez-Velasco, L. Stewart, M. Dong, X. Huang, R. Jin, I.A. Wilson, D.H. Fuller, D. Baker, Massively parallel de novo protein design for targeted therapeutics, Nature 550(7674) (2017) 74-79.
[73] P. Lu, D. Min, F. DiMaio, K.Y. Wei, M.D. Vahey, S.E. Boyken, Z. Chen, J.A. Fallas, G. Ueda, W. Sheffler, V.K. Mulligan, W. Xu, J.U. Bowie, D. Baker, Accurate computational design of multipass transmembrane proteins, Science 359(6379) (2018) 1042-1046.
[74] N.H. Joh, T. Wang, M.P. Bhate, R. Acharya, Y. Wu, M. Grabe, M. Hong, G. Grigoryan, W.F. DeGrado, De novo design of a transmembrane Zn(2)(+)-transporting four-helix bundle, Science 346(6216) (2014) 1520-4.
[75] G. Grigoryan, Y.H. Kim, R. Acharya, K. Axelrod, R.M. Jain, L. Willis, M. Drndic, J.M. Kikkawa, W.F. DeGrado, Computational design of virus-like protein assemblies on carbon nanotube surfaces, Science 332(6033) (2011) 1071-6.
[76] N.F. Polizzi, Y. Wu, T. Lemmin, A.M. Maxwell, S.Q. Zhang, J. Rawson, D.N. Beratan, M.J. Therien, W.F. DeGrado, De novo design of a hyperstable non-natural protein-ligand complex with sub-A accuracy, Nat Chem 9(12) (2017) 1157-1164.
[77] A.J. Burton, A.R. Thomson, W.M. Dawson, R.L. Brady, D.N. Woolfson, Installing hydrolytic activity into a completely de novo protein framework, Nat Chem 8(9) (2016) 837-44.
[78] F. Thomas, N.C. Burgess, A.R. Thomson, D.N. Woolfson, Controlling the Assembly of Coiled-Coil Peptide Nanotubes, Angew Chem Int Ed Engl 55(3) (2016) 987-91.
[79] N.C. Burgess, T.H. Sharp, F. Thomas, C.W. Wood, A.R. Thomson, N.R. Zaccai, R.L. Brady, L.C. Serpell, D.N. Woolfson, Modular Design of Self-Assembling Peptide-Based Nanotubes, J Am Chem Soc 137(33) (2015) 10554-62.
[80] J.M. Fletcher, R.L. Harniman, F.R. Barnes, A.L. Boyle, A. Collins, J. Mantell, T.H. Sharp, M. Antognozzi, P.J. Booth, N. Linden, M.J. Miles, R.B. Sessions, P. Verkade, D.N. Woolfson, Self-assembling cages from coiled-coil peptide modules, Science 340(6132) (2013) 595-9.
[81] C.E. Tinberg, S.D. Khare, J. Dou, L. Doyle, J.W. Nelson, A. Schena, W. Jankowski, C.G. Kalodimos, K. Johnsson, B.L. Stoddard, D. Baker, Computational design of ligand-binding proteins with high affinity and selectivity, Nature 501(7466) (2013) 212-216.
[82] D. Rothlisberger, O. Khersonsky, A.M. Wollacott, L. Jiang, J. DeChancie, J. Betker, J.L. Gallaher, E.A. Althoff, A. Zanghellini, O. Dym, S. Albeck, K.N. Houk, D.S. Tawfik, D. Baker, Kemp elimination catalysts by computational enzyme design, Nature 453(7192) (2008) 190-5.
[83] S. Gonen, F. DiMaio, T. Gonen, D. Baker, Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces, Science 348(6241) (2015) 1365-8.
[84] N.P. King, J.B. Bale, W. Sheffler, D.E. McNamara, S. Gonen, T. Gonen, T.O. Yeates, D. Baker, Accurate design of co-assembling multi-component protein nanomaterials, Nature 510(7503) (2014) 103-8.
[85] J.H. Mills, S.D. Khare, J.M. Bolduc, F. Forouhar, V.K. Mulligan, S. Lew, J. Seetharaman, L. Tong, B.L. Stoddard, D. Baker, Computational design of an unnatural amino acid dependent metalloprotein with atomic level accuracy, J Am Chem Soc 135(36) (2013) 13393-9.
[86] J.B. Siegel, A. Zanghellini, H.M. Lovick, G. Kiss, A.R. Lambert, J.L. St Clair, J.L. Gallaher, D. Hilvert, M.H. Gelb, B.L. Stoddard, K.N. Houk, F.E. Michael, D. Baker, Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction, Science 329(5989) (2010) 309-13.

[87] L. Jiang, E.A. Althoff, F.R. Clemente, L. Doyle, D. Rothlisberger, A. Zanghellini, J.L. Gallaher, J.L. Betker, F. Tanaka, C.F. Barbas, 3rd, D. Hilvert, K.N. Houk, B.L. Stoddard, D. Baker, De novo computational design of retro-aldol enzymes, Science 319(5868) (2008) 1387-91.
[88] G. Kiss, N. Celebi-Olcum, R. Moretti, D. Baker, K.N. Houk, Computational enzyme design, Angew Chem Int Ed Engl 52(22) (2013) 5700-25.
[89] S.D. Khare, Y. Kipnis, P. Greisen, Jr., R. Takeuchi, Y. Ashani, M. Goldsmith, Y. Song, J.L. Gallaher, I. Silman, H. Leader, J.L. Sussman, B.L. Stoddard, D.S. Tawfik, D. Baker, Computational redesign of a mononuclear zinc metalloenzyme for organophosphate hydrolysis, Nat Chem Biol 8(3) (2012) 294-300.
[90] S.R. Gordon, E.J. Stanley, S. Wolf, A. Toland, S.J. Wu, D. Hadidi, J.H. Mills, D. Baker, I.S. Pultz, J.B. Siegel, Computational design of an alpha-gliadin peptidase, J Am Chem Soc 134(50) (2012) 20513-20.
[91] J.B. Siegel, A.L. Smith, S. Poust, A.J. Wargacki, A. Bar-Even, C. Louw, B.W. Shen, C.B. Eiben, H.M. Tran, E. Noor, J.L. Gallaher, J. Bale, Y. Yoshikuni, M.H. Gelb, J.D. Keasling, B.L. Stoddard, M.E. Lidstrom, D. Baker, Computational protein design enables a novel one-carbon assimilation pathway, Proc Natl Acad Sci U S A 112(12) (2015) 3704-9.
[92] M.J. Bick, P.J. Greisen, K.J. Morey, M.S. Antunes, D. La, B. Sankaran, L. Reymond, K. Johnsson, J.I. Medford, D. Baker, Computational design of environmental sensors for the potent opioid fentanyl, Elife 6 (2017).
[93] J.B. Bale, S. Gonen, Y. Liu, W. Sheffler, D. Ellis, C. Thomas, D. Cascio, T.O. Yeates, T. Gonen, N.P. King, D. Baker, Accurate design of megadalton-scale two-component icosahedral protein complexes, Science 353(6297) (2016) 389-94.
[94] S. Kaltofen, C. Li, P.S. Huang, L.C. Serpell, A. Barth, I. Andre, Computational de novo design of a self-assembling peptide with predefined structure, J Mol Biol 427(2) (2015) 550-62.
[95] M.D. Peralta, A. Karsai, A. Ngo, C. Sierra, K.T. Fong, N.R. Hayre, N. Mirzaee, K.M. Ravikumar, A.J. Kluber, X. Chen, G.Y. Liu, M.D. Toney, R.R. Singh, D.L. Cox, Engineering amyloid fibrils from beta-solenoid proteins for biomaterials applications, ACS Nano 9(1) (2015) 449-63.
[96] C.J. Lanci, C.M. MacDermaid, S.G. Kang, R. Acharya, B. North, X. Yang, X.J. Qiu, W.F. DeGrado, J.G. Saven, Computational design of a protein crystal, Proc Natl Acad Sci U S A 109(19) (2012) 7304-9.
[97] G.L. Butterfield, M.J. Lajoie, H.H. Gustafson, D.L. Sellers, U. Nattermann, D. Ellis, J.B. Bale, S. Ke, G.H. Lenz, A. Yehdego, R. Ravichandran, S.H. Pun, N.P. King, D. Baker, Evolution of a designed protein assembly encapsulating its own RNA genome, Nature 552(7685) (2017) 415-420.
[98] P.B. Stranges, B. Kuhlman, A comparison of successful and failed protein interface designs highlights the challenges of designing buried hydrogen bonds, Protein Sci 22(1) (2013) 74-82.
[99] J. Karanicolas, B. Kuhlman, Computational design of affinity and specificity at protein-protein interfaces, Curr Opin Struc Biol 19(4) (2009) 458-463.
[100] K.A. Dill, J.L. MacCallum, The protein-folding problem, 50 years on, Science 338(6110) (2012) 1042-6.
[101] R. Zwanzig, A. Szabo, B. Bagchi, Levinthal's paradox, Proc Natl Acad Sci U S A 89(1) (1992) 20-2.
[102] K.A. Dill, H.S. Chan, From Levinthal to pathways to funnels, Nat Struct Biol 4(1) (1997) 10-9.
[103] C.B. Anfinsen, Principles That Govern Folding of Protein Chains, Science 181(4096) (1973) 223-230.
[104] C.B. Anfinsen, E. Haber, M. Sela, F.H. White, Jr., The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, Proc Natl Acad Sci U S A 47 (1961) 1309-14.
[105] J.D. Bryngelson, P.G. Wolynes, Intermediates and Barrier Crossing in a Random Energy-Model (with Applications to Protein Folding), J Phys Chem-Us 93(19) (1989) 6902-6915.
[106] J.D. Bryngelson, P.G. Wolynes, Spin glasses and the statistical mechanics of protein folding, Proc Natl Acad Sci U S A 84(21) (1987) 7524-8.
[107] K. Henzler-Wildman, D. Kern, Dynamic personalities of proteins, Nature 450(7172) (2007) 964-972.
[108] R. Kapon, R. Nevo, Z. Reich, Protein energy landscape roughness, Biochem Soc Trans 36(Pt 6) (2008) 1404-8.

[109] P.E. Leopold, M. Montal, J.N. Onuchic, Protein Folding Funnels - a Kinetic Approach to the Sequence Structure Relationship, P Natl Acad Sci USA 89(18) (1992) 8721-8725.
[110] J.N. Onuchic, P.G. Wolynes, Theory of protein folding, Curr Opin Struc Biol 14(1) (2004) 70-75.
[111] D.U. Ferreiro, E.A. Komives, P.G. Wolynes, Frustration in biomolecules, Q Rev Biophys 47(4) (2014) 285-363.
[112] R. Nussinov, C.J. Tsai, Free energy diagrams for protein function, Chem Biol 21(3) (2014) 311-8.
[113] J.A. Davey, R.A. Chica, Multistate approaches in computational protein design, Protein Sci 21(9) (2012) 1241-52.
[114] K. Yue, K.A. Dill, Inverse protein folding problem: designing polymer sequences, Proc Natl Acad Sci U S A 89(9) (1992) 4163-7.
[115] P. Loffler, S. Schmitz, E. Hupfeld, R. Sterner, R. Merkl, Rosetta:MSF: a modular framework for multi-state computational protein design, PLoS Comput Biol 13(6) (2017) e1005600.
[116] B.D. Allen, S.L. Mayo, An efficient algorithm for multistate protein design based on FASTER, J Comput Chem 31(5) (2010) 904-16.
[117] I. Samish, C.M. MacDermaid, J.M. Perez-Aguilar, J.G. Saven, Theoretical and computational protein design, Annu Rev Phys Chem 62 (2011) 129-49.
[118] G.H. Wei, W.H. Xi, R. Nussinov, B.Y. Ma, Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell, Chemical Reviews 116(11) (2016) 6516-6551.
[119] P. Gainza, K.E. Roberts, B.R. Donald, Protein Design Using Continuous Rotamers, Plos Computational Biology 8(1) (2012).
[120] C.A. Smith, T. Kortemme, Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction, Journal of Molecular Biology 380(4) (2008) 742-756.
[121] J.A. Davey, R.A. Chica, Improving the accuracy of protein stability predictions with multistate design using a variety of backbone ensembles, Proteins 82(5) (2014) 771-84.
[122] X.I. Ambroggio, B. Kuhlman, Design of protein conformational switches, Curr Opin Struct Biol 16(4) (2006) 525-30.
[123] J.H. Ha, S.N. Loh, Protein conformational switches: from nature to design, Chemistry 18(26) (2012) 7984-99.
[124] V. Stein, K. Alexandrov, Synthetic protein switches: design principles and applications, Trends in Biotechnology 33(2) (2015) 101-110.
[125] X.I. Ambroggio, B. Kuhlman, Computational design of a single amino acid sequence that can switch between two distinct protein folds, Journal of the American Chemical Society 128(4) (2006) 1154-1161.
[126] R. Lizatovic, O. Aurelius, O. Stenstrom, T. Drakenberg, M. Akke, D.T. Logan, I. Andre, A De Novo Designed Coiled-Coil Peptide with a Reversible pH-Induced Oligomerization Switch, Structure 24(6) (2016) 946-55.
[127] C.H. Norn, I. Andre, Computational design of protein self-assembly, Curr Opin Struct Biol 39 (2016) 39-45.
[128] E.N. Salgado, X.I. Ambroggio, J.D. Brodin, R.A. Lewis, B. Kuhlman, F.A. Tezcan, Metal templated design of protein interfaces, Proc Natl Acad Sci U S A 107(5) (2010) 1827-32.
[129] B.S. Der, B. Kuhlman, Strategies to control the binding mode of de novo designed protein interactions, Curr Opin Struct Biol 23(4) (2013) 639-46.
[130] B.S. Der, M. Machius, M.J. Miley, J.L. Mills, T. Szyperski, B. Kuhlman, Metal-mediated affinity and orientation specificity in a computationally designed protein homodimer, J Am Chem Soc 134(1) (2012) 375-85.

[131] J.D. Brodin, X.I. Ambroggio, C.Y. Tang, K.N. Parent, T.S. Baker, F.A. Tezcan, Metal-directed, chemically tunable assembly of one-, two- and three-dimensional crystalline protein arrays, Nature Chemistry 4(5) (2012) 375-382.
[132] D.S. Goodsell, A.J. Olson, Structural symmetry and protein function, Annu Rev Biophys Biomol Struct 29 (2000) 105-53.
[133] C.M. Dobson, Protein folding and misfolding, Nature 426(6968) (2003) 884-90.
[134] T.P. Knowles, M. Vendruscolo, C.M. Dobson, The amyloid state and its association with protein misfolding diseases, Nat Rev Mol Cell Biol 15(6) (2014) 384-96.
[135] N. Cremades, C.M. Dobson, The contribution of biophysical and structural studies of protein self-assembly to the design of therapeutic strategies for amyloid diseases, Neurobiol Dis 109(Pt B) (2018) 178-190.
[136] M.R. Chapman, L.S. Robinson, J.S. Pinkner, R. Roth, J. Heuser, M. Hammar, S. Normark, S.J. Hultgren, Role of Escherichia coli curli operons in directing amyloid fiber formation, Science 295(5556) (2002) 851-5.
[137] K. Si, S. Lindquist, E.R. Kandel, A neuronal isoform of the aplysia CPEB has prion-like properties, Cell 115(7) (2003) 879-91.
[138] D.M. Fowler, A.V. Koulov, C. Alory-Jost, M.S. Marks, W.E. Balch, J.W. Kelly, Functional amyloid formation within mammalian tissue, PLoS Biol 4(1) (2006) e6.
[139] K. Tsemekhman, L. Goldschmidt, D. Eisenberg, D. Baker, Cooperative hydrogen bonding in amyloid formation, Protein Sci 16(4) (2007) 761-4.
[140] J.F. Smith, T.P. Knowles, C.M. Dobson, C.E. Macphee, M.E. Welland, Characterization of the nanoscale properties of individual amyloid fibrils, Proc Natl Acad Sci U S A 103(43) (2006) 15806-11.
[141] A. Arora, C. Ha, C.B. Park, Insulin amyloid fibrillation at above 100 degrees C: New insights into protein folding under extreme temperatures, Protein Science 13(9) (2004) 2429-2436.
[142] C. Nordstedt, J. Naslund, L.O. Tjernberg, A.R. Karlstrom, J. Thyberg, L. Terenius, The Alzheimer a-Beta-Peptide Develops Protease Resistance in Association with Its Polymerization into Fibrils, Journal of Biological Chemistry 269(49) (1994) 30773-30776.
[143] G. Wei, Z. Su, N.P. Reynolds, P. Arosio, I.W. Hamley, E. Gazit, R. Mezzenga, Self-assembling peptide and protein amyloids: from structure to tailored function in nanotechnology, Chem Soc Rev 46(15) (2017) 4661-4708.
[144] I. Cherny, E. Gazit, Amyloids: not only pathological agents but also ordered nanomaterials, Angew Chem Int Ed Engl 47(22) (2008) 4062-9.
[145] T.P. Knowles, M.J. Buehler, Nanomechanics of functional and pathological amyloid materials, Nat Nanotechnol 6(8) (2011) 469-79.
[146] M.J. Buehler, NANOMATERIALS Strength in numbers, Nature Nanotechnology 5(3) (2010) 172-174.
[147] Y. Wang, J. Pu, B. An, T.K. Lu, C. Zhong, Emerging Paradigms for Synthetic Design of Functional Amyloids, J Mol Biol (2018).
[148] C. Zhong, T. Gurry, A.A. Cheng, J. Downey, Z.T. Deng, C.M. Stultz, T.K. Lu, Strong underwater adhesives made by self-assembling multi-protein nanofibres, Nature Nanotechnology 9(10) (2014) 858-866.
[149] T. Scheibel, R. Parthasarathy, G. Sawicki, X.M. Lin, H. Jaeger, S.L. Lindquist, Conducting nanowires built by controlled self-assembly of amyloid fibers and selective metal deposition, Proc Natl Acad Sci U S A 100(8) (2003) 4527-32.
[150] C.M. Rufo, Y.S. Moroz, O.V. Moroz, J. Stohr, T.A. Smith, X. Hu, W.F. DeGrado, I.V. Korendovych, Short peptides self-assemble to produce catalytic amyloids, Nat Chem 6(4) (2014) 303-9.
[151] Z. Botyanszki, P.K. Tay, P.Q. Nguyen, M.G. Nussbaumer, N.S. Joshi, Engineered catalytic biofilms: Site-specific enzyme immobilization onto E. coli curli nanofibers, Biotechnol Bioeng 112(10) (2015) 2016-24.

[152] D. Li, H. Furukawa, H. Deng, C. Liu, O.M. Yaghi, D.S. Eisenberg, Designed amyloid fibers as materials for selective carbon dioxide capture, Proc Natl Acad Sci U S A 111(1) (2014) 191-6.
[153] R.S. Jacob, D. Ghosh, P.K. Singh, S.K. Basu, N.N. Jha, S. Das, P.K. Sukul, S. Patil, S. Sathaye, A. Kumar, A. Chowdhury, S. Malik, S. Sen, S.K. Maji, Self healing hydrogels composed of amyloid nano fibrils for cell culture and stem cell differentiation, Biomaterials 54 (2015) 97-105.
[154] M.N. Bongiovanni, D.B. Scanlon, S.L. Gras, Functional fibrils derived from the peptide TTR1-cycloRGDfk that target cell adhesion and spreading, Biomaterials 32(26) (2011) 6099-6110.
[155] S.K. Maji, D. Schubert, C. Rivier, S. Lee, J.E. Rivier, R. Riek, Amyloid as a depot for the formulation of long-acting drugs, Plos Biology 6(2) (2008) 240-252.
[156] C.X. Li, J. Adamcik, R. Mezzenga, Biodegradable nanocomposites of amyloid fibrils and graphene with shape-memory and enzyme-sensing properties, Nature Nanotechnology 7(7) (2012) 421-427.
[157] J.T. MacDonald, P.S. Freemont, Computational protein design with backbone plasticity, Biochem Soc T 44 (2016) 1523-1529.
[158] C.A. Rohl, C.E.M. Strauss, K.M.S. Misura, D. Baker, Protein structure prediction using rosetta, Numerical Computer Methods, Pt D 383 (2004) 66-+.
[159] R.F. Alford, A. Leaver-Fay, J.R. Jeliazkov, M.J. O'Meara, F.P. DiMaio, H. Park, M.V. Shapovalov, P.D. Renfrew, V.K. Mulligan, K. Kappel, J.W. Labonte, M.S. Pacella, R. Bonneau, P. Bradley, R.L. Dunbrack, Jr., R. Das, D. Baker, B. Kuhlman, T. Kortemme, J.J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design, J Chem Theory Comput 13(6) (2017) 3031-3048.
[160] K.W. Kaufmann, G.H. Lemmon, S.L. Deluca, J.H. Sheehan, J. Meiler, Practically useful: what the Rosetta protein modeling suite can do for you, Biochemistry 49(14) (2010) 2987-98.
[161] Z. Li, H.A. Scheraga, Monte Carlo-minimization approach to the multiple-minima problem in protein folding, Proc Natl Acad Sci U S A 84(19) (1987) 6611-5.
[162] (!!! INVALID CITATION !!! {}).
[163] A.A. Canutescu, R.L. Dunbrack, Jr., Cyclic coordinate descent: A robotics algorithm for protein loop closure, Protein Sci 12(5) (2003) 963-72.
[164] D. Gront, D.W. Kulp, R.M. Vernon, C.E.M. Strauss, D. Baker, Generalized Fragment Picking in Rosetta: Design, Protocols and Applications, Plos One 6(8) (2011).
[165] K.T. Simons, C. Kooperberg, E. Huang, D. Baker, Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions, Journal of Molecular Biology 268(1) (1997) 209-225.
[166] C. Wang, P. Bradley, D. Baker, Protein-protein docking with backbone flexibility, J Mol Biol 373(2) (2007) 503-19.
[167] A. Zanghellini, L. Jiang, A.M. Wollacott, G. Cheng, J. Meiler, E.A. Althoff, D. Rothlisberger, D. Baker, New algorithms and an in silico benchmark for computational enzyme design, Protein Sci 15(12) (2006) 2785-94.
[168] F. Richter, A. Leaver-Fay, S.D. Khare, S. Bjelic, D. Baker, De novo enzyme design using Rosetta3, PLoS One 6(5) (2011) e19230.
[169] F. DiMaio, A. Leaver-Fay, P. Bradley, D. Baker, I. Andre, Modeling symmetric macromolecular structures in Rosetta3, PLoS One 6(6) (2011) e20450.
[170] A. Leaver-Fay, M. Tyka, S.M. Lewis, O.F. Lange, J. Thompson, R. Jacak, K. Kaufman, P.D. Renfrew, C.A. Smith, W. Sheffler, I.W. Davis, S. Cooper, A. Treuille, D.J. Mandell, F. Richter, Y.E. Ban, S.J. Fleishman, J.E. Corn, D.E. Kim, S. Lyskov, M. Berrondo, S. Mentzer, Z. Popovic, J.J. Havranek, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J.J. Gray, B. Kuhlman, D. Baker, P. Bradley, ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules, Methods Enzymol 487 (2011) 545-74.
[171] S. Chaudhury, S. Lyskov, J.J. Gray, PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta, Bioinformatics 26(5) (2010) 689-691.

