Form and Function:

Two computational protein design studies

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Kaspar Konrad Feldmeier

aus Tübingen

Tübingen

2016

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation: 11.05.2016

Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Birte Höcker
2. Berichterstatter:	Prof. Dr. Thilo Stehle

Contents

Deutsche Zusammenfassung

Proteine sind nicht nur grundlegende Bausteine unserer Zellen, sondern auch verantwortlich für die Katalyse der meisten Reaktionen die unser Leben ausmachen. Aufgrund ihres einfachen Aufbaus aus nur zwanzig verschiedenen Aminosäuren und der trotzdem hohen Variabilität in Größe, Form und Funktion sowie der gut skalierbaren Produktion in Bakterien haben Proteine aber auch ideale Voraussetzungen um gerichtet Katalysatoren für chemische Prozesse zu konstruieren. Hierbei sind besonders zwei Gebiete von Interesse: Das Design von Proteinfaltungen sowie das Design neuer Aktivitäten.

Ein Teil der vorgestellten Arbeit beschäftigt sich mit dem Design eines *de novo* TIM-barrels, der wohl wichtigsten Faltung für Enzyme. Dieses Projekt entstand aus einer Zusammenarbeit mit Possu Huang und der Arbeitsgruppe von David Baker an der Universität Washington, Seattle. Verschiedene Designs, welche nach gemeinsam festgelegten Prinzipien von Possu Huang am Computer generiert worden waren, wurden dabei von mir charakterisiert. Eine Variante kristallisierte und lieferte eine Kristallstruktur die den Erfolg des Designs bestätigte. Diese Ergebnisse wurden Anfang des Jahres in *Nature Chemical Biology* publiziert [1].

Der zweite Teil beschäftigt sich mit dem rationalen Design einer neuen enzymatischen Aktivität basierend auf der klassischen katalytischen Serintriade. Mehrere Proteinstrukturen wurden mittels Computer-gestützter Suche als mögliche Träger der Triade identifiziert und basierend darauf verschiedene Varianten mit optimierten

Bindungstaschen *in silico* konstruiert und evaluiert. Die Konstrukte wurden im Labor hergestellt und analysiert. Eine Variante zeigte eine schwache katalytische Aktivität. Diese Variante wurde hinsichtlich Struktur und Funktion im Detail charakterisiert.

Diese beiden komplementären Studien, die beide versuchen die Grenzen des Computer-gestützten Design der Form und Funktion von Proteinen zu erweitern, tragen bei zu einem besseren Verständnis von Proteinsequenz-, Struktur- und Funktionsbeziehungen.

Chapter 1

Introduction

1.1 Understanding life

Modern multicellular organisms like ourself have always been in the focus of human research. However our emphasis over the history of natural sciences shifted with the progress of methods and technical possibilities. Early approaches tried to explain the machine of life on a macroscopic level, discovering the major building parts, the organs, and their function. The development of the microscope led to discoveries that left the boundaries imposed by our naked eye. Suddenly structures present at a microscopic level were visible and could be characterized (e.g. Antonie van Leeuwenhoek, 17th century). It took another two hundred years until the rise of organic chemistry made it possible to address the question of what makes life work at an even smaller level, leading to discoveries that extended our understanding of the mechanisms involved to a molecular and atomic level.

In the last century this molecular understanding was led to new heights with the description of structures of two of the most important and complex classes of macromolecules in our cells: DNA, which stores the cells information, and proteins, the molecules responsible for most of the work necessary to keep us alive.

1.2 Proteins

What is fascinating about proteins is that in contrast to their wide variety in form, function and size they all derive from a few small building blocks, the amino acids. In most organisms there are only twenty different amino acids, and during the expression of a protein these small building blocks are connected through peptide bonds, forming a long linear string. It is the sequence of the string that is determined by the gene responsible for the expression of a particular protein, and it is this mechanism that is responsible for the usage of the vast majority of information stored in the DNA.

What happens next upon expression is coarsely described as protein folding: According to the physical and chemical properties of the different amino acids at different positions, the protein folds into a stable, three dimensional structure that will be the same for every newly expressed protein of this type. It is the combination of amino acid sequence and structure that determines what this proteins function will be, whether it will be a part of the cells skeleton, an enzyme degrading food or a receptor for insulin. Therefore, if we want to understand proteins, we have to understand protein folds and protein folding.

1.3 Protein structure

There are different levels of protein structure. The most basic one is the sequence of amino acids forming a specific protein, which has accordingly been named the primary structure of proteins.

Depending on the sequence, even small fragments of proteins can fold into what is called secondary structure. There are two major types of secondary structure, the α-helix and the β-strands (see figure 1.1). In α-helices the amino acids form a helix with roughly 3.6 amino acids per turn, stabilized by hydrogen bonds within

the backbone atoms. Within β-strands the amino acids lie flat to each other, with a characteristic zigzag pattern where every amino acid points in the opposite direction as its two neighbors. Several β-strands placed in a plane can form what is known as a β-sheet by interactions between the backbone atoms of neighboring strands. Different secondary structure elements are connected by so called loops, often quite flexible stretches of amino acids.

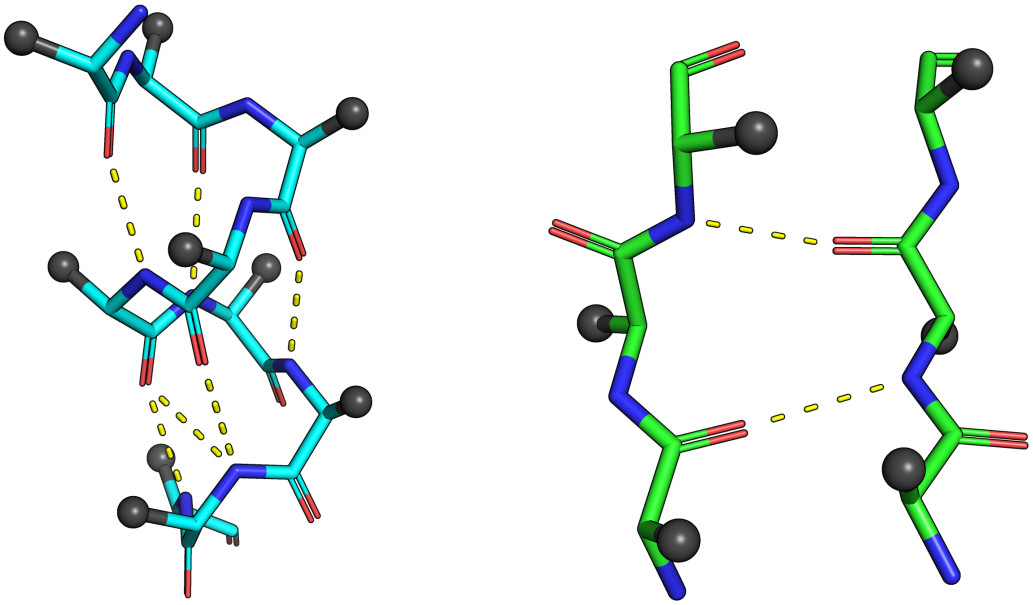


Figure 1.1: The two basic secondary structure elements of proteins: The α-helix on the left where amino acids form a helix stabilized by hydrogen bonds between backbone atoms. The other element is the β-strand on the right, where amino acids form a plane. Several (in this case parallel) β-strands form a β-sheet stabilized by hydrogen bonds between backbone atoms of neighboring strands. In both cases the Cβ atoms carrying the side chains are colored Grey.

Build from these secondary structural elements, proteins fold into a stable three-dimensional structure and for some proteins like the ones responsible for a cell's cytoskeleton, this structure is already their purpose, using their shape to give structure and stability to the cell. However, in many cases proteins have other functionalities such as the very important function of a chemical catalyst, namely enzymes.

1.4 Enzymes

If life is seen as the complex system of reactions that makes cells thrive, grow and multiply, then enzymes are its basis. Although there are indications that in early lifeforms catalysts were based on RNA [2], by now most reactions in cells are catalyzed by proteins.

In the case of a catalyst a certain structure alone is of course not sufficient for function, it has to be combined with a binding pocket where the substrates interact and a specific chemical environment depending on the reaction that is catalyzed.

A scaffold where these two different necessities, a stable structure and an active center, can be observed and even nicely structurally separated is the TIM-barrel scaffold (see figure 1.2). This scaffold is named after the triose phosphate isomerase where it was first discovered, however by now a wide variety of TIM-barrels has been described. It is not only the most variable enzyme scaffold (of the six enzyme classes TIM barrels are found in five, only lacking a member with ligase activity) but also a very old one. Members catalyze reactions that are central to modern living organisms, the already mentioned triose phosphate isomerase serves as a good example. What is quite remarkable in TIM-barrels is the separation of structure and activity. The core of the protein, build by a central barrel formed by eight β-strands and surrounded by eight α-helices is quite rigid and responsible for the overall stability of the enzyme. The functional groups are in all members, as different as their reactions are, localized on top of the barrel, either at the end of the β-strands, or in the loops connecting them.

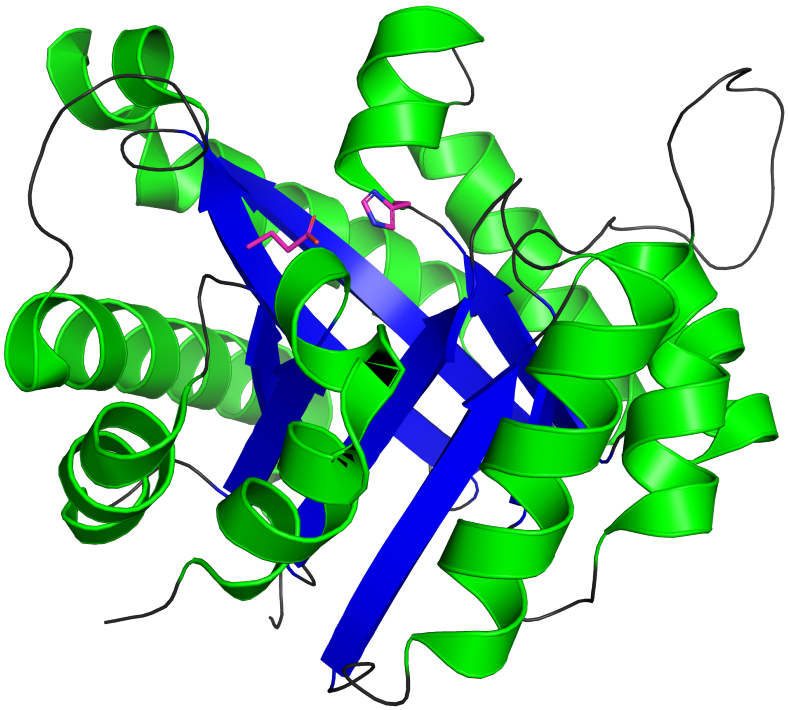


Figure 1.2: The TIM-barrel structure, arguably the most important scaffold for enzymes, is build up by elternating β-strands and α-helices. It shows a clear distinction between the main barrel which contributes stability to the scaffold and the top of the barrel which caries the different catalytic activities (catalytic residues are shown in pink). Shown here is the structure from the name-giving triose phosphate isomerase (PDB ID: 1TIM [3]).

The active centers themselves, however, are as diverse as the reactions they catalyze. As with all enzymes, they have to provide two main functions: A pocket that is likely to specifically bind the substrate and keep it in place during the reaction, and the catalytic center composed of amino acids or cofactors creating a chemical environment that leads to catalysis. While the first part strongly depends on the substrate and can vary greatly within very similar reactions, the catalytic part is often very similar within analog catalysts even if they do not share a common origin. Many of these widespread mechanisms have been analyzed quite thoroughly, and among the best understood and most famous catalytic motifs is that of the catalytic serine triad, one of the textbook examples of enzyme mechanisms.

1.5 The catalytic triad

Considering that the catalytic triad is a surprisingly small motif, it is capable of performing quite difficult reactions such as the hydrolysis of peptide bonds. It consists of three amino acids forming the name-giving acid-base-nucleophile triad and sometimes in addition one or more nitrogen atoms forming what is called the oxyanion hole. Probably due to its simplicity it evolved several times independently [4], usually within hydrolases or transferases. A variety of different triads are known today [5] with the classic one consisting of serine as the nucleophile, histidine as the base and aspartate as the acid.

These three amino acids have to be present in a very specific geometry (see figure 1.3), allowing them to form a charge-relay system that allows the nucleophile (usually serine or cysteine) to be activated and attack covalent bonds. The mechanism that leads first to the cleavage of the substrate and later to the restoration of the catalytic triad is complex and consists of several steps. After said nucleophilic attack in which a charged intermediate is stabilized by the oxyanion hole, a first

tetrahedral intermediate is formed by formation of a covalent bond between parts of the substrate and the nucleophile. The first product leaves after addition of a hydrogen from the base (histidine in the classic example). Said nucleophile is later restored, for example by hydrolysis of the second product.

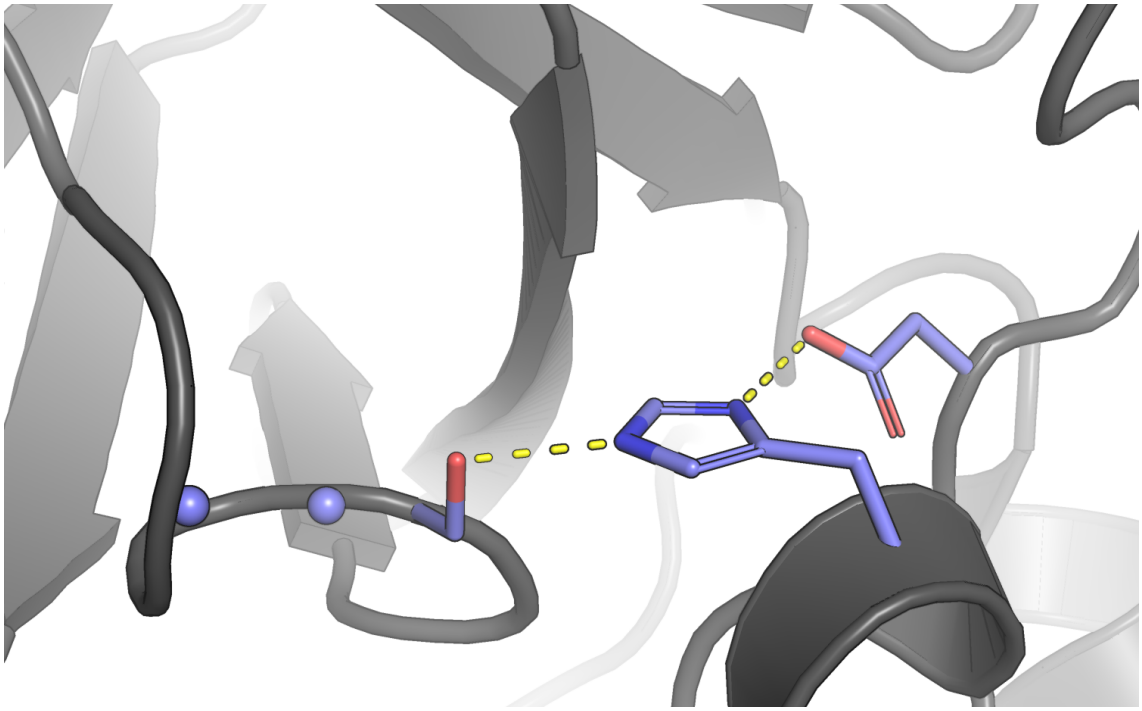


Figure 1.3: The catalytic triad as found in trypsin is comprised of a serine, a histidine and an aspartate. While the serine and histidine usually point to the substrate binding pocket, the acid can be buried within the enzyme. The blue balls represent the two backbone nitrogen atoms forming the oxyanion hole. PDB ID: 4MTB.

The number of different substrates that are processed by catalytic triads is huge. Enzymes utilize this same motif for different reactions by combining it with different scaffolds and binding pockets, leading to a variety of highly efficient and often very specific catalysts. And while enzymes catalyze a large variety of reactions for natural substrates, this mechanism should also be able to catalyze hydrolysis of substrates that might not be present in Nature yet still be beneficial for us. So the question is, with the amount of knowledge available, are we able to design whatever enzymes we might need?

1.6 Enzyme design

Just like natural enzymes need both a stable fold and a specific catalytic center, we need the ability to predict and design both if we want to create completely new enzymes with new activities. Both sides of enzyme design were in the focus of research for decades.

De novo design of a TIM-barrel

Designing a certain structure is not only important for the design of enzymes, but for the design of other protein-based tools like biosensors as well. In the past years several attempts were made to design folded proteins *de novo* [6] [7] [8]. Among the approaches more relevant to enzyme design, the *de-novo* design of a TIM-barrel has been in the focus of scientists for at least 25 years [9] and has been ongoing for a while [10] [11] [12] [13] [14]. The structure of TIM-barrels seems to be simple and natural variants often show a high stability. Nevertheless, up to now there was no confirmed *de novo* designed variant where the design could be confirmed with a structure. While early designs even seemed to lack secondary structure, the latest variants seem to be folded at least on the level of secondary structure. However, attempts to derive a structure, the crucial proof that the design goal was met, remained futile.

In collaboration with Possu Huang and David Bakers group we decided to attempt yet another trial in the race for the first successful TIM-barrel design, trying to deliver the proof that by now we are able to engineer the most widespread and important enzyme scaffold. The results are described in chapter 4.

Rational design of a new hydrolase

The design of enzymatic function, however, had quite some successful examples in the past years. Thus far there has been a variety of different reactions successfully engineered into proteins, such as Retro-Aldol reactions [15] [16], Kemp-elimination [17] [18], Morita-Baylis-Hillman [19] and Diels-Alder reactions [20] (see [21] for an overview). There were also of course attempts to incorporate the classic catalytic triad into a new scaffold [22] [23], however a complete and enzymatically active catalytic triad has not yet been engineered. Doing so would actually serve a dual purpose: On the one hand it would proof that our understanding of one of the best known catalytic motifs is thorough enough to be used as a template for design. On the other hand a successful design should be able to catalyze reactions such as the breakage of peptide bonds. Compared to prior designed enzymes, this is a quite difficult multistep reaction which requires a high activation energy to be overcome. The required geometry necessary for this reaction shows - compared with most geometries used for earlier designs - little margin for error, putting a lot of emphasis on the selection and design of protein scaffolds chosen to become enzymes.

Due to both its importance as a showcase example for enzymes and the fact that a successful design would proof that also the engineering of energetically more difficult reactions in enzymes is within our reach, I attempted to computationally design a catalytic triad based catalyst. The approach and its results are described in chapter 5.

Chapter 2

Material

2.1 Equipment

Shaker

- **Innova 411** (Eppendorf)
- **Multitron II** (Infors)

Centrifuges

- **Centrifuge 5810R** (Eppendorf)
- **Centrifuge 5424** (Eppendorf)
- **Heraeus Fresco 17** (Thermo Scientific)
- **Avanti J-26xPI** (Beckmann Coultier)

Thermocycler

- **MyCycler Thermocycler** (BioRad)
- **T3000** (Biometra)

Scales

- **572** (Kern)
- **Fine scale ALS120-4** (Kern)

Chromatography

FPLCs

- **AEkta Pure 25M** (GE Healthcare)
- **AEkta P90**0 (GE Healthcare)
- **AEkta Prime** (GE Healthcare)

Columns

- **HisTrap HP, 5ml** (GE Healthcare)
- **HisTrap HP, 1ml** (GE Healthcare)
- **Superdex 75 10/300 GL** (GE Healthcare)
- **HiLoad Superdex 75 PG** (GE Healthcare)

Spectrometer

- **Avance III NMR spectrometer 800MHz** (Bruker)
- **FP-6500 fluorescence-spectrometer** (Jasco)
- **J-810 CD-spectrometer** (Jasco)
- **ND-1000 Nanodrop** (PeqLab)
- **Cary 50 Scan UV-VIS spectrometer** (Varian)

Crystallization

- **Honeybee 963** crystallization robot (Genomic Solutions)
- **Mosquito** crystallization robot (TTP)
- **Rock Imager 182 Imaging System with UV optics** (Formulatrix)

Other Equipment

- **digital sonifier W-250** (Branson)
- **pH211 microprocessor pH meter** (Hanna)
- **Thermoblock** (Olaf Waase)
- **miniDAWN TREOS** Multi-Angle Light Scattering Detector (Wyatt technology)
- **Micropulser** electroporator (Biorad)

2.2 Enzymes

- **FD NdeI** (Fermentas/Thermo Scientific)
- **FD XhoI** (Fermentas/Thermo Scientific)
- **FastAP** (Fermentas/Thermo Scientific)
- **Taq polymerase** (Fermentas/Thermo Scientific)
- **Q5 High fidelity DNA Polymerase** (New England Biolabs)
- **T4 ligase** (New England Biolabs)

2.3 Chemicals

Loading dyes and size standards

- **Gene ruler 50bp** (Fermentas/Thermo Scientific)
- **Gene ruler 100bp** (Fermentas/Thermo Scientific)
- **Gene ruler 1kbp** (Fermentas/Thermo Scientific)
- **loading dye (6) for DNA electrophoresis** (Fermentas/Thermo Scientific)
- **2x SDS samplebuffer** 100mM TRIS pH 6.8, 20 % Glycerol, 4 % SDS, 0.2M DTT and a spatula tip of Bromphenol blue
- **Unstained Protein Molecular Weight Marker** (Thermo Scientific)

Various compounds

- **Phenylmethylsulfonyl fluoride (PMSF)** CAS: 329-98-6 (Sigma-Aldrich)
- **4-(2-Aminoethyl)benzensulfonylfluorid (AEBSF)** CAS: 30827-99-7 (Sigma-Aldrich)
- **5,5-Dithiobis-2-nitrobenzoic acid (DTNB)** CAS: 69-78-3 (Sigma-Aldrich)
- **L-Alanine 4-nitroanilide hydrochloride (ApNA)** CAS: 31796-55-1 (Sigma-Aldrich)
- **N-Benzoyl-L-tyrosine p-nitroanilide (YpNA)** CAS: 6154-45-6 (Sigma-Aldrich)
- **Val-Ala p-Nitroanilide acetate salt (VApNA)** CAS Number 108321-94-4 (Sigma-Aldrich)
- **N-Succinyl-Ala-Ala-Pro-Phe p-nitroanilide (AAPF-pNA)** CAS: 70967-97-4 (Sigma-Aldrich)
- **4-Nitrophenyl acetate (pNAc)** CAS: 830-03-5 (Sigma-Aldrich)
- **Dithiothreitol (DTT)** (Roth, No. 6908.2)
- **Coomassie Staining Solution** 400ml water, 400ml ethanol, 1g Coomassie Blue G-250

2.4 Purification kits

- **NucleoSpin Gel and PCR Clean-up** (Macherey-Nagel)
- **NucleoSpin Plasmid EasyPure** (Macherey-Nagel)

2.5 Bacterial growth media

- **LB Medium** For 1l: 10g pepton from bacteria, 5g NaCl, 5g yeast extract
- **TB medium** For 1l: 24g yeast extract, 12g tryptone, 4ml glycerol, 2.31g KH_2PO_4, 12.54g K_2HPO_4
- **Autoinduction medium** 50ml 20x NPS, 20ml 50x5052, 1ml $MgSO_4$, ZY medium ad 1l
  - **20x NPS** 6.6g $(NH_4)_2SO_4$, 13.6g KH_2PO_4, 14.2g Na_2HPO_4, water ad 100ml
  - **50x 5052** 25g glycerol, 10g α-lactose, 2.5g glucose, water ad 100ml

2.6 Vectors

- **pet21a(+)** (Novagene), Ampicillin resistance
- **pet29b(+)** (Novagene), Kanamycin resistance

2.7 Software

- **ScaffoldSelection** [24]
- **Rosetta 3.3-3.5** [25]
- **Pymol** The PyMOL Molecular Graphics System, Version 1.7 Schrdinger, LLC.
- **XDS** [26]
- **Phenix suite** [27]

- **Ape-A Plasmid Editor** (M. Wayne Davis)
- **Ugene** [28]
- **Spectrum Manager 5** (Jasco)
- **Gimp 2.8**
- **Mendeley**
- **TeXstudio**
- **Bioinformatics toolkit** [29]
- **Blender** Blender - a 3D modeling and rendering package
- **FROG2** [30]
- **CLANS** [31]

2.8 Sequences

3N8M

```
                    10        20        30        40        50
3N8M            MIEMKPHPWFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKF    50
3N8M3x          ..............h...d.............................    50
3N8M_R16H       ..............h.................................    50
3N8M_E20D       ..................d.............................    50
3N8M_H55S       ..................................................    50
3N8M_R16H_H55S  ..............h.................................    50
3N8M_E20D_H55S  ..................d.............................    50
3N8M_varA       ..............h...d...l.........f...ry..y.i.a...    50
3N8M_varB       ..............h...d...l.........f...wy..y.i.a...    50
3N8M_varC       ..............h...d.............f...ry..y.i.a...    50
3N8M_varD       ..............h...d.............f...wy..y.i.a...    50

                    60        70        80        90       100
3N8M            GNDVQHFKVLRDGAGKYFLWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI   100
3N8M3x          .....s............................................   100
3N8M_R16H       ..................................................   100
3N8M_E20D       ..................................................   100
3N8M_H55S       .....s............................................   100
3N8M_R16H_H55S  .....s............................................   100
3N8M_E20D_H55S  .....s............................................   100
3N8M_varA       .....s.l..........................................   100
3N8M_varB       .....s.l..........................................   100
3N8M_varC       .....s.l..........................................   100
3N8M_varD       .....s.l..........................................   100

                   110
3N8M            EQVPQQPTYVQAHHHHHH   118
3N8M3x          ..................   118
3N8M_R16H       ..................   118
3N8M_E20D       ..................   118
3N8M_H55S       ..................   118
3N8M_R16H_H55S  ..................   118
3N8M_E20D_H55S  ..................   118
3N8M_varA       ..................   118
3N8M_varB       ..................   118
3N8M_varC       ..................   118
3N8M_varD       ..................   118
```

Figure 2.1: Sequences of the constructs utilizing the scaffold of 3N8M, a human Grb2 SH2 Domain.

3N8M alternative insertion site

```
                    10        20        30        40        50
3N8M_wt         MIEMKPHPWFFGKIPRAKAEEMLSKQRHDGAFLIRESESAPGDFSLSVKF     50
3N8M_alt_varA   ........q......d...h..........................h.f.     50
3N8M_alt_varB   ........n......d...h...........h..a...........y.f.     50
3N8M_alt_varC   ........n......d...h...........a..s...........y.w.     50

                    60        70        80        90        100
3N8M_wt         GNDSQHFKVLRDGAGKYFLWVVKFNSLNELVDYHRSTSVSRNQQIFLRDI    100
3N8M_alt_varA   .yr.yq.....................................          100
3N8M_alt_varB   ..n.kr.r...................................          100
3N8M_alt_varC   ..r.yr.....................................          100

                    110
3N8M_wt         EQVPQQPTYVQAHHHHHH   118
3N8M_alt_varA   ..................   118
3N8M_alt_varB   ..................   118
3N8M_alt_varC   ..................   118
```

Figure 2.2: Sequences of the constructs utilizing the scaffold of 3N8M, a human Grb2 SH2 Domain, with an alternative insertion site. Selection of these variants was done together with Marcel Conrady, who also cloned the constructs.

1W54

```
                10        20        30        40        50
1W54      MSFTPANRAYPYTRLRRNRRDDFSRRLVRENVLTVDDLILPVFVLDGVNQ    50
1W54_3x   ..................................................    50

                60        70        80        90        100
1W54      RESIPSMPGVERLSIDQLLIEAEEWVALGIPALALFPVTPVEKKSLDAAE   100
1W54_3x   ................................h.................   100

                110       120       130       140       150
1W54      AYNPEGIAQRATRALRERFPELGIITDVALDPFTTHGQCGILDDDGYVLN   150
1W54_3x   ......................s...........................   150

                160       170       180       190       200
1W54      DVSIDVLVRQALSHAEAGAQVVAPSDMMDGRIGAIREALESAGHTNVRIM   200
1W54_3x   ..................................................   200

                210       220       230       240       250
1W54      AYSAKYASAYYGPFRDAVGSASNLGKGNKATYQMDPANSDEALHEVAADL   250
1W54_3x   ..................................................   250

                260       270       280       290       300
1W54      AEGADMVMVKPGMPYLDIVRRVKDEFRAPTFVYQVSGEYAMHMGAIQNGW   300
1W54_3x   ..................................................   300

                310       320       330
1W54      LAESVILESLTAFKRAGADGILTYFAKQAAEQLRRGR    337
1W54_3x   ....................d..............    337
```

Figure 2.3: Sequences of the constructs utilizing the scaffold of 1W54, a Porphobilinogen Synthase from *Pseudomonas aeruginosa*.

