UNIVERSITY OF LIVERPOOL

Development of novel gelatin-binding proteins for targeting therapeutics to cartilage lesions in the osteoarthritic joint

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Gemma Amy Paige Jackson

October 2022

Abstract: Development of novel gelatin-binding proteins for targeting therapeutics to cartilage lesions in the osteoarthritic joint

Osteoarthritis (OA) is the most common chronic joint disease appearing to be increasing in prevalence amongst a population that is now on average living significantly longer. Age is a well-established risk factor for OA. Type II (TII) gelatin, derived from TII collagen is found abundantly in damaged regions of the OA joint, making it an ideal target for binding to target a therapeutic. Retention and integration of mesenchymal stem cells (MSCs) within the damaged regions of the OA joint could facilitate not only joint repair, a prolonged improvement in pain and mobility for OA patients, but also could be defined as the first disease modifying OA therapeutic. The collagen binding domain (CBD) of Matrix-metalloproteinase-2 (MMP-2) binds to TII gelatin and has previously been used by the group as a starting point for designing mutants with increased affinity for TII gelatin. 222 is a chimeric protein previously developed by the Hollander group and proven to bind with an affinity fourteen times greater than CBD to TII gelatin. However, it was concluded that 222 would be challenging to exploit therapeutically because of poor solubility and variable efficacy, therefore the aim of this thesis was to further enhance binding efficacy and/or to improve solubility of 222. Subsequent designed mutant proteins here were intended to be used to coat mesenchymal stem cells (MSCs), to promote potent adherence to TII gelatin in the OA joint.

In this work a combination of *in silico* and *in vitro* experiments were conducted with the aim of developing proteins that bind with a greater affinity to TII gelatin than 222. Firstly, binding residues were assessed *in silico for* surface exposure and stability, to impact upon binding and be recoverable *in vitro*. Four mutants were selected to take to in vitro experiments. Asn (N) 11, 69 and 127 (equivalents in the three modules of 222) were, identified as the most important binding residues. As

their mutation to Ala (A) caused the greatest decrease in binding affinity, when assessed *in vitro* using a TII gelatin binding assay.

All three important residues were then mutated to every alternate *in silico.* Docking predicted 222W, with Trp (W) (substituted at position 11, 69 and 127) as the only such mutant predicted *in silico* to have stronger binding affinity than 222. Use of a Maltose-Binding Protein (MBP) tag was successful in aiding soluble expression of this mutant 222W. However, this mutation seemed also to alter the characteristics of the protein, preventing 3C protease cleavage, meaning protein of interest (POI) alone could not be isolated.

Alongside, the CamSol webserver was utilised to design six mutants with increasing solubility. The CamSol tool predicted mutant protein CamSol6 (CS6) to be the most soluble with three substitutions Val (V) 4 to Glu (E); Phe (F) 6 to E; Tyr (Y) 9 to E and an insertion of EEE between Gly (G) 97 and Y98. CS6 was expressed, purified, and characterised for solubility and binding to TII gelatin. An amorphous precipitation assay with polyethylene glycol (PEG) and ammonium sulfate was used to give a measure of apparent solubility. CS6 was more soluble than both CBD and 222, however binding to TII gelatin was reduced compared to 222.

The work presented in this thesis has identified key residues important for the binding of 222 to TII gelatin, as well as those important for enhancing solubility. To be used as a therapeutic a protein must be shown to be stable, soluble, demonstrate minimal heterogeneity, minimal contamination as well as being suitable to scale, with consistent and reproducible expression, purification and physiochemistry. Further work is warranted to achieve these properties and develop a strongly binding and more soluble mutant of 222, with a balance of both characteristics for optimal therapeutic potential.

Acknowledgements

First and foremost, I must thank my supervisors Professor Anthony Hollander and Dr Jill Madine. Two people I am very indebted to for sharing the wisdom that comes from years of research experience. Questioning me, spurring me on, challenging me, helping but also making me squirm ever so slightly on occasion as necessary.

Special thanks to Dr Anna Salerno and Dr Anais Dabbadie, early on you helped me find my feet in the laboratory. Dr Amy Wood thank you for helping especially when the Äkta systems went awry. Also, thanks to Professor Dan Rigden for helping with all things bioinformatics you helped me expand my project in a way I never foresaw at the outset. Dr Michael Batie, Dr Mark Frost and Dr Dominic Byrne three people who helped me a lot with troubleshooting. I know you all know the pain when a result continues to elude, thank you for giving me help and advice when I needed it most. To lab colleagues too numerous to name who shared my misery, a laugh or a kind word of encouragement, thanks for being part of my PhD journey.

Eternal gratitude to Sally Dickinson Crawshaw and family, without the funding left in Sally's name I would not have had a chance to undertake a PhD. I have learnt so much and really am not the same naïve scientist I was when I began.

Andrew, my husband who has been the most patient of all thanks for carrying me these last few years, I will not be deadweight much longer. My mum for maintaining belief in me even when I didn't always maintain it in myself. Finally special mention goes to my many pets, a PhD doesn't seem quite as stressful when you have fluffy family members waiting at home, the best comfort.

Finally, a more general reflective acknowledgement, thanks to all the scientists out there working tirelessly on science that really does change lives.

Contents

List of abbreviations

aa	Amino acid
ACI	Autologous chondrocyte implantation
ADAMTS	A disintegrin and metalloproteinase with thrombospondin motifs
AIR	Ambiguous interaction restraints
Amp	Ampicillin
Anti-IL-1	Anti-interleukin 1
AT-MSCs	Adipose tissue- derived MSCs
BLAST	Basic local Alignment Search Tool
BMI	Body mass index
BM-MSCs	Bone marrow derived MSCs
bp	Base pairs
BSA	Buried surface area
CAPRI	Critical assessment of predicted interactions
CASP	Critical assessment of structural prediction
CBD	Collagen binding domain
CCM	Stem cell derived extract
CD	Circular dichroism
CI	Confidence interval
CKs	Cytokines
Cm	Chloramphenicol
CPR	Centre for proteome research
CRFT	Concentrated reverse flowthrough
Cryo-EM	Cryo electron-microscopy
CS1	CamSol 1
CS2	CamSol 2
CS3	CamSol 3
CS4	CamSol 4
CS5	CamSol 5
CS6	CamSol 6
CVs	Colum Volumes
DMOADs	Disease-modifying OA drugs
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleoside triphosphates
dsDNA	Double-stranded DNA
DSF	Differential scanning fluorimetry
E value	Expect Value
E. coli	*Escherichia coli*
ECM	Extracellular matrix
EDS	Ehlers-Danlos syndrome
EDTA	Ethylenediaminetetraacetic acid
EQTL	Expression quantitative trait loci
ER	Elution reverse
ESCs	Embryonic stem cells
EVs	Extracellular vesicles
FBS	Fetal bovine serum
FGF	Fibroblast growth factor
FPLC	Fast protein liquid chromatography
FT	Flowthrough
FTIX	Flowthrough ion exchange
Fwd	Forward
GAGs	Glycosaminoglycans
GF	Gel filtration
GFs	Growth factors
GWAS	Genome wide analysis study
HA	Hyaluronic acid
HADDOCK	High ambiguity protein-protein driven docking
HCL	Hydrochloric acid
Hom	Homogenisation
hPESCS	Human progenitor endothelial stem cells
HSCs	Haemopoietic stem cells
HTA	HisTrap buffer A
HTB	HisTrap buffer B
Hyp	Hydroxyproline
I	Insoluble
IA-HA	Intraarticular hyaluronic acid
I-ARTIC	Intraarticular
IEC	Ion exchange chromatography
IECA	Ion exchange chromatography buffer A
IECB	Ion exchange chromatography buffer B
Ihh	Indian hedgehog
IL	Interleukin
IMAC	Immobilised metal affinity chromatography
IPSCs	Induced pluripotent stem cells
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITF	Intrinsic tryptophan fluorescence
JA	Joint arthroplasty
Kan	Kanamycin
KGN	Kartogenin
LB	Luria Broth
M1	Mutant 1
M2	Mutant 2
M5	Mutant 5
M8	Mutant 8
MACI	Matrix assisted chondrocyte implantation
MBP	Maltose-binding protein
MEPE	Matrix extracellular phosphoglycoprotein
MHC	Major histocompatibility complex
MK	Marker
MMP	Matrix-metalloproteinase
MMP-13	Matrix-metalloproteinase-13
MMP-14	Matrix-metalloproteinase-14
MMP-2	Matrix-metalloproteinase-2
MMP-3	Matrix-metalloproteinase-3
MMP-8	Matrix-metalloproteinase-8
MMP-9	Matrix-metalloproteinase-9
MQ	MilliQ

MRE	Molar residue ellipticity
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MS	Mass spectrometry
MSCs	Mesenchymal stem cells
MWCO	Molecular weight cut off
NGF	Nerve growth factor
NMR	Nuclear magnetic resonance
NPs	Nanoparticles
NR	Non-reducing
NRMSD	Normalised root mean square deviation
NSAIDs	Non-steroidal anti-inflammatory drugs
O/N	Overnight
OARSI	Osteoarthritis Research society International
OA	Osteoarthritis
OD	Optical density
OPPF	Oxford protein [production facility
PB-MSCs	Peripheral blood derived MSCs
PBS	Phosphate buffered saline
PCL	Polycaprolactone
PCR	Polymerase chain reaction
PDB	Protein Data Bank
PDI	Post dialysis insoluble
PDS	Post dialysis soluble
PEG	Polyethylene glycol
Pel	Pellet
PES	Polyethersulfone
PGs	Proteoglycans
PLDDT	Predicted local distance difference test score
PLGA	Poly(lactide-co-glycolide)
PLLA	poly(l-lactic acid
PNPP	P-Nitrophenyl Phosphate
POI	Protein of interest
POPS	Parameter Optimized Surfaces
PRP	Platelet rich plasma
PTM	Post-translational modification
QMEAN	Qualitative Model Energy Analysis
QSASA	Quotient of SASA
R	Reducing
RO	Reverse osmosis
RFT	Reverse flowthrough
RMSD	Root mean square deviation
RNA	Ribonucleic Acid
RO	Reverse osmosis
ROS	Reactive oxygen species
RT	Room temperature
S	Soluble
SASAs	Solvent accessible surface areas
SC	Subcutaneous
SEC	Size exclusion chromatography
SDS	Sodium dodecyl sulfate
SDS-PAGE	Sodium dodecyl sulfate polyacrylamide gel electrophoresis
	Size-Exclusion Chromatography
SEC-MALS	Coupled to Multi-Angle Light Scattering
SH	Sulfhydryl
SNPs	Single nucleotide polymorphisms
SOB	Super optimal broth
SOC	Super optimal broth with catabolite repression
Son	Sonication
STZ	Superficial/tangential zone
SUMO	Small Ubiquitin like Modifier
Sup	Supernatant
TEAD	TEA domain
Tet	Tetracycline
TGF	Transforming growth factor
TII	Type 2
TIX	Type 9
TJA	Total Joint arthroplasty
Tm	Melting temperature
TVI	Type 4
TX	Type 10
UV	Ultra-violet
VGLL4	Vestigial like family member 4
WT	Wildtype
XRD	X-ray diffraction

1 Introduction

1.1 Joint Anatomy

A joint is the point at which two bones make contact, joints are classified either histologically based upon the dominant type of connective tissue or functionally based upon the amount of movement permitted. Histologically there are three joints in the body: fibrous, cartilaginous, and synovial. Functionally there also three types of joints; synarthrosis (immovable), amphiarthrosis (slightly moveable), and diarthrosis (freely moveable). The two classification schemes also correlate: synarthroses are fibrous, amphiarthroses are cartilaginous, and diarthroses are synovial [1]. Diarthrodial joints include the hip, ankle, elbow, shoulder, and knee. Common structural features shared by diarthrodial joints are shown in **Figure 1**.

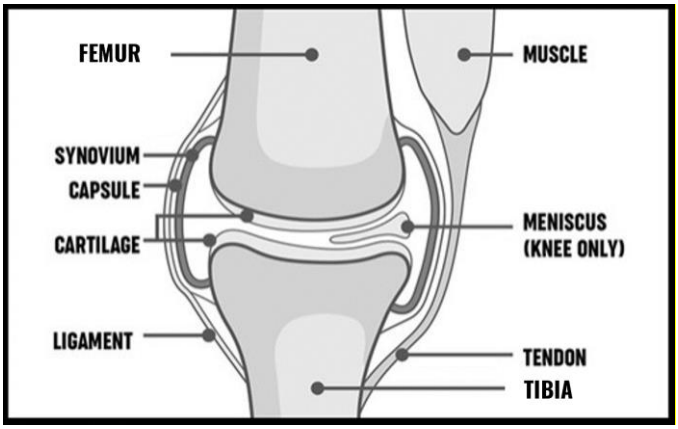


Figure 1: Diarthrodial joint anatomy. A simple representation of the knee joint adapted from [2]. Key features include the synovium, capsule, cartilage, and meniscus.

The Knee joint has been the focus of many studies involving mesenchymal stem cells (MSCs) already. It is the site of primary interest for the therapeutic pursued here. It can be better described as two joints, the tibiofemoral (tibia-femur)

joint and the patellofemoral (patella-femur) joint; these joints work together to achieve flexion, extension, and rotation of the lower legs [3]. The knee is the largest joint in the body and is essential for walking, running, and jumping. It is the only joint to have a meniscus. Menisci are avascular fibrocartilage structures commonly torn in sports trauma. This typically necessitates a meniscectomy a surgical intervention in which the meniscus is completely or partially removed, which leaves an increased risk of osteoarthritis (OA). Recently an alternative to this was developed in the form of a cell bandage made up of a collagen scaffold and undifferentiated autologous MSCs, (Azellon Ltd). The cell bandage is implanted into the meniscal tears at the time of surgical repair instead of the more conventional meniscectomy [4-6]. In a proof-of-concept trial three out of the five patients reported asymptomatically with no re-tear at two years follow-up [5]. Given that the implanted cells were undifferentiated, it was considered that any palliation was a result of trophic effect rather than engraftment, and differentiation [7]. This trial concluded that it is reasonable to repair instead of remove the meniscal tissue, preventing or significantly delaying OA development.

1.2 Cartilage

Cartilage is a connective tissue that has three forms: hyaline, fibrocartilage, and elastic [8-12]. Hyaline means “glassy” and it is this smooth tissue at the ends of articulating bones that is degraded in OA. Its function is to absorb/dissipate the strain of musculoskeletal forces and provide a smooth lubricated surface for ‘gliding’ articular movements. This function gives this type of cartilage its other name of “articular cartilage”. It is a highly specialised (osteochondral unit [13]) connective

tissue, typically 2–4 mm thick in a healthy individual, contingent on its skeletal site [14]. It is composed of four distinct zones (**Figure 2**); the superficial, middle, deep and calcified zones. The superficial (also known as tangential) region forms the smooth articular surface. The middle (also known as transitional) zone is the largest; containing very dense extracellular matrix (ECM), (see **section 1.3**) and is the region containing the highest type II (TII) collagen content. Followed by the deep (also known as radial) zone rich with aggrecan and low in collagen. Then finally the calcified zone which contains the tidemark and interfaces with the subchondral bone.

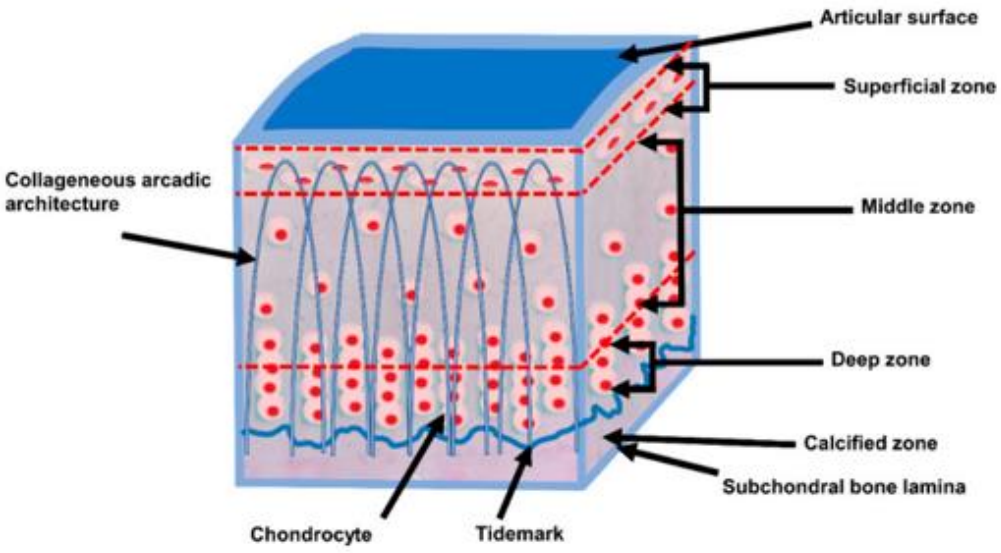


Figure 2: Cross sectional diagram of healthy mature articular cartilage. Taken from [15], outlines the four zones, collagen orientation (arcades of Benninghoff ([16, 17]), chondrocyte (the only cell type found within cartilage) shape density and distribution. Articular cartilage is made up of four distinct regions: Superficial/ tangential zone (STZ), middle zone, deep zone, and calcified zone. **Table 1** below further elaborates upon and outlines the composition and function of each zone.

Table 1 outlines the composition of each of these zones within articular cartilage, describing them and specifying key zonal functions. Cartilage is composed of chondrocytes, water (which make up 75% of its wet weight), a complex

extracellular matrix composed of collagen (90-95% of which is type II collagen [15]), proteoglycans (the most abundant and largest of them being aggrecan), and assorted non-collagenous proteins and glycoproteins. Aggrecan provides cartilage with its osmotic properties critical to its function of dissipating forces through compressive resistance [18-22]. Cartilage is anisotropic and exhibits time (gradual exudation of fluid from proteoglycan units surrounded by collagen networks) [18] and depth dependent property behaviour (depth determined local strain) [23, 24].

Table 1: Details of the zonation found in healthy mature articular cartilage. Zone no., zone name, composition, description, and function. Water concentration decreases from superficial to calcified zones, whilst proteoglycan concentration increases conversely as you progress through the zones [15].

Zone no.	Zone name	Composition	Description	Function	Ref
I	Superficial/ tangential (STZ)	Predominantly TII collagen, some TIX collagen, chondrocytes, zone with lowest proteoglycan content, contains lubricin, lamina splendens, high water content.	Collagen fibers are packed tightly within this zone making a highly cross-linked network. This limits the size of molecules which can penetrate cartilage. Collagen fibers are aligned parallel to the articular surface. Flattened chondrocytes are at a low density.	Protects deeper layers from shear forces, this zone is in contact with synovial fluid and is responsible for the tensile properties of cartilage.	[18]
II	Middle/ transitional	Higher concentration of proteoglycans, a lower water content than the superficial zone. Fewer chondrocytes than STZ, decorin, predominantly TII collagen fibers, some TVI collagen.	Contains thicker collagen fibrils and proteoglycans, collagen is organised obliquely, spherical chondrocytes are present at a low density, Chondrocyte distribution is random.	Provides an anatomic bridge between the superficial and deep zones.	[6]
III	Deep/ radial	Highest proteoglycan content, lowest water content. Aggrecan, collagen fibers, lowest cell density with few chondrocytes.	High proteoglycan content. Collagen fibrils are arranged perpendicular to the articular surface. Largest diameter collagen fibrils found here, in a radial disposition. Highest proteoglycan content, lowest water concentration, chondrocytes are typically arranged in a columnar orientation, parallel to the collagen fibers and perpendicular to the joint line.	Responsible for providing greatest resistance for compressive forces.	[6]
Zone IV	Calcified zone	Rich in TX collagen, hypertrophic chondrocytes.	In this zone, the cell population is scarce.	Integral role in securing the cartilage to bone, by anchoring the collagen fibrils of the deep zone to subchondral bone.	[6]

1.3 Extracellular matrix

The ECM is a complex three-dimensional network of macromolecules, an important feature of articular cartilage composed of three main constituents; water, collagen (organised into fibrils) and proteoglycans (**Figure 3**). There are also other proteins, glycoproteins, and lipids present but in comparably sparse concentrations [18, 25]. The homeostasis and integrity of the ECM is critical to cartilage tissue's function [26]. With advancement though the zones of cartilage the water content decreases from ~80% in the superficial zone to ~65% in the deep zone [18, 27]. Water's main function throughout the ECM of the entire cartilage layer is to hydrate proteoglycans, which along with water molecules themselves expand the collagen network (providing cushioning), lubricate the joint, and aid in the flow of nutrition to the cartilage. Water is in turn maintained within the matrix by the network of proteoglycans and collagens. The water content of articular cartilage generally diminishes over its lifetime but is confirmed to rise to ~90% in OA [28]. An increase in the water content of articular cartilage is one of the changes that leads to a decrease in strength and increase in the permeability of the cartilage layer, ultimately leading to its functional failure [14].

Mechanical damage and/or age-related wear/tear are thought to trigger systematic inflammatory responses in tissues of the joint including articular cartilage, the synovial membrane, subchondral bone, and ligaments [29, 30]. Chondrocytes, the only cell type residing in cartilage, respond to inflammation participating in the catabolism that ultimately leads to the degradation of the cartilaginous ECM in OA [31, 32].

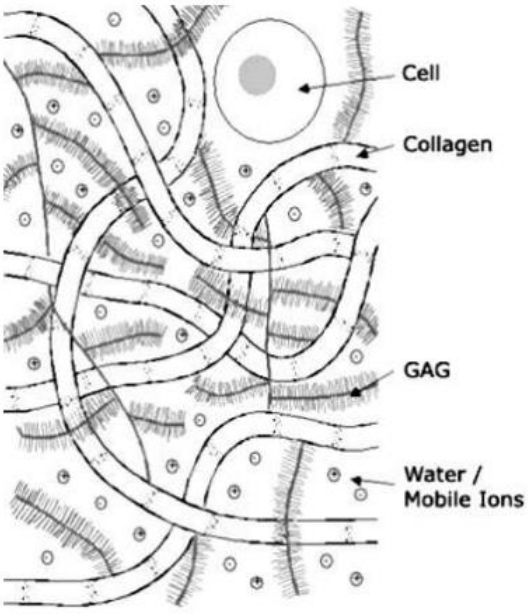


Figure 3: The Articular cartilage (AC) matrix. Contains chondrocytes, collagen fibres, glycosaminoglycans (GAGs) which form proteoglycans(PGs) and water. The properties, abundance, network integrity and relationship of these components determines the mechanical response of the cartilage [15]. The predominant PG in cartilage is aggrecan, comprised of a protein core decorated by GAG side chains consisting mainly of chondroitin sulfate (CS) and keratan sulfate (with a characteristic bottle-brush assembly) [33]. Additionally, hyaluronic acid (HA) induces aggrecan to self-assemble into complexes having up to 100 aggrecan molecules attached to each HA chain backbone [33, 34]. Taken from [35].

1.4 Collagen Structure

TI also known as tendon collagen was the first collagen to be described in 1935 [36] with full structural elucidation achieved in 1954 [37]. TII collagen was then the first collagen to be isolated from articular cartilage, in 1969 [38]. The collagen superfamily is now much larger, composed of 28 different protein species [39]. Collagens are the most abundant proteins within mammals and play important structural roles contributing to the mechanical properties of tissues, as well as to their organization and shaping. Fibrillar collagens share the same basic structural features as each other, each being organised into triple helix structures. They are

composed of three polypeptide alpha chains, rich with glycine, proline and hydroxyproline (Hyp) [40]. These three residues confer stability to the triple helix, which is reinforced by intra-chain hydrogen bonds with the Hyp and electrostatic interactions between aspartate and lysine residues [18, 41]. The helical structure of collagen molecules coupled to their organisation in a fibrillar meshwork provides cartilage with resistance to both shear and tensile forces [15].

An understanding of TII collagen's structure and arrangement within the ECM is key when considering targets accessible intraarticularly within the OA joint. The degradation products that reside abundantly here are an ideal target binding site to home in on. The fact that TII collagen is the most abundant collagen in articular cartilage, representing 80 to 85% of the total makes its degradation products present in abundance at the articular surface a good candidate target for adherence [42, 43]. This project aims to utilise and target attachment to TII gelatin as a means of delivering MSCs where needed most. Allowing time for integration into damaged regions of cartilage specifically. By adhering strongly, integration chances are increased whereupon MSCs could evoke a genuine regeneration of tissue not just trophic symptomatic relief effect.

1.5 Osteoarthritis

OA is the most common chronic joint arthropathy and its prevalence is increasing in first world populations that are now, on average, living significantly longer than previously [44, 45]. OA is characterised by progressively worsening pain, stiffness, and immobility. It has a multi-factorial etiology and can be considered the product of interplay between both systemic and local factors [44].

1.5.1 Epidemiology

Epidemiology is the study of disease in populations, specifically collating and examining information relating to how, why, and when a disease occurs.

1.5.1.1 Prevalence and Incidence

Prevalence and incidence are both valuable, widely utilised epidemiological measures but they each provide distinctly differing information. Prevalence refers to the number of cases of a disease in a specific population at a specific timepoint or during a specified period. Incidence refers to the rate of new cases of a disease in a specific population over a particular period. Prevalence includes all cases (new and pre-existing) whereas incidence is limited to only new cases. A reported measure of OA prevalence, is that in the UK 8.75 million people aged 45 and over sought treatment for OA in 2017 [46]. OA is also reported to affect an estimated 240 million people worldwide [47]. According to The Arthritis Foundation in the UK approximately one in ten adults have symptomatic, clinically diagnosed OA, with the knee being the commonest site, followed by the hip, wrist/hand and ankle/foot [48]. A study published by Swain et al 2020 looked at trends in OA, in UK patients, ≥20 years of age, from 1997 to 2017, utilising a national primary care database. In this publication prevalence is shown to have increased at a rate of 1.4% per year since 1998, then became static in 2008. Incidence however was found to be slowly declining from 2008 [48]. A recent meta-analysis estimates that OA global incidence is 203 per 10,000 person-years in individuals aged 20 and over [49]. This study also reports that there were 86.7 (95% CI, 45.3-141.3) million individuals (20 years and older) with incident knee OA in 2020.

Within the ageing population, prevalence of OA will steadily increase and is expected to be the single greatest cause of disability in the general population by 2030 [59].

1.5.1.2 Risk factors

Local factors include abnormal mechanical load and previous joint injury malignment or disease [51]. All are well-established within the literature as risk factors of OA [12, 44, 52-57]. Systemic factors include age, obesity, gender, occupation type/ sedentary lifestyle and genetics. Despite the large number of people affected with OA, there are limited modifiable risk factors for incident OA [58]. Excess weight loss, occupation, sport, injury prevention and joint alignment are amongst the limited modifiable risk factors, perhaps the largest risk factor however, ageing, and cellular senescence, in cartilage is something that cannot be avoided.

In recent years it has become more widely accepted that underlying joint shape is a strong risk factor for OA e.g. pathologic developmental abnormalities of the hip seen in Perthes disease give a predisposition to hip OA development [50]. This realisation has led to the use of surgical interventions to ameliorate OA risk by restoring the joint to a more anatomically normal shape. In contrast patients with certain genetic traits or inheritable conditions, such as Ehlers-Danlos syndrome (EDS), may have an increased risk of developing OA [51, 52]. EDS patients have deficient collagen levels, in connective tissues, meaning their ability to support muscles and joints is impaired. Which can lead to unstable and hypermobile joints that may contribute to OA [50].

1.5.1.2.1 Genetics

OA has been estimated based on studies with monozygotic versus dizygotic twins to have a heritability of between 40% (knee) [51] and 60% (hip) [52]. It has a complex non-Mendelian inheritance pattern consistent with its multifactorial nature and the phenotype heterogeneity amongst patient populations .

Table 2 outlines some of the genetic loci known to be implicated in OA risk and development. OA is polygenic, a recent OARSI review article lists 124 single nucleotide polymorphisms (SNPs) across 95 independent loci in the human genome associated with OA [52]. It is a polygenic condition, whereby multiple OA risk alleles contribute in a liability threshold model, whereby the cumulative effect means alone a gene doesn't cause OA but together with risk exerted by others the impact is that once over a threshold a person will develop OA [53[53]]. There are two ways genetics influence phenotype. Firstly by causing a change in the amino acid sequence of a protein, which implicates protein function. Or by altering the regulation of gene expression termed an expression quantitative trait loci (EQTL). Only 10% of the 124 SNPS in the OSARI review are found to change protein coding sequence, meaning it is a complex condition with multiple different heritable risks [54].

Effector genes with SNPs are thought to have a role in OAs complex heterogeneous pathogeny [53]. High confidence genes are of particular interest as potential targets for drug intervention. With most effector genes associated with skeletal development, joint degeneration, neuronal function, and development (e.g. ALDH1A2 [55], BHLHA9 [57], C2orf40 [57] and ERF [56]). Also, a minority of these effector genes are involved in adipogenesis (e.g., FTO [56]), muscle function

(CHRM2 [56]), the immune response and inflammation (e.g. APOE [56]) .Several genes involved in signalling pathways have also been implicated in OA predisposition (see **section 1.5.2** for more details of how awry signalling causes OA symptoms). Vestigial like family member 4 (VGLL4 [52]) is one such gene involved in the interactions of TEA domain (TEAD [54]) transcription factors [53]. Changes in TGF-β, Wnt/β-catenin, Indian Hedgehog (Ihh), Notch and fibroblast growth factor (FGF) pathways have been shown to contribute to OA development and progression by primarily inducing catabolic responses in chondrocytes [55].

Table 2: OA susceptibility established genetic loci summary. Human chromosome, gene, abbreviation, SNP reference and literature reference

Abbreviation	Gene	Chromosome	SNP reference	Ref
ALDH1A2	Aldehyde dehydrogenase 1 family, member A2	15	rs12915901	[56]
APOE	Apolipoprotein E	19	rs8112559	[57]
BHLHA9	Basic helix-loop-helix family member A9	17	rs216175	[58]
BMP5	Bone Morphogenetic Protein 5	6	rs921126	[59]
C2orf40	Chromosome 2 open reading frame 40	2	rs66989638	[58]
CHRM2	Cholinergic Receptor Muscarinic 2	7	rs571734653	[57]
CHST3	Carbohydrate sulfotransferase 3	10	rs3740129	[60]
COL2A1	Collagen Type II Alpha 1	1	rs143383	[61]
COL6A6	Collagen type VI alpha 6	3	rs200274210	[57]
COL9A1	Collagen Type IX Alpha 1	6	rs148350640	[57]
COLGALT2	Collagen Beta(1-O)Galactosyltransferase 2	1	rs11583641	[62]
CRHR1	Corticotropin Releasing Hormone Receptor 1	17	rs62063281	[63]
CTSK	Cathepsin K	1	1:150214028 _CT_C	[63]
DIABLO	Diablo IAP-Binding Mitochondrial Protein	12	rs11059094	[63]
DPEP1	Dipeptidase 1	16	rs1126464	[63]
DUS4L	Dihydrouridine Synthase 4 Like	7	rs3815148	[57]
ERF	ETS2 Repressor Factor	19	rs75621460	[57]
FGF18	Fibroblast growth factor 18	5	rs3884606	[63]
FIGNL1	Fidgetin-Like Protein 1	7	rs200453649	[57]
FTO	Fat mass and obesity associated	16	rs9940278	[57]
GDF5	Growth differentiation factor 5	20	rs143383	[56]
GSDMC	Gasdermin C	8	rs4733724	[64]
GNL3	Guanine nucleotide-binding protein-like 3	3	rs6976	[65]
HSPG2	Heparan Sulfate Proteoglycan 2	1	rs199899258	[57]

IL11	Interleukin 11	19	rs4252548	[52]
ITGA8	Integrin Subunit Alpha 8	10	rs371802080	[57]
MAPT	Microtubule-associated protein tau	17	rs62063281	[63]
MCF2L	Guanine Nucleotide Exchange Factor	13	rs11842874	[66]
MOB3B	MOB Kinase Activator 3B	9	rs116882138	[67]
NOTCH2	Neurogenic locus notch homolog protein 2	1	/	[68]
NLRP6	Pyrin domain containing 6	11	rs373174851	[57]
PXN	Paxillin	12	rs371118243	[57]
RWDD2B	RWD Domain Containing 2B	21	rs6516886	[54]
SELP	Selectin P	1	/	[57]
SLBP	Stem-loop binding protein	4	rs11732213	[69]
SMAD3	Smad family member 3	15	rs12901499	[59]
SMO	Smoothened, frizzled class receptor	7	rs143083812	[52]
SPN	Sialophorin	16	rs200681097	[57]
SUSD5	Sushi Domain Containing 5	3	rs377664152	[57]
TEAD1	TEA domain family member 1	11	/	[60]
TGFA	Transforming Growth Factor Alpha	2	rs2862851	[70]
TGFB1	Transforming Growth Factor Beta 1	19	rs75621460	[63]
TMEM241	Transmembrane Protein 241	18	rs10502437	[71]
TNRC6B	Trinucleotide repeat containing 6B	22	rs201057205	[57]
USP36	Ubiquitin specific peptidase 36	17	rs112843316	[57]
VGLL4	Vestigial Like Family Member 4	3	rs2276749	[52]

Genetic differences have been identified between weight bearing (knee, hip, and spine) and non-weight bearing joints (hand, finger, and thumb) in a recent genome wide analysis study (GWAS) [60]. OA development appears to be a result of a complex set of interactions between mechanical, biological, biochemical, and molecular factors that destabilise the normal coupling of degradation and synthesis of articular cartilage, chondrocytes, ECM and subchondral bone [72].

Genomics alone is unlikely to be able to lead to rollout of means to reliably identify individuals who will develop OA, but it might reveal new insights into disease pathogenesis particularly in individual joints. SNPs have been associated with several known risk factors, including hip shape, body-mass index, and bone mineral density which can be good ways to identify those more at risk [73].

1.5.2 Etiology, cellular and molecular changes seen in OA

The etiology of the condition involves the progressive degradation of articular cartilage in diarthrodial joints, as well as changes to the sub-chondral bone, creating an inflammatory microenvironment with overall decreasing tissue functionality as the disease progresses. It is a degenerative disease affecting the entire joint [74]. **Figure 4** shows some of the pathologies seen in OA versus healthy cartilage.

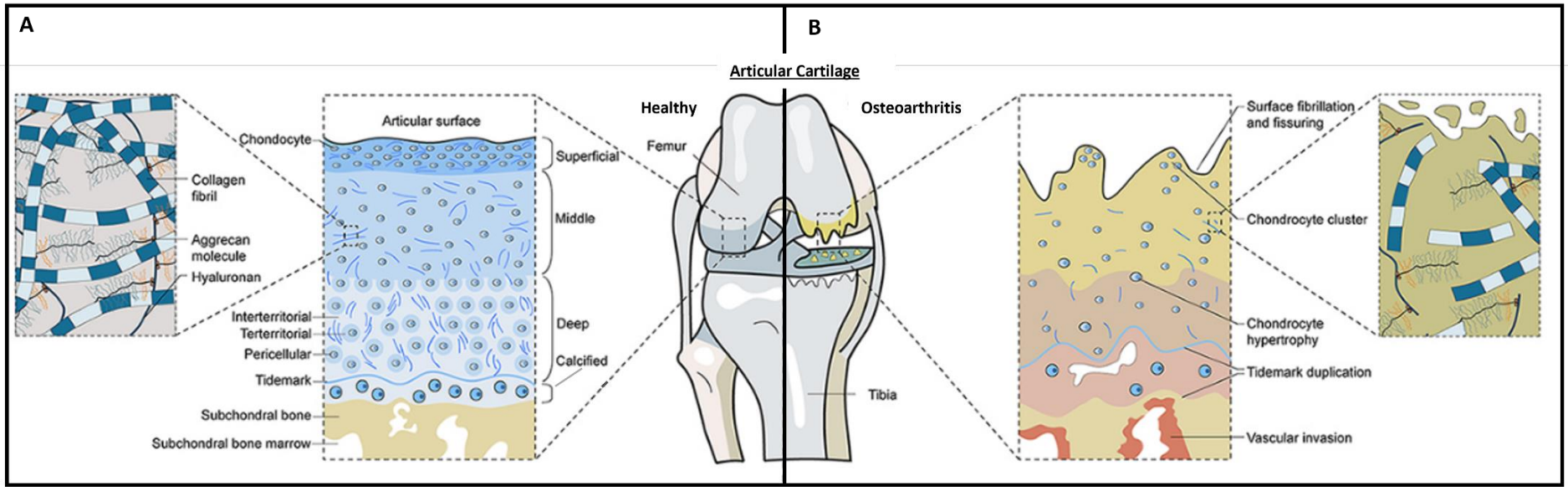


Figure 4: Schematic of articular cartilage and the major ECM components outlining differences between healthy (A) and OA (B) joint cartilage. The articular surface has four main structural elements labelled in **(A)** (as described in detail in **section 1.2**. The only cells present in articular cartilage are chondrocytes, present in varying densities dependent on the zone of articular cartilage. Middle and deep zones varying, water, protein content and collagen. Calcified cartilage, and sub-chondral bone. Within OA tissue **(B)** there are some typical presenting pathologies seen namely; fissuring and fragmentation of the articular cartilage, both chondrocyte proliferation and hypertrophy, duplication and advancement of the tidemark, expansion of the calcified cartilage zone, vascular invasion of the bone and calcified cartilage, subchondral bone thickening, subchondral bone sclerosis, osteophyte formation (osteophytosis), cartilage erosion, meniscal damage, ligament tears, synovitis, thickening of the joint capsule (swelling). All these pathologies contribute to the chronic pain, hallmark progressive joint space narrowing (JSN) and loss of joint function seen in OA. Adapted from [75].

Table 3 shows how joint histopathology progresses from healthy and intact to early and then late stage OA.

Table 3: OA progression from healthy to late/advanced. Including Disease stage/ duration, OARSI grade, description and histopathology [76] [77].

OA duration	OARSI Grade	Description	Histopathology
/	0	Healthy Intact cartilage	Smooth surface, intact, normal tissue architecture
Early	1	Surface Intact	Mild surface abrasion
			Focal or generalised swelling/ oedema
			Evidence in superficial zone of some chondrocyte death, proliferation (clusters) and hypertrophy
			Disorientation of chondron columns
	2	Surface discontinuity	Loss of small portions of superficial ECM parallel to the surface
			Deep fibrillation may be evident, whereby ECM cracks extend completely through the superficial zone
			ECM staining depletion may be present through the superficial zone
			Focal increased staining around the chondrons may be present
	3	Vertical fissures (clefts)	Vertical ECM fissures present that extend into the mid zone
			Branched fissures start to be seen
			Cationic stain depletion into deep zone
Late	4	Erosion	Mid zone excavation
			Delamination
			Mid zone cyst formation can be seen
	5	Denudation	Complete erosion of hyaline cartilage to level of mineralized cartilage or bone
			Some fibrocartilage repair extending from the surface may be evident
			Sclerotic bone
			Bone surface intact
	6	Deformation	Bone remodelling and Osteophyte formation seen
			Change in contour of articular surface
			Articular plate fractures
			Microfracture evident

The most prominent presenting clinical symptoms of OA: joint pain and stiffness do not always correspond with changes in the joints that can be seen on radiograph [30, 40, 41]. However as imaging capabilities increase in sensitivity, changes in joints indicative of OA are identified much earlier and more easily and as

a result more frequently [42-44]. With earlier diagnosis there is greater potential for early intervention before pain and disability have had time to progress to severe/ profound. Worldwide however, it has been observed that there is considerable discordance seen between reported knee pain in patients and perceptible radiographic knee OA evidence [52] adding further complication to examination of the epidemiology of OA. Although OA can develop in any synovial joint it is most diagnosed in the knees, which must with stand extreme stresses, twists, and turns throughout an individual's lifespan [12, 48, 78-80]. For this reason, the knee has been the focus of much OA research and this project focusses on a therapeutic approach that would be ideal for treating knee OA but could readily be utilised for treatment in other joints

1.5.2.1 Cell signalling

Cartilage architecture and biochemical composition are regulated by chondrocytes, which react to changes in both chemical and mechanical environment [18, 81, 82]. Producing several inflammatory response proteins, known as cytokines (namely interleukin 1β, interleukin 6, tumour necrosis factor (TNF) α, and matrix-degrading enzymes including the metalloproteinases and a disintegrin and metalloproteinase with thrombospondin-like motifs (ADAMTS) [83]. **Figure 5** shows some of the signalling pathways that undergo changes in OA whilst reiterating some of the structural changes/ pathologies seen in OA joint cartilage (refer back to **Figure 4B)**.

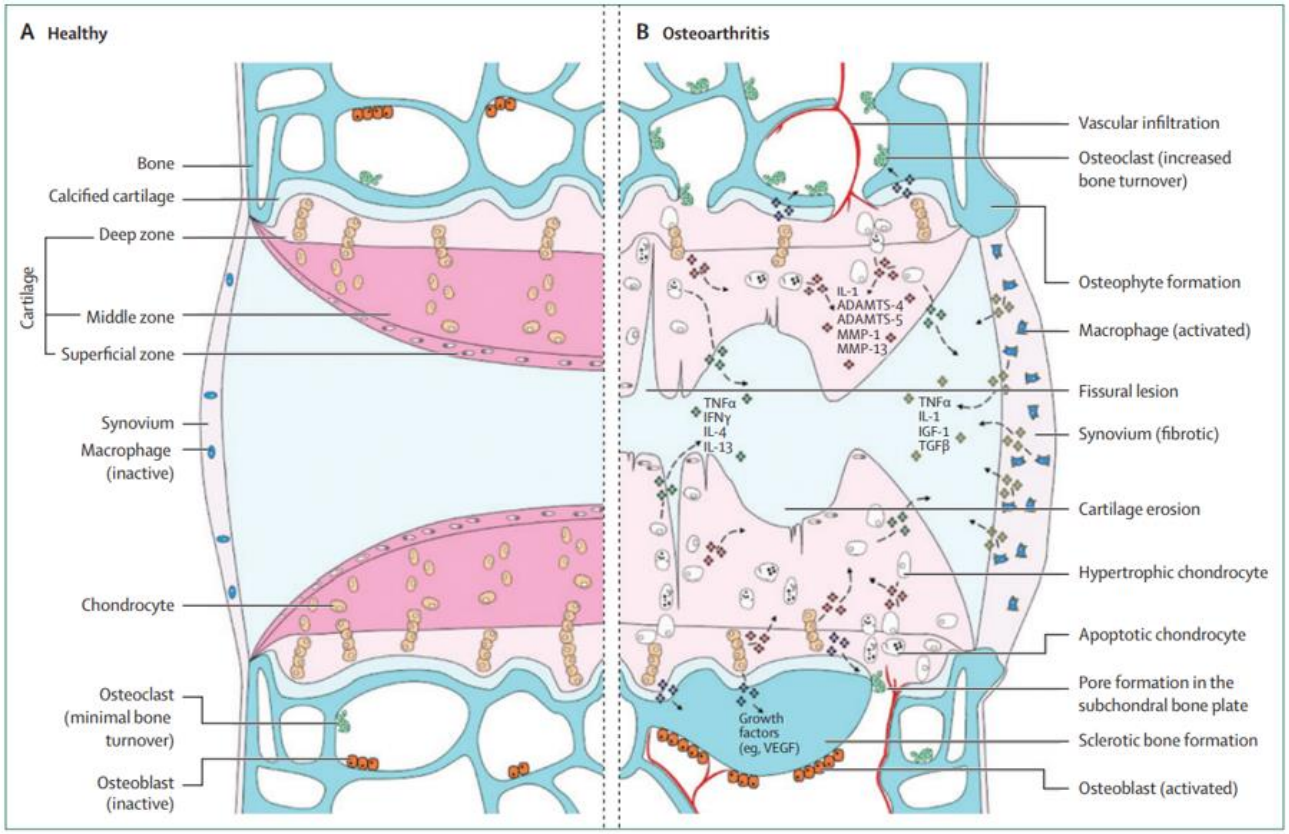


Figure 5: Summary representation of some of the signalling pathways and structural changes involved in OA development. Healthy **(A)** and OA **(B)** joint cartilage. ADAMTS=a disintegrin and metalloproteinase with thrombospondin-like motifs. IL=interleukin. MMP=matrix metalloproteinase. TNF=tumour necrosis factor. IFN=interferon. IGF=insulin-like growth factor. TGF=transforming growth factor. VEGF=vascular endothelial growth factor. Taken from [83].

Chondrocytes express many toll-like receptors [84], which are activated by damage-associated molecular patterns triggered during OA development. Consisting of extracellular matrix molecules (specifically hyaluronan [85], calcium pyrophosphate and sodium urate crystals have all been shown to bind chondrocyte toll-like receptors and might also therefore play a part in the aetiology of osteoarthritis [86]).

Expression and activation of complement has also been observed to be abnormally high in OA joints [87]. Cartilage oligomeric matrix protein (COMP) is a potent activator of the alternative complement pathway [88] whereas proteoglycans i.e. fibromodulin target the classic pathway [89].Chondrocytes also express receptors for advanced glycation end products (RAGE), which bind these products that accumulate in ageing joint tissue [90]. Stimulation of RAGE signaling can cause MAP kinase pathway activation and increased NF-κB activity. Resulting in a phenotypic shift towards catabolism perhaps partially explaining the increasing prevalence of osteoarthritis with age [91]. This myriad of responses to extracellular matrix components might simply reflect amplification of established cartilage degradation. Alternatively there is now evidence that chondrocytes could initially be activated by inflammatory signals originating from other joint structures i.e. synovium or subchondral bone [92, 93].

Figure 6 summarises how proteases and cytokines act together within the joint to cause OA pathology. Both the degeneration and remodelling of tissues in the joint require activity of several different proteases.

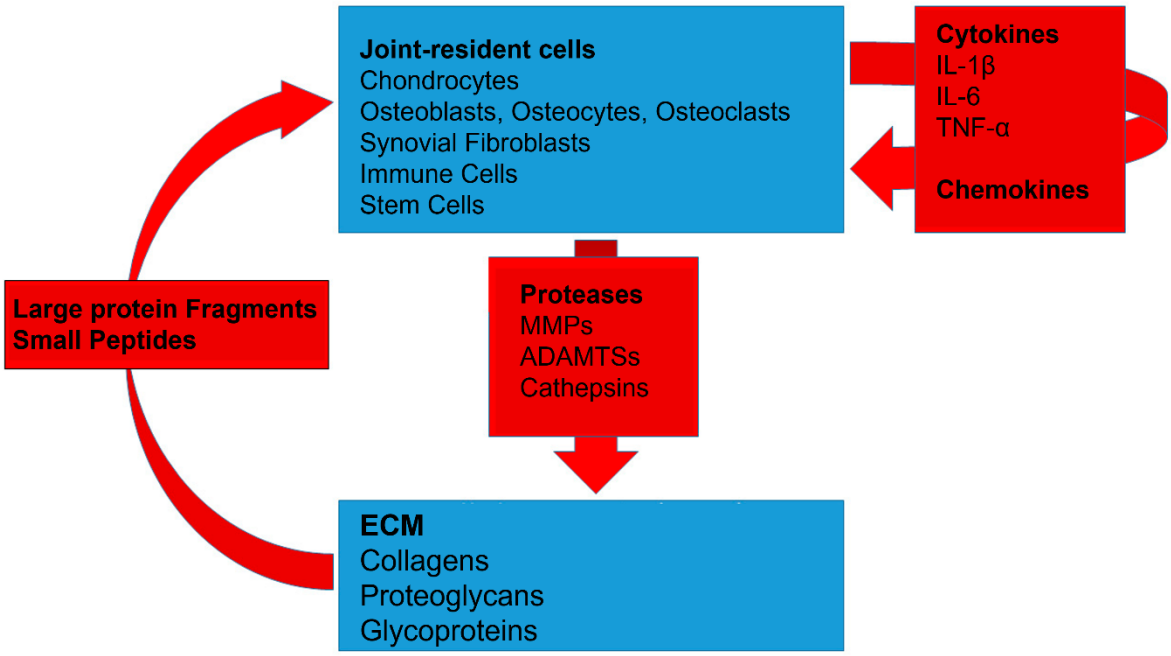


Figure 6: Proteases and cytokines in the joint involved in OA pathogenesis. Taken from [80].

The signalling involved in OAs pathology and development is clearly very complex, further study and elucidation is warranted, as therapeutic interventions are more likely to be effective when acting early rather than late in the disease process.

1.5.3 Treatment

1.5.3.1 Difficulties with treating OA

There are few treatment options available for OA and only limited pain management strategies. As a result, OA represents a significant burden to both society and healthcare services globally [79, 94, 95]. Articular cartilage is a uniquely isolated tissue, it is avascular so when it is damaged as in OA, none of the body's normal inflammatory and reparative processes are able to migrate into the articular surfaces to assist in repair processes [44, 96-102]. The only cells present in articular

cartilage are chondrocytes. They can produce all the major proteins of the cartilage ECM and so have some capacity to repair the tissue. However, because they are embedded within lacunae in the ECM, they are restricted in movement so unable to migrate from healthy to damaged regions.

In OA, with the loss of healthy cartilage, bones no longer move easily against each other and are subject to greater forces. Cartilage in affected joints loses its low coefficient of friction, imperative to its functioning. This may contribute to the chronic joint pain that is a hallmark feature of OA [55, 74, 103, 104], although the exact mechanisms leading to pain remain unconfirmed. There is evidence that in degrading OA cartilage some of the chondrocytes undergo apoptosis, reducing further the capacity for regeneration of the tissue, leading to a gradual loss of function [105-113]. As OA progresses to the later more severe and final stages, surgery such as joint arthroplasty (JA) (**see section 1.5.3.2.4**) is the only means of restoring function [114].

1.5.3.2 Surgical OA interventions

Articular cartilage defects lead to progressive change, and they become irreversible if no intervention is applied. Therefore, early treatment to prevent OA is now the favoured modality where it is feasible. Many techniques have been trialled and adopted to achieve this including abrasion, drilling, tissue autografts, allografts, and cell transplantation. Due to our increased understanding of cartilage biology and its pathologies, along with advances in imaging and arthroscopy capabilities treating chondral defects is much more popular and widely pursued [115]. However, these surgical interventions are only appropriate for treatment of focal cartilage defects to

prevent onset of OA. They cannot be used to intervene during the far more common, generalised idiopathic OA, where there is no obvious initiating injury.

1.5.3.2.1 Microfracture

Microfracture is a surgical management technique employed in early isolated full thickness articular cartilage focal defects of the knee [115, 116]. These lesions have the potential to progress into larger and higher-grade disease, ultimately progressing to OA. Microfracture is a marrow stimulating technique, involving subchondral bone perforation to recruit MSCs to a cartilage defect [117-119]. It is a technically simple procedure, considered generally safe, and cost-effective compared to other interventions, although the published clinical outcomes are limited to predominantly young patients with traumatic defects [120]. In a clinical study from 2006 comparing meniscectomy alone to microfracture with meniscetomy, it was reported that microfracture did not confer any additional benefit to meniscectomy alone [116]. There have been some mixed short- to mid-term clinical and radiographic improvements reported in some studies [121, 122]. But there have also been some concerns raised regarding longevity, suboptimal repair, fibrocartilage infill, subchondral osseous overgrowth, and deterioration of clinical improvement over long-term follow-up [118, 119, 122-124].

A recent advance in the field is in the identification of candidate synovial biomarkers that have been shown to predict patient outcome after microfracture. Two biomarkers (YWHAQ and LYVE-1) were identified as differentially abundant between the clinical responders/improvers and nonresponders after microfracture. Lower activity levels of A disintegrin and metalloproteinase with thrombospondin

motifs-4 (ADAMTS-4) were also identified in the synovial fluid preoperatively, which was identified as a predictor of better patient-reported knee function after microfracture [125]. Applied in the clinic, these could be screened for, to identify patients suited and more likely to benefit from the intervention.

1.5.3.2.2 Autologous chondrocyte implantation

Autologous chondrocyte implantation (ACI) was first explored and implemented clinically to repair cartilage lesions and circumvent OA development in the 1980s [126]. With the first patient treated in 1987 [7], the results of this clinical trial reported fully in 1999 [127]. This was a landmark and significant study, making a critical contribution and step forward for the application of cell biology to the treatment of cartilage lesions with the aim of preventing onset of OA. ACI first involves the biopsy of healthy tissue during an initial arthroscopy procedure. Chondrocytes from this biopsy are then expanded through *in vitro* culture, to reach high densities upon which they are re-implanted during a second arthroscopy procedure and covered with a flap of periosteal tissue or, more recently, with a biodegradable membrane. ACI has been shown to provide clinical benefits with significantly improved Lysholm scores [128, 129]. Which is a commonly used measure of knee instability employed by physicians to evaluate the progression of OA in pre-clinical and clinical trials postoperatively [130]. ACI was initially only implemented following microfracture failure and is generally considered superior but it is the more expensive of the two techniques due to the cell culture requirement [131]. However, in many countries it is now the first choice rather than being used after microfracture [132].

Since its implementation, advancement of this modality has emerged in the form of matrix assisted chondrocyte implementation (MACI), utilising a TI/TII chondrocyte seeded collagen scaffold [133]. This seeding of chondrocytes onto the scaffold is undertaken three days before implantation in an attempt to prevent cell de-differentiation[126]. The matrix scaffold introduction aimed to promote chondrocyte infiltration on one side and lubrication on the other to facilitate the distribution of cells more widely within the defect site [134]. At one year follow up superiority of MACI over ACI was not seen, both treatments resulted in comparable improvement in clinical, arthroscopic, and histological outcome measures . The mean Cincinnati knee score a common clinical measure derived from a patient functional outcome questionnaire, where higher scores indicate better function was utilised. At one year follow up the score increased by 17.6 in the ACI group and 19.6 in the MACI group ($p = 0.32$) [135]. There was no significant difference seen arthroscopically in appearance of the graft after both ACI and MACI. Histological biopsy results showed hyaline-like cartilage or hyaline-like cartilage with fibrocartilage in six of the 14 (42.9%) ACI graft biopsies and four of the 11 (36.4%) MACI biopsies [135]. A five year follow up of MACI versus microfracture with a larger sample size (65 MACI and 63 microfracture treated patients), highlights MACI's superiority. Comparing Knee injury and Osteoarthritis Outcome Score (KOOS), a score developed to incorporate patients' opinion about their knee and associated issues [106], improvements were maintained over the five years and statistically significant ($p=0.022$). However, in this study Magnetic Resonance Imaging (MRI) evaluation showed no significant structural changes, tissue regeneration or repair [136].

Interestingly in a 20 year follow up of first-generation ACI, 15 out of 24 knees had retained grafts, and demonstrated significant improvement according to the modified Cincinnati knee score and this improvement was sustained to 20 years. This study used several other scores including the Western Ontario and McMaster Universities Arthritis Index (WOMAC) score of pain, stiffness, and physical function. WOMAC is a disease-specific tool used for evaluating condition of patients with hip or knee OA, with lower scores indicating lower levels of symptoms and/or physical disability. However, improvements in WOMAC reported here lacked significance, with high preoperative scores potentially preventing significant improvement measures being attained post-operatively. Shortform-36 (SF-36) another widely utilised health-related quality-of-life measure was also reported, where higher scores represent better health and condition. SF-36 outputs two different summary scores whilst one; physical component summary (PCS) improved significantly; the other mental component summary (MCS) did not attain significant improvement (except at 10 years follow-up). Consistent with a study from 2013 that showed that the SF-36 MCS was the least responsive summary outcome score of cartilage repair [137]. This was the first report of follow-up data of such longevity, and it provides an important standard for comparison as newer-generation ACI techniques are developed [138].

1.5.3.2.3 Osteochondral autografts and allografts

Graft tissue implantation are a popular intervention, employed to repair chondral and osteochondral defects. Cells harvested autologously or as allografts from healthy explanted regions of donor tissue are isolated (often from the patellar or

posterior femoral condyle), expanded *in vitro,* then implanted into cartilage defects in patients [139]. Issues with anatomical mismatch, and non-congruence with the surrounding tissue, initially limited the adoption of such procedures [123]. However, these were addressed by Hangody and Bobic, through the development of a method utilising multiple cylinder osteochondral grafts concomitantly (a technique known as mosaicplasty) [140]. Osseous integration was seen with this method, but the repaired tissue was determined to be fibrocartilage [139]. A subpar outcome as fibrocartilage is not equipped to withstand the repeated variable forces that the knee joint is subject to, thereby creating instability. This has imposed restrictions on the osteochondral defect cases for which this is considered a viable treatment option. Success is highly dependent on the patient demographics; size of defect, defect site, age, return to sport, generally it has been found that smaller the defect better the outcome [141]. A shorter postoperative recovery period compared to ACI make it a sometime preferable option.

Fresh osteochondral allografts are now commercially available from tissue banks, which has accelerated their more widespread use [142]. Tissue allografts however bring with them the issue of immunogenicity being ‘foreign’ donor tissue they can trigger an immune response or transmit disease [143].

1.5.3.2.4 Joint arthroplasty

JA is the surgical replacement of joints; it is an invasive procedure with serious associated risks including infection (requiring implant revision) and pulmonary embolism. So, it is considered that JA is the last line of treatment, aimed at resolving the chronic pain and limited mobility associated with OA. A total or

partial (for isolated compartmental OA) replacement is a common end stage treatment, for OA unmanaged by alternate interventions. A partial JA involves only replacing the affected part of the knee (the medial compartment) with an artificial surface (a metal, plastic, or ceramic device known as a prostheses) built to mimic the natural structure of the joint articular surfaces. Whereas in a total JA (TJA) the entire damaged region of bone is removed and replaced. Just like natural joint structures prostheses have a durability recommended lifespan to consider due to the expected wear and fatigue of the materials. A systematic review and meta-analysis from 2019 showed that approximately 82% of TJA protheses last for 25 years [144].

Hips and knees are the most replaced joints, data from the National Joint Registry for England, Wales, and Northern Ireland (NJR) show that the number of primary TJA performed is still increasing with 101,384 hip and 108,506 knee primary TJA procedures performed in the 12 months to April 2019. Projections of these figures for 2035 taking into consideration population risk factor changes (i.e. BMI and gender) are 95,877 hip and 118,666 knee TJA procedures [145]. Therefore, a less invasive treatment to restore function, negating or substantially delaying the requirement for surgery, would be a great strive forward for patients with OA.

Ultimately there is a trade-off between the currently available OA surgical interventions with the efficacy, risks and therapeutic longevity of each, important considerations for patients and surgeons to be aware of. Failure of one or a combination of them in sequence ultimately leads a patient to a TJA. TJA is the last resort solution often considered more suitable only for older patients (>60 years) [28].

1.5.3.3 Drugs and intra-articular injections

Current treatments that aim to reduce damage and repair the OA joint itself, result in inferior tissue regeneration with comparably poor integrity and mechanical load bearing functionality restoration compared to the cartilage they are aiming to restore [146]. Pain relief has been reported in some trials, but these are short-lived requiring repeat therapeutic administration. The holy grail in terms of OA treatment is a treatment that relieves pain and inflammation, is targeted to damaged joint tissue and retained where needed to allow integration, repair, and potentially sustained improvements.

1.5.3.3.1 Non-Steroidal Anti-Inflammatory Drugs, lifestyle modification, corticoids, and visco-supplementation

The first line of treatment for mild to moderate OA features predominantly pain management and lifestyle modification e.g. losing weight [147] or changing to less intense sport such as swimming which is non weight bearing [148]. Analgesics utilised include non-steroidal Anti-inflammatory drugs (NSAIDs) such as diclofenac, ibuprofen and naproxen, acetaminophen (paracetamol) [149, 150] and opioids. Opioids are reserved for the more severe cases and are often used as a bridging management option whilst awaiting surgical intervention [151]. Analgesics only manage the consequences of OA, without addressing the underlying causes.

Intra-articular administration of corticoids are another early stage OA treatment offering moderate relief that has been used for a long time clinically, unfortunately efficacy reduces over time and with repeated use [152]. Visco-supplementation is another option for pain relief in OA, such as intra-articular

injection of hyaluronic acid (IA-HA) [153]. A recent meta-analysis presented strong evidence that IA-HA leads to a small decrease in pain compared to placebo [154].

1.5.3.3.2 Disease modifying drugs

Currently available pharmacological OA treatments mostly aim to relieve the symptoms associated with inflammation and pain. However, with our increasing understanding of OA pathology, several new therapeutic targets have been identified [155-157], enabling the development of a number of potential new drugs including some candidate disease-modifying OA drugs (DMOADs) [158]**,** **Table 4**. A DMOAD is a drug that would inhibit or even reverse the progression of OA, a highly sought-after classification for many new OA drugs under development. There are currently no approved DMOADs available, with treatments limited to analgesia for early-stage disease and surgical intervention for late-stage disease [159]. The development of effective drugs to treat OA is thus of utmost importance. Most research in this area is focused on cartilage and aims to identify approaches for either stopping cartilage degradation or promoting cartilage repair. There has been a recent influx of promising candidate OA therapeutics, not only seeking to alleviate pain in patients but to reduce progression and repair damage.

Invossa is a particularly promising *ex vivo* gene therapy being trialled for OA, utilising allogeneic chondrocytes transformed and transduced with a retrovirus expressing Transforming growth factor beta 1 (TGF-β1), a growth factor which is known to be involved in cartilage development and maturation [160, 161]. Results of a phase 3 clinical trial showed Invossa acts through paracrine effects, which have a beneficial effect against the underlying pathogenic mechanisms of OA, modifying the

micro-environment of the joint to be amenable for regeneration, with reduced pain and improved function reported [162, 163]. Early results from this trial seem to show significant structural improvement, as cartilage thickness was increased assessed via MRI. Further studies are subsequently required to probe this and explore whether it is correct to classify this therapy as the first DMOAD. At this point some issues have since been encountered regarding the cell type and possible cell contamination. Issues that raise concerns and need investigating before licensing and regulatory bodies will allow reinstatement and progression of clinical studies with Invossa [164].

OA is particularly suited for gene therapy as a local strategy, as the joint is a relatively contained site, with no extra-articular or systemic components to consider. Combinatorial and inducible strategies are also emerging which are an interesting approach to addressing OAs multifactorial complexity, pathology and etiology [165].

Table 4: Clinical trials summary of drugs including candidate disease modifying OA drugs (DMOADs). Adapted from [47, 166, 167] and modified to cover all current relevant trials (July 2022)**.** Treatment/ drug name, trial ID, manufacturer, structure, route of administration (I-ARTIC (Intraarticular), SC (Sub-cutaneous) or oral), targeted tissue, mechanism of action, stage of development, study outcome and references.

Treatment/ drug name	ClinialTrials.gov ID	Manufacturer	Structure	Route of admin	Targeted tissue	Mechanism of action	Stage of development	Study outcome	Ref
Sprifermin	NCT01919164	Merck KGaA (Darmstadt, Germany)	Recombinant human FGF-18	I-ARTIC	Cartilage regeneration and repair	Recombinant human, fibroblast growth factor-18 (FGF-18), potential DMOAD.	Phase 2 completed	Increased cartilage thickness, and substantially reduced cartilage loss.	[168-170]
Platelet rich plasma (PRP)	NCT03491761	Northshore University HealthSystem (Evanson, IL, USA)	Human PRP from patient whole blood samples	I-ARTIC	Cartilage regeneration and repair	Directs local mesenchymal stem cells (MSCs) to migrate, divide and increase collagen matrix synthesis.	Phase 2 in progress	Still ongoing but, first results show PRP did not result in a significant difference in symptoms or joint structure at 12 months. These findings do not support use of PRP for the management of knee OA.	[171, 172]

Invossa TissueGene-C (TG-C)	NCT03291470 NCT03203330	TissueGen, Inc (Duncansville, PN, USA)	Allogeneic human chondrocytes modified to express Transforming growth factor-β1 (TGF-β1)	I-ARTIC	Cartilage regeneration and repair	Cell and gene therapy composed of non-transformed and transduced chondrocytes (3:1). Retrovirally transduced to overexpress TGF-β1. Shown to possess immunosuppressive and anti-inflammatory properties by regulation inflammatory cytokines release [173, 174]. Additionally, it positively regulates chondrocyte proliferation, differentiation, ECM synthesis and deposition [175-177].	Phase 3 completed	Results suggest TG-C exerts a beneficial effect on OA by inducing a M2 macrophage-dominant micro-environment. Cell therapy using TG-C may be a promising strategy for targeting the underlying pathogenic mechanisms of OA, reducing pain, improving function, and creating a pro-anabolic micro-environment. This environment supports cartilage structure regeneration and warrants subsequent evaluation in future clinical trials.	[178, 179]
KA34	NCT03133676	Calibr, a division of Scripps Research (La Jolla, CA, USA)	Analog of Kartogenin with improved potency and chemical stability	I-ARTIC	Cartilage regeneration and repair	KA34 exhibits chondrogenic activities. Improving OA outcomes. Induces a 2-4-fold increase in mRNA expression of chondrogenic genes (SOX9, PRG4 and COMP).	Phase 1 completed	Preclinical Results characterized KA34 as a novel and safe OA drug candidate with disease modifying, cartilage regenerative and pain modulating activities. Phase 1 trial Results not yet published.	

Lorecivivint (LOR/ SM04690)	NCT03122860 (Phase II) NCT03928184 (Phase III)	Samumed LLC (San Diego, CA, USA)	N-(5-(3-(7-(3-Fluorophenyl)-3H-imidazo[4,5-c]pyridin-2-yl)-1H-indazol-5-yl)pyridin-3-yl)-3-methylbutanamide	I-ARTIC	Cartilage catabolism	Wnt/β-catenin signalling pathway inhibition. Inhibition of the intranuclear kinases CDC-like kinase 2 (CLK2) and dual-specificity tyrosine phosphorylation-regulated kinase 1A (DYRK1A).	Phase 2 completed Phase III now ongoing	Demonstrated the efficacy of LOR on pain scores (Western Ontario and McMaster Universities Arthritis Index, WOMAC score) and maintained radiographic joint space width in knee OA subjects. The optimal dose for future studies was identified (0.07 mg LOR). Larger and longer studies are now needed to further assess efficacy.	[180, 181]
Teriparatide	NCT03072147	University of Rochester (Rochester, NY, USA)	Recombinant amino acid fragment (amino acids 1-34 of human parathyroid Hormone (PTH))	SC	Subchondral bone	PTH receptor is found to be up regulated in chondrocytes in OA. Thus, a recombinant human PTH(1–34) (teriparatide) is hypothesised will inhibit aberrant chondrocyte maturation and associated articular cartilage degeneration.	Phase 2 in progress	The clinical trial has completed, but the results have not yet been published.	[182]

TPX-100	NCT03125499 NCT01925261	OrthoTrophix (Covina, CA, USA)	Matrix extracellular phosphoglycoprotein (MEPE)	I-ARTIC	Subchondral bone	Regulate subchondral bone mineralization.	Phase 2 completed	MRI measures failed to detect any difference in cartilage thickness or volume after one year, but various patient-reported measures such as knee daily function demonstrated statistically significant improvements in TPX-100-treated knees.	[183]
Lutikizumab (ABT-981)	NCT02087904	AbbVie (Chicago, IL, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Neutralising antibody IL-1α and IL-1β with a dual variable domain. Blockades IL-1. In a mouse model, reduced OA progression and increased the threshold for evoked pain more than inhibition of either antibody alone [184]. In phase I trial with knee OA, patients showed reductions inflammation markers, less inflammation would cause reduced pain [185].	Phase 2 completed	The limited improvement in the pain score and the lack of synovitis improvement, together with published results from trials of other IL-1 inhibitors, suggest IL-1 inhibition is not an effective analgesic/anti-inflammatory therapy in most patients with knee OA.	[186, 187]

Tanezumab (UBX0101)	NCT02709486	Pfizer (Brooklyn, NY, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Humanised monoclonal antibody, senolytic agent, interferes with the binding of nerve growth factor (NGF) to its corresponding receptors. NGF is increased in all chronic nociceptive and neuropathic pain states. NGF is produced as a response to any noxious stimuli that produce inflammatory cytokines such as interleukin-1 (IL-1) [188].	Phase 3 completed	Failed to meet 12-week primary endpoint. Analgesic benefit seen but also identified occurrence of rapidly progressive OA and subsequent need for total joint replacement.	[189-192]
Fasinumab	NCT03285646	Regeneron Pharmaceuticals, Inc (Tarrytown, NY, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Also, an anti-NGF monoclonal antibody. Interferes with OA nociception reducing the pain response.	Trial terminated	Number of treatment-emergent adverse events.	[193]

Skeletal stem cells	NCT05288725	Next Generation Regenerative Medicine LLC (Brevard County, FL, USA)	Minimally Manipulated Autologous Bone Marrow Aspirate containing bone marrow derived mesenchymal stem cells and other endogenous acellular components.	I-ARTIC	Cartilage regeneration and repair	Change the OA joint microenvironment to support in tissue healing and facilitate tissue regeneration.	Phase 2 ongoing (not yet recruiting)	n/a	n/a
Stem cell-derived extract (CCM)	NCT04971798	General Therapeutics Ltd, Shepherds Bush, London. UK)	Cell-free stem cell-derived extract formulation	I-ARTIC	Cartilage regeneration and repair	Novel cell-free stem cell-derived extract (CCM) from human progenitor endothelial stem cells (hPESCs). A preliminary study demonstrated the presence of essential components of regenerative medicine, namely GFs, CKs, and EVs, including exosomes, in CCM.	Early Phase 1, estimated study start date 2023	n/a	[194]

GLPG1972/ S201086	NCT03595618	Galapagos and Servier	Orally Bioavailable ADAMTS-5 Inhibitor	ORAL	Cartilage regeneration and repair	ADAMTS-5 is key in the degradation of human aggrecan (AGC), a component of cartilage. A potent and selective ADAMTS-5 inhibitor.	Phase 2 completed	Primary objective was to demonstrate the efficacy of GLPG1972/S201086 compared to placebo after 52 weeks in reducing cartilage loss in knee via quantitative MRI. The trial failed to meet the primary objective. In participants with OA, mean Aggrecan neoepitope fragments levels decreased from baseline between day 3 and day 15, followed by a floor effect between day 15 and day 29 in all GLPG1972 treatment groups. How and if this translates to a clinical meaningful effect on cartilage thickness and/or pain and function remains to be determined.	[195, 196]

Lutikizumab (ABT-981)	NCT02087904	AbbVie (Chicago, IL, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Neutralising antibody IL-1α and IL-1β with a dual variable domain. Blockades IL-1. In a mouse model, reduced OA progression and increased the threshold for evoked pain more than inhibition of either antibody alone [184]. In phase I trial with knee OA, patients showed reductions inflammation markers, less inflammation would cause reduced pain [185].	Phase 2 completed	The limited improvement in the pain score and the lack of synovitis improvement, together with published results from trials of other IL-1 inhibitors, suggest IL-1 inhibition is not an effective analgesic/anti-inflammatory therapy in most patients with knee OA.	
Tanezumab (UBX0101)	NCT02709486	Pfizer (Brooklyn, NY, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Humanised monoclonal antibody, senolytic agent, interferes with the binding of nerve growth factor (NGF) to its corresponding receptors. NGF is increased in all chronic nociceptive and neuropathic pain states. NGF is produced as a response to any noxious stimuli that produce inflammatory cytokines such as interleukin-1 (IL-1) [188].	Phase 3 completed	Failed to meet 12-week primary endpoint. Analgesic benefit seen but also identified occurrence of rapidly progressive OA and subsequent need for total joint replacement.	

Fasinumab	NCT03285646	Regeneron Pharmaceuticals, Inc (Tarrytown, NY, USA)	Anti-Interleukin-1 (Anti-IL-1)	SC	Inflammation	Also, an anti-NGF monoclonal antibody. Interferes with OA nociception reducing the pain response.	Trial terminated	Number of treatment-emergent adverse events.	
Fisetin	NCT04210986 NCT04815902	Steadman Philippon Research Institute	3,3′,4′,7-tetrahydroxyflavone	ORAL	Inflammation	Potential senolytic and anti-inflammatory action.	Phase 2 ongoing Phase 2 not yet recruiting	n/a	[197, 198]

1.6 Stem cells

Stem cells are unspecialised cells first identified and defined by McCulloch and Till in the 1960s [199], they are defined by two key traits; the ability to self-renew and to differentiate into various cell types [200]. There are several different kinds of stem cells summarised below in **Table 5**.

Table 5: Different types of stem cell summary. Including, abbreviation, description, sources, differentiation potential and ref.

Stem cell type		Abbreviation	Description	Sources	Differentiation potential	Ref
Embryonic		ESCs	Able to differentiate into cells of the three germinal layers. Considered risks are tumor formation and immune rejection.	- Inner cell mass of blastocyst	Totipotent	[201, 202] [203]
Induced-pluripotent		IPSCs	Adult somatic cells that have been reprogrammed back into an embryonic-like pluripotent state. Like ESCs in many aspects [204].	- Skin fibroblasts - Peripheral blood	Pluripotent	[203, 205-207]
Adult	Mesenchymal	MSCs	Exhibit multilineage differentiation, they are stromal cells that can be isolated from a variety of tissues. play a significant role in the body's natural healing capability.	- Bone marrow - Adipose - Spleen - Lungs - Dental pulp	Multipotent	[203, 208-216]
	Haemopoietic	HSCs	Precursor/ primitive cells that can develop into all the different types of blood cells, including those of myeloid and lymphoid lineage.	- Bone marrow - Peripheral blood - Cord blood		[203, 217-219]

1.6.1 MSCs

This project focuses specifically on protein development with an application of improving targetting and adherence of MSCs. The name MSCs was first coined by Caplan [210].

1.6.2 Differentiation

Figure 7 shows a summary of the different fates that can be attained by MSCs, for cartilage regeneration chondrogenesis shown in purple is the desired cell commitment. MSCs are multipotent adult stem cells, derived most often from bone marrow tissue. They possess the capacity to differentiate into multiple progenitor cells namely: osteocytes, chondrocytes, adipocytes, tenocytes, and myocytes [207, 217, 222, 223].

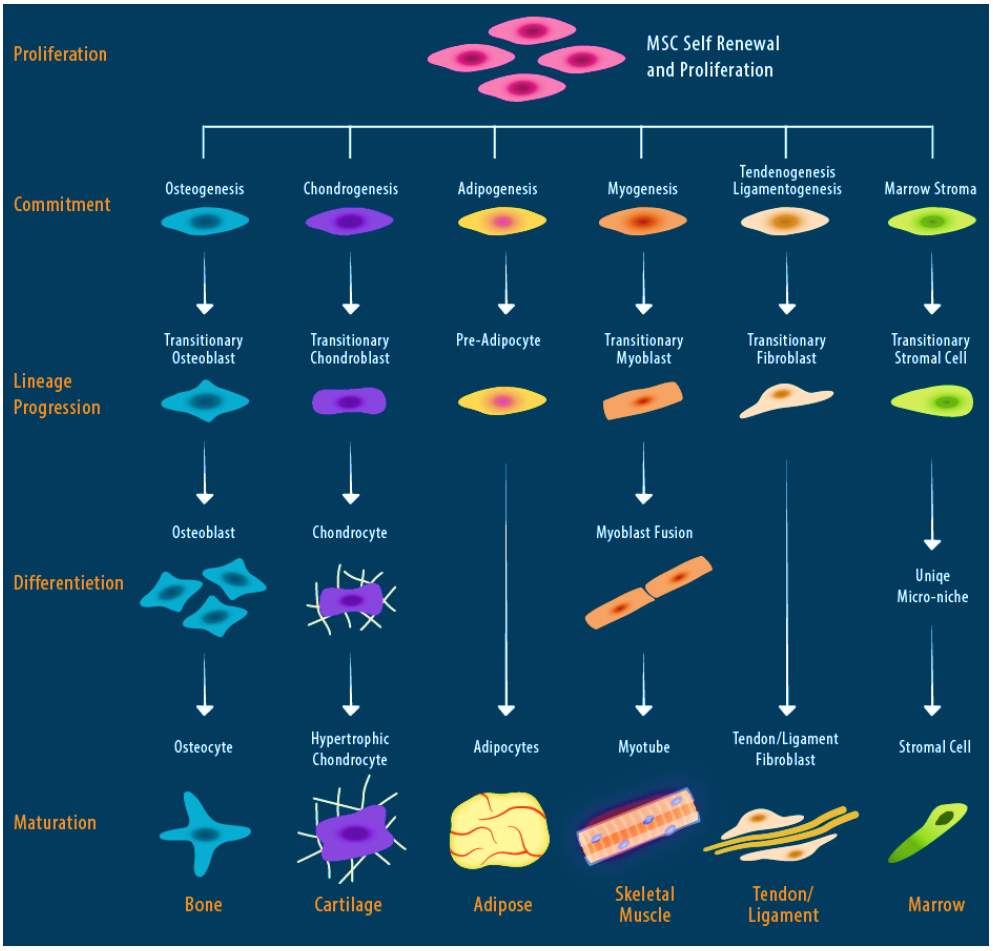


Figure 7: MSC (Mesenchymal stem cell) fates. Schematic showing the spectrum of MSC fates, transversing many different cellular niches. MSCs can be described as quiescent, are capable of self-renewal, proliferation, and commitment to different lineages in response to different inductive signals. Such signals (e.g., physiological, injury or disease) can elicit differentiation of the progenitor cells to mature cells which eventually develop into bone, cartilage, adipose, skeletal muscle, tendon, ligaments, and marrow tissues. Taken from [220].

Standard trilineage MSC differentiation conditions are widely used but additional modifications to *in vitro* conditions can promote differentiation to additional tissue types including muscle, cardiac and liver [212]. Once differentiated, the former MSCs express most of the hallmark genes expected of the differentiated cell type [212]. Currently, the more prominent MSC therapeutic uses take advantage of the MSC's production of factors and the responsiveness of other interacting cells, such as cells of the immune system (**See Introduction section 1.6.4**).

1.6.3 MSC nomenclature

As MSC use has become more extensive, with increasing utilisation in biomedical research several concerns have arisen regarding their proper identification. Criteria originally proposed by the ISCT for the identification of MSCs are no longer sufficient, reliance on these criteria alone has led to the common misconception that cells meeting the criteria are equivalent [221]. Although all MSCs exhibit the same phenotypic characteristics, irrespective of their source, their proliferation and differentiation potential do depend upon their anatomical source. It is now well documented that MSCs isolated from different tissues and cultured with varying methods represent a heterogeneous group of cells in terms of differentiation, proliferation abilities, and cell surface expression [222] [223, 224]. The large variety of tissues they can be isolated from has led to the suggestion that MSCs could originate from a perivascular niche [225, 226]. Heterogeneity of morphology and function has been seen, even from colonies expanded from single cells [227].

Nomenclature and defining characteristics have been debated for years, yet the generic term ‘MSC’ has actually been used in a non-standardised way to cover a wide range of cellular phenotypes (not just cells with stemness) [228]. Biologic properties of the unfractionated/ crude population of cells commonly employed do not meet generally accepted criteria for stem cell activity, making the name scientifically inaccurate and potentially misleading [229]. ISCT now suggest that fibroblast-like plastic-adherent cells, (which includes stem, progenitor, and differentiated cells) regardless of the tissue from which they are isolated, be referred to as multipotent mesenchymal stromal cells [229]. Whilst the term mesenchymal

stem cells is reserved only for cells that meet specified stem criteria [230]. The two defining hallmarks of cell stemness are the ability to self-renew, and to differentiate into multiple lineages [231].

True MSCs are usually described as plastic adherent, which is one of the key features used to characterise and distinguish them from other cell types [232]. They should demonstrate at least tri-lineage differentiation potential, able to become chondrocytes, osteoblasts adipocytes [233]. Although MSCs are very easily expanded they are a heterogeneous population of cells and biological variability translates to give inherent culture inconsistencies. Easily becoming biologically distinct from the *in vivo* populations from which they were originally obtained [234]. MSCs must also demonstrate a specific antigen marker profile. They must be positive for CD105, CD73 and CD90, and negative for markers of CD45, CD34, CD14 or CD11b, CD79α or CD19 and HLA-DR surface molecules [230].

Cells planned for use initially with this therapeutic etiology and available for future development beyond this project were collected from the iliac crest of patients undergoing surgical repair following traumatic injury. They have been confirmed to have positive markers for CD105 and CD90 and the absence of negative markers CD34 and CD45 [235]. However markers can be donor, isolation and passage-dependent and may not actually represent the true in vivo MSC population adding further complexity [231]. As the project progresses beyond the scope of this thesis further consideration towards characterising and maintaining cellular populations will be required.

1.6.4 Immunoprivilige & Immunomodulation

MSCs have been classed as both hypo-immunogenic, (because they do not express class II major histocompatibility complex (MHC) antigens) and immunosuppressive (because they inhibit proliferation of T-Lymphocytes [236-245]). This is hoped will make it possible to use allogeneic (donated) MSCs from one donor to treat many patients, which is a practical attribute and would aid scale-up of such a therapy as a more direct route to achieving a disease modifying treatment. Unfortunately, more study is needed to fully understand MSC immunogenicity, particularly *in vivo*.

There is some contention regarding allogenic MSC use as despite evidence of limited MHC expression there is some evidence for immune responses with mismatched MHC haplotype donors [246]. Recent studies document generation of antibodies against and immune rejection of allogeneic donor MSCs suggesting that MSCs may not actually be immune privileged but actually immune evasive [247]. *In vitro* studies into allogenic MSC administration with mixed donor lymphocyte reactions reported that MSCs prevented lymphocyte proliferation, and did not cause apoptosis of T cells, instead the T cells only responded to subsequent lymphocyte challenge when the MSCs were removed [248, 249].

MSCs have been shown capable of modulating immune responses in situations where T, DC, macrophage and NK cell proliferation would normally cause a cytokine storm. Immunomodulation is highly desirable property that clinical trials are now also working to utilise. MSCs have been to shown to produce the antibacterial agents PGE2 [250, 251] and LL-37 peptide [252], that may work in vivo by influencing

hematopoietic cells. Modulation of the immune system presents an additional means to improve functionality of the damaged tissue in the OA joint.

1.6.5 Stem cells as OA therapeutics

MSCs are considered the clinical gold standard in tissue bioengineering where regeneration and reconstruction are required to resolve disease pathology such as in OA [253]. The promise of MSCs as a regenerative therapeutic modality has been well demonstrated in much preclinical data, yet this has not always translated to consistent, successful clinical trial results [254]. The use of stem cells therapeutically to repair cartilage tissue is based upon their ability to act as chondroprogenitors, able to replace injured cartilage and as regenerative cells stimulating cartilage repair by endogenous joint cells [255].

Watkitani et al, performed the pioneering first in-patient trial implanting autologous bone-marrow derived MSCs (BM-MSCs) to repair articular cartilage in collaboration with Arnie Caplan in 1988 [256]. In this groundbreaking study MSCS were seeded onto collagen gels to act as a carrier and transplanted into induced OA knee defects in rabbits. MSC treated defects healed in a manner that was much closer in nature to normal articular cartilage and was more rigid and compliant when compared to the repair seen by the defects that were not treated. This strategy was confirmed to give significantly greater improvement compared to ACI [257]. Although not a perfectly refined therapeutic, further developments have amassed over the years since this first attempt and promise remains. BM-MSCS are commonly harvested/ sourced from the iliac crest of patients, which is a highly invasive and painful procedure [258]. Therefore, alternative more easily accessible sources have

been investigated including both adipose tissue–derived mesenchymal stem cells (AT-MSCs), and peripheral blood-derived mesenchymal stem cells (PB-MSCs) [208].

Importantly MSCs have been shown to ‘remember’ a stimulus after transitioning to a new environment [259, 260]. Therefore, priming MSCs instils a ‘short-term-memory’ (with MSCs retaining effects of in vitro stimuli *in vivo*), avoiding the need for separate *in vivo* activation when aiming towards a specific end therapeutic phenotype i.e. chondrogenesis to facilitate cartilage replacement. To promote chondrogenic differentiation MSCs can be cultured with transforming growth factor–β_3 (TGF- β_3) [261] [262].

MSCs can be primed using an array of mechanisms and signals, **Figure 8** shows summary of the most common including; hypoxia , media, matrix mechanics, 3D environment, non-coding RNA, cytokines and hormones [221].

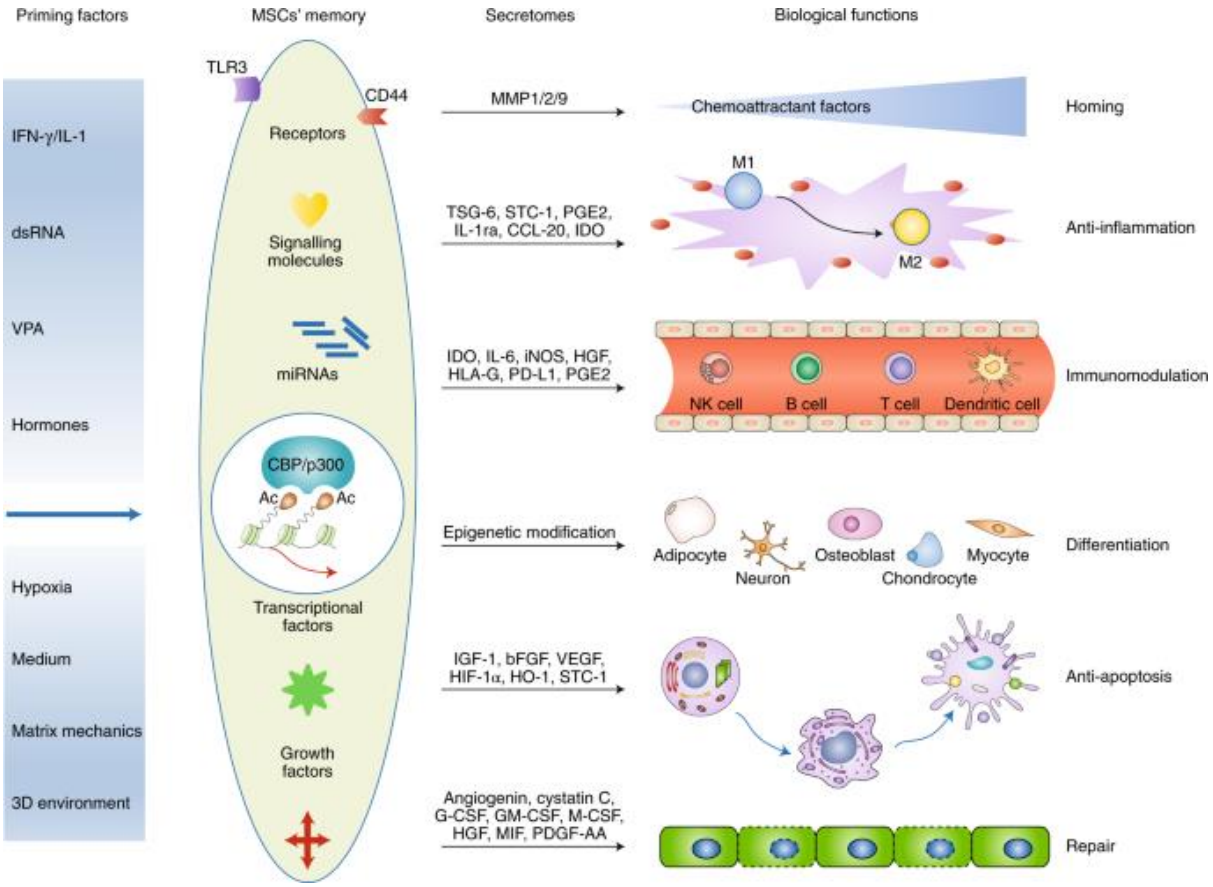


Figure 8: Summary of the mechanisms that can be utilised in MSC-priming. Taken from [221]. MSCs can be primed via different signals (including hypoxia, matrix mechanics, 3D environment, non-coding RNA, cytokines and hormones) to acquire and retain phenotypes relevant to the intended therapeutic application.

In effect *in vitro* priming and conditioning can be used to coach/ tune MSCs in advance of their administration to patients. Mesenchymal cell replacement in the large numbers needed to treat tissue injury such as repair/ replacement of cartilage lost and compromised/degraded in the OA joint requires engraftment, structural organization then cellular differentiation to allow regeneration, which is a complex sequence of events for which our understanding has progressed but still remains

unperfected. Engraftment improvement/ facilitation is sought through the coating of MSCs with a potent damaged collagen binding protein in the work here of this thesis.