[172] S.J. Fleishman, A. Leaver-Fay, J.E. Corn, E.M. Strauch, S.D. Khare, N. Koga, J. Ashworth, P. Murphy, F. Richter, G. Lemmon, J. Meiler, D. Baker, RosettaScripts: a scripting language interface to the Rosetta macromolecular modeling suite, PLoS One 6(6) (2011) e20161.
[173] M.R. Sawaya, J. Rodriguez, D. Cascio, M.J. Collazo, D. Shi, F.E. Reyes, J. Hattne, T. Gonen, D.S. Eisenberg, Ab initio structure determination from prion nanocrystals at atomic resolution by MicroED, Proc Natl Acad Sci U S A 113(40) (2016) 11232-11236.
[174] C.K. Smith, J.M. Withka, L. Regan, A Thermodynamic Scale for the Beta-Sheet Forming Tendencies of the Amino-Acids, Biochemistry 33(18) (1994) 5510-5517.
[175] D.L. Minor, P.S. Kim, Measurement of the Beta-Sheet-Forming Propensities of Amino-Acids, Nature 367(6464) (1994) 660-663.
[176] Y. Yoshimura, Y.X. Lin, H. Yagi, Y.H. Lee, H. Kitayama, K. Sakurai, M. So, H. Ogi, H. Naiki, Y. Goto, Distinguishing crystal-like amyloid fibrils and glass-like amorphous aggregates from their kinetics of formation, P Natl Acad Sci USA 109(36) (2012) 14446-14451.
[177] B. Alies, C. Hureau, P. Faller, The role of metal ions in amyloid formation: general principles from model peptides, Metallomics 5(3) (2013) 183-92.
[178] K.T. Simons, I. Ruczinski, C. Kooperberg, B.A. Fox, C. Bystroff, D. Baker, Improved recognition of native-like protein structures using a combination of sequence-dependent and sequence-independent features of proteins, Proteins-Structure Function and Genetics 34(1) (1999) 82-95.
[179] P.C. Kahn, Defining the Axis of a Helix, Comput Chem 13(3) (1989) 185-189.
[180] V.B. Chen, W.B. Arendall, 3rd, J.J. Headd, D.A. Keedy, R.M. Immormino, G.J. Kapral, L.W. Murray, J.S. Richardson, D.C. Richardson, MolProbity: all-atom structure validation for macromolecular crystallography, Acta Crystallogr D Biol Crystallogr 66(Pt 1) (2010) 12-21.
[181] C. Wang, R. Vernon, O. Lange, M. Tyka, D. Baker, Prediction of structures of zinc-binding proteins through explicit modeling of metal coordination geometry, Protein Sci 19(3) (2010) 494-506.
[182] U. Ryde, Carboxylate binding modes in zinc proteins: a theoretical study, Biophys J 77(5) (1999) 2777-87.
[183] R. Aurora, G.D. Rose, Helix capping, Protein Sci 7(1) (1998) 21-38.
[184] C. Hu, P. Koehl, Helix-sheet packing in proteins, Proteins 78(7) (2010) 1736-47.
[185] J.J. Dong, J.E. Shokes, R.A. Scott, D.G. Lynn, Modulating amyloid self-assembly and fibril morphology with Zn(II), Journal of the American Chemical Society 128(11) (2006) 3540-3542.
[186] K. Patel, A. Kumar, S. Durani, Analysis of the structural consensus of the zinc coordination centers of metalloprotein structures, Biochim Biophys Acta 1774(10) (2007) 1247-53.
[187] T.R. Jahn, S.E. Radford, The Yin and Yang of protein folding, FEBS J 272(23) (2005) 5962-70.
[188] R. Mimna, M.S. Camus, A. Schmid, G. Tuchscherer, H.A. Lashuel, M. Mutter, Disruption of amyloid-derived peptide assemblies through the controlled induction of a beta-sheet to alpha-helix transformation: application of the switch concept, Angew Chem Int Ed Engl 46(15) (2007) 2681-4.
[189] R.A. Kammerer, D. Kostrewa, J. Zurdo, A. Detken, C. Garcia-Echeverria, J.D. Green, S.A. Muller, B.H. Meier, F.K. Winkler, C.M. Dobson, M.O. Steinmetz, Exploring amyloid formation by a de novo design, Proc Natl Acad Sci U S A 101(13) (2004) 4435-40.
[190] A. Paul, B. Sharma, T. Mondal, K. Thalluri, S. Paul, B. Mandal, Amyloid β derived switch-peptides as a tool for investigation of early events of aggregation: a combined experimental and theoretical approach, MedChemComm 7(2) (2016) 311-316.
[191] J. Greenwald, R. Riek, Biology of amyloid: structure, function, and regulation, Structure 18(10) (2010) 1244-60.
[192] S.E. Boyken, Z. Chen, B. Groves, R.A. Langan, G. Oberdorfer, A. Ford, J.M. Gilmore, C. Xu, F. DiMaio, J.H. Pereira, B. Sankaran, G. Seelig, P.H. Zwart, D. Baker, De novo design of protein homo-oligomers with modular hydrogen-bond network-mediated specificity, Science 352(6286) (2016) 680-7.

[193] J. Meiler, D. Baker, ROSETTALIGAND: protein-small molecule docking with full side-chain flexibility, Proteins 65(3) (2006) 538-48.
[194] J. Moult, K. Fidelis, A. Kryshtafovych, T. Schwede, A. Tramontano, Critical assessment of methods of protein structure prediction (CASP)-Round XII, Proteins 86 Suppl 1 (2018) 7-15.
[195] R. Das, I. Andre, Y. Shen, Y.B. Wu, A. Lemak, S. Bansal, C.H. Arrowsmith, T. Szyperski, D. Baker, Simultaneous prediction of protein folding and docking at high resolution, P Natl Acad Sci USA 106(45) (2009) 18978-18983.
[196] E.A. Coutsias, C. Seok, Kinematic view of loop closure., Abstr Pap Am Chem S 228 (2004) U534-U534.
[197] W. Yang, L. Lai, Computational design of ligand-binding proteins, Curr Opin Struct Biol 45 (2017) 67-73.
[198] C. Xue, T.Y.W. Lin, D. Chang, Z.F. Guo, Thioflavin T as an amyloid dye: fibril quantification, optimal concentration and effect on aggregation, Roy Soc Open Sci 4(1) (2017).
[199] M. Lee, T. Wang, O.V. Makhlynets, Y. Wu, N.F. Polizzi, H. Wu, P.M. Gosavi, J. Stohr, I.V. Korendovych, W.F. DeGrado, M. Hong, Zinc-binding structure of a catalytic amyloid from solid-state NMR, Proc Natl Acad Sci U S A 114(24) (2017) 6191-6196.
[200] L. Gremer, D. Scholzel, C. Schenk, E. Reinartz, J. Labahn, R.B.G. Ravelli, M. Tusche, C. Lopez-Iglesias, W. Hoyer, H. Heise, D. Willbold, G.F. Schroder, Fibril structure of amyloid-beta(1-42) by cryo-electron microscopy, Science 358(6359) (2017) 116-+.

Appendices

A Extended Discussion

A.1 Approach 1: General *de novo* αβ-protein design combined with zinc site design

The general approach for αβ-protein (see for instance reference [54, 58-61]) starts out by building the protein BB structure (either in 1 or 2 steps) according to a blueprint with the SCs represented as centroids. This is then followed by a rigorous sampling of the BB using a low-resolution score function. The low-resolution score function consists mostly of knowledge-based score terms that focusses on sampling compact structures (such as radius of gyration (rg), packing between α-helices and β-strands (hs_pair) or centroid packing (cenpack)), and other terms such as BB hydrogen bonds (short and long range terms: hbond_sr_bb and hbond_lr_bb). See reference [158] for more information on the score terms.

For designing the amyloid structures, symmetry had to be applied to the αβ-proteins. In this project the amyloid structure was represented by 1 monomer with 5 symmetry mates, or simulated with 1 monomer only, but both strategies seemed to face the same challenges as described below.

Using the general αβ-protein design strategy does not yield BB structures that are suitable for designing a zinc site into, because zinc and its residue ligands does not generally fit into those structures. This is mainly because of the score terms that focusses on tight packing between the α-helix and β-strands. To design BB structures where a zinc site could fit into, a modification using harmonic constraints was applied. This was applied to specific atoms between the α-helix and β-strands to enforce separation between them. These constraints are applied directly to the score function as an energetic penalty. Using the constraints together with the compact score terms did not yield good structures. Huge penalties to the score function needed to be applied to enforce a sufficiently large distance between two SS elements.

Even though it seems obvious, reweighing the compact score terms to 0.0 was not tried out. Potentially, a good strategy might be applied turning the compact score terms off, and then

applying suitable constraints, that is, finding a good linear combination of suitable score terms.

But there are still some problems with this approach. The available constraints in Rosetta are mostly just constraints that can be applied to a couple of atoms. Many constraints can be defined at the same time and to different atoms of the helix and sheet to force them away from each other. But this is many constraints to apply at once, and it could be that sampling could be slow. Furthermore, to increase the chance of generating successful designs, one has to sample the BB of the helix (or sheet in the general case) rigorously, but this can be hard when many atoms are "frozen" by many individual constraints.

There should be a simpler approach that takes the rigorous BB sampling and the zinc volume into account. This is done more gracefully by the current algorithm presented in this thesis.

A.2 Approach 2: Fitting an ideal α-helix to inverse rotamers of zinc ligands

To fit a zinc site in between an α-helix and a β-sheet an algorithm was developed (It has been dubbed the "Sandwich Algorithm") as shown in **Figure A-1**. First (**1a** and **1b**), the coordination sphere of zinc was defined. Then following the left trajectory (**2a**-**4a**) a part of that coordination sphere (2 residue ligands) was put on the 2 β-strand scaffolds using the Rosetta Match algorithm (**3a**). To create an amyloid structure where the helix could coordinate 2 zinc sites, symmetry was applied to create 2 identical scaffolds that upheld the integrity of the amyloid fiber (when applying symmetry to generate the entire amyloid structure the asymmetric unit should be able to reproduce the entire SZ crystal structure).

Following the other trajectory (**2b**-**4b**) a part of a coordination sphere (2 residue ligands) was used to create many inverse rotamers in different geometries with respect to zinc (**3b**). Those geometries came from an analysis such as the one described in section 5.3, and the inverse rotamers were sampled around the optimal values and within 1 standard deviation from it. Then an ideal helix (ϕ=-57°, ψ=-47°) were fitted to these inverse rotamers, and this generated either an α-helix with 1 or 2 bound zinc.

Finally, the structures generated by both trajectories were then merged together (**5a** and **5b**) to create a monomer spanning 1 2 β-strand scaffold (top), or 2 2 β-strand scaffolds (bottom).

These could then be used for downstream design of a loop and sequence design.

This approach was abandoned mainly because of speed limitations. But the approach could be expanded and used in other areas of ligand binding protein design, or enzyme design. One strategy that could be interesting to explore is to create a library of structures (such as s β-sheets and or an α-helices in this case) bound to a ligand, or reaction intermediate. In a *de novo* design of ligand binding designs, these library structures could quickly be superimposed on each other (by superimposing the ligand) and then filtered based on clashes. This could then be followed by sequence design and further optimization.

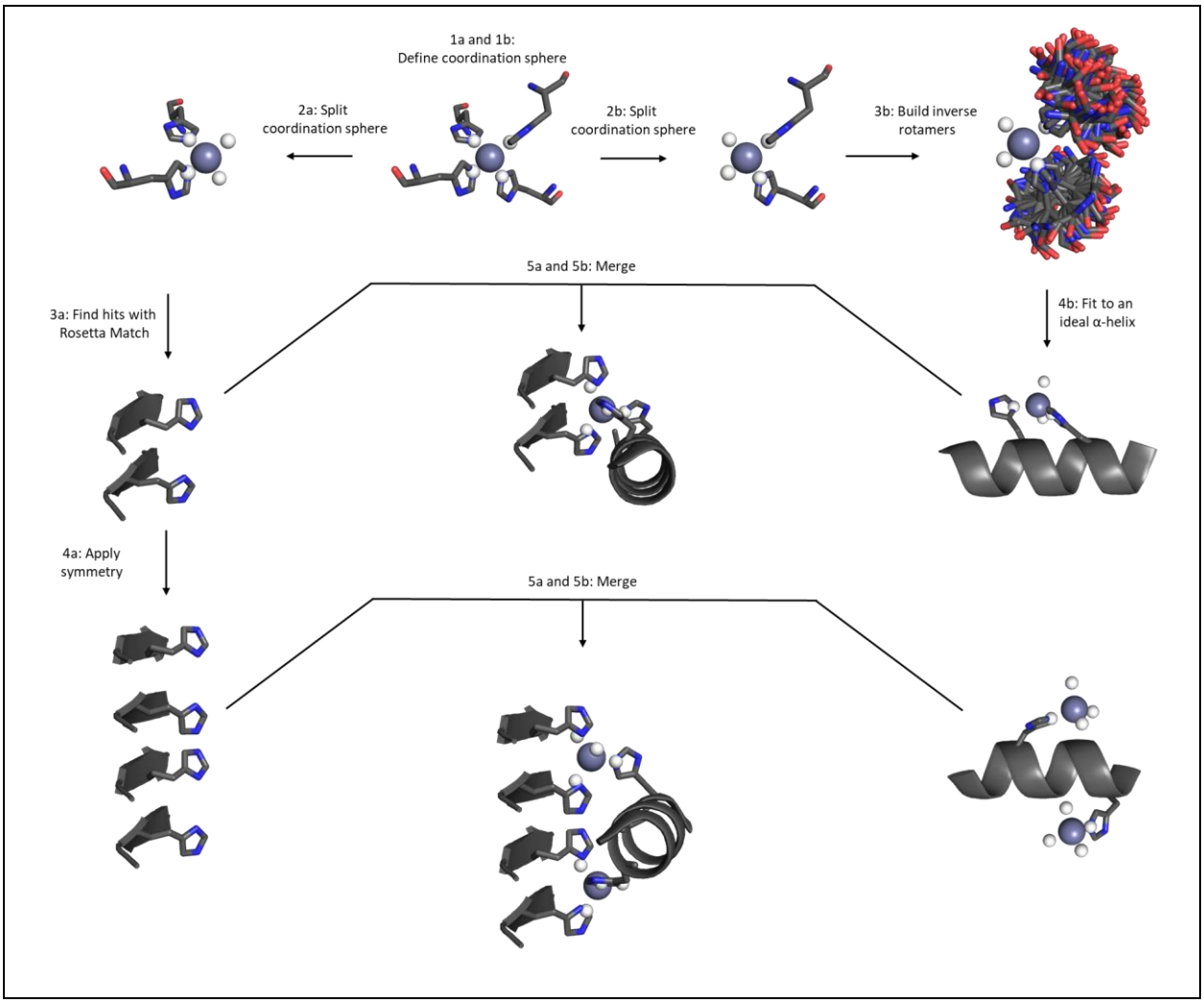


Figure A-1:The inner workings of the Sandwich Algorithm. The algorithm developed ("Sandwich Algorithm") for approach 2. Further explanation is in the main text. In this case a coordination sphere of histidine is shown, but the approach is general.

A.3 Approach 3: Grid search of an ideal α-helix on top of 2 β-strands scaffolds

The final approach that was explored was to sample an ideal α-helix on top of the 2 β-strand scaffold. Rotation, translation and tilting of the α-helix were allowed and were sampled in specific steps. This approach was quickly abandoned in favor of the current algorithm. The reason for this was that there should be constraints on where the α-helix should be allowed to move. In approach 3 only the volume of zinc was considered. But a problem arose when trying to connect loops to the α-helix. It was then thought that connecting the loops first and then doing the sampling would ease this problem, which is what the current algorithm accomplishes.

A.4 The role of the asymmetric structure in scaffold selection

Figure A-2 shows a case where the asymmetric unit could not maintain symmetry when a 2-fold rotation is applied to the 2 β-strand scaffolds. The full amyloid protofilament is shown to the left (**a**), while the asymmetric unit is shown to the right (**b**). Each unique strand is color-coded differently. If the blue and green β-strands are chosen as the scaffold, then when symmetry is applied they would land on top of the yellow and purple strand. Since these are not superimposable the symmetric representation from one monomer would not reproduce the entire SZ from the crystal structure.

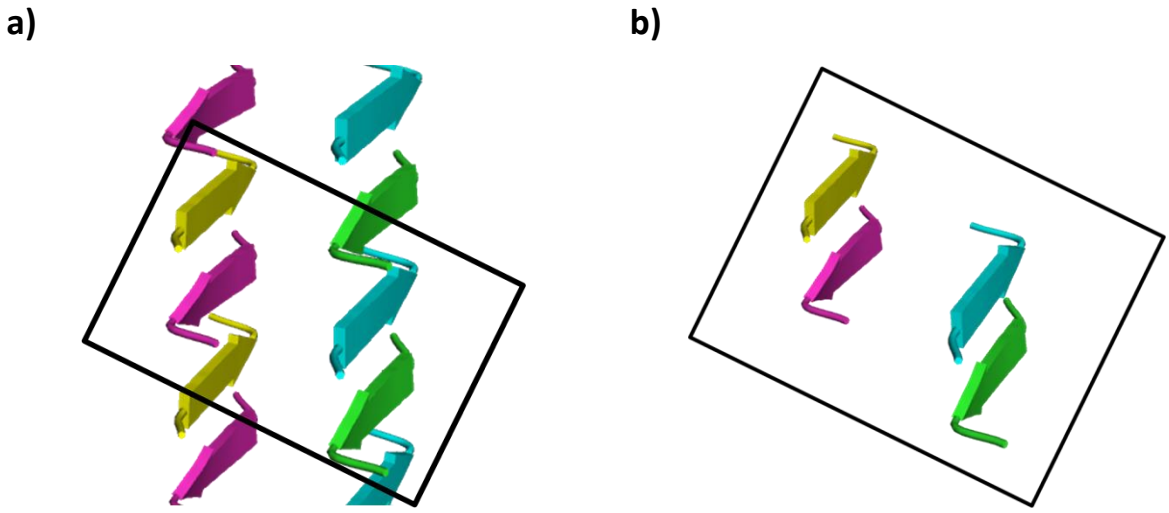


Figure A-2 The asymmetric units role in scaffold selection. a) The full protofilament of a crystal structure (PDB ID: 2OMQ). The square indicates the asymmetric unit of the crystal structure. **b)** The asymmetric unit as seen in **a**.

A.5 Other efforts in the algorithm design applied to step 1: Generation BB structures

Initially, a filter was used to separate out structures not having loops that could cap the α- helix. It turns out, in cases such as the sampling of the GBA loop shown in the Results, that this always seem to be the case (at least for the grid points sampled, taking larger grid points further out might produce non-capping results). The difference in strength of the hydrogen bonds was not looked at, only that the hbond_sr_bb and hbond_lr_bb score (long-range and short-range hydrogen bonds were above 0, but such a filter could be applied to obtain even more ideal starting structures for step 2.

The bottleneck in terms of speed in the entire design process Is the design of the zinc site. The simple choice of just searching positions having their CB atoms below the plane of the α-helices (see section 5.5.1) allows a position that clearly will never find a match with other residues, to still be searched by the Rosetta Match algorithm. For instance, the positions might be useless if a CA-CB vector (the vector from a CA to CB atom) is oriented in a way that points it away from the zinc site, or if the CB atom is way too far away from other CB atoms. Furthermore, for instance, His might coordinate from a specific position, but Cys might not, and then there is no reason to evaluate Cys for that position.