2D52

```
                    10        20        30        40        50
2D52_wt     MSSLSNSLPLMEDVQGIRKAQKADGTATVMAIGTAHPPHIFPQDTYADVY    50
2D52_3x     ..................................................    50
2D52v2      ..............................t...................    50
2D52_v3     ..............................t...................    50
2D52_v4     ..............................t...................    50
2D52_v5     ..............................t...................    50
2D52_v6     ..................................................    50
2D52_v7     ..................................................    50
2D52_v8     ..................................................    50
2D52_alt3x  ..................................................    50

                    60        70        80        90       100
2D52_wt     FRATNSEHKVELKKKFDHICKKTMIGKRYFNYDEEFLKKYPNITSYDEPS   100
2D52_3x     ..................................................   100
2D52v2      ..................................................   100
2D52_v3     ..................................................   100
2D52_v4     ..................................................   100
2D52_v5     ..................................................   100
2D52_v6     ..................................................   100
2D52_v7     ..................................................   100
2D52_v8     ..................................................   100
2D52_alt3x  ............................s..h...d.............   100

                   110       120       130       140       150
2D52_wt     LNDRQDICVPGVPALGTEAAVKAIEEWGRPKSEITHLVFCTSCGVDMPSA   150
2D52_3x     ............................s.h...d.................   150
2D52v2      ............................s.h...d.................   150
2D52_v3     ............................s.h...d.................   150
2D52_v4     ............................s.h...d.................   150
2D52_v5     ............................s.h...d.................   150
2D52_v6     ............................s.h...d.................   150
2D52_v7     ............................s.h...d.................   150
2D52_v8     ............................s.h...d.................   150
2D52_alt3x  ..................................................   150

                   160       170       180       190       200
2D52_wt     DFQCAKLLGLHANVNKYCIYMQGCYAGGTVMRYAKDLAENNRGARVLVVC   200
2D52_3x     ..................................................   200
2D52v2      ..........................................v.......   200
2D52_v3     ..........................................a.......   200
2D52_v4     ..........................................v.......   200
2D52_v5     ..........................................a.......   200
2D52_v6     ..........................................v.......   200
2D52_v7     ..........................................a.......   200
2D52_v8     ..........................................v.......   200
2D52_alt3x  ..................................................   200
```

```
                    210        220        230        240        250
2D52_wt      AELTIMGLRAPNETHLDNAIGISLFGDGAAALIIGSDPIIGVEKPMFEIV  250
2D52_3x      ..................................................  250
2D52v2       .........................................a........  250
2D52_v3      .........................................g........  250
2D52_v4      .........................................g........  250
2D52_v5      .........................................a........  250
2D52_v6      .........................................a........  250
2D52_v7      .........................................g........  250
2D52_v8      .........................................g........  250
2D52_alt3x   ..................................................  250

                    260        270        280        290        300
2D52_wt      CTKQTVIPNTEDVIHLHLRETGMMFYLSKGSPMTISNNVEACLIDVFKSV  300
2D52_3x      ..................................................  300
2D52v2       ..................................................  300
2D52_v3      ..................................................  300
2D52_v4      ..................................................  300
2D52_v5      ..................................................  300
2D52_v6      ..................................................  300
2D52_v7      ..................................................  300
2D52_v8      ..................................................  300
2D52_alt3x   ..................................................  300

                    310        320        330        340        350
2D52_wt      GITPPEDWNSLFWIPHPGGRAILDQVEAKLKLRPEKFRAARTVLWDYGNM  350
2D52_3x      ..................................................  350
2D52v2       ..................................................  350
2D52_v3      ..................................................  350
2D52_v4      ..................................................  350
2D52_v5      ..................................................  350
2D52_v6      ..................................................  350
2D52_v7      ..................................................  350
2D52_v8      ..................................................  350
2D52_alt3x   ..................................................  350

                    360        370        380        390        400
2D52_wt      VSASVGYILDEMRRKSAAKGLETYGEGLEWGVLLGFGPGITVETILLHSL  400
2D52_3x      ..................................................  400
2D52v2       ..................................................  400
2D52_v3      ..................................................  400
2D52_v4      ..................................................  400
2D52_v5      ..................................................  400
2D52_v6      ..................................................  400
2D52_v7      ..................................................  400
2D52_v8      ..................................................  400
2D52_alt3x   ..................................................  400
```

```
                          410
2D52_wt       PLMLEHHHHHH   411
2D52_3x       ..........    411
2D52v2        ..........    411
2D52_v3       ..........    411
2D52_v4       ..........    411
2D52_v5       ..........    411
2D52_v6       ..........    411
2D52_v7       ..........    411
2D52_v8       ..........    411
2D52_alt3x    ..........    411
```

Figure 2.4: Sequences of the constructs utilizing the scaffold of 2D52, a Pentaketide chromone synthase from *Aloe arborescens*.

2cfm

```
                    10        20        30        40        50
2cfmwt         MRYLELAQLYQKLEKTTMKLIKTRLVADFLKKVPDDHLEFIPYLILGEVF    50
2cfm3x         ..................................................    50
2cfm3x+F318I   ..................................................    50
2cfmv12        ..................................................    50
2cfmv34        ..................................................    50
2cfmv1314      ..................................................    50
2cfmv12_268R   ..................................................    50
2cfmv12_268C   ..................................................    50
2cfmv12_530R   ..................................................    50
2cfmv12_533K   ..................................................    50

                    60        70        80        90       100
2cfmwt         PEWDERELGVGEKLLIKAVAMATGIDAKEIEESVKDTGDLGESIALAVKK   100
2cfm3x         ..................................................   100
2cfm3x+F318I   ..................................................   100
2cfmv12        ..................................................   100
2cfmv34        ..................................................   100
2cfmv1314      ..................................................   100
2cfmv12_268R   ..................................................   100
2cfmv12_268C   ..................................................   100
2cfmv12_530R   ..................................................   100
2cfmv12_533K   ..................................................   100

                   110       120       130       140       150
2cfmwt         KKQKSFFSQPLTIKRVYQTLVKVAETTGEGSQDKKVKYLADLFMDAEPLE   150
2cfm3x         ..................................................   150
2cfm3x+F318I   ..................................................   150
2cfmv12        ..................................................   150
2cfmv34        ..................................................   150
2cfmv1314      ..................................................   150
2cfmv12_268R   ..................................................   150
2cfmv12_268C   ..................................................   150
2cfmv12_530R   ..................................................   150
2cfmv12_533K   ..................................................   150

                   160       170       180       190       200
2cfmwt         AKYLARTILGTMRTGVAEGLLRDAIAMAFHVKVELVERAYMLTSDFGYVA   200
2cfm3x         ..................................................   200
2cfm3x+F318I   ..................................................   200
2cfmv12        ..................................................   200
2cfmv34        ..................................................   200
2cfmv1314      ..................................................   200
2cfmv12_268R   ..................................................   200
2cfmv12_268C   ..................................................   200
2cfmv12_530R   ..................................................   200
2cfmv12_533K   ..................................................   200
```

```
                        210       220       230       240       250
2cfmwt          KIAKLEGNEGLAKVQVQLGKPIKPMLAQQAASIRDALLEMGGEAEFEIKY  250
2cfm3x          ..................................................  250
2cfm3x+F318I    ..................................................  250
2cfmv12         ..................................................  250
2cfmv34         ..................................................  250
2cfmv1314       ..................................................  250
2cfmv12_268R    ..................................................  250
2cfmv12_268C    ..................................................  250
2cfmv12_530R    ..................................................  250
2cfmv12_533K    ..................................................  250

                        260       270       280       290       300
2cfmwt          DGARVQVHKDGSKIIVYSRRLENVTRAIPEIVEALKEAIIPEKAIVEGEL  300
2cfm3x          ................s.................................  300
2cfm3x+F318I    ................s.................................  300
2cfmv12         ...f............s.................................  300
2cfmv34         ...n............s.................................  300
2cfmv1314       ...n............s.................................  300
2cfmv12_268R    ...f..............................................  300
2cfmv12_268C    ...f............c.................................  300
2cfmv12_530R    ...f............s.................................  300
2cfmv12_533K    ...f............s.................................  300

                        310       320       330       340       350
2cfmwt          VAIGENGRPLPFQYVLRRFRRKHNIEEMMEKIPLELNLFDVLYVDGQSLI  350
2cfm3x          ..................................................  350
2cfm3x+F318I    ................i.................................  350
2cfmv12         ..................................................  350
2cfmv34         ..................................................  350
2cfmv1314       ................i.................................  350
2cfmv12_268R    ..................................................  350
2cfmv12_268C    ..................................................  350
2cfmv12_530R    ..................................................  350
2cfmv12_533K    ..................................................  350

                        360       370       380       390       400
2cfmwt          DTKFIDRRRTLEEIIKQNEKIKVAENLITKKVEEAEAFYKRALEMGHEGL  400
2cfm3x          ..................................................  400
2cfm3x+F318I    ..................................................  400
2cfmv12         ..................................................  400
2cfmv34         ..................................................  400
2cfmv1314       ..................................................  400
2cfmv12_268R    ..................................................  400
2cfmv12_268C    ..................................................  400
2cfmv12_530R    ..................................................  400
2cfmv12_533K    ..................................................  400
```

```
                    410         420         430         440         450
2cfmwt          MAKRLDAVYEPGNRGKKWLKIKPTMENLDLVIIGAEWGEGRRAHLFGSFI   450
2cfm3x          ..................................................   450
2cfm3x+F318I    ..................................................   450
2cfmv12         ..................................................   450
2cfmv34         ..................................................   450
2cfmv1314       ..................................................   450
2cfmv12_268R    ..................................................   450
2cfmv12_268C    ..................................................   450
2cfmv12_530R    ..................................................   450
2cfmv12_533K    ..................................................   450


                    460         470         480         490         500
2cfmwt          LGAYDPETGEFLEVGKVGSGFTDDDLVEFTKMLKPLIIKEEGKRVWLQPK   500
2cfm3x          ..................................................   500
2cfm3x+F318I    ..................................................   500
2cfmv12         ..................................................   500
2cfmv34         ..................................................   500
2cfmv1314       ..................................................   500
2cfmv12_268R    ..................................................   500
2cfmv12_268C    ..................................................   500
2cfmv12_530R    ..................................................   500
2cfmv12_533K    ..................................................   500


                    510         520         530         540         550
2cfmwt          VVIEVTYQEIQKSPKYRSGFALRFPRFVALRDDKGPEDADTIERIAQLYE   550
2cfm3x          ..............................d..h..............   550
2cfm3x+F318I    ..............................d..h..............   550
2cfmv12         ..............................d..h..............   550
2cfmv34         ..............................d..h..............   550
2cfmv1314       ..............................d..h..............   550
2cfmv12_268R    ..............................d..h..............   550
2cfmv12_268C    ..............................d..h..............   550
2cfmv12_530R    ................................h..............   550
2cfmv12_533K    ..............................d................   550


                    560
2cfmwt          LQEKMKGKVESHHHHHH   567
2cfm3x          ................    567
2cfm3x+F318I    ................    567
2cfmv12         ................    567
2cfmv34         ................    567
2cfmv1314       ................    567
2cfmv12_268R    ................    567
2cfmv12_268C    ................    567
2cfmv12_530R    ................    567
2cfmv12_533K    ................    567
```

Figure 2.5: Sequences of the constructs utilizing the scaffold of 2cfm, a DNA Ligase from *Pyrococcus furiosus*.

Chapter 3

Methods

3.1 Cloning

Introduction of mutations

Mutations were introduced using primers with the desired sequence and creating fragments utilizing outside primers and the corresponding mutation primers. The fragments were used as templates in a subsequent amplification by PCR [32] utilizing only the outside primers. The program is as follows:

98°C	30 sec	
98°C	10 sec	} × 30 cycles
55°C	20 sec	
72°C	20 sec	
72°C	5 min	

Colony PCR

Colony PCRs were performed to identify correctly cloned variants. Single colonies were picked from a LB plate and used to inoculate a 5ml overnight culture in LB with the appropriate antibiotics as well as a template for a colony PCR. The tip of

the pipet used for colony picking was therefore dipped into 20μl of PCR mixture containing 2× 1μl of primer, 200μM of dNTP mix, 2μl of TAC buffer, 1.5mM of $MgCl_2$, 2 units of Taq polymerase and water up to the final volume. The mixture was put into a thermocycler and the according PCR program was run:

95°C	5 min	
95°C	20 sec	} × 30 cycles
55°C	20 sec	
72°C	30 sec	
72°C	5 min	

After the reaction finished, the solution was spun down in a tabletop centrifuge and loaded onto an agarose gel.

Agarose gel electrophoresis

Agarose gels between one and two percent were used to separate DNA fragments of different lengths. The DNA - together with SYBR green dye - was loaded into the pockets of the gel and roughly 70V of voltage were applied. After the electrophoresis, blue light was used to detect the position of the DNA bands. Pictures were taken under UV light.

Plasmid extraction

Plasmids were extracted from 5 ml over night cultures. The cells were spun down by centrifugation (10min, 4000rpm). The plasmids were purified with the NucleoSpin Plasmid EasyPure kit from Machery-Nagel.

DNA digestion and ligation

DNA fragments were digested with the appropriate enzymes in 1 × Fast Digest buffer. Vector digestions were additionally treated with fastAP. FastAP is a phosphatase, dephosphorylating the ends of the plasmid, thus preventing religation. Ligation took place in 20 μl of total volume with a mixture of digested fragment and vector, 1× fresh ligation buffer and 200 units of T4 DNA ligase. The reaction was left for one hour at room temperature or - for higher efficiency - at 4°C over night. The reaction as a whole was transformed into chemically competent cells. The cells were plated, grown and single colonies were picked and used for further testing.

Sequencing

50 ng of DNA template were mixed with 2 μl of sequencing buffer, 0.5μl of BDT mix and primer to a final concentration of 1μM. Water was added to a volume of 10μl and the reaction was put on a thermocycler utilizing the sequencing program:

$$\left.\begin{array}{ll} 96°\text{C} & 20 \text{ sec} \\ 50°\text{C} & 10 \text{ sec} \\ 60°\text{C} & 4 \text{ min} \end{array}\right\} \times 30 \text{ cycles}$$

The samples were then brought to our in-house sequencing service and the resulting sequences were analyzed.

Competent cells

Chemically competent cells

For chemically competent cells, 400ml of LB medium was inoculated with 4ml of overnight culture and grown at 37° C until the OD_{600} reached 0.5. The cells were centrifuged for 15min at 4000rpm and 4° C. The pellet was then resuspended in 20ml cooled 100mM $CaCl_2$. The cells were again centrifuged as before and the pellet resuspended in 2ml 100mM $CaCl_2$, 15% glycerol. Aliquots of 100μl were frozen and stored at -80°C.

Electrocompetent cells

50ml of LB medium was inoculated with 0.5ml overnight culture and grown at 37°C until the OD_{600} reached 0.6. The culture was then cooled on ice for 30min followed by centrifugation at 4000rpm, 4°C for 10min. The pellet was resuspended in cold, sterile water and incubated on ice for 15min. The centrifugation, resuspension and incubation on ice steps were repeated several times while the resuspension volume was decreased to 20ml, 10ml and finally 1ml. Aliquots of 100μl were frozen and stored at -80°C.

Transformation

Heat shock transformation of chemically competent cells

Chemical competent cells were thawed on ice for five minutes. Plasmid was added to the cells, mixed and incubated for ten minutes. The mixture was then heated to 42°C for 45 seconds (20 seconds if *E. coli* Arctic Express^{TM} was used) and put back on ice for an additional ten minutes. After that 900μl of LB medium was added to the mixture and the bacteria were incubated for one hour at 37°C while shaking gently. This micro culture was then used to inoculate a culture with medium and antibiotics according to the plasmid used (100μg/ml for Ampicillin and 50μg/ml for Kanamycin) or plated onto a petri dish with LB and antibiotics.

Transformation of electrocompetent cells

The sample containing the plasmid was dialysed on distilled water for 15min to remove salt. The cells were mixed with the sample and incubated on ice for 10min. The mixture was put into a chilled GE Healthcare Gene Pulser cuvette and inserted into the electroporator. The program EC2 was chosen and the charge was released. 900μl LB were added to the cells and incubated at 37°C for 1h. The cells were then plated onto a petri dish with LB and the appropriate antibiotics.

3.2 Protein expression and purification

Expression

For expression a small preculture was grown over night in the according medium and with antibiotics present. 5-10ml of that preculture were used to inoculate 2l of medium in a 5l Erlenmeyer flask. The cultures were grown at 37°C shaking at 180rpm until an OD_{600} of 0.8 was reached. The shaker was set to expression temperature and expression was induced by addition of 2 ml 1M IPTG. After a certain time dependent on the construct expressed (4 hours for 2cfm and 3N8M, around 14 hours for the other scaffolds), the cells were harvested by centrifugation for 12min at 4000rpm. The pellet was washed in 40 ml of buffer usually consisting of 150mM NaCl and 50mM potassium phosphate buffer at pH8. If the protein was used for crystallization the potassium phosphate buffer was substituted with 50mM TRIS buffer. The cells were again centrifuged at 4000rpm for 15min and the cells were then solubilized in a buffer dependent on the construct used and the experiment planned (in most cases 150mM NaCl, 50mM potassium phosphate buffer or TRIS if the protein was used for crystalization. There were two exceptions: For the early round of 3N8M designs the pH was set to 7.5, while for the 2D52 variants 100mM NaCl, 50mM HEPES pH 7 was used).

Cell rupture by sonication

E. coli cells from 2l of expression were resuspended in 40ml buffer and poured into a 50ml falcon tube. These volumes were adjusted for larger expression volumes. The tube was put into a beaker filled with water and ice and placed under the sonifier. Sonification took place for ten minutes at an amplitude of 45% and with a pulse sequence of 1:2 pulse:pause (usually 1s:2s).

Freeze/thaw lysis

Freeze/Thaw lysis was used as a high throughput method for lysing of small volumes of many constructs in parallel. The cells were mixed with buffer to a final volume of 1ml and poured into 1ml eppendorf tubes. The tubes were closed and repeatedly frozen in liquid nitrogen followed by immediate thawing in warm water. Depending on the experiment, at least five cycles were applied.

Purification

Affinity chromatography

The basis of separation in this method is the affinity of different compounds to the matrix of the column. For the proteins purified in this work, a C-terminal tag consisting of six histidines was added. This leads to a strong affinity towards metals such as cobalt or in our case Nickel which was attached to the sepharose resin.

Size exclusion chromatography

Size exclusion chromatography or gel filtration is a chromatographic technique separating chemical compounds such as proteins according to their physical size. Separation takes place in a column filled with an adsorbent material. Particles of this material have pores of a defined diameter on their surface. Depending on the size of the protein it will enter such pores with a certain probability and if a flow is applied to the column thus remain longer in the stationary phase. Larger compounds will have a smaller probability to enter these pores and therefore remain longer in the mobile phase. If a mixture of proteins is loaded onto the column larger proteins will therefore elute first.

Using the elution time of certain standard proteins with known masses, one can obtain a characteristic curve for a given column (see Figure 3.1) and can use the elution time of a protein to estimate its mass or - for proteins with known mass - its oligomerization state.

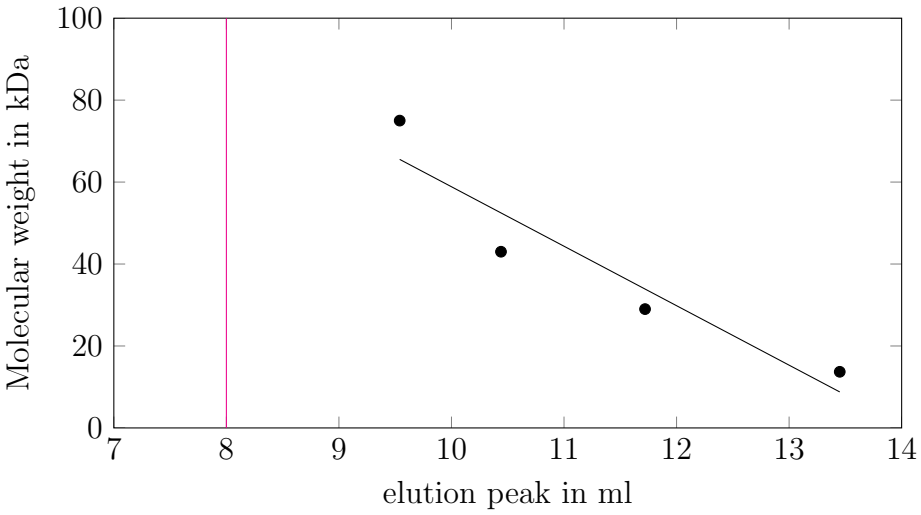


Figure 3.1: Example for a calibration curve of a size exclusion column, in this case an analytical Superdex 75. The magenta line represents the void volume of the column. The black line is the linear regression of the results from the calibration proteins. Here the LMW gel filtration calibration kit from GE Healthcare was used in several consecutive gel filtration runs, resulting in datapoints for ribonuclease A (13.7 kDa), carbonic anhydrase (29kDa), ovalbumine (44kDa) and conalbumine (75kDa). The void volume was determined with blue dextran 2000.

3.3 Protein characterization

Polyacrylamide gel electrophoresis

Polyacrylamide gel electrophoresis is a method to determine the amount and size of different protein compounds in a mixture. The protein mixture is heated to 95 °C for 5 min in the presence of Sodium dodecyl sulfate (SDS). SDS attaches to the protein, both denaturing it up to the point where it is present in a linear shape as well as leading to a homogeneous negative surface charge. The linearized fragments with the attached charged SDS is then loaded onto a polyacrylamide gel and a current is applied. The charged fragments now move to the anode with a speed that is dependent on the size of the linearized proteins and smaller proteins moving faster than large proteins. The current is stopped before the smallest proteins reach the border of the gel and protein bands are stained with coomassie blue followed

by destaining in warm water or 10% acetic acid. The protein size corresponding to positions of certain bands can then be determined by comparison with a standard mixture of proteins with defined and known masses.

Multiangle light scattering

Multiangle light scattering (or MALS) makes use of different scattering characteristics of a protein sample in solution. Since the amount of scattering strongly depends on the (known) wavelength and the size of the particles in solution, this can - under the assumption that the particles are spherical [33] - be used to determine the particles average size. Experimentally, we used an analytical gel filtration column to separate possibly different protein populations. The MALS measurements were done in real time directly after the column. A tight laser beam is send through a measurement cell with the filtered efflux of the gel filtration column. Light scattering is then measured at a variety of different angles from the measuring cell.

Circular dichroism

In order to determine the amount and type of secondary structure present in a protein, UV circular dichroism measurements can be utilized. Circular dichroism describes a property of many chiral substances. If a solution of an optical active chiral molecule is measured with left- and right handed polarized light there will be a difference in absorption dependent on the wavelength and properties of the molecule (see figure 3.3). The difference between absorption at a given wavelength is measured in degrees of ellipticity (see figure 3.3). In a circular dichroism spectrum, this ellipticity is measured for a range of wavelengths and plotted against them.

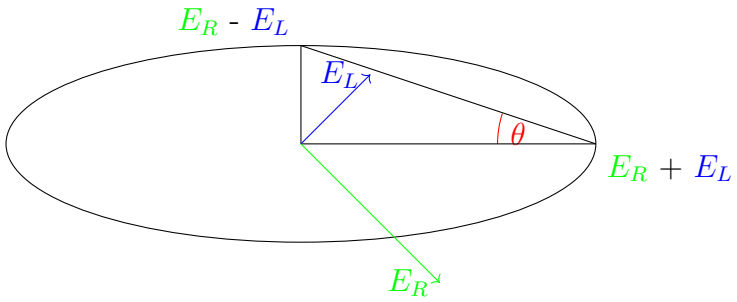


Figure 3.2: Graphical illustration of the measure of ellipticity θ used to compare the secondary structure content of different proteins

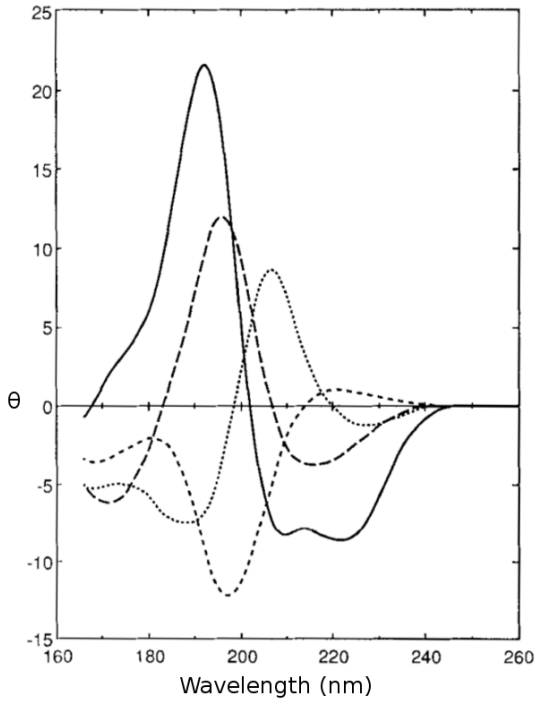


Figure 3.3: Typical CD spectra of different secondary structure content. Solid line: α-helix, coarse dashes: β-sheet, points: β-turn and fine dashes: random coil. From Johnson, W C, Protein secondary structure and circular dichroism: a practical guide. Proteins, 1990, Vol. 7 Issue 3 p. 205-214

For protein solutions, unordered structure and loops as well as α-helices and β-sheets will show specific circular dichroism spectra in the far UV region. Depending on the form of the spectrum one obtains from a protein solution, it is possible to estimate the amount of secondary structure present and thus have an estimate of the amount of folded protein [34] [35].

If not specified, all samples were measured at room temperature in a 1mm cuvette in a J-810 CD-Spectrometer (Jasco).

Melting curves

Proteins are stable only within a certain temperature range and most in fact unfold in aqueous solution before the boiling point of water is reached. The thermal stability is an important factor and determines sample handling as well as the possibility to perform certain experiments such as NMR. In order to compare different proteins with different melting characteristics, the melting point is introduced and defined as the temperature where half of the protein is still in a folded state.

Experimentally the melting point is determined by heating up a protein sample at a constant rate (usually 1°C per minute) while continuously measuring the change in circular dichroism at a wavelength that shows a large change upon unfolding (for proteins with a strong α-helical content this is usually at 222nm). The result for a single step unfolding will ideally be a sigmoid curve, which is partially constant in both the lower (colder) region and the upper (hotter) region resembling the completely folded and completely unfolded state. The melting point is then the temperature where the plot passes the arithmetic mean of the CD signal of the folded and unfolded state.

Tryptophan fluorescence

The indole ring of tryptophan has a strong fluorescence when excited with a wavelength of 280nm. Typically emission is measured between 300 and 400nm. Tryptophan is a solvatochromic fluorophore, which means that it changes both its emission maximum and its intensity dependent on the polarity of its environment. Tryptophans within folded proteins are often buried in the hydrophobic core. However, upon unfolding they will be exposed to solvent, leading to an increasingly polar environment. Changes in tryptophan fluorescence can therefore be utilized to monitor loss of tertiary structure within a protein.

If not specified, all samples were measured in a 1cm cuvette in a FP6500 fluorescence Spectrophotometer (Jasco) at 25°C.

Chemical denaturation and Gibbs energy

There are a variety of compounds that have the ability to denature folded proteins, leading to a loss of structure. Among these are reagents that change the pH or organic solvents. The changes in secondary and tertiary structure can be measured using CD spectroscopy and tryptophan fluorescence. Two of the most common denaturants are urea and guanidinium hydrochloride. They belong to the group of chaotropic agents, signifying that they increase entropy and in high concentrations lead to the disruption of hydrogen bonds as well as van der Waals interactions.

Since the destabilizing effect of these reagents is nearly linear to its concentration, titration experiments can be used to determine the actual stability of a protein by calculating the difference in Gibbs energy between the folded and the unfolded state [36].

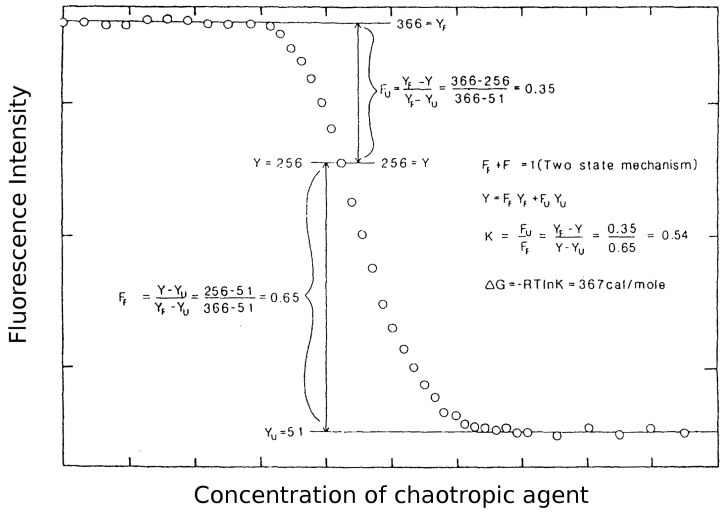


Figure 3.4: Determination of the fraction of unfolded protein from an unfolding curve. From [37], page 302

In order to do so the signal for a completely folded and a completely unfolded protein is determined and all measurements are normalized according to these values resulting in a plot signifying the fraction of folded protein at a given concentration of denaturing agent. The change in Gibbs energy is then calculated based on a concentration range where the plot is both linear and at its steepest (see figure 3.4). For every point in this range, the change in Gibbs energy at the according concentration of denaturand is calculated as

$$\Delta G = -RT \ln K$$

where R is the gas constant, T the temperature of the sample and K the fraction of unfolded/folded protein. The different ΔG values are then plotted against the concentration of denaturing agent and extrapolated to the y-intercept resembling a concentration of no chaotropic reagent. The energy at this point gives the change in Gibbs energy upon unfolding (see chapter 4, figure 4.16).

NMR

Nuclear magnetic resonance spectroscopy uses the spin of atomic nuclei to determine their state. Atoms where the sum of protons and neutrons is odd have a spin that can with a certain probability be aligned in a magnetic field. The fraction of atoms where this alignment takes place is dependent on the strength of the magnetic field applied with stronger fields resulting in better alignment and higher signal. Using electromagnetic pulses one can now induce rotation into the spin axis of the aligned nuclei. Depending on the frequency of the pulse, different nuclei can be induced specifically. The rotation and its decay over time can be measured. This so called free induction decay is different for atoms in different environments. In this work one dimensional ^{1}H NMR was used to determine the presence of tertiary structure in proteins which results in specific peaks in the Fourier-transformed NMR spectrum.