The therapeutic effect of MSCs relies upon on their ability to reach the injured or diseased site (Here we are hoping injection and targeted adhesion to TII gelatin will provide direct delivery of the cells where required). Several factors affecting the therapeutic efficacy of MSCs' need consideration; culture conditions, the number of passages, and MSC donor age will all be significant to the success. It has been shown that freshly isolated cells compared with in vitro-cultured cells have a higher engraftment efficiency.

Culture conditions also have a significant impact on homing capacity, as they can modify the expression of the surface markers involved in this process (this is not such a large consideration as we are modifying the cell surface ourself via coating). As an example, CXCR4, a chemokine receptor, is involved in the migration of MSCs. It has been shown that CXCR4 expression is lost on BM-MSCs during culture, whereas the presence of cytokines (e.g., HGF, IL-6), hypoxic conditions, or direct introduction using viral vectors allow for restoration of its expression.

Stem cells have broad therapeutic potential within the OA joint. They have been shown to provide paracrine, trophic and immunomodulatory effects giving pain relief, along with modification and restoration of homeostasis of the cartilage ECM microenvironment restoring its functionality and the balance between anabolic and catabolic processes [263]. Useful biomolecules released by stem cells include inflammatory cytokines, growth factors, and anti-inflammatory interleukins, some of which have demonstrated symptomatic relief and benefit in the inflamed painful OA

joint [264, 265]. There has been somewhat of a paradigm shift in that MSCs are now seen as a paracrine provider as well as a means of cell replacement. Emphasis has shifted toward harnessing MSCs' ability to produce factors and cytokines that stimulate innate tissue repair and modulate inflammation and immune responses [212].

The general safety of BM-MSCs has been established over the years with many attempts made already to utilise them in cartilage injury defects and OA treatment [266, 267]. The selection of stem cell-based treatments available to a patient is dependent on the specific cartilage pathology. MSCs are a strategy that had been shown to achieve significant positive patient outcomes particularly for pain and increased mobility. There are still several challenges to be overcome before MSC implantation becomes a practical OA therapeutic and cartilage repair approach within the clinical setting [263]. Particularly, targeting and engrafting into defects rather than the almost by chance approach, which is the only option currently, whereby there is a hope that MSCs engraft at the correct location for their intended application. Another key area of development lies with combination therapies, whereby small molecule drugs are administered alongside MSCs. This co-administration is used to direct MSC differentiation down the chondrogenic path. Kartogenin (KGN) is a commonly used chondrogenic factor found to be a chondrogenic promoter of BM-MSCs towards cartilage regeneration [268, 269]. Other example molecules under investigation are curcumin [270] and resveratrol [271], both of which have also similarly demonstrated induction of chondrogenic differentiation.

Tissue engineering technology combined with MSCs has interesting application in OA, in that the scaffolds and carriers (biodegradable and biocompatible) can provide support for the MSCs but also to direct delivery. During surgery, scaffolds seeded with MSCs can be placed directly within defects [126]. Whether they adhere and stay where placed is another important consideration and engineering challenge. Biomimetic materials are greatly sought [272, 273]. Examples of some of the materials developed and trialled so far are collagen hydrogels e.g. NOVOCART® 3D (TETEC Tissue Engineering Technologies, Reutlingen, Germany), Fibrin, chitosan, hyaluronic acid (HA), hydroxyapatite, gelatin, alginate, agarose, cellulose, poly(l-lactic acid) (PLLA), poly(lactide-co-glycolide) (PLGA), polycaprolactone (PCL) and a cellulose-silk composite to name just a few [266, 274-279]. The 3D environment dispensed by the scaffold has a crucial role in maintaining the chondrogenic phenotype of MSCs. The scaffold also enables the homogeneous distribution of MSCs, supplying an appropriate substrate for cell growth and mechanical integrity for post-surgical implantation [263]. 3D printing adds extra precision to manufacturing capabilities, presenting an exciting future as far as scaffolds are concerned [275].

MSC exosomes are secreted extracellular vesicles that could provide an alternate source of cytokines and trophic factors [280]. Exosomes are imperative to intercellular communication. They facilitate the transfer of bioactive lipids, nucleic acids (DNA, mRNAs, and non-coding RNAs) [39], and proteins between cells to elicit biological responses such as gene-regulation [40], proliferation, apoptosis [41], and immunomodulation [42] in recipient cells [43]. Recent studies have elucidated a crucial role for MSC-derived exosomes in the regulation of cell migration,

proliferation, differentiation, and ECM synthesis. Recently it has also been confirmed that MSC exosomes may supress OA development [280, 281], making them an interesting alternate MSC-derived therapeutic modality.

1.6.6 ECM degrading enzymes

In OA cartilage there is increased expression of matrix metalloproteinases (MMPs) and ADAMTS [280]. Overexpression of MMP-2, MMP-3, MMP-8, MMP-9, MMP-13, and MMP-14 [282] is seen, which can degrade most of the matrix molecules present in the ECM. Also ‘Aggrecanase’, ADAMTS4 and ADAMTS5 are thought to be involved specifically in the degradation of OA cartilages most prominent ECM proteoglycan aggrecan [282]. Degradation ot these critical ECM components increases tissue permeability and alters the biomechanical properties of the OA joint [283]. Arthroscopic joint lavage has traditionally been performed in milder OA cases as a minimally invasive intervention, to remove proinflammatory mediators, cells, destructive enzymes and joint debris [284]. An important point to make of particular significance to this project is that the degradative products of OA are present in abundance in diseased joints and particularly diseased regions. Instead of removing the debris, targeting it would be an interesting approach. To target degradation products is in effect to target diseased joints specifically, a strategy that has not yet been achieved clinically in OA.

1.7 Targeting MSCs into cartilage lesions

Attempts have already been made to inject undifferentiated MSCs into OA joints and have been shown to reduce pain giving improvement to patients but the longevity of such improvements is questionable, as there is little evidence as yet of

any engraftment of the MSCs into the cartilage or regeneration of the articular surface [242, 285-289]. If a method could be developed to encourage the accumulation of these injected cells into the cartilage lesions, they may then could drive a tissue repair response in addition to their pain-reducing effect.

Nanotechnology offers advantages to OA therapeutics. Nanotechnology includes liposomes, micelles, dendrimers, and nanoparticles (NPs) both polymeric (PNPs) and inorganic. NPs can be used to improve targeting and efficient delivery, enhance drug stability and solubility along with preventing dispersion and degradation extending retention in the diseased site [290]. Aptly referred to as nanozymes, nanomaterials with enzyme-like characteristics have been developed such as; iron oxide (Fe_3O_4), cerium oxide (CeO_2), manganese dioxide (MnO_2), gold (Au), and platinum (Pt) NPs [291] [290, 292]. These nanozymes mimic natural antioxidant enzymes activity and possess reactive oxygen species (ROS) scavenging activities to improve the proliferation and subsequent survival of chondrocytes [291, 293]. Magnetic targeting is another interesting similar avenue recently considered, whereby MSCs were cultured then magnetized with ferucarbotran. The coated MSC were then injected and guided magnetically into the OA defect. Follow-up results showed complete coverage of the defects, along with an infill of cartilage-like tissues and significant improvement in clinical outcomes 48 weeks after treatment [294]. Another similar magnetic nanoparticle based therapeutic strategy is using temperature- responsive release of bioactive molecules [295]. In another recent study functional nanoparticles were injected into a rat model of OA. Functional nanoparticles were employed as a means to first recruit and then to promote the therapeutic action of resident MSCs within the joint, it showed

reduction in destruction and some evidence of cartilage generation [296]. It will be interesting to follow such work through further developments as it gives a contrast to all the OA work focussed on exogenous MSCs as an OA treatment [297].

One other method of achieving targetting to OA lesions would be to modify the MSC cell surface to make it more adherent to the lesion sites. In healthy articular joints, TII collagen is the most prevalent protein, however in the OA joint this collagen is degraded, causing the accumulation and abundance of TII gelatin (degraded TII collagen) instead, particularly at the articular surface [42, 43, 286, 287, 298]. Targeting MSCs to the gelatin in articular cartilage lesions would encourage their accumulation at exactly the right location for promoting tissue regeneration where needed. **Figure 9** shows the rationale for why the TII gelatin peptide is an ideal target for binding of a protein coating stem cells. Collagen is usually organised within the arcades of Benninghoff that transcend all four zones of the ECM from superficial to calcified. When TII collagen is degraded there are two dominant products, the ¾ and ¼ fragments [42].

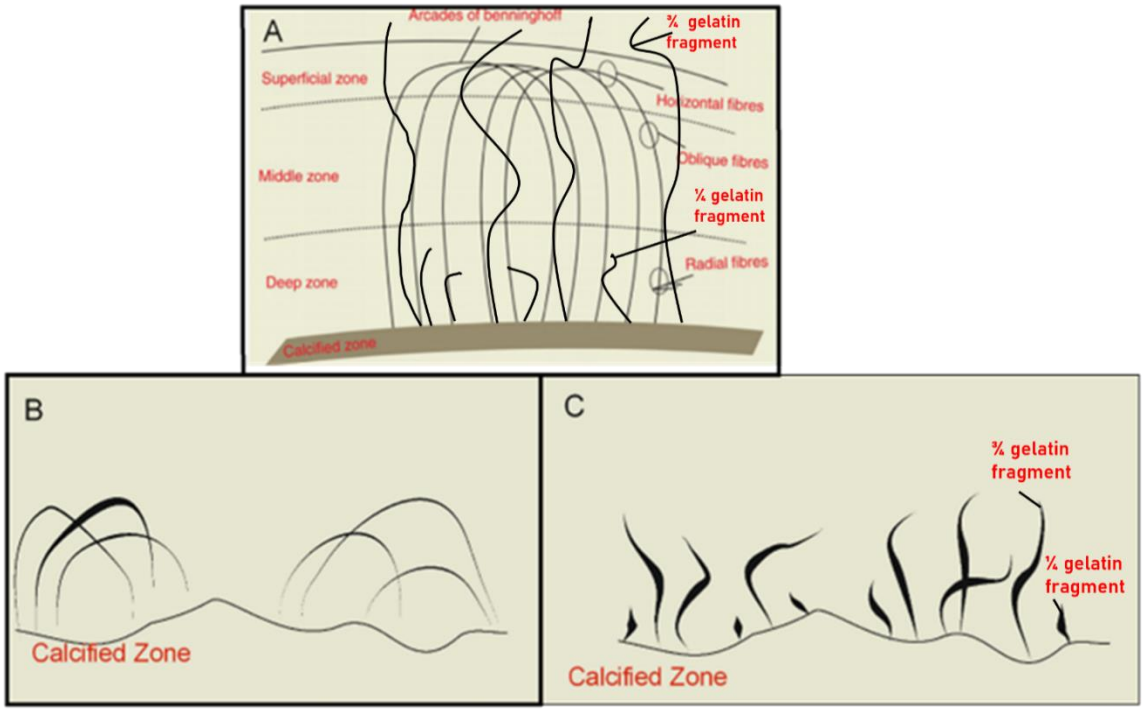


Figure 9: Gelatin degradation in articular cartilage. (A) Arcades of Benninghoff taken from [299]. (B) Intact collagen fibers anchored to the calcified zone. (C) Gelatin fragments anchored to the calcified zone.

1.8 Aims

The aim of this project is to build on the previous work of the Hollander group by continuing to characterise and improve the efficiency of a protein to target mesenchymal stem cell (MSC) adhesion to damaged cartilage when injected into the knee joint as a treatment for osteoarthritis. In the first instance, this will be done by exploring mutants of a protein that has been developed, has undergone *in vitro* testing, and has undergone initial *in vivo* testing in a post-traumatic osteoarthritis (OA) mouse model, as a method of targeting MSCs to damaged areas of cartilage. This work focuses on further *in vitro* work only. At the onset of this project, it was already clear that exploiting the previously described recombinant protein with enhanced binding, named 222 as it's made up of three repeats of module 2 of the

collagen binding domain (CBD), requires substantial further optimisation work, some of which has been undertaken and described in this thesis but also now extending beyond this project.

1.8.1 Previous related work of the Hollander group

The Hollander group hypothesised that coating MSCs with a protein segment from gelatinase A (also known as MMP-2), would lead to this targeted accumulation of MSCs where most necessary. Gelatinase A contains a segment called the collagen binding domain, (CBD) which is the basis for the work here in this strategy [235]. CBD is made up of three fibronectin-like modules that are all involved in the binding of CBD to gelatin. The CBD of MMP-2 has been shown using an *in vitro* plate binding assay, to bind with a Kd of 50nM to TII gelatin [300], Importantly CBD has also been shown to have a low capacity for binding to intact, native TII collagen [300]. This selective property is an inherently exploitable specificity towards denatured TII collagen (TII gelatin) (**Figure 10**). MMP-9 binds with a slightly higher affinity to TII gelatin (Kd of 8.0 nM), but also shows strong affinity to native TII collagen (Kd of 50nM) [301]. Which justifies the choice made to work with MMP-2 related CBD rather than using MMP-9.

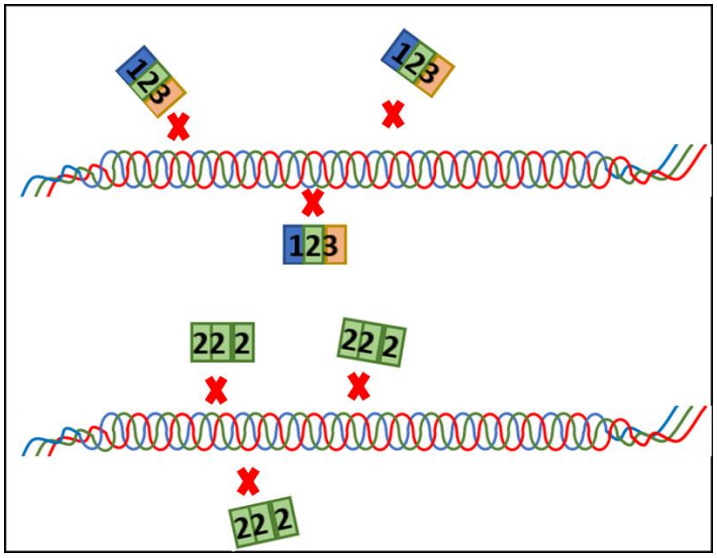


Figure 10: Lack of binding of CBD and 222 to intact TII collagen (triple helix). CBD is made up of three modules 1, 2 and 3. This is the native protein taken from MMP-2 from which module two was identified as the most potent in binding to TII gelatin. Hence a recombinant chimeric mutant known as 222 with three modules 2s was designed. Neither CBD nor 222 show any binding affinity for intact TII collagen only for TII gelatin the degraded form, available in OA lesions. Meaning targetting with this protein is to damaged cartilage tissue only.

Soluble expression and recovery of CBD in *E.coli* has been achieved however not without initial difficulty; the CBD protein contains six disulfide bonds, a post-translational modification (PTM) which proved problematic for the initial *E. coli* strain trialled only insoluble (misfolded) CBD was expressed in early attempts. It was only in subsequent trials involving switching to use a Shuffle (DE3) *E. coli* strain with a more oxidising cytoplasmic environment (better suited for disulfide bond formation) [302] that soluble CBD recombinant was attained [235]. Module 2 was then identified via the combined results of a TII gelatin binding assay [303] and nuclear magnetic resonance (NMR) spectroscopy as the CBD module that binds with the strongest affinity to type II gelatin [235]. Previous work to further improve binding engineered and expressed an alternate mutant chimeric recombinant CBD protein consisting of three module 2s (named 222).

It was found that this protein bound to TII gelatin fourteen times more strongly than CBD. 222 was then further optimised for expression within the Shuffle *E. coli* strain, attaining an average yield of 3.5mg/L of culture [235]. This yield was sufficient to enable further experiments to coat MSCs with 222 to target adhesion (**Figure 11**).

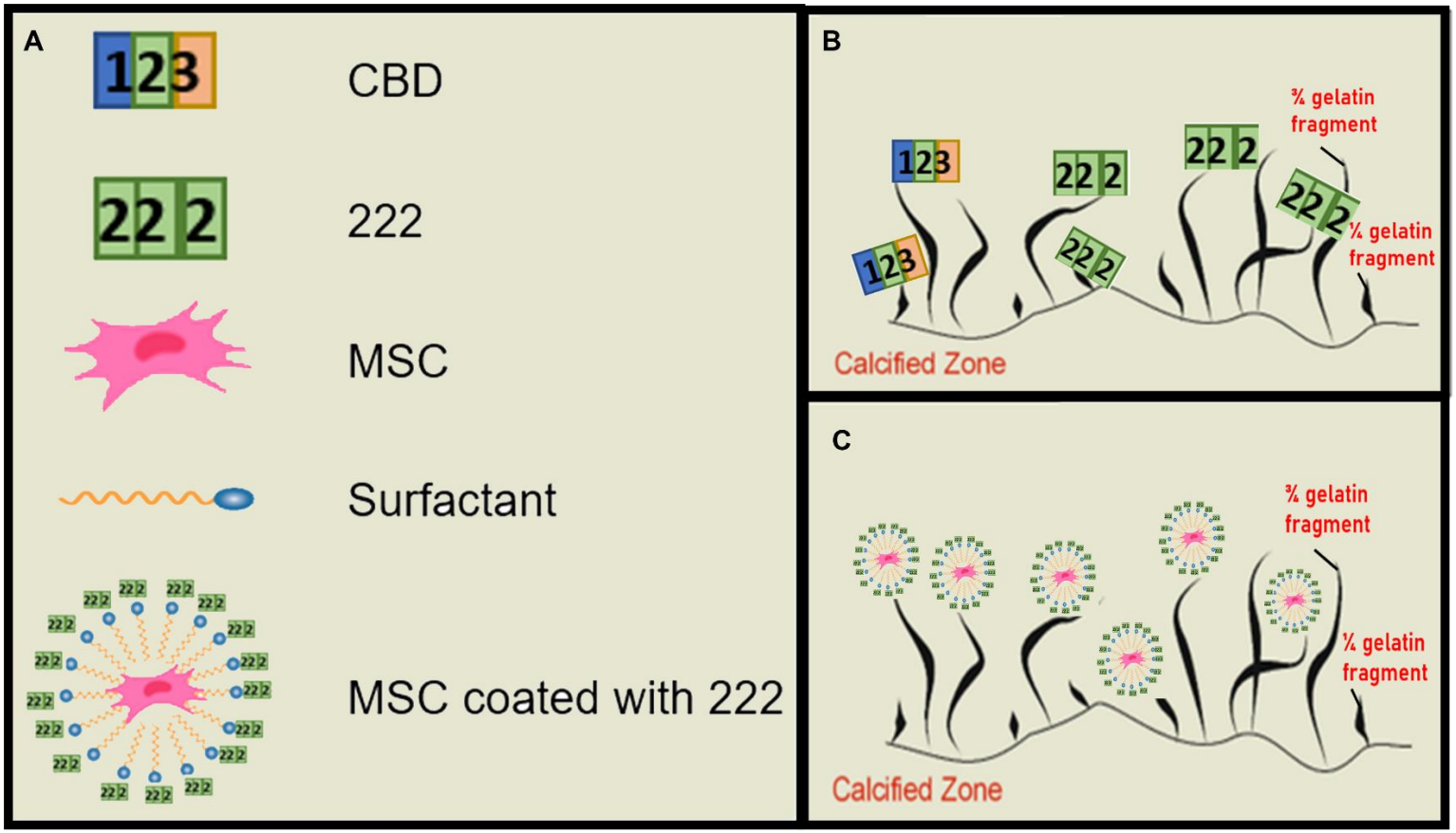


Figure 11: TII gelatin targetting rationale. A shows a key for this figure. **B** shows that both CBD and 222 bind to the gelatin fragments exposed at the surface of the OA joint, an ideal target for cell adhesion. **C** shows that MSCs are to be coated with a protein using a surfactant corona method (outlined in the next section **1.8.1.1)** to target adhesion. 222 is the most potent protein for binding to TII gelatin that the group has.

1.8.1.1 MSC coating

A surfactant corona method of coating proteins onto the surface of MSCs was reported by the Hollander group previously [304] and has now been adapted for coating of MSCs in suspension rather than when bound to plastic [235]. The method involves surrounding the protein with an anionic PEG based surfactant corona, meaning the protein can then be incorporated into the membrane of stem cells via hydrophobic interactions. The steps in the protein conjugation process are outlined below:

1. The carboxylic residues first need to be activated using a carbodiimide, N-(3-Dimethylaminopropyl)-N′-ethylcarbodiimide hydrochloride (EDC)(**Figure 12 A, B**).
2. Cationization of protein by covalent coupling of N,N'-dimethyl-1,3-propanediamine (DMPA) to the carboxylic residues of the protein. (aspartate and glutamate) (**Figure 12 C , D)** can then proceed.

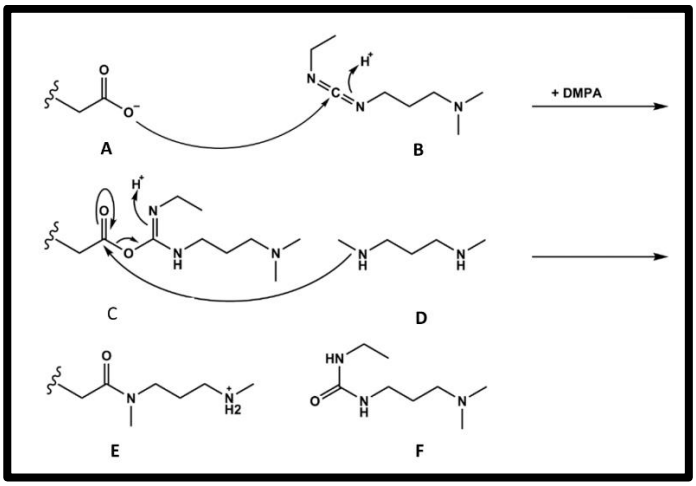


Figure 12: Covalent coupling of DMPA to the carboxylic groups of amino acids. The addition of DMPA enables the conversion of negatively charged aspartate and glutamate residues (**A**), into synthetic amino acids possessing amine side-chains (**Figure 12 E, F)**).

3. The cationised protein is the electrostatically coupled to a negatively charged PEG-based surfactant, glycolic acid ethoxylate 4-nonylphenyl ether (obtained from the oxidation of Igepal CO-890, **Figure 13**) . This surfactant then surrounds the protein like a corona.

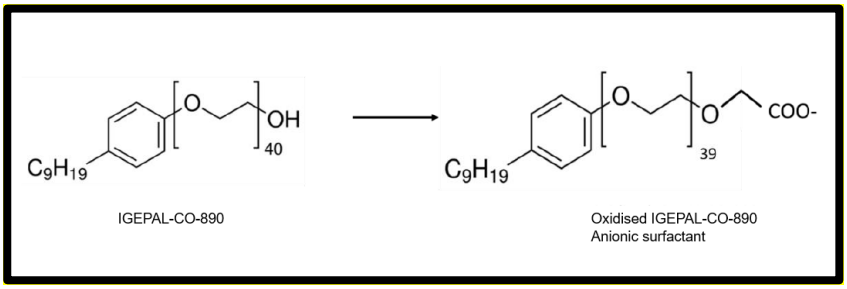


Figure 13: Surfactant oxidation reaction. Complete oxidation of the terminal hydroxyl group of Igepal CO-890. Yielding a desired anionic glycolic acid ethoxylate 4-nonylphenyl ether.

4. The conjugated protein can then be delivered into the cell membrane of the MSCs, with the surfactant anchoring the protein to the cell membrane coating the MSC in protein (**Figure 11)**.

The successful coating of MSCs with 222 has been achieved using this improved methodology; a significant result for proof of concept of the modality [235].

1.8.2 Modelling of 222, a devised rational strategy to selecting binding mutants to take to *in vitro* expression and then characterisation experiments

The first aim of this project was to model mutations in 222 to binding site residues from NMR data [235]. Alanine scanning mutagenesis was applied, whereby binding amino acids were substituted sequentially and systematically for alanine. This was done *in silico* using the Pymol mutagenesis wizard. Alanine is the smallest chiral amino acid; it is chemically inert making it a suitable substitution that can lead to functional loss whilst maintaining secondary protein structure.

Traditionally alanine scanning libraries are constructed via sequential substitution of each amino acid in a protein sequence however here we were looking to reduce as much as possible the number of modelled mutant proteins taken to *in vitro* testing. We knew from the NMR studies using 222, which residues collectively are involved in its binding to type II gelatin, so the alanine mutant *in silico* design was narrowed to explore only those residues. Computational assessment could be carried out to evaluate how mutant proteins differed in terms of stability and solvent exposure, with only the strongest mutant candidates selected for *in vitro* testing. Resulting computationally designed mutants, were then to be expressed using Shuffle cells and assayed for altered function using the existing binding assay methodology. Such an approach allowed quick determination and confirmation of each previously identified implicated residue's importance to binding functionality, by determining if the residue's replacement with alanine reduced or altered binding capacity.

1.8.3 Molecular docking, computational design, subsequent expression, and characterisation of higher binding affinity mutants

Following the identification of critical binding residues further mutant design was undertaken using computational methods including molecular docking. Here, mutant design was aimed at improving binding to TII gelatin, not disrupting it. As in the alanine mutant design described above only the strongest mutant candidates were selected for *in vitro* testing. Resulting computationally designed mutants, were then to be expressed using Shuffle cells and assayed for altered function using the existing binding assay methodology.

1.8.4 Design, expression, and characterisation of solubility mutants

Secondary to improving binding affinity a separate line of work explored computational design of 222 mutants with improved solubility, followed by *in vitro* experiments to express and characterise them. Solubility is by definition the maximum concentration of a solute that can be dissolved in a solvent at a given temperature [305]. Solubility is an important attribute of a recombinant protein particularly one to be used, as 222 was intended, therapeutically as an adjunct to target and adhere MSC to OA lesions. Designed mutants were assessed for solubility using a precipitant based plate assay. Followed by assessment of binding to TII gelatin to check for any consequential alteration to binding affinity.

1.8.5 General characterisation of CBD mutants

Given the amount of work that goes into attaining any soluble mutant protein it was considered important also to undertake as extensive as possible a biochemical

characterisation of each mutant, to improve our understanding of their fundamental biology, as well as potentially enhancing a future therapeutic development pathway.

2 Materials and Methods

This chapter covers the more general materials and methods used. The majority of the materials and methods are presented as first utilised within each results chapter then referred to as appropriate.

2.1 Growth media

All media was autoclaved for 15 minutes to sterilise as per manufacturer's instructions.

- Agar plates: 20g of Miller's LB Agar (Melford) was reconstituted with 500mL of reverse osmosis (RO) water. Where antibiotic was added (for selective growth) LB agar was cooled in a waterbath to 50°C. Antibiotic was added when required to the working concentrations shown in **Table 6.** Plates were then poured aseptically.

Table 6: Antibiotics, stock and working concentrations used throughout this work. When making agar antibiotic selection plates or when culturing bacteria for construct isolation, small scale, and large-scale expression trials. All antibiotics were purchased from Generon.

Antibiotic	Stock solution	Working concentration
Ampicillin sodium salt (A)	50mg/mL in H_2O	100µg/mL
Kanamycin sulfate (K)	50mg/mL in H_2O	50µg/mL
Kanamycin sulfate (K)	50mg/mL in H_2O	50µg/mL
Chloramphenicol (C)	25mg/mL in Ethanol	25µg/mL

- Lysogeny/ Luria Broth (LB): 25g of Millers LB powder (Melford) was reconstituted with 1L of RO water, stirred and gently heated to 37°C until fully dissolved, then pH adjusted to 7.2. LB was then sterilised for use in culture.
- Super optimal broth (SOB) media with catabolite repression (SOC): 28g SOB Broth (Melford) powder was reconstituted with 980mL RO water. SOB was autoclaved, cooled to 50°C, then 20mL of filter sterilised (0.22µm sterile PES syringe filter, Starlab) 1M glucose (Sigma Aldrich) was added aseptically to make the SOB media into SOC. SOC media is composed of salt, magnesium, and glucose that work to stabilize the bacteria, promote plasmid uptake, faster growth and increased transformation efficiency [306].
- 2x YT Media: 31g of 2x YT Powder (Melford) was dissolved in 1L RO water, this was autoclaved to sterilise for 15 minutes, as per manufacturer's instructions. 2x YT is a different nutritionally enriched media very similar to LB but with twice the amount of yeast extract [306]. This media was used for Small Ubiquitin-like Modifier (SUMO) and 3C protease production.

2.2 Proteases

SUMO protease (storage buffer: 25mM Tris-HCL, pH8.0, 1% Igepal, 250mM NaCl, 500µM DTT, 50% glycerol w/v) and 3C protease (storage buffer: 50mM Tris-HCL pH 8.0, 150mm NaCl, 1mM EDTA, 0.5mm Tris(3-hydroxypropylphosphine). 10% glycerol w/v) were both expressed and purified as outlined in the papers of Lau et al 2018 [307] for SUMO protease and Abdelkader et al 2021 [308] for 3C protease. Proteases were produced in house using stock vector constructs from a colleague (Dr Amy Wood). Purified proteases were stored at -80°C until required.

2.3 Sources of protein

Table 7 shows the sources of protein or constructs used to produce protein in this work.

Table 7: Sources of protein in this work. Whether they were provided or generated as novel constructs.

Gene/ Protein	Abbreviation	Source	Provided as/ Produced
CBD		Anais Dabbadie, University of Liverpool	Purified protein
222		Anais Dabbadie, University of Liverpool	Sequence confirmed, pOPINS construct
Mutant 1	M1	GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
Mutant 2	M2	GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
Mutant 5	M5	GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
Mutant 8	M8	GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
CamSol 6	CS6	GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
222W		GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed
222W-CS1		GeneArt, Thermofisher	Cloned in this work, construct sequence confirmed

2.4 Measurement of protein and nucleic acid concentration

Protein and DNA concentration were determined spectrophotometrically using a NanoDrop 2000 instrument (Thermo scientific).

2.4.1 Protein concentration

Absorbance readings were made at 280nm, and concentrations calculated using the Beer-Lambert formula, **Equation [1].**

$$\boldsymbol{A = C \times \varepsilon \times l} \qquad \textbf{[1]}$$

Where A is absorbance, C is concentration (mol/L), ε is molar absorption coefficient (L.mol-1cm-1) and l the path length 0.1cm. Where required, molar concentration was converted into a concentration in mg/mL, using the molecular weight. Values for ε and molecular weight were obtained for each protein using the protein sequences (outlined in each chapter) and the prot param online tool from Expasy (**Table 23**, **Table 42**, **Table 57**).

2.4.2 Nucleic acid concentration & purity

Nucleic acid concentration and purity was determined following vector linearisation, construct isolation and before sending constructs for sequencing. For nucleic acid quantification, a modified Beer-Lambert equation was used with absorbance readings taken at 260nm, **Equation [2].**

$$\boldsymbol{C = (A \times \varepsilon)/b} \qquad \textbf{[2]}$$

Where C is concentration (ng/µL), A is absorbance, ε is molar absorption coefficient (L.mol-1cm-1) and b the path length (cm). Where required, molar concentration was converted into a concentration in mg/mL, using the molecular weight. The generally accepted extinction coefficient for nucleic acids is: 50 ng-cm/µL for double-stranded DNA (dsDNA) [309].

To assess the purity of nucleic acid the ratio of absorbance at 260/280nm was utilised. A ratio of ~1.8 is generally accepted as pure [310]. A ratio below 1.8 indicates protein contamination and above 2.0 indicates RNA contamination [310].

2.5 Monitoring optical density

Optical density (OD) was monitored at 600nm during expression using a WPA Biowave CO8000 cell density meter (Biochrom). The meter was blanked with fresh media before any culture samples were measured.

2.6 SDS-PAGE

Sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) was used to assess protein expression, purity, and monitor purification steps. Gels: 15% resolving, 4% stacking, were made in house using Bio-Rad gel casting equipment, following the recipe outlined by Harlow and Lane 1988 [323, 324].

Prior to loading, 5µL samples were mixed with 5µL 4x SDS loading buffer (0.05M Tris HCl pH 6.8, 2% SDS (w/v), 10% glycerol (v/v), 0.1% bromophenol blue (w/v), 1.4M β-mercaptoethanol). Except for insoluble pellet samples which were first resuspended in ~30µL 8M urea via vortex, then had 10µL 4x SDS loading buffer added. All samples were then heated to 100°C on a heat block for 5 mins. SDS-PAGE gel tank apparatus (Bio-Rad) was assembled with the gel submerged in 1x SDS running buffer (diluted 1 in 10 with RO water from a 10X SDS running buffer stock: 248mM Tris, 793mM glycine, 35mM SDS). 5µL of these heated samples were loaded into the wells of the gel. The page ruler or page ruler plus marker was loaded into the first well (ThermoFisher) to act as a reference in identifying protein band

molecular weights. An electric current of 180V was applied for 1 hour to allow protein migration through the Gel (PowerPac 300 from Biorad). The gel was then rinsed with RO water, submerged in RO water, microwaved for 1 minute, then left rocking for 5 minutes before draining the water and submerging the gel in Coomassie brilliant blue magic stain (60mg Coomassie G-250, 996.6mL, RO water, 3.4mL concentrated hydrochloric acid (HCl)) [325]. Gels were scanned using an Image Scanner III (GE Healthcare).

2.7 Protein storage

All proteins once confirmed as sufficiently pure via sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE **materials and methods section 2.6**) assessment were aliquoted, labelled with protein concentration (**materials and methods section 2.4.1**), flash frozen in liquid nitrogen then stored frozen at -80°C. Samples were thawed in cold water and stored on ice when required.

2.8 Protein buffer exchange

2.8.1 Small dialysis cups

Slide-A-Lyzer MINI Dialysis Devices were used where buffer exchange of protein volumes less than 1mL were required. Dialysis was always conducted O/N with stirring of the target buffer, at 4°C.

2.8.2 PD-10

A PD-10 Sephadex G-25 column (Cytiva) column was utilised for buffer exchange of protein volumes >1mL but ≤2.5mL. The column was first rinsed with 30

mL of milliQ (MQ) water before being equilibrated with 25 mL of final target buffer. 2.5mL of protein in the initial buffer was applied to the column and allowed to flow through. Then 3.5 mL of final buffer was added and flow through collected to elute the bound protein in the final desired buffer.

2.8.3 Dialysis membrane

Cellulose membrane with a molecular weight cut-off of 3500 Daltons (Medicell) was used where larger volume protein buffer exchange was required during large scale purification to remove the imidazole after the initial his purification step or to swap the proteins into a 3C cleavage compatible buffer (500mM Tris-HCL, pH 7.5, 1.5M NaCl). Dialysis of proteins was always conducted O/N with stirring of the target buffer, at 4°C.

2.9 Concentrating protein

To concentrate protein Amicon ultracel regenerated cellulose centrifugal filter units (Merck) with a 15mL, 4mL or 0.5 mL capacity and a 3kDa molecular weight cut-off were used. The protein was placed in pre-equilibrated (with whichever buffer the protein was in) centrifugal units which were then spun in a SIGMA 3-18K centrifuge with swing out buckets or a benchtop Eppendorf Minispin, for 15 minutes intervals at 4000G. The flow through was disposed and the concentrated protein retained in the filter segment, protein was pipetted between each spin to mix sample and prevent filter blocking. This process was repeated until volume reduced, and concentration increased, to a desired level.

2.10 Construct and plasmid isolation

The QIAprep Spin Miniprep kit (Qiagen) was used to isolate constructs and empty plasmids when required. Using the Minispin centrifuge from Eppendorf, where required.

3 Results Chapter: *In silico* alanine mutagenesis of binding residues in a chimeric CBD protein

3.1 Introduction

Proteins have evolved to generally encode one characteristic (native) functional folded tertiary structure, typically soluble [311]. Misfolded (including entirely unfolded, denatured, and partially folded) proteins typically form insoluble aggregates, so it is important when working with recombinant protein expression systems that every effort is made to maximise correctly folded native state protein.

This is why optimisation of culture conditions is commonplace when troubleshooting insoluble protein expression, slowing down protein expression can remedy ‘delays’ in folding caused by demand for the saturated cellular protein folding machinery (i.e. ribosomes and chaperone proteins, only a fixed/ limited number available in the cell cytoplasm [281]). This delay period means polypeptides remain unfolded ‘waiting’ for folding to proceed. Proteins in the unfolded state at high concentration, are prone to aggregation (hence insolubility) due to exposure of hydrophobic surfaces that are normally buried in the native state [285].

Also the rate of protein biosynthesis in prokaryotes is significantly faster than in eukaryotes, comparison of the rates of in vitro refolding of orthologous prokaryotic and eukaryotic proteins indicates that the former refolds six times faster [282]. This suggests that the rate of folding correlates with the rate of elongation of polypeptide chains [283]. So when expressing eukaryotic proteins in bacteria issues pertaining to incorrect folding and insoluble POI recovery might be due to this combination of fast

synthesis and delayed folding, which facilitates/favours aggregation [284]. A fine balance exists between protein synthesis and protein folding.

Protein folding is determined by the physicochemical properties encoded in a protein's amino acid sequence. The chemistry of the amino acid side chains is critical to protein folding. Side chains can interact with other side chains via weak noncovalent bonds [312] e.g. forming hydrogen or ionic bonds and van der Waals attractions, driving the folding and intramolecular binding of the linear polypeptide (amino acids joined via peptide bonds) chain ultimately determining the protein's shape. These interactions place constraints on the protein and although any single one of these bonds is weak, many of them collectively act in parallel to hold two regions of a polypeptide together in a strong bond arrangement [312]. Another important factor governing the folding of a protein is the distribution of polar and nonpolar amino acids. The nonpolar (hydrophobic) side chains in a protein e.g. leucine, phenylalanine, tryptophan and valine have a tendency to cluster in a hydrophobic core region [312]. Steric limitations exist on the bond angles permitted in a polypeptide chain (**Figure 14**).

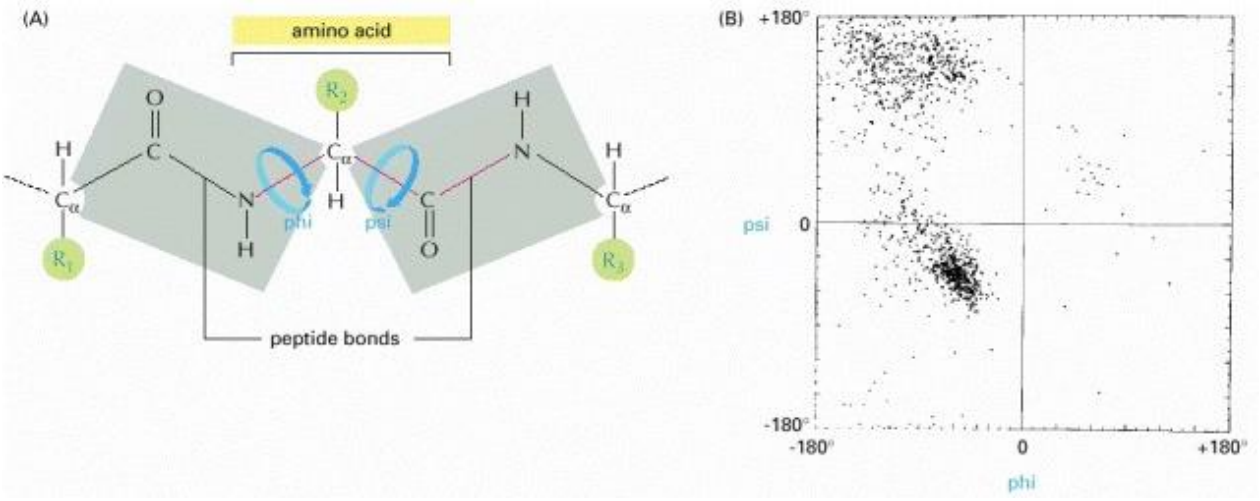


Figure 14: Polypeptide bond angles and steric limitations. A polypeptide chain has an amino terminus and a carboxy terminus. When a peptide bond forms the carboxy group of one amino acid condenses with the next in an oxidative environment. **(A)** From [312] In the peptide bond rotation is not allowed, but at the C_α–C bond is permitted, known as the psi ψ angle of rotation. Rotation is also seen at the N–C_α bond, known as the phi ϕ angle of rotation. An R group (shown in green circles*)* is commonly used to denote an amino acid side chain. **(B)** A typical Ramachandran plot, taken from Richardson, 1981 [313]. A Ramachandran plot is an assessment of model validity employed later in this chapter. Many of the angle combinations, and therefore the conformations of residues, are not possible because of steric hindrance. A Ramachandran plot shows which torsional angles are permitted and can obtain insight into the structure of proteins [314].
The conformation of the main-chain atoms in a protein is determined by one pair of ϕ and ψ angles for each amino acid; (because of steric collisions between atoms within each amino acid, most pairs of ϕ and ψ angles do not occur). Each dot in the plot represents an observed pair of angles in a protein. The plot is a means of checking for allowed structural configuration, biophysical feasibility and identifying any unfeasible/ disallowed torsional angles.

The prevailing accepted theory of protein folding is the thermodynamic hypothesis. This hypothesis states that, it is thermodynamic principles that dictate protein folding that will naturally gravitate to the state of lowest free energy (ΔΔG), and the final structure adopted by a protein is therefore typically the most energetically favourable one [315]. More recently an alternative non-equilibrium hypothesis of protein folding has been postulated, under which the native state of

most proteins does not occupy the global free energy minimum, instead a local minimum on a fluctuating free energy landscape [316].

In the process of folding, a protein adopts a range of conformations before reaching its final, stable, and unique native form. This folding occurs within microseconds to minutes [317], Levinthal's paradox states that it would not be possible by a random search of the enormously large number of possible structures for a protein to reach its native functional state in this timeframe [318, 319]. Giving rise to the typical funnel shaped energetic bias folding diagram widely accepted and adopted within structural biology as the accepted model of protein folding [320] (**Figure 15**). It is the functional native form protein science is most concerned with [321].

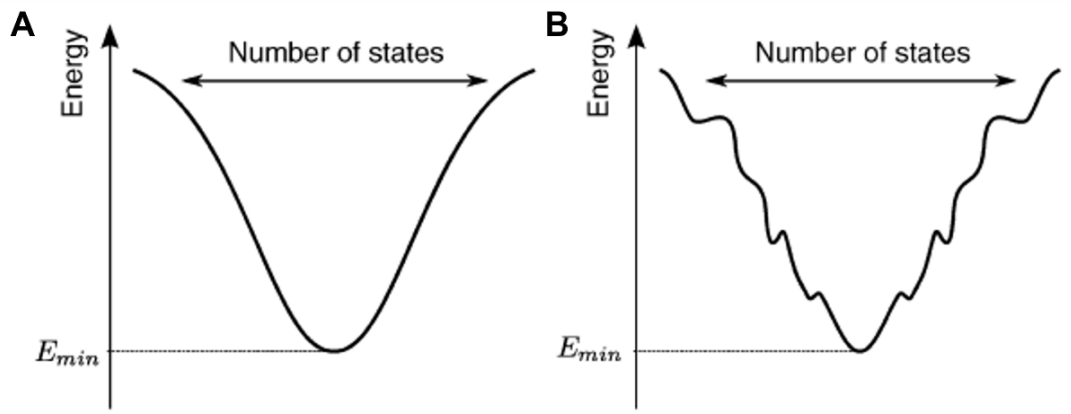


Figure 15: Energy driven protein folding funnel models. The funnel schematic is now the most widely accepted protein folding model available. The figure here is adapted from [318]. At the mouth of the funnels in both models **A** and **B** the polypeptide chain starts as an ensemble of unfolded conformations with high free energy, the number of possible states or different conformations is high (also here there is high entropy). A thermodynamic result is that proteins in the folding pathway can pass via different routes and form different intermediates, but ultimately all pathways lead to the same goal, which is the natural native (lowest energy) state of the protein. **A** shows the idealised energy landscape folding models. This is a simplistic two state folding process whereby the unfolded polypeptide chain rapidly adopts the lowest energy native state represented by the bottom of the funnel. **B** shows the alternate rugged energy landscape folding model, with multiple routes through a funnelled energy landscape. This landscape features energy troughs, which are fold states that proteins can become trapped in, delaying time taken for proteins to reach their native state and can contribute towards aggregation states [322].

Experimental determination of protein 3D structures is time consuming, labour-intensive and requires use of complicated expensive high-resolution techniques, namely protein crystallisation followed by X-ray diffraction, isotopically enriched protein NMR [323] or cryo electron-microscopy (cryo-EM) [324]. The structures of around 100,000 unique proteins have been determined to date experimentally [325], only a fraction of the billions of known protein sequences. Where an experimentally determined 3D model is lacking as here with 222, computational approaches can be utilised to more quickly fill the gap and give high-resolution increasingly accurate models [326, 327].

The number of different protein fold families appearing in nature is limited and there are well documented trends in 3D structure and folds [328]. This combined with our improved understanding of the principles of protein structure e.g. hydrophobic patterning, local residue interactions that bias short stretches of the chain towards forming specific secondary structures or rotamers and side chain torsional preferences [329], is why accuracy of models is improving and modelling progressing [330]. It is now well documented and known there is structural similarity between homologous proteins [331, 332]. Therefore structure can be inferred from sequence similarity; this process is known as homology modelling [333, 334] or comparative protein modelling [335]. Providing models comparable to low resolution X-ray crystallography or medium resolution NMR derived structures.

Homology modelling is a method whereby 3D models of proteins are determined from amino acid sequence and no physical experimental work, beginning with sequence alignment with a protein homologue (with a known structure). This known structure is utilised as the modelling template. Protein structures are more conserved than amino acid sequences so a small change in sequence often has little effect on structural conformation, meaning homologues can often be used to give accurate models of proteins [328].

Homology or comparative modelling, is reliant on detectable similarity spanning most of the target sequence and at least one known structure [336]. The necessary similarity between protein sequences for successful homology modelling is generally placed at a >30% identity threshold [337, 338]. It is limited to those sequences that can be confidently mapped to already experimentally derived structures available in

the Protein Data Bank (PDB). PDB is the largest database of experimentally resolved protein structures [339].

The scheme of the work in this chapter is summarised below in **Figure 16**.

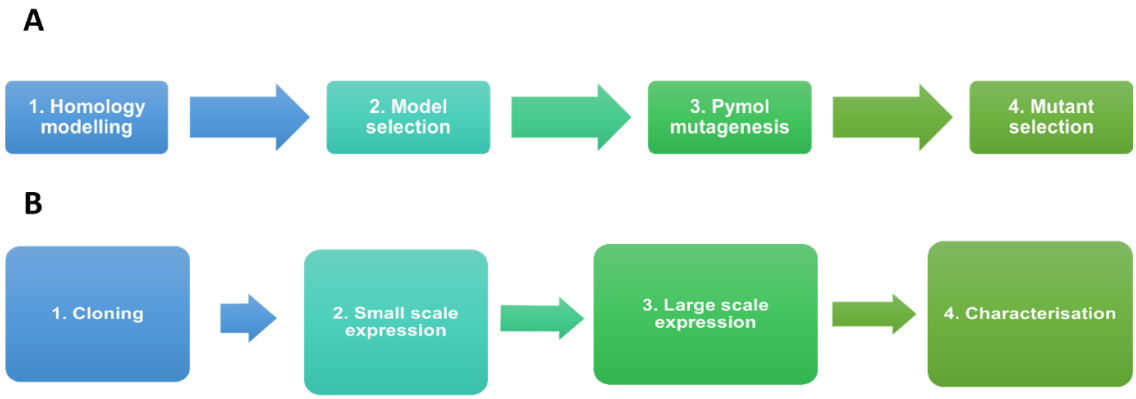


Figure 16: A, *In silico* pipeline for this chapter. B, *In vitro* pipeline for this chapter.

3.2 Methods

3.2.1 Modelling 222

Figure 17 shows a summary of the homology modelling process used in the work here. Homology modelling is a multistep process that can be summarised into six steps.

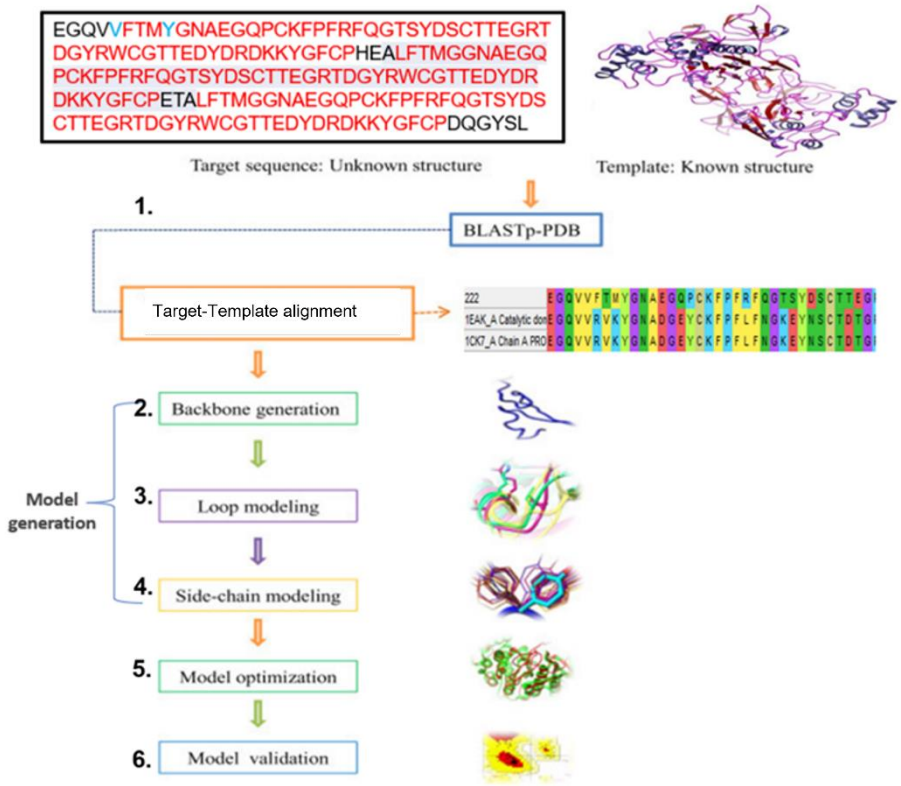


Figure 17: Homology modelling summary (adapted from [328]). 222 sequence and 2 best aligning structures 1EAK and 1CK7 shown partially (full alignment shown in **results section** 3.3.1). Homology modelling is a multistep process, that can be summarized in six steps: template identification and alignment, backbone generation, loop modeling, side chain modeling, model optimization and model validation.

3.2.2 Template identification & alignment

The protein sequence of 222 (**Figure 18**) was submitted to the BLASTp (Basic Local Alignment Search Tool, protein-protein) online search tool (available at https://blast.ncbi.nlm.nih.gov/Blast.cgi , accessed 12/2018), with the PDB database selected [340]. This proceeds via pairwise comparison, aligning the target sequence with all the sequences in the database of known structures. This search assigns the likely fold of the target sequence.

(A) **LFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTT EDYDRDKKYGFCP**

(B) **EGQVVFTMYGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRW CGTTEDYDRDKKYGFCPHEALFTMGGNAEGQPCKFPFRFQGT SYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPETALFTMGG NAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRD KKYGFCPDQGYSL**

Figure 18: Module 2 and 222 protein sequences. (A) Shows the sequence of module 2. **(B)** Shows the amino acid sequence of 222. The two residues shown in blue text were conserved from the native CBD protein, as they were known to be involved in key intramolecular interactions [235]. The blue valine residue is part of module 2 (in the position of CBD module 1) in place of what would have been a leucine had the interaction not needed to be maintained. The three module 2s, that make up 222 are shown in red (first module 2 in the position of CBD module 1), green (module 2 in the position of CBD module 2) and yellow text (module 2 in the position of CBD module 3). Then the linker region residues are shown in black text. Linker regions are short amino acid sequences separating multiple protein domains/ modules, in a single protein [341]. Most linker regions are rigid and function to prohibit unwanted interactions between the discrete domains/ modules [342].

The strongest BLASTp match was then selected for use as the template structure (downloaded as a .PDB file) using the expect (E) value. E value is a parameter that describes the number of hits "expected" due to chance alone when searching a database of a particular size. The lower the E value, the more significant the score and the alignment. The best identified template structure and alignment here were then taken to the next step and used in model generation.

3.2.3 Model generation & optimization

Next models of 222 were generated using the Modeller software package. The preceding target-template alignment step was necessary to generate the .ali file (**Figure 19**). Here two additional templates were constructed from the identified 1EAK alignment, to augment the modelling process maximising chance of model accuracy. The first 1eak_fnII provides modeller with details of how the three domains making up the 222 protein are arranged with respect to each other 1-2-3. Then to make 222 the middle domain is already correct but rather than simply instruct modeller to make models of 222 based just upon the structure of module 2 a three-template approach was utilised. Whereby two extra copies of module 2 superimposed onto the positions of module 1 and 3 in the original 1EAK structure were used. These are outlined in the .ali file shown in shown in 1eakdom2_fit1 (providing a fit onto domain 1) and 1eakdom2_fit1 (providing a fit onto domain 3). Modellers automodel class was used as it is the simplest means to build models, with automation of many steps meaning minimal user intervention was necessitated and only basic knowledge of the Python scripting language was sufficient. The alignment file was submitted to the Modeller software to define the relative

orientation of the three module 2 domains in 222 models. Using the alignment file (.ali) shown in **Figure 19** and a python file shown in **Figure 20** (.py created to drive the modelling), five models (.pdb files) were generated with Modeller 9.21 [343] (available at http://salilab.org/modeller/download_installation.htmL accessed 12/2018).

A

```
>P1;222
sequence:222::::::::
EGQVVFTMYGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPHE
ALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPETALF
TMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPDQGYSL
*
```

B

```
>P1;1eak_fnII
structureX:1eak_fnII:217:A:397:A::::
EGQVVRVKYGNADGEYCKFPFLFNGKEYNSCTDTGRSDGFLWCSTTYNFEKDGKYGFCPHE
ALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPETAMS
TVGGNSEGAPCVFPFTFLGNKYESCTSAGRSDGKMWCATTANYDDDRKWGFCPDQGYSL
*
```

C

```
>P1;1eakdom2_fit1
structureX:1eakdom2_fit1:278:A:333:A::::
---ALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCP--
------------------------------------------------------------
------------------------------------------------------------
*
```

D

```
>P1;1eakdom2_fit2
structureX:1eakdom2_fit2:278:A:333:A::::
------------------------------------------------------------
------------------------------------------------------------
ALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCP-----
-
*
```

Figure 19: .ali file used during model generation. Provides the identified template structure and target 222 sequence in a PIR database format, readable by the modeller software. The .ali file maps the sequence of the target onto the template structure. The second line of each entry contains information necessary to extract the atomic coordinates of the segment from the original PDB coordinate set. The fields in this line are separated by colon characters. Each pair of lines in this file is known as a sequence entry, providing Important alignment details. The * symbol shows the end of each entry sequence. **A** provides the query/ target sequence. StructureX tells modeller that the structure was determined by X-ray crystallography. **B, C** and **D** are 222 alignment variations, assembled from the Blast alignment results (**Figure 28**).

```
# Comparative modeling with multiple templates
# Load standard Modeller classes from modeller.automodel
from modeller import *

# Load the automodel class
from modeller.automodel import *

log.verbose()
env = environ(rand_seed=-122)
# directories for input atom files
env.io.atom_files_directory = [".", "../atom_files"]

a = automodel(
    env,
    # alignment filename
    alnfile="3templates.ali",
    # codes of the templates
    knowns=("1eakdom2_fit1", "1eakdom2_fit2", "1eak_fnII"),
    sequence="222",
)  # code of the target

a.starting_model = 1
a.ending_model = 5
a.make()
```

Figure 20: .py file. MODELLER is a command-line only tool, with no graphical user interface; so, a script file needed to be provided containing MODELLER commands. This .py file is a simple script providing modeller with the parameters of the 222 sequence, alignment, and instructions on how to proceed with modelling specifying that the automodel class be used and to create an environment, automodel parameters, number of models to generate and a.make() tells modeller to initiate the modelling.

3.2.4 *In silico* percentage secondary structure assessment

The models generated using Modeller (pdb output files, **results section 3.3.2**) and later 222 alanine mutants (**methods section 3.2.6**, shown in **results section 3.3.3**) were assessed for differences in secondary structure using the 2StrucCompare web server (https://2struccompare.cryst.bbk.ac.uk/index.php accessed 12/2018) [344].To highlight any differences between the models in terms

of percentage of each secondary structure type (not necessarily immediately apparent visually in pymol if subtle).

3.2.5 Validation

After the five models were built, it was important to check them for errors. Sequence identity above 30% is a relatively good predictor of the expected accuracy of a model [336]. The five generated models were then assessed for quality using QMEAN (Qualitative Model Energy Analysis) Version 2.5.1, QMEAN is an online comprehensive scoring function tool (https://swissmodel.expasy.org/qmean/, accessed 12/2018) [345]. The QMEAN tool performed a ranking of inputted models and highlighted potentially problematic regions for each model. QMEAN uses several different structural descriptors to assess model quality. The descriptors are local geometry, solvation potential assesses the burial status of residues, the agreement between model secondary structure and solvent accessibility [345]. A QMEAN Z-score around zero indicates a good agreement between the model structure and experimental structures of similar size.

Ramachandran plots were also generated as a secondary validation check, this was completed using the online Rampage Ramachandran plot analysis tool (http://mordred.bioc.cam.ac.uk/~rapper/rampage.php, accessed 12/2018). The Ramachandran principle on which plots are built states that alpha helices, beta strands, and turns are the most likely conformations for a protein to adopt, because most other conformations are impossible due to steric collisions between atoms [346]. Ramachandran plots provide a means to visualise energetically allowed regions for backbone amino acid dihedral (torsion) angles ψ (psi) against φ (phi)

within a protein structure [347]. From both these analyses the strongest 222 model was selected and used as the basis for designing mutants.

3.2.6 Generation of 222 binding residue alanine mutants

The modelled structure was used to conduct *in silico* alanine scanning mutagenesis with the pymol mutagenesis wizard. Previous NMR ligand binding studies (looking at chemical shift upon binding) carried out by the group with 222 identified fifteen residues that are involved in its binding to TII gelatin (**Table 8**) [235]. Five of these fifteen identified residues were glycine, which despite binding involvement were not considered for mutation here given that glycine is a smaller residue than alanine. Hence switching these residues to alanine, was unlikely to meet the primary aim of alanine mutagenesis, which was to identify residues important for protein function (swapping these to alanine would reduce binding functionality/ affinity most drastically). The remaining ten residues were substituted concurrently for alanine in all the equivalent positions, in all three module 2's of 222. Each generated mutant was saved in .pdb format for further bioinformatic use.

Table 8: 222 Binding Residues. Residues identified previously by the Hollander group using NMR as being involved in binding of 222 to TII gelatin [235]. The glycine residues that were excluded from mutagenesis are highlighted in red text.

Residue	Residue Number		
	1st module	2nd module	3rd module
G	10	68	126
N	11	69	127
R	22	80	138
F	23	81	139
G	25	83	141
Y	28	86	144
G	35	93	151
G	39	97	155
Y	40	98	156
W	42	100	158
T	46	104	162
E	47	105	163
Y	49	107	165
Y	55	113	171
G	56	114	172

3.2.7 Selecting mutants

Computational assessment was carried out to evaluate differences in stability and residue solvent exposure between the ten alanine mutant proteins. Only the strongest mutant candidates were selected to take forward to *in vitro* testing. **Figure 21** summarises the numbers and names of mutants generated and selected *in silico* for subsequent *in vitro* testing.

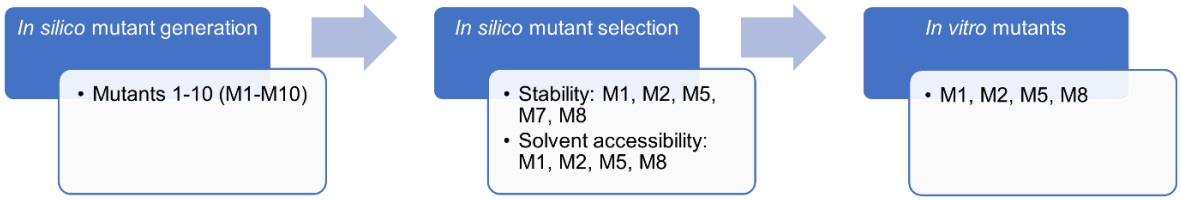


Figure 21: Summary of number and names of mutants developed during the in silico and in vitro work. Ten mutant's mutant 1-10 (M1-10) were initially generated *in silico.* Five mutants selected based on stability assessment made *in silico.* Then only four of these mutants were selected based upon solvent accessibility assessment *in silico*. These four double selected mutants M1, M2, M5 and M8 were then taken forward to *in vitro* experiments.

3.2.8 Assessing mutants for stability

Stability assessment was made using two different online software packages Popmusic (http://dezyme.com/, accessed 01/ 2019) and Maestro (https://biwww.che.sbg.ac.at/maestro/web, accessed 01/2019) [348]. Both tools required submission of a native (non-mutant, 222) PDB file and input of the mutation sites, wild-type, and mutant residues.

With Popmusic manual mode was chosen and then the residue number, natural amino acid (aa) and mutant aa inputted. With Maestro, evaluate specific mutations was selected as the task. Next the specific mutations for each mutant were listed and the evaluate combined mutations option selected. With both tools each mutant 1-10, with three substitutions (one per each of the three module 2's in 222), was assessed individually. For each mutant once the three mutations were inputted the query was processed.

3.2.9 Assessing mutants for solvent accessibility

The remaining mutants selected based on stability were then further assessed for solvent exposure to ensure the mutated residues were not buried in any of the three

domains. There are no exposure thresholds defined for buried or exposed residues for either tool used, a threshold of 20% was adopted taken from a recent publication [292] To assess solvent exposure of mutants Popmusic and Parameter Optimized Surfaces (POPS) (https://mathbio.crick.ac.uk/wiki/POPS accessed 01/2019) [349] tools were used. POPS is an online tool that assesses solvent exposure of the residues in a submitted protein structural model (pdb file), by calculating solvent accessible surface areas (SASAs). Q(SASA) is the quotient of SASA and Surf i.e., the fraction of SASA. Surf being the surface area of an isolated atom, which in this case are the mutant alanine residues. The residue numbers shown in **Table 8** were used to identify the QSASA scores of the mutations only. As all 3 modules were mutated at the same residue in each module the QSASA scores were averaged across the three modules in each mutant.

3.2.10 Expression of mutants

3.2.10.1 Codon optimisation

The synthetic genes for all mutants and cloning undertaken in this thesis were ordered from GeneArt (Thermofisher). As part of the ordering procedure all were optimised for *E.coli* expression using the GeneArt GeneOptimizer tool. The genes ordered for the work in this chapter are shown in **Table 9**.

Table 9: 222 alanine binding residue mutant genes. GeneOptimizer codon optimised DNA sequences ordered to construct pOPINS binding mutants selected for *in vitro* testing. Nucleotide codons highlighted in yellow are those that encode the mutant ala in the equivalent position of all three module 2's that make up the chimeric protein 222.

Name	DNA Sequence
Mutant1 (M1)	GAAGGTCAGGTTGTGTTTACCATGTATGGTGCAGCCGAAGGTCAGCCGTGTAAATTTCCGTT TCGTTTTCAGGGCACCAGCTATGATAGTTGTACCACCGAAGGTCGTACCGATGGTTATCGTT GGTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCC CTGTTTACAATGGGTGGCGCAGCAGAGGGCCAGCCTTGCAAATTCCCTTTTCGCTTCCAGGG TACATCTTATGATTCATGCACAACGGAAGGTCGCACAGATGGCTACCGCTGGTGCGGCACCA CAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACCGCACTGTTCACCATG GGAGGTGCTGCGGAAGGCCAACCATGCAAATTTCCATTCAGATTTCAGGGTACAAGCTACGA TTCATGTACTACTGAAGGCAGAACGGATGGATATAGATGGTGCGGTACAACCGAGGACTACG ATAGAGATAAGAAATATGGTTTCTGTCCCGATCAGGGTTATAGCCTG
Mutant 2 (M2)	GAAGGTCAGGTTGTGTTTACCATGTATGGTAATGCCGAAGGTCAGCCGTGTAAATTTCCGTTT GCATTTCAGGGCACCAGCTATGATAGTTGTACCACCGAAGGTCGTACCGATGGTTATCGTTG GTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCCC TGTTTACAATGGGTGGCAATGCAGAGGGCCAGCCTTGCAAATTTCCTTTCGCCTTCCAGGGT ACATCTTATGATTCATGCACAACGGAAGGTCGCACAGATGGCTACCGCTGGTGCGGCACCA CAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACCGCACTGTTCACCATG GGTGGTAATGCGGAAGGACAACCATGCAAATTCCCATTTGCGTTTCAAGGTACGTCATATGA TAGCTGCACAACAGAGGGACGTACGGATGGATACAGATGGTGCGGTACAACCGAGGACTAC GATAGAGAAAGAAATATGGTTTCTGTCCCGATCAGGGTTATAGCCTG
Mutant 5 (M5)	GAAGGTCAGGTTGTGTTTACCATGTATGGTAATGCCGAAGGTCAGCCGTGTAAATTTCCGTTT CGTTTTCAGGGCACCAGCTATGATAGTTGTACCACCGAAGGTCGTACCGATGGTGCACGTTG GTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCCC TGTTTACAATGGGTGGCAATGCAGAGGGCCAGCCTTGCAAATTCCCTTTTCGCTTCCAGGGT ACATCTTATGATTCATGCACAACGGAAGGTCGCACAGATGGCGCTCGCTGGTGCGGCACCA CAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACCGCACTGTTCACCATG GGTGGTAATGCGGAAGGACAACCATGCAAGTTTCCATTCCGCTTTCAGGGAACCTCATATGA TAGCTGCACAACAGAGGGACGTACGGACGGTGCTAGATGGTGCGGTACAACCGAGGACTAC GATAGAGATAAGAAATATGGTTTCTGTCCCGATCAGGGTTATAGCCTG
Mutant 8 (M8)	GAAGGTCAGGTTGTGTTTACCATGTATGGTAATGCCGAAGGTCAGCCGTGTAAATTTCCGTTT CGTTTTCAGGGCACCAGCTATGATAGTTGTACCACCGAAGGTCGTACCGATGGTTATCGTTG GTGTGGCACCACCGCAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCCC TGTTTACAATGGGTGGCAATGCAGAGGGCCAGCCTTGCAAATTCCCTTTTCGCTTCCAGGGT ACATCTTATGATTCATGCACAACGGAAGGTCGCACAGATGGCTACCGCTGGTGCGGTACAAC AGCCGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACCGCACTGTTCACCATGG GTGGTAATGCGGAAGGACAACCATGCAAGTTTCCATTCCGCTTTCAGGGAACCTCATATGAT AGCTGCACAACAGAGGGACGTACGGATGGATACAGATGGTGTGGTACGACTGCGGATTACG ATAGAGATAAGAAATATGGTTTCTGCCCAGATCAGGGTTATAGCCTG

3.2.10.2 Primer design

Forward (fwd) and reverse (rev) primers (**Table 10**) were designed, ordered from sigma, and used to subclone each gene into the pOPINS vector via infusion cloning (Takara Bio). Cloning tags for the forward and reverse primers were available on the OPPF pOPIN suite (https://www.oppf.rc-

harwell.ac.uk/OPPF/protocols/cloning.jsp accessed 02/2019). For the forward primer the first 20 nucleotides from the start of the insert sequence were taken. For the reverse primer the reverse complement of the last 18 nucleotides of each insert sequence were used. To make the infusion cloning primers the two sequences (cloning tag/ vector nucleotides followed by insert nucleotides) were combined. For all mutants in this chapter the overlap regions were the same so the same primers could be used to clone all four mutants.

When designing primers there are some general considerations outlined below which were applied here to improve chances of cloning success. The melting temperature (Tm) of the insert specific sequence was kept between 50-65°C or as near to as possible (Tm determined using the Genscript Oligo Calculation Tool available at https://www.genscript.com/tools/oligo-primer-calculation accessed 02/2019). Difference in Tm between forward and reverse primer Tms were kept to ≤4°C. A GC content between 40–60% was used and primers were always designed to end in either a G or C nucleotide.

Table 10: Infusion Cloning primers used with genes in results chapter 3.
pOPINS vector region (black text) and insert region (red text).

Mutant	Vector	Forward (fwd) Primer	Reverse (rev) Primer
M1	pOPINS	GCGAACAGATCGGTGGTG AAGGTCAGGTTGTGTTTAC	ATGGTCTAGAAAGCTTTA CAGGCTATAACCCTGATC
M2			
M5			
M8			

3.2.10.3 Gene amplification

The gene inserts were amplified via the polymerase chain reaction (PCR) in preparation for the Infusion cloning reaction. The PCR reaction required: forward and reverse primers (**Table 10**) (1µL of each, at 30µM), DNA insert (0.25µL at a concentration of 2ng/µL), 12.5µL of CloneAmp HiFi PCR Premix (a 2x PCR master mix, included with the infusion cloning kit, Takara Bio). The premix contains the DNA polymerase enzyme, optimized PCR buffer, and deoxynucleoside triphosphates (dNTPs). The remainder of the 25µL reaction volume is then made up of 10.25µL sterile MQ water. PCR was carried out using a SureCycler 8800 (Agilent technologies) following the procedure shown in **Table 11**. Following the PCR a 1% Agarose gel was ran (as described next in **methods section 3.2.10.4**, with 10uL PCR product mixed with 1.7uL 6x DNA loading buffer, NEB) to confirm that the resultant insert product size was correct and amplification had been successful (as indicated by band intensity when imaged).

Table 11: PCR gene amplification thermocycler procedure. Including temperatures, duration, explanation, and No. of cycles. For the primer annealing step the lower Tm of the pair is used.

Temperature	Duration	Explanation	No. of cycles
98°C	2min	Activates DNA polymerase	1
95°C	30sec	Denatures dsDNA	29
Primer Specific	30sec	Primer annealing	
72°C	1min 30sec	Extension and elongation	
72°C	2min	Final extension	1

3.2.10.4 Agarose gel electrophoresis

A 1% agarose gel was made with 1g of agarose (Bioline) dissolved in 100mL of 1x TAE buffer (2M Tris, 0.1M ethylenediaminetetraacetic acid (EDTA), 1M acetic acid), heated in a microwave with intermittent pauses for mixing, until fully dissolved (1-3min). The agarose was then cooled to ~50°C, and 5µL of Midori Green (NIPPON genetics) was added and mixed. This solution was used to cast the gel in a tray (Bio-Rad), a large tooth comb was inserted to form wells and the gel allowed to set at RT for ~30mins. The gel was placed in the gel box unit and filled with 1xTAE buffer (until gel covered). The gel was run at 120V in 1x TAE buffer for 1h to allow sufficient migration and separation of the DNA (PowerPac 300 from Biorad). The gel was examined and imaged using a Biorad Chemidoc MP (UV) gel imaging system.

3.2.10.5 Vector linearisation

The vector map for the pOPINS vector used in the work in this chapter is shown in **Figure 22**.

Figure 22: pOPINS vector used in the work in this chapter. pOPINS 5906bp, Kanamycin resistance, T7-Lac promoter [8]. Linearised by KpnI recognition sequence GGTACC and HindII recognition sequence AAGCTT. Digest products were a 333bp cut out (lac z gene) and 5573bp linearised vector. The parent vector of pOPINS is pET28a. This figure was created using the SnapGene software (from Insightful Science; available at www.snapgene.com).

Table 12 gives further details of the pOPINS vector, pOPINS was selected as the vector as it means resultant expressed protein of interest (POI) is both His and SUMO tagged easing purification and potentially aiding solubility [350, 351].

Table 12: pOPINS vector details. Including source, restriction enzymes and recognition sites, parent vector/ antibiotic resistance, digest products, digest products, promoter, Inducer, expression product, cleavage enzyme and references.

Vector	Source	Restriction enzyme 1 and recognition site	Restriction enzyme 2 and recognition site	Parent vector/ Antibiotic resistance	Digest products	Promoter	Inducer	Expression product	Cleavage enzyme	Ref
pOPINS	OPPF	HindIII AAGCTT	KpnI GGTACC	pTriEx2/Amp	333bp cut out, 5573bp linear vector	T7	IPTG	His-GST-POI	3C protease	[352, 353]

Vector linearisation was carried out using two restriction endonucleases, in a double digest removing an unrequired 333bp segment and linearizing the vector in the process (**Figure 23**).

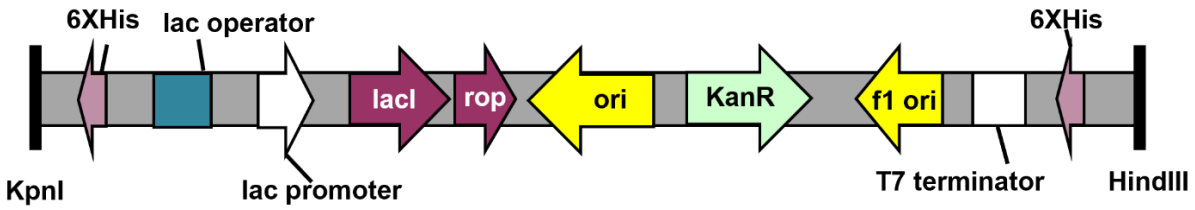


Figure 23: Linearised pOPINS vector. 333bp fragment removed during linearisation 5573bp linear vector remaining. Kanamycin resistance, required for transformed clone screening. Lac promoter and terminator required for expression of insert/POI; expression inducible by Isopropyl β-D-1-thiogalactopyranoside (IPTG).

For this restriction digest the following was added to a clean eppendorf; empty vector (1µg DNA), 10x cutsmart buffer (5µL), both restriction endonucleases (NEB, High-fidelity,1µL of each) for the double digest or just one for the single cut reference, then MQ water (to make total reaction volume up to 50µL). This reaction was then incubated for 1h at 37˚C and the reaction stopped by adding 10µL of 6x Loading buffer (NEB, composed of 50% glycerol, 50 mM EDTA (pH 8.0) and 0.05% bromophenol blue). Gel electrophoresis (**methods section 3.2.10.4**) was used to confirm linearisation of the vector, before proceeding with linear vector DNA extraction. Hyperladder marker (Bioline) was loaded as a size reference, followed in the adjacent wells by samples of both single cut and double digested vector mixed with (10µL) 6x Loading buffer. Double digested linearized plasmid bands were cut from the gel with a gel cutting tip (Corning life sciences) on a UVP benchtop UV transilluminator. Then linearised vector was isolated from the cut gel using the QIAquick gel extraction kit (Qiagen).

3.2.10.6 Infusion cloning

Infusion cloning is a ligase independent cloning method utilised here due to its high accuracy, easy and versatile application workflow. **Figure 24** shows an overview of the infusion cloning procedure. Firstly, enhancer from the infusion cloning kit was used to clean up the PCR product (2µL of cloning enhancer added to 5µL of amplified PCR product) and incubated on a thermal cycler for 15 mins at 37°C, followed by 15 mins at 80°C, held at 4°C then placed on ice. Then to a clean eppendorf linearised plasmid at a minimum concentration of 20ng/µL (3µL), Infusion enzyme (1µL) and cloning enhanced PCR product (1µL) were added. This cloning reaction was incubated on a thermal cycler at 50°C for 15 mins to activate the infusion enzyme and allow cloning to proceed. Infusion cloning is a type of homology-based cloning as homologous overlaps were necessary to allow recombination to occur between the vector and mutant insert resulting in a circularised desired mutant plasmid construct.

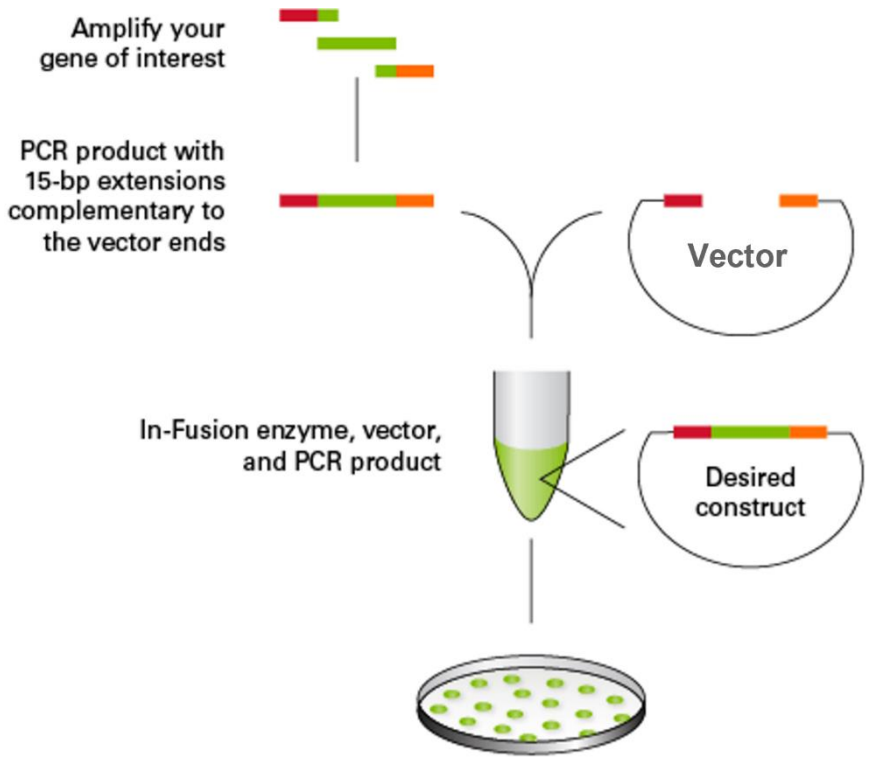


Figure 24: Overview of the infusion cloning procedure. Starting with gene amplification, proceeding to cloning and then transformation. Image taken from [354].

After the cloning reaction stellar cells (Takara Bio) were transformed, onto LB agar plates containing 50µg/µL kanamycin antibiotic using the transformation method outlined below**.** Construct-containing cells formed colonies on the agar, which were individually selected and used to inoculate three separate 50mL falcons containing 5mL of LB media and 5µL antibiotic, these were then incubated at 37°C, 180RPM O/N. Clones of the assembled construct were isolated from these three separate cultures using the miniprep II Kit from Qiagen, 10µL of each of these isolated constructs were sent for sequencing confirmation (Eurofins GATC, Germany). Once sequencing was confirmed, small scale expression testing and optimisation was commenced.

3.2.10.7 Transformation

1.5µL of construct was added to 50µL of *E. coli* cells partially thawed on ice for approx 3-5mins, (the different *E. coli* strains used in the work of this chapter are shown in **Table 13**) incubated on ice for 30mins before being heat shocked on a heatblock at 42°C for 30secs. Then heat shocked cells were replaced on ice for 3mins, 450µL of SOC media added and eppendorfs moved to a shaking incubator at 37°C, 180RPM for 1h. Following growth these cultures were spun down in a benchtop centrifuge (Minispin, Eppendorf) at 4000RPM for 30secs. 450uL of supernatant was disposed and cell pellet was resuspended by thorough pipetting the 50µL of remaining supernatant. Cells were transferred and spread onto a kanamycin LB agar plate, inverted, and incubated O/N at 37°C.

Table 13: Details of different E.coli cell strains used in this chapter. Including application, supplier, key features, and references. All cells were purchased first then replenished using the rubidium chloride method of competent cell production.

Cells/ Strain	Application	Supplier	Key features	Ref
Stellar	Cloning	Takara Bio	*E. coli* HST08 strain providing high transformation efficiency	[355]
Shuffle T7	Expression	NEB	*E. coli* K12 strain engineered to promote multiple disulfide bond formation in the cytoplasm	[302, 356, 357]

3.2.10.8 Starter Culture

For small scale expression 10mL of LB with 10µL of kanamycin antibiotic, or for large scale expression 50mL of LB with 50µL of kanamycin at a concentration of

50mg/mL (both to achieve a working concentration of 50µg/µL kanamycin) were aliquoted aseptically into a sterile 50mL falcon or a 250mL Erlenmeyer flask. Then a single bacterial colony from the transformation plate was added using a sterile pipette tip. This culture was incubated O/N with shaking at 37°C, 180RPM, to act as a starter culture to inoculate larger cultures the next day.

3.2.10.9 Small scale expression trials

Two different temperatures and durations for expression, and a range of Isopropyl β-D-1-thiogalactopyranoside (IPTG) concentrations were trialled for each protein, shown in **Table 14**. This table shows the initial test conditions used for all proteins in this work. During these trials 100µL of overnight starter culture was added to 10mL LB containing 10µL Kan (stock concentration of 50mg/mL, working concentration of 50µg/µL Kan) in a sterile 50mL falcon (eleven samples in total, with one allocated as surplus for OD checks only). These cultures were Incubated at 37°C, 180RPM shaking. OD was monitored and when OD reached 0.6 expression was induced by adding the volumes of IPTG as shown in **Table 14**. Half of the samples (five cultures) were grown after induction incubated at 25°C for 5h before harvesting. The other half of these samples (five cultures) were grown after induction O/N at 18°C. Cell harvesting was performed by centrifugation for 5mins at 8000RPM, at 4°C, disposing supernatant, and retaining pellet at -80°C, until purification was conducted.

Table 14: Small scale expression trial optimisation conditions. Expression duration, expression temperature, range of IPTG concentrations and the volume of 1M IPTG added to the 10mL small scale cultures.

Time	Temp (◦C)	IPTG Concentration	20µM	200µM	400µM	800µM	1000µM
5h	25	Volume of 1M IPTG to add (µl)	0.2	2	4	8	10
O/N	18		0.2	2	4	8	10

3.2.10.10 Purification

The first stages of purification exploit the histidine tags affinity for nickel metal, using a technique called immobilised metal affinity chromatography (IMAC) [358]. The binding of nickel to histidine is shown in **Figure 25,** it is this binding that is utilised to separate the His- fusion-POI from the initial crude mixture including E.coli proteins that are non target. This nickel affinity was utilised in the small scale using nickel-NTA-agarose resin and then large scale using a nickel sepharose affinity column format.

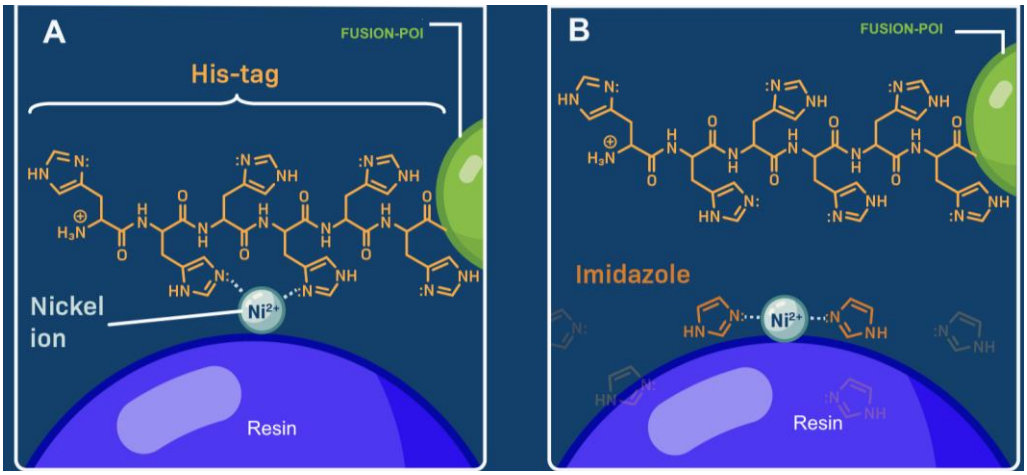


Figure 25: (A) Nickel IMAC purification summary. (A) Mode of binding of histidine to nickel, **(B)** and its subsequent elution with HTB buffer (high imidazole) **(B)**. adapted from [359]. All pOPIN vectors used in this thesis encode a His-fusion tag-POI product so utilise IMAC.

3.2.10.11 Small scale purification

Cell pellets from small scale expression trials were thawed in cold water then moved to ice, resuspended using a vortex in 1mL lysis buffer and transferred to a clean 1.5mL eppendorf tube. Lysis buffer was composed of HisTrap A buffer (HTA: 20mM Na_2HPO_4, 20mM Imidazole, 500mM NaCl, pH 7.4), 1mg/mL lysozyme, 1% Tween (v/v), DNAse (1mg/mL) and Proteoloc protease inhibitor solution (Abcam, 50µL per 20mL of lysis buffer). Samples were then rocked on ice for 30mins. Sonication on ice was subsequently used to lyse the cells, using a small probe, (MSE SoniPrep,150 plus) with the sonicator set to amplitude 13.0. Three 10 second pulses with 30 seconds pause for cooling, between each pulse were utilised to achieve cell lysis.

Resultant lysate was cleared of insoluble proteins and cellular debris via centrifugation at 13,300RPM for 10 minutes (Minispin, Eppendorf benchtop centrifuge), at 4°C. The supernatant (soluble protein) was removed and, aliquoted into clean 1.5mL eppendorf tubes and retained on ice. The insoluble pellet was resuspended (via vortex) in 1mL lysis buffer containing 8M urea before again being centrifuged at 13,300RPM for 10 minutes.

Soluble proteins were purified using Nickel-NTA agarose resin (Qiagen). All centrifuge steps were performed at 3000RPM for 30 secs. 60µL of Ni-NTA agarose resin was aliquoted into 1.5mL eppendorf tubes, prepared by removing storage ethanol, followed by washing in MQ water and then equilibrating in HTA buffer. 800µL of the soluble protein, was then applied to the resin and centrifuged. Non-specific binding was removed by washing the resin twice with 60µL HTA buffer. The

bound fraction was then recovered by adding 30µL 4x SDS (Sodium dodecyl sulfate) loading buffer. SDS-PAGE, see **materials and methods section 2.6**) was then used to identify optimal expression conditions to take to large scale expression (shown in **results section 3.3.6**).

3.2.10.12 Large scale expression

2.5L Thomson ultra-yield flasks were utilised, specially designed to enhance the aeration of cultured bacterial cells. Flasks containing 1L of LB were autoclaved to sterilise then cooled to RT. 25mL of starter culture was added to 1L of LB with 1mL of 50mg/mL kanamycin. Cultures were grown at 37°C with 180RPM shaking until OD of 0.6 was reached, when incubator temperature was reduced to 18°C. Whilst incubators cooled cultures were placed at 4°C. Once cooling completed cultures were induced by adding the optimised IPTG concentration outlined in **Table 15**. At the end of this expression period cells were harvested via centrifugation at 8000RPM for 5mins (Sorvall Lynx 6000 Centrifuge, Thermo scientific). Supernatant was discarded and pellets stored frozen at -80°C until purification.

Table 15: Expression conditions for all proteins produced in this chapter. IPTG concentration, expression temperature and duration. These were determined using small scale expression trials, outlined in **results section 3.3.6**. 222 expression was conducted using the same method as in[235].* indicates 10mM sucrose was added to cultures (following the results of additional optimisation trials with M1 outlined in **methods section 3.2.10.17**, and in **results section 3.3.8.1**).

Gene	IPTG concentration µM	Expression temperature °C	Expression duration
222	1000	18	O/N
M1	800*	16	O/N
M2	400*	16	O/N
M5	400*	16	O/N
M8	400*	16	O/N

3.2.10.13 Large scale purification

All buffers used in purification were prepared fresh, filtered through 0.2µm filter (Millipore) and degassed prior to use on the ÄKTA system. Purification was undertaken in the first instance as previously for 222 [235].

3.2.10.14 Lysis

It is common to see two or more methods being used in tandem to obtain the desired lysis result and maximise recovery, this was the approach here. All pellets were frozen before purification to utilise freeze-thaw as the first purification method [360]. Pellets were thawed in cool water then moved to store on ice. Each 1L culture pellet was then resuspended in 20mL lysis buffer. This was to enact a chemical lysis step; samples were rocked in lysis buffer on ice for 30mins. *E. coli* cells were then lysed using a MSE SoniPrep,150 plus) with the sonicator set to amplitude 13.3.

