Author:
Bochel, Alice J

Title:
Structural characterisation of a carbohydrate binding domain of the human cation-independent mannose 6-phosphate/ IGF2 receptor

Structural characterisation of a carbohydrate binding domain of the human cation-independent mannose 6-phosphate/ IGF2 receptor

Alice J Bochel

A dissertation submitted to the University of Bristol in accordance with the requirements for award of the degree of PhD in Chemistry in the Faculty of Science

School of Chemistry

September 2020

Word Count: 62,294

Abstract

The cation-independent mannose 6-phosphate/ Insulin-like growth factor-2 receptor (CI-MPR/ IGF2R) is a ~300 kDa transmembrane glycoprotein that is critical for intracellular protein trafficking, lysosome biogenesis and regulation of cell growth. The extracellular region consists of fifteen domains homologous to one another including mannose 6-phosphate (M6P) binding domains (D) 3, 5, 9 and 15, and IGF2 binding domain 11. To date, high-resolution structures have been determined for human D1-5 and D11-14. Although low resolution cryo-EM structures of bovine CI-MPR have recently been determined at pH 4.5 and 7.4, a structure of the full extracellular region of human CI-MPR has yet to be determined.

This thesis details structural studies on the central, uncharacterised region of human CI-MPR with particular focus on the elusive, specific and high-affinity M6P binding domain, D9. A modular approach has resulted in crystal structures of human CI-MPR D8, D9-10 and D7-11. D9-10 forms a rigid homodimer stabilised by a bridging N-linked glycan and maintained in D7-11, whereby two penta-domains intertwine to form a dimeric helical-type coil. Remarkably the D7-11 structure closely matches an IGF2 bound state of the receptor, suggesting this may be an intrinsically stable conformation at neutral pH. Interdomain clusters of histidine and proline residues at the D9-10 and D11-12 interfaces may impart receptor rigidity and play a role in cargo dissociation and structural rearrangement at low pH.

A parallel project took an iterative, structure-based approach to engineer a synthetic lectin. The hydrophobic IGF2 binding site of CI-MPR D11 was mutated by site-directed mutagenesis to resemble the positively charged M6P binding sites of D3 and 9. Following preparation, D11 mutants were screened by ^{1}H-^{15}N HSQC NMR for binding to monosaccharides. Although chemical shift perturbations were observed following addition of M6P, further work is required to validate these preliminary results.

Acknowledgements

I would like to thank my supervisor Professor Matt Crump for all his guidance and excitement throughout this PhD, particularly in the final year, and for supporting me in my next step. Similarly, I would also like to thank Professor Bass Hassan (University of Oxford) for his discussions and invaluable insights into the project.

I must say a huge thank you to Dr Chris Williams and Dr Ash Winter for training me in the lab and for regular troubleshooting. Chris, thank you for also teaching me how to solve crystal structures, for running so many (ambiguous) NMR experiments with me and for all the ‘team IGF2R’ discussions. Ash, thank you for the data collection help at Diamond and for letting me nap on the graveyard shifts.

My thanks is also extended to current and past members of the lab: Erik, Nahida, Rowe, Goodwin, Tom, Paul, Lina, Catherine, Marija, Luoyi and Angus. Thanks also to the MSci undergraduate students who helped along the way: Sarah, Jack, Leo and Georgia.

I also wish to thank members of the Berger and Berger-Schaffitzel group. In particular, thank you Dr Satish Yadivik and Professor Christiane Berger-Schaffitzel for helping me see my protein by negative stain. Thank you Kapil and Fru for teaching me baculoviral expression and for regularly supplying me with fresh cells. I am very grateful to Professor Imre Berger for the generous use of his eukaryotic expression facility, culture media and his enthusiasm for the project.

Lastly, I would also like to thank my parents, Issy and Harvey for all of their support, interest and patience over these last four years. Particularly a special thank you to Harvey for keeping me sane and always believing in me.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Research Degree Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, the work is the candidate's own work. Work done in collaboration with, or with the assistance of, others, is indicated as such. Any views expressed in the dissertation are those of the author.

SIGNED: …A.J.Bochel........................... DATE: …25/08/20.............................