X-ray crystallography

In this chapter I give a short introduction to x-ray crystallography. In general, this is based on 'Biomolecular Crystallography' by Rupp [38].

A short introduction

In a crystal lattice the elements forming the crystal (in our case proteins) are arranged in a symmetric, periodic way. For this introduction into the classification of crystal lattices let us assume that we have an infinite lattice $\mathcal{L}$ in $\mathcal{R}$ consisting of copies of a protein. We then define the unit cell $\mathcal{U}$ as the (by volume) smallest periodic subset of $\mathcal{L}$. The size of $\mathcal{U}$ is described by the length of the edges.

For the tetragonal-trapezohedral cells in this work all three edges are orthogonal to each other. In general there are several possible unit cells which depend on the space group. For the trivial space group p1 every translation of a unit cell is again a unit cell which results in an infinite amount of possible unit cells. This example also demonstrates that in general it is not necessary to have a complete and connected structure within a single unit cell. It is sufficient to have a complete and connected structure in our unit cell when we identify each edge with its parallel counterpart resulting in a topology equivalent to that of a 3-torus or $\mathbb{R}^3/\mathbb{Z}^3$ (see figure 3.5).

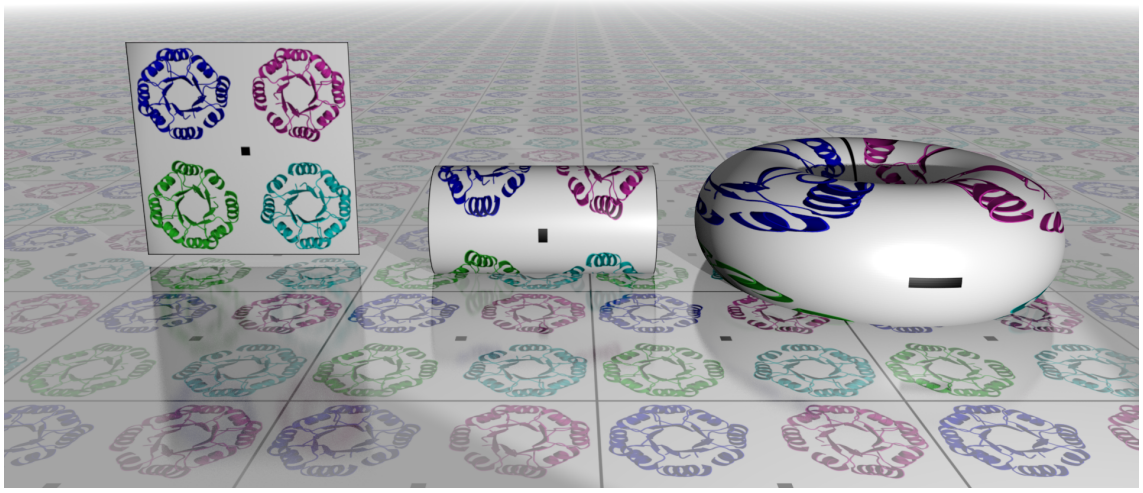


Figure 3.5: Example for the topology of a two dimensional unit cell. On the left a possible representation with the center marked by a black dot. In a crystal lattice this specific arrangement is repeated in all four directions of the crystal plane. Elements on the top of the arrangement are therefore close to elements at the bottom of the next repeat. In a unit cell representing the crystal lattice, the upper edge is therefore identified with the bottom edge, which leads to the topology of a cylinder (middle) by basically 'gluing' the edges together. For the same reason the right edge of the unit cell is identified with the left edge, and subsequently the right edge of the cylinder is identified with the left edge. Gluing these edges together leads to a torus topology (right). Three dimensional unit cells are in a similar way topologically equivalent to a 3-torus.

Point groups

Crystals lattices are classified according to their point groups. For a given unit cell the according point group $\mathcal{P}$ is defined as the set of all isometries $\mathcal{I}$ on $\mathcal{R}$ that leave the center c of the unit cell fixed, so

$$\mathcal{P} := \{i \in \mathcal{I} : i(\mathcal{U}) = \mathcal{U},\ \ i(c) = c\}$$

Since the unit cell is unchanged, the lattice is also invariant under the isometries of the unit cell. In a three dimensional space there are 32 different point groups.

Space groups

The space group $\mathcal{S}$ of a certain unit cell is the set of isomorphisms that map the unit cell on itself, so

$$\mathcal{S} := \{i \in \mathcal{I} : i(\mathcal{U}) = \mathcal{U}\}$$

Every space group can also be written as the combination of a point group with the set of discrete translations of a Bravais lattice. The set of point groups can therefore also be seen as the quotient of the set of space groups by the set of Bravais lattices. This immediately introduces a classification of the space groups according to the point group they relate to. Therefore, point groups are also called **crystal classes**. In a three dimensional space there are 230 different space groups. Determination of the correct space group is one of the first necessary steps in order to obtain a crystal structure and was in this work done with XDS [26].

Experimental setup

Protein samples were concentrated as high as possible. For each sample, Qiagen NeXtal screens were set up in 96 well sitting drop plates. The drop consisted of 0.5μl of protein sample and 0.5μl of screening buffer. The setup was pipetted automat-

ically with either the honeybee or mosquito robot by Sooruban Shanmugaratnam. The plates were stored at 20°C and imaged regularly. Grown crystals were fished (if necessary with additional cryoprotectant such as PEG400) and frozen in liquid nitrogen.

All spectra of proteins were recorded at the Swiss Light Source of the Paul Scherrer Institute in Villigen/Switzerland. For this work we used the PXII (X10SA) beamline. The defaults of the measurements are given in the appropriate tables under results.

Determination of space group and unit cell

Before beginning the actual structural work, it is necessary to determine the actual crystal lattice present and describing it in terms of its unit cell as well as its space group. This work was done using the XDS suite [26].

Molecular replacement

Since X-ray crystallographic spectra recorded with usual sensors only contain intensities ($\sim$ amplitude) and no phase information, it is impossible to directly calculate the corresponding electron density [39]. However, since the spectra should equal the Fourier transform of the electron density, the measured intensity of the spectrum can be compared with the amplitude of the Fourier transformed electron density. Historically this has been done with Patterson maps, however, in this work I used phaser utilizing a more modern method based on a maximum-likelihood estimation of the electron density. In all cases one has to give a reference structure that is used as a basis for the estimations. This reference should be as similar as possible to the expected structure and show no large rearrangements. Since the structure in the unit cell is yet unknown at the state of molecular replacement, great care should be used not to bias a possible solution too much with the reference. I therefore used

reference structures consisting only of the protein backbone and rejected solutions not showing density fitting the side chains of the newly build model.

Model building

Once an initial model is build by molecular replacement, it has to be optimized and arranged according to the density obtained. This was done alternating manual model building and optimization in coot and refinement with phenix.refine from the phenix suite vs 1.9-1692 [27]. Parts where the initial model was obviously wrong (often loops with a low density and high B values) were removed and as far as possible rebuild manually into the visible density. Also sidechains that were placed yet had no density justifying their placement were deleted.

The weights for the different optimization steps in phenix.refine were mostly chosen manually and often changed during the process of building a final model in order to address the problems most pressing at the given time.

The model was determined sufficient based on a variety of parameters, most notably the R_{free} and R_{work} but also the B factors, deviations from ideal bond angles and lengths, calculated clashscores and the occurance of ramachandran and rotamer outliers.

Mass spectrometry

Mass spectrometry is used to measure the charge to mass ratio of either complete proteins or fragments. For this work two different methods have been used.

Matrix-assisted laser-desorption/ionisation time of flight (MALDI-TOF) measurement

For this measurements a protein is mixed with another compound, the so called matrix, and quickly co-precipitated. The mixture is ablated with a laser and the ionized particles are accelerated through a vacuum tube. The time of flight t corresponds to the mass m and the charge z:

$$t \sim \sqrt{\frac{m}{z}}$$

Measuring the position of different peaks in counts/t, it is possible to determine the absolute mass of the protein used.

Electrospray ionization Mass spectrometry (ESMS)

In contrast to MALDI, electrospray ionization works with liquid protein solutions by charging the solution in a capillary until it is turned into an aerosol at the end of the capillary by electrospray. In contrast to MALDI this technique is more gentle, leading to less fragmentation of large macromolecules such as proteins.

Tryptic digest

It is possible to digest proteins of interest with proteases such as trypsin before mass spectrometry and measure the fragments obtained from this digestion. This can be used to identify proteins by comparing the spectra obtained with a calculated spectrum that would be expected for a certain protein digested with certain proteases. It is also possible to locate modifications of the protein more precisely by identifying

the fragments that show a mass shift relating to a particular modification.

3.4 Enzymatic activity assays

In order to measure enzymatic activity, chromophoric substrates were used (see figure 3.6). The typical test consisted of 100μM of substrate (usually a p-nitroanilide derivate) and 0.1 mg of purified protein in a buffer depending on the construct (usually NaCl and either potassium phosphate or TRIS buffer at pH 7.5 or pH 8.0). The reaction volume was 1ml and reactions took place in either a glass or single-use plastic cuvette. Since the released 4-nitroaniline has a much higher absorption at 410 nm than the uncut substrate (see figure 3.7) in aqueous solution, changes in absorption were measured at this wavelength. The reactions were followed spectroscopically for several hours in a Cary 50 Scan UV-VIS spectrometer. Additionally, spectra were recorded before and after the measurements to detect unspecific changes in the absorption such as precipitation of the protein or fogging of the cuvette. The reaction temperature was set to constant 20 or 25°C. Since especially the ester substrates, but to a lower degree also the peptide substrates showed a high autolysis in aqueous solution, measurements included the substrate without protein as a standard.

Enzymatic turnover speeds were derived from the changes in absorption and used to calculate parameters such as k_{cat} or k_M.

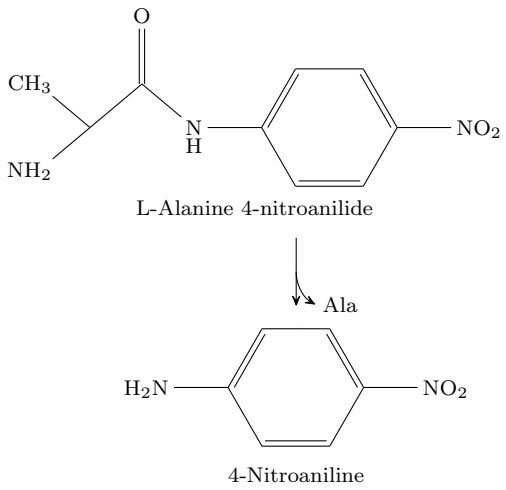


Figure 3.6: Alanine-p-nitroanilide or ApNA, the substrate used in the design study. Product formation can be followed at 410nm.

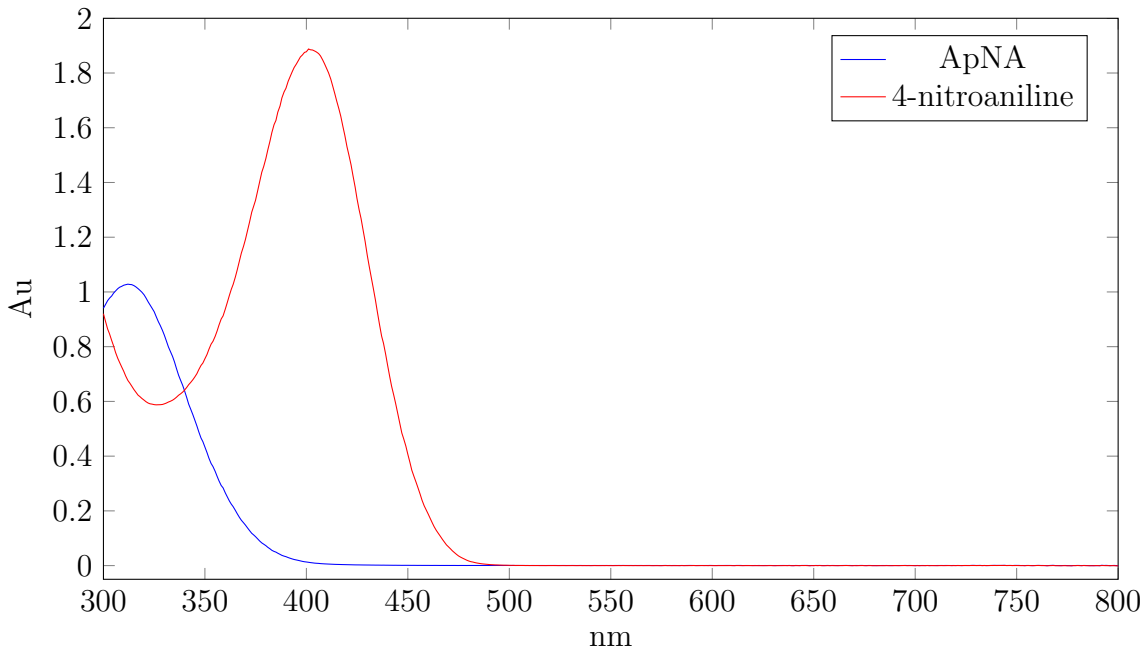


Figure 3.7: Comparison of the spectra of 100μM ApNA and 100μM 4-nitroanilide. At 410nm ApNA has a very low absorption, while 4-nitroanilide and thus the cleaved product shows a strong absorption.

3.5 Inhibition assays

There are a variety of known inhibitors of the catalytic serine triad. In this work both PMSF and AEBSF were used. The mechanism of inhibition is the same for both compounds. Activated serines from catalytic triads will cleave fluor from the inhibitor and form a covalent bond analogous to the first tetrahedral intermediate of the reaction. In contrast to the substrate the inhibitor is not released, and by occupying the active oxygen of serine the triad is not active any more (see Figure 3.8).

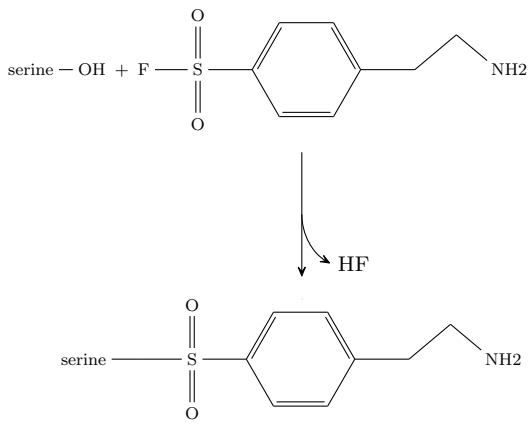


Figure 3.8: Inhibition of the catalytic triad with AEBSF. The compound functions as a substrate analog up to the point where the tetrahedral intermediate is formed. Upon formation of the first tetrahedral intermediate the reaction stops and the serine is covalently bound to the inhibitor, rendering the triad nonfunctional.

Chapter 4

De-novo design of a TIM-barrel

4.1 Introduction

TIM-barrels are probably both the most versatile and the most common fold for enzymes. Thus far, members of the TIM-barrel fold have been found in all enzyme classes except for ligases [40] [41]. They are found in all three domains of organisms alike and catalyze many basic reactions. In fact the name TIM-barrel itself derives from the triosephosphate isomerase, an enzyme essential for nearly every living organism. Due to their importance in many basic reactions these enzymes are well-studied and currently (march 2016) contribute about 2% of all the structures in the PDB [41].

TIM-barrels belong to the α/β class of the SCOP [42] classification. They consist of eight α/β elements with the β sheets forming the inner barrel and the α helices on the outside. The barrel itself is held together by hydrogen bonds between backbone atoms of the parallel β sheets. In all natural TIM-barrel enzymes, the catalytic side is at the top of the barrel and the helix-dipoles are often used to help binding negatively charged substrates [43].

Evolutionarily, TIM-barrels are likely to have evolved from an at least twofold

symmetric ancestor consisting of two half barrels [44] [45] [46] [47] [48] although also a fourfold symmetric ancestor utilizing four quarter barrels is plausible [49]. The question whether all present TIM-barrels share a common origin or whether they evolved independently is not yet completely answered, although much hints to the fact that at least the majority of the known TIM-barrels indeed have a common ancestor [50] [40] [47].

Due to the importance the TIM-barrel fold has for enzymes and to shine some additional light onto the evolutionary aspects of this fold, the decision was made to design a TIM-barrel completely *de novo*.

4.2 Design

Design principles

There are a lot of prior attempts to design a TIM-barrel [9] [10] [11] [12] [13] [14]. In contrast to many of these attempts, we did not try to redesign modern TIM-barrels, but instead go for a strategy that focuses on a minimal model of an idealized TIM barrel. As a first step to minimize our design problem we decided to create a construct that is symmetric on both the sequence as well as the structure level. We therefore looked into possible symmetries.

Since the TIM-barrels consist of eight α/β motifs, the idea of an eightfold symmetric TIM-barrel lies at hand. However, such a symmetry is structurally not feasible. The reason for this lies in the interactions of the β sheets forming the inner barrel: Since the main contributing interactions within β sheets are hydrogen bonds between the backbone atoms, neighboring β strands have to be in the same pleat pattern, thus show a register shift of multiples of two amino acids or have alternating pleat patterns (see Figure 4.1). The sum of all shifts of residues in neighboring β strands is called the shear number of the barrel. In the case of the TIM-barrel the shear number is eight, while the register shifts are usually not symmetrical over the barrel (see Figure 4.1a for triosephosphate isomerase as an example). An eightfold symmetric TIM barrel is therefore not possible since the requirement to be in the same pleat pattern would require each β strand to have a register shift of two compared to its neighbor, leading to a shear number of sixteen. For a fourfold symmetrical arrangement on the other hand there are three possibilities. A register shift of one between each β strand as shown in Figure 4.1b with an alternating pleat pattern leads to half of the loops connecting α helices and β strands with helix-facing residues which is not feasible [8]. Same goes for the possible arrangement pictured in Figure 4.1c where each β strand ends with a helix facing amino acid. This leaves

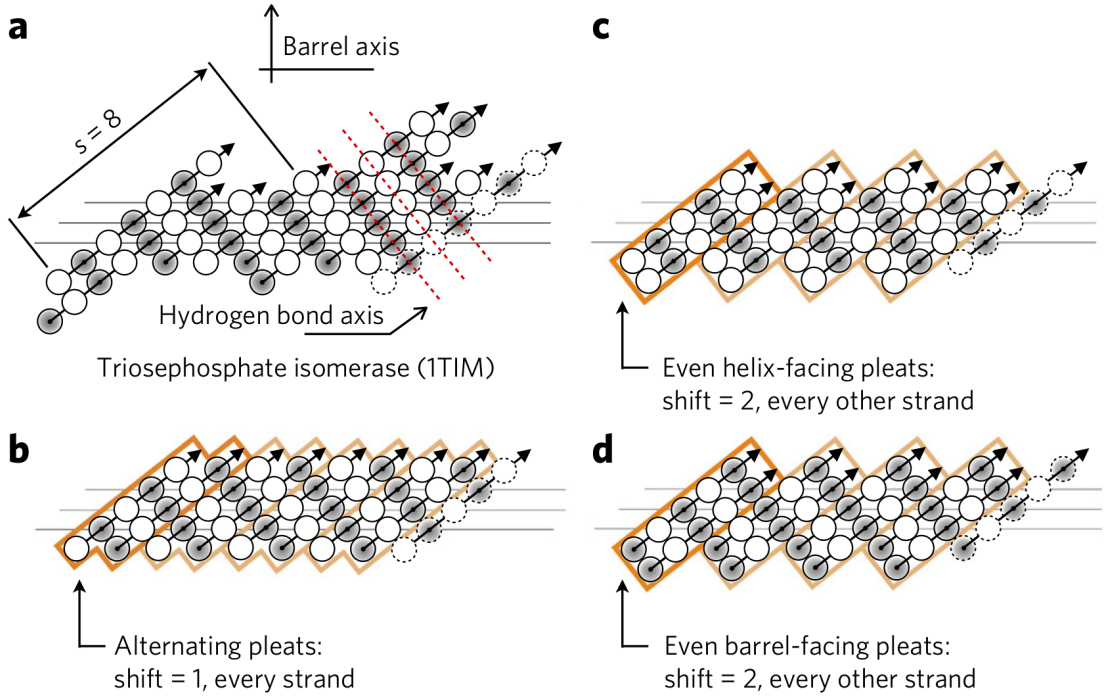


Figure 4.1: Possible pleat patterns of TIM barrels. (a) shows the non symmetric solution from Triosephosphate isomerase. (b) One possible solution with alternating pleats is not feasible due to the fact that every second lower loop has to connect a α helix to a β strand with a helix facing terminus. (c) is even worse with every lower loop now having to make such a connection. (d) shows the remaining solution we focused on in this work. From [1]

.

only one fourfold symmetric possible set of register shifts and pleat patterns, pictured in Figure 4.1d. Since all loops now have to connect helices with barrel-facing residues this design should be feasible [8].

Computational design

In order to create a starting model, the Rosetta suite was used to calculate different *de novo* designs of the quarter barrel. Each design featured β-strands with a length of five amino acids and variable lengths of the loops and α-helices. For each combination 2000 designs were calculated. The only combination for a quarter barrel which resulted in a structure with a closed barrel in the simulation consisted of 46 amino acids arranged as $5_{strand1} + 3_{loop} + 13_{helix1} + 3_{loop} + 5_{strand2} + 3_{loop} + 11_{helix2} + 3_{loop}$ (see Figure 4.2). Although other designs have been tested as well, this ultimately superior combination of secondary elements followed the principles for ideal loop lengths as described in [8]. All computations were performed by Possu Huang in David Bakers group.

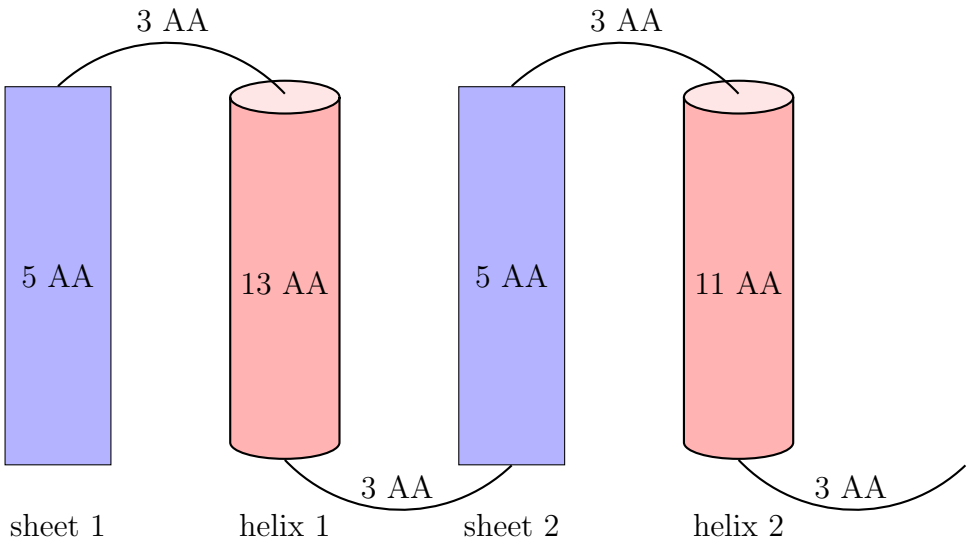


Figure 4.2: Schematic representation of the best solution of possible length of secondary structure elements found for a quarter barrel. AA = amino acids.

Additional design features

Several other design features were introduced manually to increase the chance of a correctly folded TIM-barrel as described below [1].

One of the amino acids in the loop between the different quarters was forced to be an aspartate. This design choice was made due to the exposed amide that is a result of the register shift between the two neighboring β-strands. The aspartate in the loop can create a hydrogen bond to the otherwise solvent-exposed amide and contribute to protein stability. In a similar fashion one of the amino acids in the flanking α-helix was set to an arginine to interact with a free carbonyl in the β/α-loop number 2 (see Figure 4.3a).

Hydrogen bonds to the backbone of the α/β-loop number 2 were also introduced manually by the introduction of a serine into the loop and a glutamine to the adjacent loop (see 4.3b).

Another constraint was the mandatory presence of at least one valine or leucine in the flanking α-helices interacting with the corresponding β-strands (see Figure 4.3c, position 17, 20 and 38).

One major problem could have been the correct spacing between the α-helices. While the diameter was mainly determined by the diameter of the inner barrel defined by the β-sheets, defined spacing and angles between neighboring helices was important to achieve a uniform shielding of the β-strands. This was realized by the introduction of several tryptophans at the interface of the helices (see Figure 4.3d).

In order to maintain a high degree of symmetry even in the region of the (necessary non-symmetric) termini, two cysteines were introduced in one of the last designs into the N,- and C terminus. They were supposed to form a disulfide bond and force these regions close together.

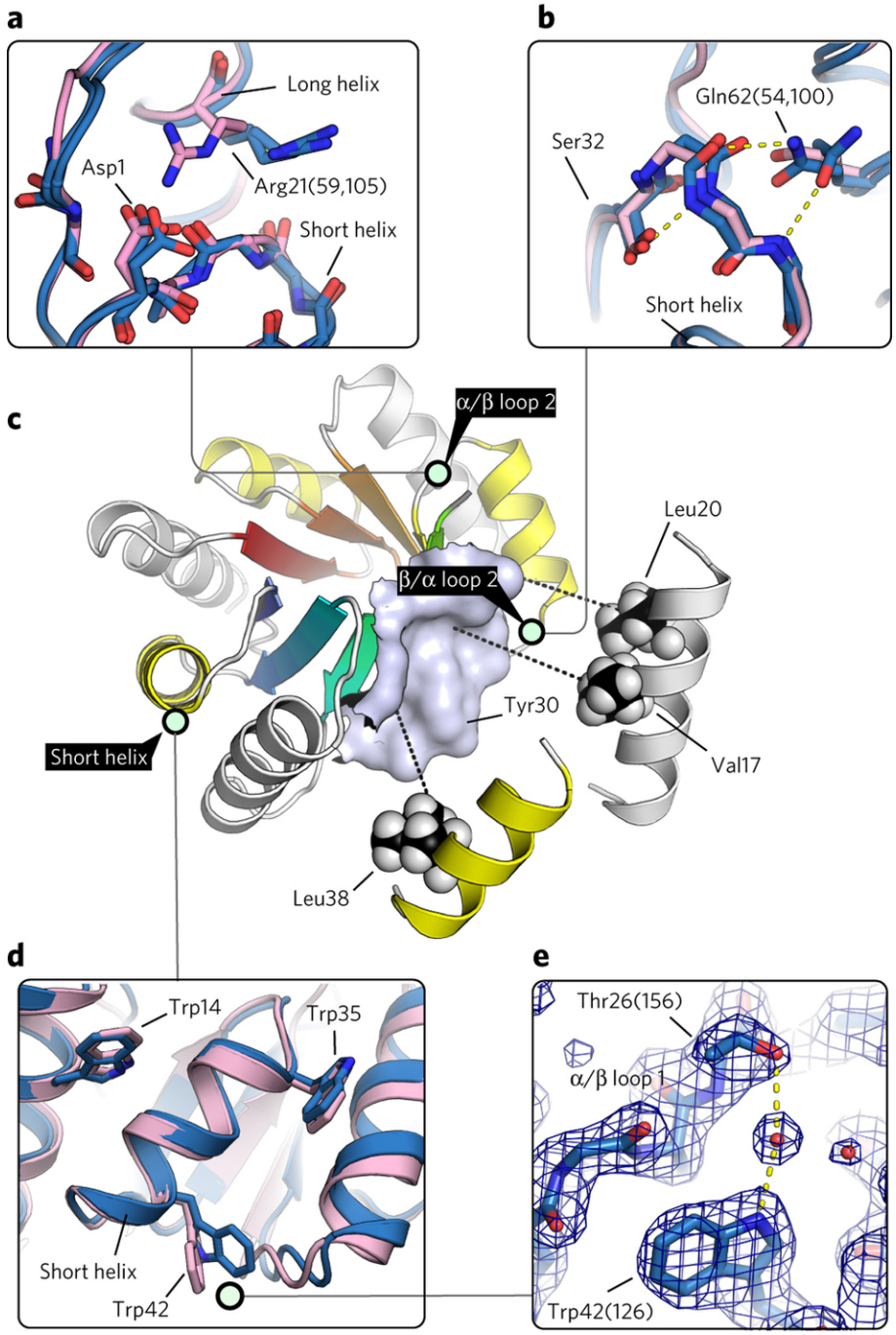
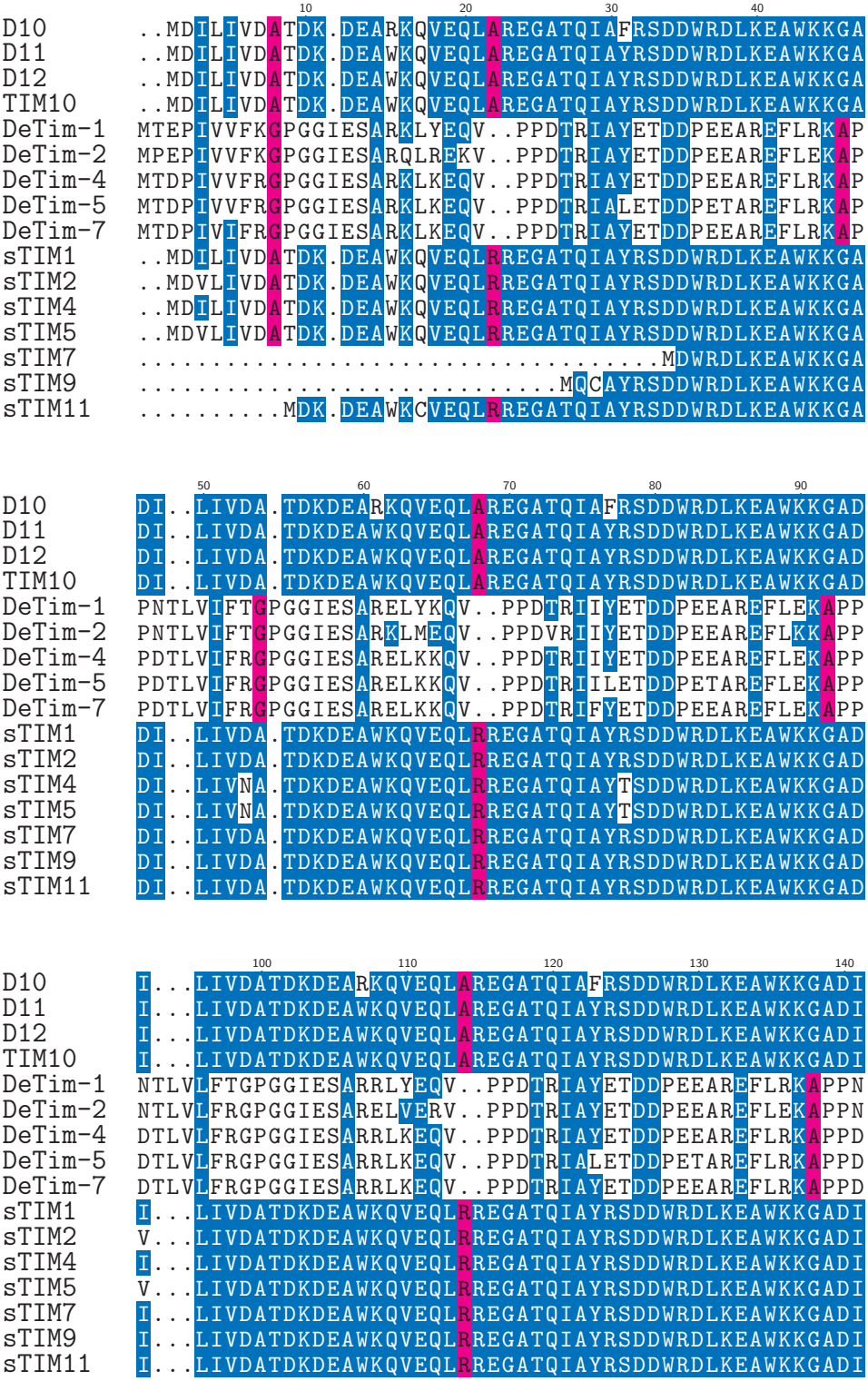


Figure 4.3: Overview of the different interactions enforced in the TIM-barrel design. **(a)** Hydrogen bonds introduced to saturate otherwise solvent-exposed features. **(b)** Interactions introduced into one of the lower loops to restrain loop flexibility and increase stability. **(c)** Different features manually introduced into the design. **(d)** Tryptophans used as spacers to control the distance between different helices. **(e)** The designed hydrogen bond between T26 and W42 seems to be water-mediated according to the crystal structure obtained. From [1].