Three 20 second pulses with 30 seconds pause for cooling, between each pulse were utilised to achieve cell lysis.

Following the M1 scale up issues an alternate lysis method to sonication was also trialled as the third process in the combination lysis process (with freeze thaw and chemical lysis). A mechanical homogenising-based lysis method was utilised, using a Stansted SPCH-EP-10-60 pressure cell homogenizer (Homogenizing Systems Ltd, UK) set to 23kpsi. This was trialled on a 1L culture scale up based on the results of the trials outlined in **methods section** 3.2.10.17 , shown in **results section** 3.3.7. Lysate was loaded 9mL at a time and passed through the precooled homogenising system twice, (using the single-shot mode) to ensure complete lysis. During this project the Stanstead system was replaced with a Continuous Flow CF1Cell Disrupter (Constant systems). For lysis in later protein preparations this cell disrupter was used in place of the Stanstead system as a homogenising method with the pressure also set to 23kpsi.

The lysate product from both methods (sonication and homogenisation) was cleared of cellular debris by centrifugation at 18000RPM for 45 minutes, at 4°C. The supernatant was then filtered through a 0.45µm sterile PES syringe filter (Starlab).

3.2.10.15 His purification, fusion tag cleavage and reverse His purification

The lysate supernatant was loaded at 2mL/min onto a HisTrap FF 5mLNi sepharose column (Cytiva), pre-equilibrated (with 5CVs of HTA buffer) using an ÄKTA Start (Cytiva) system. Column was then washed with 4CVs of HTA to remove unbound proteins, then elution was achieved using a 100% high imidazole buffer (HisTrap B, HTB: 20mM Na2HPO4, 500mM Imidazole, 500mM NaCl, pH 7.4) wash

for 6CVs. Fractions containing the POI were subsequently identified using the UV chromatograph and retained for SDS-PAGE analysis.

Fractions containing POI were pooled and dialysed O/N at 4°C using 3500Da cellulose membrane dialysis tubing into a high salt no imidazole 2L buffer, with stirring (500mM NaCl, 20mM Na2HPO4, pH 7.4). Depending on the intensity of the UV peak in the His elution step (indicative of protein concentration) a dilution was performed here with HTA buffer. To avoid precipitation observed at high protein concentrations during tag cleavage, UV peaks of ≥600mAU were diluted 1 in 2 and ≥1000mAU were diluted 1 in 3. During this dialysis step the SUMO fusion tag was cleaved by adding 1mL SUMO protease at 2mg/mL to the protein before adding it to the tubing.

Next a reverse His purification step was utilised, whereby the post dialysis soluble (PDS) solution, was His purified using a flow rate of 2mL/min. Following loading HTA buffer was flushed through with RFT collection (containing the target protein with no tag) continued until UV returned to baseline. At this point buffer flow was switched to HTB to elute cleaved tag and protease, which bound to the column, elution reverse (ER). Protein was collected this time in the reverse flowthrough (RFT) when the UV started to increase. Samples were retained to assess purity using SDS-PAGE analysis outlined in **methods section 2.6**. Fractions were collected and those determined to contain only POI pooled and retained for subsequent concentration and storage. Any fractions that were deemed to have POI but with contaminants present were pooled and further purified via gel filtration to maximise recovery.

3.2.10.16 Gel Filtration

Gel filtration (GF) sometimes referred to as size exclusion chromatography (SEC) was carried out on an ÄKTA Pure 25 system following on from the reverse His purification, utilising a HiLoad 26/600 Superdex® 200pg column (Cytiva), equilibrated with 1X phosphate buffered saline (PBS) buffer, pH 7.4. RFT from the reverse His column, concentrated to ≤5mL, was loaded using a 5mL loop, 1mL/min flow rate, followed by isocratic elution with, flow rate 2.6mL/min. 2.5 mL fractions were collected. UV was used as a means of identifying fractions containing protein. Fractions from GF peak were assessed for purity using SDS-PAGE, pure homologous fractions containing POI sized bands were pooled, concentrated, and retained for further planned characterisation experiments.

3.2.10.17 Additional optimisation

An initial move repeating small scale conditions in large scale (1L) was not successful with M1 and recovery of POI was very low. **Figure 26** gives an overview of how the replication of small to large scale expression and purification was initially trialled.

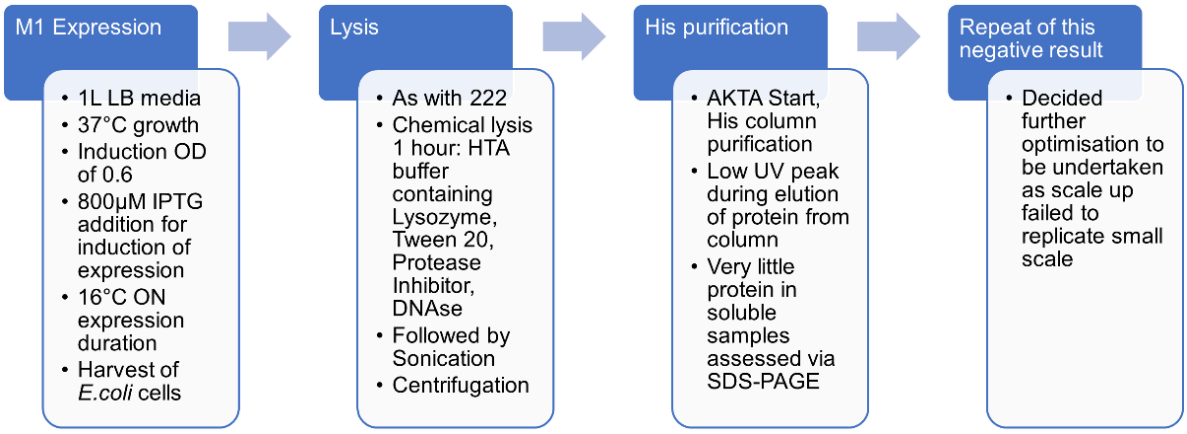


Figure 26: Overview of how small to large scale of expression conditions for M1 initially proceeded. The only change to expression conditions was culture volume from 10mL to 1L of LB (described in **methods section** 3.2.10.12). Lysis proceeded as described in **method section** 3.2.10.14. His purification proceeded to the end of the initial his column load and elution (described in **methods section** 3.2.10.15) where SDS-PAGE analysis (described in **methods section 2.6**) of the eluted fractions was made. It was at this time lack of soluble POI was identified. This was trialled a second time when further optimisation was chosen as the next best step for M1.

Following initial scale up, a range of subsequent trials were planned, to try and further optimise soluble expression and improve recovery of POI. For all these trials 100mL of media in a 250mL baffled Erlenmeyer flask was utilised (same ratio as used in **methods section** 3.2.10.12). The range of different additives, and alternative media options trialled are summarised in **Table 16**. All additives were added to LB at the point of starter culture inoculation of large-scale culture flasks. All additive stock solutions were filter sterilised using a 0.22µm sterile PES syringe filter (Starlab). All conditions were trialled once as an indicative first experiment, where an additive was identified as having a better soluble fraction a higher concentration of that additive was subsequently trialled as well as a repeat of the original

concentration. Expression temperature was also reduced to 16°C in these trials following the theory that lower temperature means slower protein folding [361]. An OD of 0.9 was always aimed for more strictly to maximise recovery from each preparation.

Table 16: Media optimisation trial conditions M1. A range of additives in LB media and alternate media options were trialled. Concentration/ composition is listed, rationale for use and reference.

Trial Condition (additive or media)	Concentration / composition	Rationale	Reference
Glycerol	1%, 5%	Chemical chaperones can influence protein stability and folding. An increase in osmotic pressure caused by these additives leads to the accumulation of osmoprotectants in the cells, which stabilise the native protein structure enhancing stability.	[362-364]
Sucrose	0.1M, 10mM, 20mM		
Sorbitol	0.5M		
Ethanol	3%, 5%	Ethanol is amphipathic and can change the membrane fluidity of the cells, membrane transport, membrane lipid composition and membrane assembly. These changes can affect activities such as DNA replication and subsequently DNA synthesis. An enhancement in DNA synthesis results in gene amplification that may enhance synthesis of inducible proteins.	[365, 366]
Glucose	1%, 3%, 5%	Catabolic repression of the lac operon.	[367, 368]
Terrific broth (TB)	2.795g TB powder (Melford) in 50mL MQ	Richer media than LB, includes glycerol as an energy source, leading to faster growth and higher yield of bacteria. TB also contains potassium phosphates to buffer the media and lower the chance of cell death due to a drop in pH during growth.	[369-371]
Minimal media	M9 minimal medium (containing only essential salts) was produced using the Cold spring harbor protocols standard recipe.	Less nutrient environment of this media reduces growth rate and expression activity, as not to overload folding machinery of cell so may result in more correctly folded soluble protein.	[372]

Once soluble expression of M1 was identified and maintained as improved in large-scale culture volume, optimised large-scale culture conditions were employed

for all subsequent mutant proteins (as outlined in **Table 15**). Adding the 10mM sucrose, aiming for a higher induction OD of 0.9 and lowering expression temp to 16°C were adopted as standard culture practice here. Homogenisation was also adopted as the preferred method of lysis for all the mutants in this chapter, more suited to larger volumes to try and maximise efficiency.

3.2.11 Characterisation of mutants

Characterisation of 222 and all mutants within this chapter included liquid chromatography mass spectrometry (LC-MS), circular dichroism (CD), Nano differential scanning fluorimetry (DSF) and assessment of disulfide bonds both *in silico* and using SDS-PAGE. The LC-MS and NanoDSF experiments were outsourced to external facilities.

3.2.11.1 Mass spectrometry

All proteins were provided to the centre for proteome research (CPR), University of Liverpool in 1XPBS buffer. In-solution trypsin digestion was first undertaken, followed by LC-MS analysis using an Ultimate 3000 RSLC™ nano-system (Thermo Scientific) coupled to a Q Exactive HF Quadrupole-Orbitrap™ mass spectrometer (Thermo Scientific).

3.2.11.2 Circular dichroism

CD is a type of light absorption spectroscopy which is an excellent tool for rapidly determining secondary structure of molecules including proteins. Providing a quick empirical determination of whether a protein is primarily α-helix, β-sheet, or unfolded which is how it was utilised in this work. Greenfield provides a very good

reference protocol, including explanation of the technique [373]. Differential absorption of left and right circularly polarised light (asymmetric absorption) can only be seen with asymmetric/chiral molecules such as proteins [373, 374]. CD signal is observed only at wavelengths where the sample absorbs radiation, i.e. under absorption bands, and the signal may be positive or negative depending on the handedness of the molecules in the sample and the transition being studied [375].

CD was carried out here on a J1100 spectropolarimeter (JASCO, UK). Far-UV CD spectra (250–180 nm) were obtained using a 0.2mm cuvette, at 4°C. 0.5mg/mL protein samples in 2.5mM HEPES buffer, pH 6.5 were used for all proteins, all readings were made in triplicate per sample. Secondary structure content values were acquired by submitting raw spectral data to the Beta Structure Selection (BeStSel) webserver available at https://bestsel.elte.hu/index.php (accessed 02/20) [376]. Molar residue ellipticity (MRE) was calculated and plotted to normalize the spectra using protein concentration using **Equation [3]**.

$$\boldsymbol{MRE} = \frac{\boldsymbol{CD}}{[\boldsymbol{peptide}]\times\boldsymbol{10}\times\boldsymbol{l}\times\boldsymbol{n}} \qquad \textbf{[3]}$$

Where *CD* is circular dichroism measure in millidegrees, *l* is the path length in cm and *n* is the number of amino acid residues. The peptide concentration is in M. The units of molar ellipticity are deg.cm^2/dmol.

3.2.11.3 Stability

3.2.11.3.1 *In silico* stability prediction

In silico prediction of the stability of all the proteins in this chapter was undertaken using the SCooP webserver available at http://babylone.ulb.ac.be/SCooP (accessed, 02/2019) [377]. Which provided a full

temperature dependent stability curve prediction. These are comparable to the data that is generated in the next section using the *in vitro* technique of NanoDSF. Giving a good point of comparison between *in silico* and *in vitro* stability measure as both determine the same comparable measure, of melting temperature (Tm).

3.2.11.3.2 NanoDSF

Protein samples were sent to the Sample Preparation and Characterisation Facility, EMBL Hamburg for analysis (Prometheus NT.48 nanoDSF, Nanotemper) or to a colleague (Dr Emily Wang) who had access to the same equipment in a collaborator's facility.

During NanoDSF thermal unfolding experiments a linear temperature ramp from 20-95°C was applied to unfold the proteins. All samples were run in triplicate, simultaneously in capillaries (10µL sample volume). Intrinsic tryptophan fluorescence (ITF) was monitored in real-time at 330nm and 350nm. Results were then processed in Microsoft Excel to give a ratio; the Tm could then be determined from a plot of this ratio against temperature. Tm is defined as the temperature at which 50% of the protein is unfolded. It can also be described as the midpoint of the transition of a protein from folded to unfolded, inflection point (IP350/330) of the transition was derived from the maximum of the first derivative of each measurement determined with GraphPad using a non-linear regression fit (four parameters).

3.2.11.4 Disulfide assessment

3.2.11.5 *In silico* disulfide assessment

Two different tools were used to assess disulfide bonds, both tools required submission of a PDB file. The first tool used was created by the Liu Lab, named PredDisulfideBond (available at: http://liulab.csrc.ac.cn:10003/index, accessed 02/2022). The second tool used was MAESTROweb, with an option to evaluate potential disulfide bonds by considering both ΔΔG and geometric constraints [348] (available at: https://www.services.came.sbg.ac.at/maestro/web, accessed 02/2022).

Tertiary structure is the three-dimensional shape of the protein determined by regions stabilized by interactions between amino acid residue side chains [378]. Structural *in silico* tools consider the conformational context of a protein, including stabilising tertiary contacts and interactions. A proteins fold has implications in determining which residues are exposed or buried, and their local chemical environment i.e., hydrophobic, and electrostatic interactions. Both tools employ structure-based assessments which means the fold and resultant residue proximity could be factored into the predictions.

3.2.11.6 *In vitro* disulfide assessment

In vitro disulfide bond assessment for all novel proteins in this work were made by comparing reducing, R (using 1.4M β-Mercaptoethanol in the loading buffer) and non-reducing, NR (no β-Mercaptoethanol in loading buffer) samples on SDS-PAGE, as described in **methods section 2.6**.

3.2.11.7 Binding Assay

Binding assay methodology as previously described in publications by Xu et al 2009 and Steffensen et al 1995 [303, 379] was used here to test the binding of all proteins to TII gelatin.

3.2.11.7.1 Plate preparation

1mg of TII collagen (Bioiberica) was reconstituted in 1mL of carbonate bicarbonate buffer (20mL of 100mM Na_2CO_3 added to 100mL of 100mM $NaHCO_3$, final buffer pH adjusted to 9.2 filtered through a 0.22µm filter). This 1mg/mL TII collagen solution was heat-denatured at 80°C for 20 minutes, with vortexing every 5 minutes. The resultant TII gelatin solution (heat denatured, collagen product) was diluted to 0.01mg/mL in 0.1M carbonate buffer, pH 9.2 . Then 50µL of the 0.01mg/mL TII gelatin solution was then added to each well of a 96 well plate (Immulon 2HB, high binding plates, ThermoScientific), except the outer wells, giving a coating of 0.5µg/well. Plates were wrapped in cling film and left at 4°C for three days. At this point plates were washed three times with PBS-Tween (0.001%), blocked with 100µL of PBS-2.5% Bovine serum albumin (Sigma) for 30 minutes at RT, washed one more time with PBS-Tween and dried in an incubator at 50 °C for 15 minutes. Plates were kept at 4°C and retained for use within 4 weeks.

3.2.11.7.2 Protein biotinylation

To begin protein aliquots were thawed on ice, dialysed overnight in 1L of 0.1M $NaHCO_3$, pH 7.4, at 4°C to buffer exchange, as per **materials and methods section 2.8**. Biotin (EZ-Link™ Sulfo-NHS-LC-Biotin, Thermo Fisher) was then added with a

20-fold molar excess, as per the manufacturer's instructions and incubated for 20 minutes at 22°C, then for 2h at 4°C for biotinylation to occur. Proteins were then dialysed in 1L of PBS, pH 7.4 overnight at 4 °C.

3.2.11.7.3 Binding assay

Biotinylated proteins were added to the gelatin coated plates with a starting concentration shown in **Table 17**, then serially diluted with 1X PBS, pH 7.4 across the plate (**Figure 27**) to give a range of concentrations, covered with cling film and incubated for 1h at 22°C. Each protein was loaded in triplicate across a 96 well plate and each plate was run in triplicate, except M2 which due to limited protein was ran in duplicate on each plate, across three plates.

Table 17: Binding assay proteins concentrations and dilutions. All proteins assessed using the plate binding assay in this chapter, starting concentration and serial dilutions used across the plates.

Protein	Concentration (µM)	Serial Dilution
222	28	1 in 7
M1	120	1 in 4
M2	40	1 in 6
M5	60	1 in 6
M8	21	1 in 4

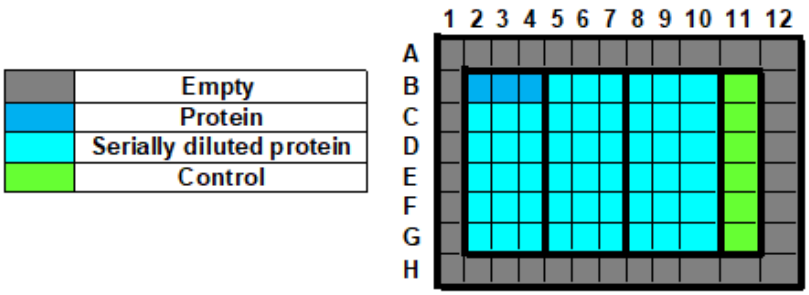


Figure 27: Binding assay 96 well plate layout. Protein wells contained protein at the concentrations shown in **Table 17**, a volume calculated to give the dilution also shown in **Table 17** was then transferred across the plate. Control wells contained 1X PBS buffer, pH 7.4 only.

Plates were washed three times with PBS-Tween 200µL, then 100µL PBS-2.5% BSA incubated to block for 30 minutes at RT. Plates were washed again three times with 200µL PBS-Tween, before alkaline phosphatase conjugated to streptavidin (diluted 1:10 000 in PBS, Pierce) was added and incubated for 30 minutes at 22°C. Plates were washed again three times with 200µL PBS-Tween and 100µL of P-Nitrophenyl Phosphate (PNPP) solution added. Plates were incubated for 8 minutes at RT before being read in a Flexstation 3 microplate reader (Molecular Devices) with a 405nm absorbance.

The average absorbance measures of control wells (1XPBS buffer only) were subtracted from the sample wells and a binding curve was generated for each plate, using a non-linear regression fit (four parameters) in GraphPad Prism. The absorbance values for each plate were averaged, then the three averages for the three plates used to generate the final binding curves, using a non-linear regression fit (four parameters) in GraphPad Prism.

3.3 Results

3.3.1 222 model generation

Two possible template protein structures were identified via the BLASTp search: 1EAK and 1CK7. Both have known crystal structures in the PDB and were equal in percentage query cover and identity (see **Table 18**). 1EAK was however determined to be the stronger match as it had the lowest E value. Therefore, 1EAK was taken forward to act as a template for modelling 222.

Table 18:Two strongest identified BLASTp results for 222. Match description, Query Cover, E value, % Identity, Accession number and reference.

Description	Query Cover	E value	% Identity	Accession	Reference
Catalytic domain of proMMP-2 E404Q mutant [Homo sapiens]	100%	7e-94	75.14%	1EAK	
Chain A, PROTEIN (GELATINASE A) [Homo sapiens]	100%	2e-92	75.14%	1CK7	[380]

Figure 28 shows the alignments of 1EAK to 222 sequence, the first match (A) was the better alignment as indicated with the lower E value, greater % identities so it was this alignment used to construct the .ali file required to build models of 222. The five models generated are shown in **Figure 29**. All models are similar in that they have similar beta strand sections in each modules secondary structure composition and there is little conformational variation between the five different models as shown in the superimposed image and showed in the *in silico* secondary structure percentage comparison (**Table 19**).

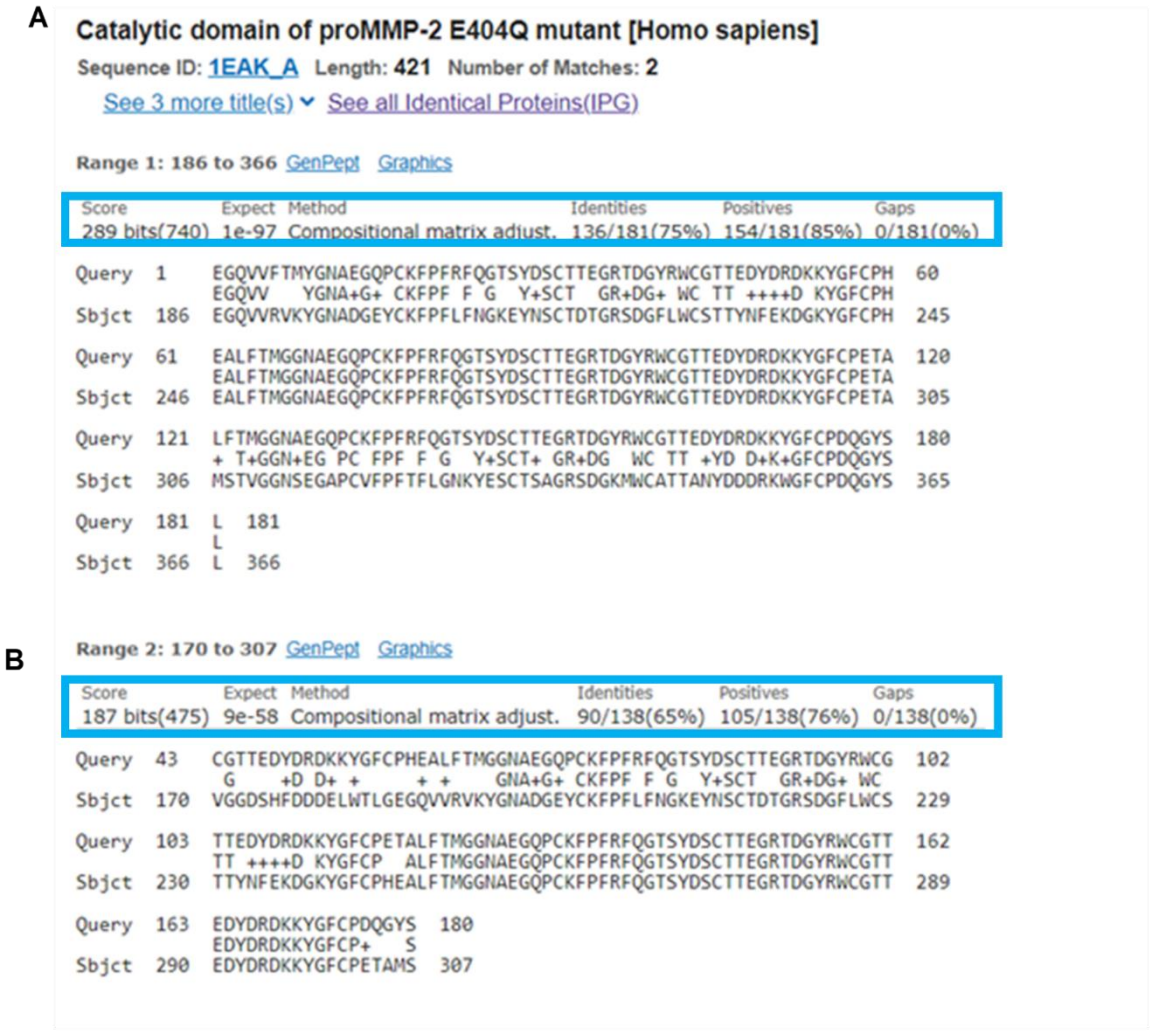


Figure 28: Two possible alignments of 222 to 1EAK template. A, first BLAST alignment match **B,** second BLAST alignment match. Highlighted in blue are alignment score, E value, method of alignment, % identities, % positives and gaps. Based on all these comparisons **A** is the better alignment, so was taken forward to be used as the template for model generation.

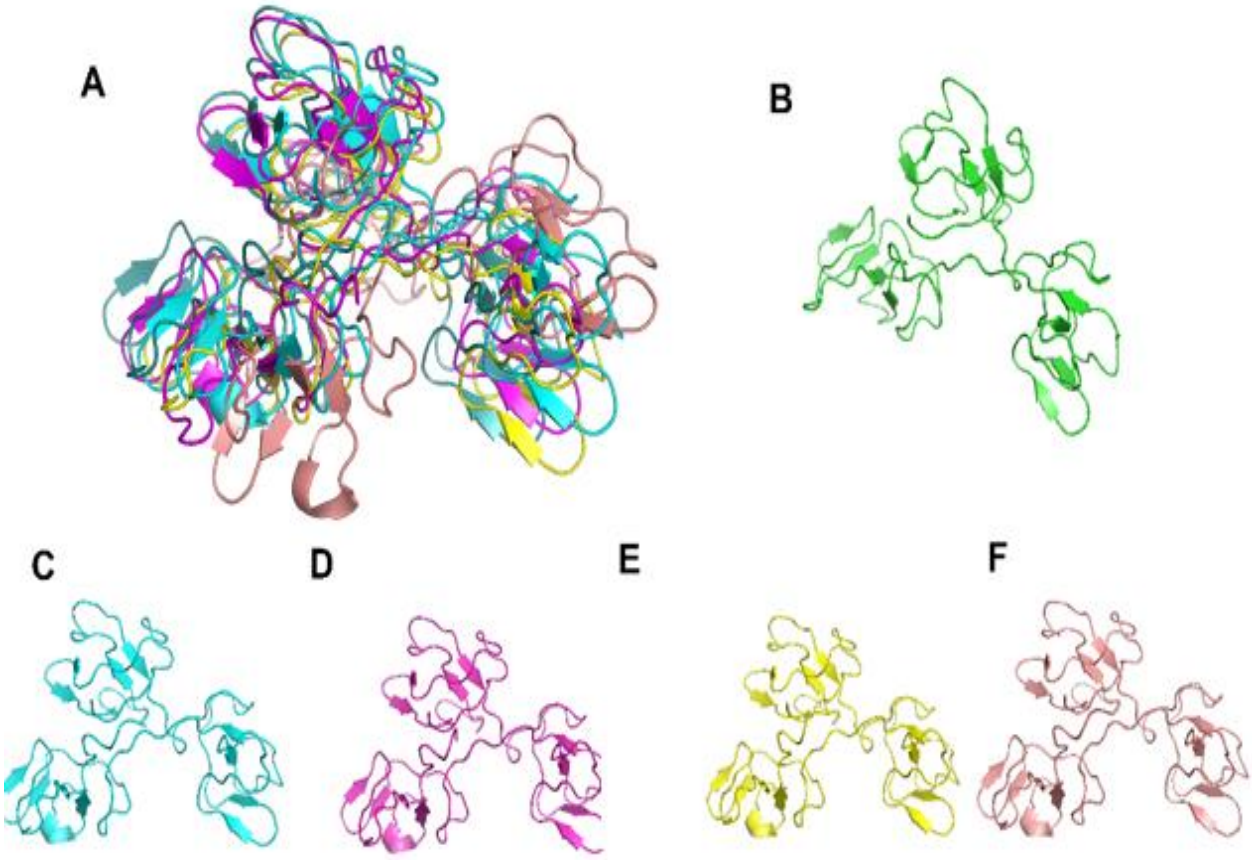


Figure 29: Five models of 222. Shown in pymol cartoon format. **A,** all five models transposed to show the conformational differences. **B,** Model 1. **C,** Model 2. **D,** Model 3. **E,** Model 4. **F,** Model 5.

Table 19: Percentage of each secondary structure type in the five models of 222. This was determined using the 2StrucCompare web server [344]. There is very little difference between the five models in terms of secondary structure.

Secondary Structure	1	2	3	4	5
Helix	0%	2%	2%	2%	2%
Sheet	20%	22%	22%	19%	18%
Other	80%	76%	76%	79%	80%
Undefined	0%	0%	0%	0%	0%

3.3.2 Model selection

QMEAN was initially used to identify the best model. The higher the "QMEAN" score the better the model [345], model 5 had the highest score of 0.29 indicating that it was the best of the five models (**Table 20**).

Table 20: QMEAN Score assessment of the five generated 222 models. Model 5 has the greatest score so was selected as the model to be used in further in silico work within this thesis.

Model no.	Q Mean score	Rank
1	-0.82	5
2	-0.02	3
3	-0.32	4
4	0.22	2
5	**0.29**	**1**

Ramachandran plot was also used as a secondary assessment to confirm model quality. The Ramachandran plot (**Figure 30**) further confirms model 5 is a conformationally sound model. This conclusion can be made as the amino acid conformation of model 5 places 95% of amino acids in favoured and 5% in allowed positioning. Importantly it also shows no outliers or disallowed residues indicative of model mistakes.

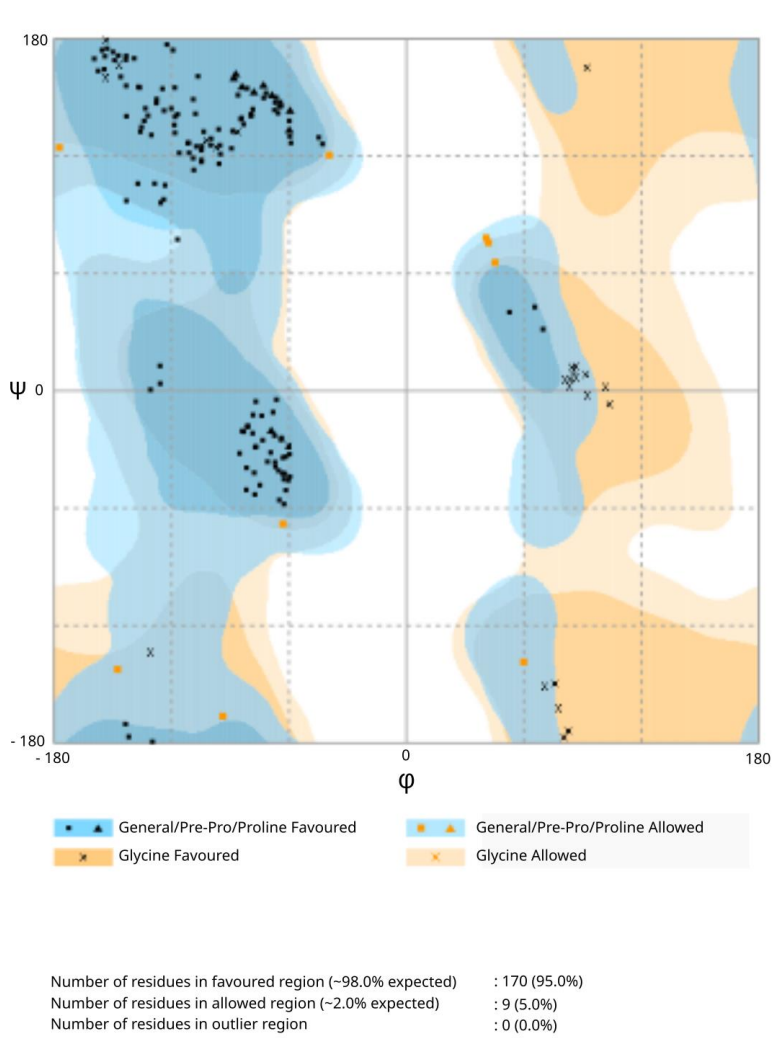


Figure 30: Ramachandran plot of Model 5. A Ramachandran plot shows the statistical distribution of the combinations of the backbone dihedral angles (ϕ and ψ). In theory, the allowed regions of the Ramachandran plot show which values of the Phi/Psi angles are possible for an amino acid, showing if models are in the allowed conformational space [381]. Providing a check of the validity and integrity of the 3d structural model. [382] The plot here confirms model 5 gives a conformation that is 95% favoured and 5% allowed. It shows no outliers or disallowed residues. This plot was generated using the rampage tool [383].

3.3.3 Alanine mutagenesis

From model 5 the ten mutants outlined in **Table 21** were computationally generated using the pymol mutagenesis wizard to mutate binding residues for alanine, shown in **Figure 31**. Looking at the mutants in pymol there is little

conformational difference most clearly shown with the superimposed structures

(**Figure 31A**).

Table 21: Mutant residue changes and numbers. Those highlighted in red are those not selected following subsequent stability (**Figure 32**) then solvent exposure assessment (**Figure 33**).

Mutant	Abbreviation	Residue mutated to A	Residue numbers 222
1	M1	N, Asn, Asparagine	11, 69, 127
2	M2	R, Arg, Arginine	22, 80, 138
3	M3	F, Phe, Phenylalanine	23, 81, 139
4	M4	Y, Tyr, Tyrosine	28, 86,144
5	M5	Y, Tyr, Tyrosine	40, 98, 156
6	M6	W, Trp, Tryptophan	42, 100, 158
7	M7	T, Thr, Threonine	46, 104, 162
8	M8	E, Glu, Glutamic acid	47, 105, 163
9	M9	Y, Tyr, Tyrosine	49, 107, 165
10	M10	Y, Tyr, Tyrosine	55, 113, 171

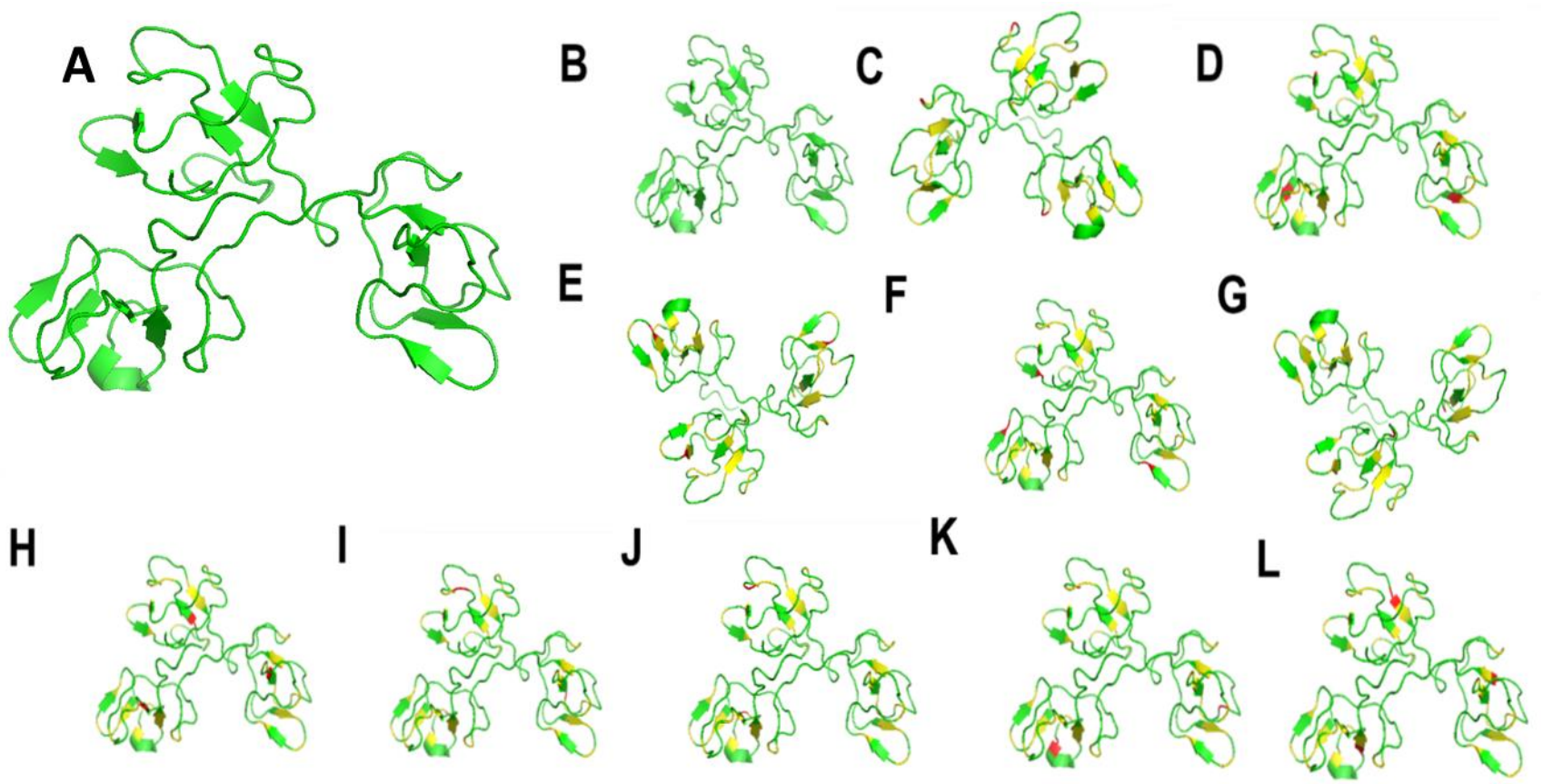


Figure 31: 222 and the 10 binding residue alanine mutants outlined in Table 21, shown in cartoon format. All mutants were generated from model 5 of 222 structure with the pymol mutagenesis wizard. **A** 222 and the 10 mutants transposed to show there was minimal change to the conformation caused by the alanine substitutions. **B** shows 222 and **C-L** show mutants (abbreviated to M) 1–10. the native residues are shown in green, binding residues are shown in yellow, mutated binding residues in the three module 2 domains of 222 are shown in red.

3.3.4 *In silico* percentage secondary structure assessment

The in silico model quantification of secondary structure type made with the 2StrucCompare sever (**Table 22**),shows no difference between any of the mutant proteins generated *in silico* in this chapter.

Table 22: Percentage of each secondary structure type in 222 and the ten alanine mutants. This was determined using the 2StrucCompare web server [344]. There is very little difference between the ten mutants in terms of secondary structure.

Secondary Structure	222	1	2	3	4	5	6	7	8	9	10
Helix	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%
Sheet	18%	18%	18%	19%	18%	18%	18%	18%	18%	18%	18%
Other	80%	80%	80%	79%	80%	80%	80%	80%	80%	80%	80%
Undefined	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%

3.3.5 Alanine mutant selection

A series of approaches were utilised to narrow down the number of mutants to progress to *in vitro* testing. First mutants were selected by using the PoPMuSiC and maestro tools to assess stability (**methods section 3.2.8**). Five out of the ten mutants were identified as the most stable by both methods of evaluation (**Figure 32**). ΔΔ*G* is defined as the difference between the free energy shift caused by mutation in the native state [384]. Higher ΔΔG values are indicative of destabilising mutations, so the lowest value mutants were most stable. Based on these results five mutants 1, 2, 5, 7 and 8 were taken forward to the next evaluation of solvent exposure.

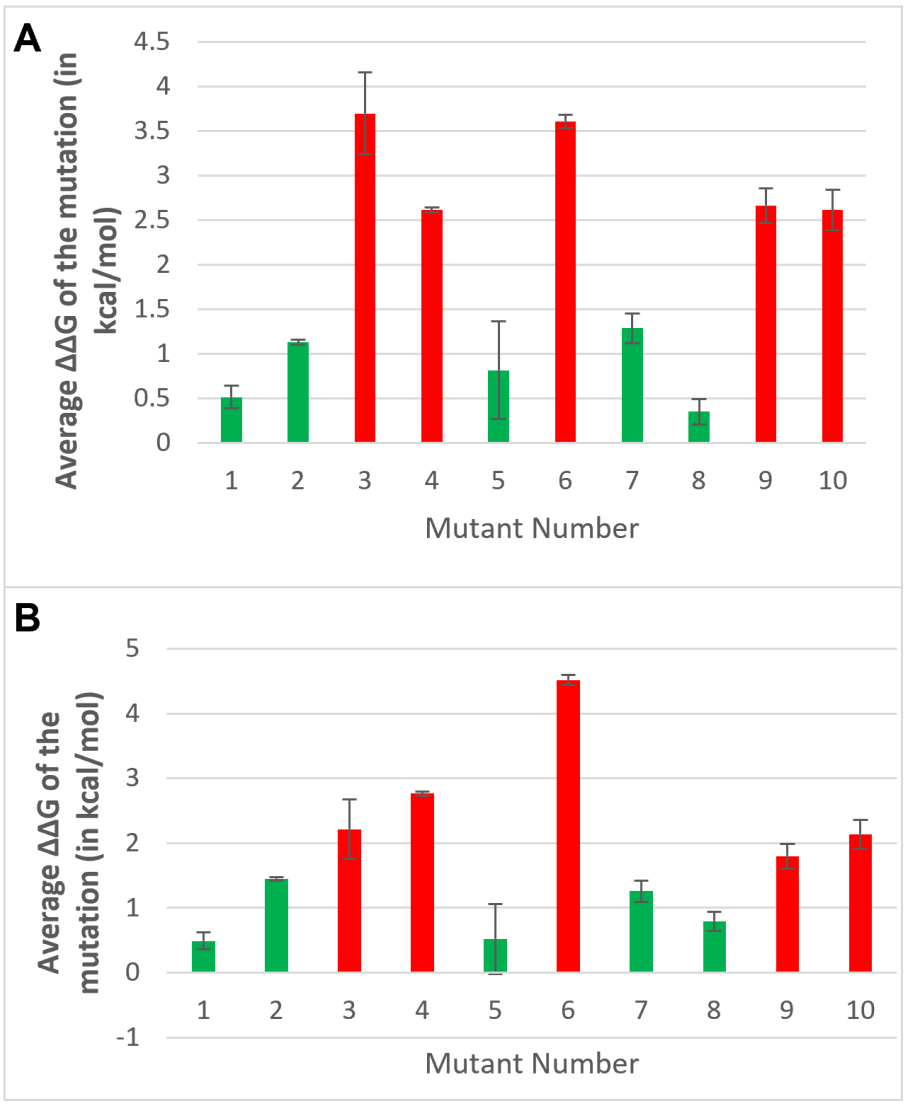


Figure 32: Stability assessment of ten binding residue alanine mutants. Two different tools were used **A**, PoPMuSiC results and **B**, Maestro results. The five mutants represented in green are the most stable of the 10 mutants as indicated by the calculated average ΔΔG.

The two evaluations of solvent exposure employed on mutants 1, 2, 5, 7 and 8 (**Figure 33**) agreed that mutants 1, 5 and 8 were the strongest candidates to take to testing based on solvent exposure. Mutant 2 however produced conflicting results in that the results of the two tools did not agree on the solvent exposure ranking, therefore here a decision was made given the mutant had positive stability assessment results to still proceed to test with this mutant. A higher QSASA

indicated the residue was more exposed which is what we would expect for a residue critical to binding. PoPMuSiC is the same online tool used previously to assess mutant stability but also capable of calculating solvent accessibility for mutant residues. Again, a higher exposure was preferable for a binding site mutation.

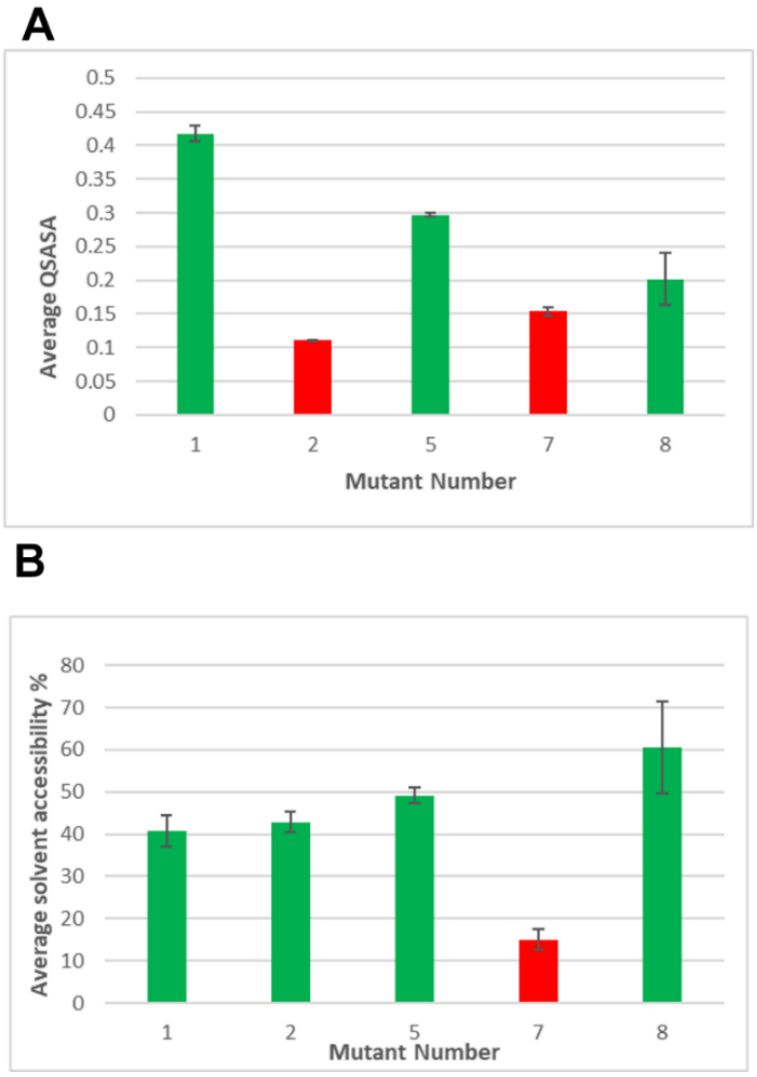


Figure 33:Solvent exposure assessments of mutants. A POPS results, the three mutants represented in green are the ones with the best solvent exposure as indicated by the calculated average QASA score. **B**, PoPMuSiC results, the four mutants represented in green are the ones with the best solvent exposure as indicated by the calculated average percentage solvent accessibility.

Table 23 shows the protein sequence for 222 and the four mutants that were selected and taken forward to *in vitro* testing.

Table 23: Protein sequences for this chapter (chapter 3), with key physical and chemical parameters. Protein names, abbreviations, alanine mutation sites stating the residue in 222 the 'native' in this chapter and the numbered position of each mutation site, as well as the protein sequences. The residues shown in blue text were left as in the native CBD protein, as they were deemed had critical involvement in intramolecular interactions. The three modules for each protein are shown in red text, and the linker regions in black text. Alanine (Ala) substitution sites for each mutant are highlighted in yellow. Abs 0.1%, extinction coefficient and molecular weight listed, determined using the Expasy ProtParam tool available at https://web.expasy.org/protparam/.

Name	Mutation site	Protein sequence	Extinction coefficient	Abs 0.1% (=1 g/l)	Molecular weight (kDa)
222 (native)	N/A	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	38110	1.845	20.65
Mutant 1 (M1)	N 11,69,127	EGQVVFTMYGAAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGAAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGAAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	38110	1.857	20.52
Mutant 2 (M2)	R 22, 80, 138	EGQVVFTMYGNAEGQPCKFPFAFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFAFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFAFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	38110	1.869	20.40
Mutant 5 (M5)	Y 40, 98, 156	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGARWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG ARWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGARWCGTTEDYDRDKK YGFCPDQGYSL	33640	1.651	20.37
Mutant 8 (M8)	E 47, 105, 163	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTAD YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTADYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTADYDRDKK YGFCPDQGYSL	38110	1.861	20.48

3.3.6 Small scale mutant expression trials

SDS-PAGE gels in **Figure 34**, **Figure 35, Figure 36** and **Figure 37** show the results of the initial small scale optimisation trials undertaken for all mutants in this chapter.

For M1 (**Figure 34**), 18°C O/N expression with an 800µM IPTG induction concentration gave the best soluble expression of approximately 50% soluble and 50% insoluble.

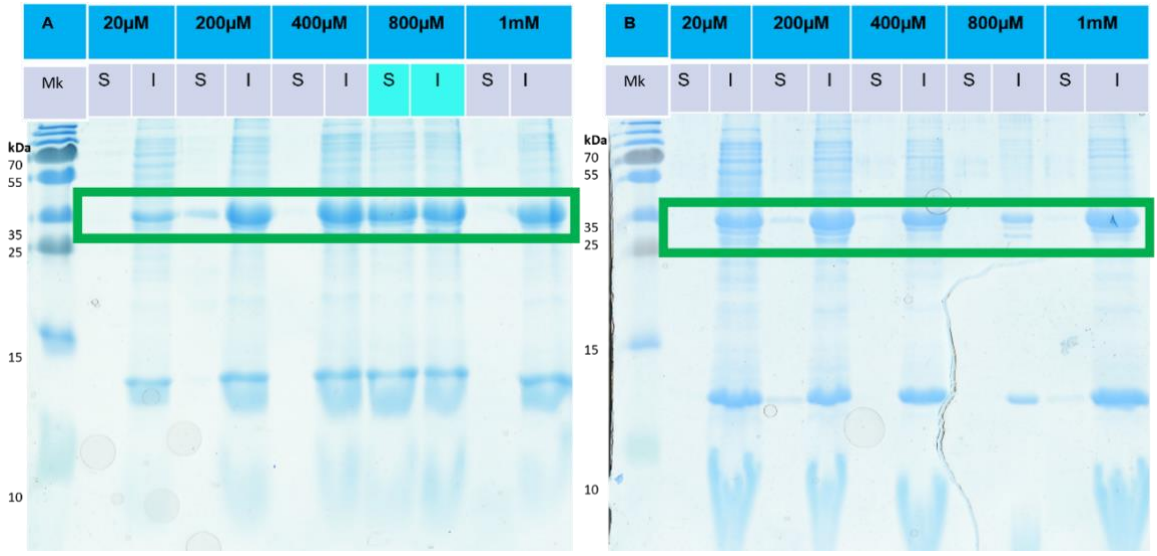


Figure 34: **Small scale expression trial results for mutant 1.** SDS-PAGE gel (n=1) showing Pageruler Marker (Mk), Soluble (S) and Insoluble (I) mutant 1 protein product which was His-SUMO tagged M1 protein with molecular weight 33.5kDa as indicated by green box. The SUMO fusion tag was utilised to try and facilitate enhanced expression and solubility aiding purification and recovery [385]. **(A)** 18°C incubation O/N or **(B)** 25°C for 5h, after IPTG induction at 20µM, 200µM, 400µM, 800µM and 1M. 800µM IPTG and O/N incubation at 18°C gave the best soluble: insoluble expression result indicated in light blue.

Figure 35, shows the results of M2 small scale expression trials. Several conditions gave comparable data here. 18°C O/N expression with an 400µM IPTG induction concentration gave a good soluble expression so was selected. With M2

200µM IPTG at 18°C and 20µM, 200µM and 400µM IPTG at 25°C there were also comparable good band ratios of soluble: insoluble expression.

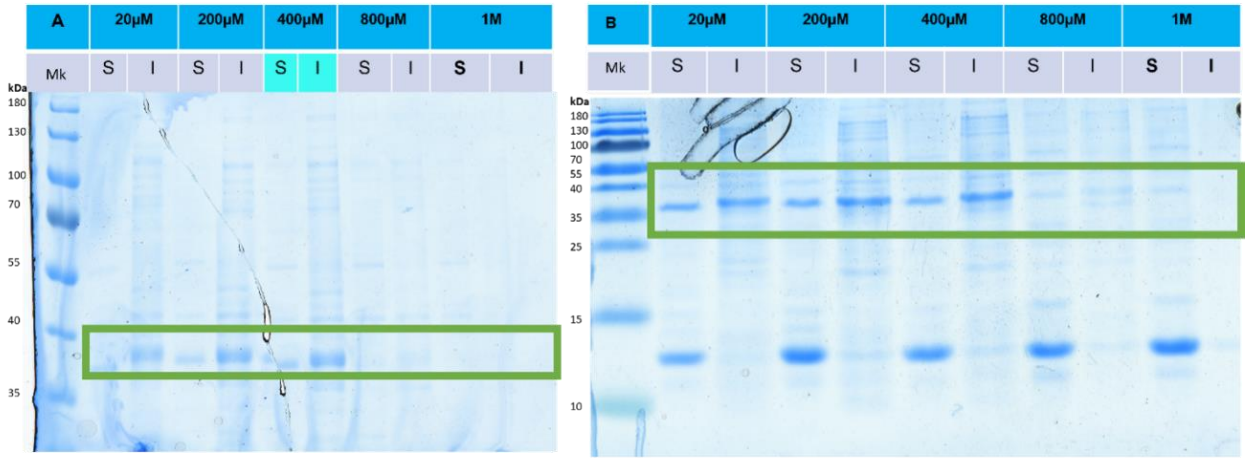


Figure 35: Small scale expression trial results for mutant 2. SDS-PAGE gel (n=1) showing Pageruler Marker (Mk), Soluble (S) and Insoluble (I) mutant 2 protein product which was His-SUMO tagged here so 33.4kDa (A) 18°C incubation O/N or (B) 25°C for 5h, after IPTG induction at 20µM, 200µM, 400µM, 800µM and 1M. 400µM IPTG and O/N incubation at 18°C was selected as a good level of soluble: Insoluble expression.

Figure 36 shows the results of M5 small scale expression trials. There were multiple conditions that gave good band ratios of soluble: insoluble expression. For ease 18°C O/N expression with an 400µM IPTG induction concentration was selected again to take to large scale.

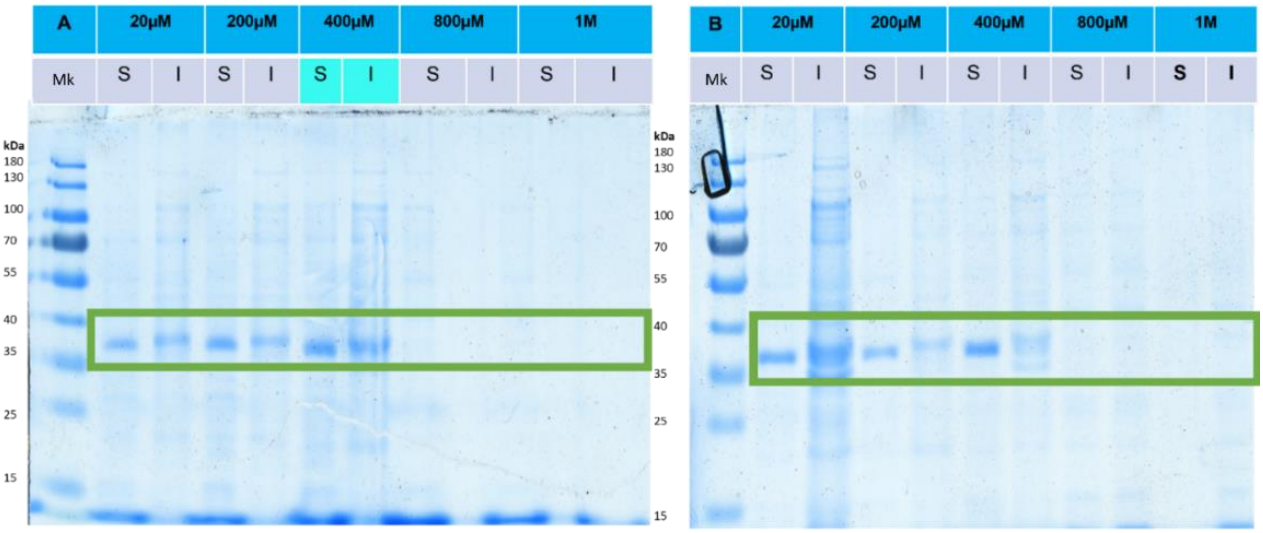


Figure 36: Small scale expression trial results for mutant 5. SDS-PAGE gel (n=1) showing Pageruler Marker (M), Soluble (S) and Insoluble (I) mutant 5 protein product which was His-SUMO tagged here so 33.4kDa **(A)** 18°C incubation O/N or **(B)** 25°C for 5h, after IPTG induction at 20µM, 200µM, 400µM, 800µM and 1M. 400µM IPTG and O/N incubation at 18°C gave the best soluble: Insoluble expression result.

Finally, **Figure 37** shows the results of M8 small scale expression trials. M8 had the best expression seen under the most different conditions indicated by the good band ratios of soluble: insoluble expression. Again, for ease 18°C O/N expression with an 400µM IPTG induction concentration was selected to take to large scale.

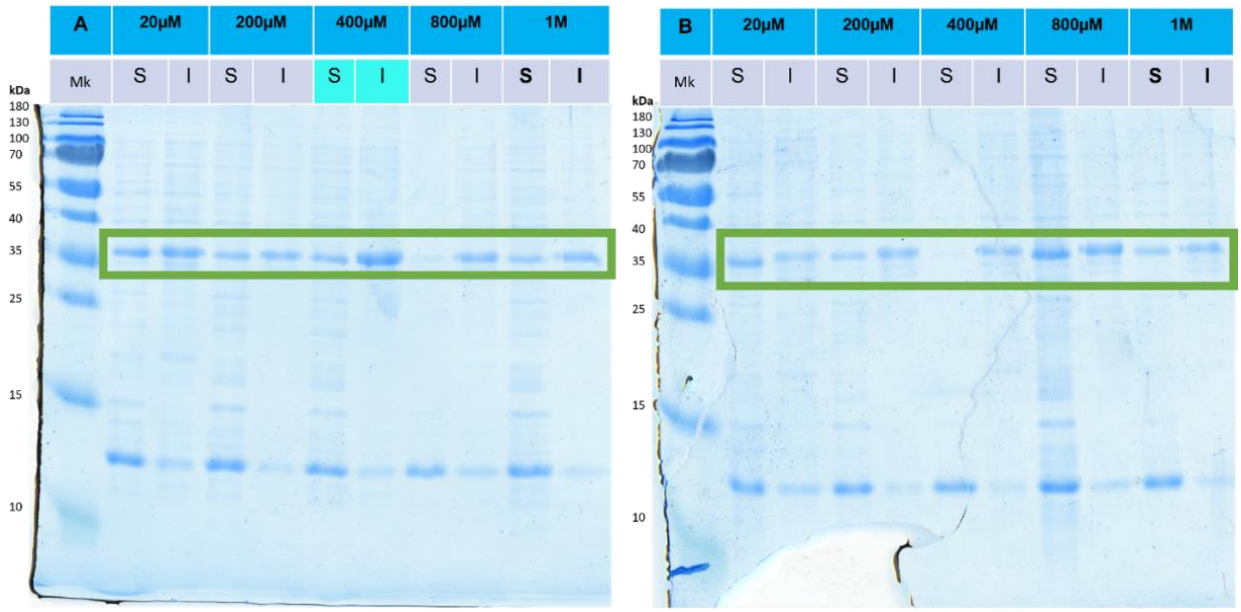


Figure 37: Small scale expression trial results for mutant 8. SDS-PAGE gel (n=1) showing Pageruler Marker (M), Soluble (S) and Insoluble (I) mutant 8 protein product which was His-SUMO tagged here so 33.5kDa (A) 18°C incubation O/N or (B) 25°C for 5h, after IPTG induction at 20µM, 200µM, 400µM, 800µM and 1M. 400µM IPTG and O/N incubation at 18°C gave the best soluble: Insoluble expression result.

These conditions 800µM IPTG for M1, then 400µM IPTG for M2, M5 and M8 with 18°C identified in small scale trials were used in the initial larger scale expression and purification to produce mutants for subsequent characterisation experiments.

3.3.7 Large scale mutant expression & purification

For M1 repeating the small-scale optimised culture conditions with large scale culture did not produce the same good level of soluble expression seen in small scale trials. This scale up was tried twice just to confirm this lack of scale up was replicable with a fresh transformation from a different shuffle *E. coli* cell stock. Therefore, additional optimisation under large scale expression conditions was carried out including altering temperature and including additives in the growth

medium. Results from additional optimisation tests with M1 are shown below in **Figure 38**, **Figure 39**, **Figure 40 and Figure 41**.

Figure 38 shows the first additive and media trial results undertaken, used to guide and inform further trials. 10mM sucrose gave the best soluble: insoluble expression ratio result when compared to control.

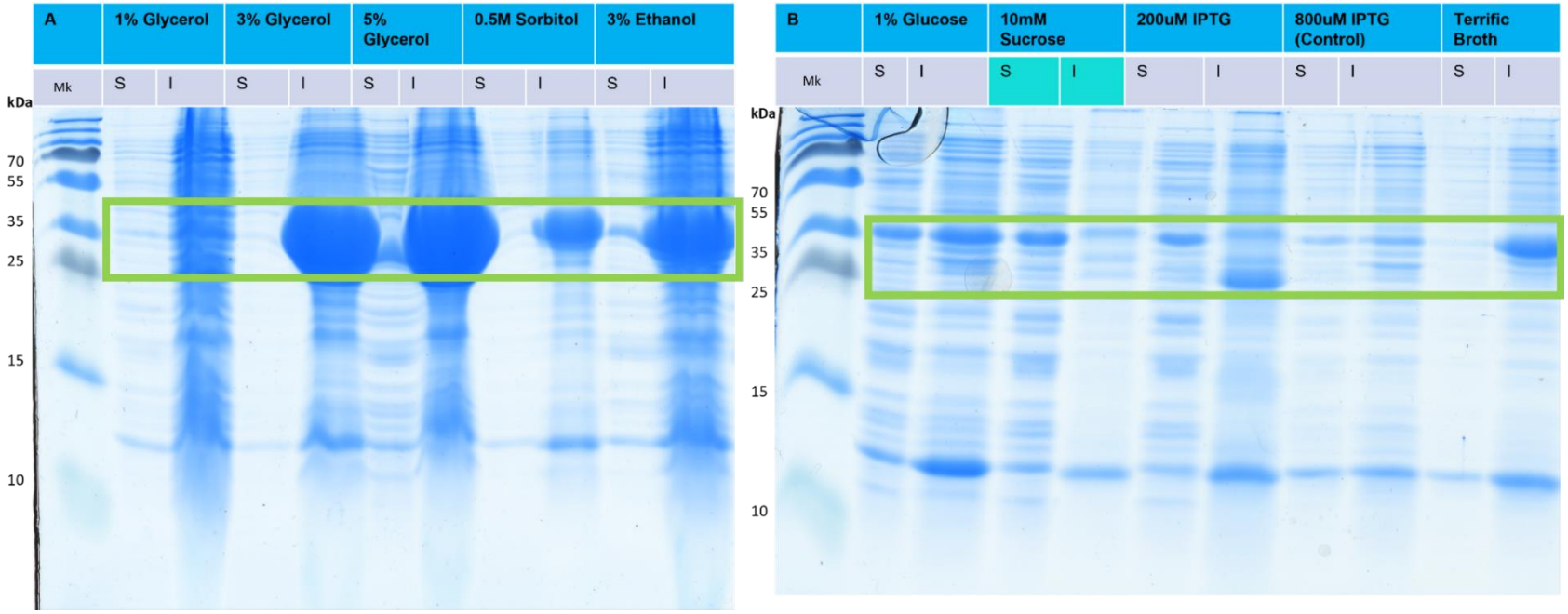


Figure 38: Mutant 1 media and additive trial results. SDS-PAGE gel showing Pageruler Marker (M), Soluble (S) and Insoluble (I) mutant 1 protein resulting product from growth and expression with several different media and additives as shown across **(A)** and **(B).** All trial cultures here had expression induced at OD 0.9, with 800µM IPTG as per small scale optimisation and O/N expression incubation at 16°C (reduced from 18°C to try and promote a more soluble expression). 10mM sucrose gave the best soluble: insoluble expression result when compared to control. Results here (n=1 for each condition) were used to guide further trials.

Following the result in **Figure 38** that 10mM sucrose improves the soluble: insoluble ratio 20mM was trialled to see if this caused greater improvement. Addition of 20mM sucrose to the growth media did appear to improve yield of soluble protein more so than 10mM sucrose, shown by the more intense band of soluble target POI size (33.5kDa) when first trialled seen in **Figure 39**.

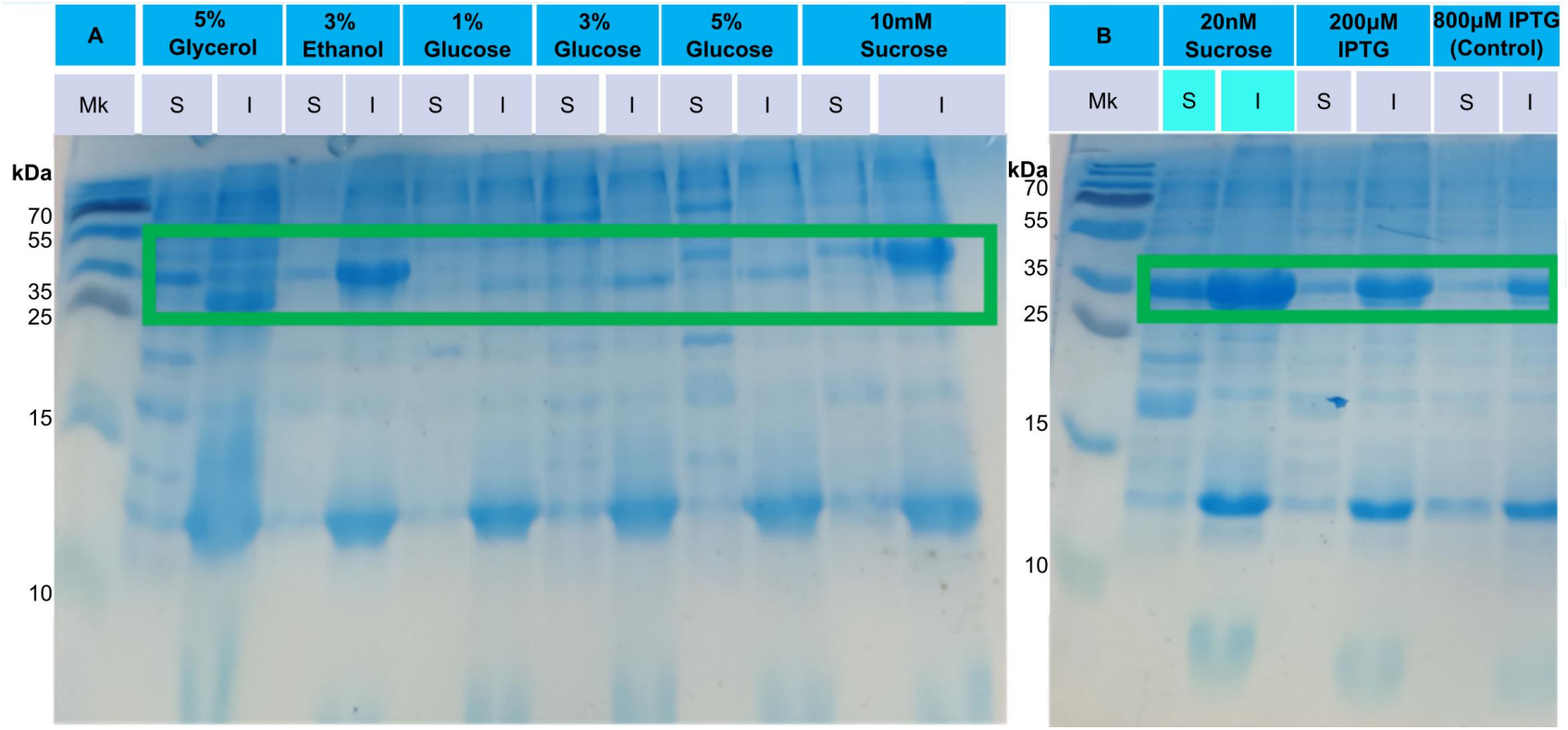


Figure 39: Mutant 1 media and additive trial results second trials and expansion. SDS-PAGE gel showing Pageruler Marker (M), Soluble (S) and Insoluble (I) mutant 1 protein resulting product from growth and expression with several different media and additives as shown across **(A)** and **(B).** All trial cultures here had expression induced at OD 0.9, with 800µM IPTG as per small scale optimisation and O/N expression incubation at 16°C (reduced from 18°C to try and promote a more soluble expression). 10mM sucrose gave the best soluble: insoluble expression result when compared to control. Results here were used as replicates and to expand up on the results of **Figure 38**. 20mM sucrose as highlighted in light blue gave the best soluble: insoluble expression result when compared to control here (n=1).

However subsequent trials revealed this improvement seen with 20mM sucrose seen in the individual sample (**Figure 39**) was not replicable, whereas the improvement with 10mM sucrose (seen in **Figure 38**) was (seen in **Figure 40**).

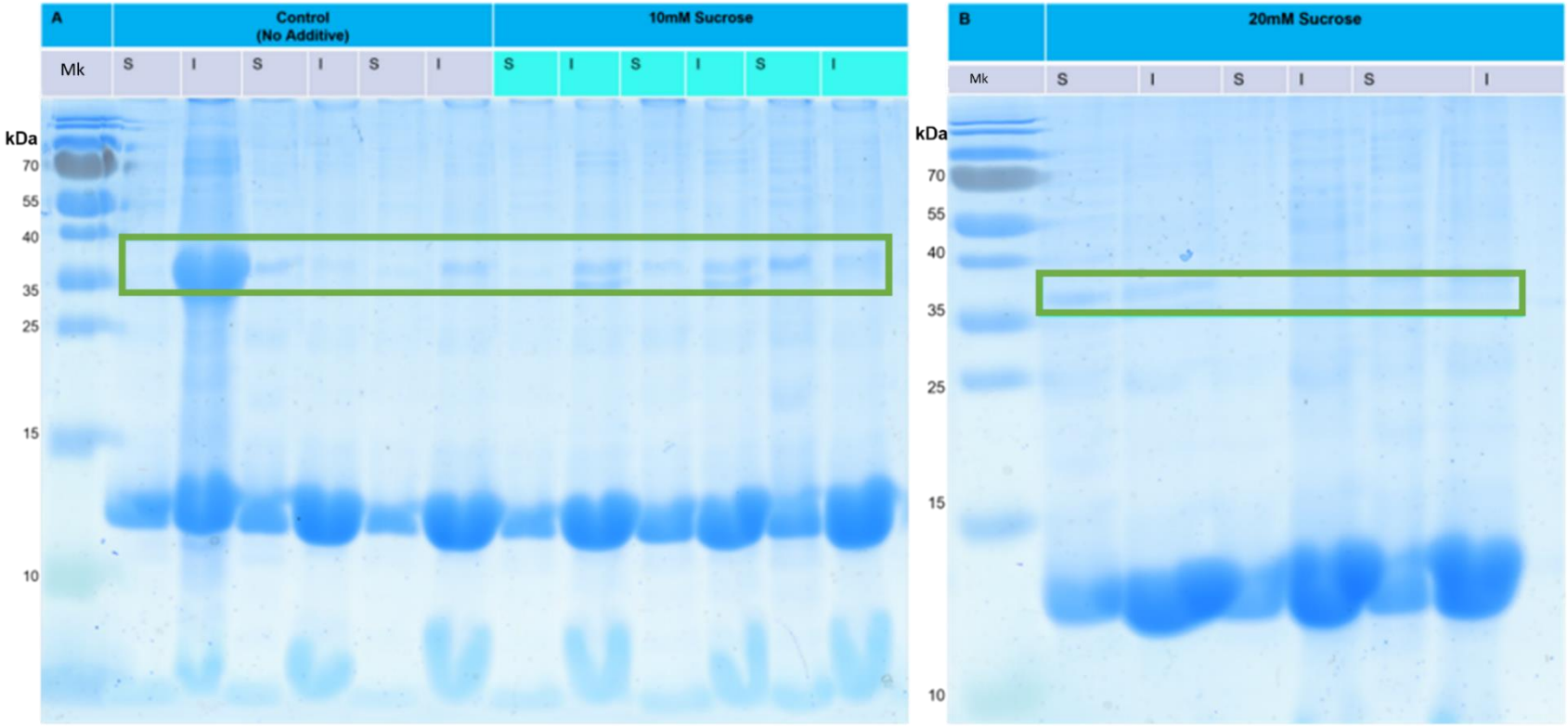


Figure 40: Mutant 1 sucrose additive repeats results. SDS-PAGE gel showing Pageruler Marker (M), Soluble (S) and Insoluble (I) mutant 1 protein resulting product from growth and expression with several different media and additives as shown across **(A)** and **(B).** All trial cultures here had expression induced at OD 0.9, with 800µM IPTG as per small scale optimisation and O/N expression incubation at 16°C (reduced from 18°C to try and promote a more soluble expression). 10mM sucrose gave the best soluble: insoluble expression result when compared to control. Results here were used as replicates and to expand up on the results of **Figure 38**. 10mM sucrose as highlighted in light blue gave the best soluble: insoluble expression result when compared to control (each condition n=3).

Both Tween 20 (a non-ionic detergent and mild lysis agent) and lysozyme (enzyme, used to break the outer membrane/peptidoglycan layer of gram-negative bacteria) were already utilised within the lysis buffer following on from the groups work with 222. All lysis variations began the same, with a one-hour chemical cell disruption period. When lysis was tried without either or both more protein was found lost in the insoluble/unrecovered (data not shown). Continuing with the exploration of lysis method, **Figure 41** shows a comparison of homogenisation and sonication in large scale culture with and without sucrose or with minimal media as an alternate. Homogenisation was implied/ indicated in the work here as the optimum secondary lysis method with M1, due to suitability of scale, a more complete lysis and better temperature control during lysis.

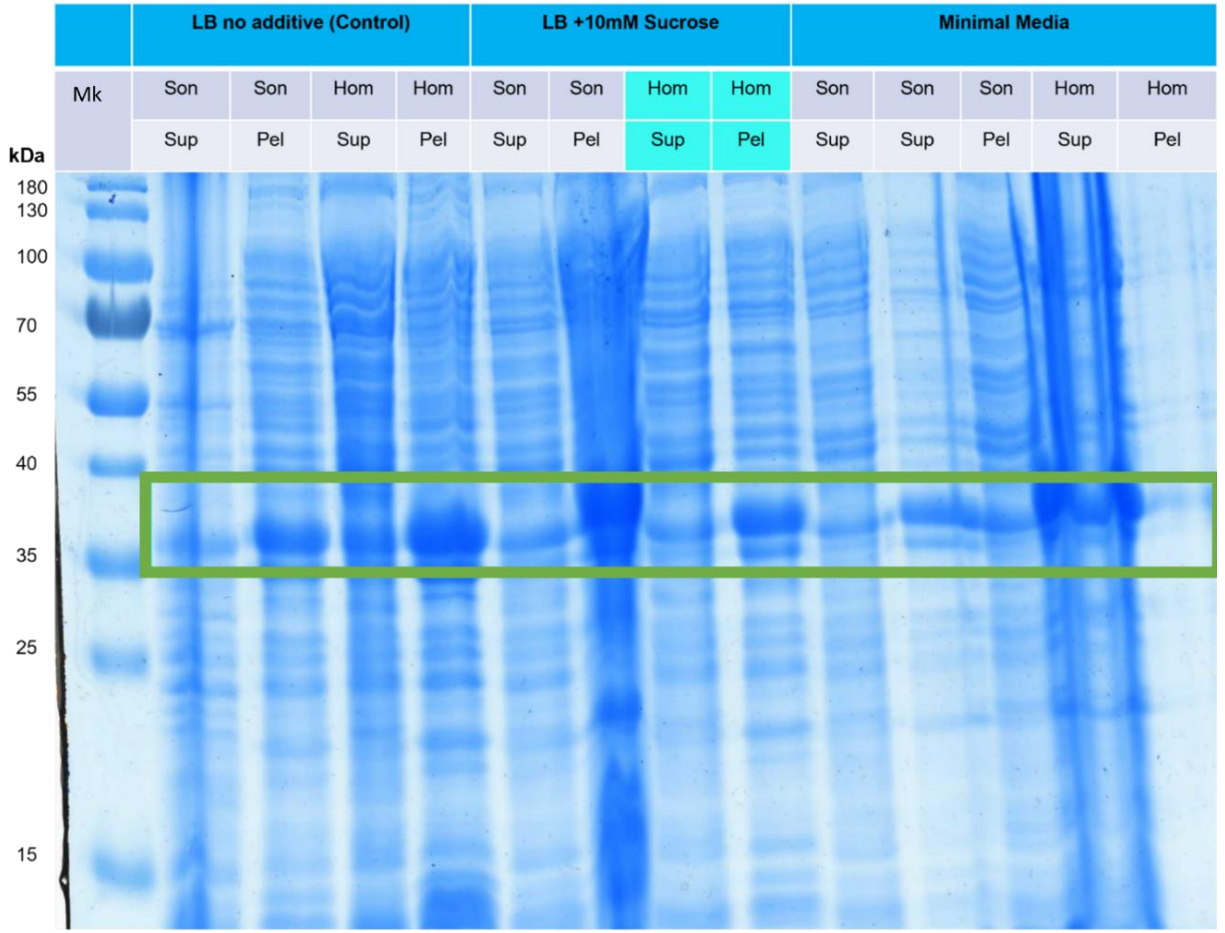


Figure 41: 1L lysis method, full scale up trials. SDS-PAGE gel (n=1) showing Pageruler Marker (M), Supernatant (Sup) and Pellet (pellet) (I). All cultures were grown to OD 0.9 then induced with 800µM IPTG and incubated overnight at 16°C for expression. Control LB no additive, identified optimal additive (10mM Sucrose) and minimal media trial.

Following the above trialling results culture in LB media with 10mM sucrose, induction at OD 0.9, followed by 16°C ON expression, and lysis via homogenisation were adopted as the optimal strategy to maximise soluble POI recovery going forward.

3.3.8 Optimised expression and purification strategy

Figure 42 shows the soluble (Sup) and insoluble (Pel)**,** M1 (POI) lysis product followed by the flow through (FT) where some protein was lost as indicated by the appropriately POI sized band (shown in a green box). Importantly there is a band present in the soluble, despite some unavoidable loss in the insoluble fraction. We were aiming to get soluble protein and this aim was met here via the optimisation changes employed. This full optimisation was only undertaken with M1 (the first mutant worked with *in vitro* and where progress was most difficult initially) due to time constraints. Given that the adopted changes weren't overly costly and difficult to employ they were adopted also with the other mutants, without further validation, to prevent delays in protein acquisition.

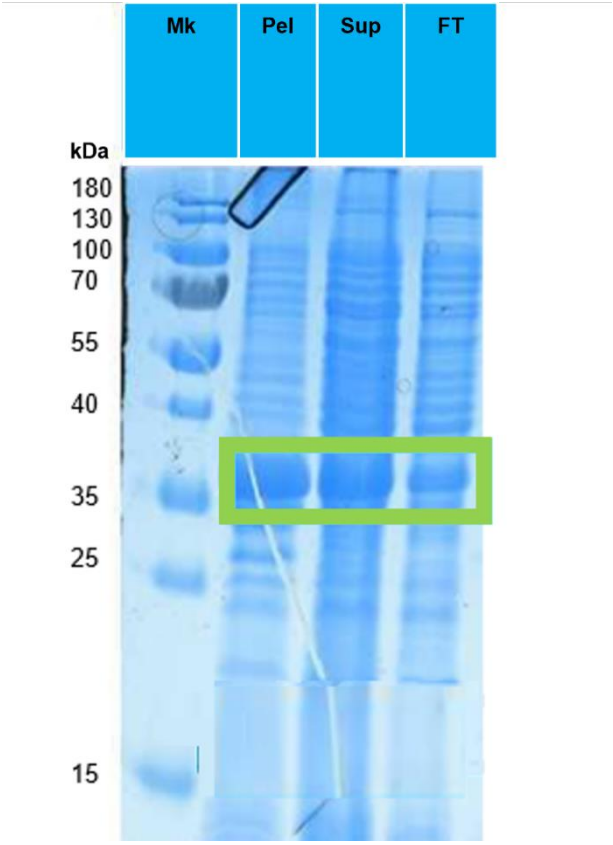


Figure 42: M1 lysis and 8L His purification summary. SDS-PAGE gel (n=1) showing Pageruler marker (Mk), pellet or insoluble (pel), supernatant (sup) and flowthrough (FT). The FT shows POI loss as band corresponding to POI is present reflecting some protein did not bind, possibly due to column overloading. To avoid this less protein was loaded in subsequent preparations at one time and protein was loaded more slowly to facilitate binding and always to a fresh recharged Nickel his column.

3.3.8.1 M1

Figure 43 is an SDS-PAGE gel showing the full range and content of fractions collected during the His purification for M1. Fractions 7-14 were pooled, tag cleaved during O/N dialysis, then subject to further gel filtration purification. This subsequent step was required as contaminating proteins were present in these fractions. Fractions 15-27 were homogeneous in content indicated by the single 33.55kDa band which coincides with what was expected for pure His-SUMO tagged M1. These fractions were therefore pooled and cleaved ON, then reverse his purified to separate protein and cleaved tag. Protein retained here (cleaved fraction 15-27 product collected in the RFT) was sufficiently pure after this reverse purification.

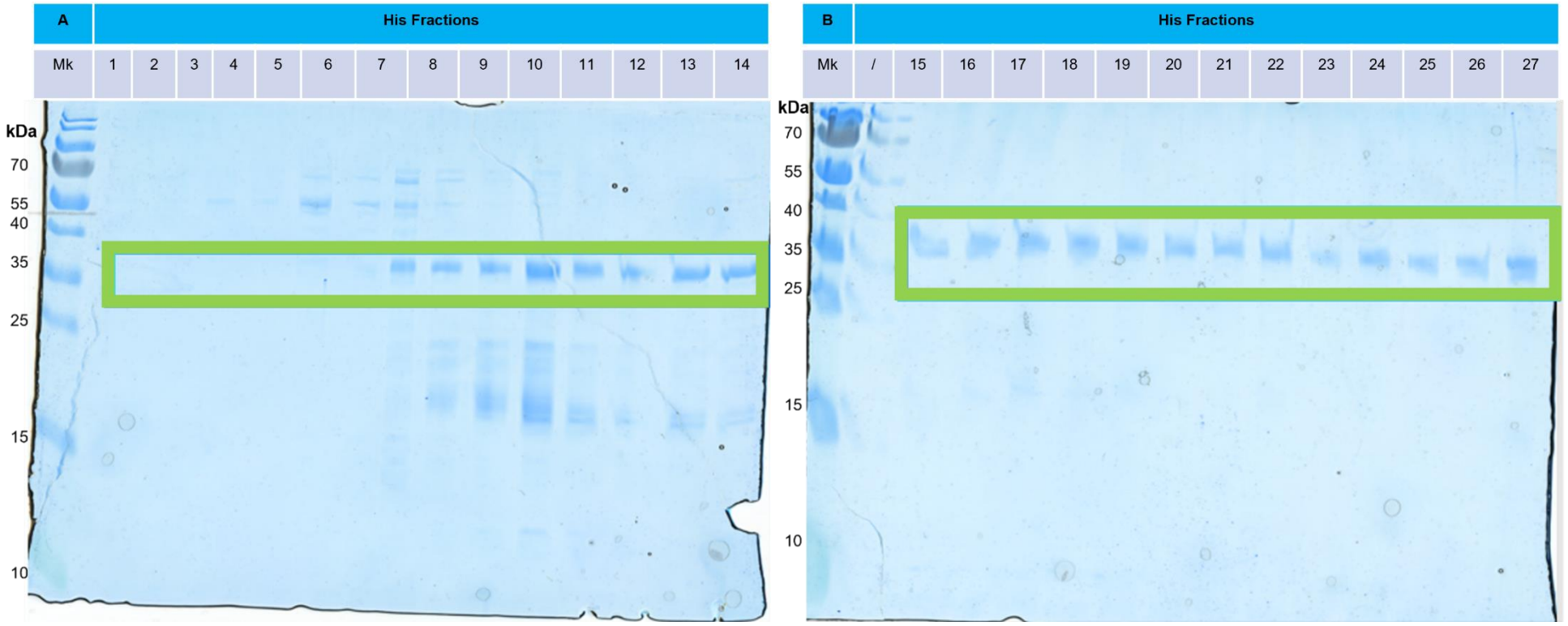


Figure 43: M1 His purification fractions. Size of M1 POI with His-SUMO-POI 33.5kDa. **A,** His fractions 7-14 were, pooled, POI had His-SUMO tag cleaved during ON dialysis with SUMO protease, then subjected to a subsequent gel filtration step to remove contaminating proteins. **B,** His Fractions 15-27 were pooled, POI had His-SUMO tag cleaved during ON dialysis with SUMO protease, reverse His purified to isolate just POI then concentrated and retained as pure for further planned experiments.

3.3.8.2 M2

Figure 44 is an SDS-PAGE gel showing the full range and content of fractions collected during the His purification for M2. Fractions 6-14 were pooled, tag cleaved during O/N dialysis, then subject to further gel filtration purification as contaminating proteins were present in these fractions. Fractions 15-26 however were pooled as significantly pure requiring only cleavage and reverse his purification. These fractions were predominantly homogeneous in content indicated by the single 35kDa band which coincides with what was expected for pure His-SUMO tagged M2. The contaminating bands in fractions 15-26 were faint, meaning not the most prevalent sample protein species. Also, these lower weighted/contaminating bands proteins were subsequently lost during dialysis and reverse his purification.

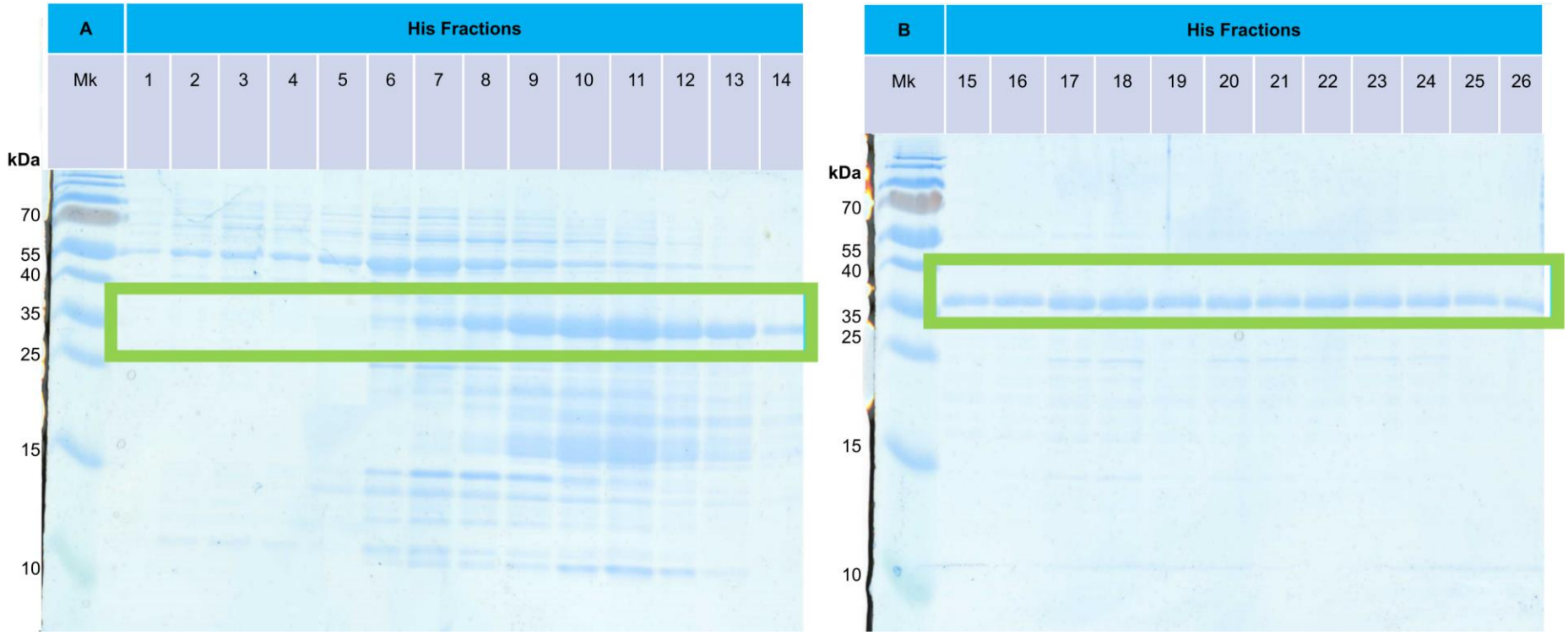


Figure 44: M2 His purification fractions. Size of M2 POI with His-SUMO-POI 33.4kDa. **A** fractions 6-14 were pooled, cleaved during dialysis with SUMO protease and subjected to subsequent gel filtration step. **B** fractions 15-26 were pooled and cleaved, reverse His purified and concentrated, then retained as pure for further planned experiments.