To quickly search through what positions and residues might be good to be used as inputs to the Rosetta Match algorithm, an algorithm was developed based on a graph structure. At each node of an outer graph structure a connection between all the residues was stored. A connection (=1) meant that the two positions could form a tetrahedral zinc site (based on minimum and maximum distances of the analysis in **Table 5-2**), and 0 if they could not. Each node then stored another inner graph structure representing possible connections between all BMs between those 2 connections of the node in the outer graph structure. A connection was 1 if 2 conditions passed: 1) the distance (based on an analysis such as in **Table 5-2**) and the angle (based on similar analysis) were within the ranges of what was observed for those BMs. The graph structure could quickly be checked for 4 residues connections for each sampled BB structure and appropriate position files and cst files (used to determine which BMs should be tried out) could be outputted.

It was never implemented in the final algorithm because it was hard to determine from CA and CB coordinates alone (for which the distance and angle measurements and analysis were

carried out) which BMs and residue positions were suitable. Most of the time all residues were chosen anyways. More development of this approach must be done in order for it to be useful. The algorithm works well, but the problem lies in the way 4 connections are accepted or rejected. More metrics than just CA and CB coordinates, with distances and angles between them must be applied.

A.6 Monomers needed to represent the full amyloid fibril

6 monomers are used to represent the full amyloid fibril for the step 3 involving full atom optimization and sequence design. It might be more efficient to use 5 or even 2 monomers instead. The two interfaces between the master subunit and the monomers next to it on the same side of the β-sheet is identical. Therefore 1 of monomer could be removed (hence 5 monomers in total). Since the SZ is not optimized during this step one could also remove the other side of the β-sheet. The space the two β-sheets is a little bit shorter than 10 Å, but the contributions they make could be negligible in the design. The other side is only used in the design process because of the residues in the loop might have interactions with the other side, that should be considered. If in future designs the SZ is also optimized, the contributions should matter.

A.7 Corrections

The values used in section 5.4 and for the Rosetta Match algorithm in section 5.6 used slightly different values that the ones presented in **Table 5-1**. These are shown “as is” in Appendix C in the constraint files. They should not change the result shown in the results significantly because the values are very similar. The constraint files should be changed, and the analysis rerun before moving on to improve the structures generated here.

A mistake was made in loop modelling using the CCD algorithm. A part of the β-sheet was remodeled, and a Val was mutated to Ala. This mutation happened in the SZ so away from other interface, and the BB was very similar to the original one. The results of the Rosetta Match algorithm (step 2) for which the generated structures were run on, should not change the results shown here significantly because of its similarity with the original structure. It was also run during BB generation (step 1) but should not change the results significantly either, because it faced the

SZ. The mistakes can be seen in the blueprint file used for the CCD modelling.

Before moving on to the optimization and sequence design, Ala was mutated back to Val, and with the same rotamer as the original. The new Val did not overlap completely with the one from the original crystal structure, but it was very close. It should not change the results of the optimization step significantly because it is facing the SZ and is not optimized.

Others that would try to improve the work presented here should change the things mentioned and should be aware of it if copying directly from Appendix C.

B Extended Figures

B.1 Density histograms and fits for the 8 other binding modes

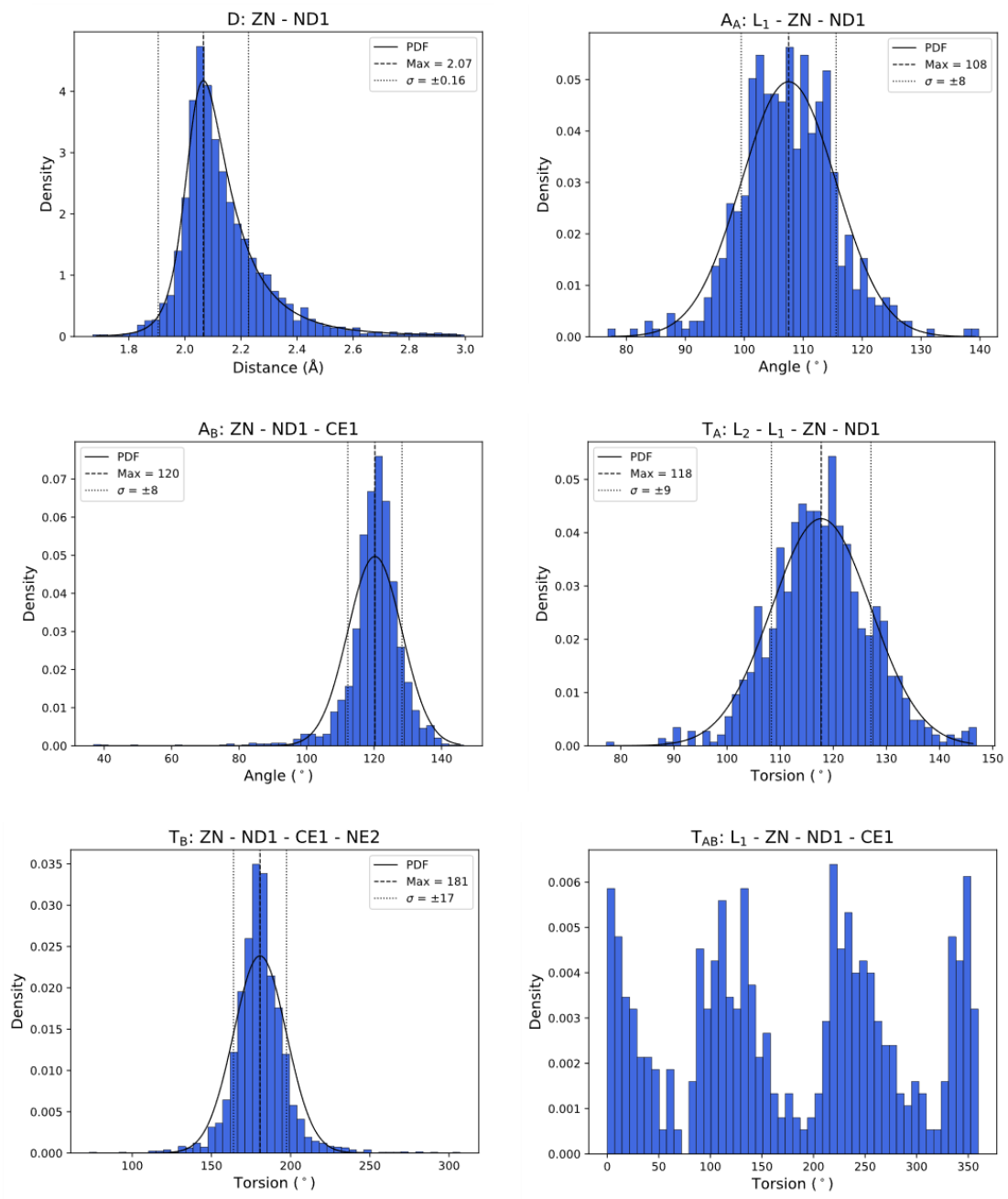


Figure B-1:Density histograms and fits for the ZN:ND1 binding mode.

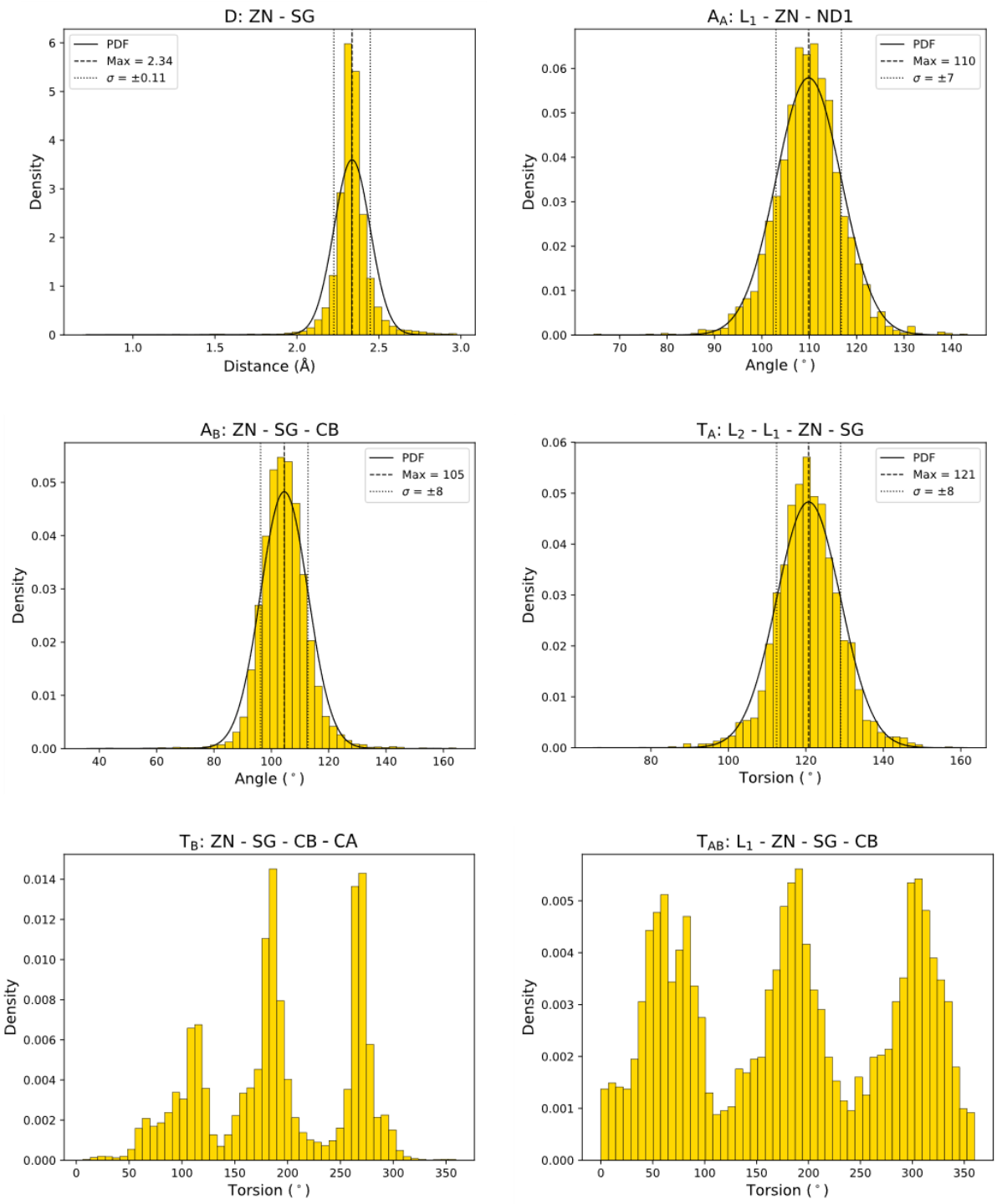


Figure B-2: Density histograms and fits for the ZN:SG binding mode.

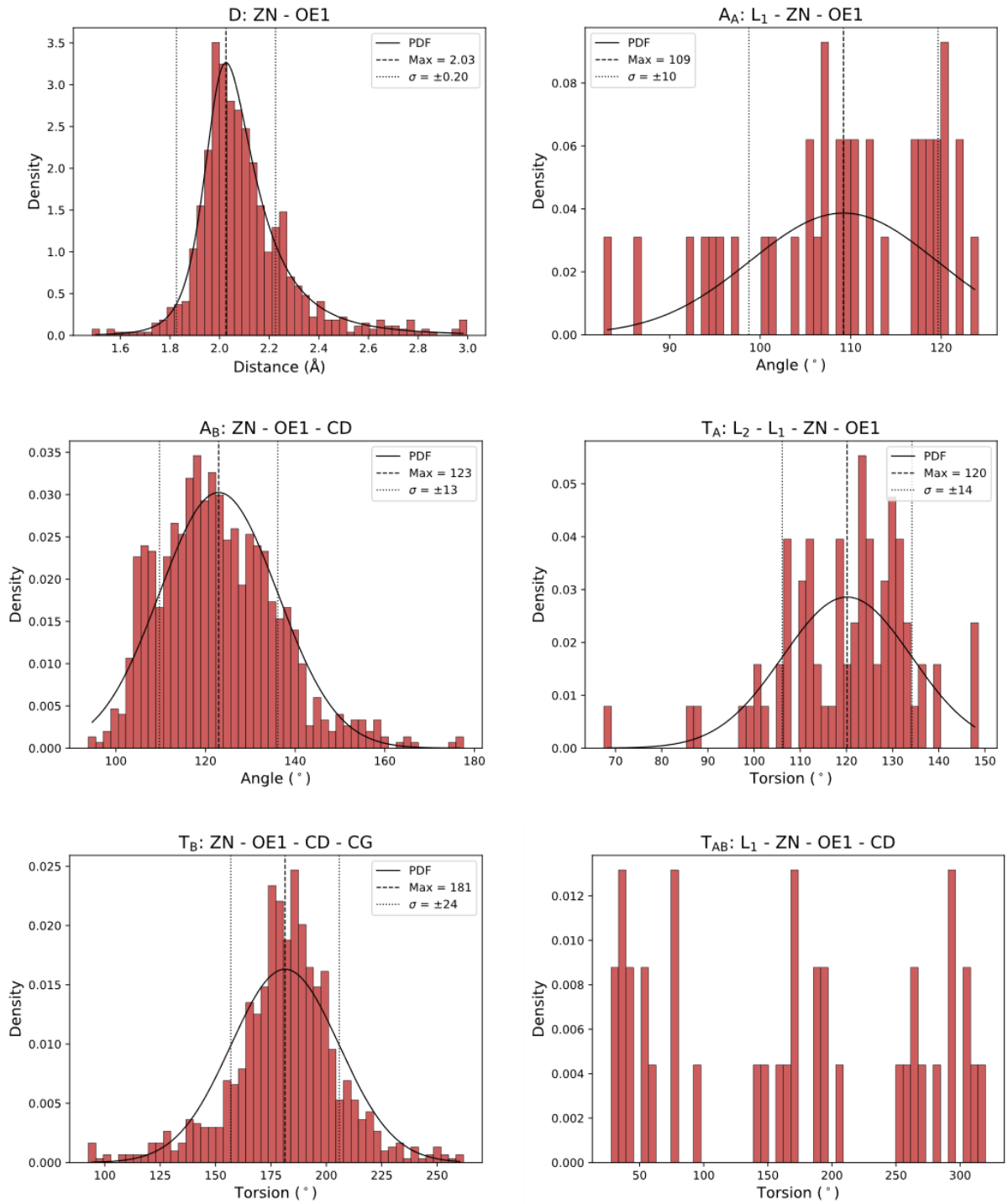


Figure B-3: Density histograms and fits for the ZN:OE1 binding mode.

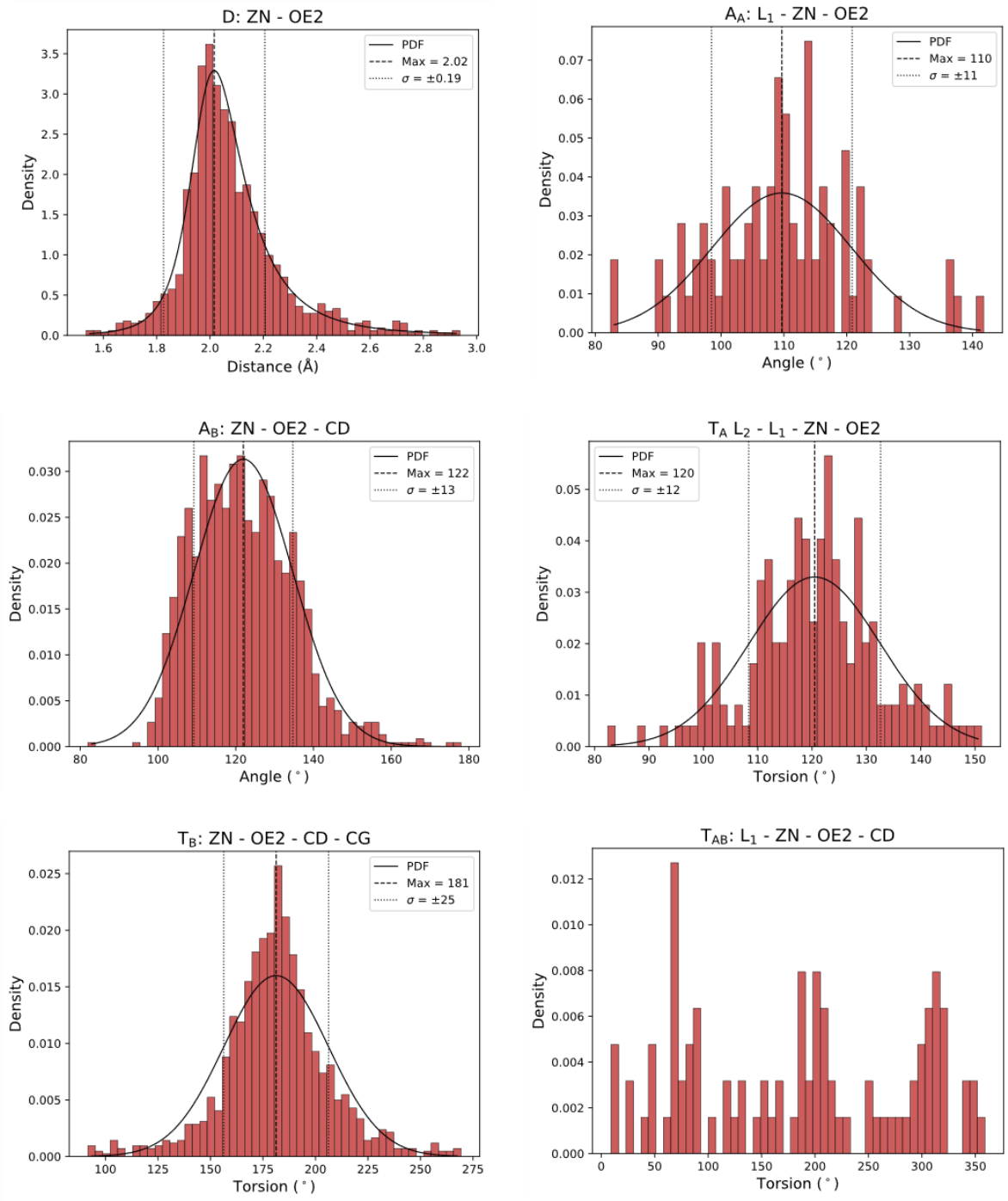


Figure B-4: Density histograms and fits for the ZN:OE2 binding mode.

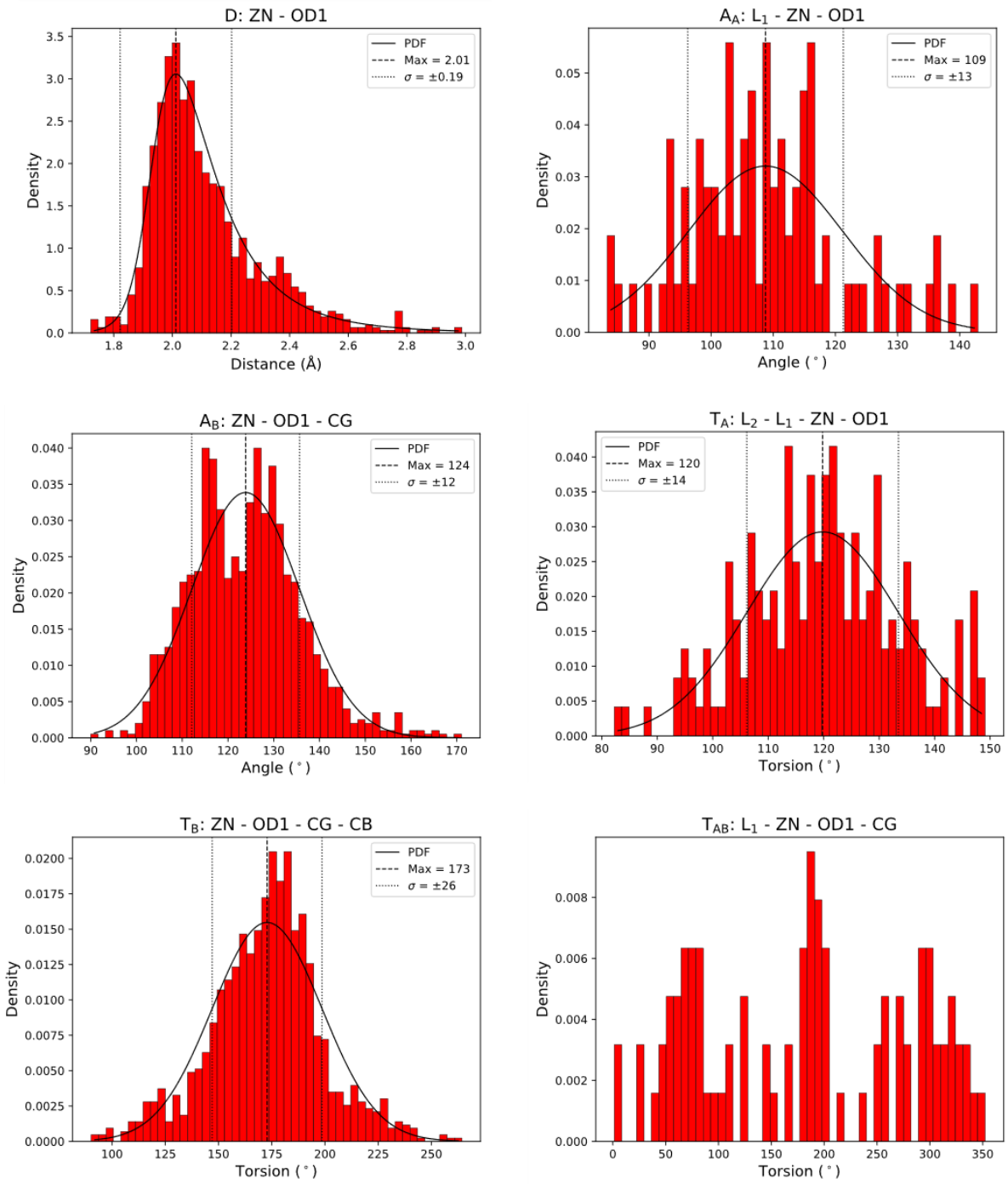


Figure B-5: Density histograms and fits for the ZN:OD1 binding mode.