4.3 List of constructs

During this work a variety of constructs were expressed and tested (see figure 4.3). Listed here is an alignment of the different sequences. First constructs were the D variants as well as TIM10. The De variants missed the Trp35 and Trp42 described in chapter 4.2. These constructs were not properly folded. Introduction of Arg21 in sTIM1 and all later sTIM variants increased the thermal stability of the constructs. The sTIM variants of the last generation of designs additionally feature different circular permutations as can be seen in the alignment.


10 20 30 40
D10 ..MDILIVDATDK.DEARKQVEQLAREGATQIAFRSDDWRDLKEAWKKGA
D11 ..MDILIVDATDK.DEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGA
D12 ..MDILIVDATDK.DEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGA
TIM10 ..MDILIVDATDK.DEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGA
DeTim-1 MTEPIVVFKGPGGIESARKLYEQV..PPDTRIAYETDDPEEAREFLRKAP
DeTim-2 MPEPIVVFKGPGGIESARQLREKV..PPDTRIAYETDDPEEAREFLEKAP
DeTim-4 MTDPIVVFRGPGGIESARKLKEQV..PPDTRIAYETDDPEEAREFLRKAP
DeTim-5 MTDPIVVFRGPGGIESARKLKEQV..PPDTRIALETDDPETAREFLRKAP
DeTim-7 MTDPIVIFRGPGGIESARKLKEQV..PPDTRIAYETDDPEEAREFLRKAP
sTIM1 ..MDILIVDATDK.DEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGA
sTIM2 ..MDVLIVDATDK.DEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGA
sTIM4 ..MDILIVDATDK.DEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGA
sTIM5 ..MDVLIVDATDK.DEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGA
sTIM7 .....MDWRDLKEAWKKGA
sTIM9 .....MQCAYRSDDWRDLKEAWKKGA
sTIM11 .....MDK.DEAWKCVEQLRREGATQIAYRSDDWRDLKEAWKKGA
50 60 70 80 90
D10 DI..LIVDA.TDKDEARKQVEQLAREGATQIAFRSDDWRDLKEAWKKGAD
D11 DI..LIVDA.TDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGAD
D12 DI..LIVDA.TDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGAD
TIM10 DI..LIVDA.TDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGAD
DeTim-1 PNTLVIFTGPGGIESARELYKQV..PPDTRIIYETDDPEEAREFLEKAPP
DeTim-2 PNTLVIFTGPGGIESARKLMEQV..PPDVRIIYETDDPEEAREFLKKAPP
DeTim-4 PDTLVIFRGPGGIESARELKKQV..PPDTRIIYETDDPEEAREFLEKAPP
DeTim-5 PDTLVIFRGPGGIESARELKKQV..PPDTRIILETDDPETAREFLEKAPP
DeTim-7 PDTLVIFRGPGGIESARELKKQV..PPDTRIFYETDDPEEAREFLEKAPP
sTIM1 DI..LIVDA.TDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGAD
sTIM2 DI..LIVDA.TDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGAD
sTIM4 DI..LIVNA.TDKDEAWKQVEQLRREGATQIAYTSDDWRDLKEAWKKGAD
sTIM5 DI..LIVNA.TDKDEAWKQVEQLRREGATQIAYTSDDWRDLKEAWKKGAD
sTIM7 DI..LIVDA.TDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGAD
sTIM9 DI..LIVDA.TDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGAD
sTIM11 DI..LIVDA.TDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGAD
100 110 120 130 140
D10 I...LIVDATDKDEARKQVEQLAREGATQIAFRSDDWRDLKEAWKKGADI
D11 I...LIVDATDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGADI
D12 I...LIVDATDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGADI
TIM10 I...LIVDATDKDEAWKQVEQLAREGATQIAYRSDDWRDLKEAWKKGADI
DeTim-1 NTLVLFTGPGGIESARRLYEQV..PPDTRIAYETDDPEEAREFLRKAPPN
DeTim-2 NTLVLFRGPGGIESARELVERV..PPDTRIAYETDDPEEAREFLEKAPPN
DeTim-4 DTLVLFRGPGGIESARRLKEQV..PPDTRIAYETDDPEEAREFLRKAPPD
DeTim-5 DTLVLFRGPGGIESARRLKEQV..PPDTRIALETDDPETAREFLRKAPPD
DeTim-7 DTLVLFRGPGGIESARRLKEQV..PPDTRIAYETDDPEEAREFLRKAPPD
sTIM1 I...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM2 V...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM4 I...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM5 V...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM7 I...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM9 I...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI
sTIM11 I...LIVDATDKDEAWKQVEQLRREGATQIAYRSDDWRDLKEAWKKGADI

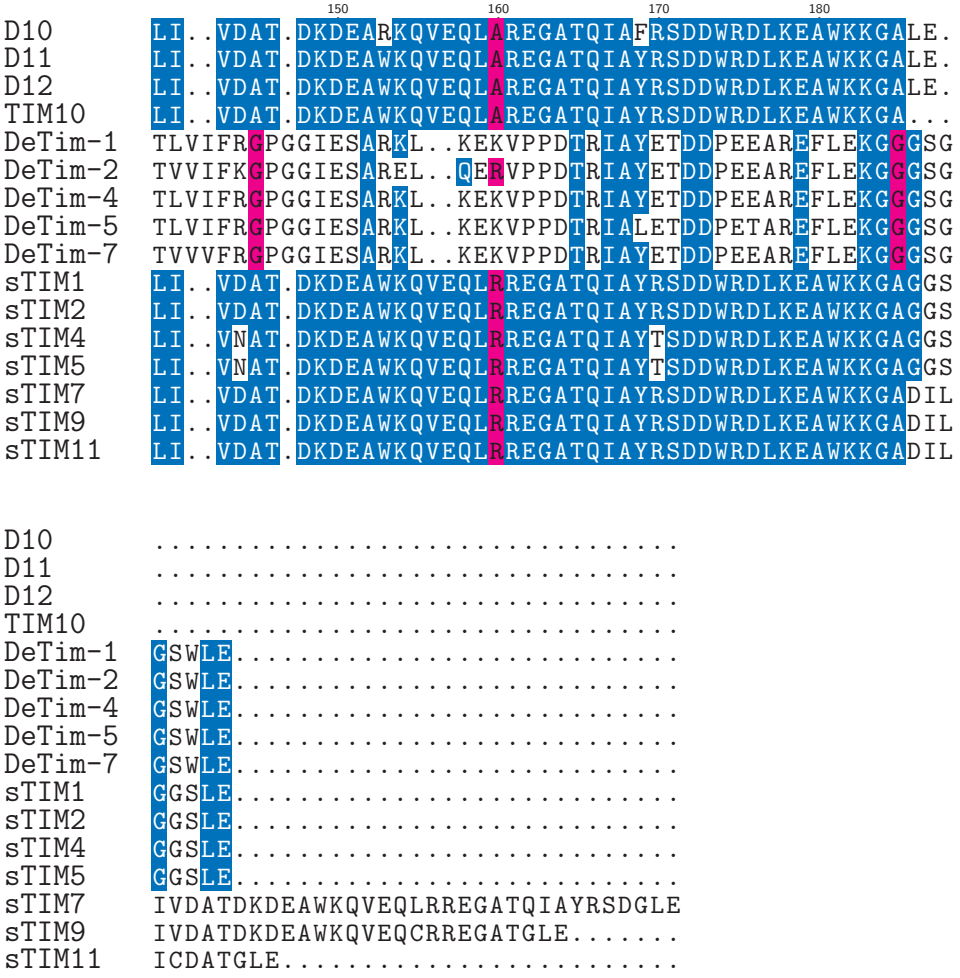


Figure 4.4: Alignment of the different variants designed and experimentally tested by me. The alignment was done with Clustal Omega [51] [52] [53], the representation was done using the TEXshade-package. the colors represent different amounts of sequence conservation.

4.4 Characterisation

Expression and purification

All genes of the different variants were send to us in either pet21 or pet29 plasmids by Possu Huang from David Bakers lab. The plasmids were used for transformations into *E. coli* BL21 cells utilizing the heat shock protocol (see chapter 3.1) and expression took place at 30°C over night according to 3.2. Interestingly the proteins only expressed in TB medium, neither LB nor self inducing media based on ZY medium led to significant amounts of protein.

Purification was done with a Ni Sepharose column (see Figure 4.5), followed by size exclusion chromatography using a preparative S75 column (see Figure 4.6).The fractions containing the construct were pooled and concentrated. For most constructs between 1 mg/l and 5 mg/l were purified.

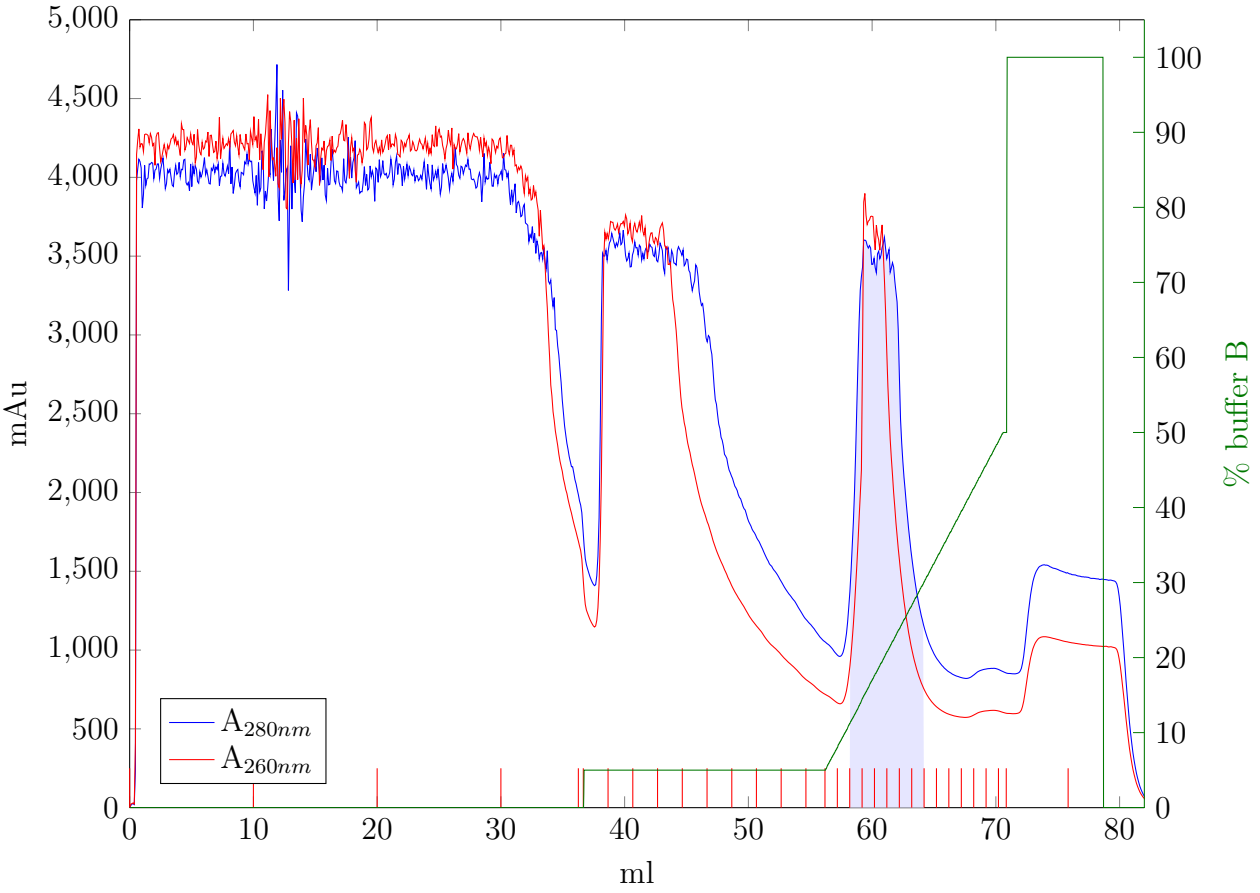


Figure 4.5: Affinity column run of sTIM11 with a 1ml Ni Sepharose column on the AEkta purifier. Elution starts at roughly 150mM Imidazole. Red lines at the x-axis signify fractionation borders from the sample collector. The blue box indicates the fractions ultimately used for further purification. The green line indicates the calculated gradient. Buffer A: 150mM NaCl, 50mM potassium phosphate pH 8.0, 10mM Imidazole. Buffer B: 150mM NaCl, 50mM potassium phosphate pH 8.0, 1M Imidazole.

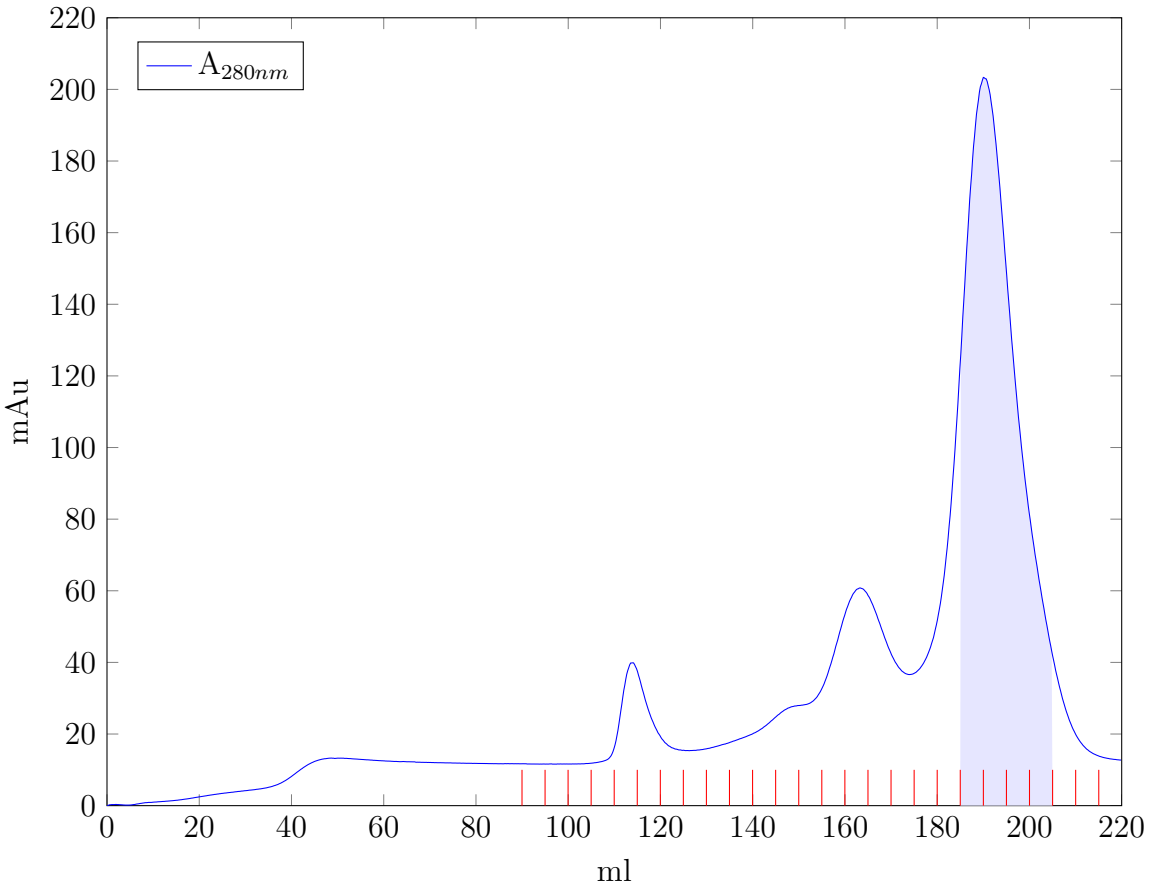


Figure 4.6: Size exclusion chromatography of sTIM11 run with a Superdex S75 column on the AEkta prime for further purification and to remove higher order oligomers. Red lines at the x-axis signify fractionation borders from the sample collector. The blue box indicates the fractions ultimately used further. The running buffer was 150mM NaCl, 50mM potassium phosphate pH 8.0.

Biophysical characterisation

As a measure of secondary structure content, CD spectra were recorded for all purified constructs. If secondary structure was present, tryptophan fluorescence was used to get an estimate of tertiary structure formation. In ambigous cases 2D NMR spectra were recorded.

If both CD and Trp fluorescence were satisfactory, melting curves were recorded measuring changes in CD signal upon heating (see figure 4.8). The resulting curves were used to estimate the melting point of the construct.

The first designs had already flaws at the level of secondary structure as shown by CD (See Figure 4.7). Later designs showed an increasing content of secondary structure.

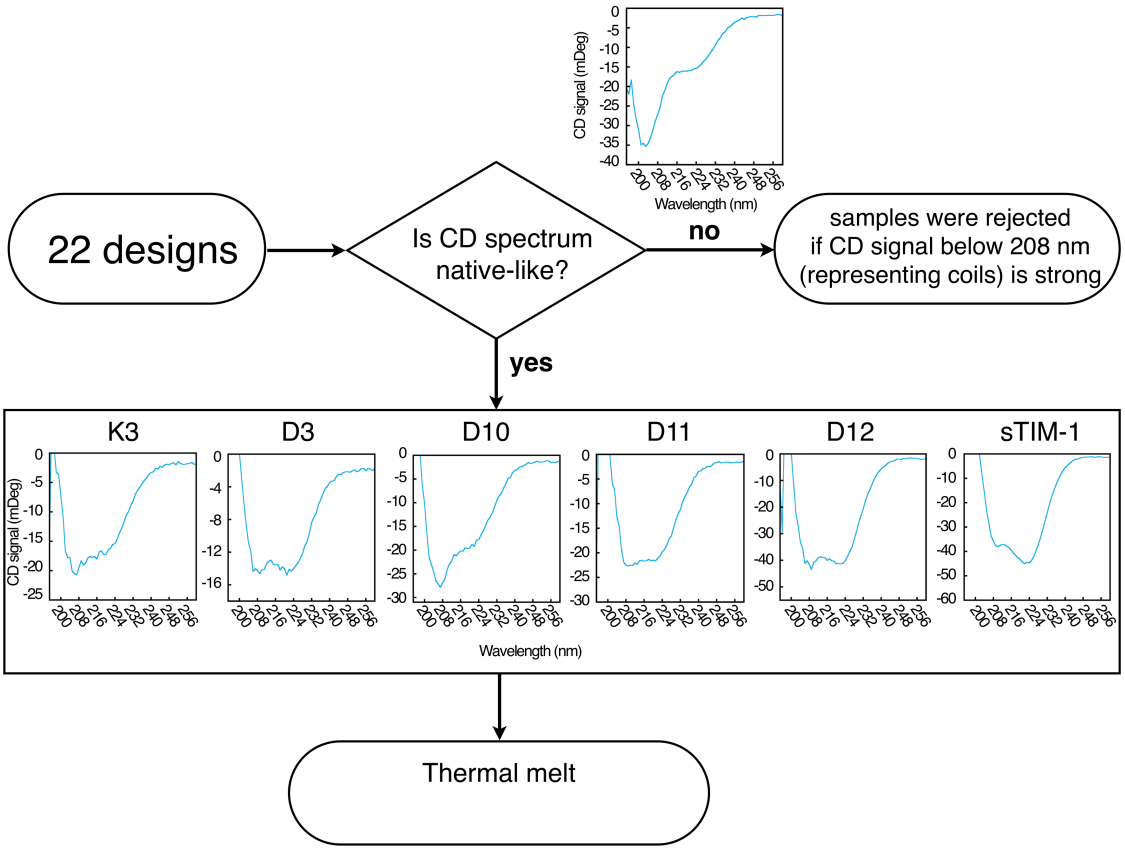


Figure 4.7: Overview of first selection based on CD spectra and examples at different stages of the design process. From [1].

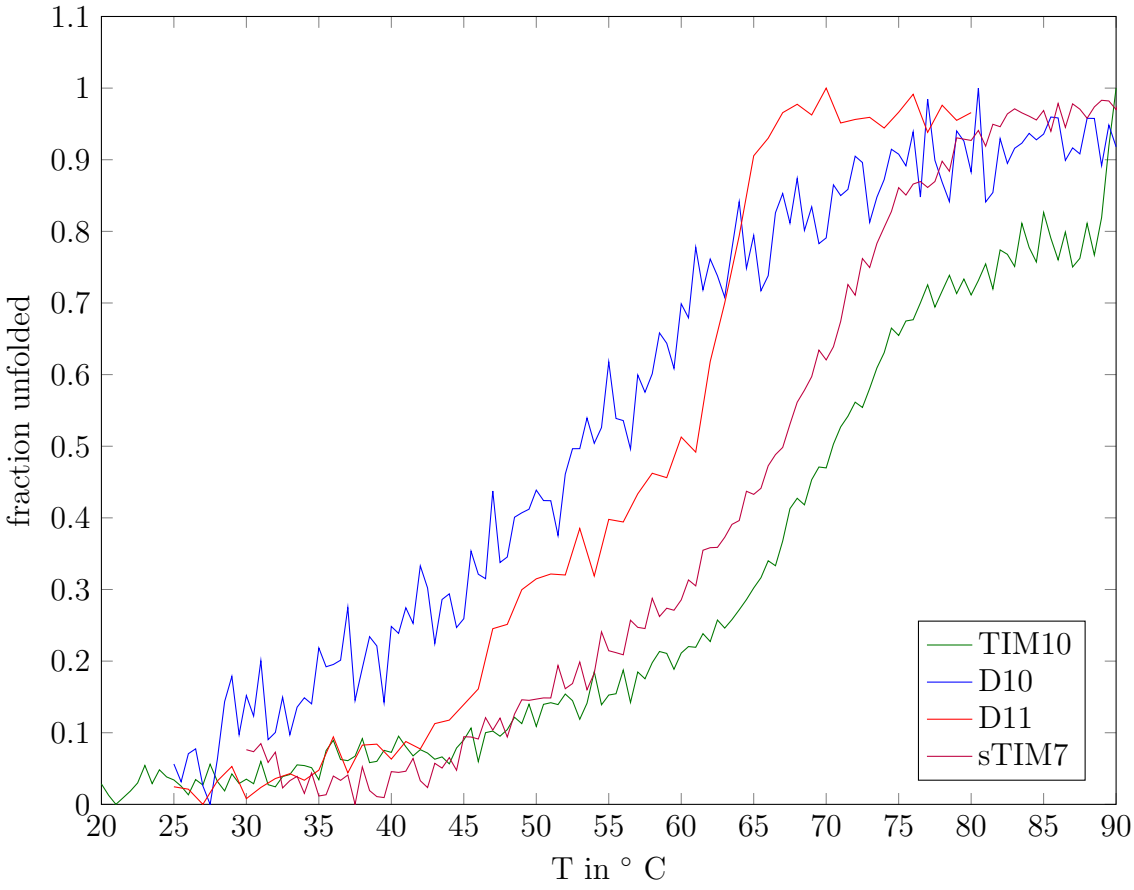


Figure 4.8: Melting curves of different constructs at a concentration of 0.3 mg/ml in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0. The fraction of unfolded protein was determined by the amount of CD signal at 222nm.

Since secondary structure content does not necessarily imply the presence of tertiary structure additional experiments had to be done to ensure that the constructs were indeed folded.

Tryptophan fluorescence is usually used to determine structural integrity. However, many of the constructs incorporate solvent-exposed tryptophans and therefore differences between folded and unfolded states might be small. Without a spectrum of a very similar and reliably folded construct as a comparison it would be difficult to determine whether the fluorescence spectrum belongs to a folded or unfolded protein. I therefore decided in critical cases to use one dimensional NMR to determine the presence of a defined tertiary structure.

In the case of a stable protein with a defined fold, a one dimensional NMR spectrum will show clear methyl peaks under 0 ppm. If - as in our constructs - tryptophans are present, one would also expect visible peaks around 9 ppm. Figure 4.9 shows the comparison of one of the constructs, D10, with a protein which is known to be folded (hRhoA from Silke Wiesner). Both features, the methyl peaks as well as the peaks resembling tryptophans, are present in the hRhoA spectrum and missing in the spectrum of the D10 variant. This leads to the conclusion that D10 is not present in a properly folded state.

Similar 1D NMR spectra were recorded for a variety of the later variants and even in the third generation of constructs none of the candidates showed any sign of a stable secondary structure (see Figure 4.10).