3.3.8.3 M5

Figure 45 is an SDS-PAGE gel showing the full range and content of fractions collected during the His purification for M5. Fractions 1-14 were pooled, tag cleaved during O/N dialysis, then subject to further gel filtration purification. This subsequent step was required as contaminating proteins were present in all these fractions. Fractions 15-26 were pooled and kept separately as they were slightly cleaner, lacking the higher weight contaminating proteins. This second protein pool also had tag cleaved during O/N dialysis, then subjected to further gel filtration purification. Importantly both His protein pools here showed the 35kDa POI band most intensely as the dominant protein present which coincides with what was expected for His-SUMO tagged M5. M5 had a much ‘dirtier’ His product with the most contaminants of all the mutants.

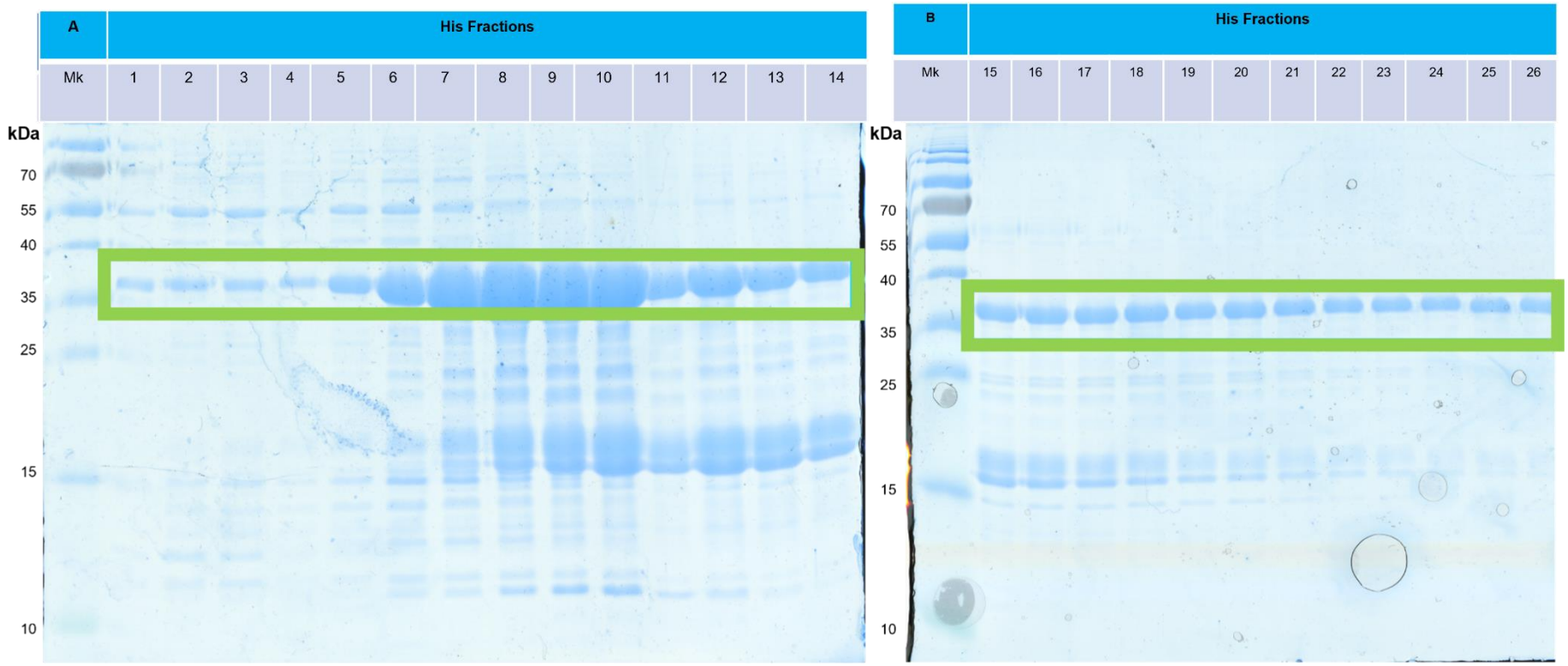


Figure 45: M5 His purification fractions. Size of M5 POI with His-SUMO-POI 33.4kDa. **A** fractions 1-14 were pooled, cleaved during dialysis with SUMO protease, reverse His purified **B,** fractions 15-26 were pooled and cleaved during dialysis with SUMO protease, reverse His purified then subjected to a subsequent gel filtration step.

3.3.8.4 M8

Figure 46 is an SDS-PAGE gel showing the full range and content of fractions collected during the His purification for M8. Fractions 1-14 were pooled, tag cleaved during O/N dialysis, then subject to further gel filtration purification. This subsequent step was required as contaminating proteins were present in all these fractions. Fractions 15-25 were pooled and kept separately as they were slightly cleaner, lacking the higher weight contaminating proteins. This second protein pool also had tag cleaved during O/N dialysis, then subjected to further gel filtration purification. Importantly both His protein pools here showed the 35kDa POI band most intensely as the dominant protein present which coincides with what was expected for His-SUMO tagged M8.

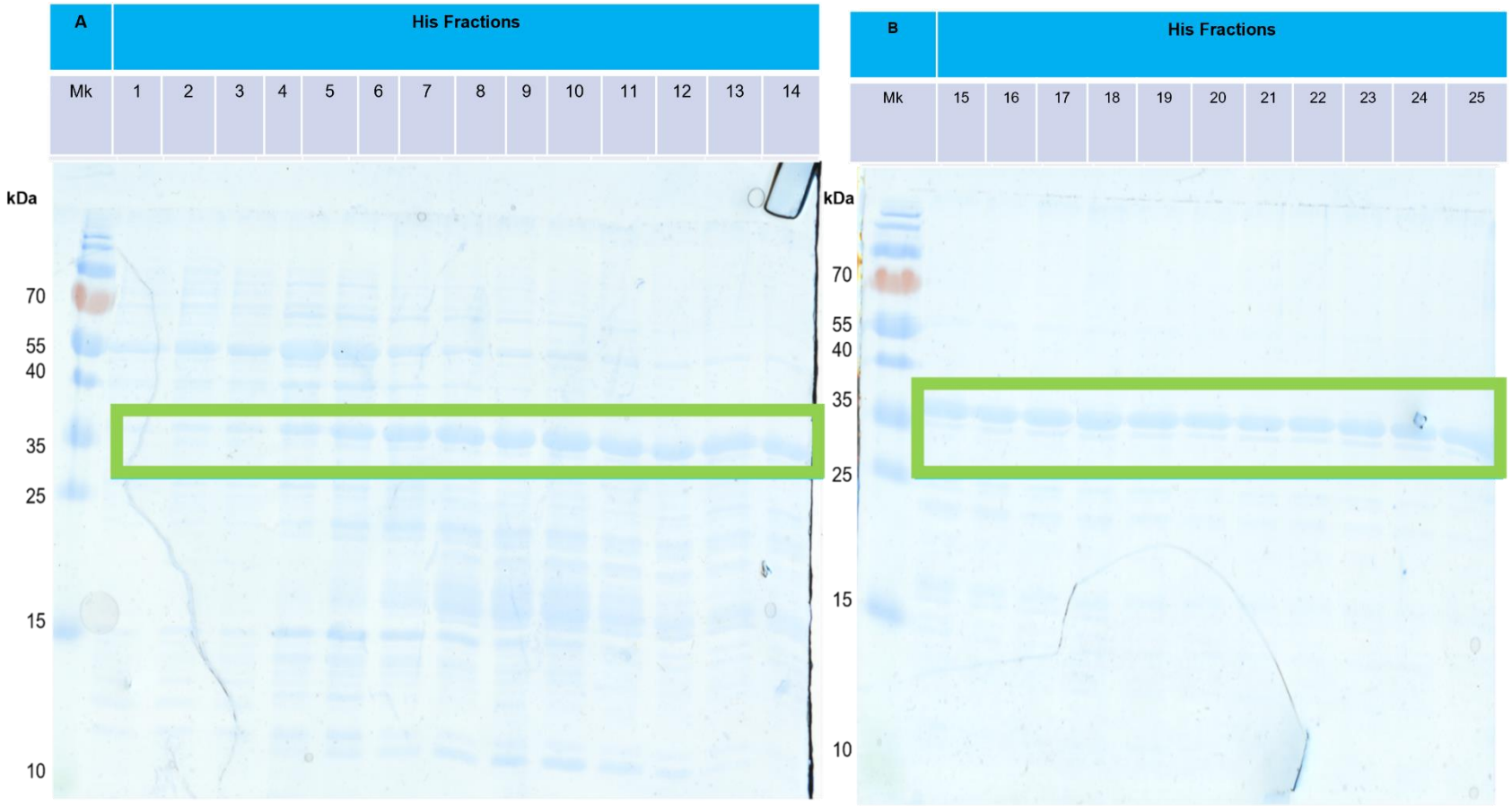


Figure 46: M8 His purification fractions. Size of M8 POI with His-SUMO-POI 33.5kDa **A,** fractions 5-14 were pooled, cleaved during dialysis with SUMO protease and subjected to subsequent gel filtration step. **B** fractions 15-25 were pooled and cleaved, reverse His purified and concentrated, then retained as pure for further planned experiments.

3.3.8.5 Reverse His purification

Figure 47 shows the reverse His purified M8 protein as an example from amongst these proteins (worked on in this chapter), with SUMO tag cleaved. The band shown in the box here on the SDS-PAGE gel is the reverse His purified M8 POI alone in collected in the Reverse flow though (RFT). Due to dilution for cleavage of the SUMO tag (actioned to prevent precipitation during dialysis) the RFT has very dilute faint bands at the appropriate size. As with 222 previously M8 runs slightly low, this is consistent for 222 and all mutants in this work. Concentration of the RFT causes the band to increase in intensity on the SDS-PAGE gel under concentrated RFT (CRFT).

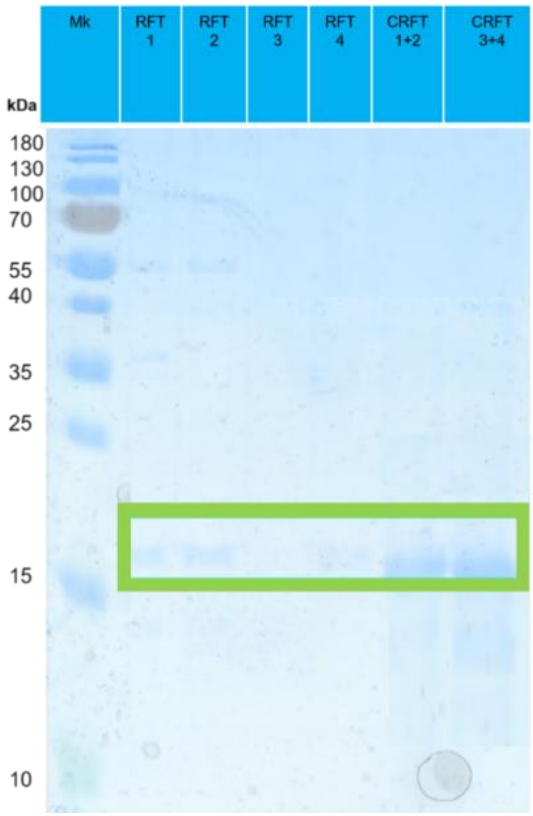


Figure 47: Reverse His M8 cleavage product gel. Size of M8 POI only is 20.52kDa. Reverse flow through (RFT) which contains the SUMO tag cleaved M8 product that no longer binds the column. Product was diluted pre dialysis and cleavage to try and counter precipitation issues encountered, in initial mutant preparations, precipitation during cleavage was particularly an issue with M8. Here multiple preparations were run consecutively to produce large scale whilst being mindful to not overload the single His column. The numbers following RFT refer to the different preparations. Concentrated reverse flow through (CRFT) is the concentrated reverse his purified POI. Preparations were pooled during concentration to achieve a pure product.

3.3.8.6 Gel filtration

For M1, M2 and M8 some sufficiently pure POI was attained from the reverse His purification step this was pooled, concentrated, and retained for characterisation experiments. M5 gave only heterogenous His fractions so all required pooled and subsequent gel filtration. A portion of the M1, M2 and M8 His fractions with heterogenous content were also pooled and subjected to gel filtration. The SDS-PAGE results shown below in **Figure 48** **(**for M1 **A** and M2 **B**) and **Figure 49** **(**for

M5 **A** and M8 **B**) show the fractions attained from the subsequent gel filtration purifications. From this additional gel filtration purification step fractions 11-19 were retained for M1, fractions 28-30 were retained for M2, fractions 21-31 were retained for M5 and fractions 25-31 for M8.

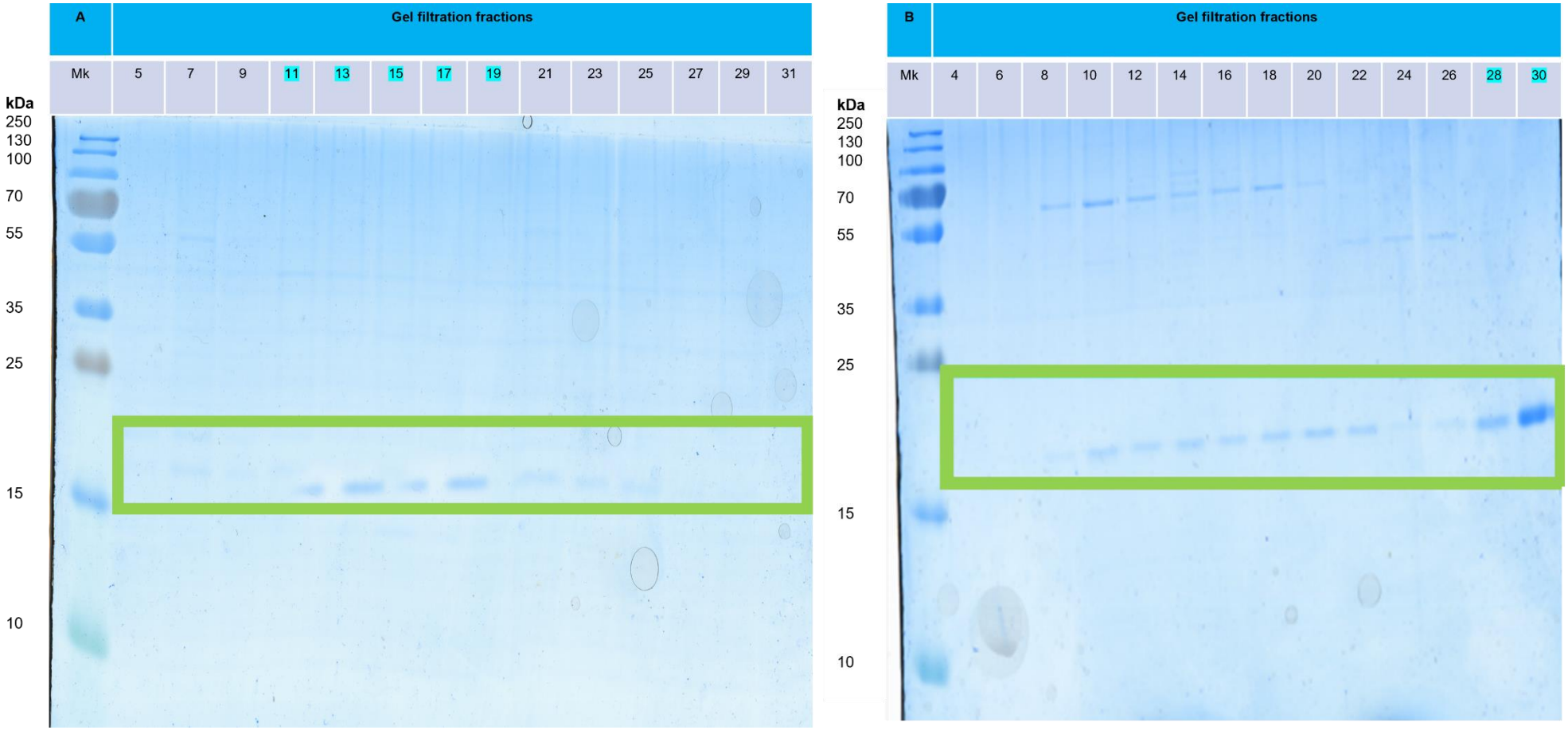


Figure 48: SDS-PAGE analysis of gel filtration fractions. (A) M1 (20.52kDa) **(B)** M2 (20.40kDa)**.** Those highlighted in light blue were pure so retained, pooled, concentrated, and stored for use in subsequent characterisation experiments.

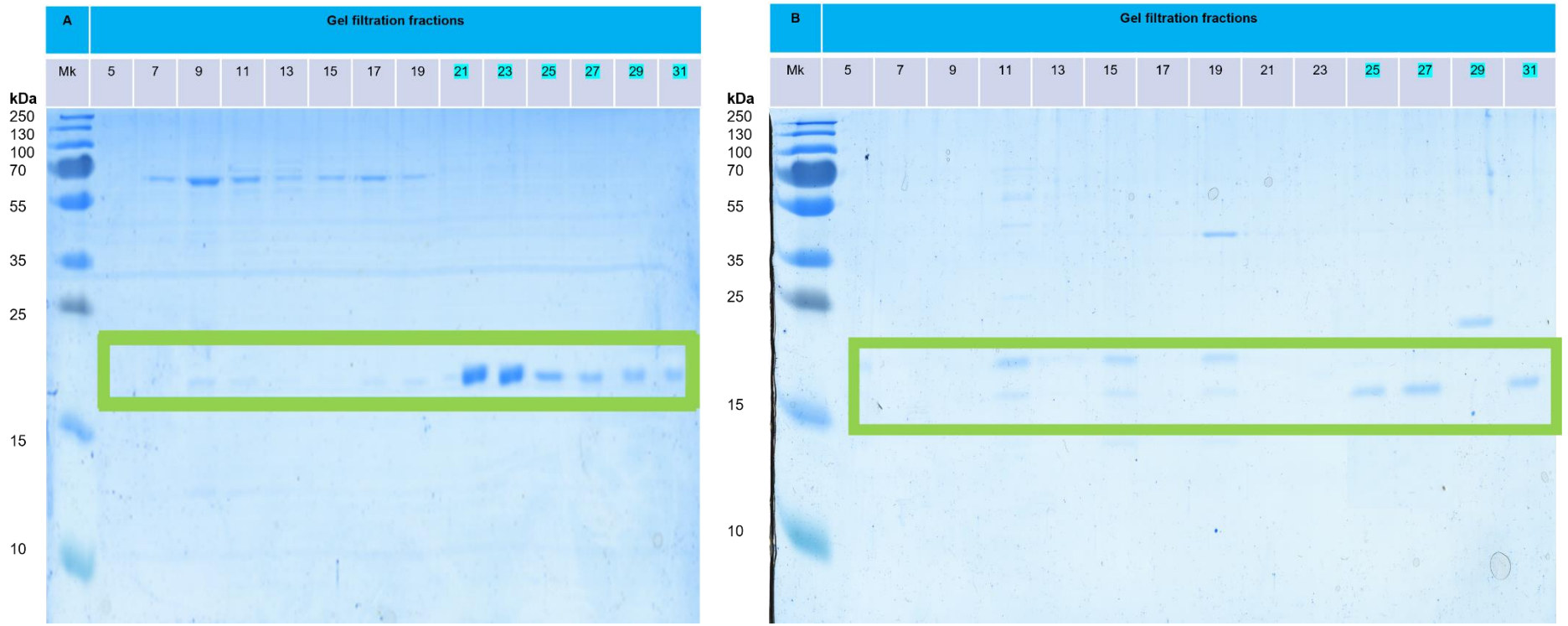


Figure 49: SDS-PAGE analysis of gel filtration fractions. (A) M5 (20.37kDa) **(B)** M8 (20.48kDa)**.** Those highlighted in light blue were pure so retained, pooled, concentrated, and stored for use in subsequent characterisation experiments.

3.3.8.7 All mutants

Figure 50 shows a final gel of all mutant proteins after all purification steps, pooling, and concentration, taken forward to *in vitro* testing. The homogeneous intense band was the sought-after result indicating good levels of appropriately sized protein and purity.

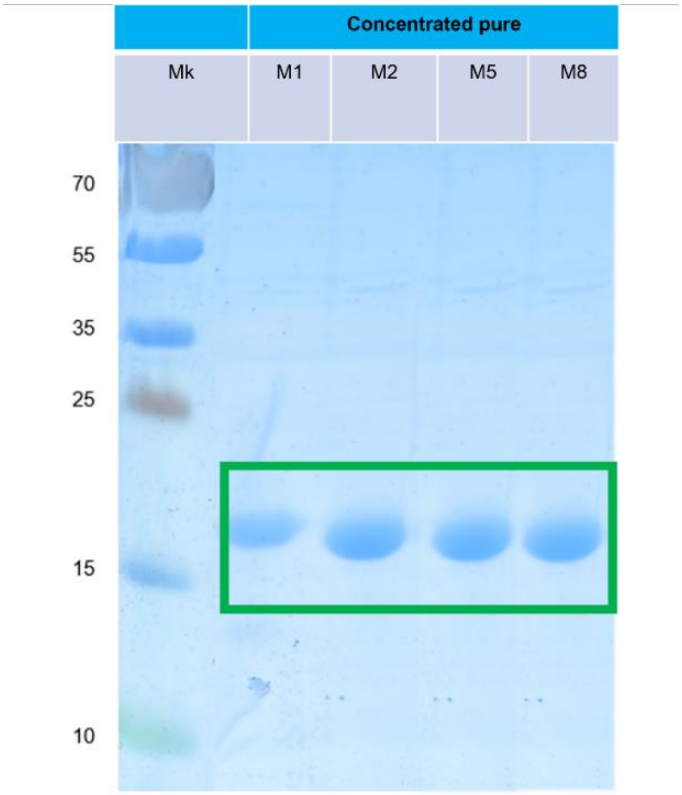


Figure 50: Final SDS-PAGE assessment of concentrated pure mutant proteins. Pageruler marker (Mk), M1 (20.52kDa), M2 (20.40kDa), M5 (20.37kDa) and M8 (20.52kDa) confirmed here to be adequately pure and corresponding to the correct molecular weights.

Table 24 gives a summary of the conditions utilised for 222 expression and mutant proteins generated for this chapter, giving details of lysis method, mean yield and standard deviation from three individual protein preparations and general notes regarding recovery. M5 was the easiest to recover and M8 was the most difficult. M2

however was the most variable in terms of yield. M1 was the mutant that underwent the most trials but was also the first worked on. M2, M5 and M8 were worked on in tandem.

Table 24: Summary of the conditions utilised for 222 expression and lysis for mutant proteins in this chapter. Details of lysis method adopted, mean yield from three separate productions (n=3), standard deviation for this, yield order and general notes regarding purification recovery.

Protein	Additive	Expression OD	IPTG concentration	Expression Temp	Shaking Speed	Lysis method	Mean Yield (mg/L), n=3	Standard Deviation	Yield order	Notes
222 (control)	N/A	0.6	1M	18°C	180RPM	Sonication	3.5	±0.2	2	Fairly standard and replicable recovery
M1	10mM Sucrose	0.9	800µM	16°C		Homogenisation	2.3	±0.6	4	Had issues with scale up initially
M2			400				2.8	±1.4	3	Very variable between preparations
M5							13.3	±0.8	1	Easy recovery, easier to express and purify than all other proteins including 222
M8							0.7	±0.2	5	Consistently low recovery, Precipitation issues seen during cleavage and dialysis

3.3.9 Protein characterisation

3.3.9.1 Mass spectrometry

Mass spectrometry (MS) was used to confirm the identity of M1, M2, M5 and M8 mutant proteins , with a fragment containing the correct mutation intended for each one (**Table 25**) being present when compared to 222.

Table 25: Protein sequences for this chapter (chapter 3) with unique MS fragments. Protein names, abbreviations, alanine mutation sites stating the residue in 222 the 'native' in this chapter and the numbered position of each mutation site, as well as the protein sequences. The residues shown in blue text were left as in the native CBD protein, as they were deemed had critical involvement in intramolecular interactions. The three modules for each protein are shown in red text, and the linker regions in black text. Alanine (Ala) substitution sites for each mutant are highlighted in yellow. Key confirmatory MS fragments were identified using pep cutter available at https://web.expasy.org/peptide_cutter/, accessed 10/21 and Microsoft Excel's match function to identify unique peptides for each protein.

Name	Mutation site	Protein sequence	Key confirmatory MS fragments	Molecular weight (kDa)
222 (native)	N/A	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	• EGQVVFTMYGN AEGQPCK • YGFCPHEALFT MGGNAEGQPC K • YGFCPETALFT MGGNAEGQPC K • FPFR • TDGYR • WCGTTEDYDR	20.65
Mutant 1 (M1)	N 11,69,127	EGQVVFTMYGAAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGAAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGAAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	• EGQVVFTMYG AAEGQPCK • YGFCPHEALFT MGGAAEGQPC K • YGFCPETALFT MGGAAEGQPC	20.52
Mutant 2 (M2)	R 22, 80, 138	EGQVVFTMYGNAEGQPCKFPFAFQ GTSYDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFAFQGTSYDSCTTEGRTDG YRWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFAFQGTSYDS CTTEGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	• FPFAFQGTSYD SCTTEGR	20.40
Mutant 5 (M5)	Y 40, 98, 156	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGARWCGTTED YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG ARWCGTTEDYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGARWCGTTEDYDRDKK YGFCPDQGYSL	• TDGAR	20.37
Mutant 8 (M8)	E 47, 105, 163	EGQVVFTMYGNAEGQPCKFPFRFQ GTSYDSCTTEGRTDGYRWCGTTAD YDRDKKYGFCPHEALFTMGGNAEG QPCKFPFRFQGTSYDSCTTEGRTDG YRWCGTTADYDRDKKYGFCPETALF TMGGNAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTADYDRDKK YGFCPDQGYSL	• WCGTTADYDR	20.48

3.3.9.2 Disulfide assessment

In silico analysis of disulfide bonds was conducted using PredDisulfideBond and MAESTRO web. Results of these assessments showed that all proteins are predicted to have the same total number of six from a possible nine disulfide bonds (**Table 26**). Results of both tools agree that the same Cys residues are involved in the disulfide bonds in 222 and mutant proteins in this chapter. Bonds shown in black are the most probable to form, bonds shown in red are the bonds that are predicted won't form due to the residues being more likely to form another disulfide bond first. Once used to form one bond via oxidation electrons of the sulfhydryl group of that cysteine residue are not then available to form another.

Table 26: Results of the PredDisulfideBond [386]and Maestro [348] webserver predictions of disulfide bonds for all proteins in this chapter. Namely: 222, M1, M2, M5 and M8. In the first column are the bond predictions, detailing the cysteine (Cys) residues involved in the possible disulfide bonds. Disulfide bonds form between the sulfhydryl (SH) side chains of two cysteine residues. The probability of each bond forming calculated by the PredDisulfideBond tool is given in the adjacent column. The higher the probability the more likely a bond is to form. The next column gives the Maestro bond score ($S_{ss)}$, which is a way to rank potential bonds. The lower the score the more likely a bond to form. Each cysteine can only be involved in one disulfide bond at a time. In instances where there are two possible bonds arising from the same residue the bond with the higher probability (bond with the lower score), is assumed to be formed. This results in three possible bonds not predicted to be formed (shown in red), with agreement across both tools used.

Cys residues forming disulfide bond	222		M1		M2		M5		M8	
	PredDisulfideBond Probability	Maestro bond score (Sss)	PredDisulfideBond Probability	Maestro bond score (Sss)	PredDisulfideBond Probability	Maestro bond score (Sss)	PredDisulfideBond Probability	Maestro bond score (Sss)	PredDisulfideBond Probability	Maestro bond score (Sss)
CYSA17-CYSA43	0.988	-1.25477	0.988	-1.26299	0.988	-1.24608	0.988	-1.2562	0.988	-1.23904
CYSA31-CYSA58	0.998	-0.35179	0.998	-0.36001	0.998	-0.3431	0.998	-0.35322	0.998	-0.34004
CYSA75-CYSA101	0.987	-1.01159	0.987	-1.01981	0.987	-1.0029	0.987	-1.01302	0.987	-0.99693
CYSA89-CYSA116	0.992	-0.77988	0.992	-0.78811	0.992	-0.77119	0.992	-0.78242	0.992	-0.76787
CYSA133-CYSA159	0.875	-0.36565	0.875	-0.37387	0.875	-0.77119	0.875	-0.3653	0.875	-0.35125
CYSA147-CYSA174	0.998	-0.77988	0.998	-0.78811	0.998	-0.77119	0.998	-0.78242	0.998	-0.76787
CYSA31-CYSA43	0.981	0.572124	0.981	0.563899	0.981	0.580815	0.981	0.572469	0.981	0.582377
CYSA89-CYSA101	0.977	0.682142	0.977	0.673917	0.977	0.690834	0.977	0.682488	0.977	0.69191
CYSA147-CYSA159	0.992	0.021001	0.992	0.012776	0.992	0.029692	0.992	0.019921	0.992	0.031608

Figure 51 shows where the disulfide bonds predicted in **Table 26** were positioned in the models for all the proteins analysed *in vitro* in this chapter. The mutations were not predicted to impact disulfide configuration.

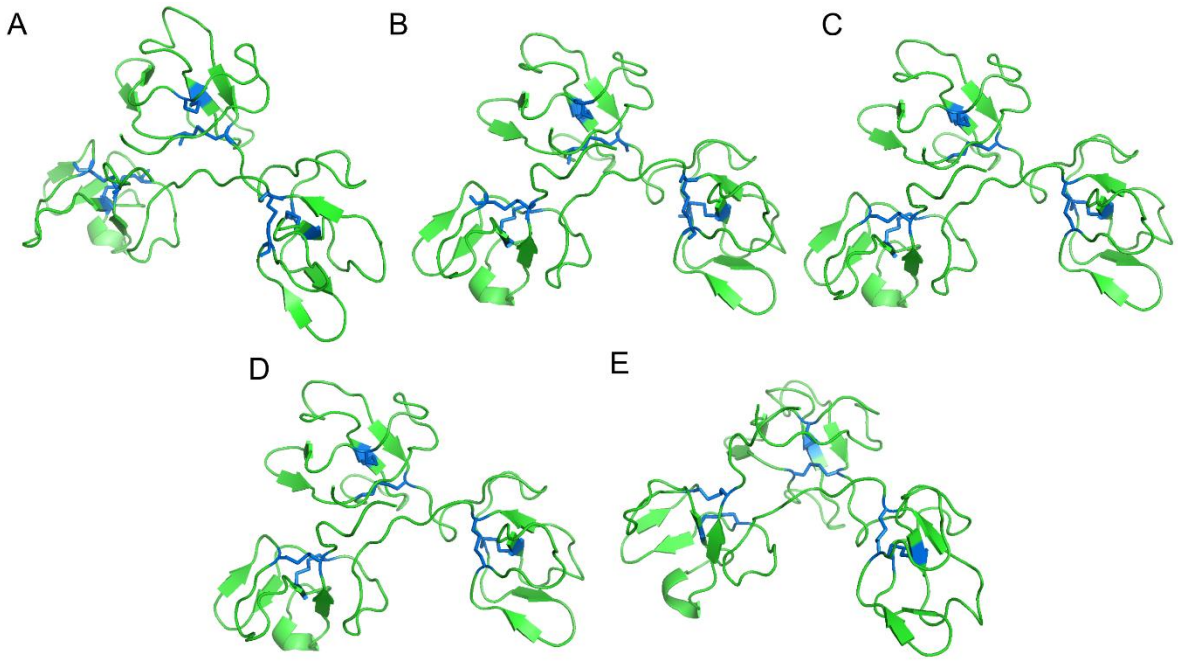


Figure 51: Disulfide bond positions in for all proteins in this chapter. (A) 222, **(B)** M1, **(C)** M2, **(D)** M5 and **(E)** M8. The bonds shown are those identified in the predictions outlined in **Table 26** highlighted in blue on the protein models generated and selected earlier in the chapter (See **results section 3.3.2 & 3.3.5**), six disulfide bonds per model 2 per module as was expected. The alanine mutations were not predicted to impact disulfide bond configuration.

In vitro analysis of disulfide bonds was undertaken using SDS-PAGE analysis, with and without reducing agent in the gel loading buffer. In non-reduced samples (NR) disulfide bonds are still present and make the proteins more compact, hence NR samples run faster on a gel resulting in lower (seemingly lighter) bands [387]. Reduced samples (R) with disulfide bonds no longer present are less compact so run slower on the gel resulting in higher (seemingly heavier) bands. This is seen as expected for 222, M5 and M8 by the band shift in (**Figure 52**) indicating disulfides

were present and alanine mutations had not changed the disulfide configuration. For M1 and M2 this shift pattern is not seen to be the same indicating that there has been some change to the disulfide configuration, meaning mutations here have impacted protein folding despite in silico predictions not identifying this.

Figure 52: In vitro disulfide assessment for all mutants in this chapter. Pageruler marker (Mk), Reducing (R) vs Non-Reducing (NR) SDS-PAGE gel disulfide assessment. The disulfide bridges present in NR samples make the proteins more compact so make it run faster on a gel [387]. 222 (20.65kDa), M1 (20.52kDa), M2 (20.40kDa), M5 (20.37kDa) and M8 (20.48kDa). 222, M5 and M8 have the same gel shift pattern with R samples as expected and running slower so appearing higher on the gel than NR. M1 and M2 do not demonstrate this same disulfide shift.

This change to the disulfide configuration was observed twice, as a replicate gel run was used to confirm this unexpected result. Closer examination of the disulfide sites *in silico* using pymol gave no explanation for this result. *In silico* results

gave no implication that M1 and M2 were any different but experimentally the gel assessment showed that disulfides were altered.

3.3.9.3 Circular dichroism

CD was used to assess the secondary structure of the proteins in this chapter, and to monitor any conformational changes arising between the mutants in this chapter and 222. Two spectral patterns were observed (**Figure 53**)**.** M1, M2 and M8 showed one spectral pattern (Group CD1). In contrast, 222 and M5 showed a different spectral pattern (Group CD2). The three proteins grouped together as CD1 displayed a spectral maxima at between 190-195nm and a minimum between 205-210nm. The proteins grouped together as CD2 showed less defined spectra, indicating a less defined secondary structure.

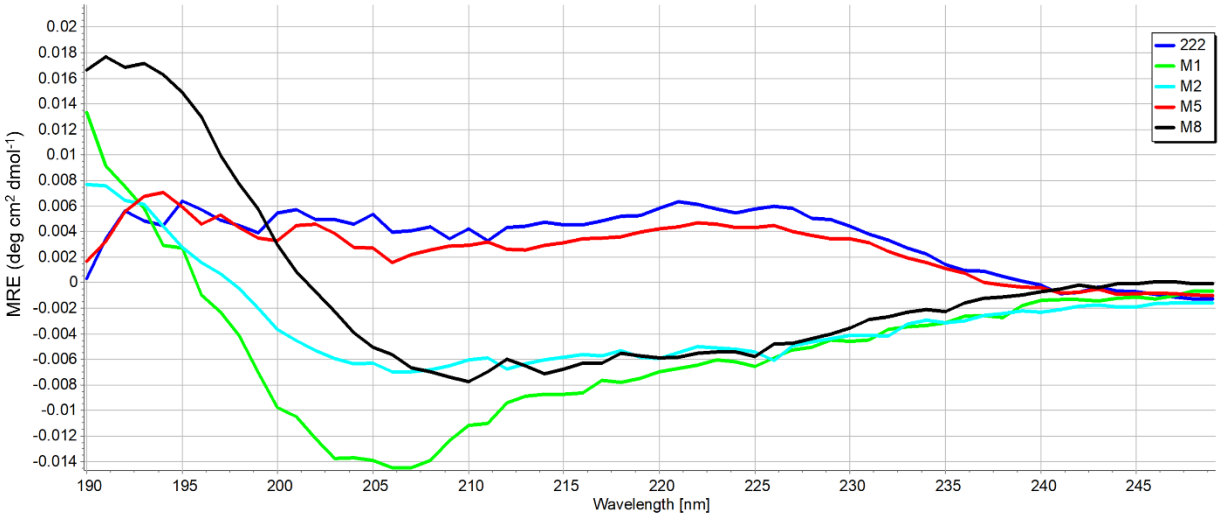


Figure 53: Far-UV CD spectra of 222, M1, M2, M5 and M8 in 2.5mM HEPES, pH 6.5. CD1 grouping composed of M1, M2 and M8. CD2 grouping consists of 222 and M5. There is a clear difference in structure in the two groupings.

The raw ellipticity CD data (delta epsilon and wavelength) for each protein was submitted to the BestSel server single spectrum analysis and fold recognition

function. This gave the estimated secondary structure content (%) table output (**Table 27**) and spectral fitting results with RMSD and NRMSD values. It is generally considered that RMSD measures the predictive power of the method, with lower values indicating better spectral fits [388]. NRMSD (normalised RMSD) facilitates the comparison between models from different CD experiments [389]. Hall et al, 2014 define good CD fits and structure estimates if the NRMSD is <0.03; reasonable for NRMSD <0.05; and variable above this.

Applying this to the data here shows a reasonable fit for M1 and M2, with a good fit for M8. Whereas 222 and M5 have a variable fit based on the NRMSD (**Table 27**). Therefore, too strong a conclusion cannot be made about 222 and M5 based on the initial Bestsel analysis, which fits with the summation above that the CD2 proteins have less defined secondary structure . Further analysis of the CD spectra using the Bestsel tool (**Table 27**) shows that CD1 proteins have no α-helix structure whereas the proteins in CD2 do**.** M5 was the easiest of the proteins to produce (highest yield, **Table 24**), it had a reduction in defined secondary structure composition looking at just the spectra (**Figure 53**) however Bestsel analysis does not fit this trend as M1 in group CD1 has a lower value of defined structure. Interestingly CD1 proteins had the highest percentage antiparallel β sheet, ≥37.2% whereas CD2 proteins all had much lower β sheet of ≤24.8%.

Table 27: Bestsel analysis of CD spectral results. Showing % composition of each secondary structure type for 222 and all mutants taken to *in vitro* experiments in this chapter. The CD1 grouping proteins are shown in black text and CD2 grouping proteins as shown in blue text (grouping based upon spectra). Root-mean-square deviation (RMSD) reported to show difference between observed (measured spectral values) and predicted spectra of best fit identified by BestSel. A lower RMSD value indicates less discrepancy, so an increased agreement/accuracy. NRMSD is reported to allow definitive quality comparison between different CD experiments. CD1 had too much variability to draw strong conclusions from this analysis.

Group	CD1	CD2	CD2	CD1	CD2
Secondary structure	222	M1	M2	M5	M8
Helix	0	22.3	21.5	0	44.4
Antiparallel (β-Sheet)	50.5	24.8	4.2	45.5	16.7
Parallel (β-Sheet)	0	0	6.9	0	0
Turn	0	0	18.2	2.4	9.5
Total defined	50.5	47.1	50.8	47.9	70.6
Others	49.5	52.9	49.3	52.1	29.3
RMSD	1.2309	0.9367	0.5343	0.8971	0.3517
NRMSD	0.07574	0.04712	0.05059	0.06313	0.1957

3.3.9.4 Stability

3.3.9.4.1 *In silico* stability prediction

The Scoop webtool was utilised to give a prediction of any differences to stability by allowing comparison of a predicted Tm by submitting the structure file models (PDB format) generated for each protein. Results of this assessment are shown in **Table 28** and predict that the alanine mutations in the mutants in this chapter made no real difference to stability with all Tms between 61.3-63.5°C.

Table 28: Scoop prediction of midpoint melting temperature (Tm). Determined by the Scoop webserver algorithm, protein structure files (PDB) and host organism *E. coli*.

SCooP	222	M1	M2	M5	M8
Stability Prediction (Tm, °C)	61.3	63.5	62.3	61.8	61.4

3.3.9.4.2 NanoDSF

NanoDSF utilises intrinsic tryptophan (trp) and tyrosine (tyr) fluorescence changes resulting from alterations in the 3D-structure of proteins as a function of temperature as proteins unfold. Tyr and trp are hydrophobic residues usually found buried in the core of proteins, so when the proteins unfold, they become more exposed and fluorescence increases. NanoDSF enables a label-free quantification and comparison of the stability of different proteins.

Typically fully unfolded protein fluoresce at 350 nm while folded protein fluoresce at ~330nm [390]. As a protein unfolds, with a temperature gradient applied to the capillaries into which the proteins are loaded, measured fluorescence at the

two wavelengths 330 and 350nm changes as the residues go from buried to exposed. A ratio was calculated from these two wavelength measures and used to derive mean midpoint Tm using the inflection points of the NanoDSF ratio data for each protein (shown in **Figure 54**).

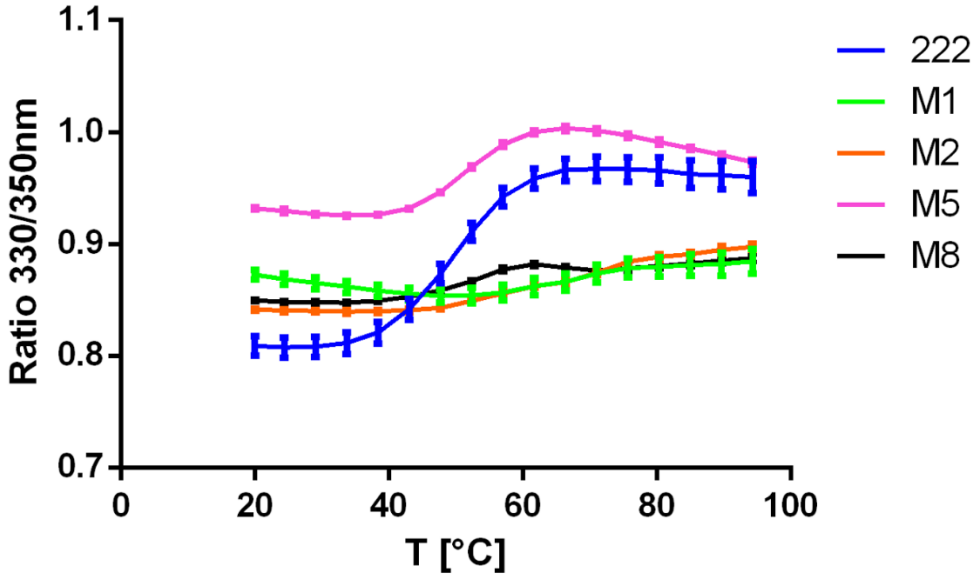


Figure 54: NanoDSF results for 222, M1, M2, M5 and M8. NanoDSF 320/350nm ratio plotted against temperature (n=3). This stability assessment utilises intrinsic tryptophan and tyrosine fluorescence changes resulting from alterations of the 3D-structure of proteins as a function of the temperature as the protein unfolds. This is a method able to monitor a proteins melt profile. These curves were used to derive mean Tm value (Melting temperature, or point at which 50% of a protein is unfolded) showing stability changes in the mutants compared to 222. Higher Tms indicate more stable variants. Curves attained for M1, M2 and M8 were atypical and there was no clear transition so Tm could not be attained for these mutants.

Table 29 shows a summary of the Tm values derived as a quantitative measure of protein thermal stability and provides a benchmark to compare the favourability of different mutant proteins with typical melt curves (with defined transition points). 222 had a Tm of 49.10°C determined from its traditional melting curve with a clear ratio transition. Only M5 gave a curve that a Tm could be derived

from. 95% confidence intervals show overlap so here no notable difference can be concluded. 222 Tm was 49.10°C, and for M5 Tm was 50.3°C, showing no notable difference in stability.

Table 29: Mean Tm value summary data for 222 and M5. Tm could only be derived for those proteins that gave a typical melts curve with clear transition indicating unfolding. 95% confidence intervals Tm range and span (of the error bars) are also shown for each protein. The Tm values show that M5 is not largely different in stability compared to 222 especially when accounting for the 95% confidence interval span. .

	222	M5
T_m [°C]	49.10	50.30
95% Confidence Intervals		
T_m [°C]	48.17 to 50.06	48.98 to 51.64
Span	0.1492 to 0.1643	0.05797 to 0.06843

3.3.9.5 Binding assay

Binding assays were carried out to elucidate which of the NMR-identified binding residues were most critical to binding, by identifying the alanine mutation leading to the most reduced binding affinity compared to the 222 positive control.

In this work the dissociation constant (Kd) is determined from characteristic sigmoidal assay curves and used to evaluate binding and strength of bimolecular interactions. The smaller the Kd value, the greater the binding affinity of the POI for its target TII gelatin. **Figure 55**, shows the binding assay curves for 222 and all four mutants expressed in this chapter. **Table 30** shows a summary table of the Kd values determine along with standard error and importantly the binding order. The data show that M1 is the mutant which had the greatest loss of binding affinity

compared with 222, with a kD as high as 1µM, meaning the non-mutant Asn residues at position 11, 69 and 127 are most critical to binding from the selection of mutation positions evaluated in this work. The 222 result here replicates binding experiments previously conducted by the Hollander group [235]. The previous work of the group gave an apparent average Kd for 222 to TII gelatin of 1.46 ± 0.53nM, whilst the work reported here, in line with this, gave a negligibly higher result of 3.99 x ±1.65nM.

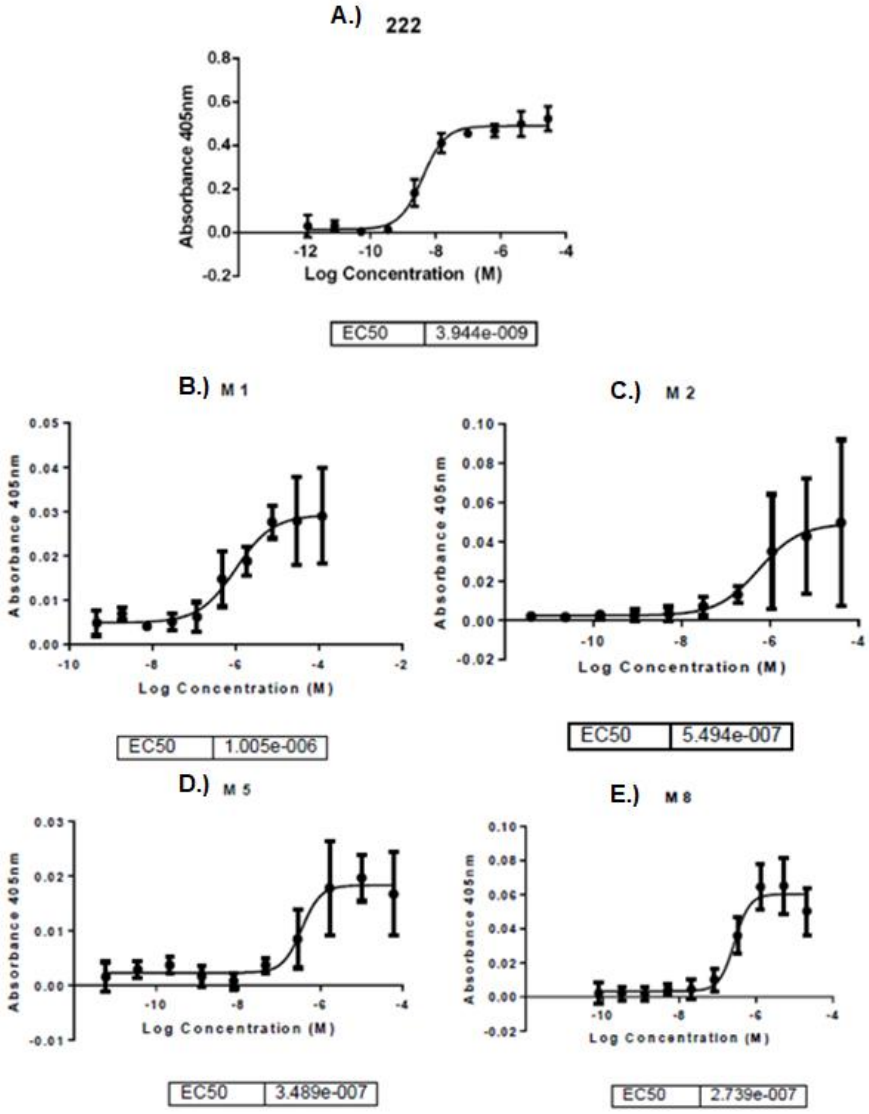


Figure 55: Binding curves for the proteins in this chapter to TII gelatin. Absorbance values shown here are the averages from across all the plates. n=3 replicates across n=3 plates, except M2 which was only n=3 replicates across n= 2 plates due to protein stock. Error bars represent standard deviation of the three plate replicates. **A** 222, **B** M1, **C** M2, **D** M5, and **E** M8. EC50 values (M) which are the Kd values derived from these curves, are shown for each.

Table 30: Binding affinities summary of all proteins in this chapter. Kd is the equilibrium dissociation constant, used to evaluate and rank order strengths of bimolecular interactions. The smaller the Kd value, the greater the binding affinity of the ligand (mutants) for its target (TII gelatin). Kd is the concentration at which half of the protein is bound and half unbound. Kd shown here in nM, along with Standard error (Std error) and binding order for 222, M1, M2, M5. Kd values were determined using the curves in **Figure 55**.

Protein	Kd (nM)	Standard error (nM)	Binding Order
222	3.9	±0.7	1
M1	1005	±490	5
M2	549.4	±614	4
M5	348.9	±145	3
M8	259.9	±57	2

3.4 Discussion

This chapter has described the computational determination of 222 structure and subsequent mutant design. Homology modelling and alanine mutagenesis successfully guided the *in vitro* work, to meet the primary aim of the chapter which was identifying the most important binding residues. *In silico* analyses of stability and solvent exposure provided a strategy to reduce/ focus efforts to mutants more likely to be successful *in vitro*. As suggested in a recent article from Marabotti et al, 2021 [391] several predictors were used as an effective way to increase reliability of

in silico stability predictions. This multiple tool logic was also applied to subsequent computational assessment of solvent exposure and disulfide bonds. Expression work was likely to be easier in mutants that were more stable. The effect of mutation on binding affinity was likely to be more profound where mutated residues were more solvent exposed to begin with.

It is important to consider that homology models are only indirectly based on experimental data so there is something of a gamble involved as to whether they are correct. In this work every effort was taken to ensure that models were strategically selected and employed using best current practice. However, there was an alternate template identified during the BLASTp search which was not investigated further. Both templates were a strong match and although the decision to go with the stronger candidate was valid, if the other was the only match, it would have been a sufficient quality match to allow modelling to proceed confidently. It will be interesting to see in the coming years whether machine learning and pattern recognition will advance the field of protein design, just as they have with structure prediction. Since the work described here was undertaken, a new and apparently highly accurate protein structure prediction tool has emerged and been widely adopted as the leading approach. Alphafold2 is the most recent version of this AI system, developed by Deepmind [392]. In 2020, Alphafold2 won the 14th critical assessment of structural prediction competition (CASP14) and has since been adopted extensively as the gold standard approach to predicting protein structures [393]. (NB had this been available at the time of this work it would have been utilised as the optimal approach; it was adopted in **chapter 5**).

Initial small scale expression of selected mutants seemed positive with conditions for soluble protein expression identified, protein expression scale up however proved to be challenging. Generating soluble protein of the first mutant M1, showed variability between repeat experiments, highlighting the temperamentality of the process and the effect of several different factors in recombinant protein expression that cannot be consistently controlled. Aeration and oxygen transfer are not always maintained between different cultures and volumes despite best efforts [394]. Sonication has a less efficient cooling than homogenisation (as a method of lysis) hence protein can be lost to denaturation more readily [395, 396], which could explain some of the issues experienced with low yield. Also sonication tip position can have an effect on mixing which can introduce variation [397]. In the work here between the limited different purification attempts, more variation in yield was seen with sonication. Homogenisation is also better suited to lysing large culture volumes due to apparatus processing capacity. These observations justify the choice to employ chemical (lysis buffer constituents) then homogenisation as a method of lysis with subsequent mutants after M1.

It was noted that M5 had the most contaminants following His-purification. It could be that the high protein concentration increased the chance of proteins sticking, together meaning other contaminant proteins co-assembled with POI and were pulled through the His purification [398]. Experimenting with an elution gradient could have cleaned up the purification here but protein was progressed to gel filtration instead. These mutants were designed to disrupt binding to TII gelatin compared to 222. Perseverance and a methodical approach paid off here, to

generate sufficient protein for all target mutants to achieve the primary aim of enabling biophysical assessment of stability and binding to gelatin.

One key attribute that was observed in the modelling results is that all proteins studied (222 and mutants), are largely unstructured with just a small amount of β sheet structure, therefore subsequently it can be hypothesised that all the proteins in this chapter adopt a range of conformations in solution preceding binding. This range of conformations hypothesis could explain the issues seen in this project and the work undertaken by the group previously, with variability and inconsistency of purification and some error range seen in binding assay characterisation. The CD results analysed using Bestsel (**Table 27**), show 222, M1, M2 and M5 have only ~50% other secondary structure, whilst M8 has the most, with 70.6% defined structure. M8 was the protein with the lowest yield. The *in vitro* secondary structure analysis in the form of CD spectra (**Figure 53**) and subsequent BestSel analysis show differences, with two spectral groupings, CD1 (222 and M5) and CD2 (M1, M2 and M8) observed. CD spectral analysis using BestSel was however notably not the best for CD1 proteins as these proteins did not have a reasonable fit based on NRMSD. The CD1 assigned proteins (2222 and M5) were the two proteins with the highest yield and easiest to recover, so repeat experiments could be more easily undertaken, with trialling of alternate buffers and protein concentrations to yield more conclusive data.

The *in vitro* CD data show a clear difference in secondary structure between those proteins in CD1 and CD2 (**Table 27**), this was not however seen in the prior *in silico* assessment of secondary structure (**Table 22**). Such discrepancy between *in*

silico and *in vitro* analyses is also observed with the disulfides assessments. Bioinformatic predictions of protein characteristics in some cases include use of approximations (due to computational cost and efficiency). Examples being varying consideration of protein backbone flexibility, solvation effects and buffer variables such as pH [329, 399]. As the predicted model of 222 shows a largely other secondary structure flexibility is likely a very significant contributor to any characterisation *in silico* (with approximations) or *in vitro* (with varying protein conformations).

As disulfide characterisation was not a primary outcome of the project this difference in M1 and M2 observed with the *in vitro* gel assessment with and without reducing agent, was not pursued further within the work here. The differences seen may relate to other subtle differences identified in structure evident here in CD traces but require further work to elucidate fully.

In the nanoDSF work (**Figure 53**), 222 and M5 have typical well-defined sigmoidal melt curves with clear transitions, whereas M1, M2 and M8 have more atypical ill-defined non-sigmoidal melt curves. A secondary method of assessment such as isothermal chemical denaturation (ICD) could be utilised to probe this further [400]. These non-typical nanoDSF curves could be attributed to improper folding [401], concentration effects, buffer pH effects or ITF exposure to begin with. If residues are already surface exposed in the starting structures of these proteins, there would be no increased exposure as unfolding proceeded [402, 403]. Computational post hoc assessment of the solvent exposure of the trp and tyr residues in all proteins in this chapter was carried out using the same tools to assess

solvent exposure as used when selecting binding residue alanine mutation sites. These analyses showed no difference in the solvent exposure *in silico* (not shown). Or it could be that the proteins are largely unstructured as native starting structures so there would not be an obvious unfolding during nanoDSF to give a Tm. This explanation is consistent with the *in silico* predictions of structure by homology modelling (**Figure 31**) and in silico secondary structure quantification (**Table 22**). Comparison of *in silico* predictions and nanoDSF attained *in vitro* Tms showed discrepancies, the predictions were higher (more stable) for both 222 and M5. *In silico* predictions agree that there is no real difference in Tm for 222 and M5.

Looking at the entire breadth of analysis conducted within this chapter there were some distinct trends and groupings present within the proteins observed within the data presented here, summarised in (**Table 31**)**.** 222 and M5 (proteins in group CD1) had defined typical melt curve transitions in the nanoDSF analysis, grouped into S1 based on Tm. Comparing other properties of this group of proteins as outlined in **Table 31**, it was hypothesised this grouping separates out the proteins in this chapter into more soluble and less soluble groupings. Based on ease of recoverability assessed by mean yields (**Table 24**), with CD1 proteins much easier to recover than CD2 proteins. Then CD1 proteins were less stable based on Tm alone but when you examine the melt curves it was CD1 proteins that had more typical melt curves.

Table 31:Characterization results grouping and trends summary. CD spectra split into CD1(222 and M5) and CD2 (M1, M2 and M8) protein groups. NanoDSF Tm split into S1 (222, M5), none of the other mutants had curves from which Tms could be derived (M1, M2). Shows if the nanoDSF results gave a typical melt curve, split into Y for yes (222 and M5) and N for no (M1, M2 and M8). This also separated out the proteins for which a Tm could or could not be calculated. Yield taken to represent stability grouping the mg/L of culture recovery split into S1 (222, M1 and M2) then S2 (M5) and S3 (M8). A more subjective but significant result was ease of recovery split into Y for yes (222 and M5) which were easy to recover in soluble form and N for no (M1, M2 and M8) which were more difficult to recover in soluble form. *In silico* stability prediction made using the SCOOP webtool split the proteins into two groups, S1 (222, M5 and M8) and S2 (M1 and M2). SDS-PAGE assessment of disulfide configuration split the proteins into D1 (222, M5 and M8) and D2 (M1 and M2). *In silico* disulfide predictions however did not separate the proteins, all were comparable and grouped under D1 (222, M1, M2, M5 and M8). Binding assay analysis results split proteins into three groupings, with the lowest most potent Kd value for B1 (222). Then the weakest affinity B2 (M1). Then B3 (M2, M5 and M8) all had Kds that were in the hundreds of nM range, so a similar magnitude of decrease in affinity. Secondary structure analysis of protein models did not show any discrimination all grouped under SS1 (222, M1, M2, M5 and M8).

Characterisation result groupings/ trend comparison			**222**	**M1**	**M2**	**M5**	**M8**
CD, Secondary structure	Spectra comparison	*In vitro*	CD1	CD2	CD2	CD1	CD2
NanoDSF, Stability	Tm	*In vitro*	S1	/	/	S1	/
NanoDSF, Stability	Typical melt curve	*In vitro*	Y	N	N	Y	N
Yield, Stability	mg/L	*In vitro*	S1	S1	S1	S2	S3
Ease of recovery	Experience	*In vitro*	Y	N	N	Y	N
Stability prediction	Tm	*In vitro*	S1	S2	S2	S1	S1
Disulfide bond assessment	SDS-PAGE	*In vitro*	D1	D2	D2	D1	D1
Disulfide bond assessment	No. of bonds, Cys residues involved	*In silico*	D1	D1	D1	D1	D1
Binding	Kd	*In vitro*	B1	B2	B3	B3	B3
Secondary structure	Type %	*In silico*	SS1	SS1	SS1	SS1	SS1

Binding assay results furthered our understanding of the criticality of specific binding residues previously identified to be involved in the binding of 222 to TII gelatin. The work of this chapter identified that the Asn residues in positions 11, 69 and 127 were most critical to binding as when mutated to alanine their loss caused

the largest loss of binding. Secondary structure predictions did not identify any differences in proteins (all grouped under SS1)

3.4.1 Summary

The tools used for selecting mutants were successful in that the four mutants identified to take to *in vitro* testing were all stable enough to be expressed, purified, and did have an impact upon binding. There were some issues in scale up with M1, but these were overcome through further trialling and optimisation. Presented here is an expression method using sucrose as an additive to improve soluble protein recovery via osmoprotectant accumulation, resulting in more soluble protein recovery (higher yield).

The *in silico* model quantification of secondary structure type prediction made with the 2StrucCompare sever (**Table 20**),shows no difference between any of the mutant proteins generated *in silico* in this chapter. 222 and M5 were also the two proteins with the most typical (potentially more valid) nanoDSF melt curves suggesting a more defined structural transition occurs. Clearly there are some structural trends between the different proteins evident in the CD, disulfide (*in vitro* SDS-PAGE), and expression (yield and ease of recovery) amongst CBD mutants identified for the first time here. There are distinct groups evident by lack of 'nice' unfolding curves and different CD spectra) pointing to a lack of structure as key feature of these proteins. Critically, binding assays showed that Asn 11, 69 and 127 were the most important residues to binding from those tested in the work here. This result was used to guide further work in results **chapter 5** to develop a mutant with a higher affinity than 222 for TII gelatin.

3.4.2 Future work

Further work to investigate the binding of 222 to TII gelatin could include additional testing of all fifteen identified binding residues and combinations of them. This would expand the alanine mutagenesis work carried out here to confirm binding residue importance and whether residues work together to form binding surfaces. However, to evaluate all possible combinations this would be a time-consuming and costly approach. If going to the effort to produce such several mutants was considered important, further NMR work would also be justified, to show how binding is altered and so maximise the information acquired from the experiments. NMR would confirm loss of residue involvement in binding or give a measure of reduction, along with any secondary or indirect changes in binding introduced by mutations. NMR structures could reveal how the mutant molecules actually achieve their function, validating or correcting hypotheses and offering insights into how subsequent improved binding designs might be achieved [404].

Given the unstructured nature of 222 which could go a long way towards explaining a lot of the inconsistencies seen with the proteins in this project it could be hypothesised that when it binds to TII gelatin, additional structure may be introduced. Specifically, through the induced fit or preexisting equilibrium/conformational selection models of binding [405-408]. Investigation into the bound state structure would therefore be a logical characterisation in future work.

The disulfide bond configuration difference seen with M1 and M2 could be further explored. There are differences in configuration between mutants not

distinguished *in silico* so further extensive examination of disulfides undertaken in vitro would be useful to clarify if a configuration difference does exist within the mutants. Despite the strategy to try and maintain the CBD backbone structure it may be that this has inadvertently changed in some mutants as per the groupings observed. This further examination of disulfide bonds could take the form of mass-spectrometry [409], infra-red spectroscopy [410], or NMR [411] [412] [413]:

4 Results Chapter: Solubility mutants of a chimeric CBD protein

4.1 Introduction

The focus of the work in this chapter was improving solubility to remedy the issues with yield and insoluble loss seen with 222, when combined with an optimised expression and purification method.

Protein solubility is a thermodynamic parameter defined as 'the concentration of protein in a saturated solution that is in equilibrium with a solid phase, either crystalline or amorphous, under a given set of conditions' [414, 415]. Solubility can be influenced by a wide range of factors both extrinsically (buffer changes and other experimental conditions) and intrinsically (amino acid exposure and sequence). Examples of extrinsic factors that influence protein solubility include pH, ionic strength, temperature, and the presence of various solvent additives e.g., glycerol [414, 416-418]. A detailed understanding of how altering intrinsic properties of a protein can increase solubility is still in its infancy.

The overall aim of this chapter was to design mutant proteins with improved solubility and assess if binding to TII gelatin is maintained. *In silico* approaches were used initially, followed by *in vitro* expression and purification of selected targets, evaluation of solubility and finally assessment of binding in comparison to 222.

Proteins are large biopolymers, macromolecules also known as polypeptides, made up of long chains of different combinations of the 20 possible amino acids, that

consist of different side chain groups. Protein folding, and structure affects the solvent accessibility of these groups. Further, protein structures are inherently flexible and have many slightly different conformations. This array of conformations or conformers a protein can adopt is one of the reasons *in silico* predictions are complex, can be challenging, or limited and not always translatable to experimental work. Proteins are critical to the full breadth of biological systems and processes [419]. Single amino acid alterations can cause a large reduction to a proteins solubility and often result in diseases e.g. cataracts [420], cancer [421], Alzheimer's disease [422], and severe complex V deficiency [423]. Conversely single amino acid residue alterations can also cause large improvements to protein solubility [424, 425]. It is therefore rational to approach improving solubility by developing mutants. Here we utilised *in silico* tools to probe and design mutant versions of 222 predicted to have improved solubility.

The difficulty in obtaining quantitative solubility measurements in the laboratory is well documented within the literature, the main issues concern accuracy and replicability [417, 426-429]. Three common methods of measuring solubility are as follows:

1. Solvent added to lyophilised protein- variable water and salt content of the lyophilized powder can be difficult to modulate, both can have a significant effect on solubility measurements [417].
2. Concentration to saturation - difficulties in replicating results are encountered; gels, aggregates and super-saturated solutions can form [430].

3. Amorphous precipitation - is the best option but still has some issues, mainly that it gives comparative and not quantitative measures [417].

Amorphous precipitation was selected as the method here to compare 222 mutants designed for increased solubility with the wild type (WT) 222. This method was selected as it is the easiest and most adopted of the methods evident within the literature [426, 427, 431-435]. Protein precipitants can be used to obtain comparative solubility measurements allowing more promising soluble proteins to be identified early on in research to prevent wasted time and resources on suboptimal candidates. A precipitant is an extraneous agent that lowers the solubility of a protein. Precipitants can be divided into three classes: salts, organic solvents, and long-chain polymers. Here we have used two common precipitants: Polyethylene glycol (PEG-8000), a long chain polymer with an average molecular weight of 8000g/mol and ammonium sulphate ($(NH_4)_2SO_4$), a chaotropic salt. Both precipitants are used extensively by crystallographers to achieve slow precipitation [436-440].

4.2 Methods

4.2.1 *In silico* solubility analysis

The first step was exploring and identifying a suitable tool to predict solubility and aid in designing rational mutants with improved solubility. **Table 32** shows the available solubility tools/webservers [441]. CamSol was selected as a structure-based predictor of solubility.

Table 32: Available solubility tools/webservers. Summary of input type, application, reference, and website address. CamSol was the web server selected to design solubility mutants for the work in this chapter.

Tool	Input	Application	Ref	Website
DSResSol	Sequence	Solubility prediction	[442]	https://tgs.uconn.edu/dsres_sol
SoluProt	Sequence	Soluble expression in E.coli prediction	[428]	https://loschmidt.chemi.muni.cz/soluprot/
CamSol	Structure	Solubility prediction	[443]	http://www-vendruscolo.ch.cam.ac.uk/camsolmethod.html
PaRSnIP	Sequence	Solubility prediction	[444]	https://github.com/RedaRawi/PaRSnIP
SODA	Sequence	Solubility prediction	[445]	http://old.protein.bio.unipd.it/soda/
SoDoPE	Sequence	Solubility prediction and optimisation	[446]	https://tisigner.com/sodope
SOLart	Structure	Solubility prediction	[447]	http://babylone.ulb.ac.be/SOLART/
AGGRESCAN 3D	Structure	Aggregation Prediction and design	[448, 449]	http://biocomp.chem.uw.edu.pl/A3D2/
Protein-Sol	Sequence	Solubility prediction	[450]	https://protein-sol.manchester.ac.uk/
GraphSol	Sequence	Solubility prediction	[451]	https://biomed.nscc-gz.cn/apps/GraphSol
SCRATCH: SOLpro	Sequence	Solubility prediction	[452]	http://scratch.proteomics.ics.uci.edu/
CCSOL	Sequence	Solubility prediction	[453]	http://s.tartaglialab.com/update_submission/438933/3d73bcb4d0

CamSol and Aggrescan 3D servers were the only structural solubility webservers available at the time of this work. There is now also SOlart [447, 454] which would have been applicable here. The CamSol tool was chosen here because of its user-friendly interface and reported accuracy [443, 455]. CamSol was developed in 2015 by Sormanni et al within the Vendruscolo lab group. It was applied in this thesis to rationally design mutants with higher predicted solubility than 222. It was important when using the method, no changes were made to critical binding residues to ensure that binding to TII gelatin was not affected.

4.2.1.1 CamSol design tool Input

- A 'cleaned' Protein Data Bank (PDB) file - a textual file format describing the three-dimensional structures of large biological molecules such as nucleic acids and proteins, 222 in this case.
- The CamSol webserver includes the option to run a script to 'clean' the PDB file - making them amenable for computational calculations. This includes but is not limited to general file tidying i.e., removing hydrogen atoms and renumbering insertions. This was completed to prepare the PDB file of 222 prior to submitting to and running the CamSol design algorithm.
- List of residues to not be mutated - these include binding residues identified through NMR chemical shift pertubation upon binding to TII gelatin [235] known to be functionally important to 222 specified as 'unmutatable' as shown in **Table 34**. These were submitted to the tool as a space-separated list of residue numbers.
    - Total number of mutations - at time of this *in silico* work a maximum of 6 mutations was permitted by CamSol (this number has now increased but

consideration should be given to conserving WT sequence). Therefore, the CamSol Design function was executed six times, with 1-6 mutations permitted in turn.

4.2.1.2 Four steps of the CamSol method:

1. **Calculation of the residue-specific intrinsic solubility profile**

   Firstly, the method considers thermodynamics, exploiting the connection between aggregation propensity and solubility. This initial calculation of solubility is based on machine learning trained on experimental databases of heterologous protein expression. An initial score is assigned to each residue in the form of a linear combination of specific physicochemical properties **eq. (4)** [456]. The CamSol method uses secondary-structure propensities calculated from the PDB using representative structures at a 50% sequence identity and a hydrophobicity scale adapted using the Wimley–White hydrophobicity scale, as shown in **Equation [4]** [456].

$$s_i = \alpha_H p_i^H + \alpha_C p_i^C + \alpha_\alpha p_i^\alpha + \alpha_\beta P_i^\beta \qquad [4]$$

Where s_i is the initial score of each residue (i is the residue), The α values are the parameters as shown in **Table 33**. p_i^H is the Hydrophobicity of residue i, p_i^C is the charge at neutral pH of residue i, p_i^α is the α-helix propensity of residue i and P_i^β is β-strand propensity of residue i.

2. **Calculation of and structural correction to the intrinsic solubility profile is made.**

   This is the defining step of the CamSol method separating it from the other sequence only based tools available. This step considers the proximity of amino acids in the three-dimensional structure not just sequence and the resulting impact on residue solvent exposure. This enables poorly soluble residues required for folding e.g., those that make up the hydrophobic core to be distinguished from those that are exposed and might therefore commence or initiate the aggregation process. The profile is smoothed over a seven-residue window and a correction is added to consider the possible presence of hydrophobic–hydrophilic patterns and the influence of charges of the same sign **Equation [5].**

$$S_i = \frac{1}{7}\left(\sum_{j=i-3}^{i+3} s_j\right) + \alpha_{pat} I_i^{pat} + \alpha_{gk} I_i^{gk} \qquad \textbf{[5]}$$

S_i is the final score for each residue. $s_j$is the initial score from **Equation [4]** I_i^{pat} accounts for the presence of specific patterns of alternating hydrophobic and hydrophilic residues (is 1 if residue i is included in a hydrophobic pattern and 0 if not) [457, 458]. I_i^{gk} brings the gatekeeping effect [459] of individual charges into the calculation as defined in **Equation [6]**. The α values are the parameters as shown in **Table 33.**

Table 33: Parameters of the linear combination [443].

Coefficient	Value
α_H	0.598
α_C	0.318
α_α	5.77
α_β	-4.807
α_{pat}	-2.816
α_{gk}	0.152

Equation [6] brings into consideration of the relative distance of charged residues along the sequence.

$$I_i^{gk} = \sum_{j=-5}^{5} exp\left(-\frac{j^4}{200}\right) C_{i+j} \quad \textbf{[6]}$$

Where C_{i+j} is the charge of the residue $i+j$.

3. **Identification of suitable mutation sites using the profile from step 2.**

   Substitutions and/or insertion mutations into the poorly soluble but solvent exposed sites are made. Mutation sites predicted to have the largest impact on overall protein solubility, are assessed by solubility score output from step 2. To be a suitable mutation site three criteria must be met; within or close to poorly soluble regions, solvent exposed and far from activity critical residues (specified by the user **Table 34**).

Table 34: Residues specified as 'unmutatable' involved in binding. Activity critical residues (specified by the user in step 3 of CamSol method)

Residue	Residue Number		
	1st module	2nd module	3rd module
G	10	68	126
N	11	69	127
R	22	80	138
F	23	81	139
G	25	83	141
Y	28	86	144
G	35	93	151
G	39	97	155
Y	40	98	156
W	42	100	158
T	46	104	162
E	47	105	163
Y	49	107	165
Y	55	113	171
G	56	114	172

4. **Screening of all possible variants to identify the most soluble, using intrinsic solubility score.**

   All possible mutations are screened systematically, again using the intrinsic solubility score to identify the most soluble variant/variants.

4.2.2 Expression

The codon optimised (**methods section 3.2.10.1**) mutant CS6 gene ordered for the work in this chapter is shown in **Table 35**.

Table 35: GeneOptimizer codon optimised DNA sequence for CS6. Nucleotide codons highlighted in yellow are those that encode the CS6 mutations.

Name	DNA Sequence
CS6	GAAGGTCAAGAGGTTGAAACCATGGAAGGTAATGCCGAAGGTCAGCCGTGTAAATTTCCGTTTCGTTTTCAGGGCACCAGCTATGATAGTTGT ACCACCGAAGGTCGTACCGATGGTTATCGTTGGTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCA CTGTTTACCATGGGTGGCAATGCAGAAGGCCAACCTTGCAAATTCCCTTTTCGCTTCCAGGGTACATCTTATGATTCATGCACAACGGAAGGTC GCACAGATGGCGAAGAGGAATATCGCTGGTGCGGCACCACAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACAGCCCTG TTCACAATGGGTGGTAATGCGGAGGGACAGCCATGCAAGTTTCCATTCCGCTTTCAGGGAACCTCATATGATAGCTGCACAACAGAGGGACGT ACGGATGGCTACCGGTGGTGCGGAACTACGGAAGATTACGACCGGGACAAGAAGTATGGTTTCTGCCCAGATCAGGGTTATAGCCTG

Expression of CS6 proceeded mostly as previously outlined in **methods section** 3.2.10 but with a different optimisation strategy (outlined in **Figure 56**). The gene for CS6 was cloned into three pOPIN vectors; pOPINS, pOPINM and pOPINJ. This was achieved using the ligase independent In-Fusion cloning methodology described in full detail within **methods section** 3.2.10.6. Stellar competent cells (Takara Bio) were transformed via heat shock with the CS6 construct plasmids (**methods section** 3.2.10.7). Stellar cells were used as a cloning strain to produce plasmid for isolation, via QIAprep miniprep kit (Qiagen),(**Materials and methods section** 2.10). Before proceeding to expression trials all constructs were confirmed via sanger sequencing (Eurofins GATC, Germany).

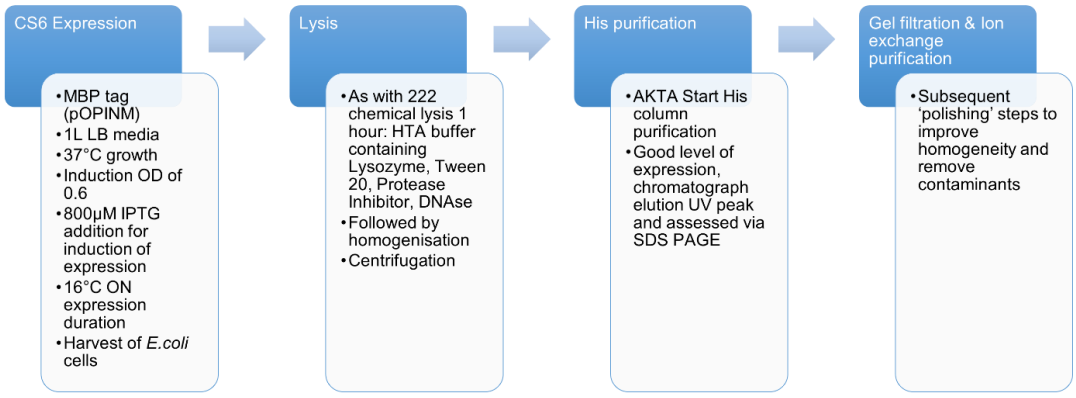


Figure 56: Overview of expression and purification conditions for CS6. Expression conditions utilised with 1L of LB. Lysis proceeded via homogenisation using the continuous Flow CF1Cell Disrupter (Constant systems) as described in **methods section** 3.2.10.14. His purification proceeded to the end of the initial His column load and elution (described in **methods section** 3.2.10.15) where SDS-PAGE analysis (described in **methods section** 2.6) of the eluted fractions was made. It was at this time a good level of soluble POI was identified. Some contaminants were present so further polishing in the form of gel filtration and ion exchange chromatography was used to maximise recovery of pure POI.

4.2.2.1 Primer design

A forward (fwd) and reverse (rev) primer (**Table 36**) were designed, ordered from sigma, and used to subclone each gene into the three pOPIN vectors via Infusion cloning (Takara Bio) as described in **methods section 3.2.10.6 .**

Table 36: Infusion Cloning primers used with CS6 gene. pOPIN vector region (black text) and insert region (red text).

Mutant	Vector	Forward (fwd) Primer	Reverse (rev) Primer
CS6	pOPINS	GCGAACAGATCGGTGGTGAAGGTCAAGAGGTTGAAAC	ATGGTCTAGAAAGCTTTACAGGCTATAACCCTGATCTG
	pOPINJ	AAGTTCTGTTTCAGGGCCCGGAAGGTCAAGAGGTTGAAAC	
	pOPINM		

4.2.2.2 Vector linearization

The vector map for the pOPINS vector used previously with the alanine mutants is shown in **Figure 22**. The vector maps for pOPINJ and POPINM used for the first time in the work of this chapter is shown in **Figure 57**.

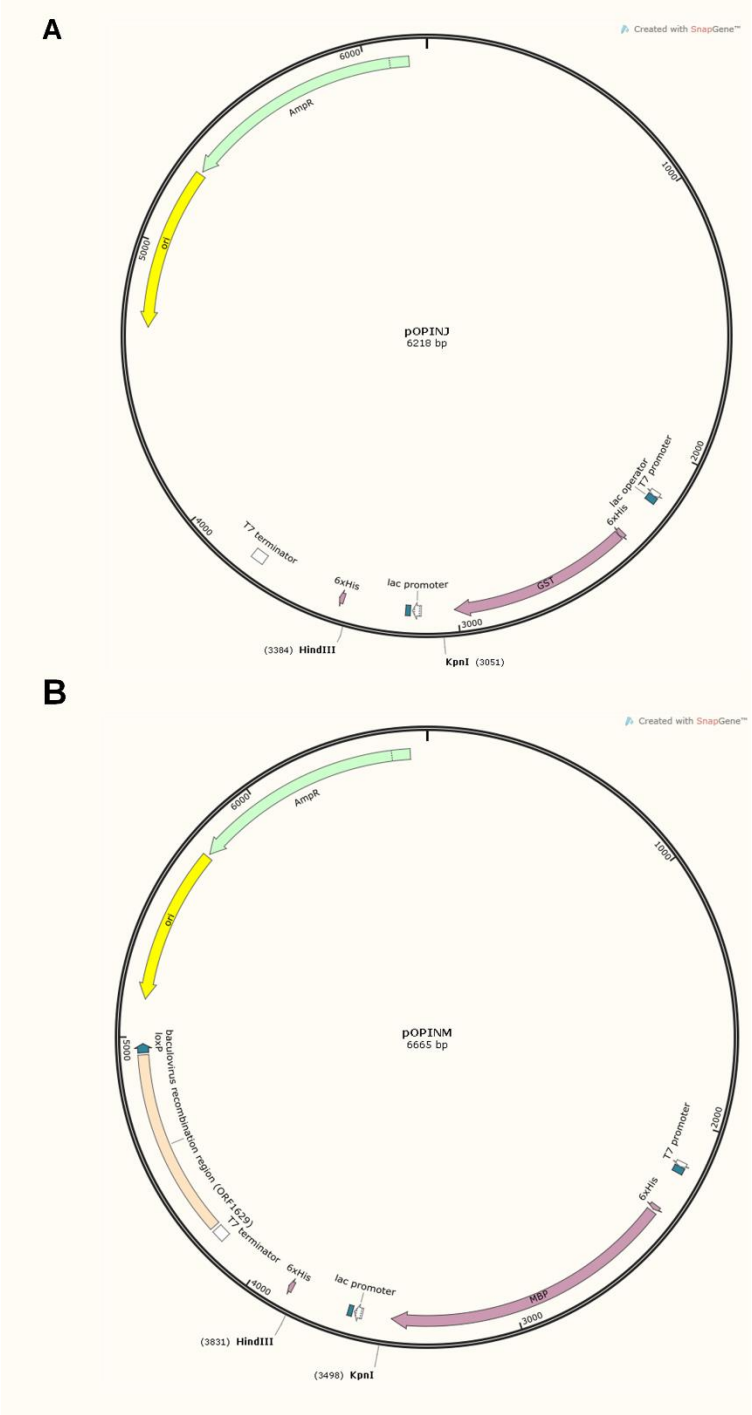


Figure 57: pOPINJ and pOPINM vectors used in the work in this chapter. **(A)** pOPINJ 6218bp, ampicillin resistance, T7-Lac promoter [8] **(B)** pOPINM 6665bp, ampicillin resistance, T7-Lac promoter [8]. Both were Linearised by KpnI recognition sequence GGTACC and HindII recognition sequence AAGCTT. Digest products were a 333bp cut out (lac z gene) and 5885bp **(A)** and 6332bp **(B)** linearised vector. The parent vector of both pOPINJ and pOPINM is pTriEx-2. This figure was created using the SnapGene software (from Insightful Science; available at www.snapgene.com).

The OPPF-UK pOPIN vector suite [353] was utilised more extensively in this chapter allowing trialling of three different fusion tags [460, 461]; SUMO, Glutathione S-transferase (GST) and Maltose binding protein (MBP). It is common to compare different fusion protein partners to determine which provides best expression and solubility for a given target. For unknown reasons, no single fusion tag is universally successful, and it is as yet impossible to predict which will perform optimally for any given target protein so trialling multiple is good practice [462]. All three tags tested here GST, MBP and SUMO can enhance solubility by increasing yield and enhancing folding [463]. All three were available here so trialled in parallel. Tandem trialling maximised chances of success and more swiftly enabled sufficient recovery of proteins for the experiments in this chapter. The additional his tag simplified purification to allow use of IMAC with only a nickel affinity column. The cleavage site for all these tags was positioned following the fusion tag, resulting in POI only after tag cleavage. **Table 12** in the previous chapter gives further details of the pOPINS vector and **Table 37** here gives further details of the pOPINJ and pOPINM vectors introduced in this chapter.

Table 37: pOPINJ and pOPINM vector details. Including source, restriction enzymes and recognition sites utilised, parent vector/ antibiotic resistance, digest products, promoter, inducer, expression product, cleavage enzyme and references.

Vector	Source	Restriction enzyme 1 and recognition site	Restriction enzyme 2 and recognition site	Parent vector/ Antibiotic resistance	Digest products	Promoter	Inducer	Expression product	Cleavage enzyme	Ref
pOPINJ	OPPF	HindIII AAGCTT	KpnI GGTACC	pTriEx2/Amp	333bp cut out, 5885bp linear vector	T7	IPTG	His-GST-POI	3C protease	[353]
pOPINM	OPPF	HindIII AAGCTT	KpnI GGTACC	pTriEx-2/Amp	333bp cut out, 6332bp linear vector	T7	IPTG	His-MBP-POI	3C protease	[353]

Vector linearisation was carried out using two restriction endonucleases, in a double digest removing an unrequired 333bp segment and linearizing the vectors in the process (**Figure 23** and **Figure 58**).

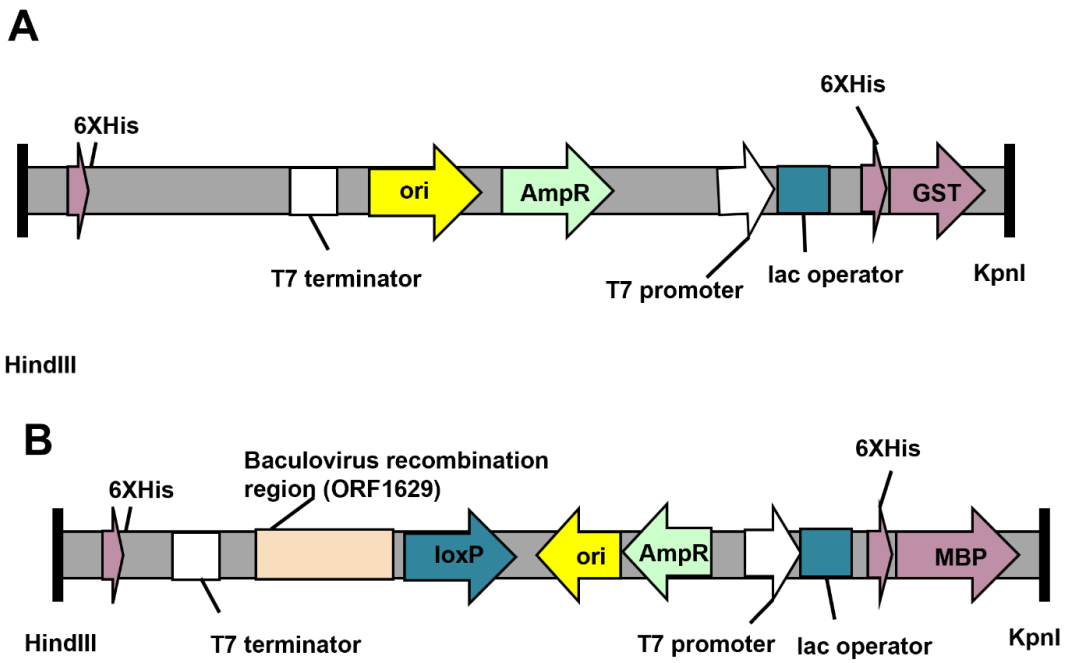


Figure 58: Linearised pOPINJ (A) and pOPINM (B) vector. 333bp fragment removed during linearisation 5885bp (A) and 6332bp (B) linear fragment. Amp Resistance, required for transformed clone screening. Lac Promoter and terminator required for expression of insert/ POI; expression inducible by IPTG.

4.2.2.3 Small scale expression & purification

Shuffle cells were again used here as an expression strain as they are optimised for disulfide bond generation [302]. There are six disulfide bonds known to be present in CBD, 222 [235], and predicted to also be present within CS6 (**Results section 4.3.3.2.1**). **Table 38** shows a summary workflow, results and progression through different conditions that were trialled at small scale for CS6**.**

Table 38: CS6 small scale mutant expression trials. Workflow order, conditions, and SDS-PAGE soluble or insoluble fraction results (See methods section **3.2.10.9** and **4.2.2.4** for further details of trials and conditions).

Trial Number	Trial Condition	Result
1	SUMO tag	Insoluble
	MBP tag	Soluble
	GST tag	Soluble
2	Chaperone co expression with SUMO tag	Insoluble

Expression was undertaken initially for all three pOPIN vector as previously outlined in **Methods section 3.2.10.9** and **3.2.10.10**. Purification was undertaken as previously outlined in **Methods section 3.2.10.10**.

4.2.2.4 Chaperone co-expression

Molecular chaperones can assist protein folding and, in some cases, this leads to increased production of soluble proteins [362, 464, 465]**.** A chaperone plasmid set (Takara bio) was used here with pOPINS to try and improve soluble protein recovery. All chaperone plasmids (**Table 39**) carried the chloramphenicol (Cm) resistance, meaning transformation for co-expression required dual antibiotic plates of Cm and Kan or Amp (depending on the mutant construct pOPIN S, J or M). Transformation was conducted as per **methods section 3.2.10.7**, but with 1.5µl of both plasmids added and dual antibiotic agar plates (Kan or Amp and Cm). Starter cultures were created as per **methods section 3.2.10.8** but with dual antibiotic to maintain selection for both plasmids.