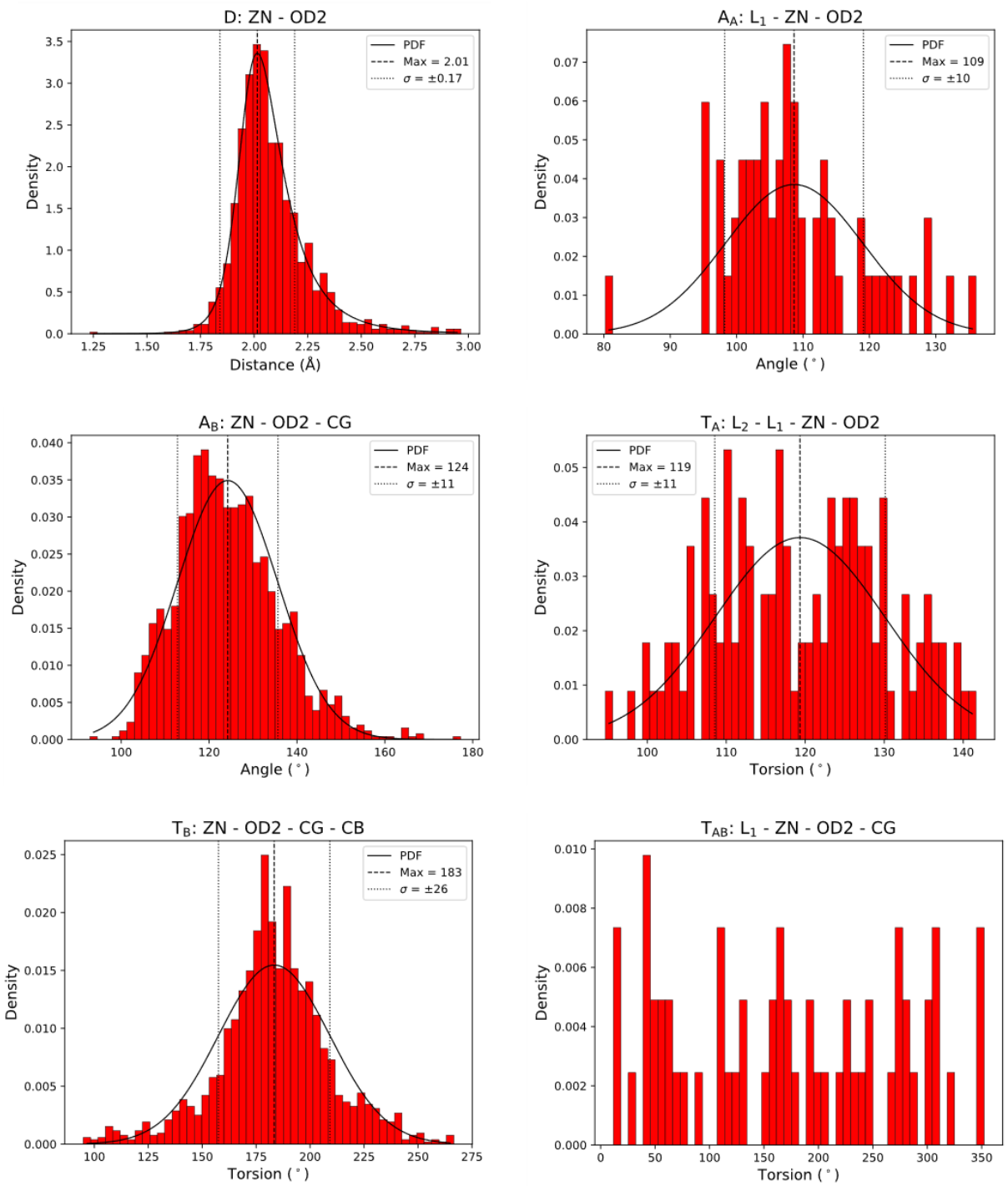


Figure B-6: Density histograms and fits for the ZN:OD2 binding mode.

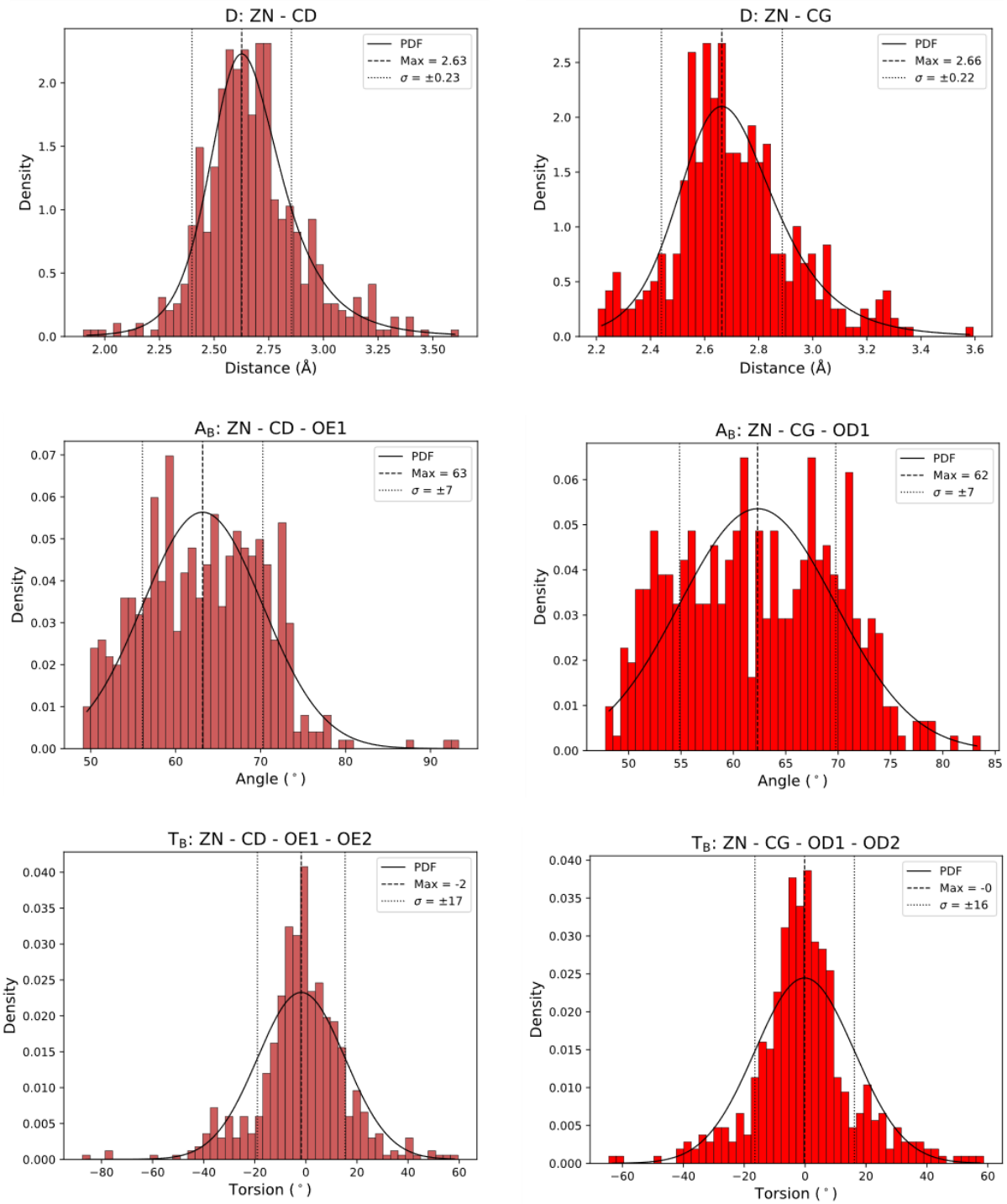


Figure B-7: Density histograms and fits for the ZN:OE1|OE2 and ZN OD1|OD2 binding mode.

B.2 Analysis and coordination parameters for mononuclear tetrahedral zinc sites

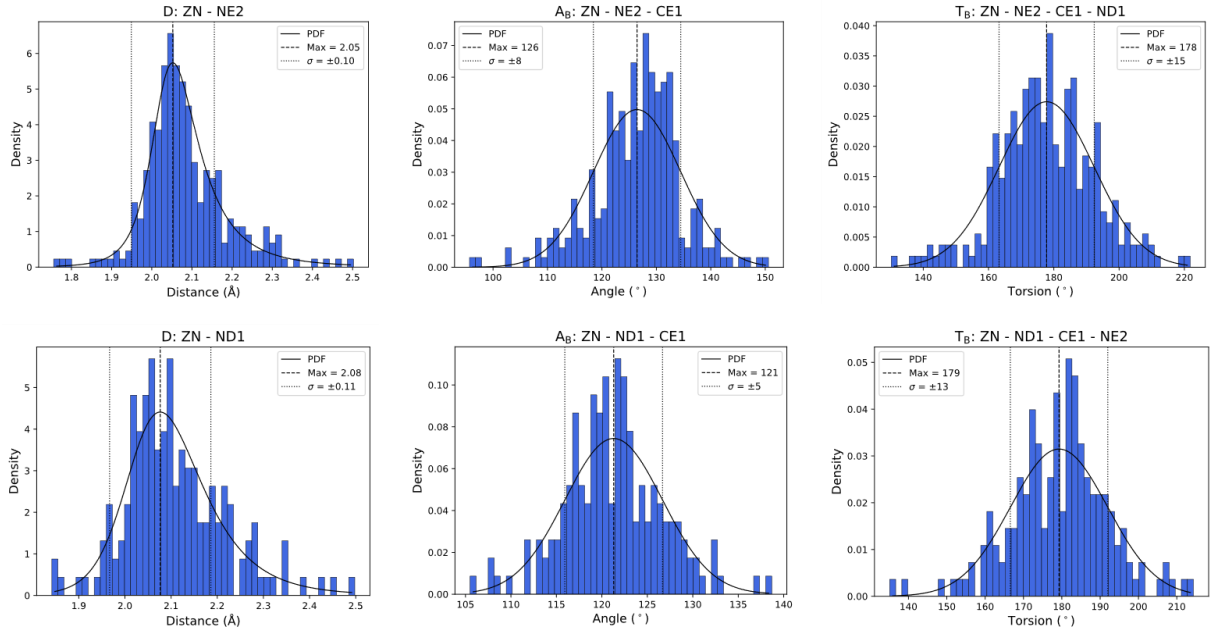


Figure B-8: Analysis the ZN:NE2 and ZN:ND1 binding mode for mononuclear tetrahedral zinc sites.

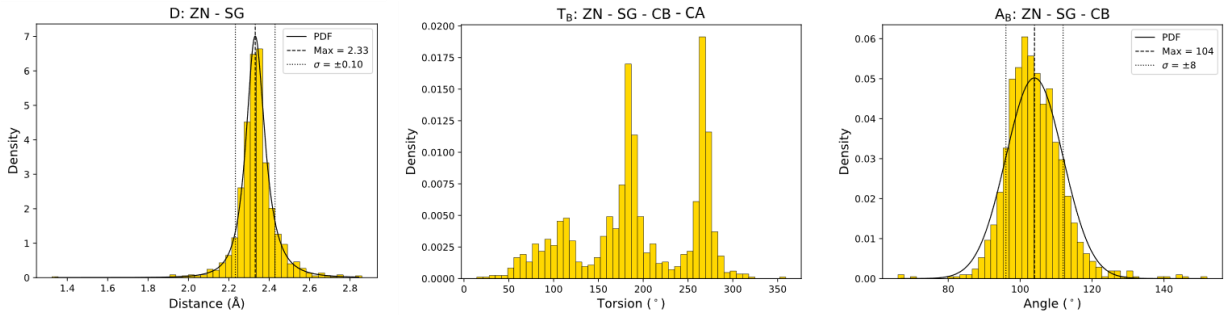


Figure B-9: Analysis of the ZN:SG binding mode for mononuclear tetrahedral zinc sites.

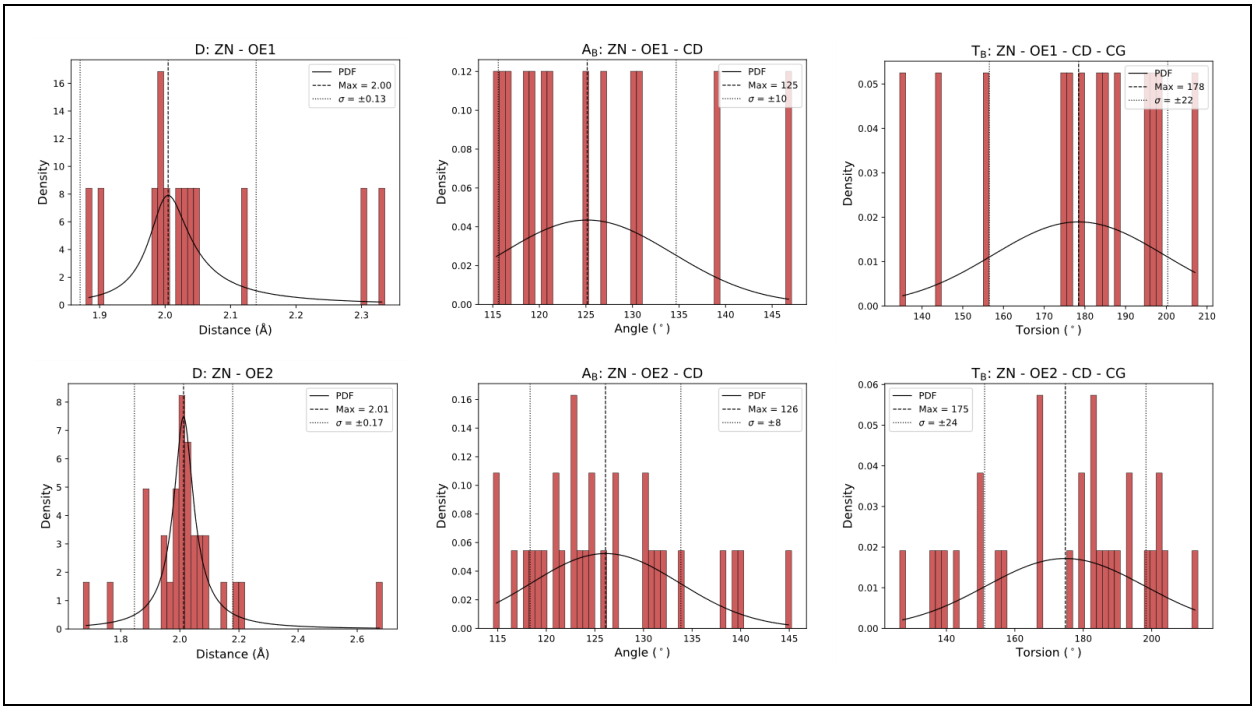


Figure B-10: Analysis of the ZN:OE1 and ZN:OE2 binding mode for mononuclear tetrahedral zinc sites.

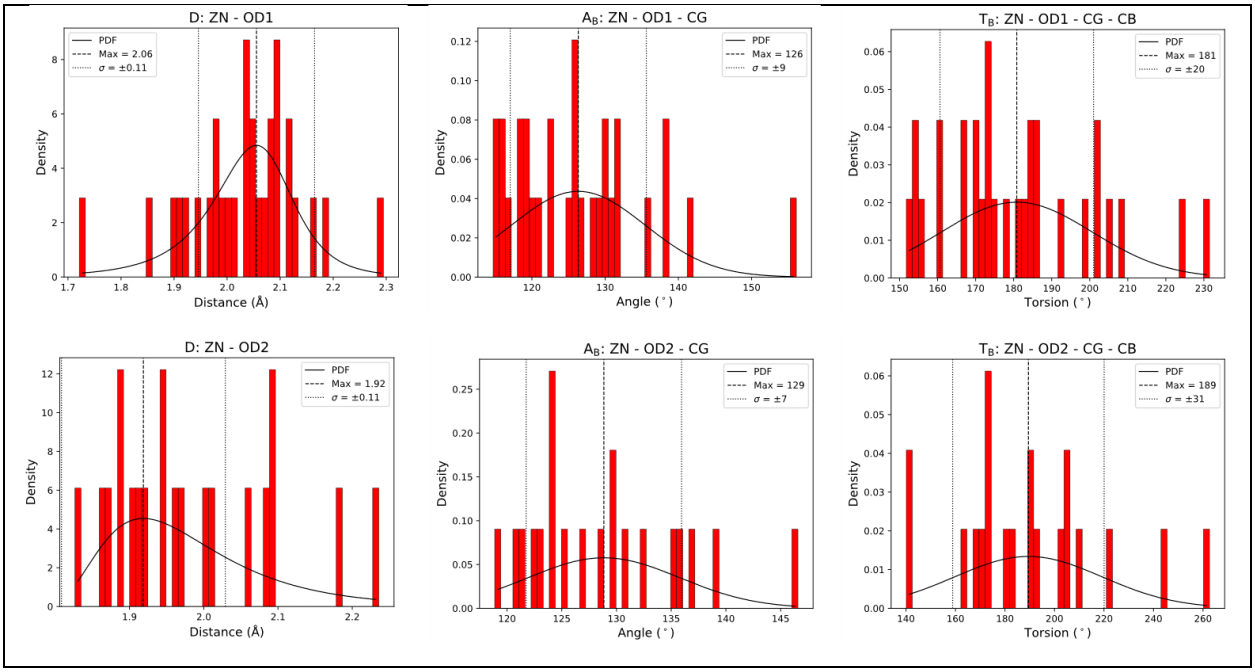


Figure B-11: Analysis of the ZN:OD1 and ZN:OD2 binding mode for mononuclear tetrahedral zinc sites.

C Scripts

All scripts presented in Appendix C are available in a file archive and can be obtained by contacting Mads Jeppesen (Author) or Ebbe Sloth Andersen (Supervisor). A readme file also exists in this file archive and it similarly describes how to run the scripts.

C.1 Zinc site analysis and fitting

Two scripts were developed to carry out the zinc sites analysis as well as the fitting as described in Methods 4.2.5, 4.2.6 and Results 5.3.

The scripts should be run in a folder containing two sub folders of the full dataset and the subset containing only mononuclear tetrahedral sites. Furthermore, a file containing information on all the coordinating atoms to the mononuclear tetrahedral sites should also be present in that folder The scripts, also present in the folders, can be run as (from the terminal):

```
Python all_analysis.py
Python mononuclear_tetrahedal_sites_analysis.py
```

This will produce all the density histograms as well as the fits. The name of the folders and paths should be changed inside the scripts.

C.2 Distance analysis

To sample 2 residues around a tetrahedral zinc atom, the following Rosetta executable was used along with the following flag file. An example of the cst file for cysteine is also shown below. The cst file would have to be changed for each of the 9 BMs. See Appendix C.8 for the ZNA.params file.

```
CstfileToTheozymePDB.mpi.linuxgccrelease @flags
```

flags file:

```
-extra_res_fa ZNA.params
-match:geometric_constraint_file <cst file>
```

```
-output_virtual true
```

Example of cst file:

```
# Match Constraints
# CYS 1:
        CST::BEGIN
        TEMPLATE::   ATOM_MAP: 1 atom_name: ZN V4 V2
        TEMPLATE::   ATOM_MAP: 1 residue3:  ZN
        TEMPLATE::   ATOM_MAP: 2 atom_name: SG CB CA ,
        TEMPLATE::   ATOM_MAP: 2 residue1:  C

        CONSTRAINT:: distanceAB:  2.34    0.11   10.0  1    1
        CONSTRAINT::    angle_A:  109.5   6.89   10.0  360. 1 # idealized
        CONSTRAINT::    angle_B:  104.51  8.27   10.0  360. 1
        CONSTRAINT::  torsion_A:  120.0   8.26   10.0  360. 1 # idealized
        CONSTRAINT::  torsion_B:  0.0     360.0  10.0  360. 4 # Free
        CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 4 # Free
        CST::END

# CYS 2:
        CST::BEGIN
        TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V3
        TEMPLATE::   ATOM_MAP: 1 residue3:  ZN
        TEMPLATE::   ATOM_MAP: 2 atom_name: SG CB CA ,
        TEMPLATE::   ATOM_MAP: 2 residue1:  C

        CONSTRAINT:: distanceAB:  2.34    0.11   10.0  1    1
        CONSTRAINT::    angle_A:  109.5   6.89   10.0  360. 1 # idealized
        CONSTRAINT::    angle_B:  104.51  8.27   10.0  360. 1
        CONSTRAINT::  torsion_A:  120.0   8.26   10.0  360. 1 # idealized
        CONSTRAINT::  torsion_B:  0.0     360.0  10.0  360. 4 # Free
        CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 4 # Free

        CST::END
```

The amount of rotamers created for each BM was:

ZN:SG: 7776

ZN:OD1|OD2: 13122

ZN:OD1: 13122

ZN:OD2: 13122

ZN:OE1|OE2: 21870

ZN:OE1: 21870

ZN:OE2: 21870

ZN:NE2: 40828

ZN:ND1: 40828

To analyze the CB distances the following script was run in parallel along with the python commands file.

```
parallel -j 9 -a python_commands python
```

python_commands file

```
all_in_one_C.py
all_in_one_D_BOTH.py
all_in_one_E_BOTH.py
all_in_one_ND1.py
all_in_one_NE2.py
all_in_one_OD1.py
all_in_one_OD2.py
all_in_one_OE1.py
all_in_one_OE2.py
```

This generates 9 datafiles that is used by the following script to obtain a matrix of all 9x9 interactions between the 4 ligands with in total 9 BMs.

```
Python get_info_from_data.py
```

C.3 Generation of the Ramachandran plots and selected grid points

Before inputting the 70% homology Top8000 into the scripts these were cleaned to only contain ATOM records (can be done with the PyRosetta Software). Following this, the first script calculates all the torsion angles with the filters as described in Methods 4.2.9. This can be run as:

```
Python phi_psi_set_from_top8000.py
```

This generates a file containing all of the torsion angles. Then the following script can be run to obtain the selected grid points as well as producing the Ramachandran plot as shown in **Figure 5-8.a**:

```
Python generate_rama_n_grids.py
```

Specifically it produces 4 files that are used in the algorithm developed for step 1.