Abbreviations

Abbreviation	Definition
AcNPV	*Autographa californica* nuclear polyhedrosis virus
BEVS	Baculovirus expression vector system
BSA	Bovine serum albumin
CBD	Carbohydrate binding domain
CD	Circular dichroism
CD-MPR	Cation-dependent mannose 6-phosphate receptor
CI-MPR	Cation-independent mannose 6-phosphate receptor
CNX	Calnexin
CREG	Cellular repressor of E1A stimulated genes
CRT	Calreticulin
CV	Column volume
DPA	Day of proliferation arrest
DTT	Dithiothrietol
EDTA	Ethylenediaminetetraacetic acid
EE	Early endosome
ER	Endoplamsic reticulum
ERAD	ER associated degradation
ERT	Enzyme replacement therapy
EndoH	Endo β-N-acetylglucosaminidase H
ESI-TOF MS	Electrospray ionisation time-of-flight mass spectrometry
FACS	Fluorescence activated cell sorting
FNII	Fibronectin type II
G6P	Glucose 6-phosphate
GAA	Acid alpha glucosidase
Gal	Galactose
GalNAc	N-Acetylgalactosamine
Glc	Glucose
GlcN	Glucosamine
GlcNAc	N-Acetylglucosamine
HEK	Human embryonic kidney cell line
HSQC	Heteronuclear single quantum coherence
IEX	Ion exchange chromatography
IGF	Insulin-like growth factor
IGF2R	Insulin-like growth factor 2 receptor
IGFBP	Insulin-like growth factor binding protein
IMAC	Immobilised metal affinity chromatography
IPTG	Isopropyl β-D-1-thiogalactopyranoside
ITC	Isothermal titration calorimetry
LacNAc	N-acetyllactosamine
LB	Lysogeny broth
LE	Late endosome
LIF	Leukaemia inhibitory factor
M6P	Mannose 6-phosphate
Man	Mannose
MRH	Mannose receptor homology
NMR	Nuclear magnetic resonance spectroscopy
NOE	Nuclear overhauser effect
PBS	Phosphate buffered saline
PCR	Polymerase chain reaction
PDB	Protein database
PM	Plasma membrane

PNGaseF	Peptide N-glycosidase F
RER	Rough endoplasmic reticulum
RMSD	Root mean squared deviation
RPTPμ	Receptor protein tyrosine phosphatase μ
SAXS	Small angle X-ray scattering
SDS-PAGE	Sodium dodecyl sulphate- polyacrylamide gel electrophoresis
SE-AUC	Sedimentation-equilibrium analytical ultracentrifugation
SEC	Size exclusion chromatography
SEC-MALS	Size exclusion chromatography multi-angle light scattering
Sf21	*Spodoptera frugiperda* cell line 21
SLe^{X}	Sialyl lewis X
SPR	Surface plasmon resonance
STD	Saturated transfer difference NMR
TBS	Tris buffered saline
TEM	Transmission electron microscopy
TGN	Trans-golgi network
UCE	Uncovering enzyme
uPAR	Urokinase receptor
WaterLOGSY	Water ligand observed gradient spectroscopy NMR
X-Gal	5-Bromo-4-Chloro-3-Indoyl β-D-galactopyranoside
YFP	Yellow fluorescent protein

Publications

Bochel, A. J. *et al.* Structure of the Human Cation-Independent Mannose 6-Phosphate/ IGF2 Receptor Domains 7–11 Uncovers the Mannose 6-Phosphate Binding Site of Domain 9. *Structure* **28**, 1300-1312 (2020). doi:10.1016/j.str.2020.08.002.

Deposited data

Coordinates and structure factors of human D9-10, D8 and D7-11 have been deposited in the PDB and have the accession codes 6Z30, 6Z31 and 6Z32. SAXS data of human D9-10 has been deposited in the SASDBD with the accession codes SASDH59,69,79 and SASDJ23.