Expression was conducted in small scale in 50mL falcons containing 10mL LB 10µl of Cm and 10µl of Kan. Cultures were grown at 37°C, 180RPM, then chaperone expression was induced with 2mg/mL L-arabinose and/or 5ng/mL tetracycline (tet) once OD reached 0.4, to give the bacteria time to accumulate chaperone before mutant expression was induced. CS6 mutant cloned into the pOPINS vector was then induced at an OD of 0.6, with 400µM IPTG (middle concentration of inducer used from those used in **methods section** **3.2.10.9**, to give a first indication of chaperone improving solubility, if so further optimisation of inducer concentration was planned to be undertaken). Expression temperature was reduced to 16°C O/N or 30°C for 3h. At the end of this expression period, cells were harvested via centrifugation at 8000RPM for 5mins (Allegra X-30R Centrifuge, Beckman Coulter). Supernatant was discarded and pellets stored frozen at -80◦C until purification (**Methods section** **3.2.10.10**).

Table 39: Takara Bio chaperone plasmid kit. Details of the five plasmids, chaperones, size (GrpE usually runs at 29kDa on a gel), promoter and inducer.

No.	Plasmid	Chaperone	Size (kDa)	Promoter	Inducer
1	pG-KJE8	Dnak-DnaJ-GrpE GroES-GroEL	70-40-22 10-60	araB Pzt1	L-arabinose Tetracycline
2	PKJE7	Dnak-DnaJ-GrpE	70-40-22	araB	L-arabinose
3	PTF16	tig	56	araB	L-arabinose
4	PG-TF2	GroES-GroEL-tig	10-60-56	Pzt1	Tetracycline
5	PGRO7	GroES-GroEL	10-60	araB	L-arabinose

4.2.2.5 Large scale expression & purification

Large scale expression was undertaken as previously outlined in **chapter 3, Methods section 3.2.10.12**. For CS6 repeating the small-scale optimised culture conditions with large scale culture did produce the same good level of soluble expression.

4.2.2.6 Lysis & His Purification

Lysis proceeded again as described previously in **methods section 3.2.10.14**

CS6 expressed using the MBP tag was then purified using a 5mL nickel column (Cytiva) as a first step, using His affinity chromatography on the ÄKTA start purification system (Cytiva). The His purification (described in **methods section 3.2.10.15**) was then followed by 3C protease cleavage to remove the His-MBP tag from the POI. POI without tag was then isolated via a subsequent reverse His purification step (outlined in **methods 3.2.10.15**).

4.2.2.7 'Gel filtration, Desalting & Ion exchange

Gel filtration was undertaken following His purification as described in **methods section 3.2.10.16.** Fractions from GF were assessed for purity using SDS-PAGE, pure fractions pooled, concentrated, and retained for further planned experiments. Any homogeneous protein content fractions were also retained following GF, pooled then subjected to Ion exchange (IE) purification, utilised as a final polishing step.

Before IE could proceed protein was buffer exchanged in a desalting step using a HiPrep 26/10 column ((Cytiva), equilibrated in low salt buffer IEC A (20mM Tris, 10mM NaCl, pH 8.0), again undertaken using the Äkta start system and a flow rate of 4mL/min. IE relies on the charge of the protein. POI was applied to a QFF IE column (Cytiva) in IEC A buffer, at a flow rate of 2mL/min. At pH 8.0 POI was negatively charged meaning it bound to the positively charged resin in the IE column. To elute a high salt buffer IEC B (20mM Tris, 1M NaCl, pH 8.0) was applied to the column at a flow rate of 2mL/min. Cl- ions have a higher affinity for the column so dissociated the POI causing its elution. UV was used as a means of identifying fractions containing POI. Fractions from IE were then assessed for purity using SDS-PAGE, pure fractions pooled, concentrated, and retained for further planned experiments.

Large scale production of CS6 was then repeated to provide enough protein for planned solubility and binding assays and to allow characterisation in comparison to 222 and CBD.

4.2.3 Protein Characterisation

Characterisation of CBD, 222 and CS6 was undertaken as in **methods section 3.2.11.7**. Characterisation included liquid chromatography LC-MS, CD, NanoDSF and assessment of disulfide bonds both *in silico* and using SDS-PAGE. The LC-MS and nanoDSF experiments were outsourced to external facilities. LC- MS was used to confirm CS6 protein was target, samples of CS6 and 222 were sent to the CPR, Liverpool, see **methods section 3.2.11.1**. CBD had been confirmed as correct in prior work of the group [235]. Disulfide bonds were assessed (as described in

methods section 3.2.11.2) to determine that mutations in CS6 had not altered the total number or configuration of disulfide bonds known to be six in 222 [235] and CBD [466]. 'Is binding maintained with enhanced solubility?', this is a question with an answer key to the OA therapeutic strategy of this thesis. Given that the primary aim of the therapeutic strategy is adherence to TII gelatin in the OA joint it was important to assess binding of CS6 using the plate assay utilised in **methods section 3.2.11.7**.

4.2.3.1 Solubility assay

An amorphous precipitation method was utilised here within a 96 well plate format adapted from Li et al, 2013 [467], to allow measures with small volumes and protein concentrations, reducing the total amount of protein required.

Samples of proteins and precipitant were loaded across two replicate 96 well plates (Grenier, clear polystyrene flat bottom). A 100µl well volume was used in triplicate as shown in **Figure 59**; made up of precipitant ,1x PBS buffer and protein in 1x PBS, pH 7.4 at a diluted final well concentration of 0.8mg/mL. The precipitant concentration increased from row A to D (down the plate). The plate was mixed for 10secs in the plate reader (Flexstation 3 microplate reader) then incubated for 24h at RT (22°C). After 24h the plate was mixed again for 10secs and absorbance measurements at 500nm taken using a Flexstation 3 microplate reader to detect turbidity resulting from precipitation**.**

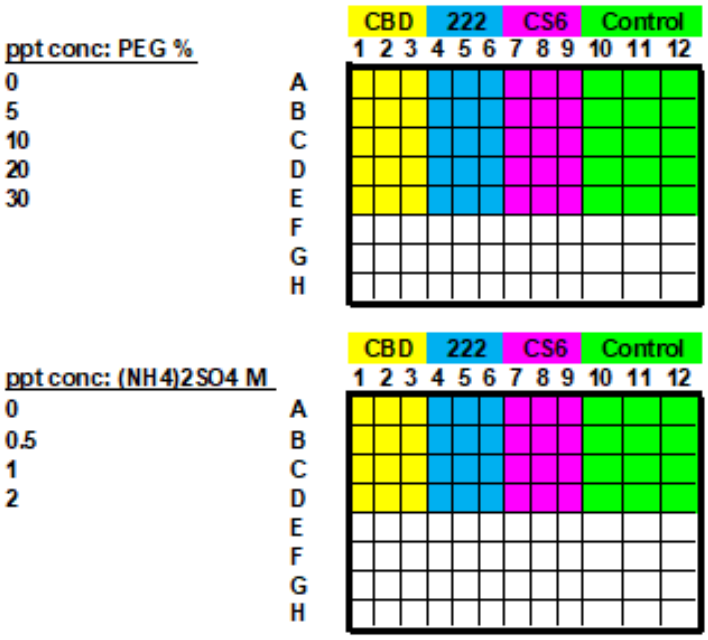


Figure 59: Solubility assay 96 well plate layout for each precipitant (ppt) and all the proteins in this chapter. CBD, 222 and CS6 were added to each of the wells to a final working well concentration of 0.8mg/mL**.** Control wells contained 1X PBS buffer, pH 7.4 only. All proteins were run in triplicate in each plate (n=3) then each plate was run in duplicate (n=2).

Protein concentration was also measured for each condition. One plate of each precipitant type retained from the A500 reads were centrifuged for 20mins at 5000G to pellet precipitant, using an Allegra X-30R Centrifuge with a swinging-bucket microplate rotor (Beckman Coulter). Protein concentration remaining in the supernatant was then determined spectrophotometrically using a Nanodrop 2000 (Thermo Scientific). This data was normalised to % of starting concentration to show loss of protein due to induced precipitation.

Additionally, this data was analysed without normalisation. With this plot a log linear relationship is expected [417, 467]. This plot allowed the calculation of apparent solubility using linear interpolation results and **Equation [7], Equation [8] and Equation [9].** Values $\log S_0$ and β were taken from these interpolations.

Apparent solubility calculation

The relationship between precipitant concentration to protein solubility is described by the following general expressions in **Equation [7]**, **Equation [8]** and **Equation [9]**

$$\log S = constant - \beta[Precipitant] \quad \textbf{[7]}$$

Where **S** is solubility (mg/mL), **constant** is the y intercept of the solubility plot (at 0M $(NH_4)_2SO_4$ or 0% PEG) and **β** is the slope of the line.

$$\log S = \log S_0 - \beta[Precipitant] \quad \textbf{[8]}$$

Where **S** is solubility (mg/mL), **β** is the slope of the Line and S_0 is the solubility in the absence of precipitant, which is the Y intercept.

$$S = 10(\log S_0 - \beta) \quad \textbf{[9]}$$

4.3 Results

4.3.1 *In silico* design of solubility mutants

Table 40 shows the mutant results from using the CamSol design algorithm. Output gave the suggested mutations and a CamSol score predicting the impact on solubility. **Figure 60** shows the CamSol scores allowing the ranking/comparison of proteins in terms of predicted solubility**.** The CamSol tool and method were also utilised here to predict how CBD the native non-mutant (start point) protein compared in terms of solubility, using this same CamSol score value (**Table 40** & **Figure 60**). Increasing CamSol score shows increasing solubility and represents a relative rather than an absolute prediction value. CBD was the least soluble protein

when assessed in this way. There was an increase in solubility with number of mutations permitted, CS6 was predicted the most soluble protein and was therefore assessed further.

Table 40: CamSol prediction results. CamSol prediction of mutations that can be made to 222 to enhance solubility. Enhanced solubility is shown by a higher CamSol score. Includes prediction of the CamSol score of CBD and 222 for comparative purposes. The first letter of the mutation refers to the protein chain, there is only one chain (A) in these proteins. This is then followed by the residue and corresponding number to be substituted and to which residue. Y9E means the tyrosine (Y) at residue number 9 be substituted for a glutamic acid (E). With CamSol 3- 6 there is also an insertion (ins) G97_Y98ins E or EEE. This means insertion of a single or triple glutamic acid between residues glycine (G) 97 and Y98.

Protein Name	Abbreviation	Mutations	CamSol score
Collagen Binding Domain	CBD	0	-0.446019
222	222	0	-0.295891
CamSol 1	CS1	A.Y9E	-0.235297
CamSol 2	CS2	A.F6E;Y9E	-0.179515
CamSol 3	CS3	A.F6E;Y9E;G97_Y98insE	-0.143773
CamSol 4	CS4	A.V4E;G97_Y98insEEE	-0.137999
CamSol 5	CS5	A.V4E;F6E;G97_Y98insEEE	-0.080165
CamSol 6	CS6	A.V4E;F6E;Y9E;G97_Y98insEEE	-0.028071

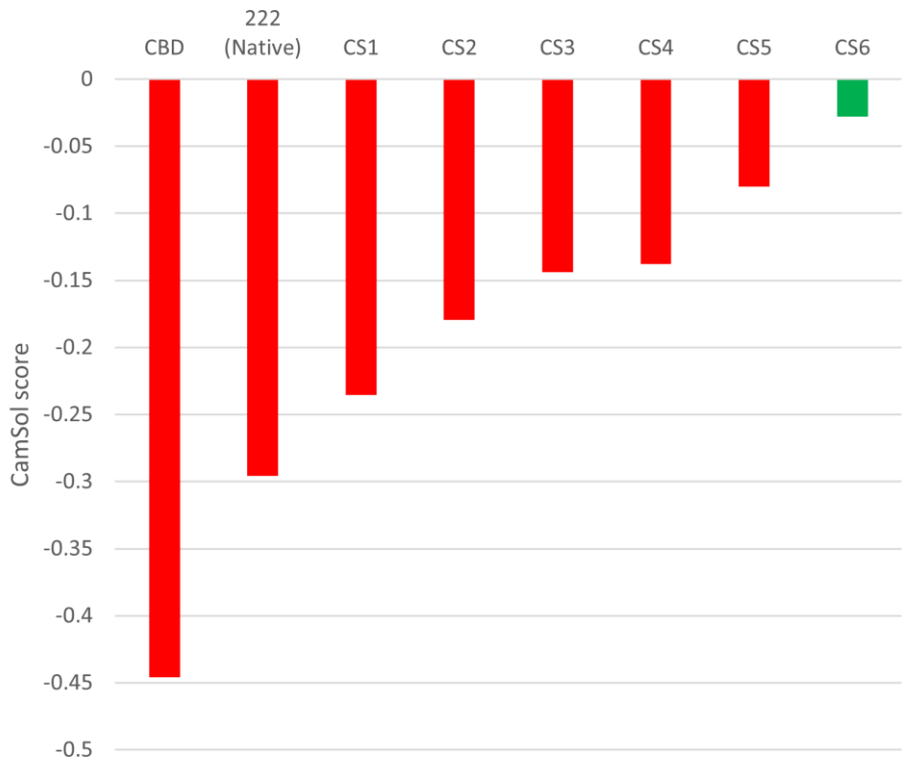


Figure 60: CamSol Scores. A higher score indicated a predicted higher solubility. CBD is the predicted least soluble protein and CS6 is predicted to be the most soluble.

CS6 has the largest change from 222 with three hydrophobic to hydrophilic substitutions V4 to E;F6 to E;Y9 to E an insertion of EEE (three additional hydrophilic residues) between G97 and Y98. **Figure 61** shows the pymol representations of CBD, 222 and CS6 highlighting the positions of mutations in CS6 compared to 222.

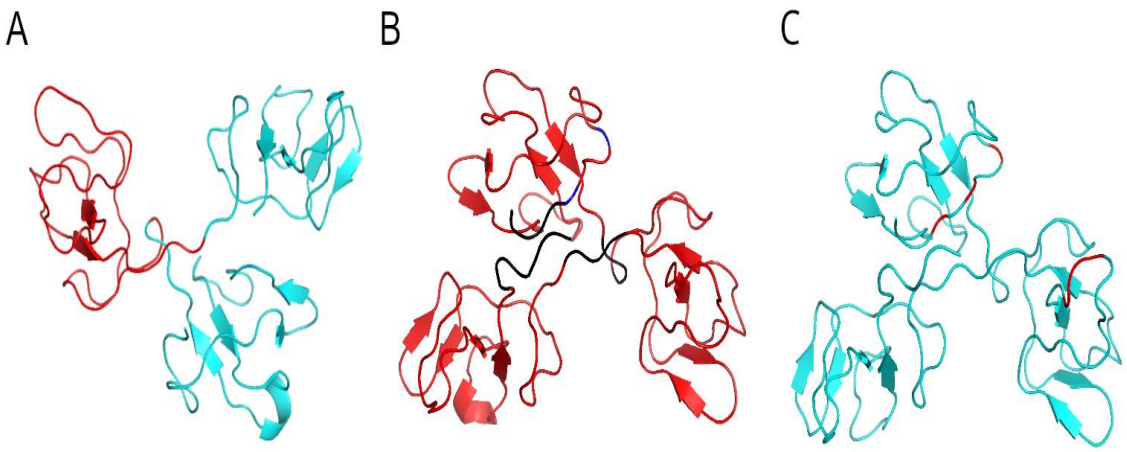


Figure 61: Structures of the three proteins in this chapter, CBD (A) PDB accession code 1CK7, 222 modelled in chapter 3 (B) and CS6 (C) modelled using https://swissmodel.expasy.org/interactive using the sequence shown below in Table 42. In CBD (**A**) module 2 is highlighted in red, in 222 (**B**) the three module 2s are shown in red, residues in blue are conserved from the original CBD to maintain intramolecular interactions and residues in black are linkers, then in CS6 **(C)** Mutations are shown in red.

Table 41 shows a comparison of the percentage of each secondary structure type across the three proteins analysed in this chapter.

Table 41: Percentage of each secondary structure type in CBD, 222 and the solubility mutant CS6. This was determined using the 2StrucCompare web server [344]. There is minimal difference between the three proteins in terms of secondary structure.

Secondary Structure	CBD	222	CS6
Helix	7%	2%	8%
Sheet	20%	18%	20%
Other	73%	80%	72%
Undefined	0%	0%	0%

CS6 was predicted as the most improved in solubility with the highest CamSol score and so was taken forward to cloning and expression. **Table 42** shows the protein sequence and key parameters for CS6.

Table 42: Protein sequence for CS6, with key physical and chemical parameters. Name of mutant, mutation sites stating the native residue and its position number in 222 followed by the mutant residue. V4E indicated that the Val (V) at position 4 in 222 be mutated to a Glu (E) residue. Ins indicated a residue insertion between two residue numbers. G97_98insEEE indicated that three Glu (E) residues be inserted after the Gly (G) at position 97 in 222. Next the entire mutant protein sequence is shown then key confirmatory mass spec (MS) fragments are shown. The residues shown in blue text were left as in the native CBD protein, as they were deemed had critical involvement in intramolecular interactions. The three modules for each protein are shown in red text, and the linker regions in black text. mutations are highlighted in yellow. Abs 0.1%, extinction coefficient and molecular weight listed, determined using the Expasy ProtParam tool available at https://web.expasy.org/protparam/ (accessed 08/21).

Name	Mutation site	Protein Sequence	Key confirmatory MS fragments	Extinction coefficient	Abs 0.1% (=1g/l)	Molecular weight (kDa)
CS6	V4E F6E Y9E G97_98insEEE	EGQEVETMEGNAEGQPCKFPF RFQGTSYDSCTTEGRTDGYRW CGTTEDYDRDKKYGFCPHEALF TMGGNAEGQPCKFPFRFQGTS YDSCTTEGRTDGEEEYRWCGT TEDYDRDKKYGFCPETALFTMG GNAEGQPCKFPFRFQGTSYDS CTTEGRTDGYRWCGTTEDYDR DKKYGFCPDQGYSL	EGQEVETMEGNAEGQPCK TDGEEEYR	36620	1.743	21.02

4.3.2 Expression & purification

SUMO tagged POI was not seen in the soluble product in any of the small-scale expression trials. Chaperone co-expression was subsequently trialled but failed to shift POI into the soluble product. pOPINJ and pOPINM small scale expression trials both had soluble expression The pOPINM (MBP tag fusion vector) however gave the best soluble expression for CS6,

Figure 62 shows the small-scale expression results for MBP-tagged CS6, with bands at the expected molecular weight of 63kDa. Target POI was present in the

soluble fraction in a higher ratio with these conditions than others, therefore 800µM IPTG induction and overnight expression at 16°C was progressed following this result into larger scale expression.

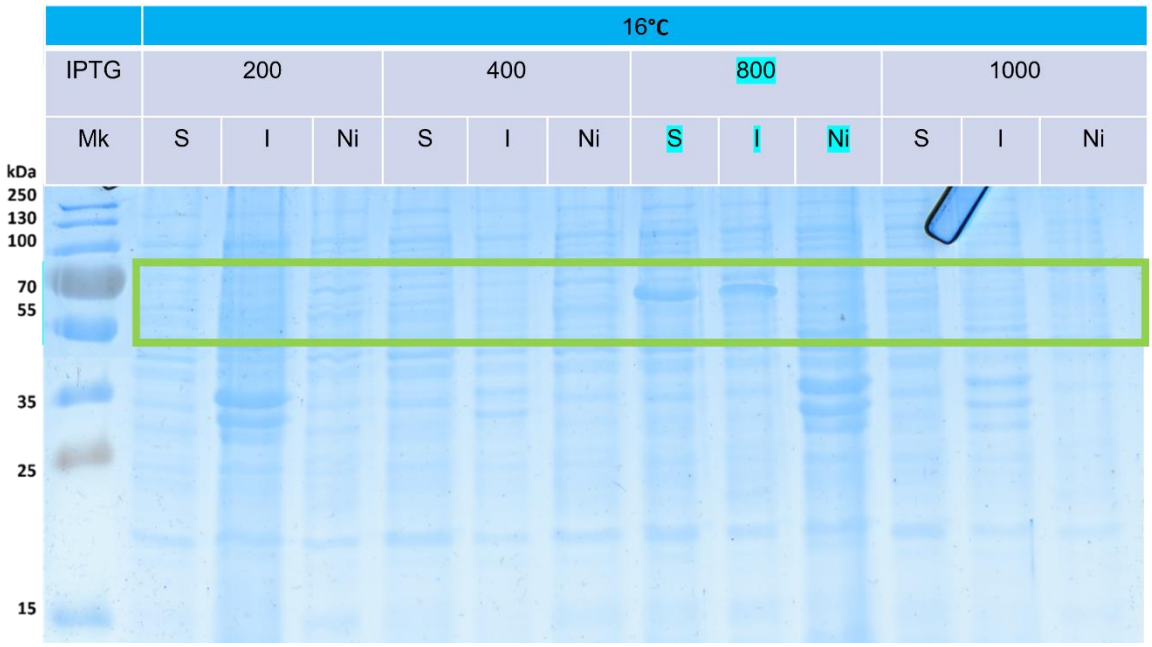


Figure 62:SDS-PAGE results of small scale MBP tagged CS6 expression trials. Expression at 16°C ON. Soluble (S) supernatant, insoluble (I) pellet, nickel resin purified (Ni) product. Looking for soluble expression with 63kDa size. 800µM IPTG was identified to be the only IPTG concentration where a band of this size was present in the S sample, highlighted in light blue. There were also target sized bands in the I and Ni samples.

The results of a subsequent 8L large scale expression beginning with lysis product (sup and pel), His purification (ft and his fractions) is shown in the light blue box in **Figure 63**. Protein was lost in the initial His column loading as shown by the presence of a His-MBP-POI sized band (65.05kDa) in the flow through (FT) sample, this indicates that not all POI bound to the column. This was followed by 3C protease cleavage (PDS) and a reverse His purification step (RFT, ER and CRFT).

Following cleavage there is a shift in band seen in **Figure 63** from the light blue box where POI had tag to the orange and yellow boxes where it is POI only with tag

cleaved. The green box shows tag only (44.03kDa). The CRFT lane shows cleaved, uncleaved and tag indicating further purification was required. The three proteins that dominate this CRFT lane are sufficiently different in size that gel filtration was the most logical subsequent step.

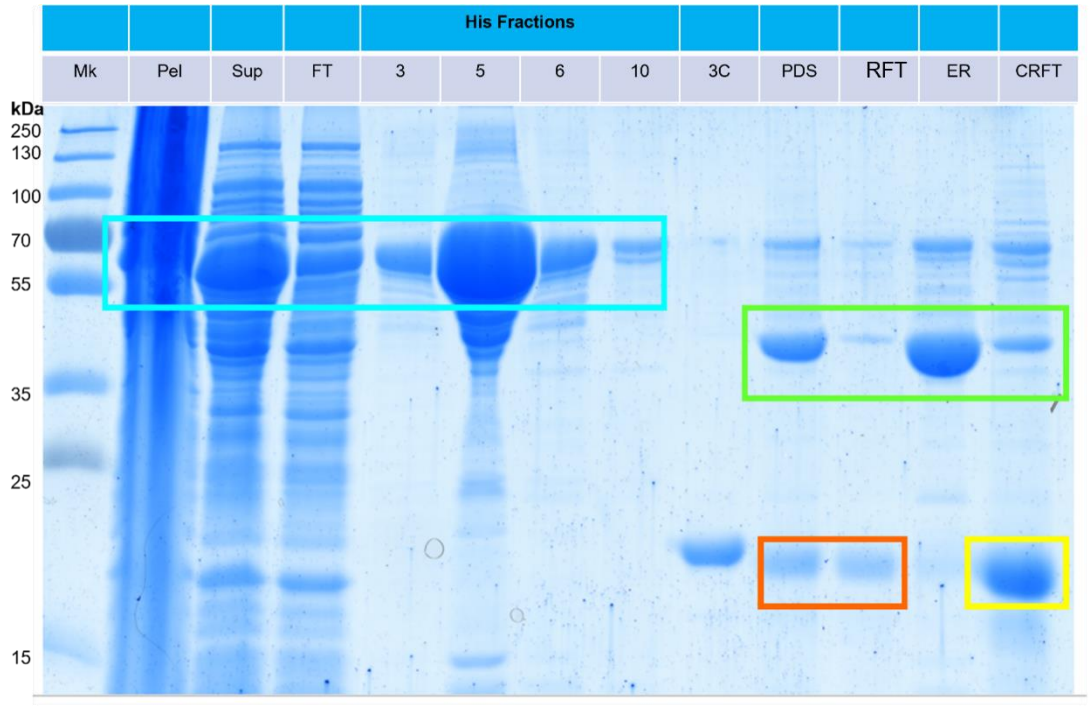


Figure 63: CS6 large scale expression product. An SDS-PAGE analysis showing MBP tagged POI Induced with 800µM IPTG expressed overnight at 16°C. Lysis product separated out by centrifugation at 18000RPM for 40minutes, at 4°C product of this shown under pellet (Pel) and supernatant (Sup). Supernatant was filtered through a 0.45µm filter then loaded onto a His column for purification, product from this shown in FT (protein that did not bind to column), and His protein fractions eluted using a 500mM (high imidazole) step elution. 3C cleavage is shown in post dialysis soluble (PDS) and reverse flow through (RFT) by the appearance of a lower band, reverse His purification, concentrated RFT (CRFT). The light blue box shows where POI that has His-MBP tag, the orange and yellow boxes show where it is POI only with tag cleaved. The green box shows tag only (44.03kDa). The CRFT lane shows cleaved, uncleaved and tag indicating further purification was required. Cleavage was achieved using 3C protease (22kDa) which can be seen as a control alone in the unboxed 3C lane. Pre-cleavage CS6 (with His-MBP tag) 65.05kDA, then CS6 size 21.02kDa.

Following His purification, a gel filtration purification step (Cytiva HiLoad™ 26/600 Superdex column) was used to clean up the still heterogeneous reverse flow through pooled and concentrated product. **Figure 64** shows the SDS-PAGE analysis of the content of the retained gel filtration fractions. Fractions containing protein with the correct molecular weight to be CS6 fractions 10-12, and 20-24 (highlighted in light blue) were pooled and concentrated (Amicon 3000 MWCO concentrators).

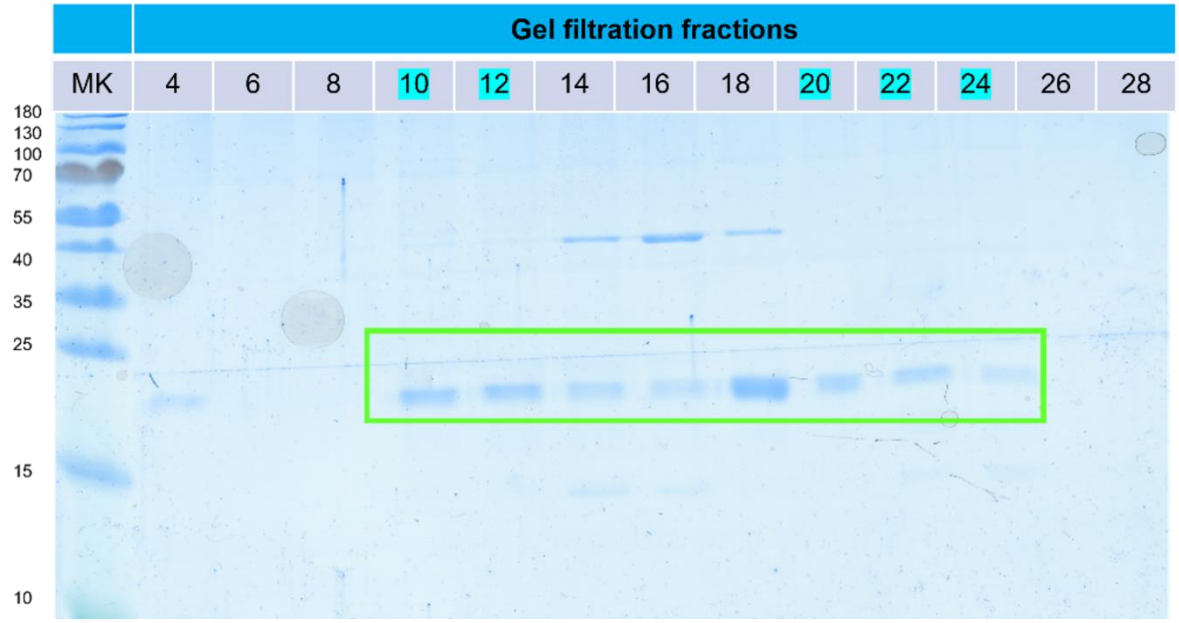


Figure 64: CS6 gel filtration SDS-PAGE results. SDS- PAGE showing the contents of gel filtration fractions for CS6. Fractions highlighted in light blue 10-12 and 20- 24 were identified as containing only target CS6 protein so were pooled and concentrated. Fractions 14-19 were pooled and taken through a subsequent anion exchange purification step.

An additional ion exchange step (Cytiva QFF column) was used to further purify pooled gel filtration fractions that still showed heterogeneity on SDS-PAGE (fractions with target sized protein shown in the green box but other contaminant proteins of a different size outside the box, fractions 14, 16, 18 and 24 in **Figure 65**. This additional step was utilised to maximise recovery of protein for subsequent analysis.

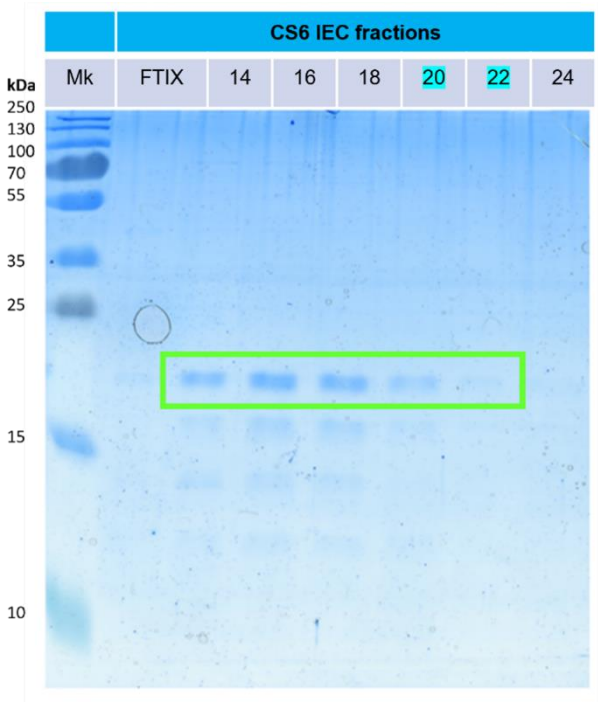


Figure 65: SDS-PAGE results of anion exchange purification of CS6 fractions. Highlighted in green 20-22 were identified as containing only CS6 (homogeneous content implied by single band) so fractions in this range were pooled and concentrated.

Fractions showing the expected molecular weight for POIs from both gel filtration and ion exchange purification were pooled and concentrated (Amicon 3000 MWCO concentrators), aliquoted, flash frozen and stored at -80°C. **Figure 66** shows the end CS6 product retained and taken forward to characterisation experiments. A single intense target POI sized (21.02kDa) band was apparent after concentration.

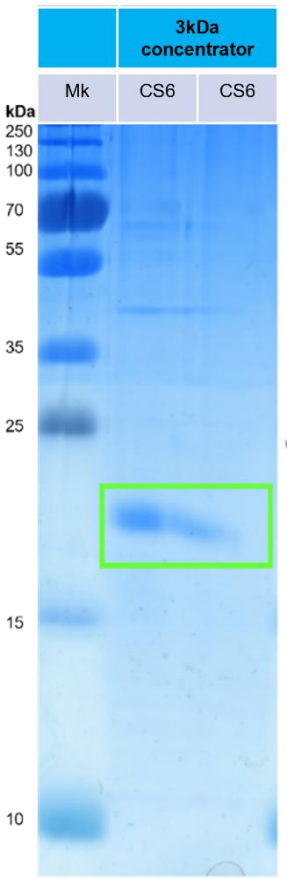


Figure 66: CS6 purification product. Purified, pooled (gel filtration and ion exchange purification product), then concentrated with a 3kDa concentrator CS6. It was this protein shown in the gel here, final purification product that was sent for LC-MS confirmation (CS6 21.02kDa).

4.3.3 Protein Characterisation

Characterisation of CBD, 222 and CS6 was the same as in previous chapters.

4.3.3.1 Mass Spectrometry (MS)

MS was used to confirm retained CS6 protein was in fact target, with fragments containing the correct mutations present when compared to 222 (**Table 42**). CBD had been confirmed as correct in the prior work of the group [235].

4.3.3.2 Disulfide bonds

Disulfide bonds were assessed to probe if the mutations in CS6 had altered the total number or configuration of disulfide bonds, known to be six in both 222 and CBD.

4.3.3.2.1 *In silico* disulfide analysis

Two different tools were used to assess *in silico* disulfide bonds predicted in CBD, 222 and CS6 as used previously for binding mutants (**methods section 3.2.11.5**). *In silico* analysis of disulfide bonds in CBD, 222 and CS6 showed that all three proteins are predicted to have the same total number of six from a possible nine disulfide bonds (**Table 43**). However, different residues are predicted to be involved in the three different proteins. Both tools agree on the six bonds in each protein. With PredDisulfideBond a probability is given for each disulfide bond forming, the higher the value the more likely a bond is to form. With Maestroweb a Maestro bond score is utilised, the lower the score the more likely bonds are to form. Bonds shown in black in **Table 43** are the most probable to form, in red are the bonds that are predicted won't form due to the residues being more likely to form another disulfide bond first.

Table 43: Shows the results of the PredDisulfideBond[386] and Maestro[348] webserver predictions for the three proteins in this chapter; CBD, 222 and CS6. In the first column for each protein are the bond predictions, detailing the residue (Cys), chain (A), residue number, followed by the same for the other residue involved in the possible disulfide bond. Disulfide bonds form between the sulfhydryl (SH) side chains of two cysteine residues. The probability of each bond forming calculated by the PredDisulfideBond tool is given in the adjacent column. The higher the probability the more likely a bond is to form. The Maestro bond score ($S_{ss)}$ which is a way to rank potential bonds. The lower the score the better. Each cysteine can only be involved in one disulfide bond at a time. In instances where there are two possible bonds arising from the same residue the bond with the higher probability is formed. Leading to the three possible bonds not predicted to be formed (shown in red). Both tools agree on these three bonds.

CBD	PredDisulfideBond Probability	Maestro bond score ($S_{ss)}$	222	PredDisulfideBond Probability	Maestro bond score ($S_{ss)}$	CS6	PredDisulfideBond Probability	Maestro bond score ($S_{ss)}$
CYSA12-CYSA38	0.994	0.010523	CYSA17-CYSA43	0.988	-1.25477	CYSA17-CYSA43	0.978	-1.70391
CYSA26-CYSA53	0.997	-0.75269	CYSA31-CYSA58	0.998	-0.35179	CYSA31-CYSA58	0.998	0.043687
CYSA70-CYSA96	0.963	-1.6984	CYSA75-CYSA101	0.987	-1.01159	CYSA75-CYSA104	0.985	-1.70391
CYSA84-CYSA111	0.99	-0.75269	CYSA89-CYSA116	0.992	-0.77988	CYSA89-CYSA119	0.993	-0.70699
CYSA128-CYSA154	0.986	-1.00048	CYSA133-CYSA159	0.875	-0.36565	CYSA136-CYSA162	0.983	-1.29277
CYSA142-CYSA169	0.995	-0.18488	CYSA147-CYSA174	0.998	-0.77988	CYSA150-CYSA177	0.997	-0.70699
CYSA26-CYSA38	0.981	0.61465	CYSA31-CYSA43	0.981	0.572124	CYSA31-CYSA43	0.987	0.096255
CYSA84-CYSA96	0.988	0.172615	CYSA89-CYSA101	0.977	0.682142	CYSA89-CYSA104	0.979	0.74701
CYSA142-CYSA154	0.976	0.727887	CYSA147-CYSA159	0.992	0.021001	CYSA150-CYSA162	0.938	1.281946

Figure 67 shows where the disulfide bonds predicted in **Table 43** were positioned in the models for all the proteins analysed in vitro in this chapter. The mutations were not predicted to impact disulfide configuration.

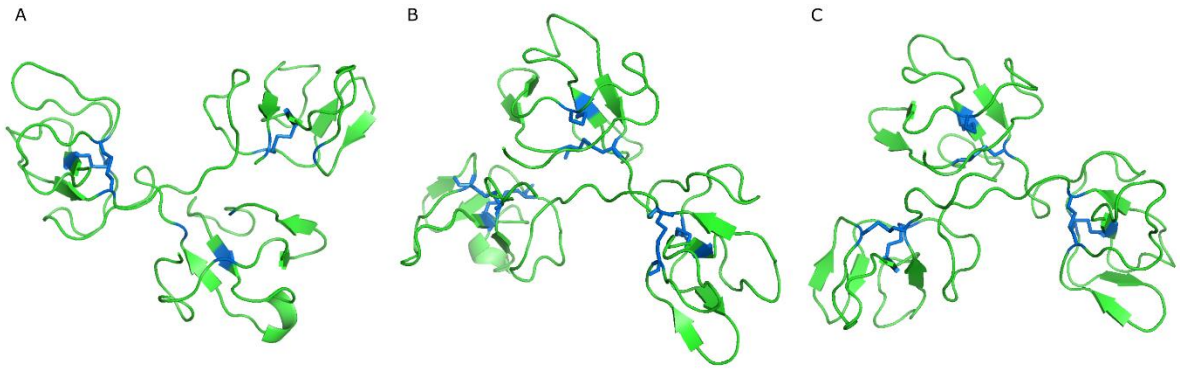


Figure 67: Disulfide bond positions shown on models for all proteins in this chapter. (A) CBD, **(B)** 222 and **(C)** CS6. The bonds shown are those identified in the predictions outlined in **Table 26** highlighted in blue on the protein models generated and selected earlier in the chapter (See **results section 3.3.2** & **3.3.5**). There are six disulfide bonds per model, 2 per module as was expected based on the CBD protein. The alanine mutations were not predicted to impact disulfide bond configuration.

4.3.3.2.2 *In vitro* disulfide assessment

In vitro disulfide bond assessment for 222 and CS6 was also carried out via SDS-PAGE analysis with (R) and without (NR) reducing agent (**methods section 3.2.11.6**). The disulfide bonds present in NR samples make the proteins more compact, hence NR samples run faster on a gel resulting in lower (seemingly lighter) bands [387]. This is seen for both 222 and CS6 indicating disulfides were present as expected the NR samples. These bonds were reduced, so no longer present in the R samples shown by the band shift (**Figure 68**). CBD had been confirmed previously to have the same configuration of six disulfide bonds as 222 so was not revaluated here [235].

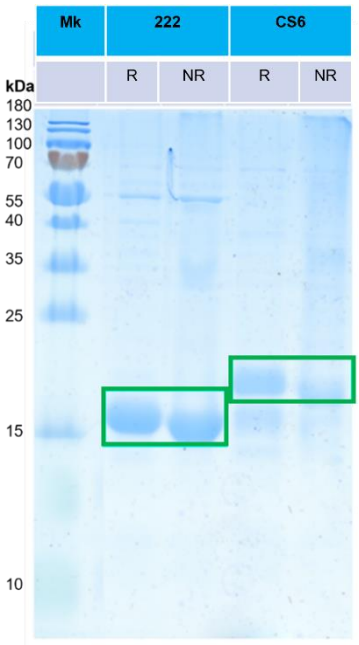


Figure 68: Reducing (R) vs Non-Reducing (NR) SDS-PAGE gel disulfide assessment. The disulfide bridges present in NR samples make the proteins more compact so make it run faster on a gel [387]. Bands corresponding to 222 (20.65kDa) and CS6 (21.02kDa) are shown in the green boxes.

4.3.3.3 Circular Dichroism

CD was used to assess the secondary structure of the three proteins in this chapter and to monitor any conformational changes between these variants. MRE calculated and plotted here normalizes the spectra using protein concentration. Two distinct spectral patterns were observed (**Figure 69**). 222 and CS6 showed one spectral pattern (Group CD1). In contrast, CBD showed a different spectral pattern (Group CD3). The two proteins grouped together as CD1 displayed spectral maxima at between 190-195nm and a minima between 205-210nm. The protein referred to as group CD3 showed characteristics consistent with a folded fibronectin domain-rich protein, with a spectral maximum at 224nm and minimum at 198nm, with a mixture of alpha turn and beta sheet [468].

222 and CBD (the less soluble proteins based on Camsol score predictions, **Table 40**) have more beta-sheet structure than CS6 (designed improved solubility mutant in this chapter). Referring to the NRMSD fit classifications outlined by Hall et al, 2014 [389] and utilised in analysing the CD results in **results section 3.2.11.2**, all proteins in this chapter have a variable fit according to Bestsel NRMSD results (**Table 44**). Therefore, too strong of a conclusion cannot be made about any based upon this analysis alone.

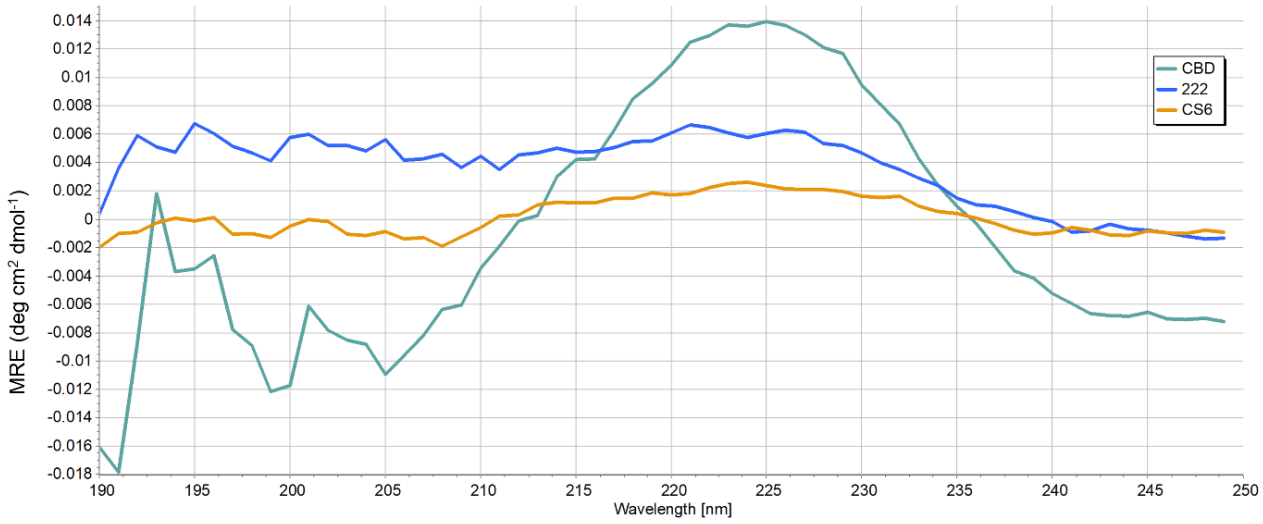


Figure 69: Far-UV CD spectra of CS6, 222 and CBD in 2.5mM HEPES, pH 6.5. CBD (0.2mg/ml, 222 and CS6 at 0.4mg/ml). CBD spectra was taken from previous work of group, Anais Dabbadie (Unpublished).

Table 44: Bestsel analysis of CD spectral results. Showing % composition of each secondary structure type for CBD, 222 and the CS6 (solubility mutant) taken to in vitro experiments in this chapter. The CD1 grouping proteins are shown in Black text and CD3 grouped protein shown in green text, grouping based upon spectra. Root-mean-square deviation (RMSD) reported to show difference between observed (measured spectral values) and predicted spectra of best fit identified by BestSel, a lower RMSD value indicates less discrepancy, so an increased agreement/accuracy (No acceptable threshold for RMSD is defined by Bestsel or within the literature. Normalised root mean square displacement (NRMSD) is also reported to allow definitive quality comparison between different CD experiments. For purposes here classifications outlined by Hall et al, 2014 of NRMSD were adapted (Considered a good fit for an NRMSD < 0.03; reasonable for an NRMSD < 0.05; and variable above this).

GROUP	**CD3**	**CD1**	**CD1**
Secondary structure	CBD	222	CS6
Helix	15.6	0	2.7
Antiparallel (β Sheet)	44.9	50.5	38.6
Parallel (β Sheet)	0.0	0	3.3
Turn	39.5	0	6.6
Total defined	**100.0**	50.5	51.2
Others	0.0	49.5	48.9
RMSD	1.7652	1.2309	0.6147
NRMSD	0.15576	0.07574	0.06607

4.3.3.4 Stability

4.3.3.4.1 NanoDSF

The stability of CBD, 222 and CS6 were assessed via nanoDSF. The results shown in **Figure 70**, show an atypical ill-defined melt curve for CS6. The melt curve shown here for CBD is most typical with a nicely defined transition, with 222 there is a transition, although less pronounced than that seen previously in nanoDSF analysis of a different batch of 222 (**Figure 54**). This analysis here was different to the previous one, all samples in either separate analysis were carried out at the same time to enable comparison. CBD is shown to be a significantly more stable protein with a much better-defined melt curve and considerably higher melt temperature (Tm) of 69.98°C. These result are consistent with the CD results showing a difference in structure for 222. The atypical curve for CS6 means Tm was not able to be calculated here.

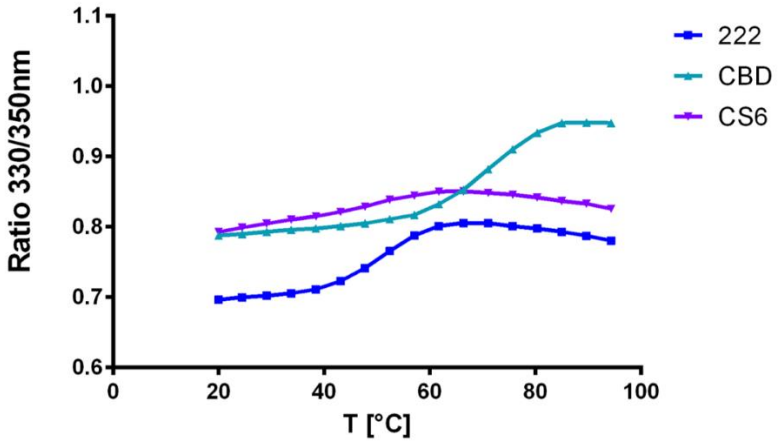


Figure 70: NanoDSF, stability assessment results for proteins in this chapter; **CBD, 222 and CS6**. NanoDSF utilises intrinsic tryptophan and tyrosine fluorescence changes resulting from alterations of the 3D-structure of proteins as a function of the temperature as a protein unfolds. NanoDSF is a way to monitor protein melt profile from this a Mean Tm can be calculated and used comparatively to show stability changes (n=3).

Table 45 shows a summary of the Tm values derived as a quantitative measure of protein thermal stability. Tm values indicate that CBD the original non mutant protein (from which TII gelatin targetting work began prior to this thesis) with a Tm of 69.98°C, was considerably more stable with a higher Tm than 222 with a Tm of 48.65°C. 222 and CBD had a traditional melt curve with a clear ratio transition, and CS6 mutant however produced an altered/ atypical in shape melt curve.

Table 45:Mean Tm value summary data for proteins in this chapter. Showing stability differences between the 222 and CBD, the two proteins in this chapter for which Tm was calculated. 95% confidence intervals Tm range and span are also shown for each protein, (n=3).

	222	CBD
Tm [°C]	48.65	69.98
95% Confidence Intervals		
Tm [°C]	47.59 to 49.73	68.96 to 71.02
Span	0.08926 to 0.1000	0.1579 to 0.1751

4.3.3.4.2 *In silico* stability predictions

In silico predictions of Tm values were also examined after the NanoDSF (**Table 46**). Values were comparable to those calculated based on the *in vitro nanoDSF* data. CBD is predicted to have a Tm of 72.7°C, consistent with the measured nanoDSF derived Tm of 70.0°C. In contrast, 222 and CS6 are predicted to have lower Tms of 61.3°C for 222 and 61.5°C for CS6. *In silico* and *in vitro* analyses of stability for 222 give similar stability profile, which is lower than the original CBD protein. Although Tm for CS6 was not determined from NanoDSF data (due to atypical melt curve) in silico prediction was that it would be similar in stability to 222.

Table 46: Scoop prediction of midpoint melting temperatures (Tm) for all proteins in this chapter. Determined by the Scoop webserver algorithm, protein structure files (PDB) and host organism *E.coli*. CBD is predicted to be significantly more stable than 222 and CS6. With a ~10°C higher Tm. Which is why they are grouped is alone in the S1 grouping as a protein with a Tm ≥72 °C whereas CBD with a Tm ≤61.5°C was grouped in S2.

SCooP	CBD	222	CS6
Stability Prediction (Tm, °C)	72.7	61.3	61.5
Grouping	S1	S2	S2

4.3.3.5 Binding assay

Analysis of the binding of CS6 to TII gelatin is shown in **Figure 71**, with a calculated Kd of 46.7nM.

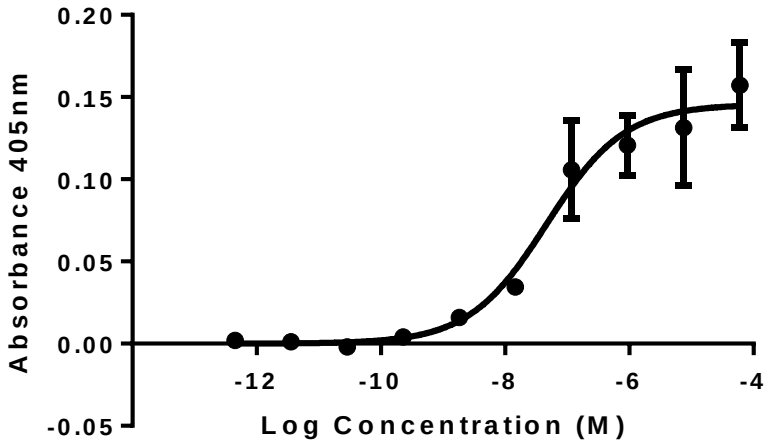


Figure 71: CS6 binding curve used to calculate the binding affinity (Kd) to TII gelatin. Each replicate was repeated in triplicate across each plate, then across 3 separate replicate plates.

Table 39 gives a summary of binding affinities of the proteins assessed in this chapter with the solubility assay. Binding affinity for 222 calculated as 3.9nM (**Figure**

55) and CBD 20.4nM were taken from previous work using the same methodology of a plate binding assay (Anais Dabbadie). The binding assay shows that CS6 binds with less affinity (46.7nM) than 222 (3.9nM) and CBD (20.4nM). A lower Kd shows higher affinity binding so 222 is confirmed still the best binding protein analysed. This means that despite being predicted more soluble CS6 does not represent a viable mutant to take forward to develop as a treatment option.

Table 47: Binding affinities of CBD, 222 and CS6. 222 has the lowest Kd of 3.9nM, reflecting the strongest affinity to TII gelatin. Lower the Kd, tighter the binding.

Protein	Kd (nM)	Sandard error (nM)
CBD	20.4	
222	3.9	±0.7
CS6	46.7	

Data presented in this chapter highlights that there is a delicate balancing act between different protein characteristics; namely binding and solubility. Based on knowledge of how hydrophobic residues often form binding pockets [469] and hydrophilic residues seem to be more abundant in more soluble protein variants [424, 470], it may be that optimising both binding and solubility in one molecule is biochemically unachievable. Improved binding may always be achieved at the cost of solubility and solubility improvement will always be achieved at the cost of binding.

The hypothesis was subsequently posed that there is an inverse relationship between binding and solubility

4.3.3.6 Solubility Assay

All three proteins (CBD, 222 and CS6) demonstrate increasing absorbance at 500nm in the presence of both PEG 8000 and ammonium sulfate. **Figure 72(A)** shows an increase in absorbance reflecting an increase in turbidity caused by the precipitation of proteins with increasing PEG concentration. **Figure 72(B)** also shows an increase in absorbance representing an increase in turbidity caused by the precipitation of proteins with increasing ammonium sulfate concentration. An expected positive correlation is present, with precipitation causing turbidity increase with increasing precipitant concentration for both precipitants [417, 427, 434, 435, 467, 471].

Theoretically the more soluble the protein the more precipitant needed to cause its precipitation. The absorbance at 500nm quantifies turbidity caused by precipitation, it is the first more crude measure attained from this solubility assay. With PEG the point of inflection (where precipitation begins) is higher for CS6 as the increase in absorbance caused by precipitation isn't seen until greater than 10% PEG is present with CS6, whereas for 222 and CBD this is seen with the lower concentration of 5% PEG (**Figure 72A)**. With ammonium sulfate the increase in absorbance is not evident until 0.5M and all three proteins show an increase in absorbance at the same precipitant concentration, so this result provides less precision and discrimination than the PEG data (**Figure 72B)**.

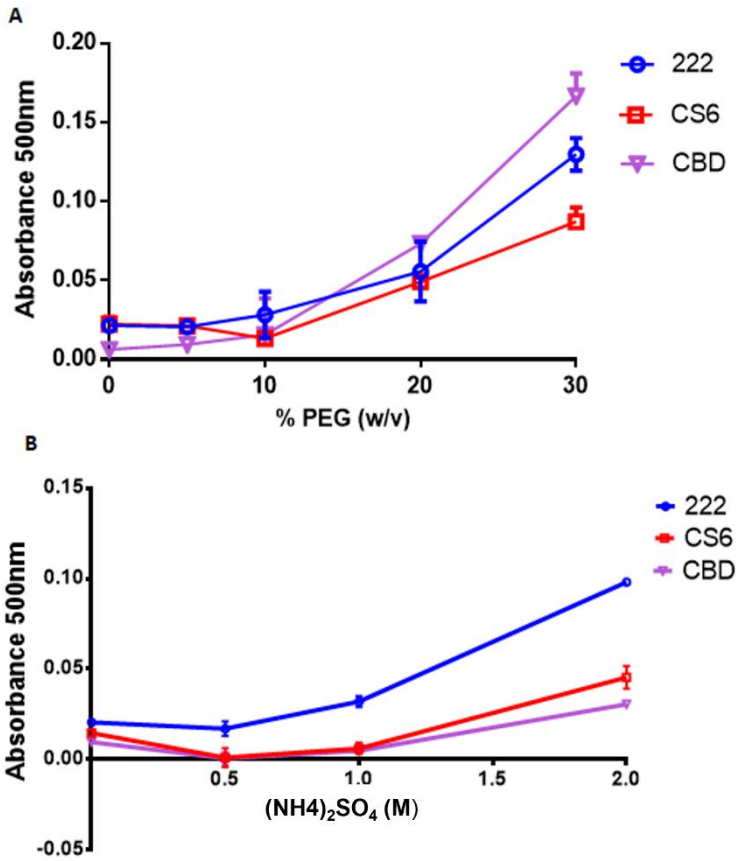


Figure 72: Turbidity increase at varying concentrations of precipitant. As proteins (222, CS6 & CBD) precipitate turbidity increases, detected by increase in absorbance at 500nm. Plot showing this increase plotted against **(A)** increasing PEG 8000 concentration, **(B)** increasing ammonium sulfate concentration (M). All samples were n=3 in two different plates.

The second measure made in the solubility assay was determining the rate of decrease in soluble protein concentration, as protein precipitation increased. **Figure 73** shows that as absorbance increased in **Figure 72** concentration of soluble protein decreased. **Figure 73(A)** shows normalised % of starting protein concentration vs PEG concentration. **Figure 73(B)** shows normalised % of starting protein concentration vs ammonium sulfate concentration. The inflection point shows the concentration where loss of protein begins (due to precipitation) , indicating CS6 is the more soluble protein, remaining in solution until a higher concentration of precipitant (inflection point of 10 with PEG and 1 with ammonium sulfate).

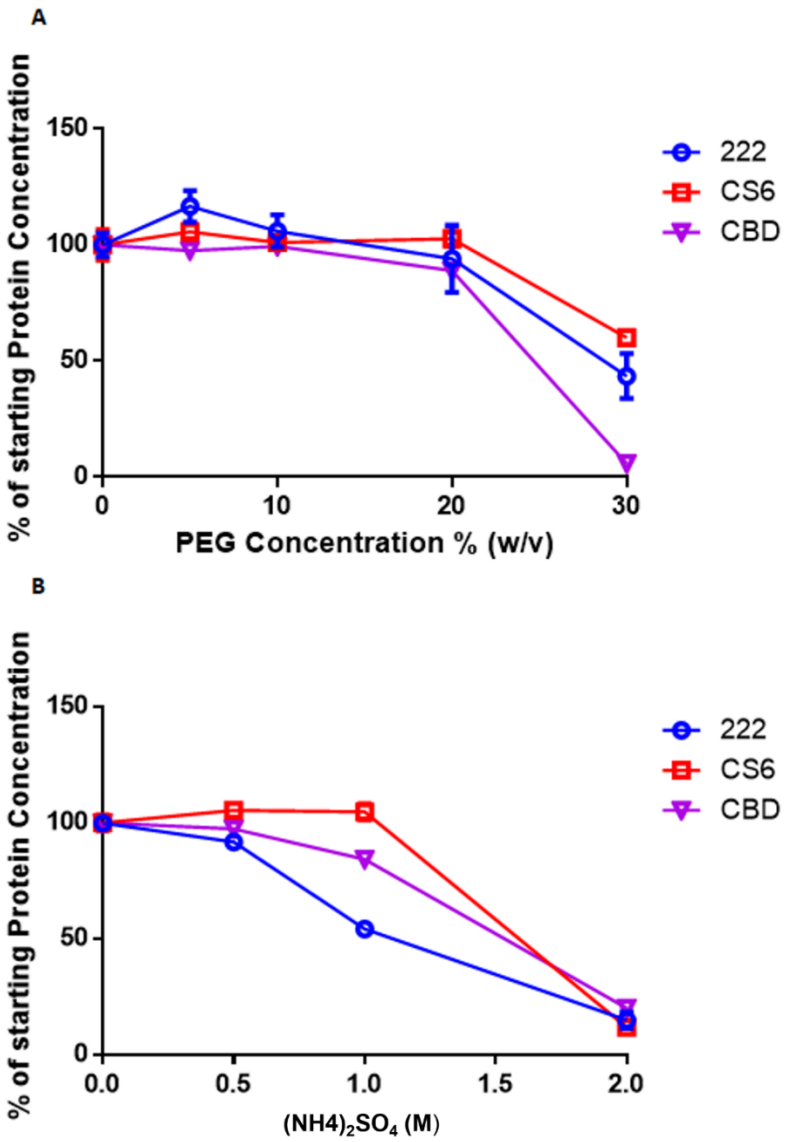
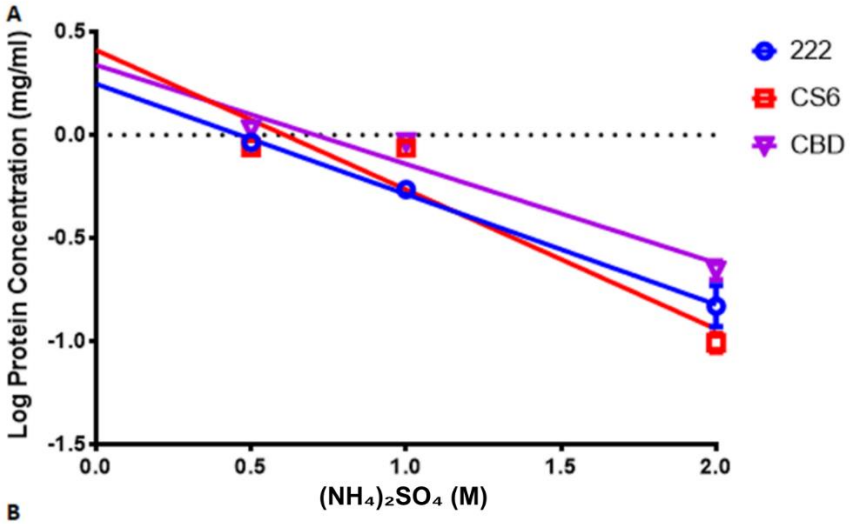


Figure 73: Protein concentration remaining at varying concentrations of precipitant. As protein precipitates out of solution protein concentration decreases.**(A)** Normalised remaining protein concentration (CBD, 222 and CS6) (%) vs precipitant concentration (PEG). **(B)** Normalised remaining protein concentration (%) vs precipitant concentration (ammonium sulfate).

The data from 0.5-2M for ammonium sulfate shown here in **Figure 73** was then used to calculate apparent solubility with a log linear plot as per **methods section 4.2.3.1.** This calculation proceeded with **Equation [10]** (slope of the line and y intercept value derived were from the linear interpolation to 0M ammonium sulfate) derived for all three proteins shown in **Figure 74A** & **C**. PEG was excluded from this

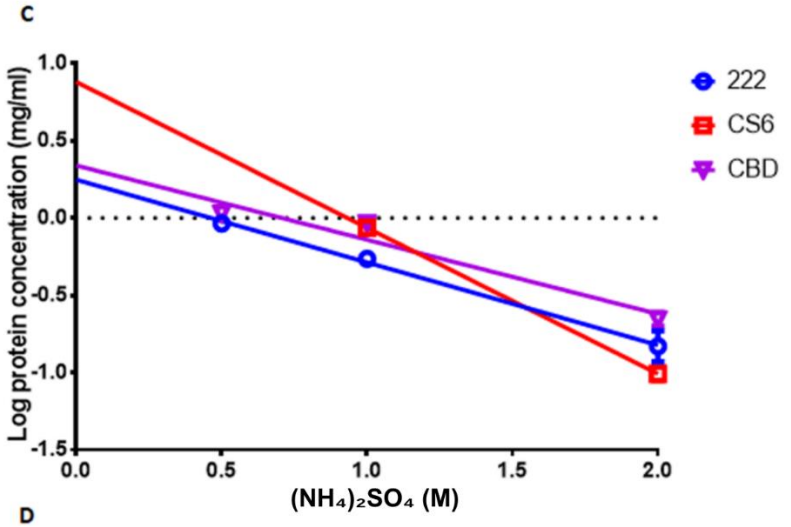
further analysis, as the log linear relationship was not observed for these proteins as expected [91, 101, 108, 109, 138, 140].

Figure 74 shows two possible log linear plots for protein concentration remaining in the solution in the wells of the 96 well plate in the presence of increasing concentrations of ammonium sulfate (precipitant) , for these proteins (CBD, 222 and CS6). It was observed that with CS6, interpolation from the complete data set with the 0.5M CS6 protein concentration outlier included didn't have the best linear fit skewing the point of interpolation when included, as noted in **Figure 74(A).** A decrease in protein concentration for CS6 wasn't observed until 1.0M ammonium sulfate. Therefore, including the 0.5M precipitant data point for CS6 as shown in **Figure 74(A)** leads to an underestimation of apparent solubility due to this skew. Although not best mathematical practice interpolating from just two data points, here it makes for a better linear interpolation for CS6 and therefore a more indicative apparent solubility determination **Figure 74(C).** It is likely that the apparent solubility of CS6 lies between the two values (12.229 and 66.819 mg/mL) determined here and reported in **Figure 74(B) & (D)**.



B

Protein	Apparent solubility (mg/ml)	Solubility order	Camsol solubility score	Camsol predicted order
CBD	6.601	2	0.968	3
222	6.063	3	1.835	2
CS6	12.229	1	2.827	1



D

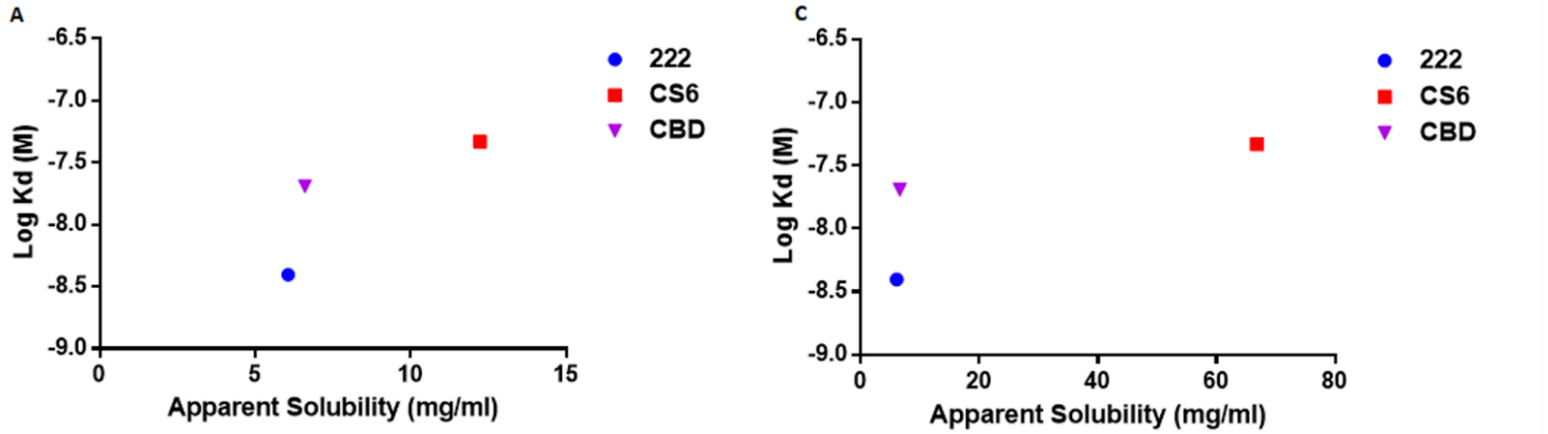
Protein	Apparent solubility (mg/ml)	Solubility order	Camsol solubility score	Camsol predicted order
CBD	6.601	2	0.968	3
222	6.063	3	1.835	2
CS6	66.819	1	2.827	1

Figure 74: Apparent solubility determination of the three proteins in this chapter. (A) Linear interpolation of data for 0.5-2M ammonium sulfate **(B)** Calculated apparent solubility inferred from (A) and order of solubility. **(C)** CS6 0.5M protein concentration (a clear outlier that skews the point of interpolation when included) excluded to give a better fitting linear interpolation to be used in calculating apparent solubility. **(D)** Calculated apparent solubility inferred from (C) and order of solubility.

Data reported in this chapter shows that CS6 is more soluble than CBD and 222 as predicted using the CamSol tool (**Figure 60**). CS6 has an apparent solubility between 12.229mg/mL - 66.819mg/mL depending on the interpolation plot used. In contrast, CBD and 222 have similar solubility of 6.601 and 6.036 mg/mL. Importantly both plots with or without CS6 linear outliers excluded show CS6 is either approximately twice or ten times more soluble than both CBD and 222. Therefore, the CamSol method is a valuable way to design and guide experimental attempts to improve solubility mutants of a protein.

4.3.3.7 Binding vs solubility

To test the inverse relationship hypothesis between solubility and binding potency plots were generated to elucidate if such a relationship existed between solubility and binding for these three proteins in this chapter. **Figure 75, A & C** show plots of binding affinity for the three proteins CBD, 222 and CS6 against apparent solubility values. **Figure 75, B & D** give a numerical breakdown of what is shown in these two plots and the calculated apparent solubility values(as shown in **Figure 74**). The apparent solubility measures for 222 and CS6 support this mutually exclusive relationship, that a protein is either a good binder or soluble, not both. However, data for CBD do not support this theory, highlighting there is not an entirely linear inverse relationship between binding and solubility.



B

Protein	Kd (nM)	Binding order	Apparent solubility (mg /mL)	Solubility order	CamSol solubility score	CamSol predicted order
CBD	20.4	2	6.601	2	0.968	3
222	3.9	1	6.063	3	1.835	2
CS6	46.7	3	12.229	1	2.827	1

D

Protein	Kd (nM)	Binding order	Apparent solubility (mg /mL)	Solubility order	CamSol solubility score	CamSol predicted order
CBD	20.4	2	6.601	2	0.968	3
222	3.9	1	6.063	3	1.835	2
CS6	46.7	3	66.819	1	2.827	1

Figure 75: Binding vs Solubility of CBD, 222 and CS6. (A) Measured Binding affinity (M) measured apparent solubility inferred from **Figure 74A** (mg /mL). **(B)** Tabular representation of the data; measured Binding affinity (M, or nM), binding order vs measured apparent solubility inferred from **Figure 74A** (mg /mL). CamSol score and predicted order also shown for comparison. **(C)** Measured binding affinity (M) vs measured apparent solubility inferred from **Figure 74C** (mg /mL). **(D)** Tabular representation of; measured binding affinity (M, or nM), binding order vs measured apparent solubility inferred from **Figure 74C** (mg /mL). CamSol score and predicted order also shown for comparison.

4.4 Discussion

This chapter has described the computational design of mutants with enhanced solubility. Utilising again here the modelled 222 structure derived in **results section 3.3.3** first shown in **Figure 29**. An extensive review of available computational tools to guide the improved solubility was conducted initially to identify the best suited to the task**.** The CamSol tool was identified, selected, and utilised here as the best available tool, employed to guide *in vitro* work towards the primary aim of this chapter which was producing a mutant of 222 (the best binding CBD mutant created thus far by the Hollander group) with enhanced solubility.

The CamSol method was used with success here, to inform what was an otherwise very challenging pursuit. All 181 residues that make up the 222 proteins could be changed to any of the alternate 19 amino acids, however, excluding the fifteen binding residues from mutation, to maintain the primary binding functionality of the protein reduces the number of residues to mutate to 166. Meaning there are $9.35X10^{215}$ possible mutant protein possibilities, a number far too large to ever consider trialling *in vitro*. CamSol was used here help identify rational options based on knowledge and theory of protein solubility, to achieve higher solubility in a specified protein structure. CamSol was chosen as a tool that makes this assessment based on structure not just sequence. Within bioinformatics structural tools are considered superior as they take into consideration inter-molecular interactions, occurring between amino acid side chains. Such interactions can have large implications on fold and residue proximity, accessibility, exposure and ultimately protein function [472, 473].

CS6 was the output generated here, predicted to be the best based on CamSol score, compared to 222. CamSol design produced a total of six possible mutants ranging from minimally changed (one mutation) in CS1 to the most changed (two substitutions and the insertion of EEE) in CS6. Originally the plan had been to test and quantify the difference in solubility for all six mutants, but due to time constraints and the disruption of the COVID pandemic the decision to trial only the most improved was made. With the option to go back and trial the other five or carry out additional optimization in the subsequent work of the Hollander group.

It is always good practice to use two contrasting bioinformatic tools where possible to strengthen the predictions and justification for taking mutants to *in vitro* testing. Aggrescan3D is also a structural assessment tool and was considered here but was deemed not suitable as it didn't have a function for designing enhanced solubility mutants. Aggrescan3D is more suited to looking into aggregation propensity of proteins which is related but does not provide suggested mutations to improve solubility as the CamSol tool does [448, 449]. Aggrescan3D has only more recently (since this work was undertaken) been updated to incorporate a functionality for engineering protein solubility mutants [448]. It was originally thought that this server could be used at a minimum as a check, for agreement (that the CS6 structural file modelled from the suggested mutants, was at the least predicted to be more soluble compared to 222). Although this tool can only look at substitutions between proteins not the combination of substitutions and the insertion present within CS6 (**Table 25**). There is now also SOLart available which would have been utilised had it been available when this work was undertaken [447].

The next part of this project moved to *in vitro* analysis, which first involved expression and purification of the CS6 protein. This time unlike with **chapter 3**, given that only the one mutant was being expressed for the first time when it came to cloning, three different fusion tags were planned to be trialled from the outset. The OPPF-UK pOPIN vector suite [353] was utilised more extensively in this chapter allowing relatively easy trialling of three different fusion tags [460, 461]; SUMO, GST and MBP. All three were trialled for soluble expression and sufficient recovery of CS6 for the experiments in this chapter. MBP was the tag that gave the best SDS-PAGE analysis results showing a good level of soluble protein expression and recovery in purification. With CS6 protein attained and reserves of CBD and 222 available the project then progressed to the characterisation stage.

Quantification methods for solubility are also not straightforward or universally accepted/ adopted [417]. Here, an amorphous precipitation method was utilised and adapted in a 96 well plate format [467]. The method developed with PEG was adapted here to also use a contrasting second precipitant, ammonium sulfate. Which achieved precipitation by a different mechanism than the PEG used in the paper this method was adopted from [467]. This method allowed the extrapolation of apparent solubility for all three proteins explored within the work of this chapter, but only following precipitation with ammonium sulfate. With PEG there may have been a requirement for a higher concentration of protein per plate well to attain . This analysis highlights an important consideration that would be worth further exploring, is ammonium sulfate a more sensitive precipitant in general? This could be assessed with other proteins not just the three mutants in this chapter or is it protein dependent. The original assay from which this work was based evaluated only

antibodies which are, by their very nature present in much higher concentration and more soluble [467].

CS6 was also observed via CD spectral analysis using Bestsel to have the most beta-sheet and turn structure of the three mutants within this chapter. Whether this links to the increased solubility that is a question that is not answered here unfortunately. Bestsel result fit analysis was classed as variable for all three proteins here according to the NRMSD classifications used in analysing the CD data in the previous chapter, [389] . So unfortunately, that strong of a conclusion cannot be made, but the idea that more soluble proteins have a distinctive secondary structure is a valid one that may be applicable to future work.

Again, looking at the breadth of analyses conducted within this chapter it was observed that there are some distinct trends and groupings present within the proteins and data presented in the work here (**Table 48**). CBD (in group CD3) and 222 (group CD1) had a typical melt curves with clear transitions when examined via nanoDSF as would be expected as a protein unfolds. The Tms however did not agree with this grouping, CS6 did not have the typical melt curve enabling Tm calculation it may indicate lack of suitability to this technique. Also, first use of DSF (data not included) preliminary to the work here showed that these proteins did have a high fluorescence to start meaning they were possibly not as well folded to begin as CBD (the only natural, native non mutant protein in this work). Again, implying that the mutant proteins (CBD and 222) are not suited to a thermal unfolding method like nanoDSF.

Three ways that this largely unfolded natural state could be confirmed amongst these CBD mutants are a subsequent comparative assessment of aggregation state using size exclusion chromatography [474] or 1D NMR (there would be a lower peak dispersion if they were unfolded to begin) [475, 476]. A more crude but easier to action test of stability if the atypical curve proteins were indeed indicative of the proteins unfolded state, they would precipitate out at a much lower concentration when concentrated [417].

Comparing other properties of these groups of proteins as outlined in **Table 48** There were two groupings in the CD results, CD3 had structure and CD1 had altered more unfolded other structure. Occam's Razor applies the principle that simpler explanation should be preferred and correct one [477] , here that would be that the mutants are more unfolded to begin. Lack of starting folded state would explains all the oddities and inconsistencies with atypical melt profiles, CD spectra, expression difficulties and some variance seen between batches of protein. This idea of a default/ native unfolded state warrants further investigation if CBD mutants are to be pursued further as intended.

Discordance in the groupings between different techniques, in silico and in vitro, is clear in some places in this this summary **Table 48**. The fact that some analysis does not separate them at all, others separate into two and others three cohorts. Discordance between *in silico* and *in vitro* methods is also evident in some assessments. For secondary structure analysis *in silico* did not predict large differences in % of each type of secondary structure (all were grouped under SS1 (222) or SS2 (CBD and CS6), but *in vitro CD* analysis showed evidence of alteration

that separated the proteins differently into CD3 (CBD)and CD1 (222 and CS6). Then in agreement for others of particular importance here is the agreement between *in silico* predictions and *in vitro* measures of apparent solubility in that *in silico* did predict what was seen *in vitro* in that CS6 was substantially more soluble than both than CBD and 222 (**Figure 74**). Interestingly binding results in this chapter CBD and CS6 as being in the same range of 20-50nM, less potent binding than the gold standard mutant 222.

The results here in this chapter highlight how stability should also be considered as a priority when designing mutants with improved solubility (CS6) or binding affinity (222). The atypical melt curve difference seen with CS6 shows the mutations to improve solubility have had an effect on its thermal stability properties. So, although we met the primary aim, we have potentially done so at the unexpected cost of stability.

Table 48: Characterization results grouping and trends summary. CD spectra split into CD1 (CS6 and 222) and CD3 (CBD) protein groups. A more subjective but significant result was ease of recovery all grouped under Y for yes (CBD, 222 and CS6) which were easy to recover in soluble form. NanoDSF determined Tm data split into S1 (CBD) and S2 (222). *In silico* stability prediction made using the SCOOP webtool split the proteins into two groups, S1 (CBD) and S2 (222 & CS6). *In silico* disulfide predictions did not separate the proteins all were comparable and grouped under D1 (CBD, 222 and CS6). SDS-PAGE assessment of disulfide configuration put proteins into one grouping D1 (222 and CS6 were assessed here).*In silico* and *in vitro* solubility results again agree on the groupings with S1 (CBD and 222) and then considerably more soluble S2 (CS6). Binding could also be grouped into B2 (low nm Kd best binding, 222) then B1 (20-50nm higher Kd than CBD but similar Kds to each other, 222 and CS6).In silico secondary structure prediction split proteins into two groups based on slight differences. SS1 for 222 and SS2 for CBD and CS6.

Characterisation result groupings/ trend comparison			**CBD**	**222**	**CS6**
CD, Secondary structure	Spectra comparison	*In vitro*	CD3	CD1	CD1
Ease of recovery	Experience	*In vitro*	Y	Y	Y
NanoDSF, Stability	Tm	*In vitro*	S1	S2	/
NanoDSF, Stability	Typical melt curve	*In vitro*	Y	Y	N
Stability prediction	Tm	*In vitro*	S1	S2	S2
Disulfide bond assessment	No. of bonds, Cys residues involved	*In silico*	D1	D1	D1
Disulfide bond assessment	SDS-PAGE	*In vitro*	D1	D1	D1
Solubility	CamSol score	*In silico*	S1	S1	S2
Solubility	Apparent solubility	*In vitro*	S1	S1	S2
Binding	Kd	*In vitro*	B1	B2	B1
Secondary structure	Type %	*In silico*	SS2	SS1	SS2

Computational protein design has undergone rapid advancements in recent years, making it a very dynamic and exciting area of biology [478, 479]. A significant issue encountered here but applicable to many strategies not just computational, is balancing the desired protein characteristics. Optimising multiple parameters at once is still beyond such tools. Here, we wanted a high affinity binding protein with improved solubility. In this work it has also become apparent stability is another important characteristic that should be considered more in future work. Making a change to improve one characteristic can have a knock-on effect on another, as was seen here. CS6 was more soluble but was both less stable and demonstrated a reduced binding affinity for TII gelatin.

4.4.1 Highlights

It is important to remember that solubility is just one parameter required for this protein therapeutic strategy. A more soluble protein would be easier to attain and work with. However, yield and solubility aren't the same thing, but it is generally accepted that if a protein is more soluble, a higher yield would be a more easily optimisable outcome.

Here, the previously developed protein 222 is aimed to be used as a means of targeting, adhering, and facilitating longevity of retention of intraarticularly delivered MSCs. The hope is to achieve adherence, integration of MSCs within damaged TII gelatin rich regions of the OA joint. Longevity of retention increases the likelihood that repair and regeneration of the damaged ECM could occur. There is no distinction between early and late-stage OA damaged TII collagen that has been degraded to TII gelatin, with TII gelatin present in both early and late-stage OA

joints. However, there would be less healthy TII collagen around at the surface of the articular joint in late-stage disease. 222 was engineered from MMP-2 rather than MMP-9 as module 2 of the CBD of MMP-2 provided a discriminatory specificity to degraded TII gelatin but not intact TII collagen, therefore cells would only adhere to damaged regions. Preparation of sufficient quality, quantity, and promotion of differentiation to a chondrogenic fate of the MSCs themselves are all separate considerations of such a therapeutic strategy that will require separate optimisation. Disease stage might have implications for dosing, i.e., more cells and repeated dosing in more advanced OA. This would have implications for protein requirements; consistent, predictable, replicable protein production is required which is something that 222 is not able to achieve currently and why optimisation and further protein development is required.

CS6 was the predicted most soluble mutant, it was expressed, purified, and tested here. Data generated using a precipitant-based plate assay confirms that it does have a significantly higher solubility, of between two and ten times higher than the native CBD protein and the already attained 222 (chimeric CBD mutant with the confirmed highest binding affinity for TII gelatin).

Binding was determined to be weaker for CS6 to TII gelatin than for 222, assessed using the coated plate binding assay from [235, 379, 480] **chapter 3**. Therefore, despite CS6 being a significantly more soluble protein which was the aim when the work in this chapter began, unfortunately it is not a suitable target to take forward as a protein to coat MSCs and be used to replace 222 in the OA treatment strategy under development by the Hollander group. The work here does show that

solubility of 222 can be improved, but subsequent work is required to maintain binding alongside increasing solubility. Linking back to the idea that some diseases are the result of single point mutations (small changes to proteins have profound impact in nature not just as a side effect of bioengineering) it was always a gamble to aim for the most altered mutant, with the most improvement in terms of solubility. The CamSol score only gives a comparative indication of improvement not an absolute value. Therefore, one of the other identified CamSol mutants with more conservative changes to native 222 may offer this sought-after property of enhanced solubility while maintaining binding potency for TII gelatin.

Finally, an initial examination of the relationship between binding and solubility is made here through a plot of measured values for Kd and interpolated apparent solubility, to look for a trend. It was thought that there was a possible linear exclusivity between the two characteristics, but this was not confirmed here. So, it is not an either/or characteristic, although not a fully resolved and understood relationship yet, the work here does leave open the prospect of attaining a protein that has both a high binding affinity as good as, if not better than 222 and with improved solubility. Further work is required to explore this relationship in more detail.

4.4.2 Future work

Further application and investigation of the solubility assay would be recommended immediately following this project to strengthen the conclusions made here. CS6 has sufficient replicates to support the conclusions but one weakness is that the plots used to extrapolate from could benefit from an additional number of

precipitant concentrations. Building on the work here an expansion of concentrations would be most applicable between 1-2M ammonium sulfate. This would reduce protein requirement as initial trials and preliminary work that was required to set up the assay and protein recovery are already conducted. It would also be worthwhile to test this assay not only using an expanded precipitant concentration range. Testing alternate precipitants could be a valuable addition to this analysis presented here, to explore different mechanisms of precipitation [417].

An expansion of the computational work and a different strategy of improving binding, solubility, and stability could be to investigate the phylogenetics of CBD [481]. If homologs (which can subcategorized into orthologs and paralogs [482]), of CBD were identified computationally amongst a diversity of species and organisms, natural optimisation may be identified. A tool such as ASPEN would be a recent method with reported accuracy, to apply to this [483]. Evolution has had far longer than we will ever have to work with such problems, so there might be homologous proteins already available to utilise in this therapeutic way.

Also applying a more general wider net approach than that used here, utilising directed molecular evolution and the construction of a mutant library, would be a valid and alternative approach in subsequent work [484]. This could be achieved using bioinformatics as a guide again, scanning point mutation libraries or using random mutagenesis kits such as the Genemorph II random mutagenesis kit (Agilent). With a binding assay, solubility assay and nanoDSF technique all now tested and validated as methods all could be employed and adopted to be high throughput, screening for the three protein properties (binding, solubility, and

stability) in tandem. The labour intensity of such an approach could be eased using what has been learnt in this project and laboratory automation/ robotics could be adopted if facilities were correctly available and or planned for [485].

Confounding variables and multivariate dependency add complexity to biological systems and experiments [486]. The most important characteristics of proteins with therapeutic applications such as the OA therapeutic targeting strategy in this work; are solubility, binding, and stability. These protein characteristics are highly coordinated and interdependent of each other. Further study is warranted in the future work of the group to elucidate such further for CBD proteins and mutants.

Because of the likely changing flux conformation state of the CBD mutants (at least until bound) in this work including CS6 it is likely that conformation is changing in solution. A key future experiment would be to test this theory (as it may be that all CBD proteins that would work for this purpose (binding to the also very flexible and changing conformation TII gelatin peptide) may be unfolded so will all have these issues.

Next steps following could then be to mutate for improvement of binding and solubility concomitantly based on the results here and the idea that even a mutant optimised for both solubility and binding affinity could still not be a therapeutically translatable development if stability was also not optimised/ considered during development as opposed to dealing with one characteristic at a time. Many more possibilities are presented with the ever emerging and improving bioinformatic tools so what isn't possible here in this thesis in silico may be possible imminently. It

would also be an option to pursue the other CamSol mutants 1-5 as solubility may still be improved but binding affinity may be maintained to be nearer that of 222.

5 Results Chapter: Improving binding of a chimeric CBD protein

5.1 Introduction

The work in this chapter used *in silico* assessment to elucidate a strategy for improving the binding of 222 to the target TII gelatin. 222 is the best binding CBD mutant developed so far and so was the starting point for the work described in this chapter. We have already showed that solubility can be improved (CS6; see **chapter 4**) but unfortunately for this mutant, high affinity binding was not maintained. Given that the primary aim of the therapeutic strategy is targeting further improvement of 222, binding was deemed an essential property as well as enhanced solubility. Stronger binding of the protein that coats the MSCs will increase the chance of integration and repair of damaged tissues, longer duration of retention in the damaged regions will allow the cells more time to differentiate down a chondrogenic path and regenerate damaged ECM including collagen. Based on the results described in **chapter 3**, it was concluded that amongst the binding residues, residues Asn 11, 69 and 127 were most critical. Here we tested effect of mutating these residues to each of the possible alternate nineteen residues in place of the native Asn in 222.

Biomolecular complexes can be considered the molecular machines of the cell integral to functioning. To fully understand how the separate components work together to fulfil their tasks, structural knowledge at an atomic level is required. Classical structural methods such as NMR, X-ray crystallography and cryo-EM

provide this knowledge, but are expensive, laborious, and often encounter difficulties when it comes to complexes. Therefore, valuable information about complexes can be obtained from a variety of experimental and predictive approaches. By combining the available information with computational approaches such as docking, insights into the biomolecular interactions can be elucidated much more readily. During recent years there has been an explosion in the number of docking methods that tackle the more complex binding interactions.

The aim of using molecular docking methods here is to give a prediction of the protein (222) target peptide (TII gelatin) complexed structure, using computational methods. This approach will provide similar information to homology modelling which can't be used for a complex made up of two molecular structures, in this case a protein and the TII gelatin peptide as we don't know how they fit together. Molecular docking as a method both saves time and limits the use of costly experimental methods. Computational docking produces a prediction/ model of the 3D-structure of a biomolecular complex, starting with the structures of the individual molecules in their free, unbound form. Docking can be achieved through two interrelated steps:

1. First a sampling of conformations of the peptide in the active site of the protein.
2. Then the ranking of these possible conformations via a scoring function. Ideally, these computational sampling algorithms should be able to reproduce the experimental binding mode.

Computational determination of the structure of protein complexes has prevailed as one of the central most challenging problems in computational structural biology,

since its beginnings in 1969 with sequence analysis [487]. Even with relatively rigid proteins it is difficult to consider and account for the 6D rotational-conformational space of assembly orientations, that can be sampled by a pair of biomolecules e.g., protein-protein, protein-peptide, or protein-ligand as they interact. Here, our biomolecules are a protein (222) and a peptide (TII gelatin). Interactions occur between the two biomolecules, through undefined (with most but not all docking tools) complementary patches on their surfaces and in distinct regions more commonly referred to as binding sites. Additionally, proteins are not static objects; they are considerably dynamic, constantly interconverting between conformers of varying energies, flexibility further adds to this complexity [488]. Therefore, when you attempt computational assembly of proteins in complex with other proteins, peptides or ligands more commonly known as molecular docking, there isn't necessarily just one correct answer available [489].

The first molecular docking algorithm was developed by Kuntz et al, 1982 [490], computationally docking heme to myoglobin and thyroid hormone to prealbumin. Since then, many advances and improvements have emerged in bioinformatics, machine learning and computing capability has improved vastly. Molecular docking is now utilised extensively in biopharmaceutical research, with vast applications in accelerating drug design. Specifically, during the recent COVID pandemic it was used to rapidly identify novel inhibitory molecules and fast track the drug development process [491, 492]. Molecular docking is now considered a key component of the translational biochemistry toolkit. Docking began with rigid body only, but now also (as will be used here and much more difficult to achieve), flexible

docking. There are also several different models of molecular docking, summarised below:

- The lock and key theory, first proposed in 1890 by Fischer, proposes that much like a key fits a lock, ‘biological locks’ require distinct stereochemically oriented ‘keys; e.g. substrate is the key that fits specifically into an active site (the lock) [493].
- Induced fit theory, first proposed in 1958 by Koshland, states that both molecules in a docking interaction adapt to one another conformationally to reach an ideal match [493].
- Conformational ensemble model, in which proteins have more recently been considered to undergo significantly greater conformational changes than first thought. This new concept states that proteins have the option to adopt multiple conformations from a pre-existing ensemble of conformational states. It is the protein's flexibility that enables it to transition between these states [494, 495].