C.4 Scripts for step 1: Generation of BB structures

The α3β2β-2γ29 model is use for the following. To mutate the 2 β-strand scaffolds and design a 14 residue α-helix and a 3 residue loop connecting it 2 β-strand scaffolds, the following script is run in the terminal:

```
rosetta_scripts.static.linuxgccrelease -parser:protocol BBgen.xml -s input.pdb
```

The input.pdb file is the 2 β-strand scaffolds (see **Figure 5-1**) and is seen below:

```
ATOM      1  N   LYS B  13       6.223  -6.367   7.940  1.00 24.11           N
ATOM      2  CA  LYS B  13       5.617  -5.255   7.193  1.00 25.75           C
ATOM      3  C   LYS B  13       5.408  -5.650   5.718  1.00 22.45           C
ATOM      4  O   LYS B  13       4.336  -6.098   5.404  1.00 22.63           O
ATOM      5  CB  LYS B  13       6.386  -3.937   7.368  1.00 36.28           C
ATOM      6  CG  LYS B  13       6.037  -3.104   8.663  1.00 37.36           C
ATOM      7  CD  LYS B  13       5.022  -1.941   8.386  1.00 43.82           C
ATOM      8  CE  LYS B  13       4.998  -0.848   9.509  1.00 38.18           C
ATOM      9  NZ  LYS B  13       3.730   0.004   9.514  1.00 37.31          N1+
```

```
ATOM     10  N   LEU B  14       6.412  -5.554   4.843  1.00 20.46           N
ATOM     11  CA  LEU B  14       6.212  -5.805   3.417  1.00 16.76           C
ATOM     12  C   LEU B  14       6.721  -7.123   2.927  1.00 15.82           C
ATOM     13  O   LEU B  14       7.928  -7.380   3.015  1.00 14.04           O
ATOM     14  CB  LEU B  14       6.848  -4.708   2.566  1.00 18.25           C
ATOM     15  CG  LEU B  14       6.720  -4.908   1.045  1.00 20.82           C
ATOM     16  CD1 LEU B  14       5.274  -4.700   0.513  1.00 18.21           C
ATOM     17  CD2 LEU B  14       7.710  -4.030   0.272  1.00 23.81           C
ATOM     18  N   VAL B  15       5.815  -7.938   2.356  1.00 14.07           N
ATOM     19  CA  VAL B  15       6.216  -9.164   1.669  1.00 14.54           C
ATOM     20  C   VAL B  15       5.648  -9.355   0.301  1.00 16.12           C
ATOM     21  O   VAL B  15       4.446  -9.581   0.125  1.00 18.28           O
ATOM     22  CB  VAL B  15       5.908 -10.410   2.471  1.00 14.30           C
ATOM     23  CG1 VAL B  15       6.409 -11.650   1.697  1.00 14.29           C
ATOM     24  CG2 VAL B  15       6.583 -10.315   3.808  1.00 13.11           C
ATOM     25  N   PHE B  16       6.537  -9.279  -0.676  1.00 19.17           N
ATOM     26  CA  PHE B  16       6.185  -9.435  -2.083  1.00 18.86           C
ATOM     27  C   PHE B  16       6.837 -10.680  -2.696  1.00 18.52           C
ATOM     28  O   PHE B  16       8.048 -10.885  -2.546  1.00 19.22           O
ATOM     29  CB  PHE B  16       6.601  -8.184  -2.863  1.00 19.73           C
ATOM     30  CG  PHE B  16       6.362  -8.293  -4.344  1.00 19.23           C
ATOM     31  CD1 PHE B  16       7.183  -9.078  -5.140  1.00 19.50           C
ATOM     32  CD2 PHE B  16       5.316  -7.598  -4.940  1.00 19.79           C
ATOM     33  CE1 PHE B  16       6.944  -9.189  -6.512  1.00 20.32           C
ATOM     34  CE2 PHE B  16       5.079  -7.703  -6.279  1.00 19.74           C
ATOM     35  CZ  PHE B  16       5.894  -8.511  -7.072  1.00 19.06           C
ATOM     36  N   PHE B  17       6.031 -11.476  -3.396  1.00 18.19           N
ATOM     37  CA  PHE B  17       6.475 -12.677  -4.121  1.00 21.35           C
ATOM     38  C   PHE B  17       5.702 -12.870  -5.415  1.00 24.27           C
ATOM     39  O   PHE B  17       4.528 -13.322  -5.404  1.00 21.75           O
ATOM     40  CB  PHE B  17       6.382 -13.995  -3.274  1.00 21.59           C
ATOM     41  CG  PHE B  17       7.260 -15.099  -3.805  1.00 23.29           C
ATOM     42  CD1 PHE B  17       8.670 -15.045  -3.642  1.00 24.72           C
ATOM     43  CD2 PHE B  17       6.720 -16.166  -4.537  1.00 22.64           C
ATOM     44  CE1 PHE B  17       9.499 -16.064  -4.191  1.00 24.66           C
ATOM     45  CE2 PHE B  17       7.553 -17.171  -5.091  1.00 19.71           C
ATOM     46  CZ  PHE B  17       8.918 -17.138  -4.920  1.00 19.51           C
ATOM     47  N   ALA B  18       6.367 -12.575  -6.532  1.00 24.61           N
```

```
ATOM     48  CA  ALA B  18       5.763 -12.785  -7.821  1.00 26.63           C
ATOM     49  C   ALA B  18       6.403 -13.992  -8.487  1.00 31.16           C
ATOM     50  O   ALA B  18       7.610 -14.032  -8.645  1.00 29.56           O
ATOM     51  CB  ALA B  18       5.896 -11.560  -8.669  1.00 29.69           C
ATOM     52  OXT ALA B  18       5.750 -14.979  -8.854  1.00 38.24          O1-
ATOM     53  N   LYS B  19       1.428 -11.738  -7.940  1.00 24.11           N
ATOM     54  CA  LYS B  19       0.822 -12.850  -7.193  1.00 25.75           C
ATOM     55  C   LYS B  19       0.613 -12.455  -5.718  1.00 22.45           C
ATOM     56  O   LYS B  19      -0.459 -12.007  -5.404  1.00 22.63           O
ATOM     57  CB  LYS B  19       1.591 -14.168  -7.368  1.00 36.28           C
ATOM     58  CG  LYS B  19       1.242 -15.001  -8.663  1.00 37.36           C
ATOM     59  CD  LYS B  19       0.227 -16.164  -8.386  1.00 43.82           C
ATOM     60  CE  LYS B  19       0.203 -17.257  -9.509  1.00 38.18           C
ATOM     61  NZ  LYS B  19      -1.065 -18.109  -9.514  1.00 37.31          N1+
ATOM     62  N   LEU B  20       1.617 -12.551  -4.843  1.00 20.46           N
ATOM     63  CA  LEU B  20       1.417 -12.300  -3.417  1.00 16.76           C
ATOM     64  C   LEU B  20       1.926 -10.982  -2.927  1.00 15.82           C
ATOM     65  O   LEU B  20       3.133 -10.725  -3.015  1.00 14.04           O
ATOM     66  CB  LEU B  20       2.053 -13.397  -2.566  1.00 18.25           C
ATOM     67  CG  LEU B  20       1.925 -13.197  -1.045  1.00 20.82           C
ATOM     68  CD1 LEU B  20       0.479 -13.405  -0.513  1.00 18.21           C
ATOM     69  CD2 LEU B  20       2.915 -14.075  -0.272  1.00 23.81           C
ATOM     70  N   VAL B  21       1.020 -10.167  -2.356  1.00 14.07           N
ATOM     71  CA  VAL B  21       1.421  -8.941  -1.669  1.00 14.54           C
ATOM     72  C   VAL B  21       0.853  -8.750  -0.301  1.00 16.12           C
ATOM     73  O   VAL B  21      -0.349  -8.524  -0.125  1.00 18.28           O
ATOM     74  CB  VAL B  21       1.113  -7.695  -2.471  1.00 14.30           C
ATOM     75  CG1 VAL B  21       1.614  -6.455  -1.697  1.00 14.29           C
ATOM     76  CG2 VAL B  21       1.788  -7.790  -3.808  1.00 13.11           C
ATOM     77  N   PHE B  22       1.742  -8.826   0.676  1.00 19.17           N
ATOM     78  CA  PHE B  22       1.390  -8.670   2.083  1.00 18.86           C
ATOM     79  C   PHE B  22       2.042  -7.425   2.696  1.00 18.52           C
ATOM     80  O   PHE B  22       3.253  -7.220   2.546  1.00 19.22           O
ATOM     81  CB  PHE B  22       1.806  -9.921   2.863  1.00 19.73           C
ATOM     82  CG  PHE B  22       1.567  -9.812   4.344  1.00 19.23           C
ATOM     83  CD1 PHE B  22       2.388  -9.027   5.140  1.00 19.50           C
ATOM     84  CD2 PHE B  22       0.521 -10.507   4.940  1.00 19.79           C
ATOM     85  CE1 PHE B  22       2.149  -8.916   6.512  1.00 20.32           C
```

```
ATOM     86  CE2 PHE B  22       0.284 -10.402   6.279  1.00 19.74           C
ATOM     87  CZ  PHE B  22       1.099  -9.594   7.072  1.00 19.06           C
ATOM     88  N   PHE B  23       1.236  -6.629   3.396  1.00 18.19           N
ATOM     89  CA  PHE B  23       1.680  -5.428   4.121  1.00 21.35           C
ATOM     90  C   PHE B  23       0.907  -5.235   5.415  1.00 24.27           C
ATOM     91  O   PHE B  23      -0.267  -4.783   5.404  1.00 21.75           O
ATOM     92  CB  PHE B  23       1.587  -4.110   3.274  1.00 21.59           C
ATOM     93  CG  PHE B  23       2.465  -3.006   3.805  1.00 23.29           C
ATOM     94  CD1 PHE B  23       3.875  -3.060   3.642  1.00 24.72           C
ATOM     95  CD2 PHE B  23       1.925  -1.939   4.537  1.00 22.64           C
ATOM     96  CE1 PHE B  23       4.704  -2.041   4.191  1.00 24.66           C
ATOM     97  CE2 PHE B  23       2.758  -0.934   5.091  1.00 19.71           C
ATOM     98  CZ  PHE B  23       4.123  -0.967   4.920  1.00 19.51           C
ATOM     99  N   ALA B  24       1.572  -5.530   6.532  1.00 24.61           N
ATOM    100  CA  ALA B  24       0.968  -5.320   7.821  1.00 26.63           C
ATOM    101  C   ALA B  24       1.608  -4.113   8.487  1.00 31.16           C
ATOM    102  O   ALA B  24       2.815  -4.073   8.645  1.00 29.56           O
ATOM    103  CB  ALA B  24       1.101  -6.545   8.669  1.00 29.69           C
ATOM    104  OXT ALA B  24       0.955  -3.126   8.854  1.00 38.24          O1-
END
```

The BBgen.xml file is seen below:

```
<ROSETTASCRIPTS>
    <RESIDUE_SELECTORS>
      <Index name="non_sz" resnums="3,5,8,10"/>
    </RESIDUE_SELECTORS>
    <MOVERS>
      <MutateResidue      name="mutate_non_sz"      residue_selector="non_sz"
new_res="ALA" />
      <DeleteRegionMover name="delete_terminal" start="1" end="1"/>
      <PeptideStubMover name="build_lnh">
            <Prepend resname="ALA" anchor_rsd="1" repeat="17"/>
      </PeptideStubMover>

      <AtomTree name="tree" fold_tree_file="tree_abb_3l.file"/>
      <SetTorsion name="idealize_loop">
```

```
      <Torsion residue="15,16,17" torsion_name="omega" angle="180" />
      <Torsion residue="15,16,17" torsion_name="rama" angle="rama_biased" />
      <Torsion residue="15,16,17" torsion_name="rama" angle="rama_biased" />
      </SetTorsion>
      <SetTorsion name="idealize_helix">
      <Torsion residue="1,2,3,4,5,6,7,8,9,10,11,12,13,14" torsion_name="omega"
angle="180" />
      <Torsion  residue="1,2,3,4,5,6,7,8,9,10,11,12,13,14"  torsion_name="phi"
angle="-57" />
      <Torsion  residue="1,2,3,4,5,6,7,8,9,10,11,12,13,14"  torsion_name="psi"
angle="-47" />
      </SetTorsion>
    </MOVERS>
    <PROTOCOLS>
      <Add mover="mutate_non_sz"/>
      <Add mover="delete_terminal"/>
      <Add mover="build_lnh"/>
      <Add mover="tree"/>
      <Add mover="idealize_loop"/>
      <Add mover="idealize_helix"/>
    </PROTOCOLS>
```

The tree_abb_3l.file is used to setup the foldtree correctly and its content is seen below:

```
FOLD_TREE EDGE 28 1 -1
```

To design a loop between the 2 β-strands with the CCD algorithm the following code is run:

```
remodel.mpi.linuxgccrelease @remodelflags
```

The remodelflags file is seen below:

```
-s 2y29_l3nh14.pdb
-nstruct 100
-remodel:blueprint 2y29_l5nh14.bp
-remodel:design:no_design 1
```

```
-remodel:generic_aa A
-remodel:RemodelLoopMover:allowed_closure_attempts 10
-remodel:RemodelLoopMover:max_linear_chainbreak 0.1
-mute true
```

The 2y29_l5nh14.pdb file is the output from the previous BBgen.xml script. The 2y29_l5nh14.bp is seen below:

```
1 A .
2 A .
3 A .
4 A .
5 A .
6 A .
7 A .
8 A .
9 A .
10 A .
11 A .
12 A .
13 A .
14 A .
15 A .
16 A .
17 A .
18 V .
19 A .
20 F .
21 A .
22 A E
23 K L
24 A L
25 F E
26 A .
27 F .
28 A .
```

When the BB structures have been generated the following command can be run which uses the algorithm as described in the Results.

```
python rama_search_newest_version.py @2y29_5lnh14_rama_flag
```

The 2y29_5lnh14_rama_flag is seen below (commented so the user can change inputs):

```
# input files (pdb id and symmetry file
-pdb 2y29_l5nh14.pdb
-sym 2y29_rama_anchor28_.sym

# number of structures to generate
-nstruct 1000

# points for defining the amyloid plane
# c is going to be the residue fartest away in the
# symmetric system in order for better accuracy
-a 26
-b 24
-c 19

# Secondary structure
-lp 15 16 17
-sz 18 20 25 27
-hx 14 13 12 11 10 9 8 7 6 5 4 3 2 1

# ABEGO sampling
-let GBA

# directories to put the pdb and pos files into
-pdb_dir rama_pdbs
-pos_dir rama_pos

# residues to match from the amyloid sheet
-ar 19 21 24 26

# to visualize in pymol
```

```
-show
-keep_history
```

This file script produces filtered structures and the respective position files (.pos) (.pos files are used in Rosetta Match – see Appendix C.5). One of the filtered structures 2y29_ln5nh14_2x.pdb are used in step 2 (see Appendix C.5).

C.5 Scripts for step 2: Design of a zinc site

The following command was used to run the Rosetta Match Algorithm:

```
match.mpi.linuxgccrelease @2y29_ln5nh14_2x.flags @match.flags
```

The match.flags file is shown below:

```
-run::preserve_header
-ignore_zero_occupancy false
-linmem_ig 10
-no_optH false
-flip_HNQ
-no_his_his_pairE
-nblist_autoupdate
-ignore_zero_occupancy false
-parser_read_cloud_pdb 1
-failed_job_exception false
-output_format CloudPDB
-updown_collision_tolerance 0.3
-bump_tolerance 0.3
-match:euclid_bin_size 0.9
-match:euler_bin_size 9.0
-match:filter_colliding_upstream_residues
```

The 2y29_ln5nh14_2x.flags file is shown below:

```
-in:file:s 2y29_l5nh14232_2x.pdb
-scaffold_active_site_residues 2y29_l5nh14232_sele.pos
```

```
-geometric_constraint_file match.cst
-match:lig_name ZNA
-extra_res_fa ZNA.params
-out:file:output_virtual true
```

The match.cst file is shown below. There are 4 blocks in total but only the first block is shown. To generate the rest of the file copy paste this and change V4 and V2, to 1) V4 and V3 2) V4 and V1 and finally 3) V2 and V1.

```
################ First Block

VARIABLE_CST::BEGIN

        CST::BEGIN
        TEMPLATE::   ATOM_MAP: 1 atom_name: ZN V4 V2
        TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

        TEMPLATE::   ATOM_MAP: 2 atom_name: SG CB CA ,
        TEMPLATE::   ATOM_MAP: 2 residue1:  C

        CONSTRAINT:: distanceAB:  2.34    0.11   10.0  1    1
        CONSTRAINT::    angle_A:  109.5   6.89   10.0  360. 1
        CONSTRAINT::    angle_B:  104.51  8.27   10.0  360. 1
        CONSTRAINT::  torsion_A:  120.0   8.26   10.0  360. 1
        CONSTRAINT::  torsion_B:  0.0       360.0  10.0  360. 4
        CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 4

        #ALGORITHM_INFO:: match
            #  IGNORE_UPSTREAM_PROTON_CHI
      #ALGORITHM_INFO::END

        CST::END

        CST::BEGIN
        TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
        TEMPLATE::   ATOM_MAP: 1 residue3:  ZN
```

```
TEMPLATE::   ATOM_MAP: 2 atom_name: CG OD1 OD2 ,
TEMPLATE::   ATOM_MAP: 2 residue1:  D

CONSTRAINT:: distanceAB:  2.66    0.22   10.0  1    1
CONSTRAINT::    angle_A:  109.5   15.0   10.0  360. 1
CONSTRAINT::    angle_B:  62.32   7.46   10.0  360. 1
CONSTRAINT::  torsion_A:  120.0   15.0   10.0  360. 1
CONSTRAINT::  torsion_B:  -0.15   16.33  10.0  360. 1
CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END

CST::BEGIN
TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

TEMPLATE::   ATOM_MAP: 2 atom_name: OD1 CG CB ,
TEMPLATE::   ATOM_MAP: 2 residue1:  D

CONSTRAINT:: distanceAB:  2.01    0.19   10.0  1    1
CONSTRAINT::    angle_A:  109.5   12.51  10.0  360. 1
CONSTRAINT::    angle_B:  123.86  11.79  10.0  360. 1
CONSTRAINT::  torsion_A:  120.0   13.67  10.0  360. 1
CONSTRAINT::  torsion_B:  172.84  25.79  10.0  360. 1
CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END


CST::BEGIN
TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

TEMPLATE::   ATOM_MAP: 2 atom_name: OD2 CG CB ,
TEMPLATE::   ATOM_MAP: 2 residue1:  D

CONSTRAINT:: distanceAB:  2.01    0.17   10.0  1    1
CONSTRAINT::    angle_A:  109.5   10.45  10.0  360. 1
```

```
CONSTRAINT::    angle_B:  124.23  11.43  10.0  360. 1
CONSTRAINT::  torsion_A:  120.0   10.79  10.0  360. 1
CONSTRAINT::  torsion_B:  183.34  25.81  10.0  360. 1
CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END

CST::BEGIN
TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

TEMPLATE::   ATOM_MAP: 2 atom_name: CD OE1 OE2 ,
TEMPLATE::   ATOM_MAP: 2 residue1:  E

CONSTRAINT:: distanceAB:  2.63    0.23   10.0  1    1
CONSTRAINT::    angle_A:  109.5   15.0   10.0  360. 1
CONSTRAINT::    angle_B:  63.18   7.09   10.0  360. 1
CONSTRAINT::  torsion_A:  120.0   15.0   10.0  360. 1
CONSTRAINT::  torsion_B:  -1.81   17.17  10.0  360. 1
CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END

CST::BEGIN
TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

TEMPLATE::   ATOM_MAP: 2 atom_name: OE1 CD CG ,
TEMPLATE::   ATOM_MAP: 2 residue1:  E

CONSTRAINT:: distanceAB:  2.03    0.20   10.0  1    1
CONSTRAINT::    angle_A:  109.5   10.45  10.0  360. 1
CONSTRAINT::    angle_B:  122.93  13.20  10.0  360. 1
CONSTRAINT::  torsion_A:  120.0   14.05  10.0  360. 1
CONSTRAINT::  torsion_B:  181.38  24.47  10.0  360. 1
CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END
```

```
  CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name:  ZN V4 V2
  TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

  TEMPLATE::   ATOM_MAP: 2 atom_name: OE2 CD CG ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  E

  CONSTRAINT:: distanceAB:  2.02    0.19   10.0  1    1
  CONSTRAINT::    angle_A:  109.5   11.18  10.0  360. 1
  CONSTRAINT::    angle_B:  121.97  12.74  10.0  360. 1
  CONSTRAINT::  torsion_A:  120.0   12.14  10.0  360. 1
  CONSTRAINT::  torsion_B:  181.29  24.96  10.0  360. 1
  CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

CST::END

  CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN V4 V2
  TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

TEMPLATE::   ATOM_MAP: 2 atom_name: NE2 CE1 ND1 ,
  TEMPLATE::   ATOM_MAP: 2 residue1:  H

  CONSTRAINT:: distanceAB:  2.07    0.16   10.0  1    1
  CONSTRAINT::    angle_A:  109.5   9.49   10.0  360. 1
  CONSTRAINT::    angle_B:  125.69  10.98  10.0  360. 1
  CONSTRAINT::  torsion_A:  120.0   10.51  10.0  360. 1
  CONSTRAINT::  torsion_B:  178.07  16.23  10.0  360. 1
  CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

  CST::END

  CST::BEGIN
  TEMPLATE::   ATOM_MAP: 1 atom_name: ZN V4 V2
  TEMPLATE::   ATOM_MAP: 1 residue3:  ZN

  TEMPLATE::   ATOM_MAP: 2 atom_name: ND1 CE1 NE2 ,
```

```
        TEMPLATE::   ATOM_MAP: 2 residue1:  H

        CONSTRAINT:: distanceAB:  2.07    0.16   10.0  1    1
        CONSTRAINT::    angle_A:  109.5   8.05   10.0  360. 1
        CONSTRAINT::    angle_B:  120.24  8.03   10.0  360. 1
        CONSTRAINT::  torsion_A:  120.0   9.36   10.0  360. 1
        CONSTRAINT::  torsion_B:  180.67  16.72  10.0  360. 1
        CONSTRAINT::  torsion_AB: 0.0     360.0  10.0  360. 6

        CST::END

VARIABLE_CST::END
```