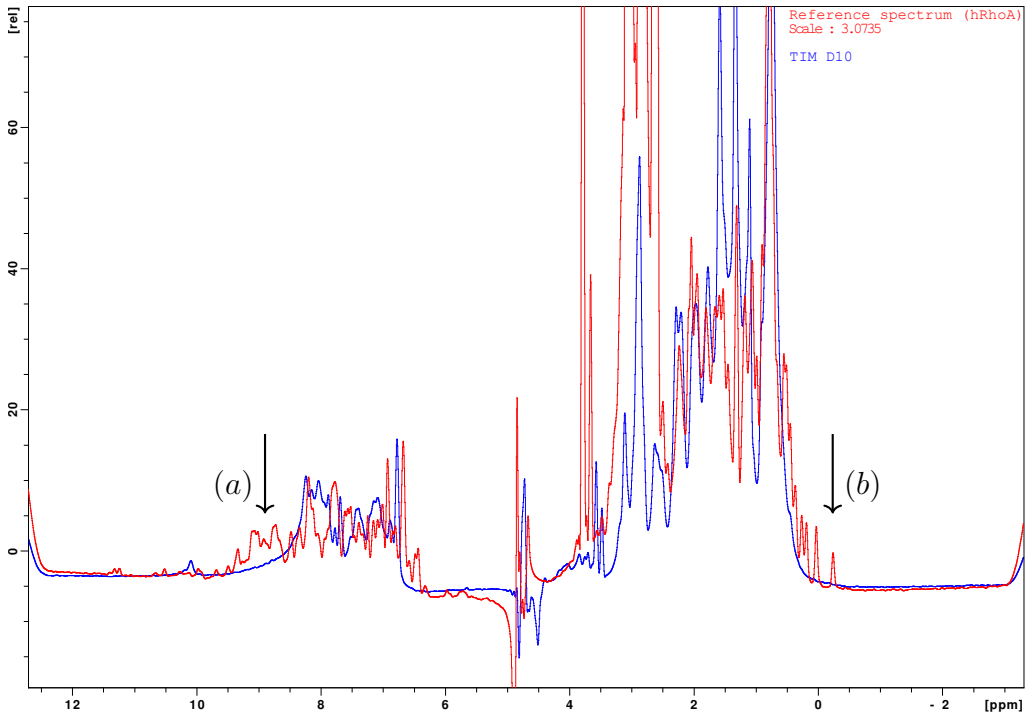
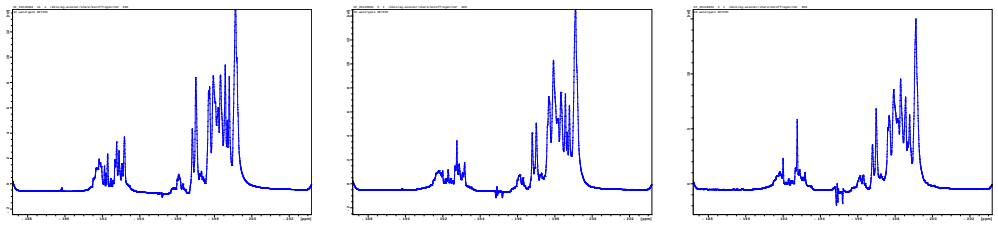
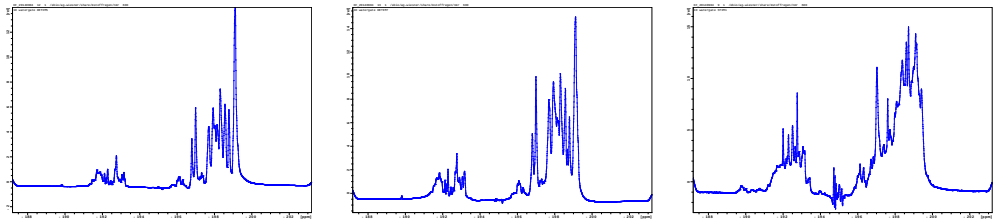


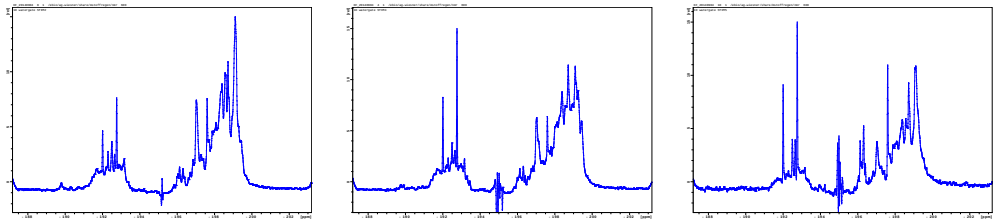
Figure 4.9: 1D-NMR spectrum of the TIM D10 variant (blue) in comparison to the spectrum of a folded protein (hRhoA, purified and recorded by Silke Wiesner). Both the methyl peaks under 0 ppm (b) and the tryptophan peaks around 9 ppm (a) are missing in the D10 spectrum. D10 spectrum recorded in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0



(a) DETIM1 (b) DETIM2 (c) DETIM4



(d) DETIM5 (e) DETIM7 (f) STIM1



(g) STIM2 (h) STIM4 (i) STIM5

Figure 4.10: 1D-NMR spectra of different constructs. The missing methyl peaks that would be expected around 0 ppm and the also expected peaks around 8-10 ppm resulting from the tryptophans present in all of the constructs hint to the fact that the proteins are probably not present in a folded state.

In contrast to most earlier designs, sTIM11 showed a strong CD signal resembling a mixed α-helix and β-sheet content (see Figure 4.11). A melting curve was recorded from sTIM11 and changes in signal were recorded at a wavelength of 222 nm. Unfolding was not completely cooperative and minor changes were already recorded at lower temperature, indicating that probably flexibility between the α helices increases already long before the protein itself unfolds. Thermal denaturation was fully reversible and, the CD signal was fully restored upon cooling to 30°C. Subsequent melting showed identical behavior (see Figure 4.12).

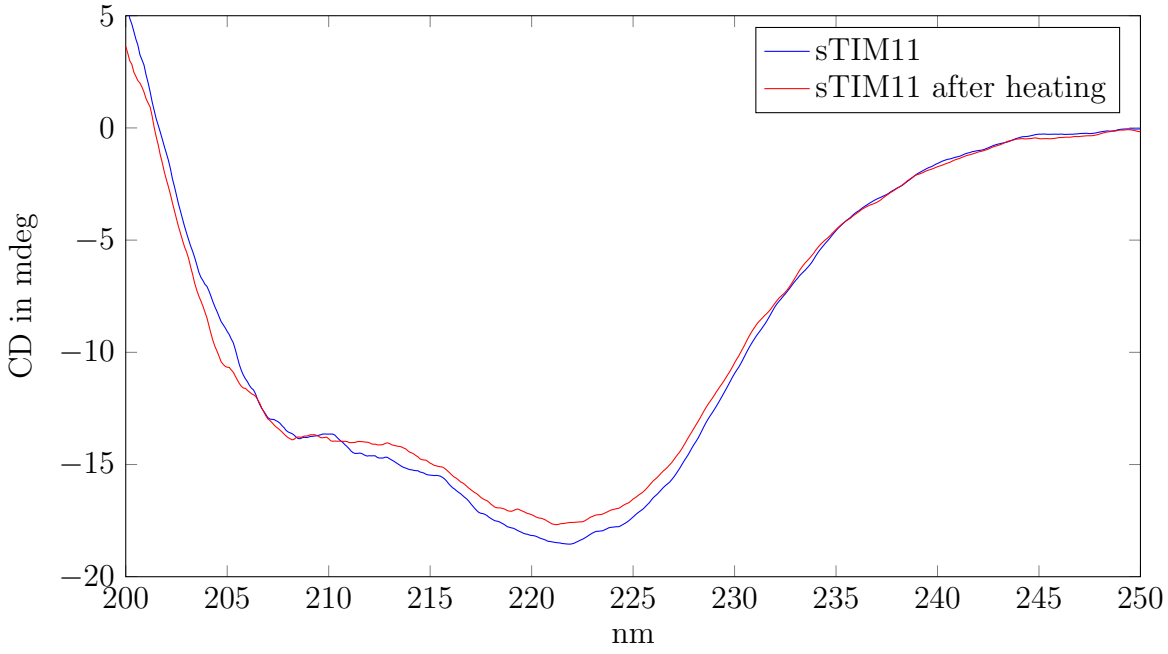


Figure 4.11: CD spectrum of sTIM11 at a concentration of 0.3mg/l in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0. The second sample was heated to 95°C for a melting curve and then cooled down again.

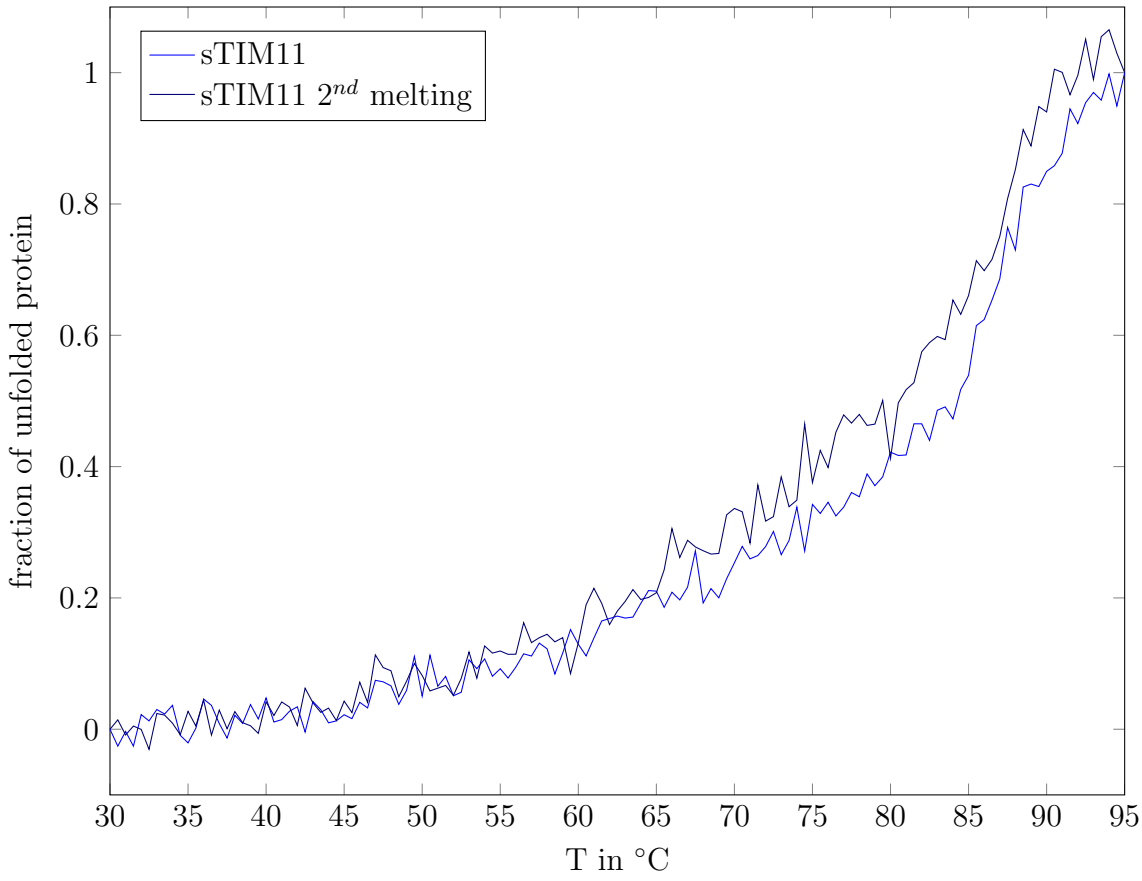


Figure 4.12: sTIM11 unfolds reversibly as shown by repetitive melting. Shown is the amount of unfolded protein recorded using the change in CD signal at 222nm upon heating with a rate of 1° C/min. The protein concentration was 0.3mg/ml in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0.

Since the design features a disulfide bond between the N and the C termini, we were interested in the contribution of the proposed disulfide bond to the stability of the construct. In order to measure this contribution melting curves with continuous measurements of the CD spectrum at 222nm were recorded in the presence of an excess of DTNB. DTNB binds to free cysteines, thus making it impossible to form disulfide bonds. Both the construct with as well as without DTNB present were heated to 95°C, cooled down to room temperature and heated again. No differences were detected between the two setups, already hinting at the fact that the disulfide bonds did not form (see Figure 4.12).

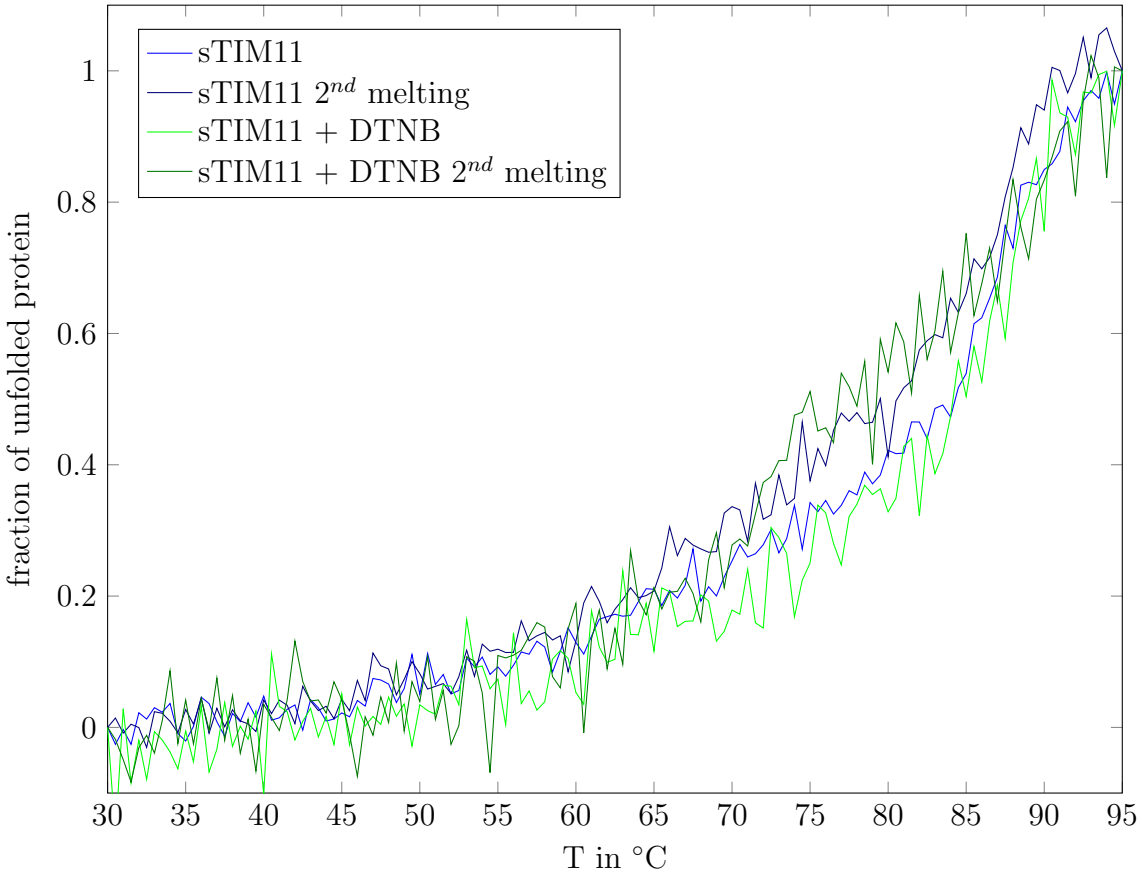


Figure 4.13: The cysteines introduced to form a disulfide bond between the N and the C termini have no effect on stability. This is shown by repetitive melting of sTIM11 in the presence and absence of an excess of DTNB, which seems to have no effect on the thermal stability. Since DTNB binds to free cysteines, thus inhibiting the formation of disulfide bonds, a contribution of the expected disulfide bond to stability should lead to a higher melting temperature in the absence of DTNB. The melting curves were recorded at a concentration of 0.3 mg/ml in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0 at a heating rate of 1°C/min.

In order to compare the stability of the construct with natural TIM-barrels, I decided to determine the change in Gibbs energy upon unfolding utilizing a chemical unfolding study. I used increasing concentrations of guanidinium chloride as a chaotropic agent to destabilize the protein. For each concentration of guanidinium hydrochloride, a tryptophan fluorescence spectrum was recorded using an excitation wavelength of 280nm. Comparing the spectra of folded and unfolded construct, the highest difference was observed at a emission wavelength of 377nm while nearly no difference was visible at 344nm (see Figure 4.14). The emission at 377nm was therefore used to determine the fraction of unfolded protein, while the emission at 344nm was used to normalize for errors resulting from slight concentration differences of the different samples. The fraction of the emission signal at these two wavelengths led to a typical unfolding curve. Compared to unfolding of the secondary structure content measured by CD signal at 222nm, the unfolding begins at a little higher concentration of guanidinium hydrochloride but corresponds to the CD signal at higher concentrations (see Figure 4.15).

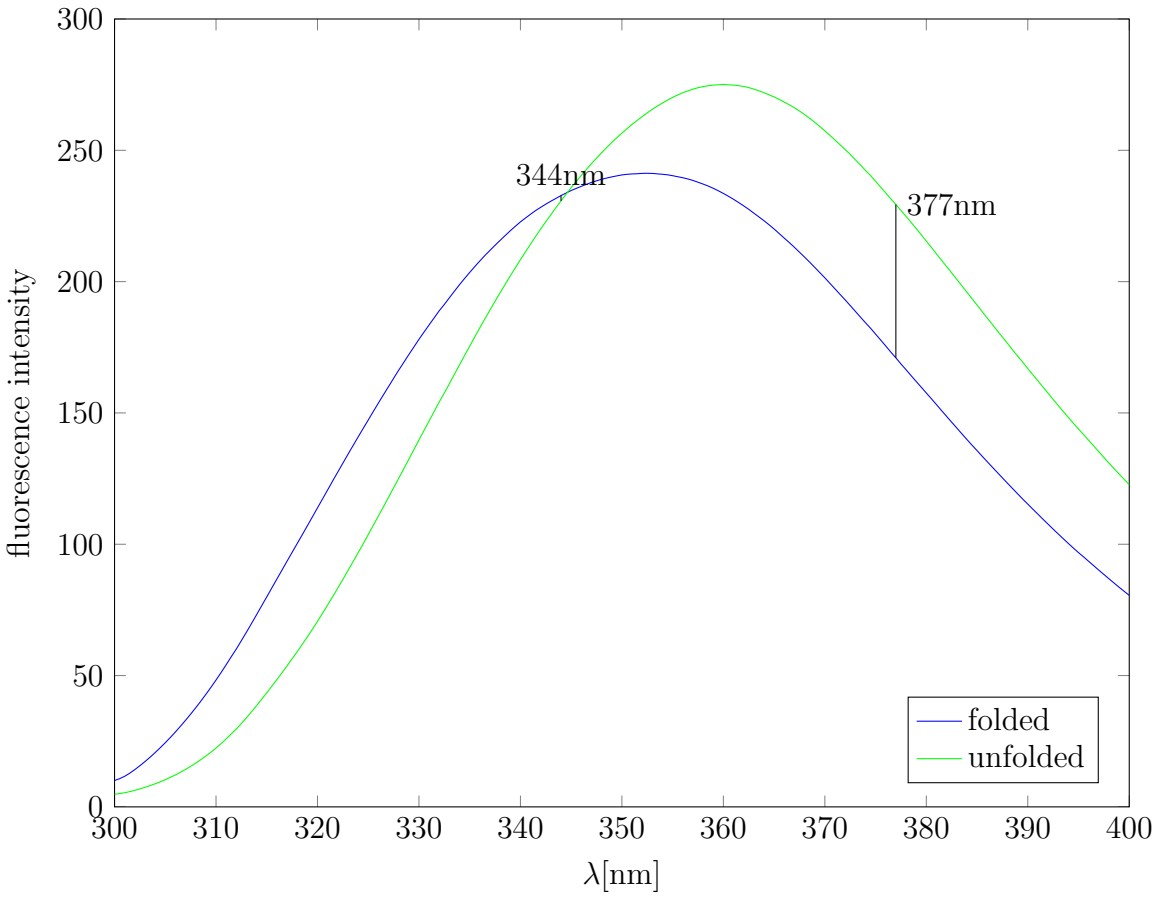


Figure 4.14: Tryptophan fluorescence of folded and unfolded sTIM11 in buffer and buffer + 5M guanidine HCL. Excitation wavelength was 280nm. Also shown are the wavelength of the largest (377 nm) and smallest (344 nm) change as determined by maximum and minimum of the mean deviation from the median of all concentrations measured.

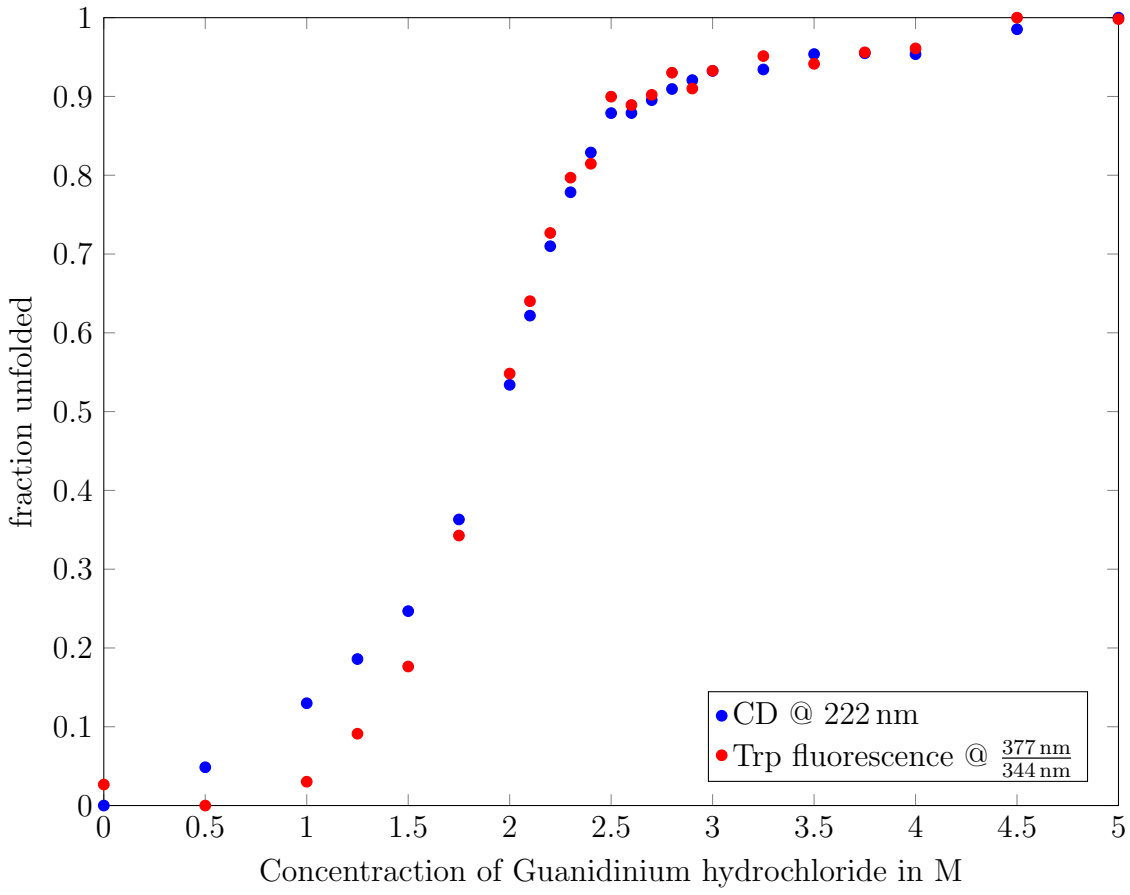


Figure 4.15: Chemical denaturation of sTIM11. The unfolding curves measured with CD and with Trp fluorescence are quite similar, indicating a simultaneous loss of secondary and tertiary structure. The only exception is the early decrease in CD signal at low concentrations of guanidinium hydrochloride. This could be due to unfolding of the probably more flexible first α-helix.

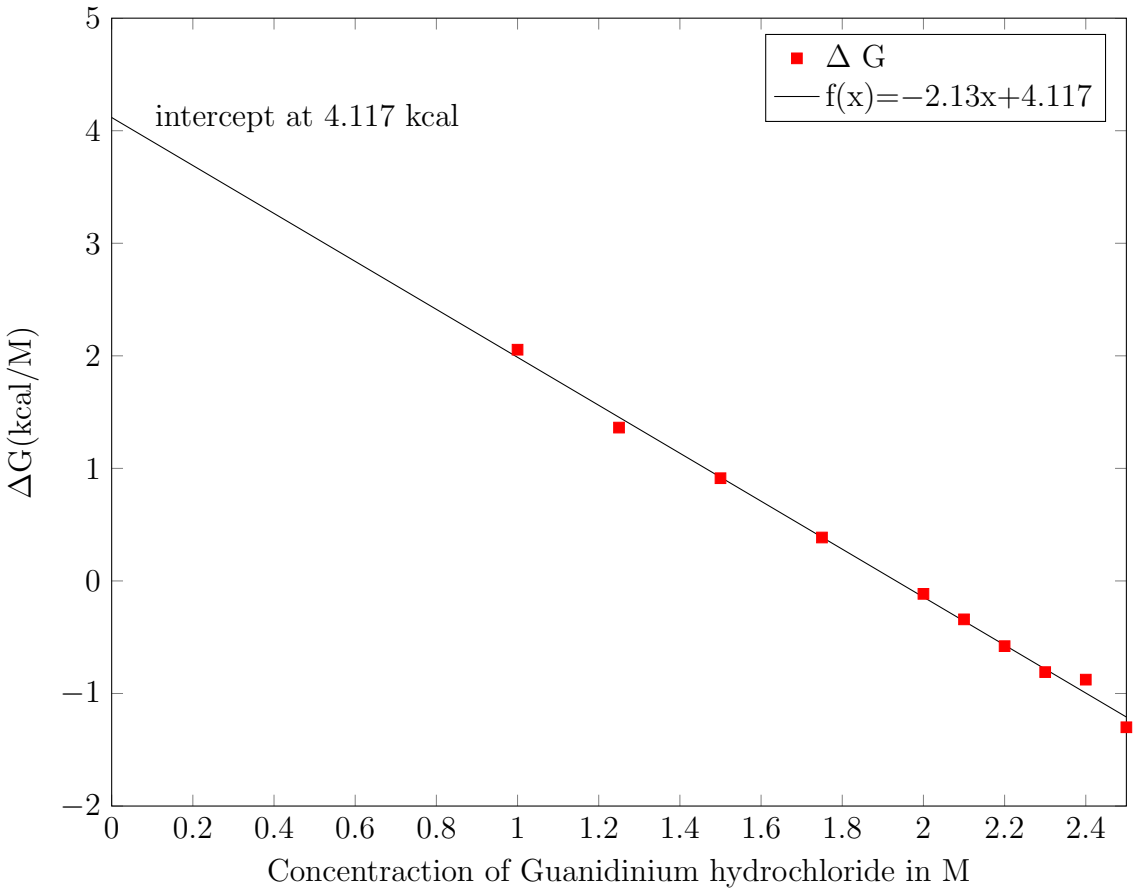


Figure 4.16: Determination of the change in Gibbs energy upon unfolding. ΔG is estimated to be 4.117 ± 0.088 kcal/mol (α=0.05) with no guanidinium hydrochloride present. Basis of the fit are the changes in Trp fluorescence in the concentration range where the change of signal was most rapid.

The unfolding data was used to calculate the change in Gibbs energy as described in Chapter 3.3. Based on the observed unfolding curve, the estimated change in Gibbs energy upon unfolding in a solution free of guanidinium hydrochloride was estimated to be 4.117 ± 0.088 kcal/M. This is less than the majority of the intensively studied (thermostable) natural TIM-barrels. For a designed protein however this represents a very good value.

X-ray structure determination

Crystallization screens were setup for the constructs TIM10, sTIM7, sTIM9 and sTIM11. These were quite willing to crystallize, however the crystals either did not show any visible diffraction or the diffraction observed was so low that no structure could be solved (worse than 5 Å). For sTIM11, however, two datasets could be recorded with the better one diffracting up to a resolution of 2 Å.

The diffraction data was processed with XDS and the space group was determined to be P 41 21 2 (92). Notably the Wilson B factor was quite high for a structure with this resolution (see table 4.1). This is already a hint that there is probably either a high flexibility within the protein or errors in the crystal lattice. In a fourfold symmetric barrel such an error could for example be an alternating orientation of the termini between different copies.

Molecular replacement was done using the backbone atoms (C, CA, O and N in the pdb nomenclature) of a Rosetta model of sTIM11.

recording parameters	
Wavelength (Å)	1
Resolution range (Å)	46.79- 1.992 (2.064- 1.992)
Space group	P 41 21 2
Unit cell	
a, b, c (Å)	50.08, 50.08, 131.28
α, β, γ (°)	90, 90, 90
Total reflections	69721 (5756)
Unique reflections	11894 (1083)
Multiplicity	5.9 (5.3)
Completeness (%)	98.26 (93.20)
Mean I/sigma(I)	15.24 (1.79)
Wilson B-factor	41.84
R-merge	0.05804 (0.6607)
R-meas	0.06359
CC1/2	0.999 (0.762)
CC*	1 (0.93)
Reflections used for R-free	595 / 5%
R-work	0.2237 (0.2882)
R-free	0.2607 (0.2989)
Number of non-hydrogen atoms	1481
macromolecules	1461
water	20
Protein residues	180
RMS(bonds)	0.013
RMS(angles)	1.21
Ramachandran favored (%)	96
Ramachandran outliers (%)	0
Clashscore	3.17
Average B-factor	61.90

Table 4.1: Parameters of the crystal structure 5BVL of sTIM11

4.5 Evaluation of sTIM11

Evaluation of the crystal structure

The crystal structure obtained shows an overwhelming similarity to the designed model. The structure superimposed to the design with a Cα-RMSD of 1.28Åand even most of the sidechains were in conformations similar or identical to the calculated model. Many of the features designed manually could be found in at least some of the quarter barrels (see figure 4.3). There were however also some differences between the model and the actual crystal structure. Most notably, the disulfide bond intended to connect the N- and C-termini did not form (see figure 4.17). This was probably due to the usage of non-optimal rotamers for the cysteines in the design. Tryptophan 42 intended to control the distance between the helices differed quite a bit from the design, however the other tryptophans in the quarter barrel introduced for the same purpose perfectly matched their predicted position (see figure 4.3d). This deviation might be due to the fact that the hydrogen bond designed between threonine 26 and tryptophan 42 is not formed directly but mediated by a water molecule (see figure 4.3e).

Relations to other TIM barrels

As interesting as a de novo design might be by itself, it is equally fascinating to compare the designed construct on a sequence level to natural proteins utilizing the same fold. This is important due to two reasons. On the one hand I had to make sure that the design was not biased too much by existing proteins. This is especially true for algorithms such as Rosetta utilizing natural fragments in the design process. On the other hand it can also give us an insight into the sequence space itself and how much of it is actually covered by known natural proteins.