222, the chimeric CBD mutant on which all the work in this thesis is based, was modelled in **chapter 3**. Therefore, one of the two molecules in the docking is already attained. Additionally, for docking to be possible a TII gelatin peptide is required, along with a review of and identification of the most suitable tool to carry out the docking. **Table 49** outlines requirements, prerequisites and output of docking, format, and source.

Table 49: Docking requirements. Structure files (PDB) for both protein (222) and peptide (TII gelatin), best docking tools identified via literature search, *in silico* match of *in vitro* binding order, M1 position mutant library generation, docking binding affinity evaluation of M1 mutants, CamSol mutants docking, dual solubility/binding mutagenesis and finally docking evaluation of dual solubility/ binding mutants.

Figure 76 ref no.	Requirement	Format	Source
1	222 Protein	Structure model (PDB)	Chapter 3
	222 Peptide	Peptide sequence	Literature search (5.2.1.1.1)
		Structure model (PDB)	Modelling tool literature search (Table 52)
	Best docking tool ID	Web, Windows, or Linux	Literature search (Table 50)
2	Experimental binding order match	Score	Docking trials/ tool validation (5.3.2)
3	Docked complexes	Downloadable docked complex prediction (PDB)	Docking trials/ tool validation (5.3.3)
4a, 4b, 6	Docked complex mutagenesis	Mutant complex (PDBs)	Pymol mutagenesis wizard (5.3.4)
5a	M1 mutant docking	Score	Docked complex refinement (5.3.4)
5b	CamSol mutant docking	Score	Docked complex refinement (5.3.5)
7	Dual solubility/binding mutants docking	Score	Docked complex refinement (5.3.5)

Within this chapter a docked complex structural prediction was sought initially. Given the inherent flexibility in both the 222 protein and TII gelatin peptide this was not straightforward. Posing a computationally intensive challenge, as a rule flexibility can't be easily handled and factored appropriately if present in both biomolecules in

a docking experiment as is so for a peptide-protein docking [496]. Every attempt was made hereto use a representative sampling from the number of conformational possibilities for the TII gelatin peptide which was generated in this chapter from only the peptide sequence. Whereas the homology modelled 222 protein flexibility was dealt with by the docking tool itself. It is important to highlight that with the plate binding assay utilised throughout all chapters of this thesis, both the protein in solution and peptide coated onto the 96 well plate is the same flexible and present in an array of conformations, as the peptide fragment would be in the OA joint itself. Previous *in vitro* binding experiments were used as a validation of any docking technique. The scheme of the work in this chapter is summarised below in **Figure 76** (*in silico*) and **Figure 77** *(in vitro)* This chapter aims to use molecular docking to identify more potent binding mutants of 222 to TII gelatin. Such mutants would then be taken to expression and further characterisation if attained in soluble form.

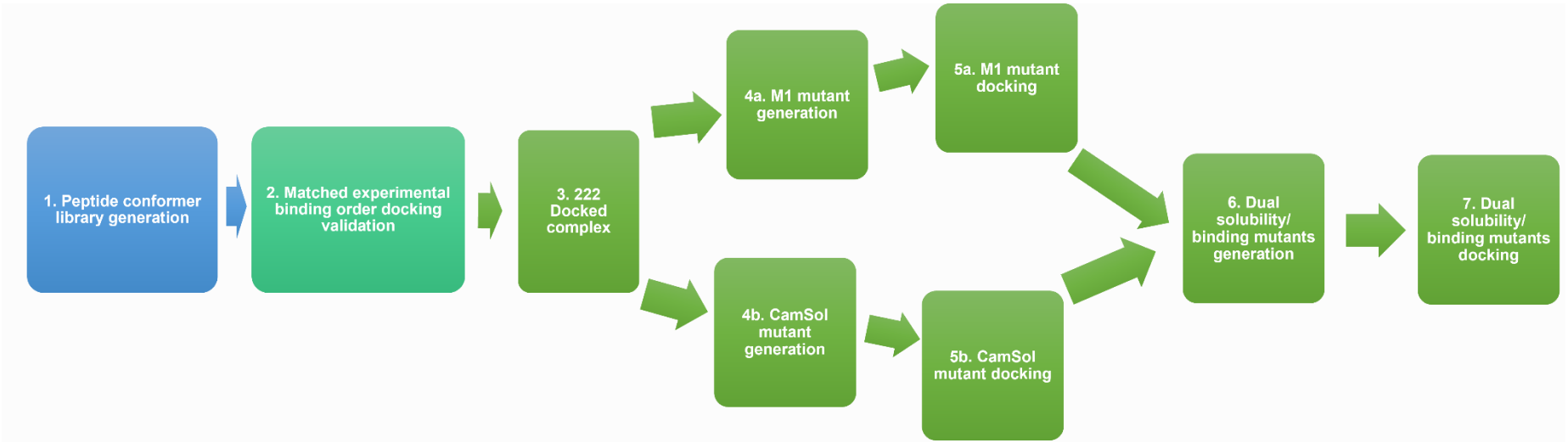


Figure 76: *In silico* pipeline for this chapter.

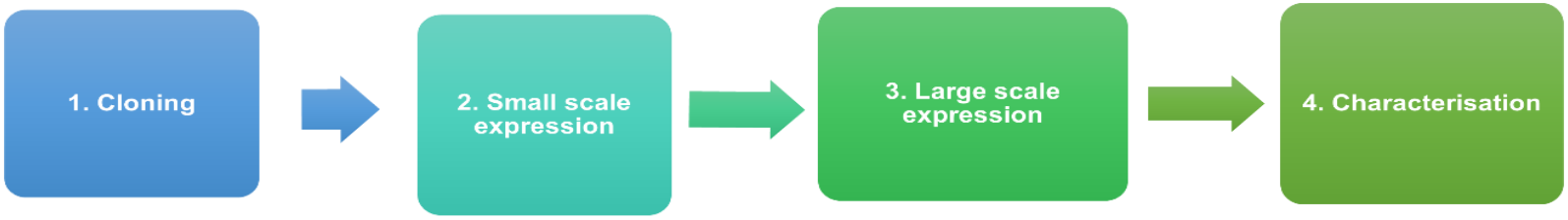


Figure 77: *In vitro* pipeline for this chapter.

5.2 Methods

5.2.1 Molecular Docking

First a suitable computational tool specifically suited to protein-peptide docking had to be identified in the literature from the range of available docking programs now available. Peptides are generally more flexible than proteins and as a result tend to adopt numerous conformations making them more challenging to predict the docking of computationally as it is difficult to sample and define their native conformations [497, 498]. There are several benchmarking studies available on protein-protein, protein-ligand and nucleic acid-ligand docking interactions, but protein-peptide docking methods specifically are not yet so rigorously validated within the literature [499]. **Table 50** shows the tools explored here and the details of why they were excluded or selected. The High ambiguity protein-protein driven docking (HADDOCK) web server was the only such tool identified here suited to protein-peptide docking that accepted the Hyp residue in the TII gelatin peptide.

As mentioned in previous chapters structural tools are considered superior to sequence-based ones and given that a structure for 222 was already attained in **chapter 3.1** only structural tools were explored here. Accuracy of docking methods has been subject to monitoring since 2004 when the critical assessment of predicted interactions (CAPRI) community wide meeting first took place (this meeting occurs every six months). It is a community-wide assessment of docking accuracy, where agreed upon predefined target protein complexes from theoretical docking and X-ray crystallography are compared [500]. A joint assembly collaboration between CAPRI and CASP, CASP-CAPRI challenge, was introduced in 2014. In CASP-CAPRI the

challenge is set to predict protein complex structures starting with just sequences of the individual component proteins rather than their crystal structures, thus requiring the use of homology modelling tools to begin. Both CAPRI and CASP-CAPRI are blind prediction experiments and hence provide unbiased information on the accuracy of docking methods. Recent progress in protein and protein complex structure prediction have necessitated such assessments to enhance integration between the array of scientific disciplines that collaborate to develop such techniques [501].

Here structures (PDB files) modelled for the two docking input molecules were required as docking input; a protein (222, which we modelled in **Chapter 3**) and a peptide (TII gelatin which was modelled here). We also used identified binding residues from NMR data to guide the docking process making it information driven rather than entirely free docking.

Table 50: Available structure-based protein-peptide docking tools explored. Tool, website, type, notes about each and references associated.

Tool	Website	Type	Notes	Ref
Autodock Vina	https://vina.scripps.edu/	Hybrid scoring function (both empirical and knowledge based)	One of the fastest and most widely used molecular docking programs. Didn't give true docking. Allows some limited flexibility of selected receptor side chains. Further explored flexibility parameters but ultimately not selected as a final tool because of this.	[502-504]
Autodock4	https://autodock.scripps.edu/	Flexible, based on AMBER force field scoring function	Better predictor of binding affinity compared to Autodock Vina. Won't accept Hyp as a residue in the peptide	[505-508]
CB dock	https://cadd.labshare.cn/cb-dock2/php/index.php	Cavity-detection guided, Blind Docking	Easy Autodock Vina interface, limited in number of submissions.	[509, 510]
ClusPro	https://cluspro.bu.edu/home.php	Direct docking, Rigid body	Not suited to Hyp, otherwise this was a front runner tool due to its accuracy and user-friendly interface.	[511]
Flexpepdock	http://flexpepdock.furmanlab.cs.huji.ac.il/	Flexible ligand, Rigid receptor	Won't accept Hyp as a residue in the peptide.	[512-514]

FRODOCK (Fast Rotational DOCKing)	http://frodock.chaconlab.org/	Rigid body docking algorithm	Won't accept Hyp as a residue in the peptide.	[515]
Galaxypepdock	https://galaxy.seoklab.org/cgi-bin/submit.cgi?type=PEPDOCK	Similarity-based docking, Energy-based optimization that allows for structural flexibility. Sampling the backbone and side-chain flexibilities of both protein and peptide.	No comparable affinity output would require further evaluation. Finds complexes already deposited in PDB. Won't accept Hyp as a residue in the peptide.	[516]
HADDOCK (High Ambiguity Driven protein-protein Docking)	https://wenmr.science.uu.nl/	Information-driven flexible docking approach	There is no correlation between HADDOCK score output and binding affinity. Indicating worse or improved binding. Another indicative but not quantitative prediction. Can accept Hyp in TII gelatin peptide. Used to generate predicted docked complex here.	[517-519]
HADDOCK refinement interface	https://wenmr.science.uu.nl/haddock2.4/refinement/1	/	Using the interface binding can be compared between substitution mutants introduced in the predicted complex structure. Lower HADDOCK score implies better binding. Although not quantitative in terms of binding affinity does give a comparable relative score measure. This was the approach deemed as most suitable and used in the work here.	[492, 520, 521]
HPepDock	http://huanglab.phys.hust.edu.cn/hpepdock/	A hierarchical flexible peptide docking approach by fast conformational modeling and orientational sampling of peptides.	Won't accept Hyp as a residue in the peptide.	[522, 523]

pepATTRACT	http://bioserv.rpbs.univ-paris-diderot.fr/services/pepATTRACT	Fully blind peptide-protein docking protocol, Flexible	Flexible protein-peptide docking algorithm which performs a rapid coarse-grained global search on the protein surface and model peptide simultaneously during docking. Won't accept Hyp as a residue in the peptide.	[524, 525]
Prodigy (PROtein binDIng enerGY prediction)	https://wenmr.science.uu.nl/prodigy/	Information-driven flexible docking approach	Collection of webservices focused on predicting binding affinity in biological complexes. Won't accept Hyp as a residue in the peptide.	[522]
Pydock	https://life.bsc.es/servlet/pydock/	Rigid body	Unstructured nature of receptor and ligand makes flexible docking more suitable.	[526]
Swarmdock	https://bmm.crick.ac.uk/~svc-bmm-swarmdock/	Flexible	Won't accept Hyp as a residue in the peptide	[527, 528]
Zdock	https://zdock.umassmed.edu/	Fast Fourier Transform based protein docking programs. Searches all possible binding modes in the translational and rotational space between the two proteins and evaluates each pose using an energy-based scoring function.	Unstructured nature of receptor and ligand makes flexible docking more suitable.	[529]

5.2.1.1.1 Peptide generation

In solution peptide conformation will not be fixed, it will be in flux so long as the peptide remains unbound and available [530]. Thus, it's akin to working with a moving target. The only way to deal with this inherent flexibility in peptides was with a conformational library. This was generated by screening multiple peptide conformations, to find a peptide that was representative of the experimental affinity result order attained in **Chapter 3**, using binding order screening as a validation step. If *in silico* binding order was matched with experimental binding order, this suggested that the docked protein-peptide complex was a valid one. When TII collagen is cleaved there are resulting ¼ and ¾ length peptide fragments [42] (shown in **Figure 9**). Here the ¾ length fragment was selected for docking, as the larger fragment available at OA the joint surface, the sequence of the ¾ length fragment is shown in **Table 51**.

Table 51: TII gelatin peptide. Sequence of ¾ length fragment was taken from [42]. The ¾ length peptide segment peptide sequence with proline (P) was first modelled, then manually hydroxylated using the PyTMs, plug in. The three proline residues 9, 15 and 18 that were hydroxylated as per [42] are highlighted in cyan. Below is the final modified peptide with hydroxyproline (Hyp) residues highlighted in yellow. The Hyp residues are known to be critical to binding [531].

Step	Peptide	PTM	Sequence
1	¾ length fragment	X	GKVGPSGAPGENGRPGPPGPQ
2		✓	GKVGPSGAHypGENGRHypGPHypGPQ

This sequence was then used as input to several different tools identified in a literature search (**Table 52**), to generate peptide PDB files to be used as docking input. Python files were used to drive the generation of peptides with two of the tools PeptideBuilder (**Figure 78**)and Fragbuilder (**Figure 79**).

```
import PeptideBuilder
import Bio.PDB

geo = Geometry.geometry("G")
structure = PeptideBuilder.initialize_res(geo)
for i in "KVGPSGAPGENGRPGPPGPQ":
    geo = Geometry.geometry(i)
    PeptideBuilder.add_residue(structure, geo)

out = Bio.PDB.PDBIO()
out.set_structure(structure)
out.save("2020-06-04\\targetnohyp-pep.pdb")
```

Figure 78: Peptidebuilder .py file used to generate TII gelatin peptide (target no hyppep). PeptideBuilder is a Python library for generating peptide PDB files. This .py file is a simple script providing PeptideBuilder with the residues of the TII gelatin sequence. This code iterates through the residues of the sequence adding each to the structure in turn. The final structure and output for this tool is saved as a PDB file. There is no Hyp at this stage as PeptideBuilder does not have the capacity to build peptides with PTMs.

```
from fragbuilder import Peptide

sequence = "GKVGPSGAPGENGRPGPPGPQ"
pep = Peptide(sequence, nterm="neutral", cterm="charged")
pep.write_pdb("2020-06-04\\targetnohypfrag.pdb", QUIET=False)
```

Figure 79: Fragbuilder .py file used to generate TII gelatin peptide (target no hypfrag). FragBuilder is a Python library for generating peptide PDB files. This .py file is a simple script providing Frag Builder with the residues of the TII gelatin sequence, used to generate a PDB output. There is no Hyp at this stage as FragBuilder does not have the capacity to build peptides with PTMs.

The other tools three tools used to generate peptides were web based and only required sequence details as input. ModPepServer was the only one of these

tools that could not modify Pro to Hyp residues automatically itself. Where post translational modification of proline to Hyp residues was required (labelled as manual in **Table 52**), a pymol plugin named PyTMs was utilised to modify Pro 9, 15 and 18 to Hyp [440]. Vienna-PTM was another method of modifying Pro residues that was utilised with the ten ModPepServer peptides selected via clustering from the 100 (Clustering is explained in the next **methods section 5.2.1.1.2**)**.** With the last tool identified, PEPstrMOD the non-natural residue module for experts was used. This was the only tool where a sequence was entered, and hydroxylation modifications specified at the outset before generation.

Table 52: Different tools used to generate peptides. Type, number of models generated, Hyp addition, website, and ref.

Tool	PeptideBuilder	FragBuilder	ModPepServer	PEPstrMOD	Vienna-PTM
Type	Manual	Manual	Manual	Automatic	Automatic
Number of models generated	1	1	100	1	10
Hyp addition	Manual, Pytms [532]	Manual, Pytms [532]	Manual, Pytms [532]	Automatic	Requires no Hyp PDB input file then modifies selected Pro residues to Hyp (used on the 10 generated manual peptides from ModPepServer)
Website	https://pypi.org/project/PeptideBuilder/	https://github.com/jensengroup/fragbuilder/	http://huanglab.phys.hust.edu.cn/modpep/	http://osddlinux.osdd.net/raghava/pepstrmod/	http://vienna-ptm.univie.ac.at/
Ref	[533]	[534]	[535]	[536]	[537]

5.2.1.1.2 Peptide clustering

Ideally all generated 113 peptide options would have been tested, however this would be computationally intensive and laborious. Instead, clustering was utilised here to reduce the number of peptide options in a systematic manner to enable trialling of a representative sample of peptides. Maxcluster is a command-line tool facilitating easy computational comparison of protein structures. It is a simple high-throughput interface enabling many common structure comparison tasks to be completed easily (either against a single reference protein or in an all-verses-all comparative approach). It was used here to cluster the 100 peptides produced using the Modpepserver (available at: http://www.sbg.bio.ic.ac.uk/maxcluster/index.htmlhttp://www.sbg.bio.ic.ac.uk/maxcluster/index.html accessed 03/20) [538], which was the only peptide generation method that produced such a large number of conformers necessitating clustering. **Figure 80** show the parameters that were used to execute clustering of the 100 ModPep peptides, all but RMSD were the Maxcluster defaults, in an all vs all Nearest Neighbour (NN) clustering comparison. Structures were considered matched if they were within a distance threshold of 4 Ångströms (Å), shared a minimum of 20 pairs of residues and a MaxSub score above 0.2.

Using the same defaults and clustering parameters a second comparison of the other 13 peptides generated by the other servers was also made to elucidate if there was any agreement between the servers, regarding conformation and identify any duplicates.

```
# Maxcluster 0.6.6

.\maxcluster.exe `
    -l states.list ` # Load the list of 100 peptides for all-vs-all
    -rmsd ` # Perform only RMSD fit
    -log all-vs-all.txt ` # Log results to file
    # Clustering options - all defaults
    -C 5 ` # Cluster method 5 = Neighbour pairs (absolute size)
    -T 4 ` # Initial clustering threshold
    -Tm 8 ` # Maximum clustering threshold
    -a 0.2 ` # Clustering threshold adjustment
    -is 50 ` # Initial cluster size
    -ms 5 ` # Minimum cluster size
    -s 0.2 ` # 3D-jury score threshold
    -P 20 # 3D-jury pair threshold
```

Figure 80: Parameters used to execute Maxcluster to 'cluster' and select from the 100 Modpep TII gelatin peptides. The all vs all approach was used which compared each peptide to all the other 99, nearest neighbour clustering whereby two proteins are considered part of the same cluster if they are closer than a cut-off threshold of 4Å, that is they are near-neighbours [539].

5.2.1.1.3 Assessing biophysical/ spatial validity of modelled peptides

Following clustering, one final check of all the selected modelled peptides (listed in **Table 53**), was undertaken reviewing their plausibility and validity before moving on to the actual docking assessment (with a confirmed representative conformer library).

Table 53: Modelled peptide conformers taken to MolProbity assessment. Hyp modification method used to sub categorise the eleven peptide into two groupings automatic or manual.

Hyp method	Peptide conformer
Automatic	Vienna 12
	Vienna 16
	Vienna 76
	Vienna 77
	Vienna Frag
	Vienna Pep
Manual	16
	40
	76
	Frag
	Pep

This was particularly necessary for the peptide models for which proline (Pro) residues were modified manually in pymol as the Hyp may have different conformational properties to the modelled Pro containing peptides. MolProbity a structure validation webservice for diagnosing problems in 3D models of proteins and peptides (available at: http://molprobity.biochem.duke.edu/index.phphttp://molprobity.biochem.duke.edu/index.php, accessed, 02/20) was used for this. PDB files were submitted to the MolProbity which gave a variety of scores and output. Of particular note is the MolProbity score which combines Clashscore, rotamer and Ramachandran evaluation into a single score [540].

5.2.1.2 HADDOCK

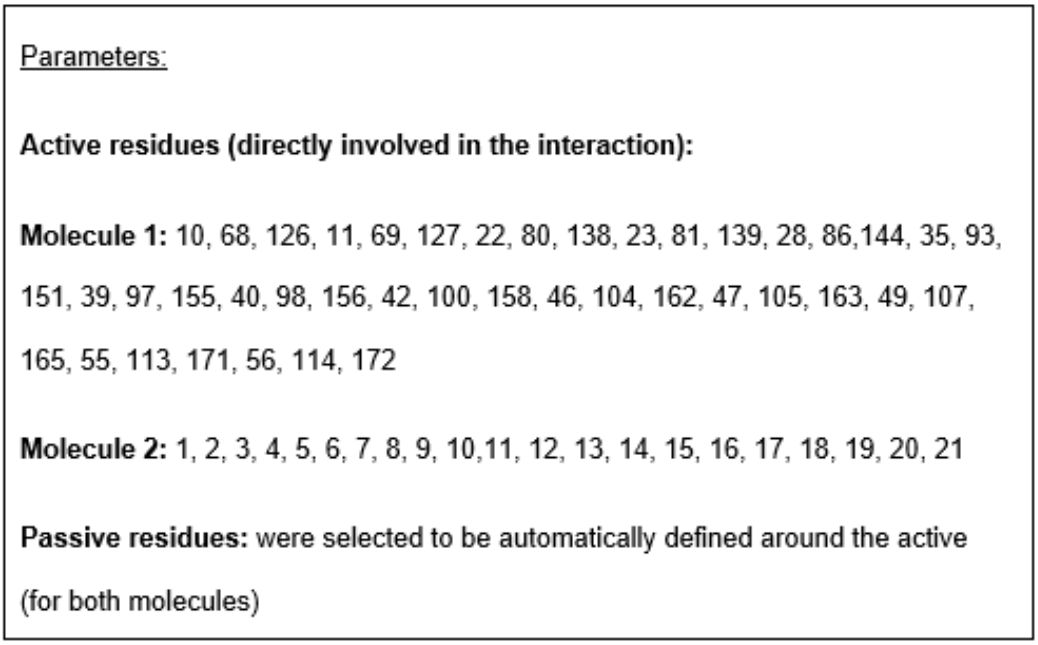
5.2.1.2.1 HADDOCK tool input

Peptides selected via both clustering and MolProbity assessment were then fed into the HADDOCK 2.2 expert interface to validate the *in silico* binding assessment using the already attained experimental binding order of 222, M1 and M2. These three were used in this assessment as they were the three most different in terms of confirmed binding affinity. HADDOCK 2.2 [518] was selected as the only tool identified in the literature search **Table 50**, suited to providing a docked complex model between the two biomolecules here namely the 222 protein and Hyp containing TII gelatin peptide. The HADDOCK output was the initial docking model in the form or a downloadable pdb file.

Input data

The 222 pdb file (molecule 1) modelled in chapter 3, and peptide conformer pdb files (molecule 2) generated in this chapter using several tools and selected through clustering **methods section 5.2.1.1.2** and then MolProbity assessment **methods section 5.2.1.1.3**. Selected peptides were uploaded as molecule 2, one conformer at a time, with 222 uploaded as molecule 1 for each. In the input menu molecule 1 was defined as a protein, then molecule 2 defined as a peptide. The C-terminus of the peptide was specified as negatively charged. All other options in the input data were left as default.

An Ambiguous interaction restraints (AIR) list was required in the input of the HADDOCK tool (**Figure 81**). The AIR input makes the docking not an entirely blind, but information driven process. Default setting to randomly exclude a fraction of the AIRs was turned off as the NMR chemical shift data used to identify binding residues was conclusive that all these residues were involved so to randomly exclude some would be incorrect



Parameters:

Active residues (directly involved in the interaction):

Molecule 1: 10, 68, 126, 11, 69, 127, 22, 80, 138, 23, 81, 139, 28, 86,144, 35, 93, 151, 39, 97, 155, 40, 98, 156, 42, 100, 158, 46, 104, 162, 47, 105, 163, 49, 107, 165, 55, 113, 171, 56, 114, 172

Molecule 2: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21

Passive residues: were selected to be automatically defined around the active (for both molecules)

Figure 81: AIR list, HADDOCK input. This AIR list composed of the NMR identified residues involved in the binding of 222 to TII gelatin (taken from the groups prior work [235]). This is what makes that the docking data-driven, by providing the list of protein binding residues we are saying we know where the binding is going to happen. We didn't have any information regarding the residues in the peptide involved in binding, so all 21 residues were selected as involved/ active.

.

5.2.1.2.2 HADDOCK docking steps

The HADDOCK score is calculated from a linear combination of energies and buried surface area (BSA). The scoring is performed according to a weighted sum (HADDOCK score) for the following terms:

- Evdw: van der Waals intermolecular energy
- Eelec: electrostatic intermolecular energy
- Eair: distance restraints energy (only unambiguous and AIR (ambig) restraints)
- Erg: radius of gyration restraint energy
- Esani: direct RDC restraint energy
- Evean: intervector projection angle restraints energy
- Epcs: pseudo contact shift restraint energy
- Edani: diffusion anisotropy energy
- Ecdih: dihedral angle restraints energy
- Esym: symmetry restraints energy (NCS and C2/C3/C5 terms)
- BSA: buried surface area
- dEint: binding energy (Etotal complex – Sum [Etotal components])
- Edesol: desolvation energy

The default scoring function settings of HADDOCK used here were for protein-protein complexes and implement the following weights shown in **Equation [11], Equation [12] and Equation [13]**.

$$\boldsymbol{HADDOCKscore_{it0} = 0.01\,E_{vdw} + 1.0E_{elec} + 1.0E_{desol} + 0.01E_{air} - 0.01BSA} \quad \textbf{[11]}$$

Where $it0$ refers to the rigid body docking stage, E_{vdw} is the van der Waals intermolecular energy, E_{elec} is the electrostatic intermolecular energy, E_{desol} is the desolvation energy, E_{air} is the distance restraints energy (only unambiguous and AIR (ambig) restraints) and BSA is the buried surface area.

$$\boldsymbol{HADDOCKscore_{it1} = 1.0\,E_{vdw} + 1.0E_{elec} + 1.0E_{desol} + 0.1E_{air} - 0.01BSA} \quad \textbf{[12]}$$

Where $it1$ refers to the semi-flexible refinement docking stage

$$\boldsymbol{HADDOCKscore_{water} = 1.0\,E_{vdw} + 0.2E_{elec} + 1.0E_{desol} + 0.1E_{air}} \quad \textbf{[13]}$$

Where $\boldsymbol{water}$ refers to the explicit solvent refinement docking stage

5.2.1.2.3 Docking method validation

First peptides were docked with 222, then the HADDOCK refinement interface was used to give comparable HADDOCK scores for M1 and M2 (two most reduced binding affinity mutants determined in **chapter 3** with Kd values and binding order shown in **Table 54**). Validation was carried out using this binding order data with lowest HADDOCK score matching the best binding protein (222), M1 the highest score and M2 the intermediate score. If a peptide matched this ordering in HADDOCK score it was deemed a valid and experimentally representative peptide to take forward. The docking complex 222 output/ results from any such validated docking experiment was then taken in PDB format and used in subsequent *in silico* mutant assessment discussed in the next section.

Table 54: Experimental binding results to match with *in silico* strategy.

Protein	Kd (nM)	Binding order
222	3.9	1
M1	1005	3
M2	549.4	2

5.2.1.2.4 Docked complex mutagenesis, refinement, mutant assessment

The strategy utilised here involved first using HADDOCK to generate a 222 docked complex which was downloaded as the best docked prediction (**5.2.1.2.2** and **5.3.3**) HADDOCK complex (pdb). Then using the pymol mutagenesis wizard substitutions were introduced into this generated complex in place of residue 11, 69 and 127, to all eighteen possible alternative residues (**Table 55**). Discounting Asn which is the native identified residue in this position in 222 and Ala which is the residue substitution in M1 that reduced binding and was already assessed in the previous validation step where we sought matched binding order. Each of the generated mutant complexes were then submitted to the refinement interface to assess mutants with all possible amino acid variations to see if any M1 mutants had a stronger interaction indicated by HADDOCK score. The lower the HADDOCK score the stronger the predicted binding interaction.

Table 55: Twenty amino acid details. Amino acid, abbreviation, single letter abbreviation and corresponding protein be it 222 or M1 (from chapter 3) and mutant ref with these residues in position 11, 69 and 127. Properties of the amino acids , * indicates a special case grouping. Specifically: Cysteine which has a reactive sulfhydryl R group that forms disulfide bridges (S-S) between regions of the protein chain [412]. Glycine is the smallest amino acid, with hydrogen only as its R group, so it fits into tight places within a protein's structure [541]. Proline has a cyclic ring involving the central carbon, which causes kinks to occur in a protein chain [542]. Both proline and glycine are common at the corner of turns in the protein folding [543].

Amino Acid	Abbreviation	Single letter abbreviation (Mutant ref)	Properties
Asparagine	Asn	N (222)	Polar, uncharged
Alanine	Ala	A (M1)	Aliphatic, hydrophobic, smallest amino acid
Arginine	Arg	R (222R)	Positive
Aspartic acid	Asp	D (222D)	Negative
Cysteine	Cys	C (222C)	Reactive sulfhydryl group, *
Glutamine	Gln	Q (222Q)	Polar, uncharged
Glutamic acid	Glu	E (222E)	Negative
Glycine	Gly	G (222G)	Aliphatic, * smallest amino acid with H as its R group
Histidine	His	H (222H)	Positive
Isoleucine	Ile	I (222I)	Aliphatic, hydrophobic
Leucine	Leu	L (222L)	Aliphatic, hydrophobic
Lysine	Lys	K (222K)	Positive
Methionine	Met	M (222M)	Hydrophobic
Phenylalanine	Phe	F (222F)	Aromatic hydrophobic,
Proline	Pro	P (222P)	Aliphatic, * has a cyclic ring
Serine	Ser	S (222S)	Polar, uncharged
Threonine	Thr	T (222T)	Polar, uncharged
Tryptophan	Trp	W (222W)	Aromatic, hydrophobic,
Tyrosine	Tyr	Y (222Y)	Aromatic, hydrophobic,
Valine	Val	V (222V)	Aliphatic, hydrophobic

Following the results in **chapter 4** that solubility was enhanced in CS6, but binding affinity was markedly reduced HADDOCK was also used to evaluate CamSol substitution mutants (which were CS1 and CS2 only) to see how binding was predicted to differ compared to 222. As with the M1 mutants, substitutions of CS1 and CS2 were introduced concomitantly into the docked 222 complex (pdb) along with any identified as better binding (having the lowest HADDOCK score from the

HADDOCK refinement assessment of the eighteen in **Table 55**). Combining any HADDOCK predicted improved binding only mutants with both CS1 and CS2, was trialled to see if solubility and binding could be balanced and improved concomitantly.

5.2.2 Expression

The codon optimised (**methods section 3.2.10.1**) mutant genes ordered for the work in this chapter are shown in **Table 56**.

Table 56: GeneOptimizer codon optimised DNA sequence for 222W and 222W-CS1. Nucleotide codons highlighted in yellow are those that encode the mutations.

Name	DNA Sequence
222W	GAAGGTCAGGTTGTGTTTACCATGTATGGTTGGGCTGAAGGTCAGCCGTGTAAATTTCCGTTTCGTTTTCAGGGCACCAGCTATGATAGTTGTA CCACCGAAGGTCGTACCGATGGTTATCGTTGGTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCCC TGTTTACAATGGGTGGCTGGGCAGAGGGCCAGCCTTGCAAATTCCCTTTTCGCTTCCAGGGTACATCTTATGATTCATGCACAACGGAAGGTCG CACAGATGGCTACCGCTGGTGCGGCACCACAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACCGCACTGTTCACCATGGG AGGTTGGGCAGAAGGACAACCCTGCAAGTTTCCATTCCGCTTTCAGGGAACCTCATATGATAGCTGCACAACAGAGGGACGTACGGATGGAT ACAGATGGTGCGGTACAACCGAGGACTACGATAGAGATAAGAAATATGGTTTCTGTCCCGATCAGGGTTATAGCCTG
222W-CS1	GAAGGTCAGGTTGTTTTTACCATGGAAGGTTGGGCTGAAGGTCAGCCGTGTAAATTTCCGTTTCGTTTTCAGGGCACCAGCTATGATAGTTGTA CCACCGAAGGTCGTACCGATGGTTATCGTTGGTGTGGTACGACCGAAGATTATGATCGTGATAAAAAGTATGGCTTTTGTCCGCATGAAGCAC TGTTTACAATGGGTGGTTGGGCAGAGGGCCAGCCTTGCAAATTCCCTTTTCGCTTCCAGGGTACATCTTATGATTCATGCACAACGGAAGGTCG CACAGATGGCTACCGCTGGTGCGGCACCACAGAGGATTATGACCGCGACAAAAAATACGGTTTTTGTCCGGAAACAGCCCTGTTCACCATGGG AGGCTGGGCAGAAGGACAACCCTGCAAGTTTCCATTCCGCTTTCAGGGAACCTCATATGATAGCTGCACAACAGAGGGACGTACGGATGGAT ACAGATGGTGCGGTACAACCGAGGACTACGATAGAGATAAGAAATATGGTTTCTGTCCCGATCAGGGTTATAGCCTG

Table 57 shows the protein sequences for the two mutants 222W and 222W-CS1, that were selected using HADDOCK score and taken forward to *in vitro* testing.

Table 57: Protein sequences for this chapter. Protein names, mutation sites stating the residue in 222 the 'native' and the numbered position of each mutation site, as well as the protein sequences. The residues shown in blue text were left as in the native CBD protein, as they had critical involvement in intramolecular interactions. The three modules for each protein are shown in red text, and the linker regions in black text. Tryptophan (Trp, W) substitution sites for each mutant are highlighted in yellow.

Name	Mutation site	Protein Sequence	Extinction coefficient	Abs 0.1% (=1g/l)	Molecular weight (kDa)
222W	N11W N69W N127W	EGQVVFTMYGWAEGQPCKFPF RFQGTSYDSCTTEGRTDGYRW CGTTEDYDRDKKYGFCPHEALF TMGGWAEGQPCKFPFRFQGTS YDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPETALFTMGGW AEGQPCKFPFRFQGTSYDSCTT EGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	54610	2.617	20.86
222W-CS1	N11W N69W N127W Y9E	EGQVVFTMEGWAEGQPCKFPF RFQGTSYDSCTTEGRTDGYRW CGTTEDYDRDKKYGFCPHEALF TMGGWAEGQPCKFPFRFQGTS YDSCTTEGRTDGYRWCGTTED YDRDKKYGFCPETALFTMGGW AEGQPCKFPFRFQGTSYDSCTT EGRTDGYRWCGTTEDYDRDKK YGFCPDQGYSL	53120	2.550	20.83

Expression of 222W and 222W-CS1 proceeded mostly as previously outlined in **methods section 3.2.10** but with a few differences (**Figure 82 & Figure 83**). The genes were cloned into three pOPIN vectors; pOPINS, pOPINM and pOPINJ. Additionally, the genes were cloned into the pMAL-p5X vector to allow periplasmic expression trialling to see if this proved an easier method of expression and recovery. All genes were cloned using the ligase independent In-Fusion cloning methodology described in full detail within **methods section 3.2.10.6**. Stellar competent cells (Takara Bio) were transformed via heat shock with the mutant construct plasmids **(methods section 3.2.10.7)**. Stellar cells were used as a cloning strain to produce plasmid for isolation, via QIAprep miniprep kit (Qiagen), **(Materials and methods section 2.10)**. Before proceeding to expression trials all constructs were confirmed via sanger sequencing (Eurofins GATC, Germany).

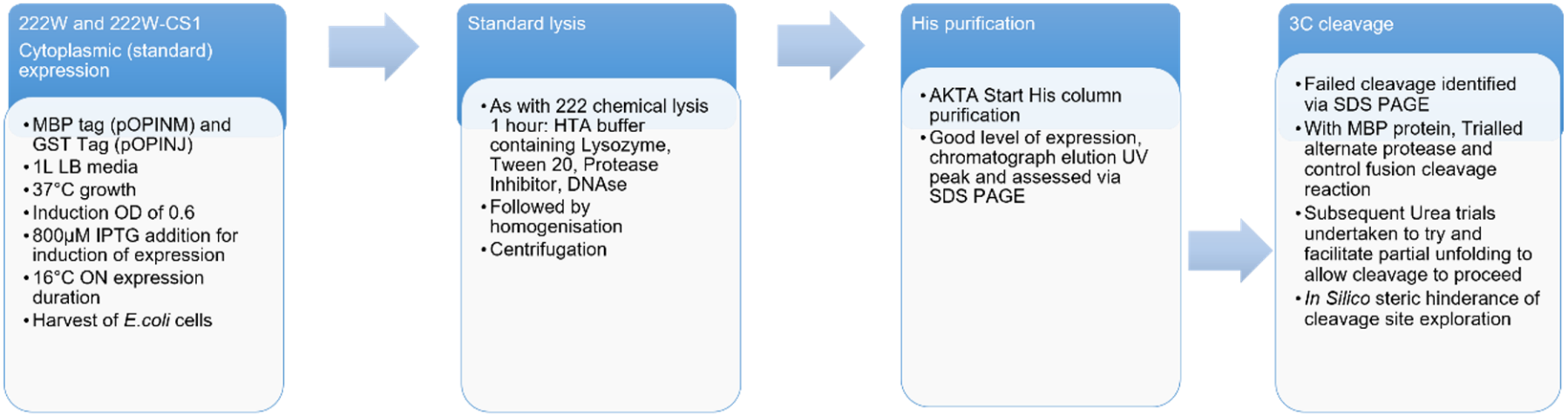


Figure 82: Overview of Cytoplasmic expression and purification conditions trialled with 222W and 222W-CS1. Standard cytoplasmic expression utilised with 1L of LB. Lysis proceeded via homogenisation using the continuous Flow CF1Cell Disrupter (Constant systems) as described in **methods section** 3.2.10.14. His purification proceeded (described in **methods section** 3.2.10.15), reverse His purification did proceed but was the purification end here due to cleavage failure, SDS-PAGE analysis (described in **methods section** 2.6) of the eluted fractions from His elution showed a good level of soluble fusion protein (His-MBP-3C-POI), PDS and Rev his collected samples (RFT and ER) indicated that cleavage had failed, with tagged protein retained wholly in the ER rather than the RFT where POI alone would be expected. Following this failure to cleave a control reaction and repeat cleavage attempt was made to no avail. Urea addition was also trialled to facilitate partial unfolding again this was unsuccessful. The fusion was then explored *in silico* for steric hinderance and inaccessibility of the cleavage site**.**

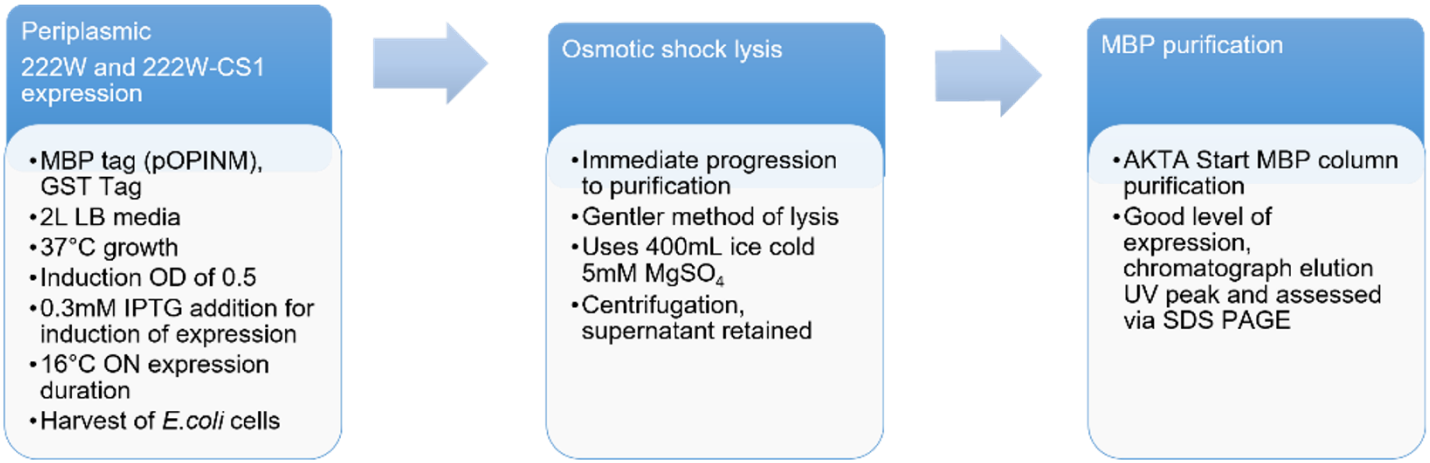


Figure 83: Overview of periplasmic expression and purification conditions trialled large scale with 222W and 222W-CS1. Periplasmic expression was trialled here with 2L LB, lysis proceeded via osmotic shock to see if it was an easier way of acquiring correctly folded soluble POI. A repeat was planned to refine protocol and progress to cleavage, but other experiments detracted from this, and cleavage was never trialled with the periplasmic expressed 222W and 222W-CS1. Given that periplasmic purification via osmotic shock was a difficult to scale and employ method and the time constraints of the project. Along with consecutive *in silico* exploration of steric hinderance and occlusion of the 3C cleavage site confirming why an MBP-3C-POI fusion would not cleave and lack of success with Urea cleavage trials with the alternate cytoplasmically expressed His-MBP-3C-POI fusion.

5.2.2.1 Primer design

Forward (fwd) and reverse (rev) primers (**Table 58**) were designed, ordered from sigma, and used to subclone each gene into the three pOPIN vectors and the additional pMAL-p5x vectors, via Infusion cloning (Takara Bio) as described in **methods section 3.2.10.6**.

Table 58: Infusion Cloning primers. Insert and vector specific portions are shown; vector region (black text) and insert region (red text), 3C protease cleavage site (green text), stop codon (Red text highlighted in yellow), positioned immediately after the insert so there will not be extra amino acids on the POI.

Mutant	Vector	Forward (fwd) Primer	Reverse (rev) Primer
222W	pOPINS	GCGAACAGATCGGTGGTGAAGGTCA GGTTGTGTTTACCATG	ATGGTCTAGAAAGCTTTACAG GCTATAACCCTGATCGGG
	pOPINJ	AAGTTCTGTTTCAGGGCCCGGAAGGT CAGGTTGTGTTTACCATG	
	pOPINM		
	pMAL-p5X	**TGTCCATGGGCGGCCGC**CTTGAAGTT CTTTTTCAAGGTCCTGAAGGTCAGGT TGTGTTTACCATG	**TACCTGCAGGGAATTC**TTATT ACAGGCTATAACCCTGATCG GG
222W-CS1	pOPINS	GCGAACAGATCGGTGGTGAAGGTCA GGTTGTTTTTACCATG	ATGGTCTAGAAAGCTTTACAG GCTATAACCCTGATCGGG
	pOPINJ	AAGTTCTGTTTCAGGGCCCGGAAGGT CAGGTTGTTTTTACCATG	
	pOPINM		
	pMAL-p5X	**TGTCCATGGGCGGCCGC**CTTGAAGTT CTTTTTCAAGGTCCTGAAGGTCAGGT TGTTTTTACCATGG	**TACCTGCAGGGAATTC**TTATT ACAGGCTATAACCCTGATCG GG

5.2.2.2 Vector linearisation

The OPPF-UK pOPIN vector suite [353] was utilised again in this chapter allowing easy trialling of three different fusion tags as first utilised with CS6 in chapter 4 [460, 461]; (SUMO, GST and MBP). All three were trialled concomitantly for soluble expression and sufficient recovery for the experiments in this chapter. The cleavage site for all these tags was positioned following the fusion tag, resulting in POI only after tag cleavage. The vector map for the pOPINS vector, used in all chapters of this thesis shown in **Figure 22** and **Table 12** provides further details of this specific vector. The vector maps for pOPINJ and POPINM first introduced in chapter 5, are shown in **Figure 57** and **Table 37**. A periplasmic expression vector was also utilised in the work of this chapter as another alternate soluble protein expression and recovery promoting strategy. In the pMAL-p5X vector the signal sequence of the malE gene allows fusion proteins to be exported to the periplasm. The periplasm is a preferential oxidising folding environment for proteins with disulfide bonds [544]. The vector map for the pMAL-p5X vector utilised in this periplasmic expression strategy is shown in **Figure 84**.

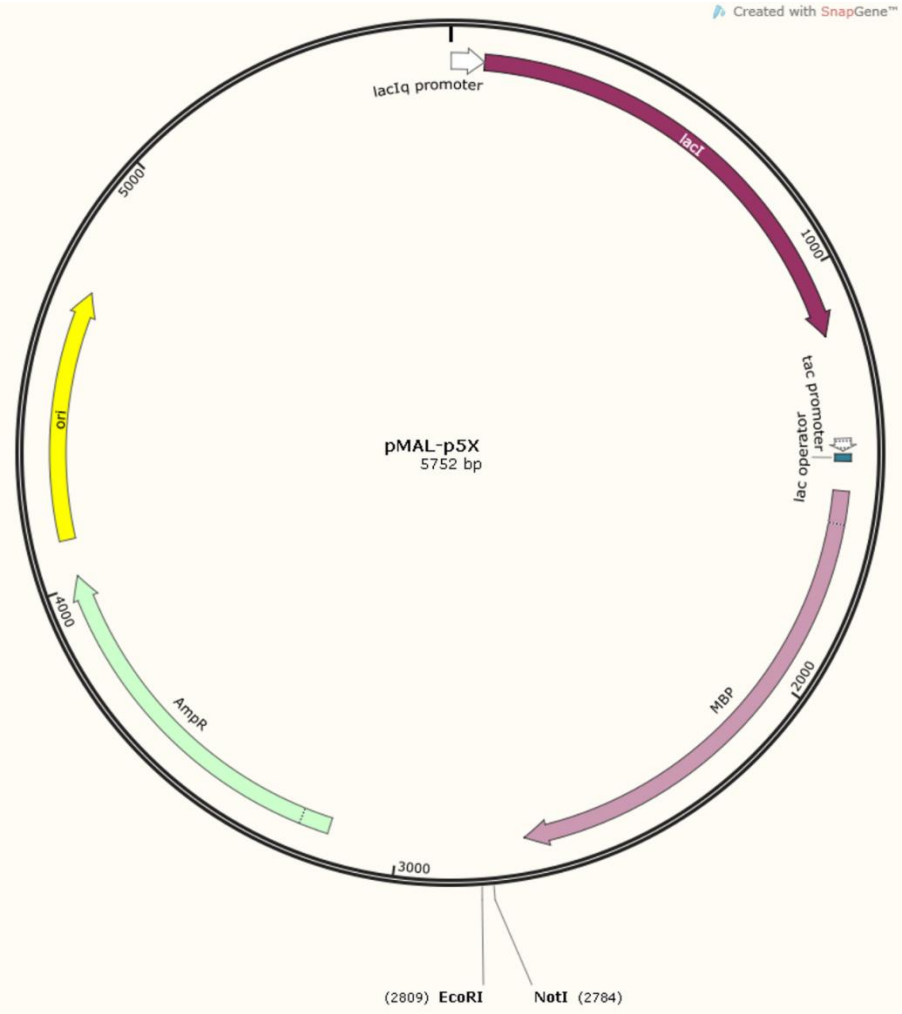


Figure 84: pMAL-p5x vector used in the work in this chapter. pMAL-p5X 5752bp, ampicillin resistance, lacIq promoter Linearised by EcoRI recognition sequence GAATTC and NotI recognition sequence GCGGCCGC. Digest products were a 25bp cut out region and 5727bp linearised vector. The parent vector of pMAl-p5X is pUC18. This figure was created using the SnapGene software (from Insightful Science; available at www.snapgene.com).

Table 59 below gives further details of this periplasmic expression vector pMAL-p5X introduced in this chapter.

Table 59: pMAL-p5X vector details. Including source, restriction enzymes and recognition sites, parent vector/ antibiotic resistance, digest products, digest products, promoter, Inducer, expression product, cleavage enzyme and references.

Vector	Source	Restriction enzyme 1 and recognition site	Restriction enzyme 2 and recognition site	Parent vector/ Antibiotic resistance	Digest products	Promoter	Inducer		Expression product	Cleavage enzyme	Ref
pMAL-p5X	NEB	NotI GCGGCCGC	EcoRI GAATTC	pUC18/Amp	25bp cut out 5727bp linear vector	Tac	IPTG		MBP-3C-POI	3C protease	[545]

Vector linearisation was carried out using two restriction endonucleases, in a double digest removing an unrequired 25bp segment and linearizing the vectors in the process (**Figure 85**).

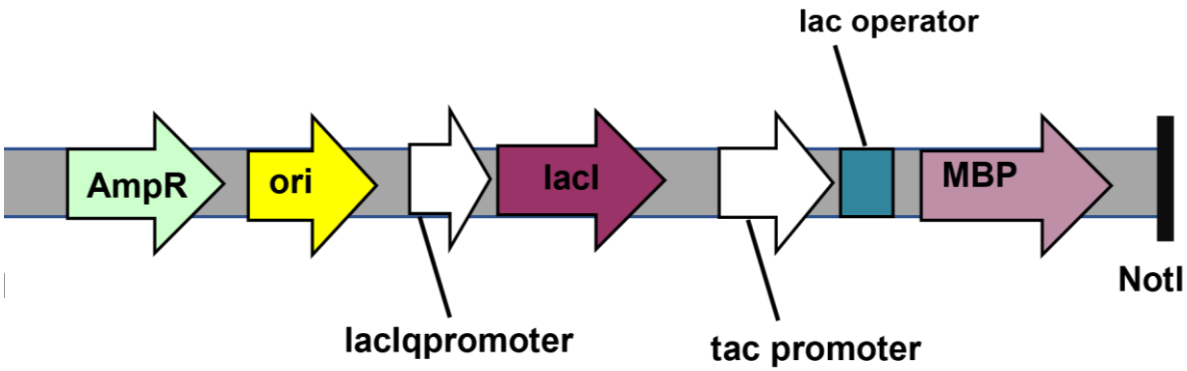


Figure 85: Linear pMAL-p5X vector. 25bp fragment removed during linearisation 5727bp linear fragment. Amp Resistance, required for transformed clone screening. Tac promoter required for expression of insert/ POI; expression inducible by IPTG.

5.2.2.3 Small scale expression & purification

Small scale expression was undertaken initially for all three pOPIN vector as previously outlined in**, Methods section 3.2.10**. Small scale purification was undertaken as preciously outlined in **Methods section 3.2.6.10.** Giving an indicative assessment of soluble (sup) and insoluble (pel) protein product. No nickel resin small scale purification was undertaken, purification ended with the separation of soluble from insol via centrifugation following lysis.

5.2.2.4 Large scale expression & purification

Large scale expression was undertaken as previously outlined in **Methods section 3.2.10.12** but with 1mL Ampicillin (50mg/mL) instead of kanamycin in the pOPINM and pOPIN J cultures. For pOPINM and pOPINJ cultures purification was

undertaken as previously outlined in **Methods section 3.2.10.13**. All buffers used in purification were prepared fresh, filtered through 0.2µm filter (Millipore) and degassed prior to use on the ÄKTA system.

5.2.2.4.1 Lysis

Lysis proceeded again as described previously in **methods section 3.2.10.14**

5.2.2.4.2 His purification

222W and 222W-CS1 expressed using the MBP tag was purified using a 5mL nickel column (Cytiva) as a first step, using His affinity chromatography on the ÄKTA start purification system (Cytiva). The His purification (described in **methods section 3.2.10.15**) was then followed by 3C protease cleavage to remove the His-MBP tag from the POI. POI without tag should habe then been recovered using a reverse His purification step (outlined in **methods 3.2.10.15**).

5.2.2.4.3 Cleavage trials

All expression and purification attempts utilised fusion tags to achieve soluble 222W and 222W-CS1 protein. Protein was acquired in soluble form in the MBP (pOPINM) large scale expression. Unfortunately, cleavage was subsequently unsuccessful, so repeat cleavage and subsequent cleavage trials were undertaken.

2M urea was added to the pooled elution reverse (ER) reverse His purification product containing tagged/ uncleaved POI expressed from large scale His purification (pOPINM), this was left rocking on ice for 1 hour. This was undertaken

with both 222W and 222W-CS1, to partially unfold/ alter the fusion protein conformation to make the cleavage site more accessible. Then 1mL of 3C protease (2mg/mL) was added and left O/N rocking at 4°C, prior to reverse His purification as outlined in **methods section 3.2.10.15**.

5.2.2.4.3.1 *In silico* cleavage site accessibility assessment

The first step to exploring *in silico,* the failure of fusion tag cleavage seen *in vitro* with proteins in this chapter here, was modelling the entire fusion tagged mutant proteins. The pOPINM vector constructs used here resulted in the production of POI with a fusion tag (His-MBP-3C-POI product). 222W, 222W-CS1 and CS6 fusion proteins were modelled using Alphafold2 [392, 393, 546] available at: https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb (accessed 08/22). Colab is a Google service that hosts the prewritten Python program, AlphaFold2 to execute the code in the cloud/online [547]. Sequences shown in **Table 60** were submitted with default settings. The unrelaxed rank 1 model 1 pdb output was taken as the best model. PLDDT score (predicted local distance difference test score, lDDT-Cα [548]), a per-residue measure of local confidence on a scale from 0–100 was provided with the generated model output giving an indication of how good AlphaFold2 considered the models were. Models with a PLDDT ≥70 are considered confident predictions therefore this was the threshold criteria used here [549]. Template modelling (TM) score and predicted TM (PTM) score was the second output given as an indication of model confidence(always lies between 0-1) with better templates having higher TM-scores, here a value of ≥0.5 was sought [550].

Table 60: Expression fusion product summary. Each protein was expressed using the pOPINM vector which means the expression fusion product was a His-MBP-3C-POI (3C highlighted in pink). Protein sequences for His tag and its linker are shown in blue text, MBP tag is shown in magenta text, 3C cleavage site is underlined and shown in red text: LEVLFQG/P (/ indicates the point of cleavage) and shown in green text is the POI 222W, 222W-CS1 or CS6 (with mutations highlighted in yellow). Molecular weight of entire fusion product shown, then His-MBP tag + POI.

Expression fusion product	Protein Sequence	Molecular weight (kDa)
His-MBP-3C-222W	HHHHHHSSGMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWA HDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDK ELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTD YSIAEAAFNKGETAMTI111NGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFL ENYLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASG RQTVDEALKDAQTSSGLEVLFQGP/EGQVVFTMYGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGT TEDYDRDKKYGFCPHEALFTMGGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYG FCPETALFTMGGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPDQGYSL	63.36 42.49+20.87
His-MBP-3C-222WCS1	HHHHHHSSGMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWA HDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDK ELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTD YSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLEN YLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQ TVDEALKDAQTSSGLEVLFQGP/EGQVVFTMEGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTE DYDRDKKYGFCPHEALFTMGGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFC PETALFTMGGWAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPDQGYSL	63.32 42.49+20.83
His-MBP-3C-CS6	HHHHHHSSGMKIEEGKLVIWINGDKGYNGLAEVGKKFEKDTGIKVTVEHPDKLEEKFPQVAATGDGPDIIFWA HDRFGGYAQSGLLAEITPDKAFQDKLYPFTWDAVRYNGKLIAYPIAVEALSLIYNKDLLPNPPKTWEEIPALDK ELKAKGKSALMFNLQEPYFTWPLIAADGGYAFKYENGKYDIKDVGVDNAGAKAGLTFLVDLIKNKHMNADTD YSIAEAAFNKGETAMTINGPWAWSNIDTSKVNYGVTVLPTFKGQPSKPFVGVLSAGINAASPNKELAKEFLEN YLLTDEGLEAVNKDKPLGAVALKSYEEELAKDPRIAATMENAQKGEIMPNIPQMSAFWYAVRTAVINAASGRQ TVDEALKDAQTSSGLEVLFQGP/EGQEVETMEGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTE DYDRDKKYGFCPHEALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGEEEYRWCGTTEDYDRDKKY GFCPETALFTMGGNAEGQPCKFPFRFQGTSYDSCTTEGRTDGYRWCGTTEDYDRDKKYGFCPDQGYSL	63.51 42.49+21.02

Models were downloaded in PDB format, then used in bioinformatic assessment of solvent accessibility, using two contrasting tools freeSASA [551] and pymol [552]. At this stage it was hypothesised that the two uncleavable fusion proteins (222W and 222W-CS1) had steric hinderance/ occlusion of the 3C protease cleavage sequence, LEVLFQGP [553, 554]. If occlusion was the issue a lower solvent accessible surface area (SASA) would be seen for the two uncleavable variants.

5.2.2.5 Periplasmic

Periplasmic expression offers several advantages over cytoplasmic expression; it helps to reduce proteolytic degradation, provides a favorable environment for disulfide bond formation and proper protein folding [555, 556].

5.2.2.5.1 Periplasmic expression

2L of LB was inoculated with 20mL of starter culture/ 1L LB flask, 1mL of Ampicillin (50mg/mL), 0.2% glucose (necessary in the growth medium to repress the maltose genes on the chromosome of the *E. coli* host, one of which is an amylase which can degrade amylose in the MBP affinity column utilised in the subsequent purification steps) and grown at 37°C, with 180RPM shaking until OD 0.5 was attained. Cultures were then induced with 0.3mM IPTG and grown O/N at 16°C. This was done for both different *E. coli* strains listed in **Table 61**.

Table 61: E. coli strains utilised in periplasmic expression trials. Strains, rationale for use and references

***E. coli* strains**	**Rationale**	**Ref**
C41	BL21 Walker strain derivative that is tolerant of toxic proteins, been shown to produce more protein consistently compared to BL21 could be due to a larger periplasmic space.	[557-559]
Lemo21	Tunable overexpression with L-Rhamnose. Good for proteins prone to insoluble expression and/ or toxic to the *E.coli* cells.	[558]

Cells were then harvested by centrifugation at 8000RPM for 20 mins, supernatant discarded and proceeded directly to purification (important that no freeze thaw because of storage was incurred).

5.2.2.5.2 Periplasmic purification

An osmotic shock protocol was used immediately following harvest to isolate proteins from the periplasm. Osmotic shock is a gentler method than other more traditional lysis methods such as homogenisation or sonication. This gentler technique leaves the cytoplasmic space intact keeping host cytoplasmic proteins separate from periplasmic.

Pelleted cells were resuspended in 400mL Buffer P (30 mM Tris-HCl, 20% sucrose, pH 8.0 (80mL/each gram of cells wet pellet weight). EDTA was added to

1mM, and the suspension incubated for 5–10minutes at RT with stirring. Again, cells were centrifuged at 8000RPM for 20 minutes at 4°C, the supernatant discarded, and pellet resuspended in 400mL of ice-cold hyperosmotic 5mM $MgSO_4$ buffer to enact osmotic shock causing a turgor pressure increase to selectively release periplasmic proteins including exported POI [437]. The resulting suspension was then left rocking for 10 minutes in an ice bath to fully resuspend soluble POI. Centrifuged at 8000RPM for 20 minutes at 4°C.The supernatant from this step was the cold osmotic shock fluid and was retained whilst pellet discarded, to this supernatant 8mL of 1M Tris-HCl, pH 7.4 was added. Then using an MBP column (amylose resin) purification was undertaken using an ÄKTA start system.

5.2.2.5.3 MBP purification, fusion tag cleavage and reverse His purification

The lysate supernatant was loaded at 2mL/min onto a MBPTrap FF, 5mL column (Cytiva) composed of dextrin Sepharose, pre-equilibrated with 5 column volumes (CVs) of MBP A buffer using an ÄKTA Start (Cytiva) system. Column was then washed with 5 CVs of MBPA to remove unbound proteins, then elution was achieved using 5 CVs of MBPTrap B elution buffer (20mM Tris-HCl, 200mM NaCl, 1mM EDTA, pH 7.4, 1mM DTT and importantly 10mM maltose) wash for 6 CVs. Fractions containing the POI were subsequently identified using the UV chromatograph and retained for SDS-PAGE analysis.

Fractions containing POI were pooled and dialysed O/N at 4°C using 3500Da cellulose membrane dialysis tubing into a high salt 2L buffer (500mM NaCl, 20mM Na2HPO4, pH 7.4) with stirring. To avoid precipitation observed at high protein concentrations during tag cleavage, UV peaks of ≥600mAU were diluted 1 in 2 and

≥1000mAU were diluted 1 in 3 with MBPA. During this dialysis step the MBP fusion tag was cleaved by adding 1mL 3C protease at 2mg/mL to the protein before adding it to the tubing.

Following dialysis, a reverse MBP purification step, whereby the post dialysis soluble (PDS) solution, was purified on a MBPTrap HP, 5mL column (Cytiva) using a flow rate of 2mL/min. Following loading MBPA buffer was flushed through with RFT collection (containing the target protein with no tag) continued until UV returned to baseline. At this point buffer flow was switched to MBPB to elute cleaved tag and protease, which bound to the column (elution reverse (ER)). Samples were retained to assess purity using SDS- PAGE analysis outlined in **methods section 2.6.**

5.3 Results

HADDOCK was identified as the best tool and only compatible tool, with the Hyp residues in the TII gelatin peptide fragment.

5.3.1 Peptide generation

Firstly 113 peptides were generated using the five tools outlined in **Table 52**. This was a large number so the decision was made to cluster the largest number generated by one tool, the ModPep generated peptides. Each of the generated 100 ModPep peptide conformers were compared individually to all the others to see if any agreed or were perfect matches using clustering. Maxcluster and a nearest neighbour all vs all assessment identified 4 clusters, and 4 outliers with no similarity meeting threshold (within 4 Ångströms (Å), sharing a minimum of 20 pairs of residues and a MaxSub score above 0.2). Results of this clustering analysis are

shown in **Table 62**, which shows the four identified clusters, the number of peptides allocated to each cluster, the central representative peptide for each cluster or just the peptide outlier alone. All eight peptides listed in the central representative row of **Table 62** were selected to take forward to the next stage of docking validation. The clustering methodology successfully reduced the number of peptides from the ModPep server (from 100 to 8) leaving a representative sampling of the peptide conformational possibilities.

Table 62: ModPep clustering results. Maxcluster successfully identified 4 clusters, 4 outliers (with no RMSD threshold meeting matches). The number of peptides allocated to each cluster is listed and a central representative peptide for each cluster or just the peptide outlier alone was specified, these were selected to take forward to the docking validation. The clustering methodology reduced the number of peptides from the ModPep server generated peptides from 100 to 8. Providing a strategic sampling of the conformational possibilities.

	Cluster 1	Cluster2	Cluster 3	Cluster 4	Outlier 1	Outlier 2	Outlier 3	Outlier 4
Number of conformers	50	20	18	8	1	1	1	1
Central representative selected	65	12	46	40	77	76	74	16

When combining these eight conformers in **Table 62** with the other thirteen generated with the other four tools used this left twenty-one different peptides as summarised in **Table 63**. These selected peptides were all considered as different and valid options to proceed with to a docking validation step.

Table 63: Overview of number of peptides generated and subsequently selected. Number of selected peptide conformers from each tool, the three manually Hyp modified tools and the two automatic tools are specified, the number of peptide models generated with each, and the subsequent number of models selected from the five different peptide model generating methods are specified.

Tool	PeptideBuilder	FragBuilder	ModPepServer	PEPstrMOD	Vienna-PTM
Type	Manual	Manual	Manual	Automatic	Automatic
Number of models generated	1	1	100	1	10
Number of models selected	**1**	**1**	**4 clusters + 4 outliers = 8**	**1**	**10**

Figure 86 and **Figure 87** show the array of different conformations produced by the two different Hyp modification types (automatic and manual), all are unstructured as would be expected for a peptide. Maxcluster showed that the Modpep generated peptides modified manually and automatically were identical (indicated by the RMSD result of 0 when comparing only the two). Looking at them in the below figures and transposed this is not always visually seen due to flexibility and pose. Thus, the text file for each was examined, this closer examination revealed that the atom configurations did show differences in the atom numbers introduced by hydroxylation (identified by comparing the before and after PDB text files). So, both were kept in the peptide library here to see if this made a difference during docking.

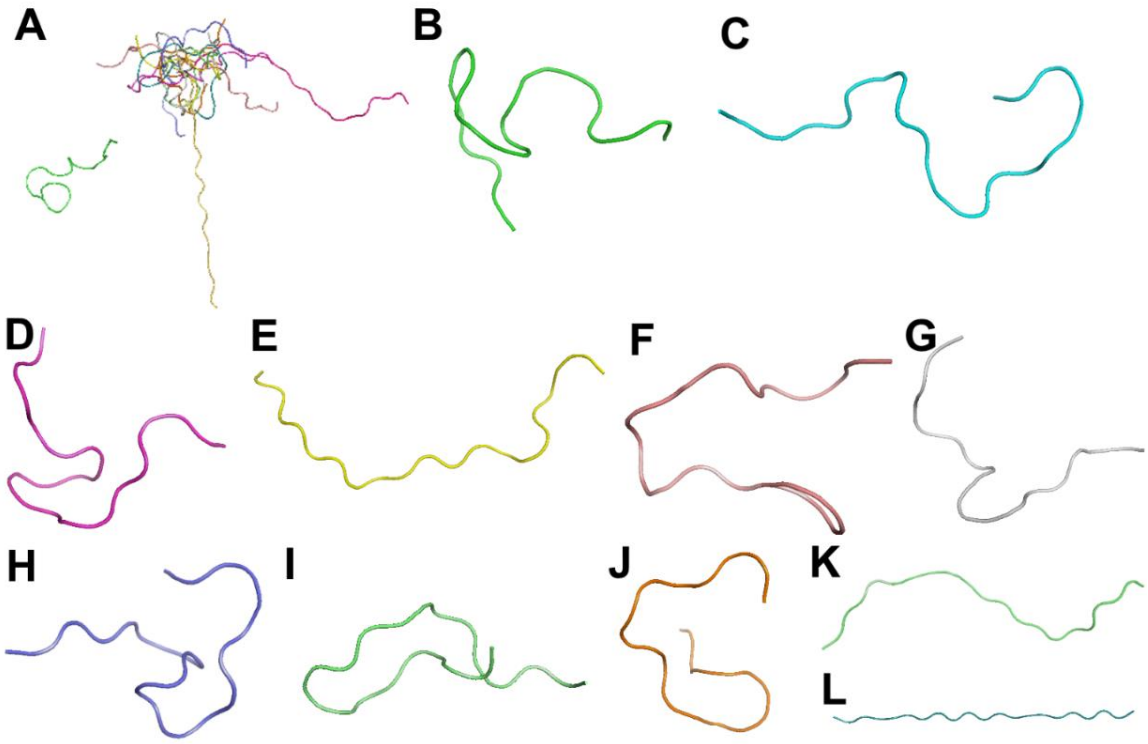


Figure 86: TII gelatin peptide conformers generated using automatic Hyp modification (A) shows the 11 transposed automatically modified peptides. **(B-L)** show the peptides modified to include the Hyp of TII gelatin using the automatic Vienna post-translational modification tool which are shown in the following order: **(B)** PepStrMod, **(C)** 12, **(D)** 16, **(E)** 40, **(F)** 46, **(G)** 65, **(H)** 74 , **(I)** 76 , **(J)** 77, **(K)** Fragment builder and **(L)** Peptide builder.

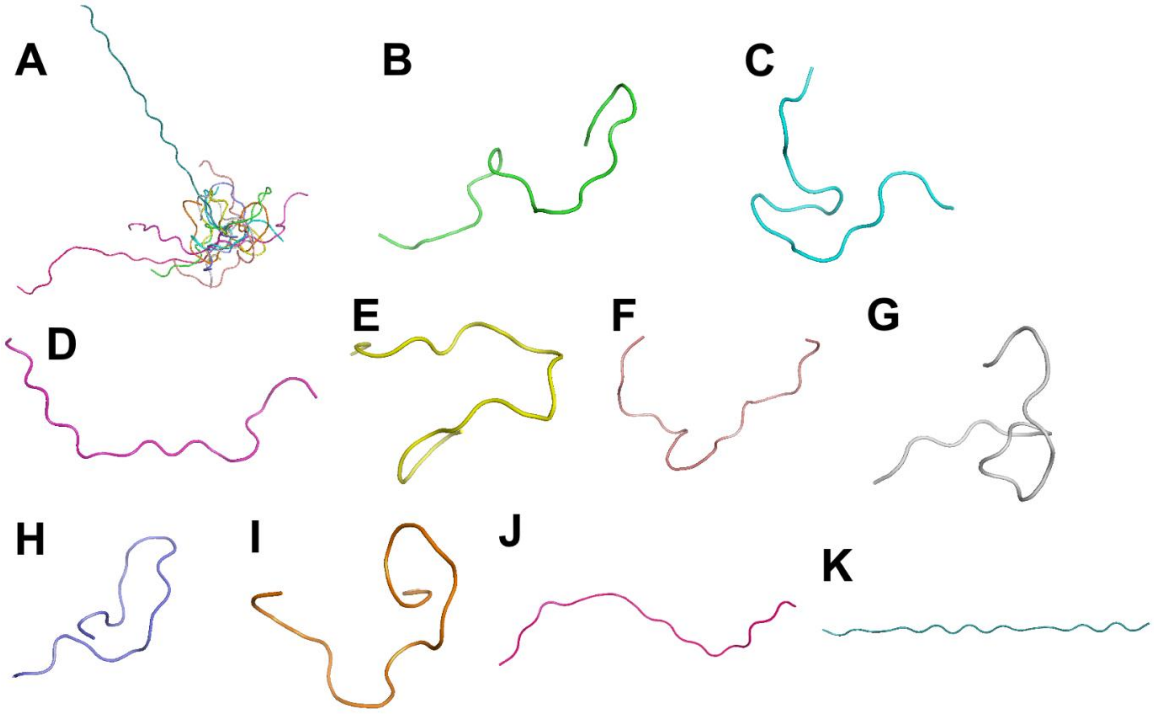


Figure 87: TII gelatin peptide conformers modified manually using the Pytms pymol plugin. (A) shows the 10 transposed manually modified peptides. **(B-K)** show the peptides modified to include the Hyp of TII gelatin, in the following order: **(B)** 12, **(C)** 16, **(D)** 40, **(E)** 46, **(F)** 65, **(G)** 74, **(H)** 76, **(I)** 77, **(J)** Fragment builder and **(K)** Peptide builder.

The main challenge here was generating and selecting a valid array of peptides to act as a conformer library for use in docking validation, increasing chances of successful *in vitro* to *in silico* binding translation (the reverse of the order that experiments usually follow).

5.3.1.1 Assessing biophysical/ spatial validity of modelled peptides

The ten peptide conformers generated and selected with manual Hyp modifying tools **Table 62** and **Table 63** were assessed using MolProbity to identify if any were better biophysical models, with more spatially allowed configurations [537]. MolProbity score combines several steric assessments, specifically Clashscore, rotamer and Ramachandran evaluation into a single score.

When reviewing the ten manual hydroxylation modified peptides based on MolProbity scores alone, it is only peptide 77 and Frag peptide conformers that were highlighted as being of concern/ not valid, the four other conformers were identified as good to proceed with.

Table 64: Manually modified peptide MolProbity results. Results are colour coded to simplify interpretation; green are good, yellow intermediate and red bad (when compared to the MolProbity listed goals shown in the top row. This color coding was taken from the MolProbity output directly). A lower MolProbity score was better.

Peptide conformer	MolProbity score (Percentile ≥66)	Clashscore (Percentile ≥66)	Ramachandran favoured (Goal >98%)	Ramachandran outliers (Goal <0.05%)	Rama distribution Z- score (Goal <2)	Bad angles (Goal 0.1%)	Bad bonds (Goal 0.0%)
12	1.34 (98th percentile)	0.00 (100th percentile)	70.00	0.00	-0.67 ±3.18	1.6	0.0
16	1.34 (98th percentile)	0.00 (100th percentile)	70.00	0.00	-3.25 ±2.36	2.6	0.0
40	1.71 (89th percentile)	3.73 (96th percentile)	90.00	0.00	-4.01 ±1.58	0.5	0.0
46	2.00 (76th percentile)	0.00 (100th percentile)	90.00	0.00	-4.42 ±1.62	0.0	0.0
65	1.05 (100th percentile)	0.00 (100th percentile)	100.00	0.00	-4.77 ±1.79	0.5	0.0
74	1.81 (85th percentile)	0.00 (100th percentile)	90.00	0.00	-0.38 ±2.49	0.5	0.0
76	0.50 (100th percentile)	0.00 (100th percentile)	100.00	0.00	-2.93 ±1.76	0.5	0.0
77	3.10 (19th percentile)	44.78 (6th percentile)	50.00	10.00	-8.86 ± 0.76	20.1	11.9
Pep	1.24 (99th percentile	0.00 (100th percentile)	80.00	20.00	-5.85 ±1.37	0.5	0.7
Frag	3.10 (19th percentile)	44.78 (6th percentile)	90.00	10.00	-8.86 ±0.76	20.1	11.9

5.3.2 Peptide binding validation

Figure 88A shows peptide conformer Vienna 76 (from the automatically modified peptides) matched the experimental binding order using HADDOCK score, so was the only automatically generated peptide that could be considered valid and taken forward to docking.

Then **Figure 88B** shows that based on HADDOCK score peptides 16, 40 and Frag (from the manually modified peptides) matched the experimental binding order (shown in **Table 54**), so could be considered valid peptides based on this HADDOCK only check. Comparing with the MolProbity results (**Table 64**) peptide 40 is considered the better model out of the three when analysing the breakdown of individual scores, specifically bond angles and Ramachandran favoured results. Although still classed as bad according to the MolProbity Ramachandran favoured goal is 98%, 90% of residues were Ramachandran favoured in both peptide 40 and Frag, with Frag being closer to the >98% goal value then the 70% observed for peptide 16. Then analysing bad angles percentage, peptide 40 is the better choice with only 0.5% of angles classified as bad whereas the frag peptide has 20.1%.

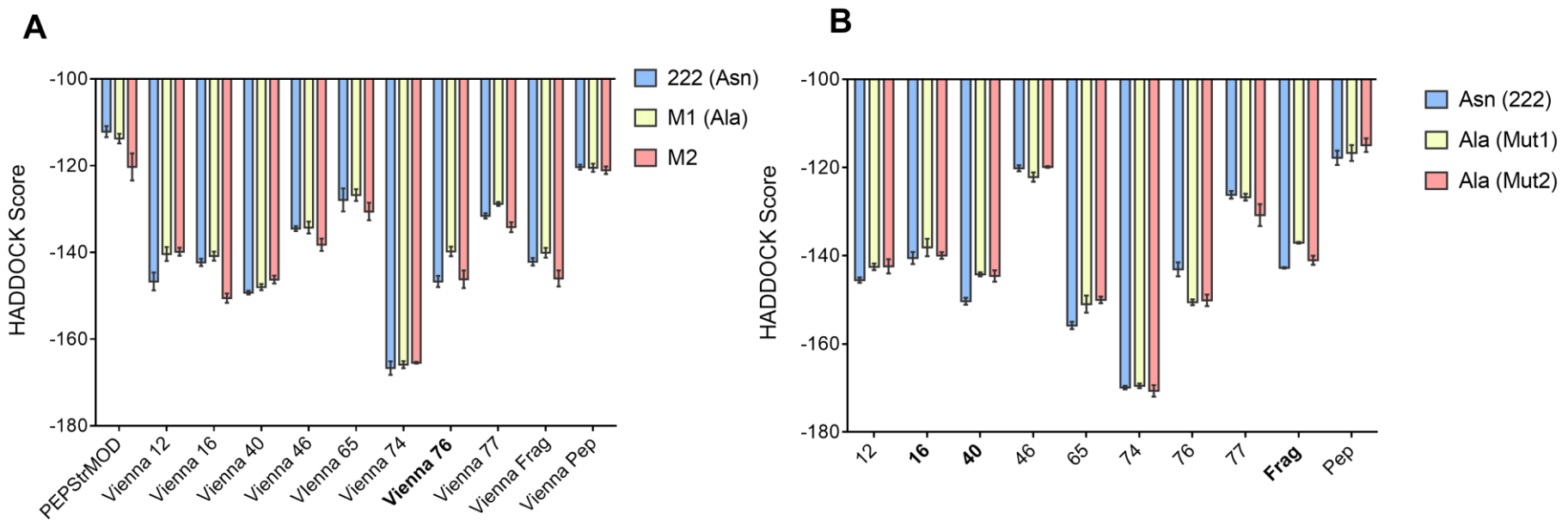


Figure 88: HADDOCK peptide validation assessment results. Lower HADDOCK score indicates better binding. **(A)** Automatic. Conformer Vienna 76 matches the experimental binding order for these three **chapter 3** protein variants (222, M1 and M2). **(B)** Manual. Conformer 16 and 40 matches the experimental binding order for these three, chapter 4 protein variants. With 222 being the lowest HADDOCK score showing it's the best binding, M1 being the highest showing it's the worst binding and M2 being in the middle of the two.

The peptides conformers selected from the possible 21 that were taken to this MolProbity, and HADDOCK matched binding order docking validation step are shown in **Table 65,** here the decision to take one manual and one automatically Hyp was made.

Table 65: Peptide choice summary. Results are colour coded to simplify interpretation as per the MolProbity color coded output; green are good, yellow intermediate and red bad. Peptide 40 was selected over peptide 16 as peptide 40 had a lower % of band angles.

Peptide Conformer	PTM (Hyp) type	Experimental binding order matched based on HADDOCK score	MolProbity Score	Bad angles (%)
16	Manual	✓	1.34	2.6
40	Manual	✓	1.71	0.5
Frag	Manual	✓	3.10	20.1
Vienna 76	Automatic	✓	1.95	1.6

5.3.3 HADDOCK complex structures

The best complexed HADDOCK outputs were downloaded as PDB files and taken to the refinement interface stage to assess binding of mutants (**Figure 89)**. With peptide 40 the peptide is in an elongated conformation spanning across a larger surface area in the 222 protein, whereas with peptide Vienna 76 there is a curvature in the docked peptide meaning it is much more compact in its docked conformation. Binding residues were also mapped onto these complex predictions. Complex 1 (**Figure 89 A & B**) does shows the peptide (pink) in proximity to the binding residues pocket (cyan). In comparison complex 2 (**Figure 89 C & D**) has more binding residues (cyan) shown across the structure not in proximity to the peptide (orange).

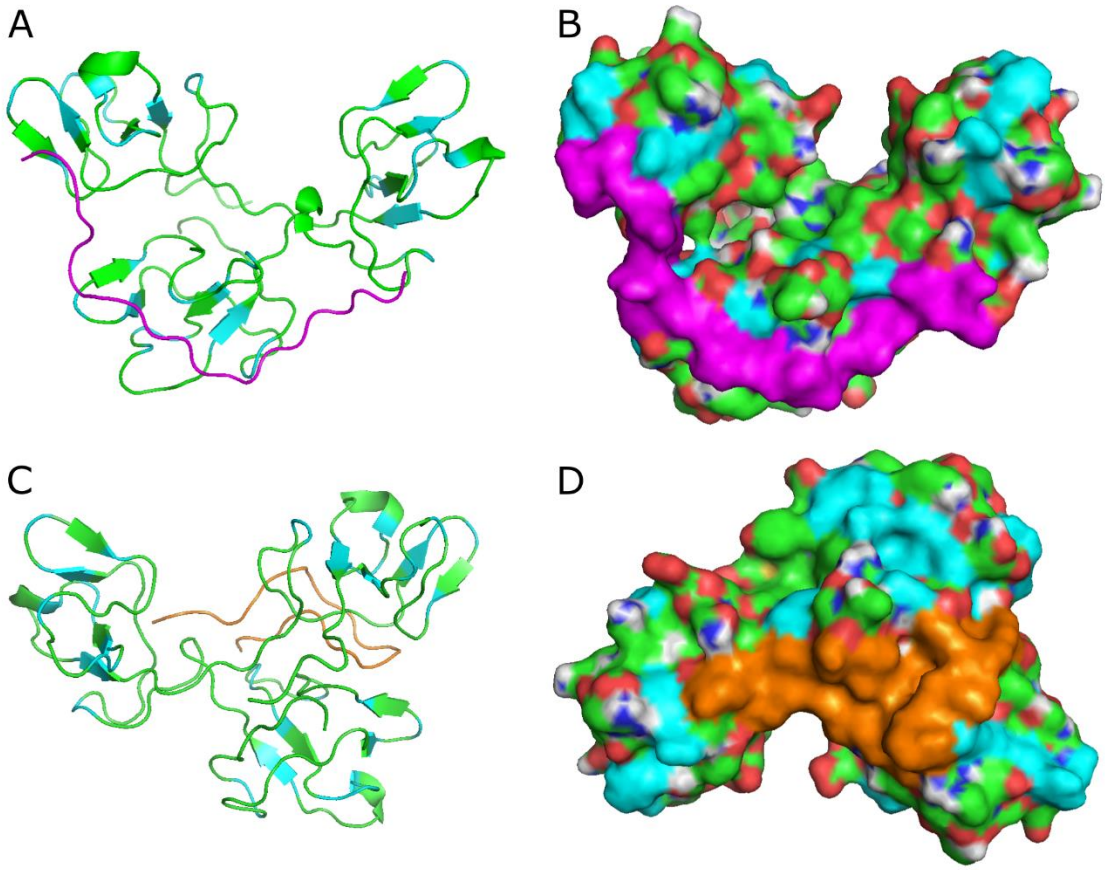


Figure 89: HADDOCK best complex structures. A & B, Complex 1 peptide 40 best docking structure output (complex structure 18). Peptide 40 highlighted in pink, 222 in green and known (NMR identified) binding residues in cyan **C & D,** Complex 2 peptide Vienna 76 docking best structure output complex structure 103. Peptide Vienna 76 highlighted in orange, 222 in green and known binding residues in cyan. **A, C** Cartoon pymol structure representation of the two best complex structures with the lowest haddock score (for the two best peptides). **B, D** Surface pymol structure of both complexes, the atom type of 222 is shown on the structure by the colour, green (carbon), red (oxygen) grey (hydrogen) and blue (nitrogen).

Previous NMR studies identified a cluster of aromatic residues forming the binding site for TI gelatin in each CBD module [560][457]. Confirming the prevalence of hydrophobic interactions in the binding of the CBD proteins to gelatin. It wasn't however until the NMR experiments conducted by the Hollander group prior to this project that CBD protein binding to TII gelatin was investigated specifically. In this work

it was shown that CBD, module 2 and 222 bind to TI and TII gelatin with the same mechanism, and through the same residues that form a hydrophobic pocket [288].

Table 66 outlines the details of the two best docked complexes from each of the validated peptide conformers**.** Of the two, complex 1 (involving peptide 40) is the better complex with the lower HADDOCK score indicating tighter binding. RMSD is comparable for both meaning both are equally strong structural predictions. Complex structures 18 and 103 were therefore downloaded and used to generate mutant pdbs.

Table 66: Peptide conformer HADDOCK statistics. HADDOCK cluster 2 was the top cluster which was the most reliable according to HADDOCK. Z-score indicates how many standard deviations from the average this cluster is in terms of score (the lower the better).

Complex ref	**Complex 1**		**Complex 2**	
Best complex structure	**18**		**103**	
Peptide conformer	40		Vienna 76	
Total number of structures clustered	194		204	
Total number of clusters	9		7	
Cluster	2		1	
HADDOCK score	2.8	±3.7	27.2	±1.2
Cluster size	57	/	82	/
RMSD from the overall lowest-energy structure	0.3	±0.2	0.4	±0.2
Van der Waals energy	-77.8	±3.5	-88.0	±4.1
Electrostatic energy	-257.4	±20.9	-237.1	±19.3
Desolvation energy	-2.1	±1.9	7.6	±2.2
Restraints violation energy	1340.8	±15.27	1551.2	±20.89
Buried Surface Area	2104.3	±18.2	2060.0	±45.8
Z-Score	-2.2	/	-1.9	-1.9

5.3.4 Generating better binding mutants

Mutants were generated through *in silico* mutation of residues Asn 11, 69 and 127 (Identified in **chapter 3** as the most critical to binding) to all other possible residues. **Figure 90** shows the HADDOCK scores generated using the refinement interface, the lower scores indicated better binding. Therefore, the mutant residues farthest left in both figures are the predicted strongest binding mutants. Although not quantitative, it was noted that the difference in HADDOCK score/binding between Trp and 222 (**A**) is greater than the difference in HADDOCK score/ binding between Met and 222 (**B**).

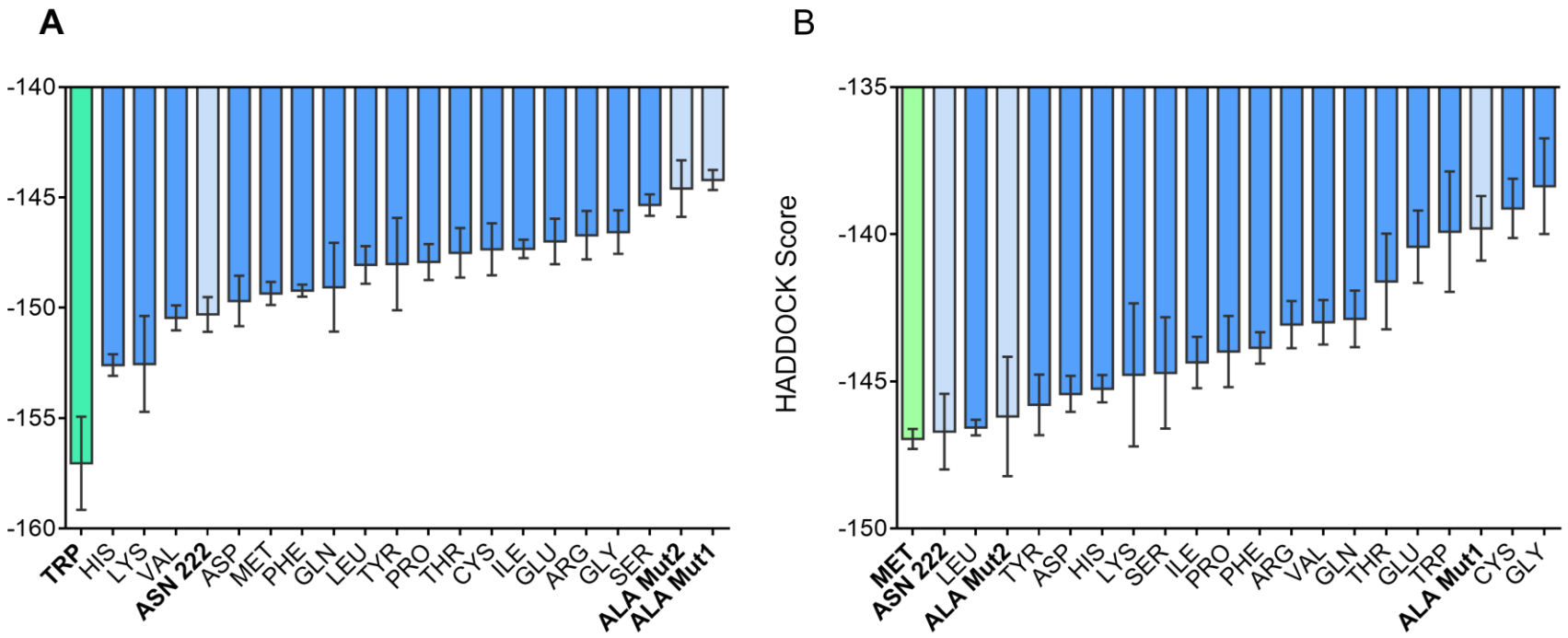


Figure 90: HADDOCK refinement interface residue substitution comparison. (A) HADDOCK score using peptide conformer 40 suggests a mutant with tryptophan (Trp) in place the of asparagine (Asn) 11, 69 and 127 (in 222) would have superior binding. **(B)** HADDOCK score using peptide conformer Vienna 76 suggests a mutant with methionine (Met) in place the of asparagine (Asn) 11, 69 and 127 (in 222) would have superior binding.