C.6 Scripts for step 3: Zinc site optimization and sequence design

To design a sequence and optimize the BB and zinc site the following command was run:

```
rosetta_scripts.mpi.linuxgccrelease   -parser:protocol   intermodel.xml   -s
intermodel.pdb -nstruct 140 -extra_res_fa ZNA.params -output_virtual true
```

This shows the design for the α3β2β-2y29 model intermonomer model (using the outputted match pdbs from Appendix C.5. See Appendix C.6 for the ZNA.params file.

The intermodel.xml script is shown below

```
<ROSETTASCRIPTS>
    <SCOREFXNS>
      # Full atom score funtion #
      <ScoreFunction name="sfx_full" weights="ref2015">

            <Reweight scoretype="atom_pair_constraint" weight="1.0" />
            <Reweight scoretype="angle_constraint" weight="1.0" />
            <Reweight scoretype="dihedral_constraint" weight="1.0" />
      </ScoreFunction>
    </SCOREFXNS>

    <RESIDUE_SELECTORS>
```

```
       <Index name="loops" resnums="109-110,102-104" />
       <Index name="coordinating_res" resnums="65,66,77,79" />
       <Index name="sz" resnums="76,78,83,85" />
    </RESIDUE_SELECTORS>

    <TASKOPERATIONS>
      <OperateOnResidueSubset name="no_design" selector="coordinating_res">
            <RestrictToRepackingRLT />
      </OperateOnResidueSubset>
      <OperateOnResidueSubset name="no_design_or_repack" selector="sz">
            <PreventRepackingRLT />
      </OperateOnResidueSubset>

      ####
      # common practice movers
      <ExtraRotamersGeneric   name="extra_chi"   ex1="1"   ex2="1"   ex1aro="1"
ex2aro="1"/>
      <IncludeCurrent name="include_rot" />
      <ConsensusLoopDesign            name="disallow_nonnative_loop_sequences"
residue_selector="loops" />
    </TASKOPERATIONS>
    <FILTERS>
    </FILTERS>
    <MOVERS>
      <SetupForSymmetry name="sym" definition="2y29.sym" />

      <FastDesign            name="fastdesign"            scorefxn="sfx_full"
task_operations="no_design,no_design_or_repack,disallow_nonnative_loop_sequen
ces,extra_chi" cst_file=" intermonomer.cst " repeats="3">

            <MoveMap name="movemap" jump="false">
                  <Span begin="76" end="76" chi="false" bb="false" />
                  <Span begin="77" end="77" chi="true" bb="false" />
                  <Span begin="78" end="78" chi="false" bb="false" />
                  <Span begin="83" end="83" chi="false" bb="false" />
                  <Span begin="84" end="84" chi="true" bb="false" />
                  <Span begin="85" end="85" chi="false" bb="false" />
            </MoveMap>
```

```
      </FastDesign>

    </MOVERS>

    <PROTOCOLS>
      <Add mover="sym"/>
      <Add mover="fastdesign"/>
    </PROTOCOLS>
</ROSETTASCRIPTS>
```

See Appendix C.7 for the 2y29.sym file. The intermonomer.cst are seen below (this is for the intermonomer model):

```
AtomPair ZN 87 NE2 77               HARMONIC 2.07  0.16
Angle    ZN 87 NE2 77 CE1 77        HARMONIC 126   11
Dihedral ZN 87 NE2 77 CE1 77 ND1 77 HARMONIC 178   16
Angle    V4 87 ZN  87 NE2 77        HARMONIC 109.5 8
Dihedral V2 87 V4  87 ZN  87 NE2 77       HARMONIC 120   11

AtomPair ZN 29 NE2 66               HARMONIC 2.07  0.16
Angle    ZN 29 NE2 66 CE1 66        HARMONIC 126   11
Dihedral ZN 29 NE2 66 CE1 66 ND1 66 HARMONIC 178   16
Angle    V4 29 ZN  29 NE2 66        HARMONIC 109.5 8
Dihedral V3 29 V4  29 ZN  29 NE2 66       HARMONIC 120   11

AtomPair ZN 87 CG  65               HARMONIC 2.66  0.22
Angle    ZN 87 CG  65 OD1 65        HARMONIC 62    7
Dihedral ZN 87 CG  65 OD1 65 OD2 65 HARMONIC 0     16
Angle    V4 87 ZN  87 OD1 65        HARMONIC 109.5 15
Dihedral V1 87 V4  87 ZN  87 CG  65       HARMONIC 120   15

AtomPair ZN 87 ND1 79               HARMONIC 2.07  0.16
Angle    ZN 87 ND1 79 CE1 79        HARMONIC 120   8
Dihedral ZN 87 ND1 79 CE1 79 NE2 79 HARMONIC 181   17
Angle    V2 87 ZN  87 NE2 79        HARMONIC 109.5 8
Dihedral V1 87 V2  87 ZN  87 NE2 79       HARMONIC 120   9
```

C.7 Symmetry setup

A script was developed to make the symmetry file used in step 3. The symmetry file used in step 1 must be modified from the one used in step 3. Starting from a file with 3 2 β-strand scaffolds from the amyloid structure of interest (**Figure C-1.a**). The following script (for the PDB ID 2Y29) can be run to generate a symmetry file containing 6 monomers as is shown in **Figure C-1.b.**

```
Python sym_file_maker_2y29.py
```

The script uses the center of mass of each of the 3 subunits to make the symmetry file. This is the only thing that should be changed inside the python file to make it general for the other PDB IDs.

a) **b)**

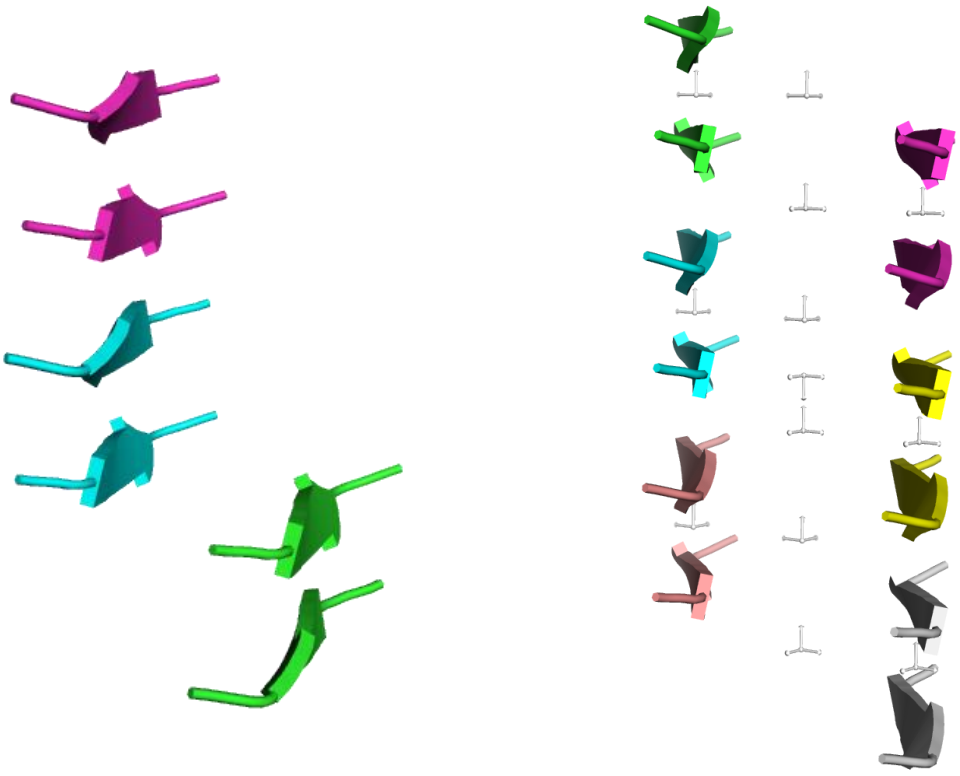


Figure C-1: Creating symmetry files. a) the script starts with a PDB file as shown. This example is for the PDB ID 2Y29. **b)** The symmetry made by the script is visually shown. It contains 13 coordinate systems.

The symmetry file used in step 1 is seen below:

```
symmetry_name 2y29_2sub1side
E = 2*VRT_0_high_base + 1*(VRT_0_high_base:VRT_0_high_base)

anchor_residue 28
virtual_coordinates_start
xyz VRT_root            1,0,0 0,1,0 6.3095,-3.0175,0.0
xyz VRT_1_high          1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    -5.678,-
3.0175,0.0
xyz VRT_1_high_base     1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    -5.678,-
9.053,0.0
xyz VRT_0_high    1.47159045597e-16,-1.0,0.0 -1.0,0.0,0.0 3.912,-3.0175,0.0
xyz VRT_0_high_base     1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    3.912,-
9.053,0.0

virtual_coordinates_stop
connect_virtual JUMP_root_0_high         VRT_root VRT_0_high
connect_virtual JUMP_0_high_1_high        VRT_0_high VRT_1_high
connect_virtual JUMP_1_high_1_high_base   VRT_1_high VRT_1_high_base
connect_virtual JUMP_0_high_0_high_base   VRT_0_high VRT_0_high_base
connect_virtual JUMP_1_high_base_SUBUNIT  VRT_1_high_base SUBUNIT
connect_virtual JUMP_0_high_base_SUBUNIT  VRT_0_high_base SUBUNIT

set_dof JUMP_root_0_high x y z
set_dof JUMP_0_high_1_high y(4.795)
set_dof JUMP_0_high_0_high_base x(-6.0355)

set_jump_group JUMPGROUP1 JUMP_0_high_1_high
set_jump_group JUMPGROUP2 JUMP_0_high_0_high_base JUMP_1_high_1_high_base
set_jump_group JUMPGROUP3 JUMP_0_high_base_SUBUNIT JUMP_1_high_base_SUBUNIT
```

The symmetry file used in step 3 is given below:

```
symmetry_name 2y29
E    =    6*VRT_0_high_base    +    4*(VRT_0_high_base:VRT_1_high_base)    +
3*(VRT_0_high_base:VRT_0_low_base)   +   2*(VRT_0_high_base:VRT_1_low_base)   +
2*(VRT_0_high_base:VRT_n1_low_base)
```

```
anchor_residue 28
virtual_coordinates_start
xyz VRT_root            1,0,0 0,1,0 6.3095,-3.0175,0.0
xyz VRT_1_high          1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    -5.678,-
3.0175,0.0
xyz VRT_1_low           0.0,1.0,0.0 -1.0,0.0,0.0 -0.883,-3.0175,0.0
xyz VRT_1_high_base     1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    -5.678,-
9.053,0.0
xyz VRT_1_low_base      0.0,1.0,0.0 -1.0,0.0,0.0 -0.883,3.018,0.0
xyz VRT_0_high    1.47159045597e-16,-1.0,0.0 -1.0,0.0,0.0 3.912,-3.0175,0.0
xyz VRT_0_low     0.0,1.0,0.0 -1.0,0.0,0.0 8.707,-3.0175,0.0
xyz VRT_0_high_base     1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    3.912,-
9.053,0.0
xyz VRT_0_low_base      0.0,1.0,0.0 -1.0,0.0,0.0 8.707,3.018,0.0
xyz VRT_n1_high   1.47159045597e-16,-1.0,0.0 -1.0,0.0,0.0 13.502,-3.0175,0.0
xyz VRT_n1_low    0.0,1.0,0.0 -1.0,0.0,0.0 18.297,-3.0175,0.0
xyz VRT_n1_high_base    1.47159045597e-16,-1.0,0.0    -1.0,0.0,0.0    13.502,-
9.053,0.0
xyz VRT_n1_low_base     0.0,1.0,0.0 -1.0,0.0,0.0 18.297,3.018,0.0
virtual_coordinates_stop
connect_virtual JUMP_root_n1_low          VRT_root VRT_n1_low

connect_virtual JUMP_1_low_1_high         VRT_1_low VRT_1_high
connect_virtual JUMP_0_high_1_low         VRT_0_high VRT_1_low
connect_virtual JUMP_0_low_0_high         VRT_0_low VRT_0_high
connect_virtual JUMP_n1_high_0_low        VRT_n1_high VRT_0_low
connect_virtual JUMP_n1_low_n1_high       VRT_n1_low VRT_n1_high

connect_virtual JUMP_1_high_1_high_base   VRT_1_high VRT_1_high_base
connect_virtual JUMP_1_low_1_low_base     VRT_1_low VRT_1_low_base
connect_virtual JUMP_0_high_0_high_base   VRT_0_high VRT_0_high_base
connect_virtual JUMP_0_low_0_low_base     VRT_0_low  VRT_0_low_base
connect_virtual JUMP_n1_high_n1_high_base VRT_n1_high VRT_n1_high_base
connect_virtual JUMP_n1_low_n1_low_base   VRT_n1_low VRT_n1_low_base

connect_virtual JUMP_1_high_base_SUBUNIT  VRT_1_high_base SUBUNIT
connect_virtual JUMP_1_low_base_SUBUNIT   VRT_1_low_base SUBUNIT
```

```
connect_virtual JUMP_0_high_base_SUBUNIT  VRT_0_high_base SUBUNIT
connect_virtual JUMP_0_low_base_SUBUNIT   VRT_0_low_base SUBUNIT
connect_virtual JUMP_n1_high_base_SUBUNIT VRT_n1_high_base SUBUNIT
connect_virtual JUMP_n1_low_base_SUBUNIT  VRT_n1_low_base SUBUNIT

set_dof JUMP_root_n1_low x y z
set_dof JUMP_n1_low_n1_high y(4.795)
set_dof JUMP_n1_high_0_low y(4.795)
set_dof JUMP_0_high_0_high_base x(-6.0355)

set_jump_group JUMPGROUP1 JUMP_n1_high_0_low JUMP_0_high_1_low
set_jump_group      JUMPGROUP2      JUMP_n1_low_n1_high      JUMP_0_low_0_high
JUMP_1_low_1_high
set_jump_group   JUMPGROUP3   JUMP_0_high_0_high_base   JUMP_n1_low_n1_low_base
JUMP_n1_high_n1_high_base      JUMP_0_low_0_low_base      JUMP_1_low_1_low_base
JUMP_1_high_1_high_base
set_jump_group JUMPGROUP4 JUMP_0_high_base_SUBUNIT   JUMP_n1_high_base_SUBUNIT
JUMP_1_high_base_SUBUNIT    JUMP_0_low_base_SUBUNIT    JUMP_n1_low_base_SUBUNIT
JUMP_1_low_base_SUBUNIT
```

C.8 Zinc parametrization

The zinc site was parametrized as below and saved as ZNA.params.

```
NAME ZNA
IO_STRING  ZNA Z
TYPE LIGAND
AA UNK
ATOM ZN   Zn2p  X    2.00
ATOM  V1  VIRT  X    0.00
ATOM  V2  VIRT  X    0.00
ATOM  V3  VIRT  X    0.00
ATOM  V4  VIRT  X    0.00
BOND ZN    V1
BOND ZN    V2
BOND ZN    V3
BOND ZN    V4
NBR_ATOM ZN
NBR_RADIUS 0.01
```

```
# tetrahedral geometry with 2.2 distance from center to vertex
# chirality is defined so that viewing from V1 to ZN, V2-V3-V4 are in clockwise
direction.
ICOOR_INTERNAL   ZN      0.000000    0.000000    0.000000  ZN    V1    V2
ICOOR_INTERNAL    V1     0.000000    0.000000    2.200000  ZN    V1    V2
ICOOR_INTERNAL    V2     0.000000   70.500000    2.200000  ZN    V1    V2
ICOOR_INTERNAL    V3   120.000000   70.500000    2.200000  ZN    V1    V2
ICOOR_INTERNAL    V4  -120.000000   70.500000    2.200000  ZN    V1    V2
```

D PDB List

D.1 PDB list for amyloids

Class 1

1YJO, 1YJP, 2BFI, 2OLX, 2OMM, 2ON9, 2ONW, 3DGJ, 3FTR, 3FTK, 3FTL, 3HYD, 3FVA, 3NVF 3NVG, 3NVH, 3OVL, 3PPD, 3Q2X, 4NIO, 4NIP, 4NP8, 4R0P, 4R0U, 4R0W, 4RIK, 4RIL, 4RP7, 4ZNN, 5KNZ, 5K2E, 5K2H

Class 2

2OL9, 2Y3J, 2Y3K, 4NIN, 4RP6, 4XFO, 5E5X, 5K7N

Class 4

2ONV, 2ONX, 3FPO, 3SGS, 4QXX

Class 5

3LOZ, 4E0K

Class 6

3FTH, 3NVE, 3PZZ, 4OLR, 4ONK, 4W5Y, 4W67, 4W71, 5E61

Class 7

2OMP, 2OMQ, 2Y2A, 2Y29, 2Y3L, 3FR1, 3OVJ, 3OW9, 4TUT, 4XFN, 5E5V, 5E5Z, 5KO0, 5TXJ

Class 8

2OKZ, 2ONA, 3FOD, 3NHC, 3NHD, 4UBY, 4UBZ, 4W5L, 4W5M, 4W5P, 4WBU, 4WBV, 5TXH

Unclassified

4E0L, 4E0M

D.2 PDB list for proteins containing zinc

PDB list used for analyzing mono tetrahedral zinc sites and all zinc sites (412 PDB files)