Contents

1. Introduction

1.1. Glycosylation

Protein glycosylation is a common and diverse post-translational modification that affects the solubility, folding, trafficking and secretion of newly synthesised glycoproteins.[1] On plasma membrane proteins glycosylation further influences protein signalling, endocytosis and half-life.[1] It is unsurprising, therefore, that defects in protein glycosylation are implicated in numerous diseases and there are over 130 congenital disorders of glycosylation.[2] Although low in incidence (approximately 1/10,000 per capita in Europe), the ubiquitous nature of protein glycosylation means that these autosomal recessive disorders exhibit a range of symptoms including neurological defects, cardiac disease and hepatopathy.[2] Changes in protein glycosylation have also been observed in patients with Alzheimer disease and cancers.[3,4]

Sugars are the most information rich macromolecule: just six monosaccharides (examples in Figure 1) can yield $>10^{12}$ unique polysaccharide glycan structures (compared to six nucleotides generating 4096 unique oligonucleotides or six amino acids forming 6×10^{7} unique peptides).[5,6] This is due to structural properties of the monosaccharide (such as ring size and modifications, epimers and anomers) and polysaccharides (such as linkage configurations and branch positions).[5]

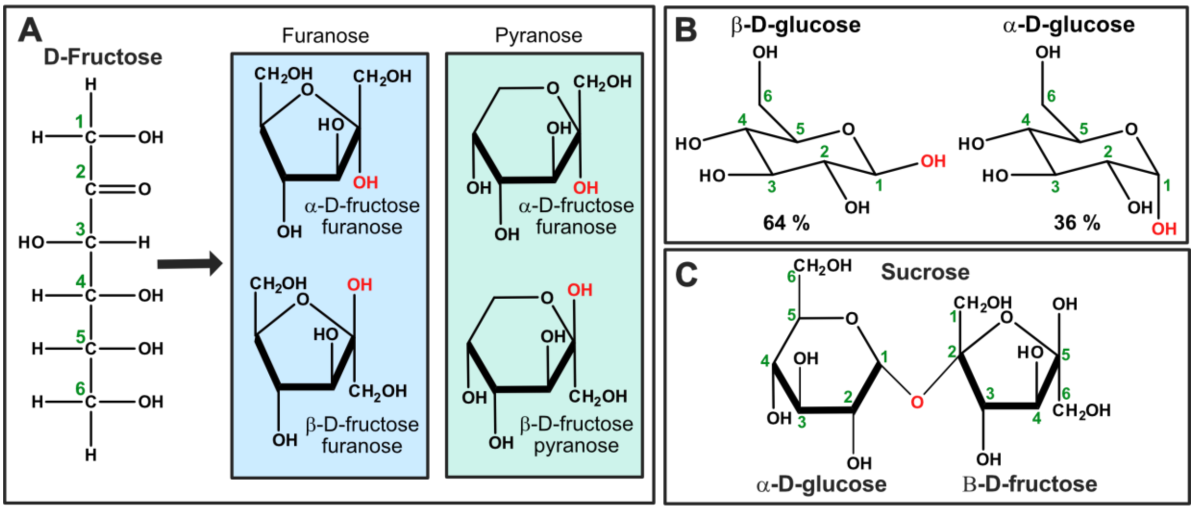


Figure 1: The structures of monosaccharides. A: The monosaccharide fructose, which is shown as a linear Fischer projection on the left, can circularise, resulting in four conformations. Two of these are 5-membered ketoses termed furanoses (blue box). The remaining two are 6-membered aldoses termed pyranoses (green box). Monosaccharides can be classified further by the position of the hydroxyl group (1'OH, red) at position C1, the anomeric carbon. When the 1'OH is below the plane of the ring, monosaccharides are termed α anomers. In β anomers the 1'OH is above the ring. The four conformations of fructose are drawn here as Haworth projections. **B:** Alternatively, monosaccharides may be drawn in skeletal form. The structures of β-D-glucose and α-D-glucose are shown here. β-D-glucose is the predominant conformer due to the equatorial positioning of all hydroxyl groups. **C:** The structure of sucrose, a disaccharide formed by condensation reaction of α-D-glucose and β-D-fructose. The monosaccharides are linked by an α1-2 O-glycosidic bond.