5.3.5 CamSol mutants HADDOCK assessment

Solubility mutants are predicted to have a weaker binding affinity for TII gelatin using HADDOCK scores, compared to 222 (the known best binding mutant from *in vitro* assessment using plate binding assay). **Figure 91** shows that CS1 and CS2 are comparable in HADDOCK score to M1 and M2 the refinement mutants screened in the previous assessment in search of a better binding mutant. This prediction fits indirectly with the observation that the most changed solubility mutant, CS6, which was tested *in vitro* in results **chapter 4**, was proven more soluble but also to have a significant loss of binding affinity compared to 222. Unfortunately, an *in silico* prediction of CS6 could not be made to compare directly with 222 using HADDOCK score as CS6 was not just a substitution mutant it contained insertions meaning refining the 222 complex was not possible. Only CS1 and CS2 were assessable using HADDOCK as CS3-CS6 were no longer refinement interface comparable. This assessment shows that solubility improvement even limited to only the lowest two CamSol mutants designed will still be at the cost of binding.

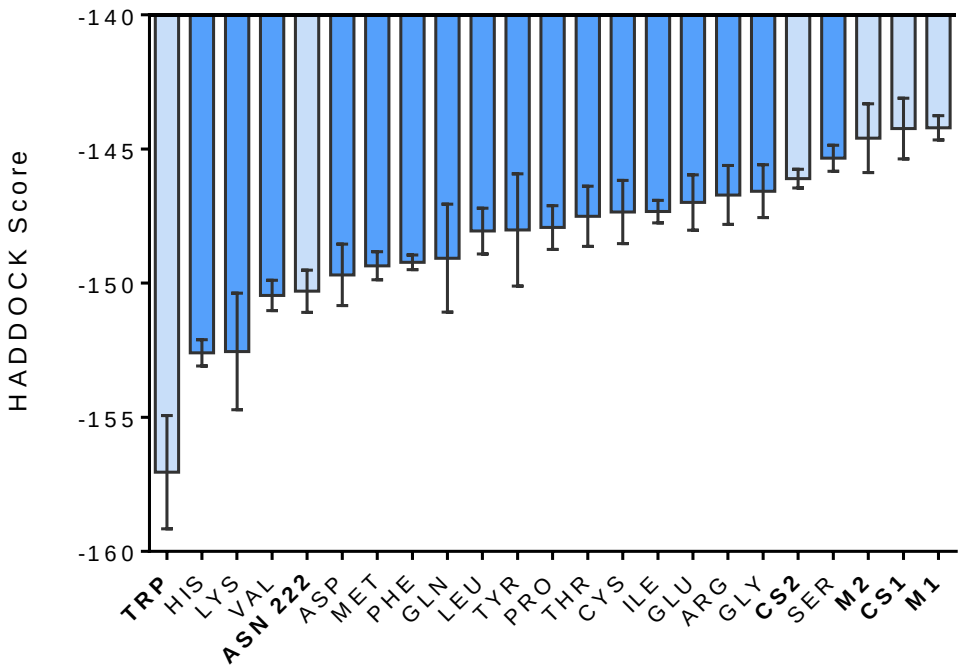


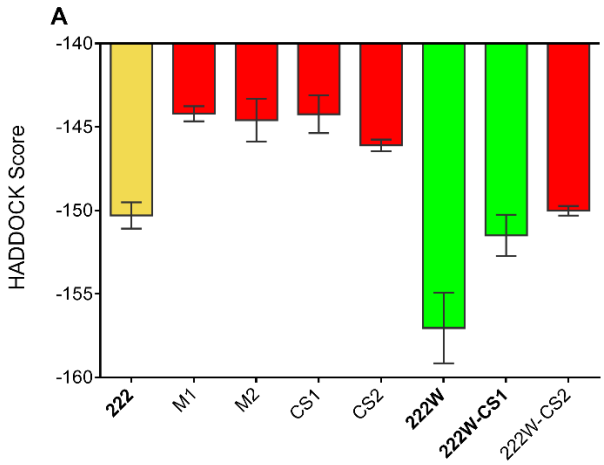
Figure 91: CS1 and CS2 mutant HADDOCK comparison. HADDOCK refinement interface residue substitution comparison with CS1 and CS2 added. HADDOCK score using peptide conformer 40 suggests a mutant with tryptophan (Trp) in place of the of asparagine (Asn) 11, 69 and 127 (in 222) will have superior binding. This shows that solubility improvement even limited to only the lowest two CamSol mutants designed will still be at the cost of binding.

5.3.5.1 Dual solubility and binding mutants

Figure 92A shows that 222W is the predicted highest potency binding mutant with the lowest HADDOCK score of -157.043.

Combining CamSol and binding mutations, to see if an improved binding and solubility mutant could be attained was attempted with 222W-CS1 and 222W-CS2. **Figure 92B** outlines the two mutations combined in these proteins before *in silico* HADDOCK assessment was utilised to predict binding affinity implications of combining the two. 222W-CS1 (HADDOCK score of -151.495) was predicted to have binding slightly more potent than that of 222 (HADDOCK score of -150.301).

This variant may offer improved solubility but improved binding when compared to 222. A recovery of binding was seen when the 222W mutation (for better binding) was combined with the HADDOCK assessable most conservative CamSol mutants CS1 or CS2. 222W-CS1 was the dual solubility binding mutant with the best HADDOCK score of -151.495, the lowest HADDOCK score hence best binding of the mutants for both traits in this work.



B

Variant	Mutation
222	Native ASN at 11, 69 and 127
M1	ALA at 11, 69 and 127
M2	ALA at 22, 80 and 138
222W	TRP at 11, 69 and 127
CS1	GLU at 9
CS2	GLU at 6 and 9
222W-CS1	GLU at 9 TRP at 11, 69 and 127
222W-CS2 (2nd batch)	Glu at 6 and 9 TRP at 11 69 127

Figure 92: Haddock assessment of the key proteins in this work. (A) HADDOCK score comparison, lower score shows a stronger affinity binding**, (B)** Summary of the key proteins in this work (HADDOCK scores of which are shown in (**A)**), and the mutations present in each. Results here allow *in silico* comparison of predicted binding within the constraints of the refinement interface Key proteins from each chapter shown: binding mutants 222, M1, M2 (**chapter 3**), CS1 and CS2 (**chapter 4**) 222W (improved binding prediction) then binding and solubility combined 222W-CS1 and 222W-CS2 (**chapter 5)**.

5.3.5.2 Small scale expression & purification

222W and 222W-CS1 were taken forward to be expressed for experimental verification of binding. Both pOPINS and pOPINJ vectors showed very limited levels of soluble expression for both 222W and 222W-CS1. With expression bands evident in the correct molecular weight size region (of 62.8kDA for both proteins with MBP tag) as assessed through SDS-PAGE analysis. The best soluble expression of both

222W and 222W-CS1 was obtained using pOPINM encoding an MBP fusion tagged POI (**Figure 76**)**.** For 222W soluble protein of target POI size was seen for all IPTG concentrations, despite only observing small levels of suspect sized expression at this stage it was worth taking to an initial 2L large scale trial to confirm if target protein was present and purified. With 222W-CS1 the slightly stronger soluble expression was seen with 800µM IPTG, but expression was seen in the soluble for all concentrations. As 800µM IPTG 16°C was selected for 222W-CS1 it was also chosen for 222W for large scale expression ease and consistency.

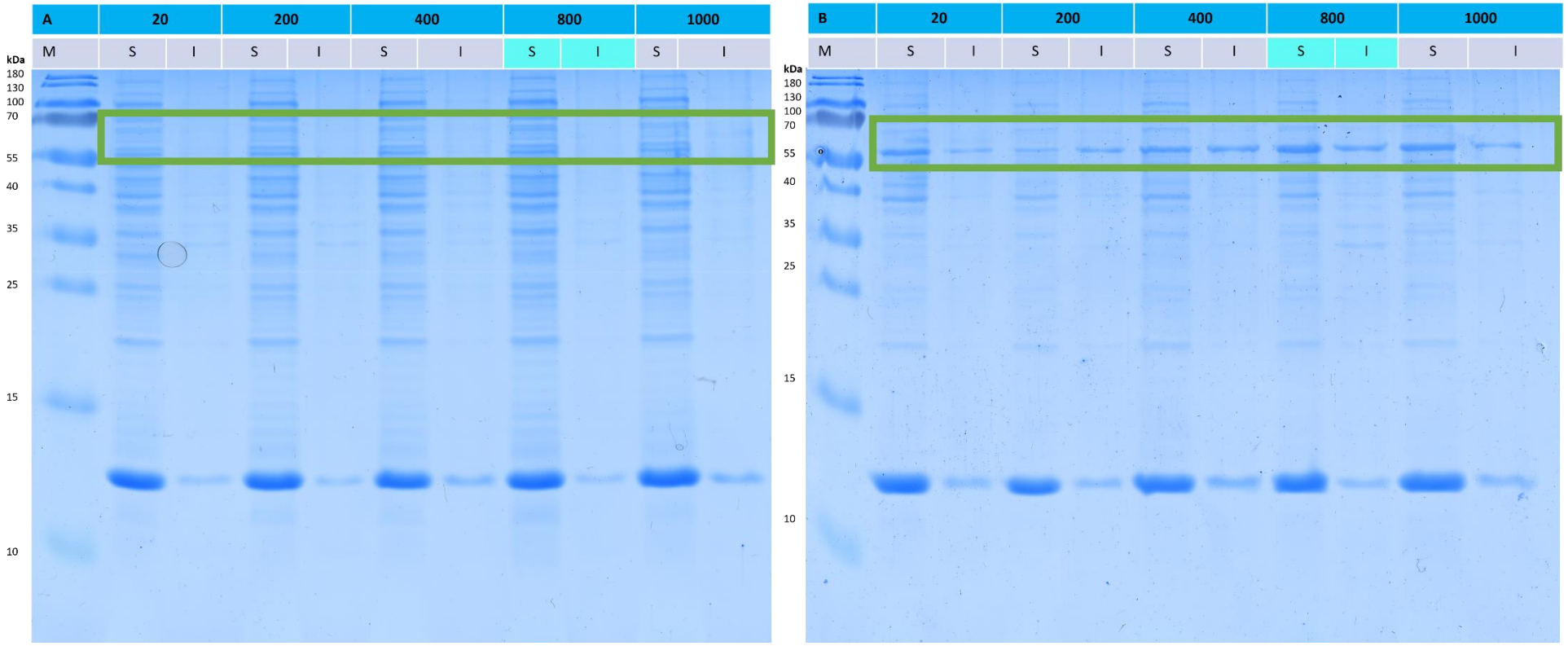


Figure 93: MBP tagged 222W (A) and 222W-CS1(B) SDS-PAGE small scale purification soluble (S) and insoluble (I) fractions. (A) 222W (20.8kDa) with MBP tag(42kDa) is 62.8kDa. **(B)** 222W-CS1 (20.8kDa) with MBP tag (42kDa) is 62.8kDa with tag. There is a small band on both gels at the appropriate size here showing expression was induced worth an Äkta trial to confirm or rule out. Because of this both proteins were taken following this gel to large scale expression and purification with an initial 2L volume, induction concentration of 800µM IPTG (highlighted in light blue), 16°C O/N expression as a first large scale full expression and purification attempt.

5.3.6 Large scale expression & purification

Figure 94 shows the results of large scale MBP tagged 222W expression and His purification. Results show the successful expression of both the pOPINM encoded His6-MBP-222W **(A**) and His6-MBP-222W-CS1 **(B)** in the soluble fraction as evidenced by the large appropriate sized band with fusion tag at 62.8kDa.There is an intense band on both gels at the appropriate size here showing expression was in fact induced and his purified (highlighted in green box in **Figure 94**). However, when 3C protease cleavage to remove the MBP tag was attempted during O/N dialysis at 4°C, using 3C protease as used in previous chapters, there was no change in size observed meaning cleavage was unsuccessful, (highlighted in the pink box, **Figure 94,** in the post dialysis soluble (PDS) and reverse flow-through (RFT)). The protein remained tagged which is why it remained at the same molecular weight in the elution reverse (ER) highlighted in yellow box in **Figure 94**. The MBP tagged 222W and 222W-CS1 proteins bound and eluted from the column during the reverse His purification as they retained fusion tag after cleavage attempt.

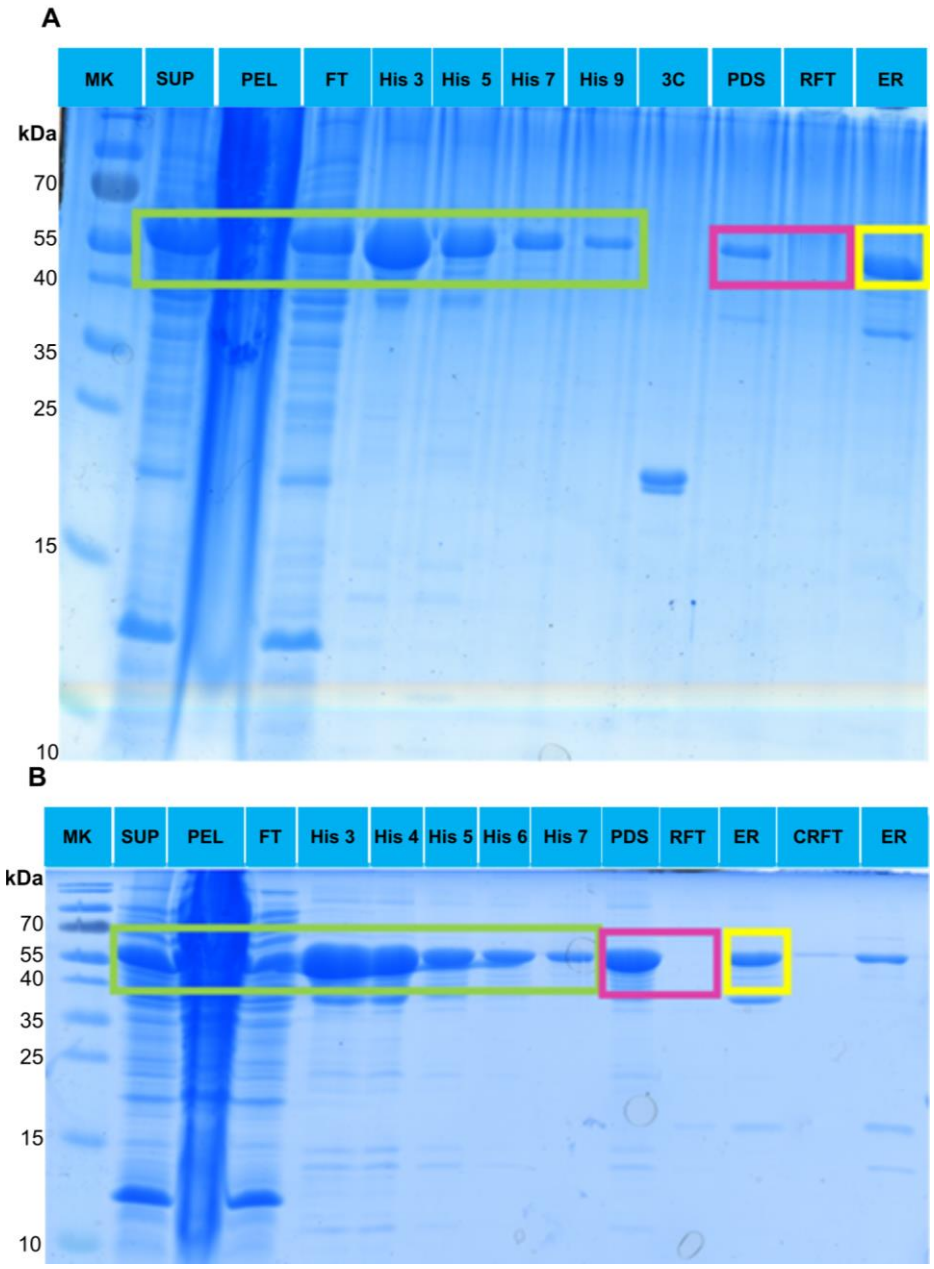


Figure 94: MBP tagged 222W and 222W-CS1 SDS-PAGE results of His purification and cleavage attempt. (A) 222W (20.8kDa) with MBP tag (42kDa) is 62.8kDa. **(B)** 222W-CS1 (20.8kDa) with MBP tag (42kDa) is 62.8kDa**.** There is an intense band on both gels at the appropriate size here showing expression was induced and his purified (highlighted in green). However, when 3C protease cleavage to remove the MBP tag was attempted during O/N dialysis at 4°C, there was no change in size observed meaning cleavage was unsuccessful, highlighted in pink in the post dialysis soluble (PDS) and reverse flow-through (RFT). The protein remained tagged which is why it remained at the same molecular weight in the elution reverse (ER) highlighted in yellow. Concentrated RFT (CRFT) and 3C protease (3C) are also shown. The MBP tagged 222W and 222W-CS1 proteins bound and eluted from the column during the reverse His purification.

5.3.6.1 Cleavage trials

Following the unsuccessful cleavage attempt as shown in **Figure 94**, cleavage was attempted with a fresh batch of 3C protease and a control protein (alternate MBP tagged protein and 3C protease from a colleague (Dr Dominic Byrne)), to check protease activity. Cleavage with the alternate batch was again unsuccessful with 222W and 222W-CS1 but successful with the control proving the enzyme was active (results not shown).

Cleavage was also attempted in the presence of 2M urea to try and induce conformational opening/ partial unfolding of the fusion site slightly to allow 3C protease cleavage to take place. Results of the urea trials are shown in **Figure 95**, urea addition did not make a difference, the protein remained uncleaved. Although protein was acquired in soluble form, MBP tag was not cleaved from either protein, so it was at this point work on these mutants ended with no testing of their binding to gelatin possible within the time available.

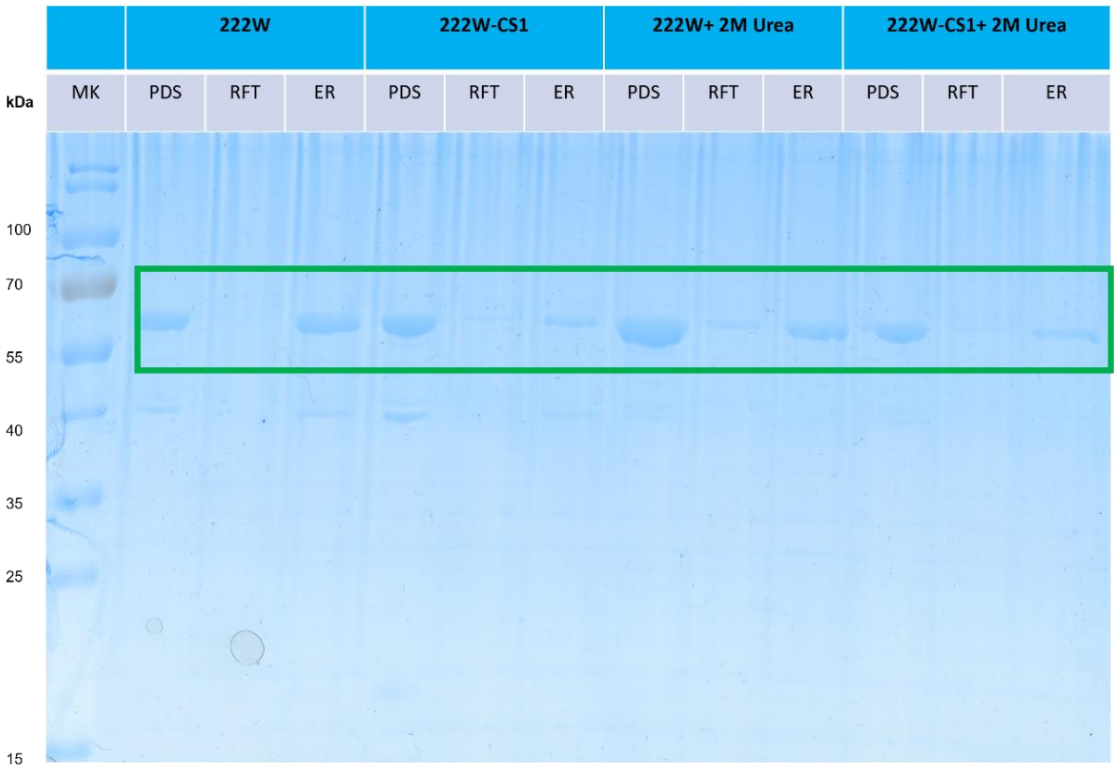


Figure 95: 222W and 222W-CS1 SDS-PAGE results of His purification and cleavage attempt with and without 2M Urea. 222W (20.8kDa) with MBP tag (42kDa) is 62.8kDa. **(B)** 222W-CS1 (20.8kDa) with MBP tag (42kDa) is 62.8kDa There is a band present at the appropriate size here showing expression was induced and His purified (highlighted in green). When 3C protease cleavage to remove the MBP tag was attempted during O/N dialysis at 4°C, there was no change of size meaning cleavage was unsuccessful (observed in the post dialysis soluble, PDS and reverse flow-through, RFT). The protein remained tagged which is why it remained at the same molecular weight in the elution reverse (ER). The MBP tagged 222W and 222W-CS1 proteins instead bound and eluted from the column during the reverse His purification.

5.3.6.2 Periplasmic expression and purification

Here was the first employment of this periplasmic expression strategy within this thesis, chosen as the periplasm is better suited to protein folding and disulfide bond formation [561]. One reason why tag cleavage was not efficient for the MBP constructs could be due to incorrect folding therefore this periplasmic method was trialled as a last effort. **Figure 96** shows the SDS-PAGE results of periplasmic pMAL-p5X expressed POIs (222W and 222W-CS1). This expression strategy gave a much cleaner initial lysate, however dilution of the osmotic shock lysed product, meant large volumes subsequently required MBP column ÄKTA purification which took a long time to process and was difficult (limited by equipment capacity). Additionally, the immediate purification progression requirement (no freezing bacterial pellet to maintain periplasm) made it a more difficult method to employ, requiring further trialling and refinement, for which there was insufficient time here. Therefore, it was judged not worth pursuing at this point. Of the two different cell types trialled Lemo21 appeared to yield greater expression levels shown by the larger band at 62.8kDa with Lemo21 compared to C41 cells (**Figure 96)**. This may be applicable if periplasmic expression is utilised in any further work.

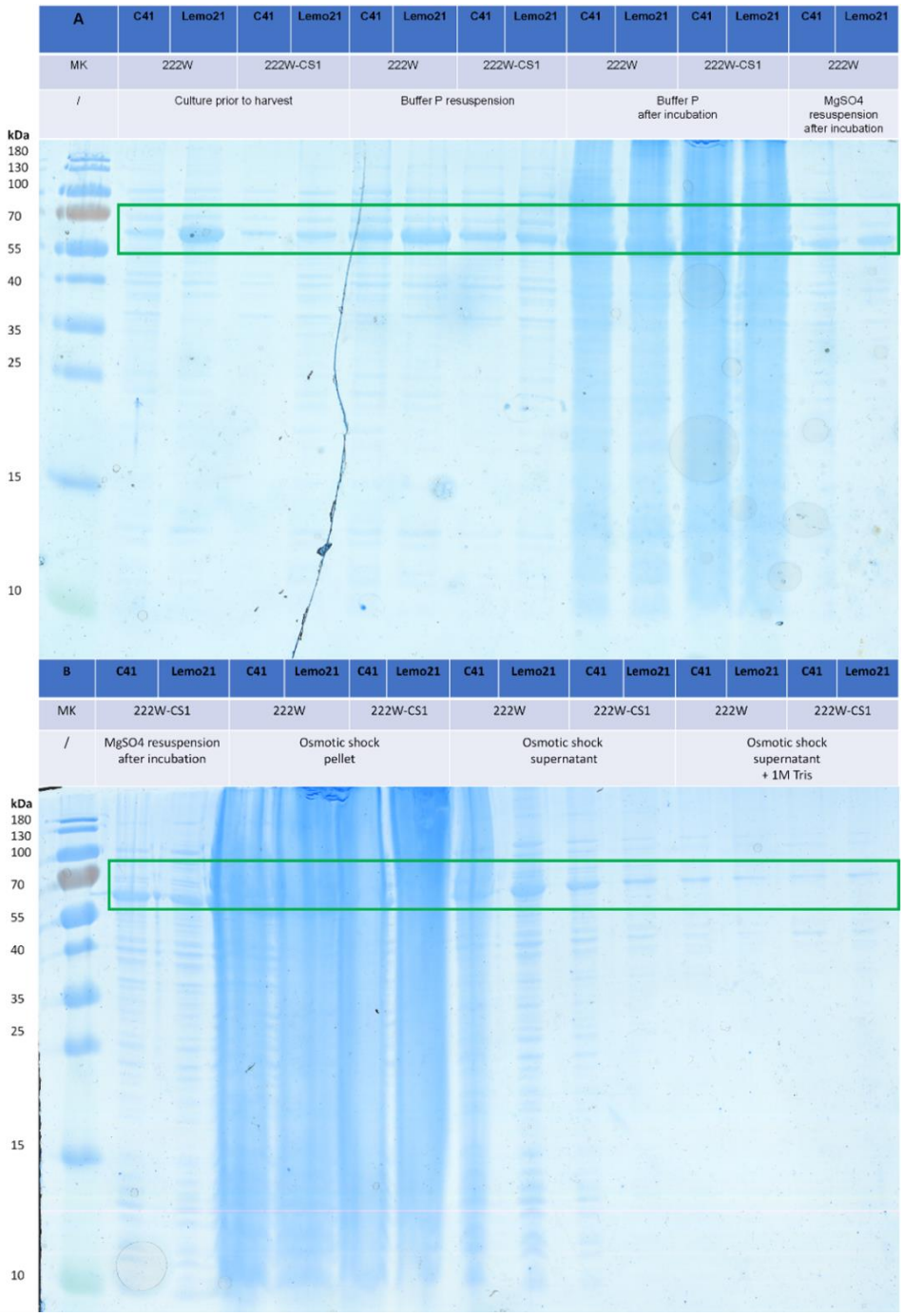


Figure 96: MBP tagged 222W and 222W-CS1 SDS-PAGE results of periplasmic expression and purification attempt. SDS-PAGE results **(A)** 222W (20.8kDa) with MBP tag (42kDa) is 62.8kDa. **(B)** 222W-CS1 (20.8kDa) with MBP tag (42kDa) is 62.8kDa There is an intense band considering the 2L culture volume on both gels at the appropriate size here showing expression was induced and osmotic shock successfully lysed and kept the periplasmic proteins separate from the cytoplasmic (highlighted in green). Significantly protein product from the periplasm was very clean compared to cytoplasmic expression. Two different cell lines Lemo21 and C41 were trialled concomitantly. Lemo21 gave a larger POI sized band.

5.3.6.3 *In silico* cleavage site modelling and accessibility assessment

To examine if occlusion could be responsible for fusion tag cleavage failure, models of the three relevant fusion products were required. A fusion product with successful cleavage attained *in vitro* His6-MBP-3C-CS6 (used as a control), then the two fusion proteins from this chapter. The best models for the three fusion proteins expressed with the MBP tag (using the pOPINM) vector, in this thesis are shown in **Figure 97**. These models were the best of those generated using Alphafold2, a more recent tool computationally producing models comparable with experimentally derived models in terms of accuracy.

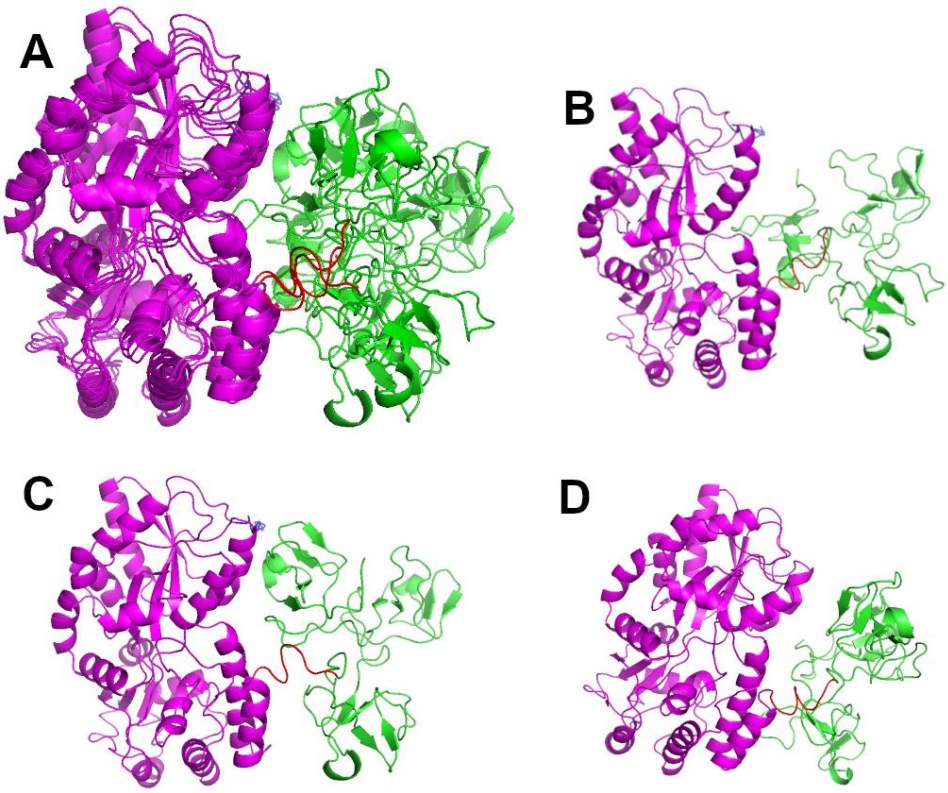


Figure 97: Alphafold2 best fusion protein product models. His-MBP fusion protein is shown in magenta, the POI in green and the 3C cleavage site in red. **A** Transposed fusion products showing some differences/ shifts in structural features. **B,** His-MBP-3C-222W. **C,** His-MBP-3C-222WCS1. **D,** His-MBP-3C-CS6.

Table 67 shows the accuracy scores given for each of these three models by Alphafold2. The pLDDT of 84.9 indicated the models are confident predictions. pTM is the other confidence metric output, here a score of 0.663 also indicated confident/ good models.

Table 67:Alphafold2 best model summary. Protein fusion product, scores reflecting local accuracy (pLDDT), global accuracy (pTM) and experimental *in vitro* cleavage experience/ success.

Protein	**pLDDT**	**pTM**	**Cleavable *in vitro***
His6-MBP-3C-222W	84.9	0.66	N
His6-MBP-3C-222W-CS1			N
His6-MBP-3C-CS6			Y

Table 68 shows the solvent accessibility assessment of the LEVLFQGP 3C protease cleavage site, with a breakdown of the solvent exposure of the individual residues that this site is composed of. For cleavage to occur the protease must bind the entire recognition sequence. When evaluated all as one (the entire recognition sequence) LEVLFQGP, the accessibility is no less for 222W and 222W-CS1 (uncleavable *in vitro*) when compared to CS6 (cleavable *in vitro*). If occlusion of the 3C protease recognition sequence in the two mutants that are uncleavable was the explanation, the recognition sequence should be less solvent exposed/ accessible. SASA is a total exposure measure so some residues could be less exposed, but this would be cancelled out via the cumulative scoring employed by both tools. Only when assessed individually does it become apparent that V382 and L383 may be less exposed in 222W and 222W-CS1 than in CS6. It was also observed in this individual residue examination that for the other residues in the cleavage site the reverse trend is shown with residues more exposed which is curious, but it is the whole site that needs exposure to cleave.

Table 68: Results of solvent exposure assessment of the 3C protease cleavage site residues. Assessment 1 was made using the free SASA tool. Assessment 2 was made using the pymol SASA calculation function. Both assessments agreed that Residues V382 and L383 were less exposed (highlighted in yellow) in 222W and 222W-CS1 (the two uncleavable fusions). In CS6 the fusion variant that did cleave these two residues were more solvent exposed.

Cleavage site residues **LEVLFQGP** **(380-387)**	**Assessment 1: Free SASA (Å²)**			**Assessment 2: Pymol SASA (Å²)**		
	His-MBP-222W	**His-MBP-222W-CS1**	**His-MBP-CS6**	**His-MBP-222W**	**His-MBP-222W-CS1**	**His-MBP-CS6**
L380	35.04	46.04	25.74	42.224	52.522	27.509
E381	113.14	138.77	91.59	119.212	147.93	97.755
V382	**79.7**	**66.32**	**111.62**	**86.254**	**73.172**	**112.481**
L383	**80.95**	**82.27**	**106.9**	**85.872**	**87.068**	**105.505**
F384	128.72	107.64	65.23	130.381	113.686	72.1
Q385	92.7	174.85	144	94.729	173.797	141.188
G386	38.01	43.47	42.91	36.236	46.37	39.686
P387	142.58	141.27	102.76	136.32	136.464	96.268
LEVLGQGP	710.83	800.64	690.76	731.24	831.033	692.501

5.4 Discussion

In silico docking is not yet as developed for binding assessment of proteins and peptides [562, 563], generally binding predictions are limited to small biomolecules which have made it very valuable in drug and small molecule therapeutic development [564]. More recently a focus has shifted to peptide docking [565], computationally difficult due to flexibility, even more complicated if you have flexibility in the binding protein as well [496, 566]. Peptides are flexible molecules that can bind to proteins even in the absence of defined binding pockets/targets. If the proposed extended binding conformation seen in the best complex prediction of peptide 40 is correct this may be the case here (**Figure 89**). Computational modeling, and particularly blind peptide–protein docking, is often hindered by the lack of known structure/ conformation for the peptide this was an issue that we tackled by using a library of options and validations. The experimental binding order from **chapter 3** was fortunately available and used here as a key validation step.

Although great strides have been made in terms of how accurate molecular docking is and capability/ strategies have evolved and improved. It is however still a field with limits and it is important to not misinterpret or apply such tools by overreaching in their use [478, 522]. For example, HADDOCK gives an indication of how binding compares, but it doesn't determine an actual Kd value that is comparable to the experimental values we have for CBD, 222 and the other proteins in this work [567]. This comparative but not absolute output is a common limitation in computational biology. It does however help steer strategies selecting lead candidate

mutants to take to *in vitro* testing. Here we have been strategic and used best tools whilst being very aware and mindful of the limitations [567].

Alphafold2 has very recently been applied to the protein-peptide docking problem in a simple implementation involving connecting the peptide to the receptor. Then using Alphafold2 to accurately identify unstructured regions and model them as extended linkers, predicting protein-peptide complexes without the need for multiple sequence alignment of the peptide [568]. Also, Patchman (Patch-Motif AligNments) is a very recent docking tool, having only just become available [569], that uses a novel peptide docking approach employing structural motifs to map peptide conformations onto the protein backbone [570]. As a method it outperforms other protein-peptide docking methods including the recent implementation of AlphaFold2 mentioned above [569]. So presents an interesting new direction to addressing the protein-peptide docking 'subproblem' and difficulties. Had it been available it would have been explored as an option for use here in this chapter.

On reflection given the emergence of Alphafold2 as the new gold standard approach a test was made to see how the model for 222 generated in **results chapter 3** compared to an Alphafold2 model of 222 generated here (at the time of the writing up this chapter). There are subtle differences between the models which can be seen in **Figure 98**. These differences were quantified by an RMSD of 12.968 (using the pymol align command).

Figure 98: Aligned models of 222 generated using a template-based approach with Modeller (considered best practice in 2018) and Alphafold2 (considered best practice now in 2022). Aligned using the pymol align command, giving this figure and a RMSD value output of 12.229 which shows agreement but does elude to some differences (quantifying how different they are in angstroms). A lower RMSD shows more agreement. Pink Model is the Alphafold2 model, green is the homology model generated in **results chapter 3** using BLAST and Modeller. Alphafold2 is an end to end pipeline where all steps take place in the one tool whereas the homology modelling method starts with a database search, template selection, alignment, modeller generation and finally separate QMEAN assessment and Ramachandran plot generation as model validation.

When exploring how the two different tools go about modelling there is good explanation for why the models produced show some differences. Alphafold2 utilises neural networking bringing together structural information from multiple known proteins found in the PDB that meet a threshold level of identity, to give the best possible prediction for the whole sequence rather than using only one single best identified template as in the older homology modelling approach. This can be useful when templates don't have good levels of identity. Here the identity level was good for the template based modelling approach used and it's important to point out that

neither are incorrect and are both justified. The only way to prove/ disprove the model would be experimentally but given the inherent flexibility evident in both and the experience of working with the 222 protein as being variable/ inconsistent it may be that there is not only one correct model.

These mutations to Trp in this chapter have clearly changed something about these proteins and introduced the cleavage issue which we are attributing (based on bioinformatic analysis) to possible inaccessibility/ occlusion of the cleavage site. Specifically, and interestingly this change in exposure is limited to residues Val 382 and Leu 383. There is no mention of such a criticality of these two 3C protease recognition sequence residues in a search of the literature. So it may be that in the work here this is an unexpected novel finding. Trp is a large residue, which makes occlusion a seem feasible explanation, especially if its introduction also constrained movement/ flexibility in the fusion protein possibly by introducing self-assembly, aggregation [571] or aromatic stacking [572]. Alternatively, it could be that these proteins form dimers or higher order structures which occludes the cleavage site, this would need to be investigated using Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS) analysis [474].

5.4.1 Highlights

Docking strategy was well researched and executed, the Hyp residues in the TII gelatin peptide posed an issue as did the flexibility in protein-peptide docking. Both 222 and TII gelatin largely lack defined structure so make flexibility even more of a challenge here. In response a logical approach and thorough use of all available peptide structure prediction generating tools was employed to generate a peptide

library to provide a valid sampling of conformer options. Then HADDOCK was identified as the only tool that could dock the TII gelatin peptide with its Hyp modifications, it also allowed input of NMR identified protein residues from earlier work known to be involved in binding. These NMR residues were used to guide the process making it information driven, not blind. Then a critical matched binding order validation step was used with any selected peptides to check agreement with the docking order from the first chapter.

Good levels of expression of tagged protein were achieved looking at the intensity of the His-MBP purified protein shown in the SDS PAGE analyses. Subsequently we were unable to remove the MBP fusion tag to yield pure POI. 222W and 222W-CS1 were therefore never tested for *in vitro* binding as intended and so could still be the better binding mutants we seek. Despite efforts including Urea cleavage we couldn't find a method to produce native untagged protein in the time available remaining of this project. Periplasmic expression was a strategy worth trialling as the periplasm is a better environment for the formation of disulfide bonds. Although it was a more difficult purification as there was no pause point where cells could be stored frozen as is common with most cytoplasmic expression strategies. With the periplasmic approach it was important to keep the periplasm intact and maintain segregation of *E. coli* contaminant proteins. However, this method too was never replicated to test cleavage despite promising first expression results larger purification volumes and time constraints prevented further work with this strategy.

5.4.2 Future work:

Ultimately an alternative expression system such as yeast could be a strategy worth pursuing with the aim of generating better binding mutant proteins. Yeast is a eukaryote, therefore may be better suited to achieving appropriate folding (in the endoplasmic reticulum) and has the appropriate cellular machinery capacity to carry out PTMs (e.g., disulfide bond formation) [555, 573]. Yeast would not be that much more effort or cost to employ as an expression system compared to *E.coli* [574]. It may be that in yeast a solubility tag is not required to attain soluble protein. This would resolve the cleavage issue that halted the work with mutants in this chapter from progressing to planned characterisation. Self-cleavable tags are another interesting alternative that may also remedy the cleavage issue here and allow acquisition of 222W and 222W-CS1 proteins [575].

De novo binding protein design library would be an alternative approach to explore to generate binding variants for testing, but well beyond the scope of the work here. A recent paper from Cao et al [576] describes a strategy for designing *de novo* binding mutants starting with a broad exploration of the vast space of possible binding modes to a selected region of a protein surface, and then intensifying the search within the vicinity of the most promising binding modes. This approach would facilitate the design of peptide binding proteins from the target structure alone [576].. This may be an alternative way to tackle the current aim of attaining a mutant with higher binding affinity to TII gelatin.

6 General discussion and conclusion

The aim of this project was to develop and characterise mutant CBD proteins to be utilised as a strategy to deliver MSCs into damaged regions of OA joints. Specifically using proteins targetting binding to TII gelatin, a TII collagen degradation product available in abundance at the articular surface in OA damaged joints, representing an ideal binding target. TII gelatin is available in OA joints from mildly affected to those more severely affected with endstage OA, making it a non-discriminatory targeting strategy. This means that, if such a protein was taken forward in further work to develop it as a therapeutic, it would potentially be beneficial to the full range of severity/ progression stages seen in OA patients.

Importantly though, MMP-2 from which the 222 protein was developed by the group and the starting point for this project discriminates in binding only to TII gelatin and not intact functional TII collagen. This is an important distinction, significant to the treatment modality and worth highlighting again here. MMP-9 binds TII gelatin with a higher affinity than MMP-2, but it doesn't discriminate in this way between TII collagen and its degraded form of TII gelatin [301]. This strategy aims to target MSCs to damaged degraded cartilage lesions, not intact regions where MSC engraftment and differentiation to facilitate regeneration are not required therefore MMP-2 derived proteins are a better starting point for this therapeutic approach.

6.1 Disease summary

OA is the most common chronic degenerative joint disorder in the world and one of the most common sources of pain and disability reported in the elderly [577].

Whilst considerable heterogeneity exists in defining OA epidemiologically, the evidence is clear and conclusive in defining age as the single greatest risk factor. OA presents a significant burden to healthcare and society globally. OA is a complex, heterogeneous, multifactorial condition with a range of systemic, genetic, biomechanical, and environmental contributing factors. It is a dynamic disease, characterised by a disruption in cartilage homeostasis.

6.2 OA treatments summary

Most OA treatments to date have had limited successes, reporting pain relief and functional improvement rather than regeneration of the damaged articular cartilage or inflammation alleviation. These limited improvements are also relatively short lived. Without repair and regeneration, the debilitating damage and inflammation causing lack of cartilage functionality in the OA joint is not directly addressed meaning one of the main features of the disease is not modified/ remedied. In recent years new therapeutic strategies for OA treatment have shifted focus towards the regeneration of damaged cartilage and inflammation reduction [578]. Therapeutics achieving this would be the first and long awaited DMOADs with the potential to provide lifechanging amelioration to OA patients.

Current therapies insufficiently meet clinical need, there are currently no effective pharmacological or biological therapies that can restore the original structure and function of the OA joint.

6.2.1 MSCS in OA treatment

Caplan was the first to develop the concept of MCS as a therapy in 1991 [210]. He then followed this with early animal studies involving their use [579]. Since MSCs were first trialled as an injectable OA treatment by Centeno et al in a clinical trial of MSC injection into the human knee in 2008 [265] there have then been significant developments made in our understanding of MSCs potential as simply therapeutic chondroprogenitors.

An obvious question is why not deliver chondrocytes directly instead of MSCs that are intended to follow a chondrogenic differentiation path once administered into OA lesions. Chondrocytes have been used for years via autologous surgical implantation rather than injection in pursuit of this. Surgical implantation of chondrocytes from a donor however is complicated by immunogenicity and by MHC expression [580]. MSCs however do not present this problem [581, 582]. This however is not the primary limitation and issue with chondrocytes, it is their lack of trophic repair properties which makes them the inferior joint therapy. MSCs can act through trophic repair even if they attach only to soft tissues in the joint. To give symptomatic relief MSCs don't need to engraft after injection just to be present in the vicinity of the joint to act via trophic factors and provide symptom relief [583-586].

One development example that could be applied when pursuing MSCs as an OA therapeutic is the identification of a method of predicting MSCs with an enhanced capacity for chondrogenesis. The Receptor Tyrosine Kinase-Like Orphan Receptor 2 (ROR2) cell surface marker was found to be upregulated in the most chondrogenic cells distinguishing them from within the mixed population of MSC cells often utilised

[587]. So, using cells with this marker would increase the chances of chondrogenesis.

MSCs have progressed into clinical use in OA and thus far have shown good results in terms of pain management and improved motor function [263, 264, 267, 588]. However, biomechanically matched boundary-less regenerated cartilage is still sought following their use [589, 590]. Targetting and adhesion/ retention of MSCs specifically where needed into sites of OA damage remains an outstanding challenge. Although preclinical attempts have been made to address this [591, 592], no such strategies have translated to the clinic successfully yet. Until resolved the potential for MSCs to modify the course of OA disease is limited as evidenced by clinical trials so far, which have produced unclear and inconsistent results [593]. No MSC treatment currently trialled has managed to produce a neocartilage with biomechanical properties to match that of the cartilage which it aims to replace. A question remains as to whether regenerated cartilage is sufficiently integrated with any healthy/ functioning tissue pre-existing in the joint, improving adherence of MSCs to damaged OA seems a good way to promote this.

Invossa warrants special mention as a particularly promising alternative therapy to MSC injection, recently trialled for OA, utilising allogeneic chondrocytes as a delivery vehicle for a gene therapy. The therapy involves using chondrocytes transformed and transduced with a retrovirus to express a growth factor (TGF-β1) known to be involved in cartilage development and maturation, here it is the growth factor driving the joint repair rather than the chondrocytes [122, 123]. Results of a phase 3 clinical trial showed Invossa modifies the micro-environment of the joint to

be amenable for regeneration, with reduced pain and improved function reported [124, 125]. Significantly early results from this trial also seem to show structural improvement, as cartilage thickness was increased, determined via MRI Further studies are subsequently required to probe this and explore whether it is correct to classify this therapy as the first DMOAD. Presently regulatory concerns have halted progress, if clinical studies are reinstated this could be a very exciting therapeutic option for OA patients.

Combinatorial and inducible strategies are now also being considered more extensively, which are an interesting approach to addressing OAs multifactorial complexity, pathology and etiology [594]. As a multifactorial disease with more than one etiology more than one strategy will likely be required in its treatment. MSCs have the potential to provide pain relief, functional improvement, regeneration of the damaged ECM, reduced inflammation and immunomodulation which is a considerable breadth.

6.3 Summary of major findings, interesting results, and aspects of this work

In the first results chapter of this thesis, **chapter 3**, alanine mutagenesis revealed N 11, 69 and 127 were the most important residues in binding interaction for 222 with TII gelatin. This was determined using the plate binding assay to test alanine binding mutant proteins, M1 in which these three equivalent binding N residues were mutated to alanine was identified as the most reduced in binding potency. M1 had a reduced binding affinity measured at 1005 ±490nM compared to 222 which had a kD of 3.9 ±0.7nM, the largest loss of affinity of the alanine mutants tested here.

This finding was used to inform the work in **chapter 5** where mutants with enhanced binding were sought. A mutation of these residues to W rather than N, was predicted to infer a tighter binding than that of 222 to TII gelatin, using the molecular docking tool HADDOCK. 222 had a HADDOCK score of -150.301 ±0.783 whereas 222W had a HADDOCK score of 157.301 ±2.112, a lower score indicated better binding [520].

In **chapter 4** the Camsol tool [443] was used successfully to design a more soluble novel mutant named CS6, with a CamSol score of -0.028071 compared to 222 which had a CamSol score of -0.295891 (with a higher score indicating a more soluble protein). CS6 as the most soluble mutant predicted using the tool was then taken to *in vitro* work, being expressed, and recovered from the cytoplasm of *E. coli* shuffle cells. A precipitant based solubility assay [467] was then adapted and utilised to give a measure of apparent solubility, giving an indicative comparable measure of how soluble the protein was. This assay was employed with CS6, 222 and CBD, CBD and 222 have had binding affinity quantified previously but are also proteins with binding affinity that has been quantified here in this thesis as well. Then in a solubility assay, ammonium sulfate was determined as the precipitant that gave better resolution of apparent solubility and distinguished more clearly between mutants. Apparent solubility for CS6 was between 12-66mg/mL depending on the interpolation plot used. CBD and 222 have similar solubility of 7 and 6 mg/mL.

Clear identification of essential residues for binding and improved solubility were made. In this work we have designed and produced CS6, a mutant of 222 with confirmed superior solubility. Then the idea that it is very likely a balancing act to

achieve both potent binding and improved solubility was highlighted. A mutant protein optimised for binding, solubility, and stability with improvement to all three properties would be the optimal outcome but improvement of binding and maintenance of solubility and stability would be enough to progress or improvement of solubility and stability but maintained binding potency to the level of 222. This will be a key consideration going forward beyond the project here. Suitability of nanoDSF to deal with variable structured proteins largely unfolded to begin proteins was also a key finding. So although nanoDSF as used here and is a widely applied in the literature, as it relies on loss of structure an alternate assessments would need to be used in future.

The predicted improved binding affinity mutants that came out of the HADDOCK predictions in **chapter 5** were unfortunately not confirmed due to lack of fusion tag cleavage when they were expressed *in vitro*. So 222W mutant was never actually tested with the plate binding assay as planned, to elucidate if improved binding as predicted was attained *in vitro*. In future, self-cleaving tags could be explored as a possible means to get around this issue [575].

The work here has brought to attention the unfolded starting state of 222 and other CBD mutants in this work which possibly points to a key feature and aspect of their functioning.

6.3.1 Bioinformatics value and limitations

6.3.2 To protein engineering

Bioinformatics is a very powerful scientific subdiscipline which employs computational methods to collect, store, analyse and disseminate biological data and information, such as DNA and protein sequences along with annotations. Bioinformatics has vast applications with some of the largest medical challenges at present, evidenced most recently during the COVID-19 pandemic [492, 521]. Implementation of *in silico* studies in COVID-19 research allowed timely sequencing of the SARS-CoV-2 genome at a time when it was needed most but also properly analysed the sequencing errors, evolutionary relationship, genetic variations and putative drug candidates within a very short time period [595-597].

In silico binding affinity prediction for larger biomolecules such as the protein to peptide here in this thesis is still an imperfect bioinformatic application, with tools still under development/ not emerged as entirely fit for purpose yet. This is largely due to the conformational selection paradigm, which assumes that the relevant conformation of the target will be sampled in the unbound species structural file inputs used in docking [598]. Docking currently only works to give an assessment of binding (that is indicative not quantitative). Very much like the CamSol scoring, that gave a determination that the six mutations were increasingly more soluble compared to each other but didn't predict empirical differences that would be observed *in vitro*. Neither *in silico* prediction gave a quantification just a ranking comparable to each other, relative not absolute.

6.3.3 To this project specifically

This thesis has described the successful application of an array of bioinformatic tools to direct *in vitro* efforts in developing novel gelatin binding proteins to target therapeutics in the OA joint. The aims of this project were ambitious and without the bioinformatic tools used here to guide and provide a rational approach in all three chapters the results in this project would not have been attained. Within this project, structure-based tools were selected as superior to sequence-only based ones. This, as a rule, is best practice given that how a protein is folded can have large implications for residue exposure and proximity [496, 497].

Reducing the number of mutants to produce and study in **chapter 3** was key in directing the work. The previous NMR work of the Hollander group identified fifteen residues known to be involved in binding of 222 to TII gelatin. Alanine mutagenesis was utilised as the start point in elucidating the individual importance of selected binding residues equivalent in all three modules of 222. Given that identifying key residues that direct binding was not the only aim of this thesis, fifteen alanine mutants would have been too many proteins to work on within the timescale of this PhD project. Therefore, first and foremost a strategy to select alanine mutants from the possible fifteen was required. The tools used in this chapter reduced the number of possible alanine mutant proteins to test *in vitro* to four a much more amenable number than fifteen.

In silico to *in vitro* experimental translation is also a big point of consideration relevant to work presented within this thesis and when using *in silico* analyses and tools/ predictors in the wider discipline of protein engineering. *In silico* tools often

provide predictions giving relative comparisons, so understandably discordance exists, and tools are advancing or new alternates emerging constantly. PPI-Affinity is a new web tool for the prediction and optimization of Protein−Peptide and Protein−Protein Binding Affinity [599] which would have been a definite contender for use in **chapter 5** had it been available when the work was undertaken. Usually when utilising bioinformatics multiple tools are recommended for use to generate reliability in results, but this wasn't possible here due to no other suitable tool being available. The main issue encountered when exploring docking strategies was the Hyp PTM in TII gelatin. Only the HADDOCK tool was compatible with this feature of the peptide so it would be interesting to determine if this new tool PPI-Affinity had the capacity for use with our peptide.

Also Alphafold2 which was used to model the entire MBP-3C-222W fusion protein in **chapter 5** [392] would have been the now best practice choice for modelling 222 in **chapter 3**. But both are valid approaches and rely upon good scientific theory and evidence.

In **chapter 3** predictions of the selected mutants being stable, and solvent exposed enough to affect binding when mutated to alanine were correct as all alanine mutants were expressed in a soluble recoverable form and did have reduced binding affinity compared to 222.

6.3.4 Recombinant protein expression: victories, trials, and tribulations within this project

In vitro optimisation was laborious and extensive particularly with M1 in **chapter 3**. Even though options for trial conditions were not exhausted, it wouldn't

have been feasible to try everything as there are so many variables at play meaning that not all could ever have been tested. An extensive range of factors were trialled, e.g., altering growth conditions such as media, temperature and inducer concentration, use of culture additives, different *E.coli* cell lines, employ of alternate fusion tags and vectors, switching lysis method, changing the cellular expression site and chaperone co-expression. The work was systematic, rational, and based upon sound theory/ evidence. This did pay off to generate protein for *in vitro* analysis, but it is also worth noting that the proteins in **chapter 3** were never intended to go any further than these binding experiments. The alanine mutations were designed to decrease binding and thereby reveal individual binding importance of specific residues. Importantly, enough mutant protein was acquired to complete experiments and answer questions regarding binding and solubility in **chapters 3 and 4.** In future work automation would streamline the process, allow more high-throughput trialling of multiple proteins or conditions in tandem for optimisation. The use of multiwell plates limiting volumes to ration finite resources such as protein saves time and was already used in this project, automation would only further improve upon and could be adapted to suit already established and validated experimental methods [435].

Issues were faced throughout this project with soluble protein expression and recovery. Six disulfides were challenging for the *E. coli* expression system. Moving to a yeast expression system such as *Pichia. Pastoris* would be a relatively simple modification that potentially could see large improvements in the feasibility of any future CBD mutant protein binding work [600]. Yeast may be better suited to dealing with the PTMs such as disulfide bond formation than the more basic and readily used *E. coli* bacterial expression system. Engineered cell lines such as the shuffle *E.*

coli cells utilised in this project can be helpful with this, but yeast is not that much more expensive or difficult of an expression system to work with to maximise chances of success in acquiring higher yields of soluble protein.

The oxidising cytoplasm is the conventional region for recombinant bacterial protein expression. Periplasmic expression is an interesting alternative strategy that was trialled with the **chapter 5** mutants 222W and 222W-CS1. The periplasm is a reducing environment, so is better suited to disulfide bond formation. Unfortunately, although SDS-PAGE indicated good levels of fusion protein expression (MBP-3C-POI), fusion tag cleavage planned to be trialled in a subsequent repeat never went ahead considering the *in silico* investigation of 3C protease cleavage occlusion.

6.3.5 Implications and Future directions

It has been shown in this thesis that these proteins are difficult to work with, particularly as they seem to lack defined secondary structure. Rather than trying to mutate a CBD protein and balance characteristics, *de novo* binding protein design library would be an alternative approach. Essentially building a protein from scratch would be the next direction to take this work. If an optimal protein was attained, the previous work of the Hollander group has already adapted, developed, and validated a method of coating the membrane of MSCs in suspension and intraarticular injection is confirmed as the delivery method envisioned for this MSC treatment. Coating MSCs in suspension is an important detail as it means the cells aren't subject to tryptic removal from tissue culture plastic, a step which could have damaging side effects and affect viability of the MSCs and any protein coating on them. This method was already trialled *in vitro* with 222 to target MSC adherence

and assessed using a plate binding assay. Cells were confirmed to bind significantly better to TII gelatin when coated with 222 [235]. Although 222 is only optimal for binding, any other proteins developed going forward would utilise this tried and tested methodology demonstrated in this thesis and previous work of the group [235].

6.4 Concluding Statement

In conclusion, the findings and methods presented in this thesis form a solid foundation on which future work can be based. Throughout all three results chapters in this thesis bioinformatic tools were identified from the literature and applied successfully to help direct and produce a body of work that without their employ would not have been possible. Bioinformatics is a hastily advancing field and there are tools available now at this time of thesis write up that were not available at the commencement of this project. Alphafold2 is now the gold standard approach to modelling that was not available when 222 was homology modelled using what is now the 'archaic' by comparison modeller software. Alphafold2 was adopted in the cleavage occlusion investigation undertaken as part of the results chapter where fusion proteins models were required to look for difference between proteins that cleaved (the CS6 control) and failed to cleave mutants (222W and 222W-CS1). Surface exposure of two residues of the cleavage site were identified as less exposed (V382 and L383), a potentially novel finding.

One of the most important findings of this project was that these proteins are not all folded to begin, even 222 which at the outset of this project was considered a CBD fold mimicking protein 'a folded fibronectin domain-rich protein', it now seems

much more likely when you bring together some of the key findings of this work starting from the docking models that show very little notable defined structure and are largely flexible at least in the unbound state. Given the reported inconsistencies and sometimes not replicable findings such as the 20mM sucrose result, general variability in yield. CD with non-discriminate spectral fit patterns, none with good levels of agreement across all proteins. There is a fine balance between protein properties and to ‘fix’ one is often to ‘break another’ as was seen with the solubility mutant CS6. Which was a novel protein designed, expressed, and characterised in the work here for the first time. The design undertaken using the predicted 222 structure from the first results chapter was used to generate using an identified structural tool (CamSol), so this work proves its validity and efficacy.

The TII gelatin peptide is a changing target without fixed conformation, also 222 and its variants are also flexible/ moving. In the plate assay as in the joint. Its no fixed conformation state makes it a ‘moving target’. It is a valid target but more needs to be understood regarding how the binding occurs and to confirm as I predict at the conclusion here 222 and its derivatives have an introduction of structure upon binding.

Protein coated MSC targetting of TII gelatin is worth further pursuit as an OA therapeutic and can now bring everything together and use lessons learnt from this work to develop high-throughput screens designed to balance properties of optimised proteins designed for purpose. The findings presented in this thesis are significant and show more than anything how bioinformatics can aid in protein engineering. This outcome offers a new possible approach in OA research and

development is progressing in the right direction, with each experiment moving closer to the target, and hopefully this will lead to the emergence of a new disease modifying MSC therapeutic. Given the complexity of OA, a single therapy is not likely to be curative alone. OA is a complex, multifactorial, condition now understood to effect the entire joint not just the cartilage [80, 601] therefore the most promising strategies address both symptoms and structural changes to maximise efficacy [155, 602].

7 References

1. Juneja, P., A. Munjal, and J.B. Hubbard, *Anatomy, Joints*, in *StatPearls*. 2022, StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.: Treasure Island (FL).

2. Versus Arthritis. *Osteoarthritis (OA)*. n.d. [cited 2019 January 26]; Available from: https://www.versusarthritis.org/about-arthritis/conditions/osteoarthritis/

3. Bielajew, B.J., et al., *Knee orthopedics as a template for the temporomandibular joint.* Cell reports. Medicine, 2021. **2**(5): p. 100241-100241.

4. Pabbruwe, M.B., et al., *Repair of meniscal cartilage white zone tears using a stem cell/collagen-scaffold implant.* Biomaterials, 2010. **31**(9): p. 2583-91.

5. Whitehouse, M.R., et al., *Repair of Torn Avascular Meniscal Cartilage Using Undifferentiated Autologous Mesenchymal Stem Cells: From In Vitro Optimization to a First-in-Human Study.* Stem Cells Transl Med, 2017. **6**(4): p. 1237-1248.

6. Pabbruwe, M.B., et al., *Induction of cartilage integration by a chondrocyte/collagen-scaffold implant.* Biomaterials, 2009. **30**(26): p. 4277-86.

7. Hollander, A.P. and A. Salerno, *The biology of mesenchymal stem/stromal cells in the treatment of osteoarthritis.* Journal of Cartilage & Joint Preservation, 2022. **2**(1): p. 100035.

8. Dang, A.C. and A.C. Kuo, *Cartilage Biomechanics and Implications for Treatment of Cartilage Injuries.* Operative Techniques in Orthopaedics, 2014. **24**(4): p. 288-292.

9. Buckwalter, J.A. and J.A. Martin, *Osteoarthritis.* Advanced Drug Delivery Reviews, 2006. **58**(2): p. 150-167.

10. Aigner, T., et al., *Osteoarthritis: Pathobiology—targets and ways for therapeutic intervention.* Advanced Drug Delivery Reviews, 2006. **58**(2): p. 128-149.

11. Hunter, D.J. and F. Eckstein, *Exercise and osteoarthritis.* Journal of anatomy, 2009. **214**(2): p. 197-207.

12. Heidari, B., *Knee osteoarthritis prevalence, risk factors, pathogenesis and features: Part I.* Caspian journal of internal medicine, 2011. **2**(2): p. 205-212.

13. Lepage, S.I.M., et al., *Beyond Cartilage Repair: The Role of the Osteochondral Unit in Joint Health and Disease.* Tissue Eng Part B Rev, 2019. **25**(2): p. 114-125.

14. Baumann, C.A., et al., *Articular Cartilage: Structure and Restoration*, in *Joint Preservation of the Knee: A Clinical Casebook*, A.B. Yanke and B.J. Cole, Editors. 2019, Springer International Publishing: Cham. p. 3-24.

15. Eschweiler, J., et al., *The Biomechanics of Cartilage-An Overview.* Life (Basel, Switzerland), 2021. **11**(4): p. 302.

16. Benninghoff, A., *Form und Bau der Gelenkknorpel in ihren Beziehungen zur Funktion.* Zeitschrift für Zellforschung und Mikroskopische Anatomie, 1925. **2**(5): p. 783-862.

17. Klika, V., et al., *An overview of multiphase cartilage mechanical modelling and its role in understanding function and pathology.* Journal of the Mechanical Behavior of Biomedical Materials, 2016. **62**: p. 139-157.

18. Sophia Fox, A.J., A. Bedi, and S.A. Rodeo, *The basic science of articular cartilage: structure, composition, and function.* Sports health, 2009. **1**(6): p. 461-468.

19. Hall, B.K., *Chapter 4 - Invertebrate Cartilages, Notochordal Cartilage and Cartilage Origins*, in *Bones and Cartilage (Second Edition)*, B.K. Hall, Editor. 2015, Academic Press: San Diego. p. 63-78.

20. Rojas, F.P., et al., *Molecular adhesion between cartilage extracellular matrix macromolecules.* Biomacromolecules, 2014. **15**(3): p. 772-780.

21. Han, E., et al., *Contribution of proteoglycan osmotic swelling pressure to the compressive properties of articular cartilage.* Biophysical journal, 2011. **101**(4): p. 916-924.

22. Nia, H.T., C. Ortiz, and A. Grodzinsky, *Aggrecan: approaches to study biophysical and biomechanical properties.* Methods Mol Biol, 2015. **1229**: p. 221-37.

23. Hosoda, N., et al., *Depth-Dependence and Time-Dependence in Mechanical Behaviors of Articular Cartilage in Unconfined Compression Test under Constant Total Deformation.* Journal of Biomechanical Science and Engineering, 2008. **3**: p. 209-220.

24. Gastaldi, D., et al., *Effect of the anisotropic permeability in the frequency dependent properties of the superficial layer of articular cartilage.* Computer Methods in Biomechanics and Biomedical Engineering, 2018. **21**(11): p. 635-644.

25. Chang, L.R., G. Marston, and A. Martin, *Anatomy, Cartilage*, in *StatPearls*. 2022, StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.: Treasure Island (FL).

26. Gao, Y., et al., *The ECM-cell interaction of cartilage extracellular matrix on chondrocytes.* BioMed research international, 2014. **2014**: p. 648459-648459.

27. Buckwalter, J. and H. Mankin, *Articular cartilage: part I.* Journal of Bone and joint surgery, 1997. **79**(4): p. 600.

28. Bhosale, A.M. and J.B. Richardson, *Articular cartilage: structure, injuries and review of management.* British Medical Bulletin, 2008. **87**(1): p. 77-95.

29. Loeser, R.F., *Age-related changes in the musculoskeletal system and the development of osteoarthritis.* Clinics in geriatric medicine, 2010. **26**(3): p. 371-386.

30. de Rezende, M.U. and G.C. de Campos, *Is osteoarthritis a mechanical or inflammatory disease?* Revista Brasileira de Ortopedia (English Edition), 2013. **48**(6): p. 471-474.

31. van den Bosch, M.H.J., *Inflammation in osteoarthritis: is it time to dampen the alarm(in) in this debilitating disease?* Clinical & Experimental Immunology, 2019. **195**(2): p. 153-166.

32. Maldonado, M. and J. Nam, *The role of changes in extracellular matrix of cartilage in the presence of inflammation on the pathology of osteoarthritis.* BioMed research international, 2013. **2013**: p. 284873-284873.

33. Horkay, F., et al., *Structure and Properties of Cartilage Proteoglycans.* Macromolecular symposia, 2017. **372**(1): p. 43-50.

34. Chandran, P.L. and F. Horkay, *Aggrecan, an unusual polyelectrolyte: review of solution behavior and physiological implications.* Acta biomaterialia, 2012. **8**(1): p. 3-12.

35. Poole, R., et al., *Composition and Structure of Articular Cartilage: A Template for Tissue Repair.* Clinical Orthopaedics and Related Research, 2001. **391**: p. S26-S33.

36. Wyckoff, R.W.G., R.B. Corey, and J. Biscoe, *X-Ray Reflections of Long Spacing from Tendon.* Science, 1935. **82**(2121): p. 175-176.

37. Ramachandran, G.N. and G. Kartha, *Structure of Collagen.* Nature, 1954. **174**(4423): p. 269-270.

38. Miller, E.J. and V.J. Matukas, *Chick cartilage collagen: a new type of alpha 1 chain not present in bone or skin of the species.* Proc Natl Acad Sci U S A, 1969. **64**(4): p. 1264-8.

39. Ricard-Blum, S., *The collagen family.* Cold Spring Harbor perspectives in biology, 2011. **3**(1): p. a004978-a004978.

40. Bhowmick, M. and G.B. Fields, *Stabilization of collagen-model, triple-helical peptides for in vitro and in vivo applications.* Methods in molecular biology (Clifton, N.J.), 2013. **1081**: p. 167-194.

41. Motooka, D., et al., *The triple helical structure and stability of collagen model peptide with 4(s)-hydroxyprolyl-pro-gly units.* Peptide Science, 2012. **98**(2): p. 111-121.

42. Hollander, A.P., et al., *Increased damage to type II collagen in osteoarthritic articular cartilage detected by a new immunoassay.* Journal of Clinical Investigation, 1994. **93**(4): p. 1722-1732.

43. Hollander, A.P., et al., *Damage to type II collagen in aging and osteoarthritis starts at the articular surface, originates around chondrocytes, and extends into the cartilage with progressive degeneration.* J Clin Invest, 1995. **96**(6): p. 2859-69.

44. Zhang, Y. and J.M. Jordan, *Epidemiology of Osteoarthritis.* Clinics in geriatric medicine, 2010. **26**(3): p. 355-369.

45. Wallace, I.J., et al., *Knee osteoarthritis has doubled in prevalence since the mid-20th century.* Proceedings of the National Academy of Sciences of the United States of America, 2017. **114**(35): p. 9332-9336.

46. Arthritis Research UK. *State of musculoskeletal health 2017,*. 2017 [cited 2018 July, 01]; Available from: https://www.arthritisresearchuk.org/arthritis-information/data-and-statistics/state-of-musculoskeletal-health.aspx.

47. Katz, J.N., K.R. Arant, and R.F. Loeser, *Diagnosis and Treatment of Hip and Knee Osteoarthritis: A Review.* JAMA, 2021. **325**(6): p. 568-578.

48. Swain, S., et al., *Trends in incidence and prevalence of osteoarthritis in the United Kingdom: findings from the Clinical Practice Research Datalink (CPRD).* Osteoarthritis and Cartilage, 2020. **28**(6): p. 792-801.

49. Cui, A., et al., *Global, regional prevalence, incidence and risk factors of knee osteoarthritis in population-based studies.* eClinicalMedicine, 2020. **29**.

50. Shirley, E.D., M. Demaio, and J. Bodurtha, *Ehlers-danlos syndrome in orthopaedics: etiology, diagnosis, and treatment implications.* Sports Health, 2012. **4**(5): p. 394-403.

51. MacGregor, A.J., et al., *The genetic contribution to radiographic hip osteoarthritis in women: results of a classic twin study.* Arthritis Rheum, 2000. **43**(11): p. 2410-6.

52. Manek, N.J., et al., *The association of body mass index and osteoarthritis of the knee joint: an examination of genetic and environmental influences.* Arthritis Rheum, 2003. **48**(4): p. 1024-9.

53. Dempster, E.R. and I.M. Lerner, *Heritability of Threshold Characters.* Genetics, 1950. **35**(2): p. 212-36.

54. Aubourg, G., et al., *Genetics of osteoarthritis.* Osteoarthritis and Cartilage, 2022. **30**(5): p. 636-649.

55. Chen, D., et al., *Osteoarthritis: toward a comprehensive understanding of pathological mechanism.* Bone research, 2017. **5**: p. 16044-16044.

56. Liu, J., et al., *Rs143383 in the growth differentiation factor 5 (GDF5) gene significantly associated with osteoarthritis (OA)-a comprehensive meta-analysis.* Int J Med Sci, 2013. **10**(3): p. 312-9.

57. Gari, M.A., et al., *Identification of novel genetic variations affecting osteoarthritis patients.* BMC Medical Genetics, 2016. **17**(1): p. 68.

58. Paththinige, C.S., et al., *Split hand/foot malformation with long bone deficiency associated with BHLHA9 gene duplication: a case report and review of literature.* BMC Med Genet, 2019. **20**(1): p. 108.

59. Sharma, A.C., et al., *Association between Single Nucleotide Polymorphisms of SMAD3 and BMP5 with the Risk of Knee Osteoarthritis.* J Clin Diagn Res, 2017. **11**(6): p. Gc01-gc04.

60. Boer, C.G., et al., *Deciphering osteoarthritis genetics across 826,690 individuals from 9 populations.* Cell, 2021. **184**(18): p. 4784-4818.e17.

61. Knowlton, R.G., et al., *Genetic linkage of a polymorphism in the type II procollagen gene (COL2A1) to primary osteoarthritis associated with mild chondrodysplasia.* N Engl J Med, 1990. **322**(8): p. 526-30.

62. Kehayova, Y.S., et al., *Genetic and Epigenetic Interplay Within a COLGALT2 Enhancer Associated With Osteoarthritis.* Arthritis & Rheumatology, 2021. **73**(10): p. 1856-1865.

63. Tachmazidou, I., et al., *Identification of new therapeutic targets for osteoarthritis through genome-wide analyses of UK Biobank data.* Nat Genet, 2019. **51**(2): p. 230-236.

64. Reynard, L.N. and M.J. Barter, *Osteoarthritis year in review 2019: genetics, genomics and epigenetics.* Osteoarthritis and Cartilage, 2020. **28**(3): p. 275-284.

65. Gee, F., et al., *Allelic expression analysis of the osteoarthritis susceptibility locus that maps to chromosome 3p21 reveals cis-acting eQTLs at GNL3 and SPCS1.* BMC Med Genet, 2014. **15**: p. 53.

66. Day-Williams, A.G., et al., *A variant in MCF2L is associated with osteoarthritis.* Am J Hum Genet, 2011. **89**(3): p. 446-50.

67. Zengini, E., et al., *Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis.* Nature Genetics, 2018. **50**(4): p. 549-558.

68. Zieba, J.T., et al. *Notch Signaling in Skeletal Development, Homeostasis and Pathogenesis*. Biomolecules, 2020. **10**, DOI: 10.3390/biom10020332.

69. Brumwell, A., et al., *Identification of TMEM129, encoding a ubiquitin-protein ligase, as an effector gene of osteoarthritis genetic risk.* Arthritis Research & Therapy, 2022. **24**(1): p. 189.

70. Castaño-Betancourt, M.C., et al., *Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis.* PLOS Genetics, 2016. **12**(10): p. e1006260.

71. Steinberg, J., et al., *Decoding the genomic basis of osteoarthritis.* bioRxiv, 2020: p. 835850.

72. Man, G.S. and G. Mologhianu, *Osteoarthritis pathogenesis - a complex process that involves the entire joint.* J Med Life, 2014. **7**(1): p. 37-41.

73. Hochberg, M.C., et al., *Genetic epidemiology of osteoarthritis: recent developments and future directions.* Curr Opin Rheumatol, 2013. **25**(2): p. 192-7.

74. Loeser, R.F., et al., *Osteoarthritis: a disease of the joint as an organ.* Arthritis and rheumatism, 2012. **64**(6): p. 1697-1707.

75. Peng, Z., et al., *The regulation of cartilage extracellular matrix homeostasis in joint cartilage degeneration and regeneration.* Biomaterials, 2021. **268**: p. 120555.

76. Pritzker, K.P.H., et al., *Osteoarthritis cartilage histopathology: grading and staging.* Osteoarthritis and Cartilage, 2006. **14**(1): p. 13-29.

77. Waldstein, W., et al., *OARSI osteoarthritis cartilage histopathology assessment system: A biomechanical evaluation in the human knee.* Journal of Orthopaedic Research, 2016. **34**(1): p. 135-140.

78. Felson, D.T., *EPIDEMIOLOGY OF HIP AND KNEE OSTEOARTRRITIS 1.* Epidemiologic Reviews, 1988. **10**(1): p. 1-28.

79. Litwic, A., et al., *Epidemiology and Burden of Osteoarthritis.* British medical bulletin, 2013. **105**: p. 185-199.

80. Mobasheri, A. and M. Batt, *An update on the pathophysiology of osteoarthritis.* Annals of Physical and Rehabilitation Medicine, 2016. **59**(5): p. 333-339.

81. Wang, N., et al., *Mechanotransduction pathways in articular chondrocytes and the emerging role of estrogen receptor-α.* Bone Research, 2023. **11**(1): p. 13.

82. Zheng, L., et al., *The role of metabolism in chondrocyte dysfunction and the progression of osteoarthritis.* Ageing Research Reviews, 2021. **66**: p. 101249.

83. Glyn-Jones, S., et al., *Osteoarthritis.* Lancet, 2015. **386**(9991): p. 376-87.

84. Kim, H.A., et al., *The catabolic pathway mediated by Toll-like receptors in human osteoarthritic chondrocytes.* Arthritis Rheum, 2006. **54**(7): p. 2152-63.

85. Liu-Bryan, R. and R. Terkeltaub, *Chondrocyte innate immune myeloid differentiation factor 88-dependent signaling drives procatabolic effects of the endogenous Toll-like receptor 2/Toll-like receptor 4 ligands low molecular weight hyaluronan and high mobility group box chromosomal protein 1 in mice.* Arthritis Rheum, 2010. **62**(7): p. 2004-12.

86. Liu-Bryan, R., et al., *TLR2 signaling in chondrocytes drives calcium pyrophosphate dihydrate and monosodium urate crystal-induced nitric oxide generation.* J Immunol, 2005. **174**(8): p. 5016-23.

87. Wang, Q., et al., *Identification of a central role for complement in osteoarthritis.* Nat Med, 2011. **17**(12): p. 1674-9.

88. Happonen, K.E., et al., *Regulation of complement by cartilage oligomeric matrix protein allows for a novel molecular diagnostic principle in rheumatoid arthritis.* Arthritis Rheum, 2010. **62**(12): p. 3574-83.

89. Sjöberg, A.P., et al., *Short leucine-rich glycoproteins of the extracellular matrix display diverse patterns of complement interaction and activation.* Mol Immunol, 2009. **46**(5): p. 830-9.

90. Loeser, R.F., et al., *Articular chondrocytes express the receptor for advanced glycation end products: Potential role in osteoarthritis.* Arthritis Rheum, 2005. **52**(8): p. 2376-85.

91. Rasheed, Z., N. Akhtar, and T.M. Haqqi, *Advanced glycation end products induce the expression of interleukin-6 and interleukin-8 by receptor for advanced glycation end product-mediated activation of mitogen-activated protein kinases and nuclear factor-κB in human osteoarthritis chondrocytes.* Rheumatology (Oxford), 2011. **50**(5): p. 838-51.

92. Liu-Bryan, R., *Synovium and the innate inflammatory network in osteoarthritis progression.* Curr Rheumatol Rep, 2013. **15**(5): p. 323.

93. Hu, Y., et al., *Subchondral bone microenvironment in osteoarthritis and pain.* Bone Research, 2021. **9**(1): p. 20.

94. Woolf, A.D., T. Vos, and L. March, *How to measure the impact of musculoskeletal conditions.* Best Pract Res Clin Rheumatol, 2010. **24**(6): p. 723-32.

95. Chen, A., et al., *The Global Economic Cost of Osteoarthritis: How the UK Compares.* Arthritis, 2012. **2012**: p. 698709-698709.

96. Kwong, F.N.k. and M. Spector, *45 - Articular Cartilage A2 - Atala, Anthony*, in *Principles of Regenerative Medicine*, R. Lanza, J.A. Thomson, and R.M. Nerem, Editors. 2008, Academic Press: San Diego. p. 766-781.

97. Danišovič, Ľ., et al., *The tissue engineering of articular cartilage: Cells, scaffolds and stimulating factors*. Vol. 237. 2011. 10-7.

98. Nelson, L., J. Fairclough, and C.W. Archer, *Use of stem cells in the biological repair of articular cartilage.* Expert Opinion on Biological Therapy, 2010. **10**(1): p. 43-55.

99. Jeng, L., F. Ng kee Kwong, and M. Spector, *Chapter 42 - Articular Cartilage*, in *Principles of Regenerative Medicine (Second Edition)*, A. Atala, et al., Editors. 2011, Academic Press: San Diego. p. 761-777.

100. Kramer, W.C., K.J. Hendricks, and J. Wang, *Pathogenetic mechanisms of posttraumatic osteoarthritis: opportunities for early intervention.* International journal of clinical and experimental medicine, 2011. **4**(4): p. 285-298.

101. Camarero-Espinosa, S., et al., *Articular cartilage: from formation to tissue engineering.* Biomaterials Science, 2016. **4**(5): p. 734-767.

102. Temenoff, J.S. and A.G. Mikos, *Review: tissue engineering for regeneration of articular cartilage.* Biomaterials, 2000. **21**(5): p. 431-440.

103. Neogi, T., *The Epidemiology and Impact of Pain in Osteoarthritis.* Osteoarthritis and cartilage / OARS, Osteoarthritis Research Society, 2013. **21**(9): p. 1145-1153.

104. Sofat, N., V. Ejindu, and P. Kiely, *What makes osteoarthritis painful? The evidence for local and central pain processing.* Rheumatology, 2011. **50**(12): p. 2157-2165.

105. Hwang, H.S. and H.A. Kim, *Chondrocyte Apoptosis in the Pathogenesis of Osteoarthritis.* International Journal of Molecular Sciences, 2015. **16**(11): p. 26035-26054.

106. Zamli, Z. and M. Sharif, *Chondrocyte apoptosis: A cause or consequence of osteoarthritis?* Vol. 14. 2011. 159-66.

107. Grogan, S.P. and D.D. D'Lima, *Joint aging and chondrocyte cell death.* International journal of clinical rheumatology, 2010. **5**(2): p. 199-214.

108. Oldershaw, R.A., *Cell sources for the regeneration of articular cartilage: the past, the horizon and the future.* International journal of experimental pathology, 2012. **93**(6): p. 389-400.

109. Mobasheri, A., et al., *Chondrocyte and mesenchymal stem cell-based therapies for cartilage repair in osteoarthritis and related orthopaedic conditions.* Maturitas, 2014. **78**(3): p. 188-98.

110. Damia, E., et al., *Adipose-Derived Mesenchymal Stem Cells: Are They a Good Therapeutic Strategy for Osteoarthritis?* International Journal of Molecular Sciences, 2018. **19**(7).

111. Cui, D., et al., *Mesenchymal Stem Cells for Cartilage Regeneration of TMJ Osteoarthritis.* Stem Cells International, 2017. **2017**: p. 11.

112. Baghaban Eslaminejad, M. and E. Malakooty Poor, *Mesenchymal stem cells as a potent cell source for articular cartilage regeneration.* World journal of stem cells, 2014. **6**(3): p. 344-354.

113. Blanco, F.J., et al., *Osteoarthritis chondrocytes die by apoptosis: A possible pathway for osteoarthritis pathology.* Arthritis & Rheumatism, 2004. **41**(2): p. 284-289.

114. de l'Escalopier, N., P. Anract, and D. Biau, *Surgical treatments for osteoarthritis.* Annals of Physical and Rehabilitation Medicine, 2016. **59**(3): p. 227-233.

115. Steadman, J.R., et al., *Microfracture technique forfull-thickness chondral defects: Technique and clinical results.* Operative Techniques in Orthopaedics, 1997. **7**(4): p. 300-304.

116. Lee, J.J., et al., *Results of microfracture in the osteoarthritic knee with focal full-thickness articular cartilage defects and concomitant medial meniscal tears.* Knee surgery & related research, 2013. **25**(2): p. 71-76.

117. Redondo, M.L., A.J. Beer, and A.B. Yanke, *Cartilage Restoration: Microfracture and Osteochondral Autograft Transplantation.* J Knee Surg, 2018. **31**(3): p. 231-238.

118. Song, S.J. and C.H. Park, *Microfracture for cartilage repair in the knee: current concepts and limitations of systematic reviews.* Annals of Translational Medicine, 2019: p. 34.

119. Aae, T.F., et al., *Microfracture is more cost-effective than autologous chondrocyte implantation: a review of level 1 and level 2 studies with 5 year follow-up.* Knee Surg Sports Traumatol Arthrosc, 2018. **26**(4): p. 1044-1052.

120. Bae, D.K., K.H. Yoon, and S.J. Song, *Cartilage healing after microfracture in osteoarthritic knees.* Arthroscopy, 2006. **22**(4): p. 367-74.

121. Mithoefer, K., et al., *Clinical efficacy of the microfracture technique for articular cartilage repair in the knee: an evidence-based systematic analysis.* Am J Sports Med, 2009. **37**(10): p. 2053-63.

122. Erggelet, C. and P. Vavken, *Microfracture for the treatment of cartilage defects in the knee joint - A golden standard?* J Clin Orthop Trauma, 2016. **7**(3): p. 145-52.

123. Welton, K.L., et al., *Knee Cartilage Repair and Restoration: Common Problems and Solutions.* Clin Sports Med, 2018. **37**(2): p. 307-330.

124. Hoemann, C.D., et al., *Chondroinduction Is the Main Cartilage Repair Response to Microfracture and Microfracture With BST-CarGel: Results as Shown by ICRS-II Histological Scoring and a Novel Zonal Collagen Type Scoring Method of Human Clinical Biopsy Specimens.* Am J Sports Med, 2015. **43**(10): p. 2469-80.

125. Hulme, C.H., et al., *Identification of Candidate Synovial Fluid Biomarkers for the Prediction of Patient Outcome After Microfracture or Osteotomy.* Am J Sports Med, 2021. **49**(6): p. 1512-1523.

126. Davies, R.L. and N.J. Kuiper, *Regenerative Medicine: A Review of the Evolution of Autologous Chondrocyte Implantation (ACI) Therapy.* Bioengineering, 2019. **6**(1): p. 22.

127. Brittberg, M., *Autologous chondrocyte transplantation.* Clin Orthop Relat Res, 1999(367 Suppl): p. S147-55.

128. Barié, A., et al., *Prospective Long-term Follow-up of Autologous Chondrocyte Implantation With Periosteum Versus Matrix-Associated Autologous Chondrocyte Implantation: A Randomized Clinical Trial.* Am J Sports Med, 2020. **48**(9): p. 2230-2241.

129. Lysholm, J. and Y. Tegner, *Knee injury rating scales.* Acta Orthop, 2007. **78**(4): p. 445-53.

130. Collins, N.J., et al., *Measures of knee function: International Knee Documentation Committee (IKDC) Subjective Knee Evaluation Form, Knee Injury and Osteoarthritis Outcome Score (KOOS), Knee Injury and Osteoarthritis Outcome Score Physical Function Short Form (KOOS-PS), Knee Outcome Survey Activities of Daily Living Scale (KOS-ADL), Lysholm Knee Scoring Scale, Oxford Knee Score (OKS), Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), Activity Rating Scale (ARS), and Tegner Activity Score (TAS).* Arthritis Care Res (Hoboken), 2011. **63 Suppl 11**(0 11): p. S208-28.

131. Saris, D.B.F., et al., *Matrix-Applied Characterized Autologous Cultured Chondrocytes Versus Microfracture Two-Year Follow-up of a Prospective Randomized Trial.* The American journal of sports medicine, 2014. **42**.

132. Belk, J.W. and E. McCarty, *Editorial Commentary: Autologous Chondrocyte Implantation Versus Microfracture for Knee Articular Cartilage Repair: We Should Focus on the Latest Autologous Chondrocyte Implantation Techniques.* Arthroscopy, 2020. **36**(1): p. 304-306.

133. Shi, W.J., et al., *Biologic injections for osteoarthritis and articular cartilage damage: can we modify disease?* Phys Sportsmed, 2017. **45**(3): p. 203-223.

134. Kon, E., et al., *Matrix assisted autologous chondrocyte transplantation for cartilage treatment: A systematic review.* Bone & joint research, 2013. **2**(2): p. 18-25.

135. Bartlett, W., et al., *Autologous chondrocyte implantation versus matrix-induced autologous chondrocyte implantation for osteochondral defects of the knee: a prospective, randomised study.* J Bone Joint Surg Br, 2005. **87**(5): p. 640-5.

136. Brittberg, M., et al., *Matrix-Applied Characterized Autologous Cultured Chondrocytes Versus Microfracture: Five-Year Follow-up of a Prospective Randomized Trial.* Am J Sports Med, 2018. **46**(6): p. 1343-1351.

137. Ebert, J.R., et al., *A comparison of the responsiveness of 4 commonly used patient-reported outcome instruments at 5 years after matrix-induced autologous chondrocyte implantation.* Am J Sports Med, 2013. **41**(12): p. 2791-9.

138. Ogura, T., et al., *A 20-Year Follow-up After First-Generation Autologous Chondrocyte Implantation.* Am J Sports Med, 2017. **45**(12): p. 2751-2761.

139. Patil, S. and S.R. Tapasvi, *Osteochondral autografts.* Current reviews in musculoskeletal medicine, 2015. **8**(4): p. 423-428.

140. Hangody, L., et al., *Clinical Experiences With Autologous Osteochondral Mosaicplasty in an Athletic Population: A 17-Year Prospective Multicenter Study.* The American Journal of Sports Medicine, 2010. **38**(6): p. 1125-1133.

141. Branam, G.M. and A.Y. Saber, *Osteochondral Autograft Transplantation*, in *StatPearls*. 2022, StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.: Treasure Island (FL).

142. Torrie, A.M., et al., *Osteochondral allograft.* Current reviews in musculoskeletal medicine, 2015. **8**(4): p. 413-422.

143. Beer, A.J., et al., *Use of Allografts in Orthopaedic Surgery: Safety, Procurement, Storage, and Outcomes.* Orthopaedic Journal of Sports Medicine, 2019. **7**(12): p. 2325967119891435.

144. Evans, J.T., et al., *How long does a knee replacement last? A systematic review and meta-analysis of case series and national registry reports with more than 15 years of follow-up.* The Lancet, 2019. **393**(10172): p. 655-663.

145. Culliford, D., et al., *Future projections of total hip and knee arthroplasty in the UK: results from the UK Clinical Practice Research Datalink.* Osteoarthritis and Cartilage, 2015. **23**(4): p. 594-600.