In contrast to structure-based search procedures, detecting similarities based on

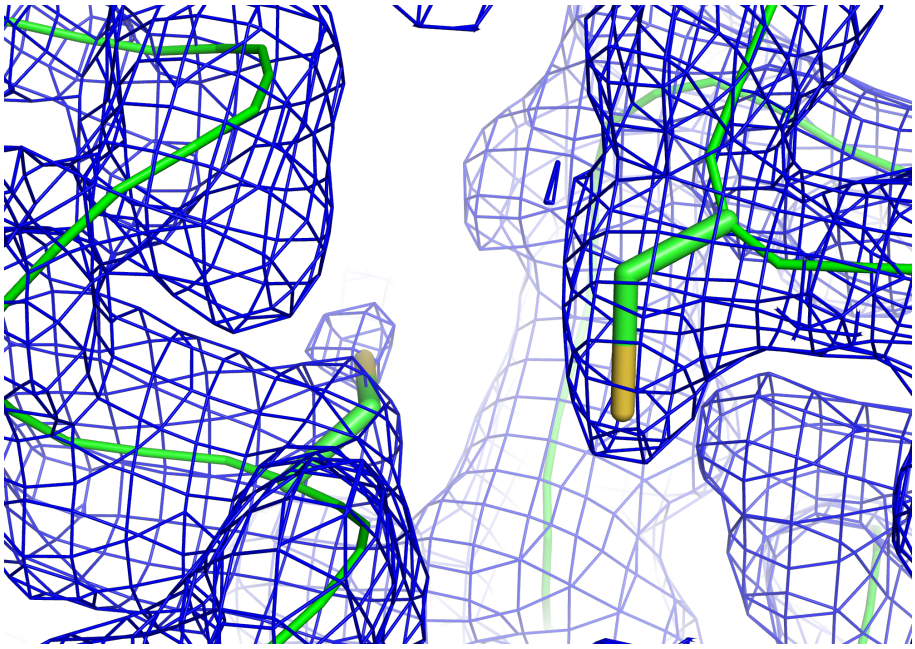


Figure 4.17: Closeup on the designed disulfide bridge intended to connect the N and the C termini. The disulfide bridge can not form with the geometry observed.

the sequences was more difficult. Standard psi-blast searches using three iterations did not detect any similarity to other TIM-barrels. More sensitive profile-based searches with HHsearch [54] however were able to find some commonalities with low probabilities (the best p-value to another member of the TIM-barrel family was 8.4E-06 with a quarter and 1.8E-08 with the full-length protein, both against the SCOPe95 2.06, no secondary structure scoring, 3 iterations). I therefore created a database with the sequences of all TIM-barrels in the SCOPe as well as sTIM11. Similarities between all pairs of sequences were calculated and used for clustering.

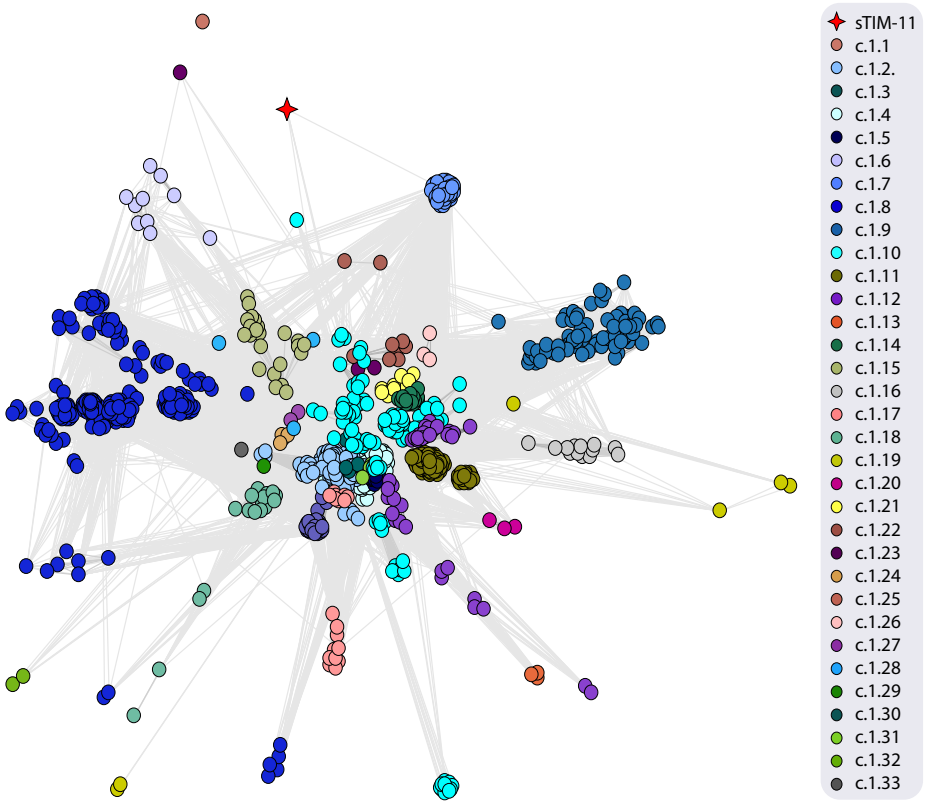


Figure 4.18: Cluster of representatives of all different TIM-barrel families present in the SCOPe. The sequence similarity scores were calculated with HHblits. sTIM11 is represented as a red star. Clustering was done with CLANS [31]

The natural TIM-barrels are highly connected, not only within the families but also between different families and even superfamilies (see Figure 4.18). In fact borders between different superfamilies are often not clearly defined and the existing SCOP classification is difficult to justify just based on the similarity scores observed. This is especially true for many of the superfamilies of c.1.1-c.1.10.

If we now compare these highly interconnected sequences to the sequence of sTIM11, the latter shows only a few, quite improbable similarities to natural TIM barrels (compare Figure 4.18). Although there are sequences that are even more isolated (the only member of superfamily 25, Monomethylamine methyltransferase from *Methanosarcina barkeri* does not show a single significant sequence similarity), sTIM11 seems not only to have a quite unique sequence but also a much lower similarity to other TIM-barrels than most other members of this superfamily would have.

4.6 Discussion

The design goal, a *de novo* TIM-barrel, was met as shown by the crystal structure. Moreover, the resulting protein is remarkably stable for a designed protein and shows several beneficial attributes such as the capability to refold completely. Independent of the fact that this protein did not evolve naturally it is quite interesting. It features an extremely minimalistic structure which probably puts it amongst the smallest possible TIM-barrels. It also incorporates an almost perfect fourfold symmetry, giving rise to the possibility that the ancestral TIM-barrels have evolved from quarter barrels. As expected for a true *de novo* design, the sequence used shows no close relations to any natural TIM-barrel, in fact only the most sophisticated profile-based search algorithms are able to find any similarities with quite low significance.

This design expands our landscape of man-made proteins on quite some frontiers. It is so far the biggest *de novo* designed globular protein, its stability without any experimental optimization is remarkable and it is the first successful design of the most important fold for enzymes. Its stability, symmetry on both the sequence as well as the geometric side and the fact that the protein was thus far only optimized for its geometry and does not bare any evolutionary deadweight should help develop it into enzymes with very different demands.

The design was published in Nature chemical biology [1] and is already subject to further research on both its structural as well as functional potential. Gregor Wiese from the University of Tübingen started working on designs based on sTIM11 and will continue to do so in his Master Thesis.

Chapter 5

Rational design of a new protease

5.1 Overview

There are two main motivations for the design of new enzymes. On the one hand there are a number of beneficial reactions that no natural enzymes are known to catalyze. To be able to design enzymes for these cases would be beneficial for a variety of applications. On the other hand enzyme design can also be seen as the ultimate proof of our understanding of a specific reaction mechanism.

Independent of the motivation for the design of new enzymes, at the current state the successful designs catalyze energetically easy reactions. Many interesting reactions for both application as well as investigation of the mechanism however are quite challenging, often needing several catalytic steps to complete.

In order to extend the field of enzyme design towards more difficult reactions, this part deals with the design of a protease. The ability to design new proteases would on the one hand open a variety of applications. The catalytic mechanism I intend to use, the catalytic serine triad, is on the other hand one of the best-studied enzymatic mechanisms and being able to design a functional triad would ensure us that our understanding of its mechanism is quite thorough.

The design strategy used is based on the insertion of a catalytic serine triad into an existing host scaffold. This strategy has been successfully used for enzyme designs before [15] [17] [20] [19], however it imposes some additional problems for the insertion of the catalytic triad. Since the geometric definition of the catalytic triad only allows for a very small conformational space, the probability to find an insertion site in a given number of possible host scaffolds is much lower than for the more loosely defined geometries of most prior designs. It is therefore necessary to drastically increase the number of screened host scaffolds and while most prior studies used Rosetta Match to screen databases of around 200 structures [21] this work aimed for a system able to screen the whole protein database.

In order to screen large databases effectively, certain simplifications have to be performed in the early stage of the design. Since energy calculations are very resource intensive and also of limited use before the actual pocket design takes place, only geometric considerations are used to initially find possible insertion sites. After the amount of insertion sites is reduced to a manageable number, more complex calculations can be made. More specifically I used consecutive runs of Rosetta relax and Roseta enzyme design to figure out which insertion sites can be made into a functional enzyme (see later chapters for a more detailed description). These reduce the amount of possible designs further until in the end only a small number of designs has to be evaluated experimentally. Figure 5.1 gives an overview of the amount of data dealt with at different stages of the design procedure and the tools involved.

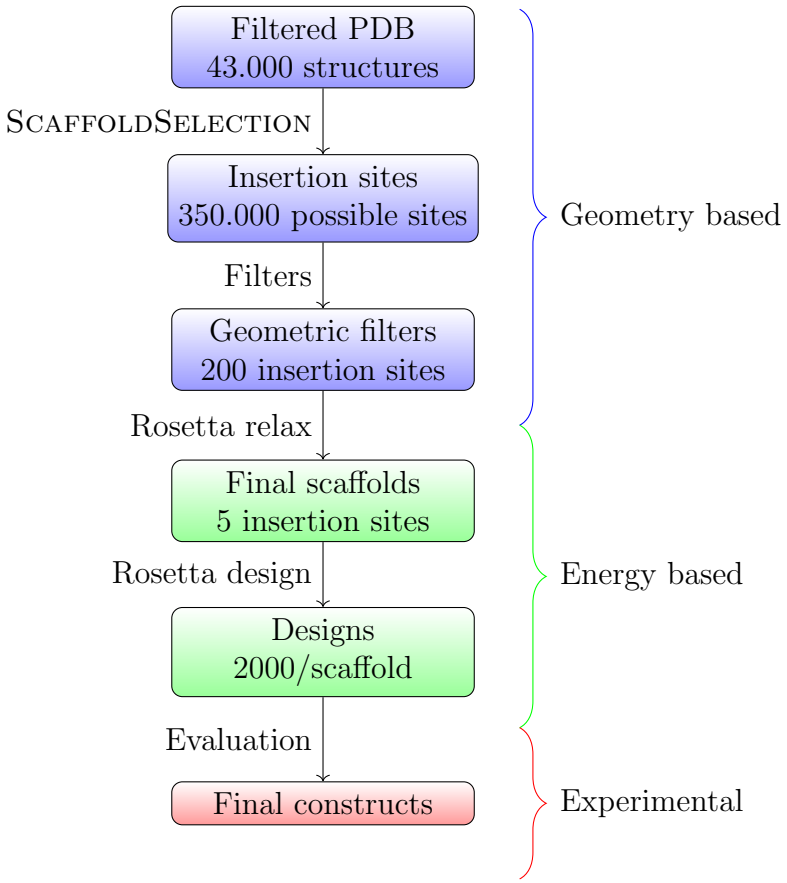


Figure 5.1: Overview of the design strategy and usage of the different tools. In blue are tools making only use of the geometry without considering energy. These parts are based on ScaffoldSelection and the result of ScaffoldSelection runs. Green are calculations that use energy terms. These are based on Rosetta relax (for a first overview of the energetic feasibility of an insertion site) as well as Rosetta Enzyme design (for optimization of the binding pocket). Red is actual experimental work.

5.2 ScaffoldSelection

In order to screen large quantities of possible host scaffolds a program called ScaffoldSelection was developed in our group several years ago by Christoph Malisi in collaboration with Oliver Kohlbacher [24] [55]. ScaffoldSelection uses a geometric definition of a desired motif and searches a database of structures for possible insertion sites. Such a motif usually consists of the amino acids that have to be placed, a placeholder for a ligand or substrate that has to be bound in a geometry relative to the introduced sidechains and additional conditions on the insertion sites such as the presence of specific atoms at certain positions. The set of all of these elements and the desired geometry they should appear in is called a theozyme [56]. ScaffoldSelection creates an inverse rotamer tree of said amino acids that is superimposed on their functional atoms. Every final node of the rotamer tree represents a possible insertion site into the protein backbone. This method is similar to other programs developed for the same task (like Rosetta Match), however is much faster, allowing the screening of the whole pdb within hours on a small cluster.

In order to reduce the computational effort, ScaffoldSelection first searches for cavities on the protein surface that might support a binding site. ScaffoldSelection then tries to fit the members of the inverse rotamer trees the representation of the ligand or substrate into these pockets. Clashes of the protein backbone with both the placed amino acids as well as the ligand are evaluated. In contrast clashes of the placed motif with sidechains of the host scaffold are neglected since ScaffoldSelection assumes that sequence optimization is taken care of in later optimization steps. If a requirement for the presence of certain atoms at specific positions was set, ScaffoldSelection will filter all insertion sites that do not meet these criteria in an additional step.

Next to the position of the catalytic triad's active atoms, the existence of a nitrogen atom at a position where it could function as a oxyanion hole was also set as a requirement. Since the position of the oxyanion hole can be quite variable in natural proteases, I started several runs of ScaffoldSelection in parallel, each with a different required oxyanion hole position. The results of these different ScaffoldSelection runs were then united and evaluated together.

The evaluation itself was based on four different criteria:

- **Backbone clash score,** which evaluates clashes of the inserted amino acids with the backbone of the host scaffold
- **Catalytic geometry penalty,** which evaluates differences between the inserted amino acids and the template
- **Rotamer frequency score,** which evaluates how common the rotamers of the inserted amino acids are
- **Substrate clash score,** which evaluates clashes of the substrate with the backbone of the host scaffold

One of the problems when working with ScaffoldSelection is that a single insertion site is evaluated independently for all of these criteria. If it fails to meet the requirements for one of these scores, the site will still be present in all of the other lists. Therefore, if ScaffoldSelection is allowed to select for the highest scoring insertion sites, this will typically lead to a very low amount of complete hits where all scores are present. This is especially problematic for the Backbone clash score, which is in most cases zero. If ScaffoldSelection filters a certain amount of insertion sites as implemented, this leads to a random selection of insertion sites that score good in said score.

This was the reason I disabled initial sorting by ScaffoldSelection and instead used all found insertion sites for further analysis. For this purpose scripts were

written that collect and allocate the scores for all insertion sites. Based on these collated quadruples of scores selections of credible positions were made.

The definition of a credible position according to the ScaffoldSelection scores is a position that scores similar to a natural protease. Since ScaffoldSelection gives not one score but four, I started by optimizing possible sets of weights and filters for these scores to identify the best scaffolds for the intended design.

The goal of these optimizations was to find a set of filters (thresholds for each score type that cannot be undercut) and weights (a linear weight multiplied with the score in order to give it more or less relevance) that perform well in the sense that natural proteases score very good. I therefore performed ScaffoldSelection searches on a set of pdb structures that not only contained possible host scaffolds but also structures that reportedly already contain the catalytic triad as a test set. Insertion sites in the test set were identified by my scripts and I evaluated how they scored under different evaluation criteria.

The goal was to find a set of parameters that enriches the natural proteases in the top scoring insertion sites. Due to the different spread of scores it was however not possible to find weights that would lead to the desired result. I therefore derived new scores by calculating the rank of each score compared to all other scores. Filters were still used on the basis of the raw scores, but weights were applied to the ranks instead. Especially strict filters were applied to the Backbone clash score. This was due to the fact that most insertion sites had a Backbone clash score of zero resulting in a rank of one. These ranks were therefore quite meaningless and hence were not used for further evaluation. Using different sets of weights or in some cases only the Catalytic geometry penalty ranks lead to the desired enrichment of natural proteases in the top percentiles (see Figure 5.2).

Insertion sites that scored in the area dominated by positions found in natu-

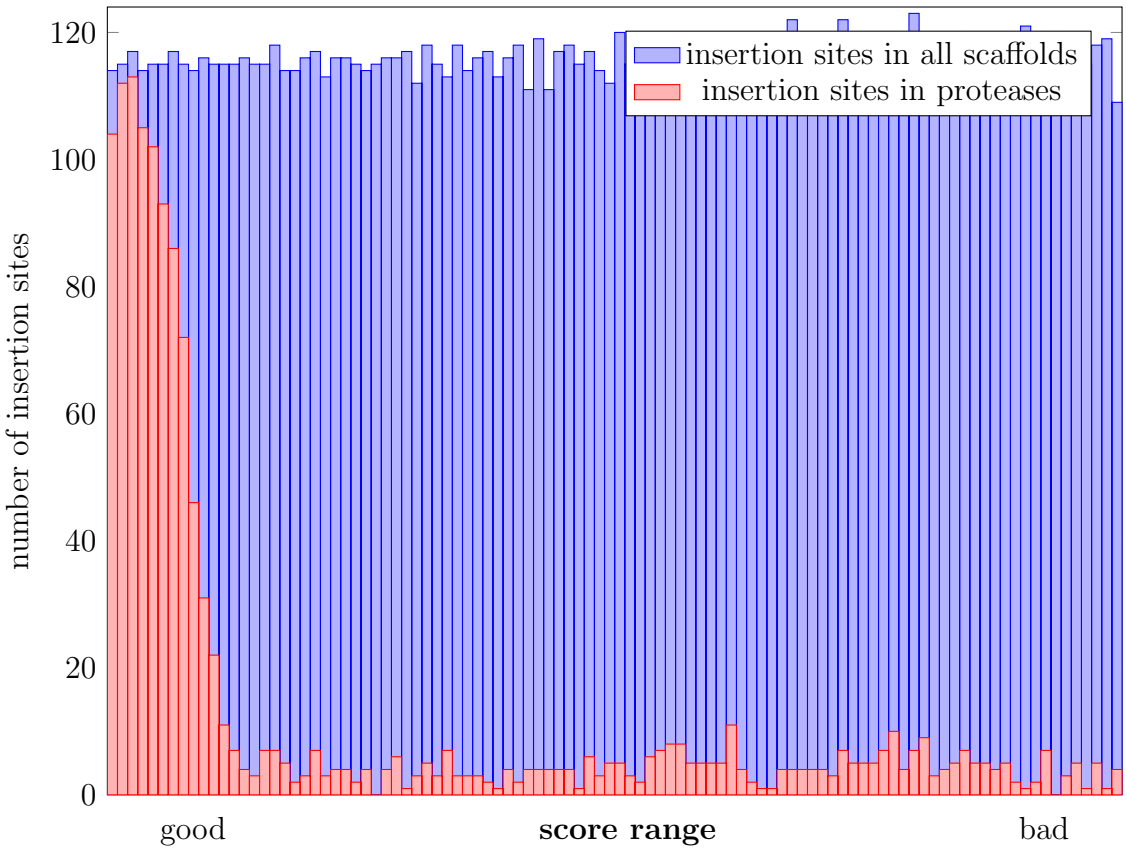


Figure 5.2: Graphic representation of the best 4500 insertion sites found in one ScaffoldSelection run. The red bars indicate the amount of positions found in the test cases consisting of structures of roughly 2500 serine proteases. The blue bars indicate positions found in the 43.000 PDB structures searched for potential insertion sites.

ral proteases were extracted. These have a geometric feasibility similar to already existing triads and were therefore classified as "geometrically credible". However, further steps were needed to ensure that they were also energetically probable.

5.3 Preliminary energetic evaluation

Before starting a complete pocket optimization with the introduction of mutations in the proposed binding pocket, the general energetic feasibility of the insertion was evaluated utilizing the Rosetta relax application.

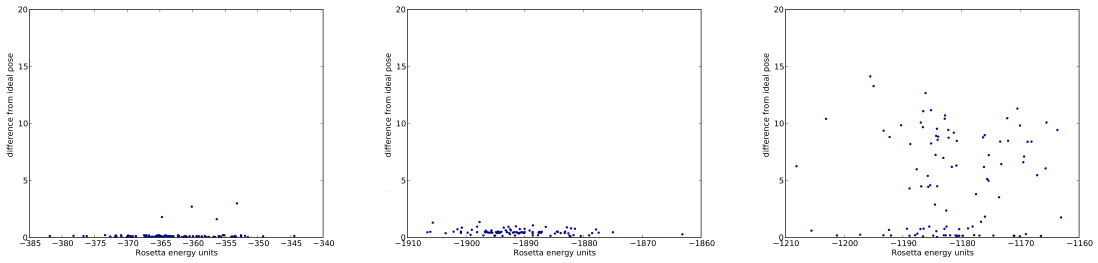
For this purpose, each insertion site was placed with the motif construction tool from ScaffoldSelection and prepared for Rosetta runs. Around 100 unconstrained relaxes were calculated for each model. The resulting relaxed files were then loaded and differences to the theozyme were calculated. In order to reduce these differences to a single number, I calculated the spacial deviation of the distances between the atom OG from serine and the atom NE2 from histidine (d1) and the atom ND1 from histidine and the atom CG from aspartate (d2) and compared them to the theozyme distances (D1 and D2 respectively):

$d1 := ||\mathrm{OG}_{Ser} - \mathrm{NE2}_{His}||_2$

$d2 := ||\mathrm{ND1}_{His} - \mathrm{CG}_{Asp}||_2$

$$K = \sqrt{(d1 - D1)^2 + (d2 - D2)^2}$$

These plots were used to quickly compare the integrity of the catalytic triad after the relaxes. If large deviations were observed in the majority of cases the insertion site was rejected (see figure 5.3).



(a) 1CHG (b) 2D52 (c) 1W2C

Figure 5.3: Evaluation of the result of a ROSETTA relax experiment. Plotted is the deviation of the distances of the members of the catalytic triad relative to each other from the distances in the theozyme. (a) shows chymotrypsin as a positive control. The catalytic triad maintains its geometry in most relaxes. (b) shows an example of a good candidate. Again, the designed triad maintains its geometry. (c) is an example for a candidate failing this test. In many relaxes the catalytic triad deviates from the design goal as indicated by the large amount of structures with a high distance difference.

The remaining insertion sites (around 20) were evaluated manually. Several candidates were rejected due to criteria such as insertion of the triad into a hinge region or geometries that were only observed in proteins with bound ligands. Those that seemed viable were optimized using the program ROSETTA Enzyme design.

5.4 Pocket optimization

In order to not only optimize for the formation of the catalytic triad but also for substrate binding, I included ApNA as the desired substrate into the calculations. I decided to optimize for the state of the first tetrahedral intermediate. The formation of this state is arguably the most difficult part of the reaction although also later stages have been reported to be troublesome to design [22]. This, however, requires a model of ApNA in the first tetrahedral intermediate state.

First, I tried to define a noncanonical amino acid resembling serine and the bound ApNA, intending to exchange the active serine with this amino acid and optimizing the sequence for this construct. Unfortunately, Rosetta 3.3 builds the rotamers based on rotamer trees and seems to be incapable of handling a branching tree, which is required to create all possible rotamers of the serine-ApNA construct (see Figure 5.4). Therefore, I decided to include a library of all possible rotamers created

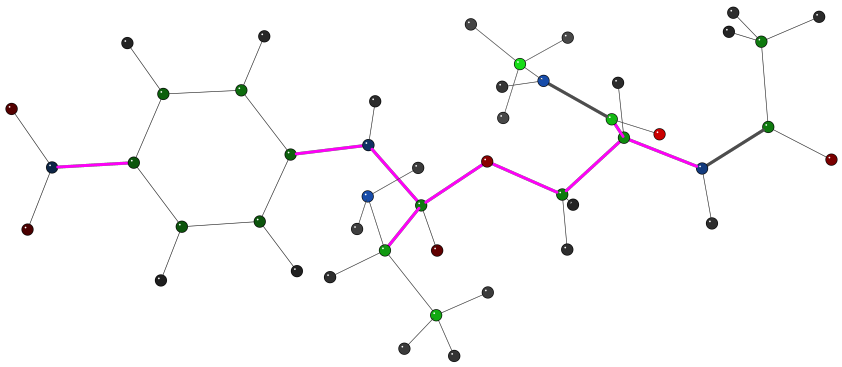


Figure 5.4: Model of Serine with ApNA covalently bound. Pink bonds indicate the edges of the rotamer tree. Note the branching of the rotamer tree at the central carbon atom.

with FROG2 [30] of ApNA and constrain it to the serine to a position mimicking the tetrahedral intermediate. This was either achieved by the introduction of a virtual

atom (which is an atom with a radius of zero and no van-der Waals interactions with other atoms) bound to the carbon that would bind to serine. This virtual atom was then subsequently forced to superimpose with the serine oxygen during the design. Alternatively, the distances and angles of the cleavage site were constrained relative to the oxygen of serine. Additional constraints were set up between the functional atoms of serine and histidine on the one hand and histidine and aspartate on the other hand to ensure that the catalytic triad would stay intact.

Depending on the host scaffold, either all positions except for the triad were allowed to mutate, only positions in the supposed binding pocket, or a combination of both where first all positions were allowed to mutate and the results were used to identify important positions, which were in a second run exclusively allowed to change. Up to several thousand runs were started with different parameters. One of the initial problems was that Rosetta introduced many mutations, most of them at positions distant to the insertion site. In order to control this behavior the FAVORNAT option of ROSETTA was used. This option introduces a specific penalty for each introduced mutation and thereby forces ROSETTA to reject introduced mutations that result in an energy improvement worse than the specified threshold. The exact penalty required to keep the amount of mutations in a reasonable range differs from scaffold to scaffold and was determined individually for each host structure. By combining design runs with different FAVORNAT penalties I was able to produce sets of designs with a wide range of numbers of mutations (see figure 5.5). These were combined and evaluated together for each scaffold.

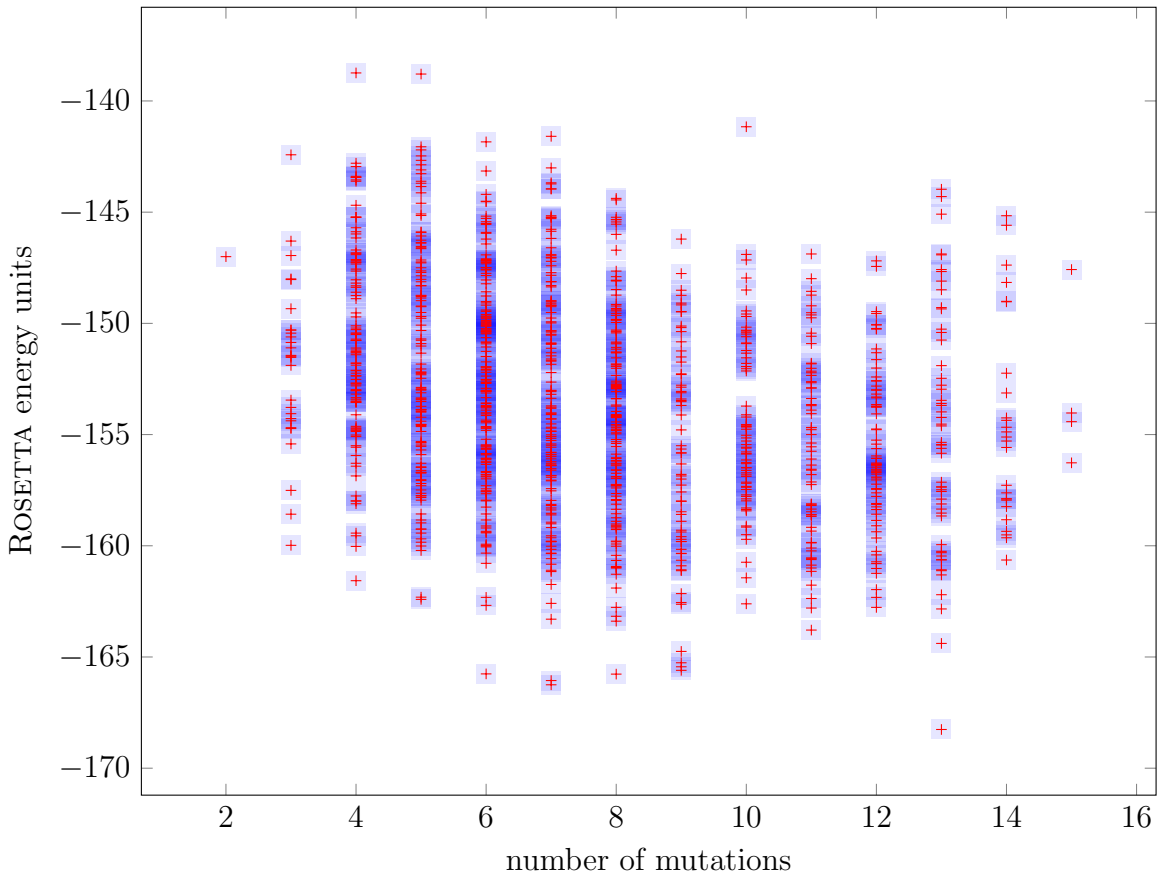


Figure 5.5: Collection of different design runs of one of the proposed insertion sites (3N8M, insertion into residue 51 (Serine), residue 17 (histidine) and residue 13 (aspartate)). Each cross represents one design with a specific set of mutations. As expected a larger number of mutations allows Rosetta to find more optimal solutions resulting in a lower total energy. In this specific case a low amount of mutations (around six) is enough to reach an energy plateau of ≈ -165 REU (Rosetta Energy Units) that is maintained even if more mutations are allowed.