1. **Introduction**

Protein glycosylation is a complex modification with ~700 proteins required to generate the ~7000 different glycan structures found in mammalian cells.[1] Where present, glycans make up an average of ~20 % of a proteins molecular weight.[7] Protein glycosylation is classified by the glycan-peptide linkage [1] into the six classes described below: N-linked, O-linked, C-linked, P-linked, G-linked and S-linked glycosylation.[7]

N-linked glycosylation

The most common class of protein glycosylation is N-linked glycosylation, with an estimated 76 % of eukaryotic proteins being potentially derivatised in this way.[8] N-linked glycosylation is sequence specific, with only asparagine residues in the conserved sequence N-X-S/T (termed NST sequon, whereby X is any amino acid except P) being modified.[9] Statistical analysis of glycoprotein structures in the PDB revealed that ~70 % of N-linked glycans occur on N-X-T sequons, with only ~30 % on N-X-S sequons.[10] However, in total, only ~66 % of NST sequons are glycosylated suggesting that glycosylation status is also more subtly affected by sequence and structural constraints.[7] For example, the presence of cysteine, bulky or charged residues (C, P, W, D, E, R, K) in the centre of the NST sequon inhibits N-linked glycosylation.[11,12] Similarly, small non-polar amino acids are often found at the +1 and -2 positions adjacent to the asparagine of NST sequons, with larger hydrophobic amino acids at +3 and +5 positions.[10,13] The majority (~44 %) of N-linked glycans are found on flat, convex surfaces of the protein with ~27 % found on β-turns and only ~10 % found on α-helices.[10]

N-linked glycans are derived from a single parent glycan precursor (Figure 2) consisting of two N-acetylglucosamine (GlcNAc), nine mannose (Man) and three glucose (Glc) residues ($GlcNAc_2Man_9Glc_3$).[14] This glycan precursor is synthesised at the membrane of the rough endoplasmic reticulum (RER).[14] The first seven sugars ($GlcNAc_2Man_5$) of this glycan are assembled on the cytosolic side of the RER attached to a dolichol-pyrophosphate lipid anchor.[15] This glycan precursor is transferred to the luminal surface of the RER by a flippase enzyme, where the remaining four mannose and three glucose residues are added.[16] The assembled glycan precursor ($GlcNAc_2Man_9Glc_3$) is transferred en bloc to an asparagine residue in the NST sequon of a nascent polypeptide by an oligosaccharyltransferase (OST) enzyme.[17] Once attached to the protein, the glycan precursor is trimmed by Glucosidase I, II and α1-2 Mannosidase before transport to the Golgi body for further trimming and processing.[15]

Mammalian N-linked glycans fall into three categories (Figure 2): high mannose, hybrid and complex glycans.[15] Each category shares the same chitobiose core of $GlcNAc_2Man_3$.[15] Addition of a GlcNAc residue to a mannose of the chitobiose core initiates biosynthesis of hybrid and complex glycans.[18] Hybrid and complex glycans are characterised by the presence of extended arms/ branches consisting of GlcNAc-Gal (β1-4 linked), termed type 2 N-acetyllactosamine (LacNAc), capped with sialic acid, fucose or an additional galactose (Figure 2C, 2D).[18] Sialic acid is a family of ~40 nine-carbon monosaccharides characterised by an anomeric carboxylic acid group, an amino group at C5 and a three-carbon extension at C6 (Figure 2E).[19,20] The chitobiose core of complex glycans may also be modified, most frequently by the addition of an α1-6 linked fucose residue to the first GlcNAc (Figure 2D).[15,18]