146. Kohli, N., et al., *An In Vitro Comparison of the Incorporation, Growth, and Chondrogenic Potential of Human Bone Marrow versus Adipose Tissue Mesenchymal Stem Cells in Clinically Relevant Cell Scaffolds Used for Cartilage Repair.* Cartilage, 2015. **6**(4): p. 252-63.

147. Bliddal, H., A.R. Leeds, and R. Christensen, *Osteoarthritis, obesity and weight loss: evidence, hypotheses and horizons - a scoping review.* Obesity reviews : an official journal of the International Association for the Study of Obesity, 2014. **15**(7): p. 578-586.

148. Lo, G.H., et al., *Evidence that Swimming May Be Protective of Knee Osteoarthritis: Data from the Osteoarthritis Initiative.* PM & R : the journal of injury, function, and rehabilitation, 2020. **12**(6): p. 529-537.

149. Graham, G.G. and K.F. Scott, *Mechanism of action of paracetamol.* Am J Ther, 2005. **12**(1): p. 46-55.

150. Towheed, T.E., et al., *Acetaminophen for osteoarthritis.* Cochrane Database Syst Rev, 2003(2): p. Cd004257.

151. Ivers, N., I.A. Dhalla, and G.M. Allan, *Opioids for osteoarthritis pain: benefits and risks.* Canadian family physician Medecin de famille canadien, 2012. **58**(12): p. e708-e708.

152. Kijowski, R., *Risks and Benefits of Intra-articular Corticosteroid Injection for Treatment of Osteoarthritis: What Radiologists and Patients Need to Know.* Radiology, 2019. **293**(3): p. 664-665.

153. Conrozier, T., et al., *Viscosupplementation for the treatment of osteoarthritis. The contribution of EUROVISCO group.* Therapeutic advances in musculoskeletal disease, 2021. **13**: p. 1759720X211018605-1759720X211018605.

154. Pereira, T.V., et al., *Viscosupplementation for knee osteoarthritis: systematic review and meta-analysis.* BMJ, 2022. **378**: p. e069722.

155. Grässel, S. and D. Muschter, *Recent advances in the treatment of osteoarthritis.* F1000Res, 2020. **9**.

156. Grässel, S., F. Zaucke, and H. Madry, *Osteoarthritis: Novel Molecular Mechanisms Increase Our Understanding of the Disease Pathology.* J Clin Med, 2021. **10**(9).

157. Tong, L., et al., *Current understanding of osteoarthritis pathogenesis and relevant new approaches.* Bone Research, 2022. **10**(1): p. 60.

158. Cho, Y., et al., *Disease-modifying therapeutic strategies in osteoarthritis: current status and future directions.* Experimental & Molecular Medicine, 2021. **53**(11): p. 1689-1696.

159. McClurg, O., R. Tinson, and L. Troeberg, *Targeting Cartilage Degradation in Osteoarthritis.* Pharmaceuticals (Basel, Switzerland), 2021. **14**(2): p. 126.

160. Wang, W., D. Rigueur, and K.M. Lyons, *TGFβ signaling in cartilage development and maintenance.* Birth defects research. Part C, Embryo today : reviews, 2014. **102**(1): p. 37-51.

161. MacFarlane, E.G., et al., *TGF-β Family Signaling in Connective Tissue and Skeletal Diseases.* Cold Spring Harbor perspectives in biology, 2017. **9**(11): p. a022269.

162. Lee, B., *INVOSSA, a first-in-class of cell and gene therapy for osteoarthritis treatment: the phase III trial.* Osteoarthritis and Cartilage, 2018. **26**: p. S43-S44.

163. Lew, S., et al., *Long-term follow-up assessment of the safety and efficacy of INVOSSA-K INJ., a novel cell mediated gene therapy for treatment of osteoarthritis.* Osteoarthritis and Cartilage, 2019. **27**: p. S212.

164. Evans, C.H., *The vicissitudes of gene therapy.* Bone & joint research, 2019. **8**(10): p. 469-471.

165. Grol, M.W. and B.H. Lee, *Gene therapy for repair and regeneration of bone and cartilage.* Curr Opin Pharmacol, 2018. **40**: p. 59-66.

166. Oo, W.M., et al., *The Development of Disease-Modifying Therapies for Osteoarthritis (DMOADs): The Evidence to Date.* Drug Design, Development and Therapy, 2021. **Volume 15**: p. 2921-2945.

167. Cai, X., et al., *New Trends in Pharmacological Treatments for Osteoarthritis.* Frontiers in Pharmacology, 2021. **12**.

168. Hochberg, M.C., et al., *Effect of Intra-Articular Sprifermin vs Placebo on Femorotibial Joint Cartilage Thickness in Patients With Osteoarthritis: The FORWARD Randomized Clinical Trial.* Jama, 2019. **322**(14): p. 1360-1370.

169. Eckstein, F., et al., *Intra-articular sprifermin reduces cartilage loss in addition to increasing cartilage gain independent of location in the femorotibial joint: post-hoc analysis of a randomised, placebo-controlled phase II clinical trial.* Ann Rheum Dis, 2020. **79**(4): p. 525-528.

170. Guehring, H., et al., *The effects of sprifermin on symptoms and structure in a subgroup at risk of progression in the FORWARD knee osteoarthritis trial.* Semin Arthritis Rheum, 2021. **51**(2): p. 450-456.

171. Dório, M., et al., *Efficacy of platelet-rich plasma and plasma for symptomatic treatment of knee osteoarthritis: a double-blinded placebo-controlled randomized clinical trial.* BMC musculoskeletal disorders, 2021. **22**(1): p. 822-822.

172. Bennell, K.L., et al., *Effect of Intra-articular Platelet-Rich Plasma vs Placebo Injection on Pain and Medial Tibial Cartilage Volume in Patients With Knee Osteoarthritis: The RESTORE Randomized Clinical Trial.* JAMA, 2021. **326**(20): p. 2021-2030.

173. Becker, C., M.C. Fantini, and M.F. Neurath, *TGF-beta as a T cell regulator in colitis and colon cancer.* Cytokine Growth Factor Rev, 2006. **17**(1-2): p. 97-106.

174. Zhang, F., et al., *TGF-β induces M2-like macrophage polarization via SNAIL-mediated suppression of a pro-inflammatory phenotype.* Oncotarget, 2016. **7**(32): p. 52294-52306.

175. Blaney Davidson, E.N., P.M. van der Kraan, and W.B. van den Berg, *TGF-beta and osteoarthritis.* Osteoarthritis Cartilage, 2007. **15**(6): p. 597-604.

176. van Beuningen, H.M., et al., *Transforming growth factor-beta 1 stimulates articular chondrocyte proteoglycan synthesis and induces osteophyte formation in the murine knee joint.* Lab Invest, 1994. **71**(2): p. 279-90.

177. Yang, X., et al., *TGF-beta/Smad3 signals repress chondrocyte hypertrophic differentiation and are required for maintaining articular cartilage.* J Cell Biol, 2001. **153**(1): p. 35-46.

178. Kim, M.K., et al., *A Multicenter, Double-Blind, Phase III Clinical Trial to Evaluate the Efficacy and Safety of a Cell and Gene Therapy in Knee Osteoarthritis Patients.* Hum Gene Ther Clin Dev, 2018. **29**(1): p. 48-59.

179. Lee, H., et al., *TissueGene-C promotes an anti-inflammatory micro-environment in a rat monoiodoacetate model of osteoarthritis via polarization of M2 macrophages leading to pain relief and structural improvement.* Inflammopharmacology, 2020. **28**(5): p. 1237-1252.

180. Yazici, Y., et al., *A Phase 2b randomized trial of lorecivivint, a novel intra-articular CLK2/DYRK1A inhibitor and Wnt pathway modulator for knee osteoarthritis.* Osteoarthritis Cartilage, 2021. **29**(5): p. 654-666.

181. Tambiah, J.R.S., et al., *Comparing Patient-Reported Outcomes From Sham and Saline-Based Placebo Injections for Knee Osteoarthritis: Data From a Randomized Clinical Trial of Lorecivivint.* Am J Sports Med, 2022. **50**(3): p. 630-636.

182. Sampson, E.R., et al., *Teriparatide as a chondroregenerative therapy for injury-induced osteoarthritis.* Science translational medicine, 2011. **3**(101): p. 101ra93-101ra93.

183. McGuire, D., et al., *Study TPX-100-5: intra-articular TPX-100 significantly delays pathological bone shape change and stabilizes cartilage in moderate to severe bilateral knee OA.* Arthritis Research & Therapy, 2021. **23**(1): p. 242.

184. Lacy, S.E., et al., *Generation and characterization of ABT-981, a dual variable domain immunoglobulin (DVD-Ig(TM)) molecule that specifically and potently neutralizes both IL-1α and IL-1β.* mAbs, 2015. **7**(3): p. 605-619.

185. Wang, S.X., et al., *Safety, tolerability, and pharmacodynamics of an anti-interleukin-1α/β dual variable domain immunoglobulin in patients with osteoarthritis of the knee: a randomized phase 1 study.* Osteoarthritis and Cartilage, 2017. **25**(12): p. 1952-1961.

186. Fleischmann, R.M., et al., *A Phase II Trial of Lutikizumab, an Anti-Interleukin-1α/β Dual Variable Domain Immunoglobulin, in Knee Osteoarthritis Patients With Synovitis.* Arthritis Rheumatol, 2019. **71**(7): p. 1056-1069.

187. Cao, Z., et al., *Is Lutikizumab, an Anti-Interleukin-1α/β Dual Variable Domain Immunoglobulin, efficacious for Osteoarthritis? Results from a bayesian network meta-analysis.* BioMed research international, 2020. **2020**: p. 9013283-9013283.

188. Nair, A.S., *Tanezumab: Finally a Monoclonal Antibody for Pain Relief.* Indian journal of palliative care, 2018. **24**(3): p. 384-385.

189. Schnitzer, T.J., et al., *Effect of Tanezumab on Joint Pain, Physical Function, and Patient Global Assessment of Osteoarthritis Among Patients With Osteoarthritis of the Hip or Knee: A Randomized Clinical Trial.* Jama, 2019. **322**(1): p. 37-48.

190. Berenbaum, F., et al., *Subcutaneous tanezumab for osteoarthritis of the hip or knee: efficacy and safety results from a 24-week randomised phase III study with a 24-week follow-up period.* Annals of the Rheumatic Diseases, 2020. **79**(6): p. 800.

191. Hu, R., et al., *Clinical Outcomes of Tanezumab With Different Dosages for Patient With Osteoarthritis: Network Meta-Analysis.* Frontiers in Pharmacology, 2021. **12**.

192. Katz, J.N., *Tanezumab for Painful Osteoarthritis.* JAMA, 2019. **322**(1): p. 30-32.

193. Dakin, P., et al., *Efficacy and safety of fasinumab in patients with chronic low back pain: a phase II/III randomised clinical trial.* Annals of the Rheumatic Diseases, 2021. **80**(4): p. 509.

194. Gupta, A., et al., *Cell-free stem cell-derived extract formulation for treatment of knee osteoarthritis: study protocol for a preliminary non-randomized, open-*

label, multi-center feasibility and safety study. J Orthop Surg Res, 2021. **16**(1): p. 514.

195. Brebion, F., et al., *Discovery of GLPG1972/S201086, a Potent, Selective, and Orally Bioavailable ADAMTS-5 Inhibitor for the Treatment of Osteoarthritis.* J Med Chem, 2021. **64**(6): p. 2937-2952.

196. Bernard, K., et al., *FRI0393 BASELINE CHARACTERISTICS OF THE STUDY POPULATION IN ROCCELLA, A PHASE 2 CLINICAL TRIAL EVALUATING THE EFFICACY AND THE SAFETY OF S201086/GLPG1972 IN PATIENTS WITH KNEE OSTEOARTHRITIS.* Annals of the Rheumatic Diseases, 2020. **79**(Suppl 1): p. 794-795.

197. Chen, T., et al., *Update on Novel Non-Operative Treatment for Osteoarthritis: Current Status and Future Trends.* Frontiers in pharmacology, 2021. **12**: p. 755230-755230.

198. Zheng, W., et al., *Fisetin inhibits IL-1β-induced inflammatory response in human osteoarthritis chondrocytes through activating SIRT1 and attenuates the progression of osteoarthritis in mice.* Int Immunopharmacol, 2017. **45**: p. 135-147.

199. Till, J.E. and E.A. McCulloch, *A Direct Measurement of the Radiation Sensitivity of Normal Mouse Bone Marrow Cells.* Radiation Research, 1961. **14**(2): p. 213-222.

200. Zakrzewski, W., et al., *Stem cells: past, present, and future.* Stem Cell Research & Therapy, 2019. **10**(1): p. 68.

201. Hanna, J., et al., *Human embryonic stem cells with biological and epigenetic characteristics similar to those of mouse ESCs.* Proceedings of the National Academy of Sciences, 2010. **107**(20): p. 9222-9227.

202. National Research, C., B. Institute of Medicine Committee on the, and R. Biomedical Applications of Stem Cell, in *Stem Cells and the Future of Regenerative Medicine*. 2002, National Academies Press (US) Copyright 2002 by the National Academy of Sciences. All rights reserved.: Washington (DC).

203. Eridani, S., *Types of Human Stem Cells and Their Therapeutic Applications.* Stem Cell Discovery, 2014. **04**: p. 13-26.

204. Takahashi, K. and S. Yamanaka, *Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors.* Cell, 2006. **126**(4): p. 663-76.

205. Sharma, R., *iPS Cells-The Triumphs and Tribulations.* Dent J (Basel), 2016. **4**(2).

206. Martin, U., *Therapeutic Application of Pluripotent Stem Cells: Challenges and Risks.* Frontiers in medicine, 2017. **4**: p. 229-229.

207. Alonso-Goulart, V., et al., *Mesenchymal stem cells from human adipose tissue and bone repair: a literature review.* Biotechnology Research and Innovation, 2018. **2**(1): p. 74-80.

208. Berebichez-Fridman, R. and P.R. Montero-Olvera, *Sources and Clinical Applications of Mesenchymal Stem Cells: State-of-the-art review.* Sultan Qaboos University medical journal, 2018. **18**(3): p. e264-e277.

209. Minguell, J., A. Erices, and P. Conget, *Mesenchymal Stem Cells.* Experimental biology and medicine (Maywood, N.J.), 2001. **226**: p. 507-20.

210. Caplan, A.I., *Mesenchymal stem cells.* J Orthop Res, 1991. **9**(5): p. 641-50.

211. Caplan, A.I., *Mesenchymal Stem Cells: Time to Change the Name!* Stem Cells Transl Med, 2017. **6**(6): p. 1445-1451.

212. Pittenger, M.F., et al., *Mesenchymal stem cell perspective: cell biology to clinical progress.* npj Regenerative Medicine, 2019. **4**(1): p. 22.

213. Murphy, M.B., K. Moncivais, and A.I. Caplan, *Mesenchymal stem cells: environmentally responsive therapeutics for regenerative medicine.* Experimental & Molecular Medicine, 2013. **45**(11): p. e54-e54.

214. Gomez-Salazar, M., et al., *Five Decades Later, Are Mesenchymal Stem Cells Still Relevant?* Frontiers in Bioengineering and Biotechnology, 2020. **8**.

215. Rodríguez-Fuentes, D.E., et al., *Mesenchymal Stem Cells Current Clinical Applications: A Systematic Review.* Archives of Medical Research, 2021. **52**(1): p. 93-101.

216. Ullah, I., R.B. Subbarao, and G.J. Rho, *Human mesenchymal stem cells - current trends and future prospective.* Bioscience reports, 2015. **35**(2): p. e00191.

217. Lee, J.Y. and S.-H. Hong, *Hematopoietic Stem Cells and Their Roles in Tissue Regeneration.* International journal of stem cells, 2020. **13**(1): p. 1-12.

218. Ng, A.P. and W.S. Alexander, *Haematopoietic stem cells: past, present and future.* Cell Death Discovery, 2017. **3**(1): p. 17002.

219. Eva, M.-R. and F. Maria Carolina, *Understanding intrinsic hematopoietic stem cell aging.* Haematologica, 2020. **105**(1): p. 22-37.

220. LifeMap Sciences. *Mesenchymal Stem Cells fate potentials*. n.d. [cited 2019 March, 11]; Available from: https://discovery.lifemapsc.com/library/images/mesenchymal-stem-cells-fate-potentials.

221. Yin, J.Q., J. Zhu, and J.A. Ankrum, *Manufacturing of primed mesenchymal stromal cells for therapy.* Nature Biomedical Engineering, 2019. **3**(2): p. 90-104.

222. Musiał-Wysocka, A., M. Kot, and M. Majka, *The Pros and Cons of Mesenchymal Stem Cell-Based Therapies.* Cell Transplantation, 2019. **28**(7): p. 801-812.

223. Filardo, G., et al., *Mesenchymal stem cells for the treatment of cartilage lesions: from preclinical findings to clinical application in orthopaedics.* Knee Surg Sports Traumatol Arthrosc, 2013. **21**(8): p. 1717-29.

224. Wagner, W. and A.D. Ho, *Mesenchymal Stem Cell Preparations—Comparing Apples and Oranges.* Stem Cell Reviews, 2007. **3**(4): p. 239-248.

225. Barry, F. and M. Murphy, *Mesenchymal stem cells in joint disease and repair.* Nat Rev Rheumatol, 2013. **9**(10): p. 584-94.

226. Oh, M. and J.E. Nör, *The Perivascular Niche and Self-Renewal of Stem Cells.* Frontiers in physiology, 2015. **6**: p. 367-367.

227. Rennerfeldt, D.A. and K.J. Van Vliet, *Concise Review: When Colonies Are Not Clones: Evidence and Implications of Intracolony Heterogeneity in Mesenchymal Stem Cells.* Stem Cells, 2016. **34**(5): p. 1135-41.

228. Wilson, A.J., et al., *Characterisation of mesenchymal stromal cells in clinical trial reports: analysis of published descriptors.* Stem Cell Research & Therapy, 2021. **12**(1): p. 360.

229. Horwitz, E.M., et al., *Clarification of the nomenclature for MSC: The International Society for Cellular Therapy position statement.* Cytotherapy, 2005. **7**(5): p. 393-5.

230. Dominici, M., et al., *Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement.* Cytotherapy, 2006. **8**(4): p. 315-7.

231. Cook, D. and P. Genever, *Regulation of mesenchymal stem cell differentiation.* Adv Exp Med Biol, 2013. **786**: p. 213-29.

232. Sart, S., et al., *Three-dimensional aggregates of mesenchymal stem cells: cellular mechanisms, biological properties, and applications.* Tissue engineering. Part B, Reviews, 2014. **20**(5): p. 365-380.

233. Kfoury, Y. and D.T. Scadden, *Mesenchymal cell contributions to the stem cell niche.* Cell Stem Cell, 2015. **16**(3): p. 239-53.

234. Tolar, J., et al., *Concise review: hitting the right spot with mesenchymal stromal cells.* Stem cells (Dayton, Ohio), 2010. **28**(8): p. 1446-1455.

235. Dabbadie, A., et al., *Development of chimeric forms of the matrix metalloproteinase 2 collagen binding domain as artificial membrane binding proteins for targeting stem cells to cartilage lesions in osteoarthritic joints.* Biomaterials, 2022. **285**: p. 121547.

236. Haddad, R. and F. Saldanha-Araujo, *Mechanisms of T-Cell Immunosuppression by Mesenchymal Stromal Cells: What Do We Know So Far?* BioMed Research International, 2014. **2014**: p. 216806.

237. Meesuk, L., et al., *The immunosuppressive capacity of human mesenchymal stromal cells derived from amnion and bone marrow.* Biochemistry and Biophysics Reports, 2016. **8**: p. 34-40.

238. Schu, S., et al., *Immunogenicity of allogeneic mesenchymal stem cells.* Journal of cellular and molecular medicine, 2012. **16**(9): p. 2094-2103.

239. Madigan, M. and R. Atoui, *Therapeutic Use of Stem Cells for Myocardial Infarction.* Bioengineering, 2018. **5**(2).

240. E Newman, R., et al., *Treatment of Inflammatory Diseases with Mesenchymal Stem Cells*. Vol. 8. 2009. 110-23.

241. Klinker, M.W. and C.-H. Wei, *Mesenchymal stem cells in the treatment of inflammatory and autoimmune diseases in experimental animal models.* World journal of stem cells, 2015. **7**(3): p. 556-567.

242. Freitag, J., et al., *Mesenchymal stem cell therapy in the treatment of osteoarthritis: reparative pathways, safety and efficacy – a review.* BMC Musculoskeletal Disorders, 2016. **17**: p. 230.

243. Wyles, C.C., et al., *Mesenchymal stem cell therapy for osteoarthritis: current perspectives.* Stem cells and cloning : advances and applications, 2015. **8**: p. 117-124.

244. Chen, F.H. and R.S. Tuan, *Mesenchymal stem cells in arthritic diseases.* Arthritis Research & Therapy, 2008. **10**(5): p. 223.

245. Kong, L., et al., *Role of mesenchymal stem cells in osteoarthritis treatment.* Journal of Orthopaedic Translation, 2017. **9**: p. 89-103.

246. Berglund, A.K., et al., *Immunoprivileged no more: measuring the immunogenicity of allogeneic adult mesenchymal stem cells.* Stem Cell Res Ther, 2017. **8**(1): p. 288.

247. Ankrum, J.A., J.F. Ong, and J.M. Karp, *Mesenchymal stem cells: immune evasive, not immune privileged.* Nature biotechnology, 2014. **32**(3): p. 252-260.

248. Di Nicola, M., et al., *Human bone marrow stromal cells suppress T-lymphocyte proliferation induced by cellular or nonspecific mitogenic stimuli.* Blood, 2002. **99**(10): p. 3838-43.

249. Klyushnenkova, E., et al., *T cell responses to allogeneic human mesenchymal stem cells: immunogenicity, tolerance, and suppression.* J Biomed Sci, 2005. **12**(1): p. 47-57.

250. Aggarwal, S. and M.F. Pittenger, *Human mesenchymal stem cells modulate allogeneic immune cell responses.* Blood, 2005. **105**(4): p. 1815-22.

251. Németh, K., et al., *Bone marrow stromal cells attenuate sepsis via prostaglandin E(2)-dependent reprogramming of host macrophages to increase their interleukin-10 production.* Nat Med, 2009. **15**(1): p. 42-9.

252. Galipeau, J., *The mesenchymal stromal cells dilemma--does a negative phase III trial of random donor mesenchymal stromal cells in steroid-resistant graft-versus-host disease represent a death knell or a bump in the road?* Cytotherapy, 2013. **15**(1): p. 2-8.

253. Marion, N.W. and J.J. Mao, *Mesenchymal stem cells and tissue engineering.* Methods in enzymology, 2006. **420**: p. 339-361.

254. Wright, A., M.L. Arthaud-Day, and M.L. Weiss, *Therapeutic Use of Mesenchymal Stromal Cells: The Need for Inclusive Characterization Guidelines to Accommodate All Tissue Sources and Species.* Frontiers in Cell and Developmental Biology, 2021. **9**.

255. Zhang, R., et al., *Mesenchymal stem cell related therapies for cartilage lesions and osteoarthritis.* American journal of translational research, 2019. **11**(10): p. 6275-6289.

256. Wakitani, S., et al., *Mesenchymal cell-based repair of large, full-thickness defects of articular cartilage.* J Bone Joint Surg Am, 1994. **76**(4): p. 579-92.

257. Wakitani, S., et al., *Safety of autologous bone marrow-derived mesenchymal stem cell transplantation for cartilage repair in 41 patients with 45 joints followed for up to 11 years and 5 months.* J Tissue Eng Regen Med, 2011. **5**(2): p. 146-50.

258. Lotfy, A., et al., *Comparative study of biological characteristics of mesenchymal stem cells isolated from mouse bone marrow and peripheral blood.* Biomedical reports, 2019. **11**(4): p. 165-170.

259. Bernardo, M.E. and W.E. Fibbe, *Mesenchymal stromal cells: sensors and switchers of inflammation.* Cell Stem Cell, 2013. **13**(4): p. 392-402.

260. Venkatesha, S., et al., *Soluble endoglin contributes to the pathogenesis of preeclampsia.* Nat Med, 2006. **12**(6): p. 642-9.

261. Pittenger, M.F., et al., *Multilineage potential of adult human mesenchymal stem cells.* Science, 1999. **284**(5411): p. 143-7.

262. Mackay, A.M., et al., *Chondrogenic differentiation of cultured human mesenchymal stem cells from marrow.* Tissue Eng, 1998. **4**(4): p. 415-28.

263. Le, H., et al., *Mesenchymal stem cells for cartilage regeneration.* Journal of tissue engineering, 2020. **11**: p. 2041731420943839-2041731420943839.

264. Pers, Y.M., et al., *Mesenchymal stem cells for the management of inflammation in osteoarthritis: state of the art and perspectives.* Osteoarthritis and Cartilage, 2015. **23**(11): p. 2027-2035.

265. Centeno, C.J., et al., *Increased knee cartilage volume in degenerative joint disease using percutaneously implanted, autologous mesenchymal stem cells.* Pain Physician, 2008. **11**(3): p. 343-53.

266. Wang, M., et al., *Advances and Prospects in Stem Cells for Cartilage Regeneration.* Stem Cells Int, 2017. **2017**: p. 4130607.

267. Xiang, X.-N., et al., *Mesenchymal stromal cell-based therapy for cartilage regeneration in knee osteoarthritis.* Stem Cell Research & Therapy, 2022. **13**(1): p. 14.

268. Spakova, T., et al., *Influence of Kartogenin on Chondrogenic Differentiation of Human Bone Marrow-Derived MSCs in 2D Culture and in Co-Cultivation with OA Osteochondral Explant.* Molecules, 2018. **23**(1).

269. Johnson, K., et al., *A stem cell-based approach to cartilage repair.* Science, 2012. **336**(6082): p. 717-21.

270. Chandran, B. and A. Goel, *A randomized, pilot study to assess the efficacy and safety of curcumin in patients with active rheumatoid arthritis.* Phytother Res, 2012. **26**(11): p. 1719-25.

271. Saiko, P., et al., *Resveratrol and its analogs: defense against cancer, coronary disease and neurodegenerative maladies or just a fad?* Mutat Res, 2008. **658**(1-2): p. 68-94.

272. Wang, J., et al., *Biomimetic cartilage scaffold with orientated porous structure of two factors for cartilage repair of knee osteoarthritis.* Artif Cells Nanomed Biotechnol, 2019. **47**(1): p. 1710-1721.

273. Bauza-Mayol, G., et al., *Biomimetic Scaffolds Modulate the Posttraumatic Inflammatory Response in Articular Cartilage Contributing to Enhanced Neoformation of Cartilaginous Tissue In Vivo.* Advanced Healthcare Materials, 2022. **11**(1): p. 2101127.

274. Condello, V., et al., *Use of a Biomimetic Scaffold for the Treatment of Osteochondral Lesions in Early Osteoarthritis.* Biomed Res Int, 2018. **2018**: p. 7937089.

275. Chung, J.J., et al., *Toward Biomimetic Scaffolds for Tissue Engineering: 3D Printing Techniques in Regenerative Medicine.* Frontiers in Bioengineering and Biotechnology, 2020. **8**.

276. Zhou, X.Z., et al., *Mesenchymal stem cell-based repair of articular cartilage with polyglycolic acid-hydroxyapatite biphasic scaffold.* Int J Artif Organs, 2008. **31**(6): p. 480-9.

277. Zhu, T., et al., *Engineered three-dimensional scaffolds for enhanced bone regeneration in osteonecrosis.* Bioactive materials, 2020. **5**(3): p. 584-601.

278. Mazaki, T., et al., *A novel, visible light-induced, rapidly cross-linkable gelatin scaffold for osteochondral tissue engineering.* Sci Rep, 2014. **4**: p. 4457.

279. Begum, R., et al., *Chondroinduction of Mesenchymal Stem Cells on Cellulose-Silk Composite Nanofibrous Substrates: The Role of Substrate Elasticity.* Front Bioeng Biotechnol, 2020. **8**: p. 197.

280. Mianehsaz, E., et al., *Mesenchymal stem cell-derived exosomes: a new therapeutic approach to osteoarthritis?* Stem Cell Research & Therapy, 2019. **10**(1): p. 340.

281. Cosenza, S., et al., *Mesenchymal stem cells derived exosomes and microparticles protect cartilage and bone from degradation in osteoarthritis.* Scientific Reports, 2017. **7**(1): p. 16214.

282. Mazor, M., et al., *Mesenchymal stem-cell potential in cartilage repair: an update.* Journal of cellular and molecular medicine, 2014. **18**(12): p. 2340-2350.

283. Krishnan, Y. and A.J. Grodzinsky, *Cartilage diseases.* Matrix Biol, 2018. **71-72**: p. 51-69.

284. Gerlag, D.M. and P.P. Tak, *32 - Minimally invasive procedures*, in *Rheumatology (Sixth Edition)*, M.C. Hochberg, et al., Editors. 2015, Mosby: Philadelphia. p. 242-249.

285. Richardson, S.M., et al., *Mesenchymal stem cells in regenerative medicine: Focus on articular cartilage and intervertebral disc regeneration.* Methods, 2016. **99**: p. 69-80.

286. Roelofs, A.J., J.P.J. Rocke, and C. De Bari, *Cell-based approaches to joint surface repair: a research perspective.* Osteoarthritis and Cartilage, 2013. **21**(7): p. 892-900.

287. Escobar Ivirico, J.L., et al., *Regenerative Engineering for Knee Osteoarthritis Treatment: Biomaterials and Cell-Based Technologies.* Engineering, 2017. **3**(1): p. 16-27.

288. Shafiq, M., Y. Jung, and S.H. Kim, *Insight on stem cell preconditioning and instructive biomaterials to enhance cell adhesion, retention, and engraftment for tissue repair.* Biomaterials, 2016. **90**: p. 85-115.

289. Ha, C.W., et al., *Intra-articular Mesenchymal Stem Cells in Osteoarthritis of the Knee: A Systematic Review of Clinical Outcomes and Evidence of Cartilage Repair.* Arthroscopy, 2018.

290. Jin, G.-Z., *Current Nanoparticle-Based Technologies for Osteoarthritis Therapy.* Nanomaterials (Basel, Switzerland), 2020. **10**(12): p. 2368.

291. Wu, J., et al., *Nanomaterials with enzyme-like characteristics (nanozymes): next-generation artificial enzymes (II).* Chem Soc Rev, 2019. **48**(4): p. 1004-1076.

292. Tian, R., et al., *Rational Design and Biological Application of Antioxidant Nanozymes.* Frontiers in Chemistry, 2021. **8**: p. 831.

293. Brown, S., et al., *Nanoparticle Properties for Delivery to Cartilage: The Implications of Disease State, Synovial Fluid, and Off-Target Uptake.* Molecular pharmaceutics, 2019. **16**(2): p. 469-479.

294. Kamei, N., et al., *The safety and efficacy of magnetic targeting using autologous mesenchymal stem cells for cartilage repair.* Knee Surg Sports Traumatol Arthrosc, 2018. **26**(12): p. 3626-3635.

295. Walker, M., *Development of functional nanoparticles for MSC-targeted joint repair in situ.* Osteoarthritis and Cartilage, 2018. **26**: p. S148.

296. Lu, Y., et al., *Hierarchical functional nanoparticles boost osteoarthritis therapy by utilizing joint-resident mesenchymal stem cells.* Journal of Nanobiotechnology, 2022. **20**(1): p. 89.

297. Xiao, L., et al., *Therapeutic potential of nanotechnology-based approaches in osteoarthritis.* Frontiers in Pharmacology, 2022. **13**.

298. Kean, T.J., et al., *MSCs: Delivery Routes and Engraftment, Cell-Targeting Strategies, and Immune Modulation.* Stem Cells International, 2013. **2013**: p. 13.

299. Sarraf, K.M. and D.F. Kader, *Knee oral core topics*, in *Postgraduate Orthopaedics: The Candidate's Guide to the FRCS (Tr & Orth) Examination*, D.F. Kader and P.A. Banaszkiewicz, Editors. 2017, Cambridge University Press: Cambridge. p. 292-338.

300. Steffensen, B., et al., *Human fibronectin and MMP-2 collagen binding domains compete for collagen binding sites and modify cellular activation of MMP-2.* Matrix Biology, 2002. **21**(5): p. 399-414.

301. Xu, X., et al., *Functional basis for the overlap in ligand interactions and substrate specificities of matrix metalloproteinases-9 and -2.* Biochem J, 2005. **392**(Pt 1): p. 127-34.

302. Lobstein, J., et al., *SHuffle, a novel Escherichia coli protein expression strain capable of correctly folding disulfide bonded proteins in its cytoplasm.* Microbial Cell Factories, 2012. **11**(1): p. 753.

303. Xu, X., et al., *Nuclear magnetic resonance mapping and functional confirmation of the collagen binding sites of matrix metalloproteinase-2.* Biochemistry, 2009. **48**(25): p. 5822-31.

304. Armstrong, J.P., et al., *Artificial membrane-binding proteins stimulate oxygenation of stem cells during engineering of large cartilage tissue.* Nat Commun, 2015. **6**: p. 7405.

305. Lu, J.X., C. Tupper, and J. Murray, *Biochemistry, Dissolution and Solubility*, in *StatPearls*. 2023, StatPearls Publishing

306. Lessard, J.C., *Growth media for E. coli.* Methods Enzymol, 2013. **533**: p. 181-9.

307. Lau, Y.-T.K., et al., *Discovery and engineering of enhanced SUMO protease enzymes.* The Journal of biological chemistry, 2018. **293**(34): p. 13224-13233.

308. Abdelkader, E.H. and G. Otting, *NT*-HRV3CP: An optimized construct of human rhinovirus 14 3C protease for high-yield expression and fast affinity-tag cleavage.* Journal of Biotechnology, 2021. **325**: p. 145-151.

309. Scientific, T., *Nucleic Acid Thermo Scientific NanoDrop Spectrophotometers*. 2010: Wilmington.

310. Lucena-Aguilar, G., et al., *DNA Source Selection for Downstream Applications Based on DNA Quality Indicators Analysis.* Biopreserv Biobank, 2016. **14**(4): p. 264-70.

311. Skretas, G. and S. Ventura, *Editorial: Protein Aggregation and Solubility in Microorganisms (Archaea, Bacteria and Unicellular Eukaryotes): Implications and Applications.* Frontiers in Microbiology, 2020. **11**.

312. Alberts B, J.A., Lewis J, et al., *The Shape and Structure of Proteins*, in *Molecular Biology of the Cell*. 2002, Garland Science: New York.

313. Richardson, J.S., *The anatomy and taxonomy of protein structure.* Adv Protein Chem, 1981. **34**: p. 167-339.

314. Hollingsworth, S.A. and P.A. Karplus, *A fresh look at the Ramachandran plot and the occurrence of standard structures in proteins.* Biomol Concepts, 2010. **1**(3-4): p. 271-283.

315. Anfinsen, C.B., *Principles that govern the folding of protein chains.* Science, 1973. **181**(4096): p. 223-30.

316. Sorokina, I., A.R. Mushegian, and E.V. Koonin, *Is Protein Folding a Thermodynamically Unfavorable, Active, Energy-Dependent Process?* Int J Mol Sci, 2022. **23**(1).

317. Lin, M.M. and A.H. Zewail, *Protein folding – simplicity in complexity.* Annalen der Physik, 2012. **524**(8): p. 379-391.

318. Martínez, L., *Introducing the Levinthal's Protein Folding Paradox and Its Solution.* Journal of Chemical Education, 2014. **91**(11): p. 1918-1923.

319. Zwanzig, R., A. Szabo, and B. Bagchi, *Levinthal's paradox.* Proceedings of the National Academy of Sciences, 1992. **89**(1): p. 20-22.

320. Englander, S.W. and L. Mayne, *The nature of protein folding pathways.* Proceedings of the National Academy of Sciences, 2014. **111**(45): p. 15873-15880.

321. Vila, J.A., *Thoughts on the Protein's Native State.* The Journal of Physical Chemistry Letters, 2021. **12**(25): p. 5963-5966.

322. Hartl, F.U., A. Bracher, and M. Hayer-Hartl, *Molecular chaperones in protein folding and proteostasis.* Nature, 2011. **475**(7356): p. 324-32.

323. Thompson, M., T. Yeates, and J. Rodriguez, *Advances in methods for atomic resolution macromolecular structure determination [version 1; peer review: 2 approved].* F1000Research, 2020. **9**(667).

324. Wu, X. and A. Rapoport Tom, *Cryo-EM structure determination of small proteins by nanobody-binding scaffolds (Legobodies).* Proceedings of the National Academy of Sciences, 2021. **118**(41): p. e2115001118.

325. wwPDB consortium, *Protein Data Bank: the single global archive for 3D macromolecular structure data.* Nucleic Acids Research, 2019. **47**(D1): p. D520-D528.

326. Haddad, Y., V. Adam, and Z. Heger, *Ten quick tips for homology modeling of high-resolution protein 3D structures.* PLOS Computational Biology, 2020. **16**(4): p. e1007449.

327. Sander, C. and R. Schneider, *Database of homology-derived protein structures and the structural meaning of sequence alignment.* Proteins, 1991. **9**(1): p. 56-68.

328. Agnihotry, S., et al., *Chapter 11 - Protein structure prediction*, in *Bioinformatics*, D.B. Singh and R.K. Pathak, Editors. 2022, Academic Press. p. 177-188.

329. Kuhlman, B. and P. Bradley, *Advances in protein structure prediction and design.* Nature Reviews Molecular Cell Biology, 2019. **20**(11): p. 681-697.

330. Kryshtafovych, A., et al., *Critical assessment of methods of protein structure prediction (CASP)—Round XIV.* Proteins: Structure, Function, and Bioinformatics, 2021. **89**(12): p. 1607-1617.

331. Monzon, A.M., et al., *Homology modeling in a dynamical world.* Protein science : a publication of the Protein Society, 2017. **26**(11): p. 2195-2206.

332. Chothia, C. and A.M. Lesk, *The relation between the divergence of sequence and structure in proteins.* Embo j, 1986. **5**(4): p. 823-6.

333. Pearson, W.R., *An Introduction to Sequence Similarity ("Homology") Searching.* Current protocols in bioinformatics / editoral board, Andreas D. Baxevanis ... [et al.], 2013. **0 3**: p. 10.1002/0471250953.bi0301s42.

334. Lima, I. and E.A. Cino, *SS3D: Sequence similarity in 3D for comparison of protein families.* bioRxiv, 2020: p. 2020.05.27.117127.

335. Fiser, A., *Comparative Protein Structure Modelling*, in *From Protein Structure to Function with Bioinformatics*, D. J. Rigden, Editor. 2017, Springer Netherlands: Dordrecht. p. 91-134.

336. Fiser, A., *Template-based protein structure modeling.* Methods in molecular biology (Clifton, N.J.), 2010. **673**: p. 73-94.

337. Madhusudhan, M.S., et al., *Comparative Protein Structure Modeling*, in *The Proteomics Protocols Handbook*, J.M. Walker, Editor. 2005, Humana Press: Totowa, NJ. p. 831-860.

338. Xiang, Z., *Advances in homology protein structure modeling.* Current protein & peptide science, 2006. **7**(3): p. 217-227.

339. Lange, J., et al., *Facilities that make the PDB data collection more powerful.* Protein Science, 2020. **29**(1): p. 330-344.

340. Berman, H.M., et al., *The Protein Data Bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.

341. Campbell, I.D. and A.K. Downing, *Building protein structure and function from modular units.* Trends Biotechnol, 1994. **12**(5): p. 168-72.

342. Reddy Chichili, V.P., V. Kumar, and J. Sivaraman, *Linkers in the structural biology of protein–protein interactions.* Protein Science, 2013. **22**(2): p. 153-167.

343. Webb, B. and A. Sali, *Protein structure modeling with MODELLER.* Methods Mol Biol, 2014. **1137**: p. 1-15.

344. Drew, E.D. and R.W. Janes, *2StrucCompare: a webserver for visualizing small but noteworthy differences between protein tertiary structures through interrogation of the secondary structure content.* Nucleic Acids Research, 2019. **47**(W1): p. W477-W481.

345. Benkert, P., S.C. Tosatto, and D. Schomburg, *QMEAN: A comprehensive scoring function for model quality assessment.* Proteins, 2008. **71**(1): p. 261-77.

346. Finkelstein, A.V. and O.B. Ptitsyn, *Lecture 13*, in *Protein Physics (Second Edition)*, A.V. Finkelstein and O.B. Ptitsyn, Editors. 2016, Academic Press: Amsterdam. p. 181-197.

347. Richardson, J.S., W.B. Arendall, and D.C. Richardson, *New Tools and Data for Improving Structures, Using All-Atom Contacts*, in *Methods in Enzymology*. 2003, Academic Press. p. 385-412.

348. Laimer, J., et al., *MAESTRO - multi agent stability prediction upon point mutations*. Vol. 16. 2015.

349. Cavallo, L., J. Kleinjung, and F. Fraternali, *POPS: A fast algorithm for solvent accessible surface areas at atomic and residue level.* Nucleic acids research, 2003. **31**(13): p. 3364-3366.

350. Costa, S., et al., *Fusion tags for protein solubility, purification and immunogenicity in Escherichia coli: the novel Fh8 system.* Frontiers in Microbiology, 2014. **5**.

351. Butt, T.R., et al., *SUMO fusion technology for difficult-to-express proteins.* Protein expression and purification, 2005. **43**(1): p. 1-9.

352. Assenberg, R., et al., *Expression, purification and crystallization of a lyssavirus matrix (M) protein.* Acta crystallographica. Section F, Structural biology and crystallization communications, 2008. **64**(Pt 4): p. 258-262.

353. Berrow, N.S., et al., *A versatile ligation-independent cloning method suitable for high-throughput expression screening applications.* Nucleic acids research, 2007. **35**(6): p. e45-e45.

354. TakaraBio. *In-Fusion Cloning FAQs: How does In-Fusion Cloning work?* 2019 [cited 2019 April 24]; Available from: https://www.takarabio.com/learning-centers/cloning/in-fusion-cloning-faqs.

355. TakaraBio. *Stellar chemically competent cells for cloning*. n.d. [cited 2022 July 19]; Available from: https://www.takarabio.com/products/cloning/competent-cells/stellar-chemically-competent-cells#:~:text=Stellar%20Cells%20are%20an%20E,and%20even%20methylated%20DNA%20cloning.

356. Robinson, M.-P., et al., *Efficient expression of full-length antibodies in the cytoplasm of engineered bacteria.* Nature Communications, 2015. **6**(1): p. 8072.

357. Ren, G., N. Ke, and M. Berkmen, *Use of the SHuffle Strains in Production of Proteins.* Curr Protoc Protein Sci, 2016. **85**: p. 5.26.1-5.26.21.

358. Dutta, S. and K. Bose, *Protein Purification by Affinity Chromatography*, in *Textbook on Cloning, Expression and Purification of Recombinant Proteins*, K. Bose, Editor. 2022, Springer Nature Singapore: Singapore. p. 141-171.

359. LabXchange. *Using Imidazole to Elute His-tagged Proteins from a Nickel Column*. 2021 April 10, 2023]; Available from: https://www.labxchange.org/library/items/lb:LabXchange:c2062540:lx_image:1.

360. Bakir, U. and H. Hamamci, *The effect of freeze-thawing on the release of intracellular proteins from Escherichia coli by means of a bead mill.* World Journal of Microbiology and Biotechnology, 1997. **13**(4): p. 475-477.

361. Francis, D.M. and R. Page, *Strategies to optimize protein expression in E. coli.* Curr Protoc Protein Sci, 2010. **Chapter 5**(1): p. Unit 5.24.1-29.

362. Prasad, S., P.B. Khadatare, and I. Roy, *Effect of chemical chaperones in improving the solubility of recombinant proteins in Escherichia coli.* Applied and environmental microbiology, 2011. **77**(13): p. 4603-4609.

363. Lee, J.C. and S.N. Timasheff, *The stabilization of proteins by sucrose.* J Biol Chem, 1981. **256**(14): p. 7193-201.

364. Cortez, L. and V. Sim, *The therapeutic potential of chemical chaperones in protein folding diseases.* Prion, 2014. **8**(2): p. 197-202.

365. Georgiou, G. and P. Valax, *Expression of correctly folded proteins in Escherichia coli.* Current Opinion in Biotechnology, 1996. **7**(2): p. 190-197.

366. Chhetri, G., P. Kalita, and T. Tripathi, *An efficient protocol to enhance recombinant protein expression using ethanol in Escherichia coli.* MethodsX, 2015. **2**: p. 385-391.

367. Lengeler, J.W., *Catabolite Repression*, in *Encyclopedia of Genetics*, S. Brenner and J.H. Miller, Editors. 2001, Academic Press: New York. p. 281-284.

368. Deutscher, J., *The mechanisms of carbon catabolite repression in bacteria.* Curr Opin Microbiol, 2008. **11**(2): p. 87-93.

369. Tartof, K.D. and C. Hobbs, *Improved media for plasmid and cosmid clones.* Bethesda Res. Lab. Focus, 1987. **9**.

370. Kram, K.E. and S.E. Finkel, *Rich Medium Composition Affects Escherichia coli Survival, Glycation, and Mutation Frequency during Long-Term Batch Culture.* Appl Environ Microbiol, 2015. **81**(13): p. 4442-50.

371. Sambrook, J., E.F. Fritsch, and T. Maniatis, *Molecular Cloning: A Laboratory Manual*. 1989: Cold Spring Harbor Laboratory Press.

372. *M9 minimal medium (standard).* Cold Spring Harbor Protocols, 2010. **2010**(8): p. pdb.rec12295.

373. Greenfield, N.J., *Using circular dichroism spectra to estimate protein secondary structure.* Nature Protocols, 2006. **1**(6): p. 2876-2890.

374. Andrews, S.S. and J. Tretton, *Physical Principles of Circular Dichroism.* Journal of Chemical Education, 2020. **97**(12): p. 4370-4376.

375. Rodger, A. and M.A. Ismail, *99Introduction to circular dichroism*, in *Spectrophotometry and Spectrofluorimetry: A Practical Approach*, M. Gore, Editor. 2000, Oxford University Press. p. 0.

376. Micsonai, A., et al., *BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra.* Nucleic acids research, 2018. **46**(W1): p. W315-W322.

377. Pucci, F., J.M. Kwasigroch, and M. Rooman, *SCooP: an accurate and fast predictor of protein stability curves as a function of temperature.* Bioinformatics, 2017. **33**(21): p. 3415-3422.

378. Rehman, I., C.C. Kerndt, and S. Botelho, *Biochemistry, Tertiary Protein Structure*, in *StatPearls*. 2022, StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC.: Treasure Island (FL).

379. Steffensen, B., U.M. Wallon, and C.M. Overall, *Extracellular matrix binding properties of recombinant fibronectin type II-like modules of human 72-kDa gelatinase/type IV collagenase. High affinity binding to native type I collagen but not native type IV collagen.* J Biol Chem, 1995. **270**(19): p. 11555-66.

380. Morgunova, E., et al., *Structure of human pro-matrix metalloproteinase-2: activation mechanism revealed.* Science, 1999. **284**(5420): p. 1667-70.

381. Wiltgen, M., *Algorithms for Structure Comparison and Analysis: Homology Modelling of Proteins*, in *Encyclopedia of Bioinformatics and Computational Biology*, S. Ranganathan, et al., Editors. 2019, Academic Press: Oxford. p. 38-61.

382. Aarthy, M. and S.K. Singh, *Chapter 28 - Envisaging the conformational space of proteins by coupling machine learning and molecular dynamics*, in *Advances in Protein Molecular and Structural Biology Methods*, T. Tripathi and V.K. Dubey, Editors. 2022, Academic Press. p. 467-475.

383. Lovell, S.C., et al., *Structure validation by Cα geometry: ϕ,ψ and Cβ deviation.* Proteins: Structure, Function, and Bioinformatics, 2003. **50**(3): p. 437-450.

384. Sugita, Y. and A. Kitao, *Dependence of Protein Stability on the Structure of the Denatured State: Free Energy Calculations of I56V Mutation in Human Lysozyme.* Biophysical Journal, 1998. **75**(5): p. 2178-2187.

385. Bell, M.R., et al., *To fuse or not to fuse: What is your purpose?* Protein Science, 2013. **22**(11): p. 1466-1477.

386. Gao, X., et al., *Prediction of disulfide bond engineering sites using a machine learning method.* Scientific Reports, 2020. **10**(1): p. 10330.

387. Braakman, I., et al., *Analysis of Disulfide Bond Formation.* Current protocols in protein science, 2017. **90**: p. 14.1.1-14.1.21.

388. Khrapunov, S., *Circular dichroism spectroscopy has intrinsic limitations for protein secondary structure analysis.* Analytical biochemistry, 2009. **389**(2): p. 174-176.

389. Hall, V., M. Sklepari, and A. Rodger, *Protein secondary structure prediction from circular dichroism spectra using a self-organizing map with concentration correction.* Chirality, 2014. **26**(9): p. 471-82.

390. Schlamadinger, D.E. and J.E. Kim, *Thermodynamics of membrane protein folding measured by fluorescence spectroscopy.* Journal of visualized experiments : JoVE, 2011(50): p. 2669.

391. Marabotti, A., et al., *Performance of Web tools for predicting changes in protein stability caused by mutations.* BMC Bioinformatics, 2021. **22**(7): p. 345.

392. Jumper, J., et al., *Highly accurate protein structure prediction with AlphaFold.* Nature, 2021. **596**(7873): p. 583-589.

393. Al-Janabi, A., *Has DeepMind's AlphaFold solved the protein folding problem?* BioTechniques, 2022. **72**(3): p. 73-76.

394. Garcia-Ochoa, F. and E. Gomez, *Bioreactor scale-up and oxygen transfer rate in microbial processes: an overview.* Biotechnol Adv, 2009. **27**(2): p. 153-76.

395. Mason, T. and D. Peters, *Practical Sonochemistry: Power Ultrasound Uses and Applications*. 2002.

396. Stathopulos, P.B., et al., *Sonication of proteins causes formation of aggregates that resemble amyloid.* Protein science : a publication of the Protein Society, 2004. **13**(11): p. 3017-3027.

397. Ferdous, S., J.L. Dopp, and N.F. Reuel, *Optimization of E. coli tip-sonication for high-yield cell-free extract using finite element modeling.* AIChE Journal, 2021. **67**(10): p. e17389.

398. Wells, Jonathan N., L.T. Bergendahl, and Joseph A. Marsh, *Co-translational assembly of protein complexes.* Biochemical Society Transactions, 2015. **43**(6): p. 1221-1226.

399. Gainza-Cirauqui, P. and B.E. Correia, *Computational protein design-the next generation tool to expand synthetic biology applications.* Curr Opin Biotechnol, 2018. **52**: p. 145-152.

400. Svilenov, H., U. Markoja, and G. Winter, *Isothermal chemical denaturation as a complementary tool to overcome limitations of thermal differential scanning fluorimetry in predicting physical stability of protein formulations.* European Journal of Pharmaceutics and Biopharmaceutics, 2018. **125**: p. 106-113.

401. Crowther, G.J., et al., *Use of thermal melt curves to assess the quality of enzyme preparations.* Analytical biochemistry, 2010. **399**(2): p. 268-275.

402. Gao, K., R. Oerlemans, and M.R. Groves, *Theory and applications of differential scanning fluorimetry in early-stage drug discovery.* Biophysical Reviews, 2020. **12**(1): p. 85-104.

403. Malleshappa Gowder, S., et al., *Prediction and analysis of surface hydrophobic residues in tertiary structure of proteins.* TheScientificWorldJournal, 2014. **2014**: p. 971258-971258.

404. Gilbreth, R.N. and S. Koide, *Structural insights for engineering binding proteins based on non-antibody scaffolds.* Current opinion in structural biology, 2012. **22**(4): p. 413-420.

405. Tobi, D. and I. Bahar, *Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state.* Proceedings of the National Academy of Sciences, 2005. **102**(52): p. 18908-18913.

406. Paul, F. and T.R. Weikl, *How to Distinguish Conformational Selection and Induced Fit Based on Chemical Relaxation Rates.* PLoS computational biology, 2016. **12**(9): p. e1005067-e1005067.

407. Michel, D., *Conformational selection or induced fit? New insights from old principles.* Biochimie, 2016. **128-129**: p. 48-54.

408. Orosz, F. and B.G. Vértessy, *What's in a name? From "fluctuation fit" to "conformational selection": rediscovery of a concept.* History and Philosophy of the Life Sciences, 2021. **43**(3): p. 88.

409. Guan, X., L. Zhang, and J. Wypych, *Direct mass spectrometric characterization of disulfide linkages.* mAbs, 2018. **10**(4): p. 572-582.

410. Ishikawa, H., et al., *Disulfide bond influence on protein structural dynamics probed with 2D-IR vibrational echo spectroscopy.* Proceedings of the National Academy of Sciences, 2007. **104**(49): p. 19309-19314.

411. Mobli, M. and G.F. King, *NMR methods for determining disulfide-bond connectivities.* Toxicon, 2010. **56**(6): p. 849-854.

412. Wiedemann, C., et al., *Cysteines and Disulfide Bonds as Structure-Forming Units: Insights From Different Domains of Life and the Potential for Characterization by NMR.* Frontiers in Chemistry, 2020. **8**.

413. Poppe, L., et al., *PADLOC: A Powerful Tool to Assign Disulfide Bond Connectivities in Peptides and Proteins by NMR Spectroscopy.* Analytical Chemistry, 2012. **84**(1): p. 262-266.

414. Vihinen, M., *Solubility of proteins.* ADMET and DMPK, 2020. **8**.

415. Arakawa, T. and S.N. Timasheff, *[3]Theory of protein solubility*, in *Methods in Enzymology*. 1985, Academic Press. p. 49-77.

416. Vagenende, V., M.G. Yap, and B.L. Trout, *Mechanisms of protein stabilization and prevention of protein aggregation by glycerol.* Biochemistry, 2009. **48**(46): p. 11084-96.

417. Kramer, R.M., et al., *Toward a molecular understanding of protein solubility: increased negative surface charge correlates with increased solubility.* Biophysical journal, 2012. **102**(8): p. 1907-1915.

418. Hirai, M., et al., *Direct Evidence for the Effect of Glycerol on Protein Hydration and Thermal Structural Transition.* Biophys J, 2018. **115**(2): p. 313-327.

419. Lesk, A., *Introduction to Protein Science: Architecture, Function, and Genomics*. 2010: OUP Oxford.

420. Thomas, B.C., et al., *Cataracts are caused by alterations of a critical N-terminal positive charge in connexin50.* Investigative ophthalmology & visual science, 2008. **49**(6): p. 2549-2556.

421. Liu, J.-J., et al., *The structure-based cancer-related single amino acid variation prediction.* Scientific Reports, 2021. **11**(1): p. 13599.

422. Murrell, J.R., et al., *Early-onset Alzheimer disease caused by a new mutation (V717L) in the amyloid precursor protein gene.* Arch Neurol, 2000. **57**(6): p. 885-7.

423. De Meirleir, L., et al., *Respiratory chain complex V deficiency due to a mutation in the assembly gene <em>ATP12</em&gt.* Journal of Medical Genetics, 2004. **41**(2): p. 120.

424. Trevino, S.R., J.M. Scholtz, and C.N. Pace, *Amino acid contribution to protein solubility: Asp, Glu, and Ser contribute more favorably than the other hydrophilic amino acids in RNase Sa.* Journal of molecular biology, 2007. **366**(2): p. 449-460.

425. Izard, J., et al., *A single amino acid substitution can restore the solubility of aggregated colicin A mutants in Escherichia coli.* Protein Engineering, Design and Selection, 1994. **7**(12): p. 1495-1500.

426. Wang, Y. and R.F. Latypov, *Quantitative Evaluation of Protein Solubility in Aqueous Solutions by PEG-Induced Liquid-Liquid Phase Separation.* Methods Mol Biol, 2019. **2039**: p. 39-49.

427. Trevino, S.R., J.M. Scholtz, and C.N. Pace, *Measuring and increasing protein solubility.* J Pharm Sci, 2008. **97**(10): p. 4155-66.

428. Hon, J., et al., *SoluProt: prediction of soluble protein expression in Escherichia coli.* Bioinformatics, 2021. **37**(1): p. 23-28.

429. Chai, Q., et al., *Development of a high-throughput solubility screening assay for use in antibody discovery.* mAbs, 2019. **11**(4): p. 747-756.

430. Dumetz, A.C., et al., *Protein Phase Behavior in Aqueous Solutions: Crystallization, Liquid-Liquid Phase Separation, Gels, and Aggregates.* Biophysical Journal, 2008. **94**(2): p. 570-583.

431. Gibson, T.J., et al., *Application of a high-throughput screening procedure with PEG-induced precipitation to compare relative protein solubility during formulation development with IgG1 monoclonal antibodies.* J Pharm Sci, 2011. **100**(3): p. 1009-21.

432. Hofmann, M., et al., *Limitations of polyethylene glycol-induced precipitation as predictive tool for protein solubility during formulation development.* J Pharm Pharmacol, 2018. **70**(5): p. 648-654.

433. Ingham, K.C., *[23] Precipitation of proteins with polyethylene glycol*, in *Methods in Enzymology*, M.P. Deutscher, Editor. 1990, Academic Press. p. 301-306.

434. Toprani, V.M., et al., *A Micro–Polyethylene Glycol Precipitation Assay as a Relative Solubility Screening Tool for Monoclonal Antibody Design and Formulation Development.* Journal of Pharmaceutical Sciences, 2016. **105**(8): p. 2319-2327.

435. Oeller, M., P. Sormanni, and M. Vendruscolo, *An open-source automated PEG precipitation assay to measure the relative solubility of proteins with low material requirement.* Scientific Reports, 2021. **11**(1): p. 21932.

436. Burgess, R.R., *Protein precipitation techniques.* Methods Enzymol, 2009. **463**: p. 331-42.

437. Wingfield, P.T., *Overview of the purification of recombinant proteins.* Current protocols in protein science, 2015. **80**: p. 6.1.1-6.1.35.

438. McPherson, A. and J.A. Gavira, *Introduction to protein crystallization.* Acta crystallographica. Section F, Structural biology communications, 2014. **70**(Pt 1): p. 2-20.

439. McPherson, A., *Introduction to protein crystallization.* Methods, 2004. **34**(3): p. 254-65.

440. Dessau, M.A. and Y. Modis, *Protein crystallization for X-ray crystallography.* Journal of visualized experiments : JoVE, 2011(47): p. 2285.

441. Habibi, N., et al., *A review of machine learning methods to predict the solubility of overexpressed recombinant proteins in Escherichia coli.* BMC bioinformatics, 2014. **15**: p. 134-134.

442. Madani, M., K. Lin, and A. Tarakanova, *DSResSol: A Sequence-Based Solubility Predictor Created with Dilated Squeeze Excitation Residual Networks.* International Journal of Molecular Sciences, 2021. **22**(24).

443. Sormanni, P., F.A. Aprile, and M. Vendruscolo, *The CamSol method of rational design of protein mutants with enhanced solubility.* Journal of molecular biology, 2015. **427**(2): p. 478-490.

444. Rawi, R., et al., *PaRSnIP: sequence-based protein solubility prediction using gradient boosting machine.* Bioinformatics, 2018. **34**(7): p. 1092-1098.

445. Paladin, L., D. Piovesan, and S.C.E. Tosatto, *SODA: prediction of protein solubility from disorder and aggregation propensity.* Nucleic Acids Res, 2017. **45**(W1): p. W236-w240.

446. Bhandari, B.K., P.P. Gardner, and C.S. Lim, *Solubility-Weighted Index: fast and accurate prediction of protein solubility.* Bioinformatics (Oxford, England), 2020. **36**(18): p. 4691-4698.

447. Hou, Q., et al., *SOLart: a structure-based method to predict protein solubility and aggregation.* Bioinformatics, 2020. **36**(5): p. 1445-1452.

448. Kuriata, A., et al., *Aggrescan3D (A3D) 2.0: prediction and engineering of protein solubility.* Nucleic Acids Research, 2019. **47**(W1): p. W300-W307.

449. Zambrano, R., et al., *AGGRESCAN3D (A3D): server for prediction of aggregation properties of protein structures.* Nucleic Acids Research, 2015. **43**(W1): p. W306-W313.

450. Hebditch, M., et al., *Protein–Sol: a web tool for predicting protein solubility from sequence.* Bioinformatics, 2017. **33**(19): p. 3098-3100.

451. Chen, J., et al., *Structure-aware protein solubility prediction from sequence through graph convolutional network and predicted contact map.* Journal of Cheminformatics, 2021. **13**(1): p. 7.

452. Cheng, J., et al., *SCRATCH: a protein structure and structural feature prediction server.* Nucleic Acids Res, 2005. **33**(Web Server issue): p. W72-6.

453. Agostini, F., et al., *ccSOL omics: a webserver for solubility prediction of endogenous and heterologous expression in Escherichia coli.* Bioinformatics, 2014. **30**(20): p. 2975-7.

454. Hou, Q., et al., *Solart: A Structure-Based Method To Predict Protein Solubility And Aggregation.* bioRxiv, 2019: p. 600734.

455. Sormanni, P. and M. Vendruscolo, *Protein Solubility Predictions Using the CamSol Method in the Study of Protein Homeostasis.* Cold Spring Harb Perspect Biol, 2019. **11**(12).

456. Wimley, W.C. and S.H. White, *Experimentally determined hydrophobicity scale for proteins at membrane interfaces.* Nature Structural Biology, 1996. **3**(10): p. 842-848.

457. Der-Sarkissian, A., et al., *Structural organization of alpha-synuclein fibrils studied by site-directed spin labeling.* J Biol Chem, 2003. **278**(39): p. 37530-5.

458. Ferguson, N., et al., *Rapid amyloid fiber formation from the fast-folding WW domain FBP28.* Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(17): p. 9814-9819.

459. Beerten, J., et al., *Aggregation gatekeepers modulate protein homeostasis of aggregating sequences and affect bacterial fitness.* Protein Engineering, Design and Selection, 2012. **25**(7): p. 357-366.

460. Kosobokova, E.N., K.A. Skrypnik, and V.S. Kosorukov, *Overview of Fusion Tags for Recombinant Proteins.* Biochemistry (Mosc), 2016. **81**(3): p. 187-200.

461. Marblestone, J.G., et al., *Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO.* Protein Sci, 2006. **15**(1): p. 182-9.

462. Maffitt, M., et al., *Rapid screening for protein solubility and expression.* Nature Methods, 2015. **12**(6): p. i-ii.

463. Bell, M.R., et al., *To fuse or not to fuse: what is your purpose?* Protein science : a publication of the Protein Society, 2013. **22**(11): p. 1466-1477.

464. Nishihara, K., et al., *Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in Escherichia coli.* Applied and environmental microbiology, 1998. **64**(5): p. 1694-1699.

465. de Marco, A., et al., *Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli.* BMC Biotechnology, 2007. **7**(1): p. 32.

466. Mikhailova, M., et al., *Identification of collagen binding domain residues that govern catalytic activities of matrix metalloproteinase-2 (MMP-2).* Matrix biology : journal of the International Society for Matrix Biology, 2012. **31**(7-8): p. 380-388.

467. Li, L., A. Kantor, and N. Warne, *Application of a PEG precipitation method for solubility screening: a tool for developing high protein concentration formulations.* Protein science : a publication of the Protein Society, 2013. **22**(8): p. 1118-1123.

468. Bányai, L., H. Tordai, and L. Patthy, *Structure and Domain-Domain Interactions of the Gelatin-binding Site of Human 72-Kilodalton Type IV Collagenase (Gelatinase A, Matrix Metalloproteinase 2) (∗).* Journal of Biological Chemistry, 1996. **271**(20): p. 12003-12008.

469. Stank, A., et al., *Protein Binding Pocket Dynamics.* Accounts of Chemical Research, 2016. **49**(5): p. 809-815.

470. Schwartz, R. and J. King, *Frequencies of hydrophobic and hydrophilic runs and alternations in proteins of known structure.* Protein science : a publication of the Protein Society, 2006. **15**(1): p. 102-112.

471. Yamniuk, A.P., et al., *Application of a kosmotrope-based solubility assay to multiple protein therapeutic classes indicates broad use as a high-throughput screen for protein therapeutic aggregation propensity.* J Pharm Sci, 2013. **102**(8): p. 2424-39.

472. Zhou, H.-X. and X. Pang, *Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation.* Chemical reviews, 2018. **118**(4): p. 1691-1741.

473. Hou, Q., et al., *Computational analysis of the amino acid interactions that promote or decrease protein solubility.* Scientific Reports, 2018. **8**(1): p. 14661.

474. Some, D., et al., *Characterization of Proteins by Size-Exclusion Chromatography Coupled to Multi-Angle Light Scattering (SEC-MALS).* J Vis Exp, 2019(148).

475. Puglisi, R., et al., *Quantifying the thermodynamics of protein unfolding using 2D NMR spectroscopy.* Communications Chemistry, 2020. **3**(1): p. 100.

476. Gupta, S. and S. Bhattacharjya, *NMR characterization of the near native and unfolded states of the PTB domain of Dok1: alternate conformations and residual clusters.* PloS one, 2014. **9**(2): p. e90557-e90557.

477. Gross, F., *Occam's Razor in Molecular and Systems Biology.* Philosophy of Science, 2019. **86**(5): p. 1134-1145.

478. Perkel, J.M., *The computational protein designers.* Nature, 2019. **571**(7766): p. 585-587.

479. Pan, X. and T. Kortemme, *Recent advances in de novo protein design: Principles, methods, and applications.* Journal of Biological Chemistry, 2021. **296**: p. 100558.

480. Xu, X., et al., *NMR Mapping and Functional Confirmation of the Collagen Binding Sites of MMP-2.* Biochemistry, 2009. **48**(25): p. 5822-5831.

481. Watanabe, K., et al., *Designing thermostable proteins: ancestral mutants of 3-isopropylmalate dehydrogenase designed by using a phylogenetic tree.* J Mol Biol, 2006. **355**(4): p. 664-74.

482. Gabaldón, T. and E.V. Koonin, *Functional and evolutionary implications of gene orthology.* Nature reviews. Genetics, 2013. **14**(5): p. 360-366.

483. Sloutsky, R. and K.M. Naegle, *ASPEN, a methodology for reconstructing protein evolution with improved accuracy using ensemble models.* eLife, 2019. **8**: p. e47676.

484. Wang, T.W., et al., *Mutant library construction in directed molecular evolution: casting a wider net.* Mol Biotechnol, 2006. **34**(1): p. 55-68.

485. Sormanni, P., et al., *Rapid and accurate in silico solubility screening of a monoclonal antibody library.* Scientific reports, 2017. **7**(1): p. 1-9.

486. Galas, D.J., et al., *Describing the complexity of systems: multivariable "set complexity" and the information basis of systems biology.* Journal of computational biology : a journal of computational molecular cell biology, 2014. **21**(2): p. 118-140.

487. Levitt, M., *The birth of computational structural biology.* Nature Structural Biology, 2001. **8**(5): p. 392-393.

488. Porter, K.A., et al., *What method to use for protein-protein docking?* Current opinion in structural biology, 2019. **55**: p. 1-7.

489. Ahmad, S., et al., *MOLECULAR DOCKING SIMPLIFIED : Literature Review.* 2022. **4**: p. 37-44.

490. Kuntz, I.D., et al., *A geometric approach to macromolecule-ligand interactions.* Journal of Molecular Biology, 1982. **161**(2): p. 269-288.

491. Martínez-Flores, D., et al., *SARS-CoV-2 Vaccines Based on the Spike Glycoprotein and Implications of New Viral Variants.* Frontiers in Immunology, 2021. **12**.

492. Verma, J. and N. Subbarao, *Insilico study on the effect of SARS-CoV-2 RBD hotspot mutants' interaction with ACE2 to understand the binding affinity and stability.* Virology, 2021. **561**: p. 107-116.

493. Dias, R. and W.F. de Azevedo, Jr., *Molecular docking algorithms.* Curr Drug Targets, 2008. **9**(12): p. 1040-7.

494. Raval, K. and T. Ganatra, *Basics, types and applications of molecular docking: A review.* IP International Journal of Comprehensive and Advanced Pharmacology, 2022. **7**: p. 12-16.

495. Mohammadi, S., et al., *Ensemble learning from ensemble docking: revisiting the optimum ensemble size problem.* Scientific Reports, 2022. **12**(1): p. 410.

496. Sacquin-Mora, S. and C. Prévost, *Docking Peptides on Proteins: How to Open a Lock, in the Dark, with a Flexible Key.* Structure, 2015. **23**(8): p. 1373-1374.

497. Weng, G., et al., *Comprehensive Evaluation of Fourteen Docking Programs on Protein–Peptide Complexes.* Journal of Chemical Theory and Computation, 2020. **16**(6): p. 3959-3969.

498. Ciemny, M., et al., *Protein–peptide docking: opportunities and challenges.* Drug Discovery Today, 2018. **23**(8): p. 1530-1537.

499. Agrawal, P., et al., *Benchmarking of different molecular docking methods for protein-peptide docking.* BMC Bioinformatics, 2019. **19**(13): p. 426.

500. Kangueane, P. and C. Nilofer, *Protein-Protein Docking: Methods and Tools*, in *Protein-Protein and Domain-Domain Interactions*, P. Kangueane and C. Nilofer, Editors. 2018, Springer Singapore: Singapore. p. 161-168.

501. Desta, I.T., et al., *Performance and Its Limits in Rigid Body Protein-Protein Docking.* Structure, 2020. **28**(9): p. 1071-1081.e3.

502. Eberhardt, J., et al., *AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings.* Journal of Chemical Information and Modeling, 2021. **61**(8): p. 3891-3898.

503. Trott, O. and A.J. Olson, *AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading.* J Comput Chem, 2010. **31**(2): p. 455-61.

504. Jaghoori, M.M., B. Bleijlevens, and S.D. Olabarriaga, *1001 Ways to run AutoDock Vina for virtual screening.* Journal of computer-aided molecular design, 2016. **30**(3): p. 237-249.

505. Nguyen, N.T., et al., *Autodock Vina Adopts More Accurate Binding Poses but Autodock4 Forms Better Binding Affinity.* Journal of Chemical Information and Modeling, 2020. **60**(1): p. 204-211.

506. Morris, G.M., et al., *AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility.* Journal of computational chemistry, 2009. **30**(16): p. 2785-2791.

507. Valdés-Tresanco, M.S., et al., *AMDock: a versatile graphical tool for assisting molecular docking with Autodock Vina and Autodock4.* Biology Direct, 2020. **15**(1): p. 12.

508. Bitencourt-Ferreira, G., V.O. Pintro, and W.F. de Azevedo, Jr., *Docking with AutoDock4.* Methods Mol Biol, 2019. **2053**: p. 125-148.

509. Liu, Y., et al., *CB-Dock: a web server for cavity detection-guided protein–ligand blind docking.* Acta Pharmacologica Sinica, 2020. **41**(1): p. 138-144.

510. Liu, Y., et al., *CB-Dock2: improved protein–ligand blind docking by integrating cavity detection, docking and homologous template fitting.* Nucleic Acids Research, 2022. **50**(W1): p. W159-W164.

511. Kozakov, D., et al., *The ClusPro web server for protein–protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.

512. London, N., et al., *Rosetta FlexPepDock web server--high resolution modeling of peptide-protein interactions.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W249-53.

513. Raveh, B., N. London, and O. Schueler-Furman, *Sub-angstrom modeling of complexes between flexible peptides and globular proteins.* Proteins, 2010. **78**(9): p. 2029-40.

514. Raveh, B., et al., *Rosetta FlexPepDock ab-initio: Simultaneous Folding, Docking and Refinement of Peptides onto Their Receptors.* PLOS ONE, 2011. **6**(4): p. e18934.

515. Ramírez-Aportela, E., J.R. López-Blanco, and P. Chacón, *FRODOCK 2.0: fast protein–protein docking server.* Bioinformatics, 2016. **32**(15): p. 2386-2388.

516. Lee, H., et al., *GalaxyPepDock: a protein-peptide docking tool based on interaction similarity and energy optimization.* Nucleic acids research, 2015. **43**(W1): p. W431-W435.

517. de Vries, S.J., M. van Dijk, and A.M. Bonvin, *The HADDOCK web server for data-driven biomolecular docking.* Nat Protoc, 2010. **5**(5): p. 883-97.

518. van Zundert, G.C.P., et al., *The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes.* Journal of Molecular Biology, 2016. **428**(4): p. 720-725.

519. Bonvin, A.M.J.J., *Flexible protein–protein docking.* Current Opinion in Structural Biology, 2006. **16**(2): p. 194-200.

520. Rodrigues, J.P., et al., *Insights on cross-species transmission of SARS-CoV-2 from structural modeling.* bioRxiv, 2020: p. 2020.06.05.136861.

521. Ozden, B., et al., *Benchmarking the Widely Used Structure-based Binding Affinity Predictors on the Spike-ACE2 Deep Mutational Interaction Set.* bioRxiv, 2022: p. 2022.04.18.488633.

522. Xue, L.C., et al., *PRODIGY: a web server for predicting the binding affinity of protein–protein complexes.* Bioinformatics, 2016. **32**(23): p. 3676-3678.

523. Zhou, P., et al., *HPEPDOCK: a web server for blind peptide-protein docking based on a hierarchical algorithm.* Nucleic Acids Res, 2018. **46**(W1): p. W443-w450.

524. de Vries, S.J., et al., *The pepATTRACT web server for blind, large-scale peptide-protein docking.* Nucleic Acids Res, 2017. **45**(W1): p. W361-w364.

525. Schindler, Christina E.M., Sjoerd J. de Vries, and M. Zacharias, *Fully Blind Peptide-Protein Docking with pepATTRACT.* Structure, 2015. **23**(8): p. 1507-1515.

526. Jiménez-García, B., C. Pons, and J. Fernández-Recio, *pyDockWEB: a web server for rigid-body protein–protein docking using electrostatics and desolvation scoring.* Bioinformatics, 2013. **29**(13): p. 1698-1699.

527. Torchala, M., et al., *SwarmDock: a server for flexible protein–protein docking.* Bioinformatics, 2013. **29**(6): p. 807-809.

528. Moal, I.H., et al., *A Guide for Protein-Protein Docking Using SwarmDock.* Methods Mol Biol, 2020. **2165**: p. 199-216.

529. Pierce, B.G., et al., *ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers.* Bioinformatics (Oxford, England), 2014. **30**(12): p. 1771-1773.

530. Temussi, P.A., et al., *Bioactive conformation of linear peptides in solution: an elusive goal?* Biopolymers, 1989. **28**(1): p. 91-107.

531. Ruehl, M., et al., *Hydroxyproline-containing collagen analogs trigger the release and activation of collagen-sequestered proMMP-2 by competition with*

prodomain-derived peptide P33-42. Fibrogenesis & tissue repair, 2011. **4**(1): p. 1-1.

532. Warnecke, A., et al., *PyTMs: a useful PyMOL plugin for modeling common post-translational modifications.* BMC Bioinformatics, 2014. **15**(1): p. 370.

533. Tien, M.Z., et al., *PeptideBuilder: A simple Python library to generate model peptides.* PeerJ, 2013. **1**: p. e80-e80.

534. Christensen, A.S., T. Hamelryck, and J.H. Jensen, *FragBuilder: an efficient Python library to setup quantum chemistry calculations on peptides models.* PeerJ, 2014. **2**: p. e277.

535. Yan, Y., D. Zhang, and S.-Y. Huang, *Efficient conformational ensemble generation of protein-bound peptides.* Journal of Cheminformatics, 2017. **9**.

536. Singh, S., et al., *PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues.* Biology Direct, 2015. **10**(1): p. 73.

537. Margreitter, C., D. Petrov, and B. Zagrovic, *Vienna-PTM web server: a toolkit for MD simulations of protein post-translational modifications.* Nucleic Acids Research, 2013. **41**(W1): p. W422-W426.

538. Ortiz, A.R., C.E.M. Strauss, and O. Olmea, *MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison.* Protein science : a publication of the Protein Society, 2002. **11**(11): p. 2606-2621.

539. Shortle, D., K.T. Simons, and D. Baker, *Clustering of low-energy conformations near the native structures of small proteins.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**(19): p. 11158-11162.

540. Williams, C.J., et al., *MolProbity: More and better reference data for improved all-atom structure validation.* Protein Science, 2018. **27**(1): p. 293-315.

541. Buxbaum, E., *Amino Acids*, in *Fundamentals of Protein Structure and Function*, E. Buxbaum, Editor. 2015, Springer International Publishing: Cham. p. 3-13.

542. Melnikov, S., et al., *Molecular insights into protein synthesis with proline residues.* EMBO reports, 2016. **17**(12): p. 1776-1784.

543. Hutchinson, E.G. and J.M. Thornton, *A revised set of potentials for beta-turn formation in proteins.* Protein Sci, 1994. **3**(12): p. 2207-16.

544. de Marco, A., *Strategies for successful recombinant expression of disulfide bond-dependent proteins in Escherichia coli.* Microbial Cell Factories, 2009. **8**: p. 26-26.

545. Riggs, P., *Expression and purification of maltose-binding protein fusions.* Curr Protoc Mol Biol, 2001. **Chapter 16**: p. Unit16.6.

546. Marcu, Ş.-B., S. Tăbîrcă, and M. Tangney, *An Overview of Alphafold's Breakthrough.* Frontiers in Artificial Intelligence, 2022. **5**.

547. Mirdita, M., et al., *ColabFold: making protein folding accessible to all.* Nature Methods, 2022. **19**(6): p. 679-682.

548. Mariani, V., et al., *lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests.* Bioinformatics (Oxford, England), 2013. **29**(21): p. 2722-2728.

549. Tunyasuvunakool, K., et al., *Highly accurate protein structure prediction for the human proteome.* Nature, 2021. **596**(7873): p. 590-596.

550. Zhang, Y. and J. Skolnick, *Scoring function for automated assessment of protein structure template quality.* Proteins, 2004. **57**(4): p. 702-10.

551. Mitternacht, S., *FreeSASA: An open source C library for solvent accessible surface area calculations.* F1000Research, 2016. **5**: p. 189-189.

552. Honegger, A., *Intermediate PyMOL*. 2017.

553. Ullah, R., et al., *Activity of the Human Rhinovirus 3C Protease Studied in Various Buffers, Additives and Detergents Solutions for Recombinant Protein Production.* PloS one, 2016. **11**(4): p. e0153436-e0153436.

554. Takara Bio. *HRV 3C Protease*. n.d. [cited 2022 August 01]; v201612Da:[Available from: http://www.takara-bio.com/.

555. Ma, Y., C.J. Lee, and J.S. Park, *Strategies for Optimizing the Production of Proteins and Peptides with Multiple Disulfide Bonds.* Antibiotics (Basel), 2020. **9**(9).

556. Ahmed, N., et al., *Method for efficient soluble expression and purification of recombinant human interleukin-15.* Protein Expression and Purification, 2021. **177**: p. 105746.

557. Hand, K., M.C. Wilkinson, and J. Madine, *Isolation and purification of recombinant immunoglobulin light chain variable domains from the periplasmic space of Escherichia coli.* PLOS ONE, 2018. **13**(10): p. e0206167.

558. Wagner, S., et al., *Tuning Escherichia coli for membrane protein overexpression.* Proceedings of the National Academy of Sciences, 2008. **105**(38): p. 14371-14376.

559. Gonzalez-Perez, D., et al., *Random and combinatorial mutagenesis for improved total production of secretory target protein in Escherichia coli.* Scientific Reports, 2021. **11**(1): p. 5290.

560. Xu, X., et al., *Contributions of the MMP-2 collagen binding domain to gelatin cleavage. Substrate binding via the collagen binding domain is required for hydrolysis of gelatin but not short peptides.* Matrix Biol, 2004. **23**(3): p. 171-81.

561. Berkmen, M., *Production of disulfide-bonded proteins in Escherichia coli.* Protein Expression and Purification, 2012. **82**(1): p. 240-251.

562. Antunes, D.A., et al., *Structure-based Methods for Binding Mode and Binding Affinity Prediction for Peptide-MHC Complexes.* Current topics in medicinal chemistry, 2018. **18**(26): p. 2239-2255.

563. Lee, A.C., et al., *A Comprehensive Review on Current Advances in Peptide Drug Development and Design.* International Journal of Molecular Sciences, 2019. **20**(10).

564. Pinzi, L. and G. Rastelli, *Molecular Docking: Shifting Paradigms in Drug Discovery.* International journal of molecular sciences, 2019. **20**(18): p. 4331.

565. Chang, L., A. Mondal, and A. Perez, *Towards rational computational peptide design.* Frontiers in Bioinformatics, 2022. **2**.

566. Meng, X.-Y., et al., *Molecular docking: a powerful approach for structure-based drug discovery.* Current computer-aided drug design, 2011. **7**(2): p. 146-157.

567. Kastritis, P.L., et al., *Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface.* J Mol Biol, 2014. **426**(14): p. 2632-52.

568. Tsaban, T., et al., *Harnessing protein folding neural networks for peptide–protein docking.* Nature Communications, 2022. **13**(1): p. 176.

569. Khramushin, A., et al., *PatchMAN docking: Modeling peptide-protein interactions in the context of the receptor surface*. 2021.

570. Khramushin, A., et al., *Matching protein surface structural patches for high-resolution blind peptide docking.* Proc Natl Acad Sci U S A, 2022. **119**(18): p. e2121153119.

571. Pignataro, M.F., M.G. Herrera, and V.I. Dodero, *Evaluation of Peptide/Protein Self-Assembly and Aggregation by Spectroscopic Methods.* Molecules (Basel, Switzerland), 2020. **25**(20): p. 4854.

572. Makwana, K.M. and R. Mahalakshmi, *Implications of aromatic–aromatic interactions: From protein structures to peptide models.* Protein Science, 2015. **24**(12): p. 1920-1933.

573. Mattanovich, D., et al., *Recombinant Protein Production in Yeasts*, in *Recombinant Gene Expression*, A. Lorence, Editor. 2012, Humana Press: Totowa, NJ. p. 329-358.

574. Nielsen, C.A.F., et al., *The important ergot alkaloid intermediate chanoclavine-I produced in the yeast saccharomyces cerevisiae by the combined action of EasC and EasE from aspergillus japonicus.* Microbial Cell Factories, 2014. **13**(1).

575. Li, Y., *Self-cleaving fusion tags for recombinant protein production.* Biotechnology Letters, 2011. **33**(5): p. 869-881.

576. Cao, L., et al., *Design of protein-binding proteins from the target structure alone.* Nature, 2022. **605**(7910): p. 551-560.

577. Shane Anderson, A. and R.F. Loeser, *Why is osteoarthritis an age-related disease?* Best practice & research. Clinical rheumatology, 2010. **24**(1): p. 15-26.

578. Harrell, C.R., et al., *Mesenchymal stem cell-based therapy of osteoarthritis: Current knowledge and future perspectives.* Biomedicine & Pharmacotherapy, 2019. **109**: p. 2318-2326.

579. Dennis, J.E., et al., *Targeted delivery of progenitor cells for cartilage repair.* Journal of Orthopaedic Research, 2004. **22**(4): p. 735-741.

580. Osiecka-Iwan, A., et al., *Antigenic and immunogenic properties of chondrocytes. Implications for chondrocyte therapeutic transplantation and pathogenesis of inflammatory and degenerative joint diseases.* Central-European journal of immunology, 2018. **43**(2): p. 209-219.

581. Shegos, C.J. and A.F. Chaudhry, *A narrative review of mesenchymal stem cells effect on osteoarthritis.* Annals of Joint, 2021. **7**.

582. Vinod, E., B. Ramasamy, and U. Kachroo, *Comparison of immunogenic markers of human chondrocytes and chondroprogenitors derived from non-diseased and osteoarthritic articular cartilage.* Journal of Orthopaedics, Trauma and Rehabilitation, 2020. **27**(1): p. 63-67.

583. Hofer, H.R. and R.S. Tuan, *Secreted trophic factors of mesenchymal stem cells support neurovascular and musculoskeletal therapies.* Stem Cell Research & Therapy, 2016. **7**(1): p. 131.

584. Caplan, A.I. and J.E. Dennis, *Mesenchymal stem cells as trophic mediators.* J Cell Biochem, 2006. **98**(5): p. 1076-84.

585. Afizah, H. and J.H.P. Hui, *Mesenchymal stem cell therapy for osteoarthritis.* Journal of clinical orthopaedics and trauma, 2016. **7**(3): p. 177-182.

586. Mancuso, P., et al., *Mesenchymal Stem Cell Therapy for Osteoarthritis: The Critical Role of the Cell Secretome.* Frontiers in Bioengineering and Biotechnology, 2019. **7**.

587. Dickinson, S.C., et al., *The Wnt5a Receptor, Receptor Tyrosine Kinase-Like Orphan Receptor 2, Is a Predictive Cell Surface Marker of Human Mesenchymal Stem Cells with an Enhanced Capacity for Chondrogenic Differentiation.* Stem Cells, 2017. **35**(11): p. 2280-2291.

588. Chahal, J., et al., *Bone Marrow Mesenchymal Stromal Cells in Patients with Osteoarthritis Results in Overall Improvement in Pain and Symptoms and Reduces Synovial Inflammation.* Stem Cells Transl Med, 2019.

589. Goldberg, A., et al., *The use of mesenchymal stem cells for cartilage repair and regeneration: a systematic review.* Journal of Orthopaedic Surgery and Research, 2017. **12**(1): p. 39.

590. Armiento, A.R., M. Alini, and M.J. Stoddart, *Articular fibrocartilage - Why does hyaline cartilage fail to repair?* Advanced Drug Delivery Reviews, 2019. **146**: p. 289-305.

591. Yang, J., et al., *Tropoelastin improves adhesion and migration of intra-articular injected infrapatellar fat pad MSCs and reduces osteoarthritis progression.* Bioactive Materials, 2022. **10**: p. 443-459.

592. Andersen, C., et al., *Human integrin α10β1-selected mesenchymal stem cells home to cartilage defects in the rabbit knee and assume a chondrocyte-like phenotype.* Stem Cell Research & Therapy, 2022. **13**(1): p. 206.

593. Song, Y. and C. Jorgensen, *Mesenchymal Stromal Cells in Osteoarthritis: Evidence for Structural Benefit and Cartilage Repair.* Biomedicines, 2022. **10**(6).

594. Koh, Y.G., et al., *Adipose-Derived Mesenchymal Stem Cells With Microfracture Versus Microfracture Alone: 2-Year Follow-up of a Prospective Randomized Trial.* Arthroscopy, 2016. **32**(1): p. 97-109.

595. Ray, M., et al., *Essential interpretations of bioinformatics in COVID-19 pandemic.* Meta gene, 2021. **27**: p. 100844-100844.

596. Cannataro, M. and A. Harrison, *Bioinformatics helping to mitigate the impact of COVID-19 – Editorial.* Briefings in Bioinformatics, 2021. **22**(2): p. 613-615.

597. Basit, A., et al., *Designing Short Peptides to Block the Interaction of SARS-CoV-2 and Human ACE2 for COVID-19 Therapeutics.* Front Pharmacol, 2021. **12**: p. 731828.

598. Pantsar, T. and A. Poso, *Binding Affinity via Docking: Fact and Fiction.* Molecules (Basel, Switzerland), 2018. **23**(8): p. 1899.

599. Romero-Molina, S., et al., *PPI-Affinity: A Web Tool for the Prediction and Optimization of Protein–Peptide and Protein–Protein Binding Affinity.* Journal of Proteome Research, 2022. **21**(8): p. 1829-1841.

600. Ahmad, M., et al., *Protein expression in Pichia pastoris: recent achievements and perspectives for heterologous protein production.* Applied Microbiology and Biotechnology, 2014. **98**(12): p. 5301-5317.

601. Poole, R., *Osteoarthritis as a Whole Joint Disease.* HSS journal : the musculoskeletal journal of Hospital for Special Surgery, 2012. **8**: p. 4-6.

602. Goldring, M.B. and F. Berenbaum, *Emerging targets in osteoarthritis therapy.* Current opinion in pharmacology, 2015. **22**: p. 51-63.