3MI9, 3MSU, 5IQK, 4G26, 3ALR, 3CNG, 1XKI, 3O9X, 1BI0, 2AC3, 3LKM, 258L, 5ERC, 1X6M, 1A73, 1GR0, 4HC9, 3KNV, 4U08, 2Z45, 1T8H, 4GNE, 5ELK, 3H5N, 3HCJ, 3O9P, 2FPR, 1WUR, 3FM2, 3HCS, 3SWR, 3X0D, 2XQC, 2DS5, 1L7O, 3Q7C, 2BZ1, 4LJ0, 4I1F, 5YEH, 1ZY7, 5UAM, 1NNQ, 3TBG, 4LMG, 2PVE, 1YEJ, 4MO1, 1XER, 4PYS, 1X0T, 2FEA, 3TN2, 4TXD, 4L7X, 3G1P, 4FGL, 4KFV, 3VDP, 1RYQ, 3FQM, 2OO4, 1SVM, 3OOI, 3HUG, 2A1K, 1L0Y, 2IV0, 3AVR, 3U1L, 4DR8, 3FLO, 4NM6, 5HEE, 3T9O, 4HVL, 2B9D, 2X4H, 2AS9, 1D9D, 3T92, 1IA6, 2XCM, 1Z83, 3L00, 2HVW, 3RUI, 2AYD, 4BN4, 2V0C, 4IUP, 2CJL, 3UJZ, 2CS7, 4MB7, 1W5Q, 3PJN, 3IUF, 1XRU, 2RA6, 2CJS, 3MWM, 1QF8, 2WOJ, 2FGY, 2J7J, 3H0L, 2XOC, 4IXJ, 3D00, 1HK8, 2FYG, 3L8H, 3U50, 2PPT, 5DKA, 1RUT, 2J13, 2YVR, 2J6A, 2I2X, 3QL9, 3VPB, 2FE3, 1U0A, 4XB6, 2GAG, 3NM8, 4Z4M, 1LBU, 2GVI, 1V33, 3K9T, 3FID, 3T7L, 1K3X, 4Y7L, 3MWP, 3S6L, 1KWG, 1VZY, 1A5T, 2QEE, 3W0F, 4TVR, 3ORU, 2V9K, 2F44, 1U5K, 5H3J, 1KK1, 2XB4, 1R44, 1NW2, 3MLN, 5UVM, 2CH9, 4WAI, 1P5D, 3M7K, 3RSN, 3P2A, 3O4N, 4TSD, 2F9I, 4HHJ, 3HKO, 2A5H, 5AF0, 1CG2, 5G33, 4YBG, 1EF0, 4NZG, 5KKQ, 2VF7, 1P9R, 2HU9, 4C4A, 3G9M, 4BF7, 5MLT, 3QMD, 2XZL, 5GL7, 3D2Q, 4GEL, 4GBM, 3SWN, 5IJL, 1K6Y, 4DGU, 5KZZ, 5DIN, 3VGL, 4GV3, 3SU6, 3LS9, 3ZVS, 4WTB, 2WB0, 4XXB, 1C7K, 3MJH, 3LS1, 1V4P, 3U9G, 2RIQ, 2QKD, 5AH5, 1XTM, 4ZDT, 4AXD, 4F6M, 5AWS, 3UEJ, 1E87, 5GOX, 3VQJ, 2HQH, 4K0D, 4NN2, 3P1V, 1Y02, 3EH1, 4ME3, 2D5B, 2V08, 5KAY, 2OLM, 2VAD, 1T9H, 3L8E, 2ORW, 3PY7, 1V47, 3NA7, 3IR9, 4I1S, 2ESL, 3VK6, 2GLZ, 5T77, 5URB, 4CQ1, 1D1T, 3FVZ, 1I3J, 3WUP, 2J9U, 1NO5, 4GQT, 3ASL, 2XGW, 3GA8, 2P09, 5ELH, 3BOC, 3RMQ, 3LRQ, 4YZI, 2OSO, 5EU8, 2UX1, 1Y7W, 3C37, 3A1B, 2CZR, 2XEV, 5OQD, 1PZW, 2F6S, 2BJR, 4GLX, 3VHS, 3HNA, 3IRB, 2VRS, 2ZC0, 3DGD, 3EBE, 1HXR, 3V4K, 1OQJ, 4PAW, 2ZZE, 3UNG, 1OYW, 1K24, 3EER, 1VK6, 3TS2, 3K6J, 4ESN, 4FO9, 2C1I, 5IKJ, 2XIG, 2X5R, 3L0A, 5CH7, 3AXS, 2JIG, 3C8Z, 2OFK, 4FVD, 3UMI, 4IJD, 5B2Q, 4A46, 5MMD, 1J98, 5HKO, 1Y8Q, 1CVR, 1T4W, 1Y93, 1MWQ, 3BOF, 3RAY, 1BTK, 4HY4, 5LPI, 4BOL, 3UX8, 2V1X, 2I9W, 1RUR, 1JW9, 4RUW, 1F4T, 3RPD, 1RMD, 3GOX, 4ESJ, 4N0G, 3OYM, 1DVF, 4DT3, 4KYW, 4FYY, 3RYM, 4KMB, 4QN1, 3V1E, 3ERP, 1MXG, 3C10, 4F9C, 2PSY, 1VSR, 4K98, 2G2N, 4WB7, 4R2Y, 2PQ8, 4GGJ, 3VRH, 3OD8, 3LZN, 5CUO, 2OAJ, 1XB0, 5GHA, 2ZE7, 5K2M, 1FN9, 3TTC, 2ZNR, 1NZJ, 4LQX, 3KDE, 3LCZ, 5JP2, 1E7L, 4FC8, 2F4M, 4YWQ, 1A1H, 1U8B, 1XAF, 2VQM, 4ODR, 4M8O, 2V89, 4Z7R, 3TIO, 2Z94, 2X5C, 2GFO, 2GMW, 1P6O, 4BBQ, 1TWF, 4ZLH, 4GZN, 4ADN, 1D0Q, 5KD2, 3D06, 1XCR, 1OZJ, 1HWT, 3PLW, 1LLM, 2HH5, 3IFU, 5I01, 4A2V, 2OIK, 3U52, 1XWY,

PDB list used for analyzing all zinc sites only (2533 PDB files)

5NC9, 4KXQ, 5YKN, 1K4P, 1LR0, 4GUJ, 1YKF, 5HNM, 2VAC, 4OY6, 4GKV, 5KL2, 4FM7, 3U6C, 1GVF, 4OJX, 2W0M, 1R2Z, 2OWA, 5YFB, 2Y28, 2Q4Z, 3EQN, 1VYK, 5KCP, 5W0R, 4CO9, 5AHO, 5BQ1, 2OV3, 4GER, 5AFS, 5L19, 2XWC, 1Q74, 3U6P, 1LI5, 5LYN, 3CIZ, 3PE8, 4JIX, 2QH0, 4ILK, 5TRB, 1PCX, 3Q87, 1UBD, 3O0M, 1H2B, 1BY4, 3S5M, 3QMG, 1IQ8, 3G27, 5W8I, 1YQD, 4D1L, 3BAL, 5LWH, 5HH7, 2ZTX, 2FLI, 3OG8, 2D30, 3SZQ, 3BK2, 4K9A, 3PVN, 4ZO2, 2B3J, 5B8D, 2YGT, 4MTU, 4C98, 5NAI, 3OVG, 4RH4, 1HR6, 2O5I, 1KAE, 3BB6, 3H9C, 6EQP, 4ZPJ, 4N0N, 4LEV, 6C4V, 4CN2, 2WBS, 3O7A, 4RQT, 1TO4, 4IN9, 5A0T, 4NPD, 3B6P, 4R7E, 1RTQ, 3ST1, 2O6D, 4LCQ, 3X1L, 3ZWF, 5UN7, 3IE4, 2GWG, 5TP4, 1PTQ, 4L92, 5CGZ, 4RUD, 3EB9, 3CBB, 3S9C, 1OAO, 2WWO, 3EH2, 1TAF, 4XCH, 3GIP, 4S25, 2HBV, 4TNT, 3LMC, 3VAT, 2KZZ, 4COS, 3IUU, 4WCN, 4RNA, 3P57, 3GQ5, 4J4K, 2WGQ, 5VKT, 1ZV8, 1F2I, 1ATL, 3UBG, 4HIF, 5TD5, 4UOZ, 3IU6, 4MTS, 4GUT, 4G7A, 1VJ0, 1YIX, 2DI3, 1PCA, 5EHF, 5C1Z, 4NUR, 2Y20, 1IA9, 2Z30, 4N9V, 3FJU, 2DDF, 4DYG, 3NE8, 3TWO, 5G0X, 5H4J, 3E9Q, 4MSX, 6B0P, 5MEY, 1KYS, 2DQ4, 2FAW, 2YKG, 4QGS, 1SW1, 3BON, 4HEX, 2RF3, 1GEN, 2QNW, 3MPZ, 5B3R, 4DO7, 2ZSG, 3QSU, 3W0T, 2EJC, 1DYQ, 4N3V, 2ZZW, 5M1Q, 5FD3, 4ZFZ, 2A8N, 4OFC, 3CAO, 2YIK, 4D1T, 5E7W, 1AU1, 4MN6, 3PFO, 3MDU, 4FKB, 3OJ7, 1AST, 1XB8, 5KL5, 3NEH, 4X3O, 2W5Y, 3OA4, 5CDE, 3EIP, 3H7T, 1JWQ, 4QQ4, 3IJ6, 1GKP, 4TZH, 5WZ3, 4Z3C, 2W4L, 5AUN, 3UHM, 4CXP, 4OTE, 1VFY, 1DV6, 3T65, 1I7W, 3PFS, 3FCA, 3U31, 5AB0, 4EEZ, 5W83, 3G4G, 5B5Z, 5E68, 3SPD, 2VPB, 2XEU, 5HX5, 4Z65, 2CC0, 4XQA, 3HXS, 1AA0, 3LWV, 4CN7, 4CC9, 3T9K, 3ZYQ, 5M6N, 5WRI, 2Z00, 3L1E, 3C58, 3LOT, 5XZG, 1A7W, 2QFA, 4H3S, 3QL6, 3P3C, 2PGF, 4YYE, 4ISM, 3RHG, 3GS2, 3DSS, 2ERO, 2HEK, 1TUP, 4Z6K, 4FRU, 1T92, 5LY0, 3UHJ, 1T8K, 3FU1, 4KJM, 4R9G, 2EPF, 3S2J, 3UN6, 3HAX, 3MEQ, 4O64, 1GL4, 1LBA, 5JP6, 3ORJ, 4OX3, 4XUK, 4MUQ, 2QM1, 2VS0, 3M3G, 1Q1A, 3UH0, 5VYD, 5OF9, 5NUS, 5TDR, 3Q2R, 1CTT, 4A3N, 3G97, 1OLP, 5K5M, 2J44, 3PSQ, 4HJE, 2C1C, 4PZI, 3DC8, 3M4R, 4UB9, 1XVX, 3HWP, 2HSI, 4JIV, 3PN3, 4PXY, 4OK4, 5F7R, 1U0B, 5EXH, 2Z72, 1K3W, 1XC8, 3MMK, 4UVQ, 2W3Z, 4P6P, 4MN5, 4QHJ, 2BSF, 1UDV, 3UWB, 3SEW, 1KWM, 3PU7, 1MQ0, 5KD5, 5B5O, 5NOF, 2CBI, 2AQP, 5DOR, 5A0R, 1EVX, 4HN5, 3PDT, 3LEC, 3CZX, 4CKV, 3H50, 5SYT, 3U24, 2Y7I, 2WTA, 4K1T, 3SAR, 3O94, 5C0Q, 1M4L, 3IT7, 1JQ5, 5DQ0, 3ZFK, 2V4B, 3A9L, 3GLJ, 1DTD, 3LAT, 1HKK, 3G7L, 2NXF, 1C8Y, 3GPP, 4FVK, 6B0O, 5LKT, 2ISW, 3NVO, 4YF4, 4Y4O-PDB-BUNDLE2, 3SUB, 5A2R, 5GPO, 5GND, 4R2C, 1TAQ, 3RJA, 3K6D, 5L2L, 4MT2, 1QUA, 4NT9, 4JIU, 4WJB, 4FIL, 5ODG, 3KDK, 4M3O, 3B0X, 2OB3, 4HDT, 5KE7, 4S2R,

4L6H, 3WBH, 3EA1, 4L05, 3UX3, 2OBA, 4B87, 4KJG, 5DMM, 2CDC, 1W4R, 5ABS, 4P2Y, 5G0F, 5WS2, 5TDA, 4DII, 1QUM, 5KIV, 4IE5, 2OFI, 5OL0, 5WE0, 1F3Z, 3ZBW, 5M45, 3M7P, 2VZ5, 4LA2, 3LL8, 3QGV, 4C4O, 5JLB, 4RN7, 3GJ3, 2V1Z, 5N1P, 2ZC2, 2YC5, 4LZU, 3VTG, 2Z1A, 5EMI, 3K2G, 3E3I, 3S8P, 4UCI, 4R2J, 1PGU, 2GX8, 2PG3, 2VXX, 5K3Q, 5KEA, 3HMF, 2VME, 5I4M, 4YXP, 3MK1, 3DMO, 4KOY, 5KEF, 2ZUM, 3LWR, 2OOG, 4MPH, 1Q0E, 4Z0O, 2FVK, 1R0O, 5DEU, 2EBN, 5VAB, 4RLR, 4AQJ, 3MTW, 1SHW, 2JG6, 1DU3, 5T57, 3DYN, 4GV9, 1E1H, 3Q05, 5K5I, 3S3M, 3GO9, 4F3W, 5BUO, 3NH4, 3LLX, 4EGE, 2CUA, 3U43, 4Y4P-PDB-BUNDLE1, 1ZB7, 4OLS, 4CVR, 1FA5, 4UPL, 5EGH, 4OND, 1YG9, 2VQX, 1PQ4, 4AXL, 1ZIN, 2CKL, 3DI4, 4FYT, 4DV8, 2BCO, 3GQ4, 3HL5, 2X4K, 1SU3, 3KUI, 3B08, 3LQ0, 5KDJ, 3RQZ, 5VSC, 3CP0, 3Q94, 4EJ6, 5FC1, 4YPE, 3RZV, 5FJK, 1G2F, 5V3M, 4V33, 4OBI, 5CM2, 3DHA, 1J6W, 5E69, 5WXH, 4MCA, 2R2V, 5M8S, 4P53, 3OAJ, 5DVX, 1BC8, 4JNJ, 4RQM, 3KL7, 4AF1, 2GYQ, 1F8F, 5UB8, 1WNI, 4MMO, 4CSH, 3KAO, 5GQ1, 5TNX, 2PLI, 4PUC, 5EGB, 1XEM, 2O6P, 2VM5, 2FR5, 2Q2L, 5D0I, 3IDV, 4E5V, 3LJM, 2EHS, 1Y9A, 4YO9, 4A7U, 1KSP, 1OHT, 3ZVZ, 3G8Q, 4F9W, 5M0T, 4JEA, 4EG2, 4IS1, 1HW1, 4NS5, 4YBB-PDB-BUNDLE1, 1E4C, 4I2D, 8RNT, 4ZK0, 2PMP, 5TDC, 4L58, 2RHK, 1L2D, 2QW5, 4M9E, 3R3R, 3GSH, 2X7M, 5WDX, 1R42, 4CN3, 3QM3, 3Q4R, 1Z1N, 4Z8Y, 3FED, 3N55, 3PW3, 4JY6, 4PA5, 1P47, 4QME, 5LSU, 3LJU, 4J44, 4FUK, 3T3O, 3VOW, 3U6O, 5H7W, 4A4J, 5WGI, 4YN2, 4HE2, 2QIC, 3LUB, 2Y3G, 1A1J, 4LY4, 3AII, 1K7I, 5C5T, 3KJH, 2WZN, 5KDW, 3O8C, 3UA7, 1KRP, 5XFP, 4K97, 2W8S, 2YMK, 1PV9, 4TOR, 1JAZ, 2XSC, 2QF7, 5G32, 5TCD, 1WDK, 2CF5, 2HC9, 5AH1, 3CE1, 4XWW, 4MVT, 5K5J, 4WEC, 2RAG, 5F8I, 1OAL, 2RHS, 1A8L, 3SAT, 5J47, 6F9T, 5MVW, 2JHO, 5MKW, 1UWZ, 1EE8, 2AA4, 1KL9, 5XN9, 3DWD, 4Z4P, 1I6N, 2B3Z, 3VUS, 2A0B, 5CBZ, 2A21, 4LZ5, 1Y44, 2BOA, 3L9Y, 3PKI, 4BIN, 1WCZ, 4C1G, 1A7T, 3QFK, 1IAU, 1XOC, 1J30, 3A32, 1IM5, 4O5O, 3FMU, 2AQ2, 5H9F, 2QQ4, 3CQJ, 3HYG, 4K4S, 4ENL, 3QDH, 4CQB, 4F14, 4MUR, 4GVE, 2HF9, 2ICS, 2V9L, 3UW2, 5GM3, 5XT5, 4AQL, 5HZD, 1JK2, 5HRT, 1WEJ, 2Y6D, 3LCN, 3V77, 1VLI, 1HBN, 4FAI, 1UUF, 3LX3, 5AX7, 1JVB, 5UMF, 3WS6, 4OKO, 1Y79, 2IXD, 5GV3, 2DKD, 2EV6, 1ONW, 2AB4, 4ZUR, 2E7Y, 6ENX, 5W0N, 1R23, 3ZE3, 3V43, 4CT0, 2FZW, 4CA1, 3EE6, 3UK3, 3WV6, 1WWR, 1CZ0, 3L6N, 5VM2, 3CX3, 2EY4, 4P36, 4H4L, 5W3V, 3SWF, 4ZEJ, 1Q2L, 5TS5, 2J6X, 4BE2, 4ASK, 5KL7, 5CC1, 4G09, 3TQO, 4NDF, 3T6P, 3AAL, 2WBU, 1CW0, 5FI3, 3PM6, 1F1M, 3IGL, 4LEY, 4Z1D, 2Q7S, 3IO2, 5VLC, 4RN5, 5OPZ, 3NO5, 1PL8, 5VC9, 1MKM, 5C28, 1TDZ, 2NX9, 4B6D, 4Y8A, 4RFL, 4Y93, 3MBJ, 4V20, 3LQB, 1KEQ, 3M1M, 5CHU, 5I4J, 4C7A, 4R2E, 3G8U, 3HSV, 3RCM, 1U1J, 5B8I, 3DLJ, 1E4M,

5KEB, 3FWX, 2PKP, 5CBY, 6EOM, 1VF2, 4YH8, 4R2D, 2GSO, 1YJ6, 3MKV, 3TKK, 2P9X, 6AVS, 4JLX, 3Q01, 6BLW, 4LMY, 1IWL, 4CN5, 5KZJ, 3ZGZ, 5TSA, 3CNQ, 1LFW, 1HFE, 1IJ0, 3QAY, 2XS7, 5B2P, 3WI9, 4KX7, 3AKQ, 3S2S, 5T67, 3G4H, 3DFI, 2WC8, 6AN0, 1FD9, 1NLC, 2PQ3, 5LSV, 4IUW, 2KFZ, 2Y7E, 3ML5, 1BT0, 1EB6, 1NOY, 1XAH, 5IIX, 1S2Z, 1TA9, 4OAQ, 4RP7, 1O06, 4C1Q, 5XM5, 1ZSW, 1V3W, 5DAG, 5TXU, 3W8K, 1Y7P, 1XX4, 4G9Z, 1J7N, 3PNZ, 1KQ9, 4HI8, 1NWP, 3U04, 3SKS, 3GA3, 3GQ3, 3EQT, 3E7G, 5LS4, 3GO8, 2F4L, 1LU0, 1N08, 4K08, 5KBE, 4UEJ, 1K51, 1B66, 1DDZ, 3GI1, 4AXV, 4KA7, 1BUD, 1W7V, 2DGE, 3GP1, 3Q1D, 1YB0, 2HUE, 1P0F, 5MX9, 3FGG, 3MI3, 3ZPG, 3RFN, 1V72, 5CJF, 2A97, 1GX1, 1LAT, 1F82, 2OWO, 3FL2, 5GJ3, 4OJA, 1YLO, 1JQG, 3LPE, 4N4G, 5W9Q, 2IOI, 5FB3, 3U92, 3SBJ, 5FRU, 2Y43, 1S3Q, 2V8H, 1K07, 2C36, 5WGX, 2ORV, 3QC0, 4XLI, 3GRI, 5FNP, 1VJE, 5G41, 4TMU, 2AFW, 2IW0, 5MW8, 4LX9, 2C7A, 1V8Y, 1RHF, 4W2F-PDB-BUNDLE3, 5YHY, 4O6S, 5FWJ, 1AOL, 4YKL, 2D5M, 2Y4Y, 4Y4P-PDB-BUNDLE2, 1DG6, 4F0R, 1FFY, 2R3A, 1AH7, 3G9O, 4C8I, 1WKQ, 5A3Y, 3HCT, 4F9U, 3G9P, 5JBG, 5UPD, 3II1, 2AU3, 5JAJ, 1ZFP, 4C3E, 4ATY, 4LIK, 2O36, 1ZEI, 2WXU, 5UMH, 3L9W, 5OJJ, 3D0A, 1IO0, 1Y13, 5V3J, 3G6Q, 4YBM, 5A7M, 5I81, 3ANU, 4Y4P-PDB-BUNDLE3, 1QTW, 3E7J, 4Q7R, 1T0B, 3G6P, 1VDD, 4I68, 5XZE, 2CE7, 3RDR, 4H2W, 4H2G, 3MEN, 4OLN, 4U3E, 3O0F, 1IAG, 3TTY, 4MTZ, 5M2H, 3GTT, 3BKF, 4JH2, 4QL5, 4KFU, 4WD8, 1QH5, 5HDA, 2ZU2, 3T3W, 1G5C, 1JK1, 1EI6, 5U8O, 1TSR, 3EYY, 2VU4, 5FQD, 4MHN, 4ON1, 2Y1H, 1FR9, 1Z3A, 2W9M, 3FYL, 2P2L, 2ZWS, 3DWB, 2H6T, 4MTE, 2HBA, 3CLH, 5B78, 4Y4O-PDB-BUNDLE3, 5KCI, 4FVL, 3L7T, 1U9K, 3C3U, 3QMI, 3AYU, 3E02, 2Q1Z, 2EXU, 3WWL, 4TWV, 4HX2, 4K99, 4AW8, 1RO5, 4L8H, 6B0R, 1FNO, 1MFT, 4DXB, 2D0W, 3G6T, 2Y33, 2AK1, 4K9B, 4HP3, 3BVU, 2G9H, 2IMZ, 2UVL, 1TY2, 5JMY, 4IBW, 1M55, 4GWD, 2AI2, 3CNE, 2XJY, 3LGD, 1XM8, 1VQ2, 3CSV, 1GHQ, 5JB2, 2ANU, 5J1L, 4X2U, 1A6Y, 4FVY, 4N6P, 3EDH, 2ADY, 4WK7, 3F5L, 2VNF, 2YHW, 3ZQ6, 3MK5, 3U6L, 4QNL, 3US0, 4PIE, 4FUO, 1BYF, 5N6F, 4Z40, 5TKM, 5KZ6, 2UZH, 1L6J, 2E47, 2HZC, 4OV4, 1EUC, 5G6T, 1MEY, 2ZWI, 4FUU, 3LAS, 1R61, 4DD8, 3K7L, 1LML, 1MZB, 3CXK, 3PFE, 2QPX, 6C49, 5EMC, 1Q7H, 2W3Q, 5H0M, 2I3H, 5VA4, 2HCN, 1AYE, 3CVD, 3VUV, 1ESO, 4LAD, 2GZJ, 1Z9P, 5M6E, 5UD5, 4IC3, 2G84, 2Z26, 2FBH, 5VEO, 5VDC, 4L9P, 4K96, 3M4G, 4UBQ, 5E0Y, 3WX7, 4EHC, 3R2N, 5G57, 4GGG, 4QF2, 5IN2, 3EII, 4KNN, 3S2Q, 4W2F-PDB-BUNDLE1, 2FGE, 2IYB, 5TJ3, 5UXS, 5G3Y, 4O6A, 4BZ7, 4NJ9, 2DT8, 1HI9, 4DYK, 4G4N, 3PLQ, 5CXK, 3OJ6, 1V7Z, 1IBQ, 2F5O, 5KE8, 2Y0O, 3H8G, 3BYW, 3U9W, 1CYQ, 4XNV, 2DVT, 4JQP, 4UF4, 1HP1, 3G64, 4OGE, 5GUT, 5WLK, 2EUL, 1M2K, 4B9P, 2ZNE, 3BKM, 2DFV, 5IN3, 2PIG, 2Y6E, 2POI, 4COB,