All designs for a certain insertion site were combined and used to derive parameters such as the frequency of certain mutations or the probability to be cointroduced with specific other mutations. Residues that were mutated in the majority of cases were identified as especially problematic and paid special attention to when considering final designs (see Figure 5.6 for an example).

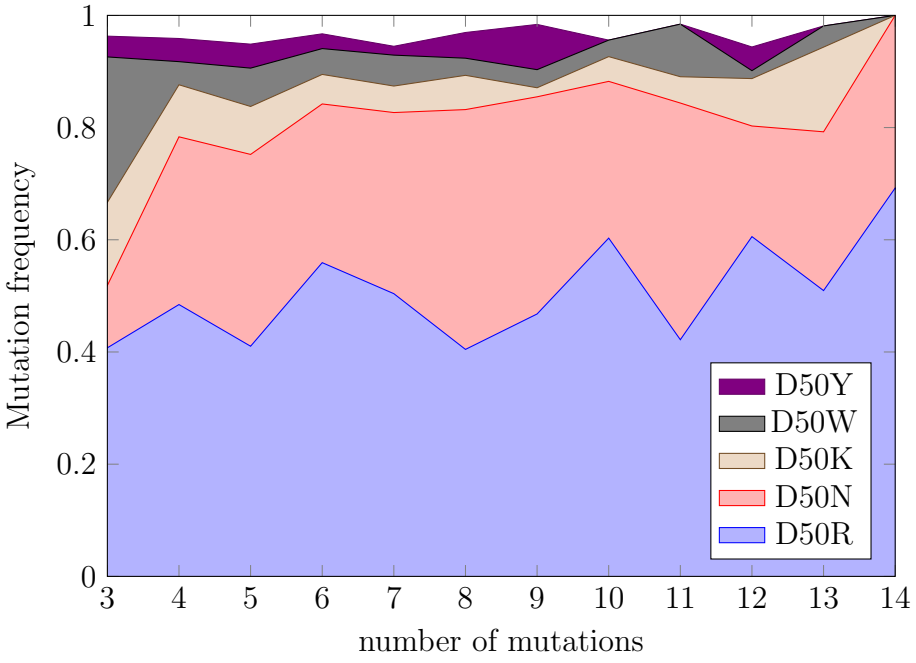
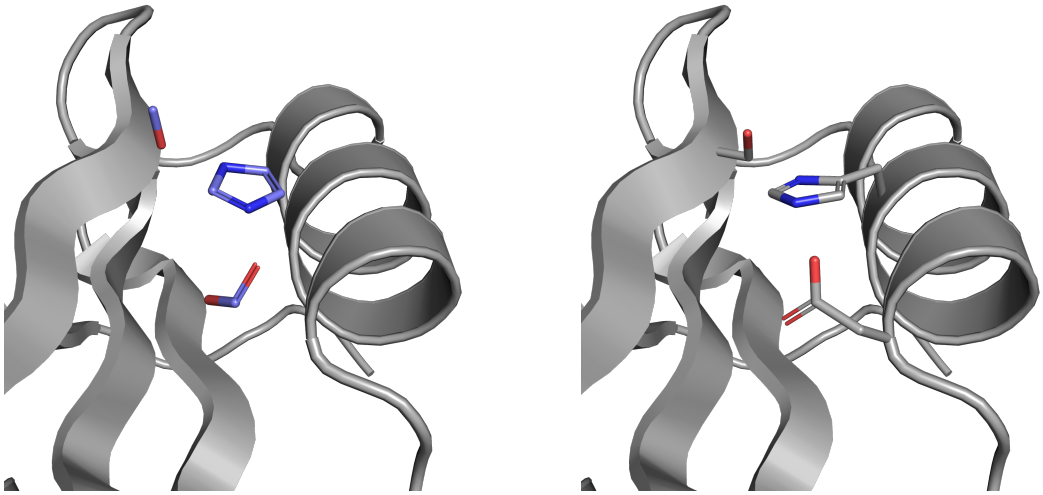


Figure 5.6: Frequencies of mutations at position 50 in one of the insertion sites (3N8M, site 51-17-13). In nearly all designs the original aspartate was mutated. This is a strong hint that the aspartate is quite problematic once the triad is inserted and should be taken care of.

After this pre-analysis, single designs that combine both low energy and a low number of mutations were inspected manually and patterns derived from the prior statistics were looked for. Since the ROSETTA design runs used constraints to enforce the formation of the catalytic triad, additional unconstrained relaxes were done with promising candidates. These relaxes did, however, not include the substrate model since the clashes between the serine and the substrates force the ApNA out of the binding pocket. The best scoring relaxes were evaluated manually and if the triads formed as expected the designs were chosen for experimental evaluation (see figure 5.7). In total several final designs were made based on five insertion sites found by SCAFFOLDSELECTION and an additional insertion site manually derived from one of these. These 6 insertion sites were placed in 4 different scaffolds.



(a) Theozyme triad positioned in a possible host scaffold

(b) relax of the same scaffold after the triad was inserted and the sequence optimized

Figure 5.7: Comparison of a part of a host scaffold (a) in blue and the placed triad after optimization of the binding pocket and relaxation (b).

5.5 Experimental construction and validation of designs

The initial set of designs was later extended by a manually designed insertion site based on an insertion found by ScaffoldSelection (see chapter 5.11 for more information). The triads were inserted into a variety of proteins from a diverse set of organisms (see table 5.1) The different variants were created by either cloning from

Host Scaffold	Organism	Insertion site	Results
3N8M / Grb2 SH2 Domain	*H. sapiens*	S53, H13, D17	Several constructs, had to be refolded, no detectable activity
3N8M / Grb2 SH2 Domain	*H. sapiens*	S51, H17, D13	Manually designed insertion site, three constructs, not yet tested
2D52 / chromone synthase	*A. arborescens*	S83, H85, D90	Eight constructs, instable, low expression, no activity
2D52 / chromone synthase	*A. arborescens*	S130, H132, D136	One construct, no expression
1W54 / Porphobilinogen synthase	*P. aeruginosa*	S125, H84, D322	No activity, low expression, Rosetta not improving energy
2CFM / DNA ligase	*P. furiosus*	S268, H533, D530	Several constructs, activity in one variant

Table 5.1: Overview over the different insertion sites and their corresponding host scaffolds. In all cases wild type proteins were cloned and expressed as well.

genomic DNA or ordering the gene with a codon-optimized sequence from Thermo Fisher GeneArt™ translating into the desired protein directly. Mutations were introduced by PCR using according primers and all constructs were introduced into the vector pet21a. The DNA insertions were sequenced and the expression strains (BL21 or BLR) were transformed with the plasmids. Expression was optimized and depending on the variant different expression temperatures, length and media were used. Expression was induced either by addition of IPTG or automatically when using self-inducing media.

5.6 Purification and enzymatic assays

After expression the cells were pelleted, washed and cracked open by sonication. After centrifugation, the desired protein in the supernatant was purified utilizing first a NiNTA column followed by analytical gel filtration. Fractions containing the

protein were pooled and concentrated. In the case of 2cfm variants, 4h of expression at 37°C, purification and measurements had to be done on the same day due to the instability of the constructs. The resulting purified proteins were characterized biophysically and used for enzymatic activity measurements.

5.7 Activity in a 2cfm variant

Activity measurements

Catalytic activity with ApNA as a substrate was measured with a variety of 2cfm constructs repeatedly. An activity over background was only detected with ApNA as the substrate and 2cfm v12, which had only one additional mutation next to the triad, a phenylalanine introduced at position 253 (see figure 5.8). All other variants showed no activity over background.

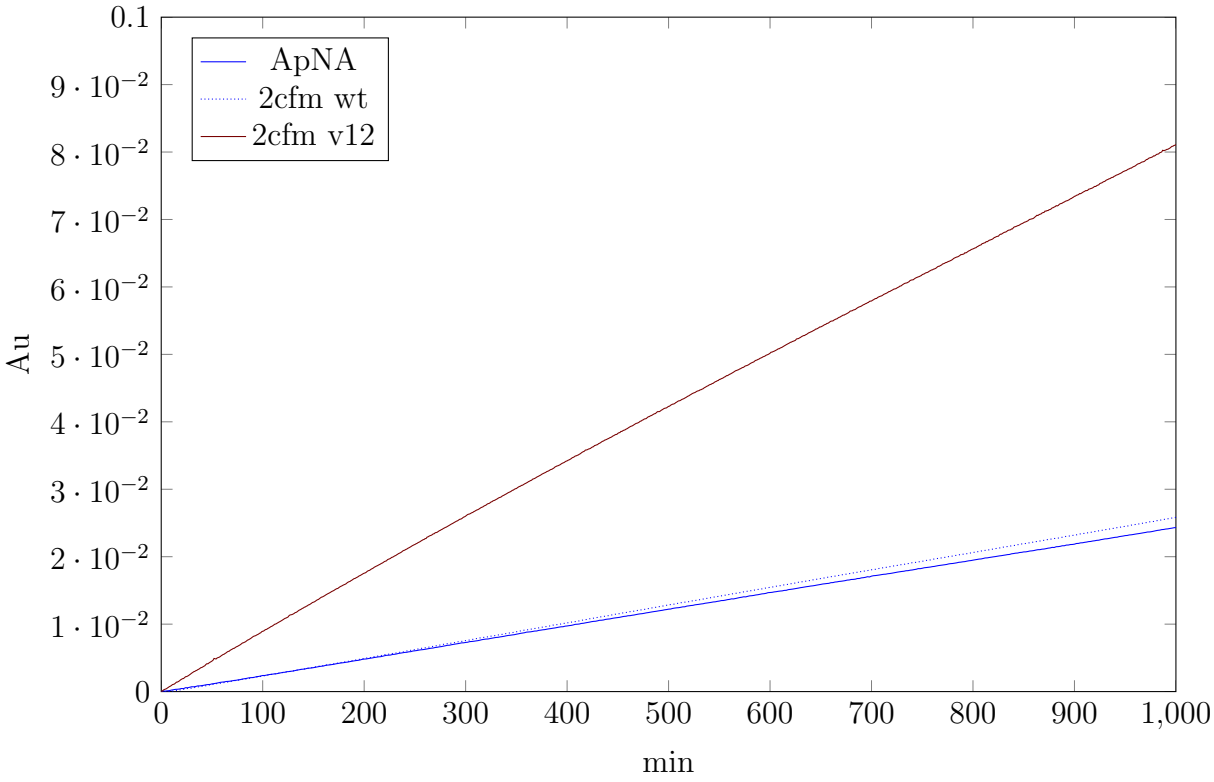


Figure 5.8: An enzymatic assay of 2cfm wild type with ApNA as a substrate shows only background activity. With the variant v12 the catalytic turnover is clearly improved. Measurements were taken with 100μM ApNA in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0.

In order to test whether the catalytic triad is responsible for the observed activity, each of the three amino acids from the catalytic triad was exchanged back to the wild type and an activity measurement was performed with each construct. Each of the knockouts led to a complete loss of activity (see Figure 5.9), leading to the conclusion that most likely all residues in triad are indeed responsible for the observed turnover.

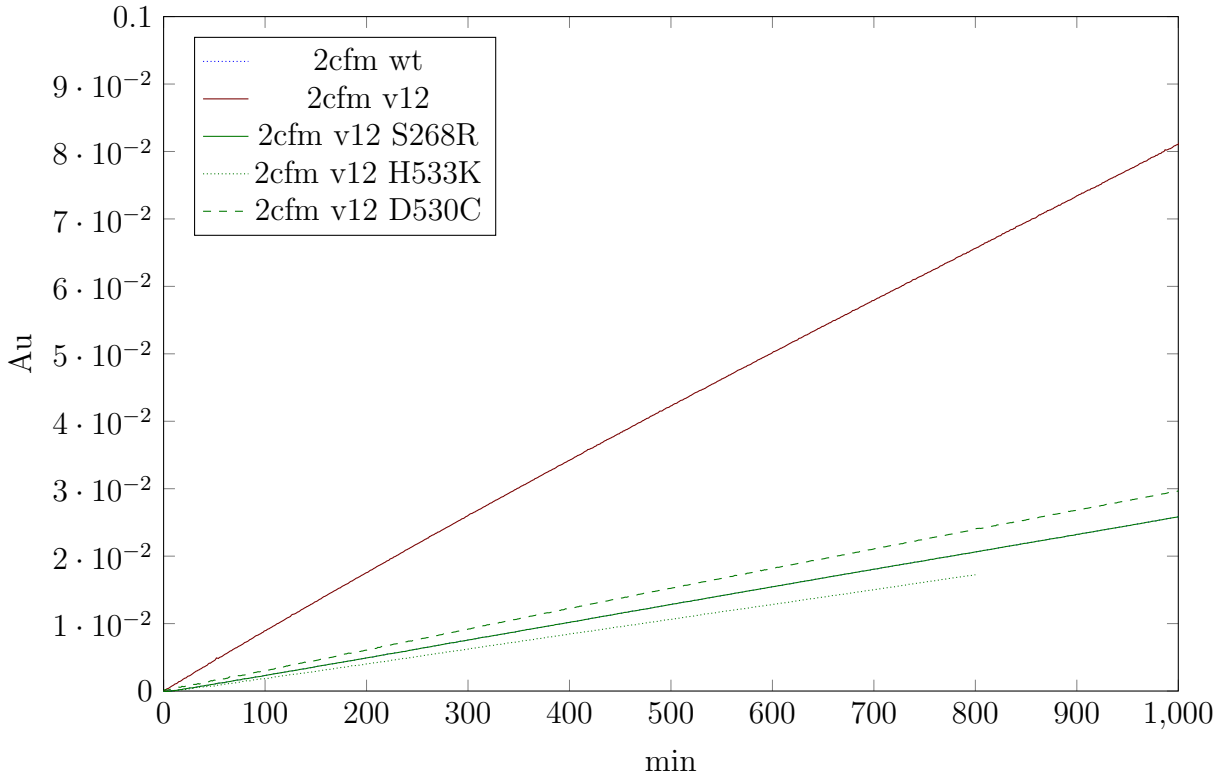


Figure 5.9: Compared to the active variant, mutations of each of the amino acids of the catalytic triad lead to a loss of function. This implies that all amino acids are important for catalysis. Measurements were taken with 100μM ApNA in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0.

In another assay both PMSF and AEBSF as inhibitors of the catalytic triad were tested for their ability to decrease the activity of the construct. In order to do so, fresh protein was incubated with either PMSF or AEBSF for one hour. After that, an activity assay with ApNA as a substrate was performed. While PMSF seems to have no effect, AEBSF is able to decrease the observed activity considerably (see figure 5.10).

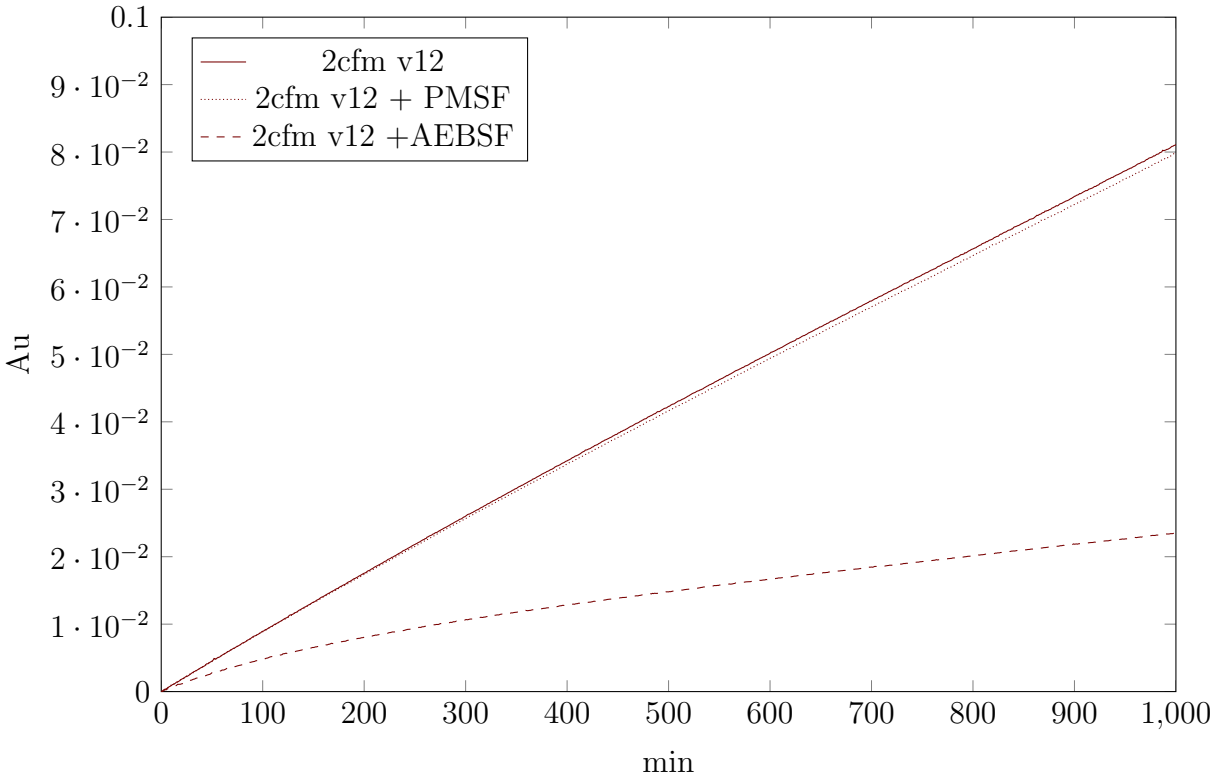


Figure 5.10: Both PMSF and AEBSF were tested for their capability to inhibit the catalytic triad. While PMSF shows no effect, AEBSF leads to a loss of function similar to background. Measurements were taken with 100μM ApNA in 150mM NaCl, 50mM potassium phosphate buffer pH 8.0.

I measured activity of 2cfm v12 at different concentrations of ApNA and derived a classical Michaelis-Menten kinetic (see figure 5.11). The K_M was estimated to be 224 μM and the catalytic efficiency was calculated to be 0.0278 molecules per min. This leads to

$$\frac{k_{cat}}{K_M} = 124 \frac{1}{M \times min}$$

The rate acceleration was rather low and calculated to be

$$\frac{k_{cat}}{k_{uncat}} = 1006$$

where k_{uncat} was calculated from the hydrolysis of free ApNA observed in the control experiments.

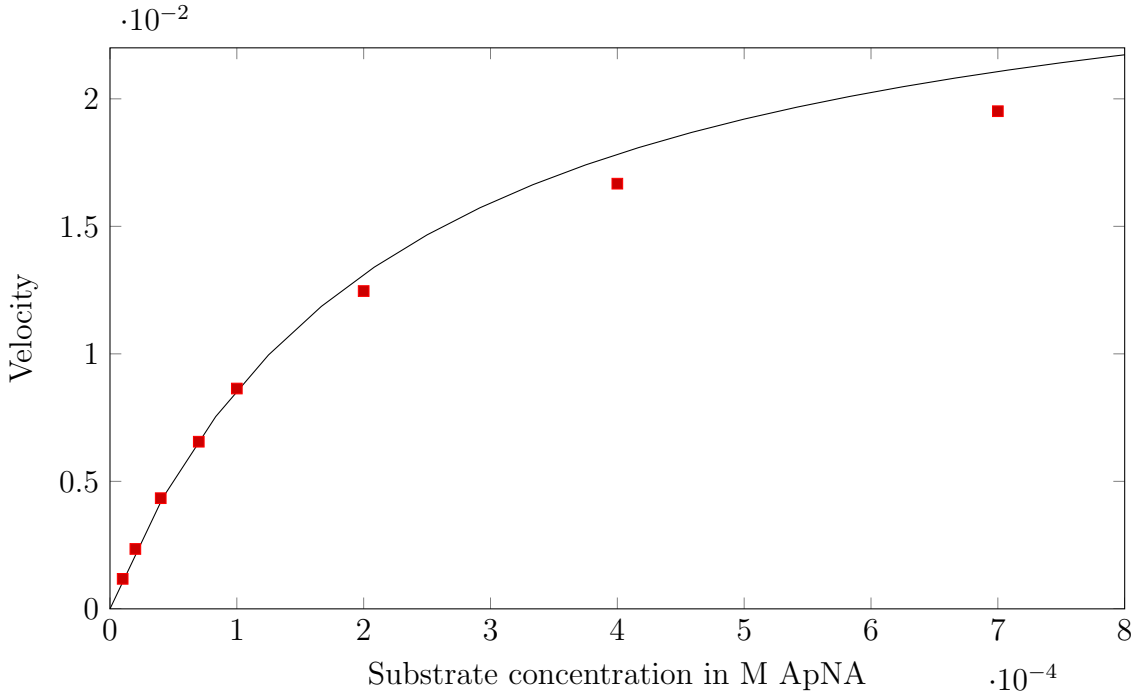


Figure 5.11: Enzymatic activity of 2cfm v12 at different concentrations of ApNA. The K_M is estimated to be 224μM, the V_{max} is estimated at 0.0278 molecules per min. Also plotted are the theoretical kinetics calculated from these parameters (black line).

5.8 Mass spectrometry

In order to test that indeed the active serine is inhibited by covalent linking to AEBSF, I decided to perform mass spectrometry experiments. For this purpose 2cfm v12 was incubated with 1mM AEBSF for at least one hour and thereafter dialyzed in a large quantity of buffer that was exchanged at least once. The first idea was to use tryptic digest followed by analysis of the digested fragments to detect the mass shift of 183 Da resulting from AEBSF bound covalently to serine. The actual mass spectrometry experiments were done by the proteome center in Tübingen. Although the experiments were repeated several times with both labeled and unlabeled samples no significant differences could be detected. I decided to try measurements of the undigested constructs, however, the fragments were too large to be detectable on the spectrometer used.

Measurements with both the labeled and unlabeled constructs were repeated at the MALDI of Hubert Kalbacher at the University Tübingen. While both the labeled and unlabeled variants could be detected and seemed to fly nicely for their size, the spectra obtained were too broad and noisy to see a mass shift as small as the expected resulting from the AEBSF label.

In order to finally obtain a good spectrum from the unfragmented protein I send samples of 2cfm v12 with and without AEBSF treatment to Anja Boumeester at the university of Utrecht. Both samples were measured on an EMR orbitrap and the spectra were send to us.

peaks from untreated sample				
	Species	Theoretical mass (Da)	Experimental mass (Da)	Δ Mas (Da)
A	Unknown		63795.9 ±0.9	
B	2cfm v12 untreated	64485.8	64481.7 ± 1.0	-4.1
C	Unknown		64547.0 ± 0.4	
peaks from sample treated with AEBSF				
	Species	Theoretical mass (Da)	Experimental mass (Da)	Δ Mass (Da)
A	2CFM V1/2	64485.8	64485.7 ± 0.2	-0.1
B	2cfm v12 + 1 AEBSF	64668.8	64666.2 ± 0.4	-2.6
C	2cfm v12 + 2 AEBSF	64851.8	64850.2 ± 0.9	-1.6
D	2cfm v12 + 3 AEBSF	65034.8	65032.6 ± 0.5	-2.2
E	2cfm v12 + 4 AEBSF	65217.8	65215.8 ± 0.9	-2.0
F	2cfm v12 + 5 AEBSF	65400.8	65399.7 ± 1.0	-1.1
G	2cfm v12 + 6 AEBSF	65583.8	65582.4 ± 1.0	-1.4
H	2cfm v12 + 7 AEBSF	65766.8	65765.4 ± 0.4	-1.4
I	2cfm v12 + 8 AEBSF	65949.8	65948.9 ± 0.7	-0.9
J	2cfm v12 + 9 AEBSF	66132.8	66131.9 ± 0.5	-0.9
K	2cfm v12 + 10 AEBSF	66315.8	66315.1 ± 0.5	-0.7
L	2cfm v12 + 11 AEBSF	66498.8	66499.3 ± 1.0	0.5
M	2cfm v12 + 12 AEBSF	66681.8	66681.3 ± 0.1	-0.5

Table 5.2: Peaks from the spectrum of the untreated and inhibited 2cfm v12. Data provided by the University Utrecht.

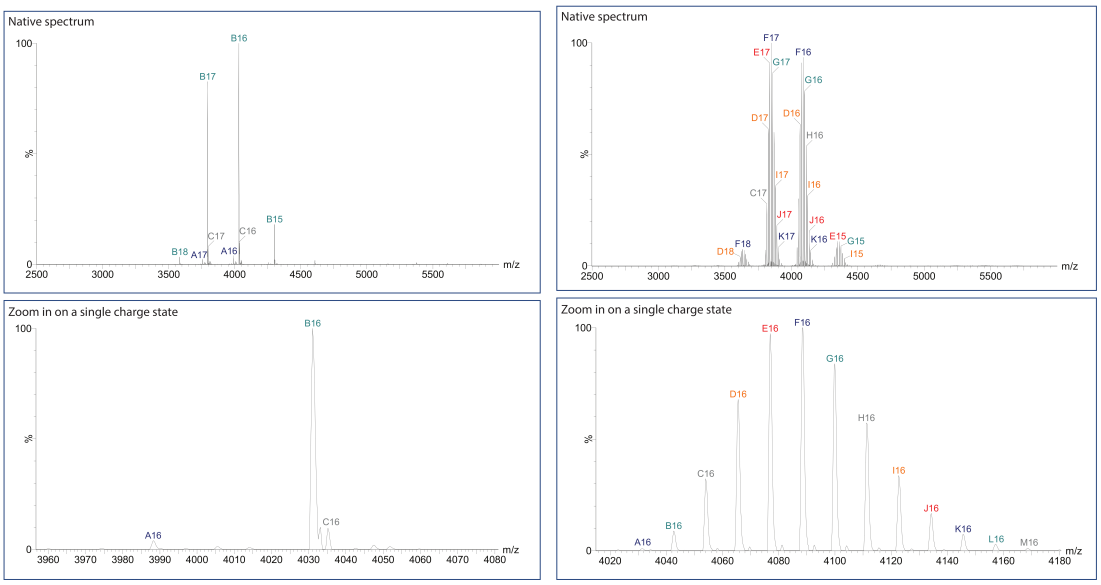


Figure 5.12: Spectra of the native and the inhibited 2cfm v12 taken with an EMR orbitrap. The inhibited sample on the right shows different mass shifts, which correspond to the masses of multiple copies of AEBSF attached to the protein.

Quite remarkably it seemed that up to twelve copies of AEBSF bound to the inhibited 2cfm v12 (see figures 5.2 and 5.12). Since the mass shift of AEBSF bound covalently is about 10 Da smaller than that of free AEBSF, it seems as if indeed twelve serines were labeled covalently. While the variant has ample serines to explain this massive labeling, sulfonyl fluorides should not be able to activate non-activated serines close to neutral pH [57]. In the future, we therefore want to repeat the tryptic digest experiments in Utrecht and try to localize the labeling sites.

5.9 Crystal structure

Several crystallization screens were set up with the 2cfm v12 construct both alone and with AEBSF or ApNA in the mixture. Only one condition formed crystals after a microseeding experiment and only in the pure protein mixture. Spectra were recorded at the PXII, SLS. Molecular replacement was challenging and was initially carried out with the backbone of 2cfm. Both phasing and autobuild lead to incomplete solutions in the beginning. I therefore iterated a circle of phasing with the latest solution, autobuild, manual building and refinements several times. After a satisfying solution was found I started refining the structure. After the first phasing was successful, in the end a structure could be solved with a resolution of 2.6Å, (see table 5.3) showing 2cfm v12 to be present in an open conformation with two copies forming a dimer and the triad being torn apart (compare figure 5.13). Since prior runs of gel filtration showed only a monomeric species present I decided to investigate whether maybe short lived dimers could be observed in solution.

Wavelength (Å)	0.97796
Resolution range (Å)	46.01 - 2.593 (2.686 - 2.593)
Space group	P 42 21 2 / 94
Unit cell	146.05 146.05 73.91 90 90 90
Unique reflections	25033 (2399)
Completeness (%)	99.79 (97.88)
Mean I/sigma(I)	11.98 (2.27)
Wilson B-factor	48.19
R-work	0.2140 (0.2941)
R-free	0.2687 (0.3665)

Table 5.3: Properties of the crystal structure obtained from 2cfm v12

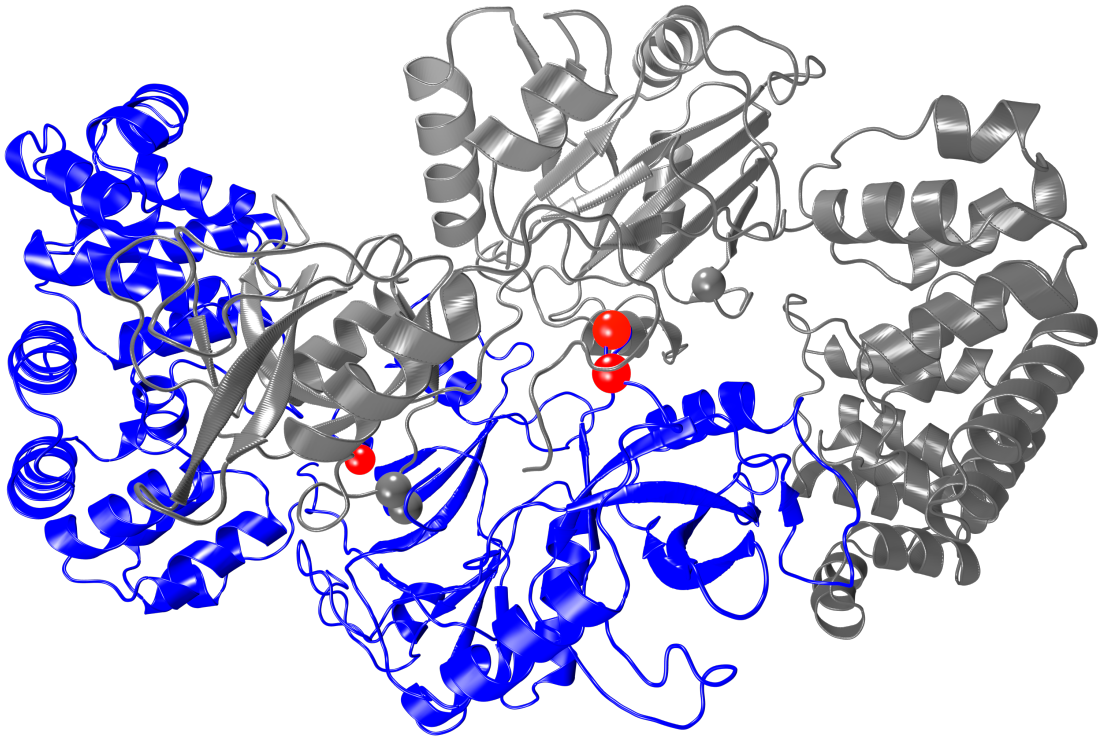


Figure 5.13: Structure of 2cfm v12 with two copies interacting with another. The balls represent the insertion sites of the catalytic triad, with the histidine and aspartate close to each other but the serine quite distant due to the two interlocked copies.

5.10 Multi angle light scattering(MALS)

In order to test whether 2cfm might form short-lived dimers in solution similar to the ones observed in the crystal structure multi angle laser light scattering experiments were set up.

Purified protein of 2cfm v1/2 was split into two aliquots, one of which was treated with AEBSF. Both were separately run over a S75 analytical gel filtration column immediately followed by MALS measurements of the filtered flowthrough. In both cases only one diffracting species with an estimated mass corresponding to the monomer could be detected, leading to the conclusion that also short lived dimers are not formed in solution (see figure 5.14).

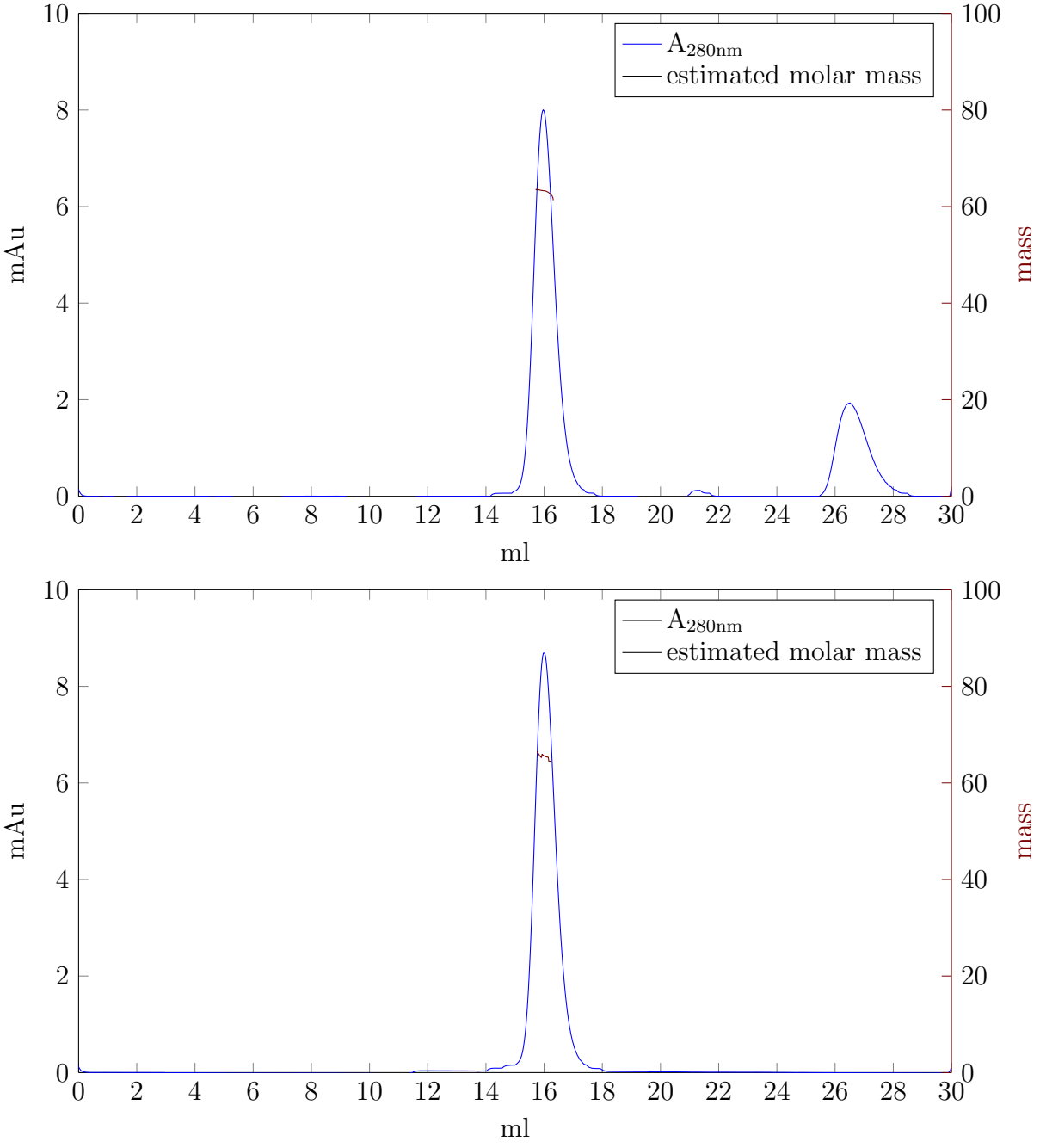


Figure 5.14: (a) MALS measurement of untreated 2cfm var12. The second peak is caused due to the equilibration of the column and is visible at every new measurement according to the operator. The measurement shows a single peak wit an estimated average particle size of 65 kDa, coinciding with the size of 2cfm. (b)MALS measurement of 2cfm var12 inhibited with AEBSF. The measurement shows a single peak with an estimated average particle size of 65 kDa, coinciding with the size of 2cfm. Measurement taken with Stefan Grüner on a Wyatt miniDAWN TREOS at 658nm.

Conclusion

The only active variant so far, 2cfm, is quite challenging. Although activity seems to depend on all members of the catalytic triad, it is hard to get any direct evidence that the enzymatic mechanism mirrors that of natural serine proteases and the design goal is therefore met. One inherent problem of the host scaffold is its large size, that makes a variety of experiments such as structure determinition by NMR unfeasible and due to the complex arrangement of three domains makes other experiments such as crystallography more likely to fail.

Since there are only two domains which carry the designed triad and the first domain which interlocks with another copy of the protein in the crystal structure is also unlikely to interact with the bound substrate, one of the next steps would be to express the protein without the first domain and repeat the enzymatic assays and crystallization.

However one of the earlier and quite small designs was also reevaluated and used for additional designs with some unique properties.

5.11 Guiding computational design into new directions

Manual redesign of the 3N8M insertion site

The insertion site into 3N8M revealed some major problems. In general, the triads sidechains were not able to come close enough to each other to create a functional geometry. Since the protein itself is quite small and the placement of aspartate and histidine was good relative to each other I tried a manual redesign and evaluated it computationally.

With the current placement of histidine and aspartate, it was impossible to find a suitable position for the serine. I therefore permuted the histidine and the aspartate. The favorable geometry of histidine to aspartate stays mainly intact after the permutation and I started looking for insertion sites for serine close to the new histidine position.

One suitable position was found for serine, however this new insertion site favors a geometry where the histidine ring is flipped, inverting the roles of the nitrogens in the indole ring. I decided to do designs with Rosetta 3.5 for both orientations and at the same side look into the feasibility of a catalytic triad with an inverted histidine. Based on the resulting structures, designs were chosen together with Marcel Conrady from the University of Tübingen. He will continue with this project in his bachelor thesis.

Inverting the Histidine ring

Since the indole ring of histidine has two nitrogen atoms and both can be protonated, there are two different tautomers in solution. Although the tautomer with the protonated N-1 is preferred, it should be possible to also create proteases with an inverted histidine where the N-1 interacts with the serine while the protonated N-3 forms a hydrogen bond with the aspartate. Since the same insertion site of histidine leads to geometrically different placements of the remaining catalytic residues when different tautomers are used, this could not only be a chemically interesting alternative but also increase the amount of possible insertion sites.

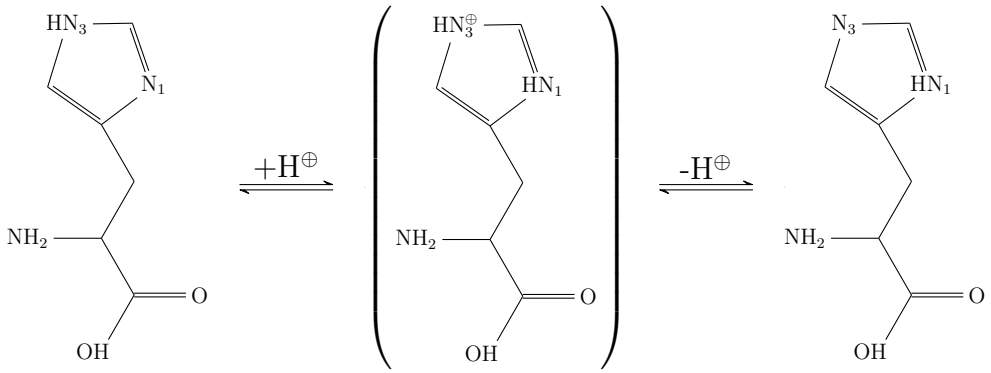


Figure 5.15: The two different tautomers of the indole ring in histidine. The transition between the two tautomers is rapid at lower pH and slows down rapidly with a pH of 9.

If the transition between the tautomers is fast (this is the case at low pH), only a single peak should be visible in NMR at a quite unique position (12ppm on the proton and 200ppm on the nitrogen scale). With a lower transition (at high pH), these should separate into two peaks that differ on the nitrogen scale. Since usually spectra are not recorded down to 200ppm, I decided to start NMR measurements to see whether I can observe them. In a first NMR experiments with N_{15}-labeled

ubiquitin no peaks were visible, which might be due to the fact that the only histidine in ubiquitin is quite close to the C terminus. It will, however, be interesting to repeat the experiments with the new variants and try to find a pH and temperature where both peaks are visible in order to be able to determine which orientation of the histidine ring is dominating.

5.12 Discussion

Although the goal to introduce catalytic activity into a host scaffold seems to be accomplished with the 2cfm v12 variant, the proof that indeed the catalytic triad is responsible for this activity and the catalytic mechanism works as intended is difficult to make. Different approaches to measure the inhibition of the active serine with mass spectrometry were inconclusive, spectra of the unfragmented protein show multiple bound inhibitor molecules. Further experiments with fragmentation of the construct are needed to give insight into the binding sites of the inhibitor.

Thus far, only one set of crystals formed without any additional compounds present. Attempts to co-crystallize the construct with either the substrate or inhibitor were unsuccessful. The resulting structure shows that the protein is present in an open form with the catalytic triad torn apart. This is likely the result of interactions with a second copy of 2cfm v12. In solution no dimers could be detected with gel filtration or multi angle light scattering.

On the other hand a new generation of designs in a manually adapted insertion site in the 3N8M scaffold is about to be tested, which also features a reversed indole ring and according to the designs have better energies and a lower flexibility than prior designs.

Publications

Feldmeier K and Höcker B (2013). Computational protein design of ligand binding and catalysis. Current Opinion in Chemical Biology, 17(6), 929-933 [21]

The review was planned and written together with Birte Höecker.

Stiel A, Feldmeier K and Höcker B (2014). Identification of Protein Scaffolds for Enzyme Design Using Scaffold Selection. Methods in Molecular Biology, 2014, 1216(1999), 117-128 [55]

This review was planned and written together with Andre Stiel and Birte Höecker.

Huang P*, Feldmeier K*, Parmeggiani F, Fernandez A, Höcker B & Baker D (2016). De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy. Nature Chemical Biology, 12, 29-34 [1] * Equal contribution

The Design was planned together with all authors involved. I characterized most of the variants described in the publication including techniques such as gel filtrations, CD spectra and melting point analysis, Trp fluorescence and NMR. For the successful design I additionally performed chemical unfolding studies and derived ΔG values, determined the non-formation of the disulfide bonds designed, set up crystallization screens (the pipetting robot was operated by Sooruban Shanmugaratnam), took crystal spectra at the SLS and solved the structure now published as 5BVL. Together with Birte Höcker I looked into the relation to other TIM-barrels, made HHsearch comparisons and created a cluster map of sTIM11 with other members of the TIM-barrel fold.

Chapter 6

Bibliography

Bibliography

[1] P.-S. Huang, K. Feldmeier, F. Parmeggiani, D. A. Fernandez Velasco, B. Höcker, and D. Baker, "De novo design of a four-fold symmetric TIM-barrel protein with atomic-level accuracy," *Nature Chemical Biology*, vol. 12, no. November, pp. 29–34, 2016.

[2] S. D. Copley, E. Smith, and H. J. Morowitz, "The origin of the RNA world: Co-evolution of genes and metabolism," *Bioorganic Chemistry*, vol. 35, no. 6, pp. 430–443, 2007.

[3] D. W. Banner, A. C. Bloomer, G. A. Petsko, D. C. Phillips, and I. A. Wilson, "Atomic coordinates for triose phosphate isomerase from chicken muscle," *Biochemical and Biophysical Research Communications*, vol. 72, no. 1, pp. 138–145, 1976.

[4] A. R. Buller and C. a. Townsend, "Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 8, pp. E653–61, 2013.

[5] G. Dodson and A. Wlodawer, "Catalytic triads and their relatives," *Trends in Biochemical Sciences*, vol. 23, pp. 347–352, sep 1998.

[6] B. Kuhlman, G. Dantas, G. C. Ireton, G. Varani, B. L. Stoddard, and D. Baker, "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy," *Science*, vol. 302, no. November, pp. 1364–1368, 2003.

[7] N. H. Joh, T. Wang, M. P. Bhate, R. Acharya, Y. Wu, M. Grabe, M. Hong, G. Grigoryan, and W. F. DeGrado, "De novo design of a transmembrane Zn-transporting four-helix bundle.," *Science (New York, N.Y.)*, vol. 346, no. 6216, pp. 1520–4, 2014.

[8] N. Koga, R. Tatsumi-Koga, G. Liu, R. Xiao, T. B. Acton, G. T. Montelione, and D. Baker, "Principles for designing ideal protein structures.," *Nature*, vol. 491, no. 7423, pp. 222–7, 2012.

[9] K. Goraj, A. Renard, and J. A. Martial, "Synthesis, purification and initial structural characterization of octarellin, a de novo polypeptide modelled on the alpha/beta-barrel proteins.," *Protein engineering*, vol. 3, no. 4, pp. 259–266, 1990.

[10] M. Beauregard, K. Goraj, V. Goffin, K. Heremans, E. Goormaghtigh, J. M. Ruysschaert, and J. A. Martial, "Spectroscopic investigation of structure in octarellin (a de novo protein designed to adopt the alpha/beta-barrel packing).," *Protein engineering*, vol. 4, no. 7, pp. 745–749, 1991.

[11] T. Tanaka, H. Kimura, M. Hayashi, Y. Fujiyoshi, K. Fukuhara, and H. Nakamura, "Characteristics of a de novo designed protein," *Protein Sci*, vol. 3, no. 3, pp. 419–427, 1994.

[12] A. Houbrechts, B. Moreau, R. Abagyan, V. Mainfroid, G. Préaux, A. Lamproye, A. Poncin, E. Goormaghtigh, J. M. Ruysschaert, and J. A. Martial, "Second-generation octarellins: two new de novo (beta/alpha)8 polypeptides designed

for investigating the influence of beta-residue packing on the alpha/beta-barrel structure stability.," *Protein engineering*, vol. 8, no. 3, pp. 249–259, 1995.

[13] F. Offredi, F. Dubail, P. Kischel, K. Sarinski, A. S. Stern, C. Van de Weerdt, J. C. Hoch, C. Prosperi, J. M. François, S. L. Mayo, and J. A. Martial, "De novo backbone and sequence design of an idealized α/β-barrel protein: Evidence of stable tertiary structure," *Journal of Molecular Biology*, vol. 325, no. 1, pp. 163–174, 2003.

[14] M. Figueroa, N. Oliveira, A. Lejeune, K. W. Kaufmann, B. M. Dorr, A. Matagne, J. a. Martial, J. Meiler, and C. Van de Weerdt, "Octarellin VI: using rosetta to design a putative artificial $(\beta/\alpha)8$ protein.," *PloS one*, vol. 8, no. 8, p. e71858, 2013.

[15] 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, D. Hilvert, K. N. Houk, B. L. Stoddard, and D. Baker, "De Novo Computational Design of Retro-Aldol Enzymes," *Science*, vol. 319, no. 5868, pp. 1387–1391, 2008.

[16] E. A. Althoff, L. Wang, L. Jiang, L. Giger, J. K. Lassila, Z. Wang, M. Smith, S. Hari, P. Kast, D. Herschlag, D. Hilvert, and D. Baker, "Robust design and optimization of retroaldol enzymes," *Protein Science*, vol. 21, pp. 717–726, may 2012.

[17] D. Röthlisberger, 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, and D. Baker, "Kemp elimination catalysts by computational enzyme design.," *Nature*, vol. 453, pp. 190–5, may 2008.

[18] M. Merski and B. K. Shoichet, "Engineering a model protein cavity to catalyze the Kemp elimination," *Proceedings of the National Academy of Sciences*, vol. 109, pp. 16179–16183, oct 2012.

[19] S. Bjelic, L. G. Nivon, N. Celebi-Olcum, G. Kiss, C. F. Rosewall, H. M. Lovick, E. L. Ingalls, J. L. Gallaher, J. Seetharaman, S. Lew, G. T. Montelione, J. F. Hunt, F. E. Michael, K. N. Houk, and D. Baker, "Computational design of enone-binding proteins with catalytic activity for the morita-baylis-hillman reaction," *ACS Chemical Biology*, vol. 8, no. 4, pp. 749–757, 2013.

[20] 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, and D. Baker, "Computational design of an enzyme catalyst for a stereoselective bimolecular Diels-Alder reaction.," *Science (New York, N.Y.)*, vol. 329, pp. 309–13, jul 2010.

[21] K. Feldmeier and B. Höcker, "Computational protein design of ligand binding and catalysis," *Current Opinion in Chemical Biology*, vol. 17, no. 6, pp. 929–933, 2013.

[22] F. Richter, R. Blomberg, S. D. Khare, G. Kiss, A. P. Kuzin, A. J. T. Smith, J. Gallaher, Z. Pianowski, R. C. Helgeson, A. Grjasnow, R. Xiao, J. Seetharaman, M. Su, S. Vorobiev, S. Lew, F. Forouhar, G. J. Kornhaber, J. F. Hunt, G. T. Montelione, L. Tong, K. N. Houk, D. Hilvert, and D. Baker, "Computational design of catalytic dyads and oxyanion holes for ester hydrolysis," *Journal of the American Chemical Society*, vol. 134, pp. 16197–16206, oct 2012.

[23] S. Rajagopalan, C. Wang, K. Yu, A. P. Kuzin, F. Richter, S. Lew, A. E. Miklos, M. L. Matthews, J. Seetharaman, M. Su, J. F. Hunt, B. F. Cravatt, and

D. Baker, "Design of activated serine-containing catalytic triads with atomic-level accuracy.," *Nature chemical biology*, vol. 10, no. 5, pp. 386–91, 2014.

[24] C. Malisi, O. Kohlbacher, and B. Höcker, "Automated scaffold selection for enzyme design," *Proteins: Structure, Function and Bioinformatics*, vol. 77, no. 1, pp. 74–83, 2009.

[25] A. Leaver-Fay, M. Tyka, S. M. Lewis, 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. A. Ban, S. J. Fleishman, E. Corn, D. E. Kim, S. Lyskov, M. Berrondo, J. J. Havranek, S. Mentzer, Z. Popovic, J. Karanicolas, R. Das, J. Meiler, T. Kortemme, J. J. Gray, B. Kuhlman, D. Baker, and P. Bradley, "ROSETTA 3 : An Object-Oriented Software Suite for the Simulation and Design of Macromolecules," *Methods in Enzymology, Volume 487*, vol. 487, no. 11, pp. 545–574, 2011.

[26] W. Kabsch, "Xds," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 2, pp. 125–132, 2010.

[27] P. D. Adams, P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L. W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger, and P. H. Zwart, "PHENIX: A comprehensive Python-based system for macromolecular structure solution," *Acta Crystallographica Section D: Biological Crystallography*, vol. 66, no. 2, pp. 213–221, 2010.

[28] K. Okonechnikov, O. Golosova, M. Fursov, A. Varlamov, Y. Vaskin, I. Efremov, O. G. German Grehov, D. Kandrov, K. Rasputin, M. Syabro, and T. Tleukenov, "Unipro UGENE: A unified bioinformatics toolkit," *Bioinformatics*, vol. 28, no. 8, pp. 1166–1167, 2012.

[29] A. Biegert, C. Mayer, M. Remmert, J. Söding, and A. N. Lupas, "The MPI Bioinformatics Toolkit for protein sequence analysis," *Nucleic Acids Research*, vol. 34, no. WEB. SERV. ISS., pp. 335–339, 2006.

[30] T. B. Leite, D. Gomes, M. A. Miteva, J. Chomilier, B. O. Villoutreix, and P. Tuffery, "Frog: A FRee Online druG 3D conformation generator," *Nucleic Acids Research*, vol. 35, no. SUPPL.2, pp. 568–572, 2007.

[31] T. Frickey and A. Lupas, "CLANS: A Java application for visualizing protein families based on pairwise similarity," *Bioinformatics*, vol. 20, no. 18, pp. 3702–3704, 2004.

[32] R. K. Saiki, D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi, G. T. Horn, K. B. Mullis, and H. A. Erlich, "Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase.," *Science (New York, N.Y.)*, vol. 239, no. 4839, pp. 487–491, 1988.

[33] M. Andersson, B. Wittgren, and K.-g. Wahlund, "Accuracy in Multiangle Light Scattering Measurements for Molar Mass and Radius Estimations . Model Calculations and Experiments Accuracy in Multiangle Light Scattering Measurements for Molar Mass and Radius Estimations . Model Calculations and Experiments," *Anal. Chem.*, vol. 75, no. 16, pp. 4279–4291, 2003.

[34] W. C. J. Jr, "Protein secondary structure and circular dichroism: a practical guide.," *Proteins: Structure, Function, and Genetics*, vol. 7, no. 3, pp. 205–214, 1990.

[35] N. Greenfield, "Using circular dichroism spectra to estimate protein secondary structure," *Nature Protocols*, vol. 1, no. 6, pp. 2876–2890, 2007.

[36] T. E. Creighton, *Protein Structure: A Practical Approach*. 1997.

[37] G. R. Grimsley, B. M. Huyghues-Despointes, C. N. Pace, and J. M. Scholtz, "Measuring the Conformational Stability of a Protein by NMR," *Cold Spring Harbor Protocols*, vol. 2006, no. 1, pp. pdb.prot4244–pdb.prot4244, 2006.

[38] B. Rupp, *Biomolecular Crystallography*. 2010.

[39] H. Hauptman, "The phase problem of x-ray crystallography," 1983.

[40] N. Nagano, C. A. Orengo, and J. M. Thornton, "One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions," *Journal of Molecular Biology*, vol. 321, no. 5, pp. 741–765, 2002.

[41] H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne, "The Protein Data Bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[42] T. J. P. Hubbard, B. Ailey, S. E. Brenner, A. G. Murzin, and C. Chothia, "SCOP: A structural classification of proteins database," *Nucleic Acids Research*, vol. 27, no. 1, pp. 254–256, 1999.

[43] W. G. J. Hol, "The role of the a-helix dipole in protein function and structure," *Progress in Biophysics and Molecular Biology*, vol. 45, no. 3, pp. 149–195, 1985.

[44] B. Höcker, C. Jürgens, M. Wilmanns, and R. Sterner, "Stability, catalytic versatility and evolution of the $(\beta\alpha)$8-barrel fold," 2001.

[45] B. Hocker, J. Claren, and R. Sterner, "Mimicking enzyme evolution by generating new (ba)8-barrels from (ba)4-half-barrels," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 47, pp. 16448–16453, 2004.

[46] L. N. Kinch and N. V. Grishin, "Evolution of protein structures and functions," *Current Opinion in Structural Biology*, vol. 12, pp. 400–408, jun 2002.

[47] M. Henn-Sax, B. Höcker, M. Wilmanns, and R. Sterner, "Divergent evolution of ($\beta\alpha$)8-barrel enzymes," *Biological Chemistry*, vol. 382, no. 9, pp. 1315–1320, 2001.

[48] L. Carstensen, J. M. Sperl, M. Bocola, F. List, F. X. Schmid, and R. Sterner, "Conservation of the folding mechanism between designed primordial ($\beta\alpha$)8-barrel proteins and their modern descendant," *Journal of the American Chemical Society*, vol. 134, no. 30, pp. 12786–12791, 2012.

[49] M. Richter, M. Bosnali, L. Carstensen, T. Seitz, H. Durchschlag, S. Blanquart, R. Merkl, and R. Sterner, "Computational and experimental evidence for the evolution of a (ba)8-barrel protein from an ancestral quarter-barrel stabilised by disulfide bonds," *Journal of Molecular Biology*, vol. 398, no. 5, pp. 763–773, 2010.

[50] D. Reardon and G. K. Farber, "The structure and evolution of alpha/beta barrel proteins.," *The FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, vol. 9, no. 7, pp. 497–503, 1995.

[51] M. Goujon, H. McWilliam, W. Li, F. Valentin, S. Squizzato, J. Paern, and R. Lopez, "A new bioinformatics analysis tools framework at EMBL-EBI," *Nucleic Acids Research*, vol. 38, no. SUPPL. 2, pp. W695–W699, 2010.

[52] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J. D. Thompson, and D. G. Higgins, "Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.," *Molecular systems biology*, vol. 7, no. 1, p. 539, 2011.

[53] H. McWilliam, W. Li, M. Uludag, S. Squizzato, Y. M. Park, N. Buso, A. P. Cowley, and R. Lopez, "Analysis Tool Web Services from the EMBL-EBI.," *Nucleic acids research*, vol. 41, no. Web Server issue, pp. W597–600, 2013.

[54] J. Söding, "Protein homology detection by HMM-HMM comparison," *Bioinformatics*, vol. 21, no. 7, pp. 951–960, 2005.

[55] F. K. Stiel A and H. B, "Identification of Protein Scaffolds for Enzyme Design Using Scaffold Selection. Methods in Molecular Biology," in *Methods in Molecular Biology*, vol. 1216, pp. 117–128, 2014.

[56] D. J. Tantillo, J. Chen, and K. N. Houk, "Theozymes and compuzymes: theoretical models for biological catalysis.," *Current opinion in chemical biology*, vol. 2, no. 6, pp. 743–750, 1998.

[57] A. M. Gold and D. Fahrney, "Sulfonyl Fluorides as Inhibitors of Esterases. II. Formation and Reactions of Phenylmethanesulfonyl α-Chymotrypsin," *Biochemistry*, vol. 3, no. 6, pp. 783–791, 1964.

Chapter 7

Acknowledgment

This work would not have been possible without a lot of both scientific and personal support.

On a scientific basis I would first like to thank my supervisor Prof. Dr. Birte Höcker, who gave me the opportunity to work in her group, both leaving me freedom to follow my own ideas as well as bringing me back if I ran too far astray.

I would like to thank the members of my thesis advisory committee, Prof. Dr. Oliver Kohlbacher and Prof. Dr. Thilo Stehle. The meetings on the progress of my work were always as friendly as they were helpful. A variety of turns this projects took would not have been possible without this input.

I would also like to thank the members of my Group, especially Sooruban Shanmugaratnam, his experience in the labratory made things much easier, and Andre Stiel, who gave me a lot of advice when solving protein structures and who lived through the heights and downs of Rosetta together with me.

There were several people working on these projects with me. Robert Rietmeijer from the University of Rochester did a RISE internship on the project of the triad design. Gregor Wiese from the University of Tübingen is working on the new TIM-barrel and wants to bring the design even further, introducing functionality into the

scaffold. Marcel Conrady from the University of Tübingen did an internship with me designing new variants for the triad with the inverted indole ring. He will likely continue this work.

The design of the TIM barrel was done in cooperation with Po-Ssu Huang, Fabio Parmeggiani, Alejandro Fernandez Velasco and David Baker (university of Washington, Seattle).

Many of the mass spectrometry experiments were performed together with Vaishnavi Ravikumar and Boris Macek at the Proteome Center Tübingen and with Sophie Stotz and Hubert Kalbacher from the University of Tübingen.

NMR experiments were done together with Vincent Truffault, Remco Sprangers and Silke Wiesner from the Max Planck Institute for Developmental Biology Tübingen.

MALS Measurements were recorded with Stefan Grüner also from the MPI Tübingen.

I would also like to thank Department 1 for making the CD and fluorescence spectrometer accessible to us and Andre Noll for keeping our Cluster running!

On a personal level I would like to thank my wife, Miriam Bombieri, her sunny mood and passion for Italian food made sure both my mind and body would stay healthy even while finishing her own PHD thesis.

Of course I also would like to thank my parents, Ulrike and Reinhard Feldmeier, who made sure I would not - to cite Rückert - "der Welt abhandenkommen", or in this case, get lost in the laboratory.