4JJJ, 5KDS, 3MVI, 3M4W, 1PTM, 4D7V, 5VI6, 1G71, 5V3G, 4Z67, 1Q3B, 5JOH, 1ZUD, 2XOD, 4Q1L, 2Z8F, 1L1Z, 4M0C, 3W20, 4PD2, 4TQT, 2F5Q, 3H68, 5VA0, 4OJ5, 2E6L, 1A1G, 4R2R, 2IGI, 4NDH, 4AO7, 2A2Q, 5H7R, 3QYN, 4C5W, 1DY1, 5E6A, 1PV8, 4ILO, 3ISZ, 5YJ0, 2BYO, 4D1P, 3NGJ, 4AR9, 3B4N, 4Z8W, 3QY1, 1GHX, 3IPJ, 1V8D, 2FZ6, 4WSQ, 3F2C, 4H8Q, 6ASD, 4RNZ, 3E38, 4GAA, 5JMU, 3NQX, 2XL9, 2UUR, 3KSV, 3VOV, 1LLU, 3FW3, 1SU0, 4Y4O-PDB-BUNDLE4, 4FAJ, 2EK0, 3OLB, 2YWW, 4A6D, 5K77, 1HFC, 3BJC, 4RDR, 1R6O, 1ILE, 2QHA, 1CDO, 6EQS, 2YHG, 2HXV, 4JBG, 5EVK, 4ROT, 3JYG, 5G0G, 3L22, 3DDT, 5OD6, 4ZTX, 5J7K, 3CSQ, 5K8C, 5IC5, 3EYX, 4X36, 5DK5, 3N9R, 3G8R, 3OL9, 3G99, 2Z8X, 2FTW, 1XV2, 3E2D, 5L44, 5KE9, 3EDI, 4YO8, 4HNO, 2I13, 5H2Q, 4X9J, 1AAY, 5CQI, 5CH8, 1EPW, 2OKL, 3NIS, 2RCN, 2E26, 1DQS, 5FD6, 2QGS, 3UW4, 2PEB, 3IBM, 4NWK, 3WRN, 3SFW, 1IQB, 1VE0, 1R55, 5T93, 1I76, 4G4Q, 1K0W, 1VQ0, 3FB4, 5XKX, 3PTM, 1Y2K, 3G9J, 1JKE, 3U6Q, 1D8Y, 1RK6, 3KEW, 1G9K, 1YW4, 5FBF, 2F3B, 3ZDR, 3ZEU, 3COS, 3PAO, 5UEJ, 1F1G, 4U5G, 4HP1, 5IJT, 1A74, 1XKF, 5G31, 4TWJ, 3WID, 3KYA, 2JLP, 5K81, 2WYH, 2HBL, 5L6Q, 2OX8, 5AH0, 4QF3, 3WE7, 5K4P, 3OJG, 5MFP, 4XWT, 1L1T, 3MP2, 5NGG, 3TA7, 1A1F, 3H0N, 4B6Z, 5TG0, 1NJF, 4N7S, 3N71, 2Q4H, 2CFU, 6CK7, 1VEC, 1OHL, 5SWC, 5VZV, 3GBO, 1SU1, 1T2T, 3U6D, 4W2F-PDB-BUNDLE4, 3JR5, 2HJH, 4H9K, 5CYZ, 2HCM, 3OCQ, 1U2W, 4RDA, 4BUU, 4C24, 3FDK, 4LY5, 2NLL, 3LWP, 2GWN, 1Z6U, 5FEY, 1GA5, 5H83, 4QEN, 5WPN, 2OU3, 3FP5, 5G2U, 4MTH, 1ZKP, 5I8G, 5OHJ, 3SXK, 1EKJ, 1KQ3, 4X5S, 4QP5, 4CCG, 3GC9, 1S5P, 2RI7, 3HPA, 4MAH, 3JZE, 3M1D, 3BPU, 3UCJ, 5G1A, 4LOC, 3K6I, 2XS2, 3OTM, 3V96, 5WLF, 4ZLA, 1PJI, 2RKN, 2WKX, 3KHI, 1N5N, 1MVH, 2CI7, 1ZZM, 3MND, 2J4X, 2ZKT, 5T00, 1Y0Y, 3POA, 2YAV, 4W6Z, 5GIQ, 2B0O, 3WUG, 1K7H, 1SRD, 3FC3, 1U19, 3WON, 1Q7L, 2APO, 4NIN, 4A0X, 1Y6H, 1TQX, 3AJ3, 3X17, 1KOP, 2P57, 4OIF, 5MXY, 4Y0X, 3G6U, 2AHJ, 4I6V, 4CJ0, 1K82, 3QE3, 4E2Z, 1HML, 3IGK, 6BNC, 3I9F, 5TZG, 2NYT, 3H1W, 1UD9, 4EGU, 1Q08, 5JU7, 2J1M, 5C22, 4G4R, 2F5S, 2BO0, 2VQE, 4CVB, 2B8T, 2HRV, 5SYB, 4WUY, 5G0H, 2G64, 4NNO, 3LRR, 3ECG, 5XEV, 4K60, 3LUM, 3RXZ, 3CT8, 2QYV, 3SNG, 5L7J, 2QSW, 2CJA, 5HLO, 2QN0, 3SZY, 5JGF, 1NPC, 5KO3, 3GIU, 4BMB, 2EH9, 2HVY, 1TAZ, 1ODZ, 4KNK, 5AMB, 4NTL, 2R2D, 5O6T, 3IEM, 2XYB, 5F6L, 5XIL, 3DOW, 1GUD, 5Z0C, 1Q14, 2VYO, 4OP4, 5NS5, 4NTK, 5GOF, 5TSH, 5XKR, 2YB5, 5A3D, 1FBL, 4JXE, 5W89, 5G34, 5AMH, 3E49, 1R4O, 1ZKL, 3V93, 1EVL, 4HTM, 1NN7, 1ONS, 2FS5, 3WKX, 4NIX, 1QWY, 3DRA, 5W0M, 3WGH, 4XCT, 5IYE, 3KV5, 1CFV, 2BIB, 1DSZ, 3B1B, 3F7L, 4Y4P-PDB-BUNDLE4, 1KFI, 1ENR, 2CEY, 1OCY, 2P46, 4QC1, 3K7N, 2ZTG, 1A1L, 4JE6, 4FC5, 4KP6, 2NLY, 2BF9, 2QYK, 1EKM, 3KVT, 3SPL, 4NRN, 4MI5,

3O14, 1A2P, 4Z3F, 5NCW, 4U4T, 3EXJ, 5K8P, 2NUT, 2QTV, 3O70, 3QMC, 5NA6, 1RJW, 4XFW, 5YXC, 2R2Z, 1YJO, 5KAR, 1VHE, 2IMS, 1AYM, 3P8B, 2WAD, 3UKO, 4R4X, 3AHN, 1A1I, 1G2D, 5JFY, 2G6Q, 1FIO, 3GYY, 5YEG, 5E2M, 2ASH, 5BR4, 2V0P, 3KWE, 2YV5, 3SIQ, 4R2S, 1BF6, 1LR5, 3I3W, 1FUK, 2ZWR, 4RKG, 3TAS, 4J4M, 5J46, 4CQN, 5XZB, 3G6R, 5EIU, 5U2J, 4YBB-PDB-BUNDLE3, 5A3A, 3M7N, 2AP1, 4OX5, 4HWT, 5HFS, 3T33, 3LUU, 4LE6, 4TOI, 1Z05, 1VCV, 5NL9, 2O6M, 2V8L, 5XP6, 1BQB, 2RGX, 3KBF, 3PPD, 4X29, 4IBU, 5K83, 4X3T, 3B6N, 3U79, 4XHV, 3SI2, 1NVT, 1EC5, 5GIV, 5EMP, 5KIA, 2PVX, 4WOK, 5UB9, 4W2F-PDB-BUNDLE2, 3COQ, 3HC4, 2EER, 2XC2, 4A1X, 1Z5H, 1RP0, 3L1O, 2NX8, 1ZAT, 3B90, 3G8X, 5UE3, 2VUN, 3BYR, 2W5W, 1Y9Q, 4QEO, 2G0D, 4NY2, 2A66, 2CLB, 2PW6, 2H6F, 2PDO, 1U10, 3SAU, 2G3F, 2ADW, 3OVL, 4LGJ, 5JVS, 3CGH, 3SAW, 2AGZ, 4QBS, 4Z04, 5JW0, 5VK3, 3B0Z, 1S9Z, 3FYM, 4DF9, 3G9I, 3R0D, 5VGM, 2H0D, 5A89, 4E4W, 2HF1, 5J72, 1RM8, 2IIM, 1EFQ, 4WPV, 5HKX, 4MZ7, 4TO8, 5E6C, 5UKI, 1TON, 3L7X, 4J1V, 5NDE, 5LD9, 1E3I, 2DPH, 5W9S, 3GPY, 1DCQ, 1BTG, 5DOC, 2OPF, 2GPY, 1M2A, 3TGN, 5TAB, 5CCL, 4NX4, 4NDI, 3GPX, 5G5Y, 4K8V, 4R2H, 2ZOG, 5GPY, 5F8N, 5AUM, 4IXN, 4WCJ, 4KKZ, 4FGM, 2XQ0, 5L0M, 1TOA, 4OWF, 5EMQ, 4Q8G, 4A69, 5TKQ, 4U10, 2YXO, 3HPS, 5E5C, 3KWO, 5XN8, 4YBB-PDB-BUNDLE2, 5F7Q, 5GNE, 1CZF, 1JQB, 3HTR, 1OI0, 3RPC, 2AKF, 4LY1, 5NJ9, 3DXS, 1JD5, 4P3X, 2D3K, 1CI3, 4C6E, 1QHW, 1XLL, 4IHM, 5ABX, 2XZF, 2CYE, 2I0M, 2NQJ, 2XZU, 3M9E, 5WLJ, 5BVR, 5U00, 4Y86, 4COQ, 3DCP, 5FUW, 5X4K, 1LAM, 3MHX, 3SUJ, 2H6E, 5HPJ, 1XSO, 5FYZ, 3MBG, 4XM2, 4UP0, 4HDH, 1FLJ, 4GXW, 2PJS, 4EVB, 3QU1, 1IUJ, 2E18, 5HWA, 4GC3, 1ET8, 3QNA, 4U9B, 2X3C, 3MIT, 3QC3, 1UE1, 4I28, 1R8Q, 3RQ4, 1R2Y, 1JOC, 2WKN, 1GUQ, 4IBV, 3GDF, 2GEQ, 4RVN, 2I7V, 4P10, 1Q9U, 4NE7, 3SAS, 4Z9K, 4ARF, 4A37, 5D7W, 1B0N, 4XLG, 5EI9, 2VQG, 2F5P, 5XFR, 3CEW, 2VH9, 1KEA, 1H1O, 3U6M, 5B1A, 4FTF, 4FKE, 2GU1, 4O6I, 1T0A, 5F9Z, 4E6R, 1L2B, 2RC5, 1ETE, 4LR2, 3LI2, 2OU2, 1ZDE, 5SVY, 4K89, 5B15, 2ZVL, 3ZD7, 3G0Z, 5EZB, 1OAH, 5Y2S, 4PXV, 3CJJ, 5XFQ, 4DZH, 2APS, 5B0H, 2Z3H, 4XXF, 3BLE, 3CWW, 1VFL, 5H82, 2HAP, 4AX1, 3AYV, 3PA0, 3T0C, 5JA5, 1PM5, 5B5L, 1ZAA, 2Q0Y, 2AZ4, 2H6L, 3F0D, 4L63, 3GPU, 1J6X, 1EAR, 3HDB, 3C6W, 4KAV, 3BI1, 3RZA, 1SLX, 4CSA, 1ML9, 1Z7H, 5V91, 2XY4, 3NG2, 1SNN, 1R85, 2WJY, 4GWM, 3QZ6, 1QU2, 5WO2, 2QYM, 3IEH, 5EW0, 5MSA, 5F6Q, 4WIG, 3LGB, 2E88, 5KL6, 5LCL, 2YX1, 1QX0, 5JLU, 3IFE, 3L11, 2KFN, 5VA7, 5IAA, 2NQ9, 5GZ5, 5TPR, 3E7M, 3IFJ, 5KEG, 4C1H, 4G6U, 5BUA, 4MYP, 1YXP, 3TC8, 5JJS, 3RF4, 4XRN, 3SFX, 5SZB, 2EO4, 3EA6, 4M9V, 1JCC, 1SLM, 4A7K, 3O8R, 5JR1, 1S0U, 1L6S, 2RCC, 1G43, 3KQI, 3CE9, 3GZK, 5PHJ, 5B2O, 3DCI, 1LMH, 3ST7, 5M3H, 4U09, 1C3R, 5PXL,

2EA0, 4M3P, 1QE3, 4Q3J, 2YR2, 4CN9, 1VHD, 3OL6, 1QHD, 3URZ, 5XGX, 5LCM, 4XBA, 2HJN, 3QMB, 5JIL, 2CKS, 3F6G, 5FKA, 2DW0, 1K9Z, 5HIF, 1YM3, 1I27, 2CB8, 4Q8R, 1KAP, 4UXH, 2PTZ, 5K7J, 1QWR, 1YN4, 2PUY, 1KOL, 5LX9, 1UVQ, 3F3Q, 5FGS, 3H7H, 5C5R, 5DKD, 1XRT, 4WVU, 4NG5, 3WAJ, 2XNJ, 3R3Q, 4I2F, 4LQ6, 4R97, 3DFF, 2AC0, 5KL3, 2A6H, 4K7T, 4RHZ, 1YT3, 1ZR6, 3R75, 4R2Q, 1DQ3, 2CCV, 5W3X, 4QBG, 4G4O, 5JF6, 4TN0, 3QBE, 4U06, 3LY0, 5VMD, 4MTD, 2O1Q, 5H4F, 5K9G, 1TBF, 1YC5, 3UW5, 1R5T, 3ZXO, 5TD3, 4RV9, 1E3J, 2F5N, 5GRQ, 1V15, 1F35, 4E45, 3U6S, 5G35, 5KL4, 4K1G, 5LAZ, 1AK0, 3L8Y, 3IEK, 1EU4, 2XBL, 5WZQ, 7MDH, 2YJP, 1SE0, 4K2H, 4B29, 4Y4O-PDB-BUNDLE1, 5D88, 1M65, 4ZYA, 3BQ5, 5MTM, 3AY2, 3CHV, 3AYF, 3D2S, 2NSF, 5GJ9, 5DZT, 4QCL, 1QX2, 2XF4, 2R8Q, 5GY6, 3ISO, 5UIW, 1L2C, 5E6D, 5JI5, 2IMR, 1EKB, 1PMI, 3KZ8, 3CSK, 5K1A, 3HT2, 5XFO, 4MKO, 4WNC, 3W95, 2ATA, 4LAN, 2EIH, 1EQW, 4WCO, 5X6J, 1XX6, 3CJP, 3EZ5, 1DVP, 1ZME, 4BR5, 2CIH, 5Y20, 4IVV, 4EBB, 3PB6, 3FAV, 5IH4, 1ZNS, 3HFT, 4XI7, 4YS4, 4YHB, 2HCV, 3ZBV, 5CZW, 4ARE, 3N3U, 3WT4, 1T3C, 4H2K, 3ZTV, 2HAN, 1PSZ, 5VWM, 5GK9, 1RO2, 1Y75, 2BNM, 5E33, 5A0L, 2FBJ, 1HP7, 3LNL, 2VO9, 2PLM, 4HVC, 3VTH, 3P24, 4X2Z, 3SAV, 4NU7, 5OPJ, 5E6T, 3KFL, 1EWC, 3IM4, 1ITU, 3A52, 1P9E, 1A0Q, 3LS6, 4AC1, 3LQH, 2FPQ, 4BT7, 5TT5, 5CC0, 4NJ5, 3PUR, 4FKD, 2XVA, 2I0O, 3R6F, 3M97, 3IO3, 5H6B, 5MFR, 3C9F, 3QU6, 3G9Y, 1R4V, 2OZU, 5GT1, 3S2E, 2OSV, 1Z9N, 3FSR, 2C2U, 3O8B, 4CNX, 3OHR, 1PP7, 1KFS, 1WY2, 4DB3, 4N0L, 1CYY, 5JVW, 2W15, 3DGV, 5V5H, 3IOF, 5CBX, 4ZN3, 1NNJ, 5U7G, 2C1D, 3WF7, 3OME, 5D9Y, 5KGN, 1HC7, 3FPC, 3QMH, 4V1T, 2WFQ, 3PNU, 4LFY, 2AC2, 1WPP, 5MR8, 3ELF, 6EKB, 5JWP, 4TWE, 3Q3Q, 3ZDU, 3P2N, 1A1K, 2JKS, 2F94, 5IB9, 4LIM, 2OH3, 5O9F, 2E4T, 4QFI, 1Y23, 4KMN, 3EB5, 2O3E, 2DJW, 4KF9, 1VS0, 5KRB, 3BO5, 2I1O, 5ACS, 3N1F, 4V2W, 4LEZ, 2Q02, 3T8W, 1EU3, 4ICS, 5T5I, 1YP1, 2FQP, 5CIO, 3U6E, 3VD6, 1SED, 2AHI, 3TG0, 1XOV, 3DZA, 5UDZ, 4GV6, 2QA1, 3SJP, 3E2I, 4O1K, 1PJJ, 3KMD, 5H4G, 3FNS, 2HQK, 4C81, 1G12, 5UQP, 4PCZ, 1U3W, 2VUT, 1HY7, 4O3M, 5EH2, 1DOS, 5F8H, 4HCG, 5E6B, 3ICJ, 4W4O, 4COG, 4CIS, 4R2A, 3P3G, 4KUJ, 3IIB, 4XGL, 5JVI, 2X98, 4LXL, 3LWU, 5MQ1, 3HJW, 3E7L, 5MTE, 2ZO4, 1F5F, 4PS2, 3E1Z, 1CPR, 2RGT, 1TKJ, 4EOG, 3GL6, 4K90, 5DKC, 1TJL, 3ADR, 4NAZ, 3HTK, 3S3N, 5H9M, 1GYT, 3F2B, 3LKV, 6BMS, 1HCQ, 2B5W, 5FGW, 1Z2L, 2IQJ, 1KHO, 4OH1, 5OJ7, 2P18, 4R2P, 4IF2, 3ECM, 4M6R, 2A03, 3LDY, 5IHE, 3MA2, 2BO9, 5IKK, 5WAM, 5KB2, 3CZT, 5NM9, 5I2B, 4PE3, 3FPL, 3QDF, 3IMI, 2RKU, 2V25, 3PVK, 5BRK, 3SP7, 2BQ8, 3EXL, 3OQ3, 4RKH, 1M2X, 4ZYB, 3R68, 2P6Y, 5NQZ, 4YIW,

4OMF, 5M2K, 4DIH, 4AQ7, 4AY7, 4ETS, 3K7I, 5DLT, 1CY5, 2P53, 5KE6, 4R34, 2XZQ, 1EH6, 2HJG, 1HW7, 1D8W, 2GC0, 4N4F, 5LB3, 3RZU, 3MCX, 2G7Z, 3EEF, 3Q02, 4UPI, 5AEB, 5VN5,