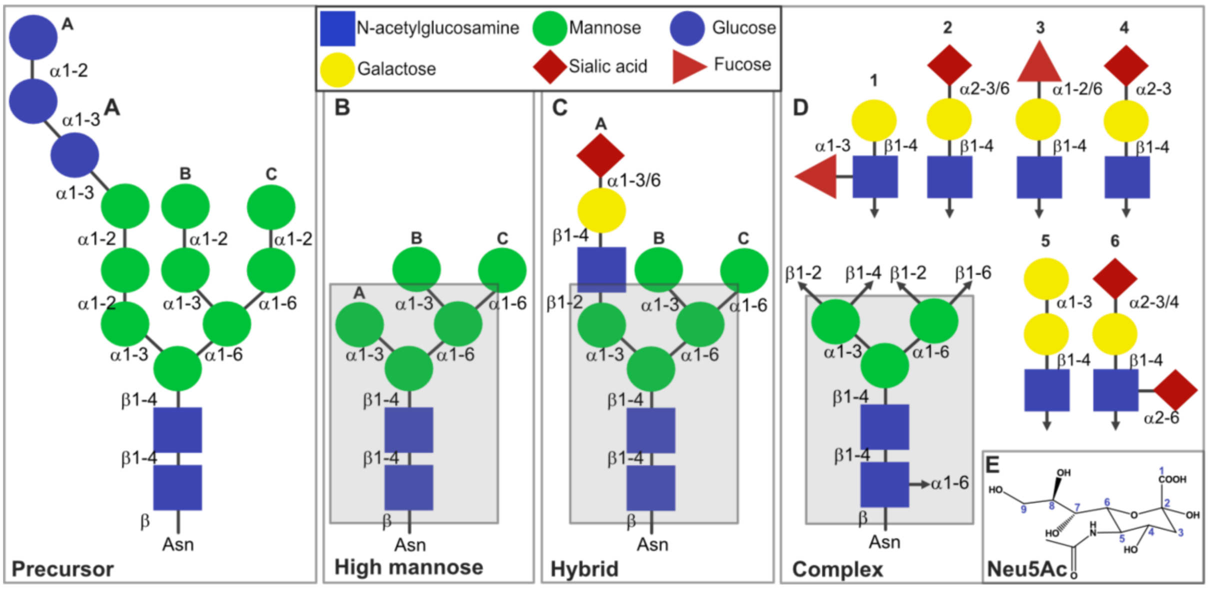


Figure 2: The structures of mammalian N-linked glycans.[15,21] **A:** The structure of the N-linked glycan precursor. **B:** The structure of the high mannose N-linked glycan. **C:** The structure of the hybrid N-linked glycan. **D:** The structure of the complex N-linked glycan, in which the terminal mannose residues of the chitobiose core (grey box) are modified with a variety of sugars. Additionally, the first GlcNAc residue of the chitobiose core may be modified by an α1-6 linked fucose. The A, B and C arms of the precursor, high mannose and hybrid glycans are labelled. All schematic glycan depictions conform to nomenclature recommended by the Consortium for Functional Glycomics. **E:** The structure of the most common sialic acid, 5-N-acetylneuraminic acid (Neu5Ac).

The importance of N-linked glycosylation was first reported in the 1970s by observing the effects of Tunicamycin, a nucleoside analog that inhibits the first step of N-linked glycosylation (transfer of GlcNAc onto the dolichol-pyrophosphate lipid anchor), causing ER stress.[14,16] N-linked glycosylation is directly linked to protein folding and quality control mechanisms. For example, Glucosidase I and II remove the two terminal glucose residues from the glycan precursor, allowing the chaperone proteins Calnexin (CNX) and Calreticulin (CRT) to recognise the single, remaining glucose residue and facilitate protein folding.[22] Only upon removal of this remaining glucose residue by Glucosidase II can the glycoprotein proceed to the Golgi body.[14] α1-2 mannosidases process the de-glucosylated glycoproteins, lowering their affinity for CNX and CRT.[23] Misfolded glycoproteins are re-glucosylated in the RER by a UDP-Glc glucosyltransferase for recognition by the chaperone proteins.[14] Terminally misfolded glycoproteins are trimmed further by α1-2 mannosidases to expose a terminal α1-6 mannose residue that signals the glycoprotein for ER-associated degradation (ERAD).[23,24]

O-linked glycosylation

O-linked glycosylation consists of the attachment of a glycan to the oxygen of a serine or threonine hydroxyl group.[7] Although O-linked glycosylation is sequence specific it is much harder to predict and typically occurs on unstructured regions of peptide that are rich in serine, threonine and proline.[21,25] These regions are termed mucin domains due to their prevalence in secreted and transmembrane glycoproteins of the mucus-secreting epithelial cells lining the tracheobronchial and gastrointestinal tracts.[26]

The most common form of O-linked glycosylation involves the mucin linkage. There are eight core O-linked mucin glycan structures, with cores 1-4 being the most common (Figure 3). Each core structure is characterised by a S/T α-linked to an N-acetylgalactosamine (GalNAc) that may be modified by the addition of glucose, galactose, N-acetylglucosamine, sialic acid, N-acetylgalactosamine or sulphate.[26]

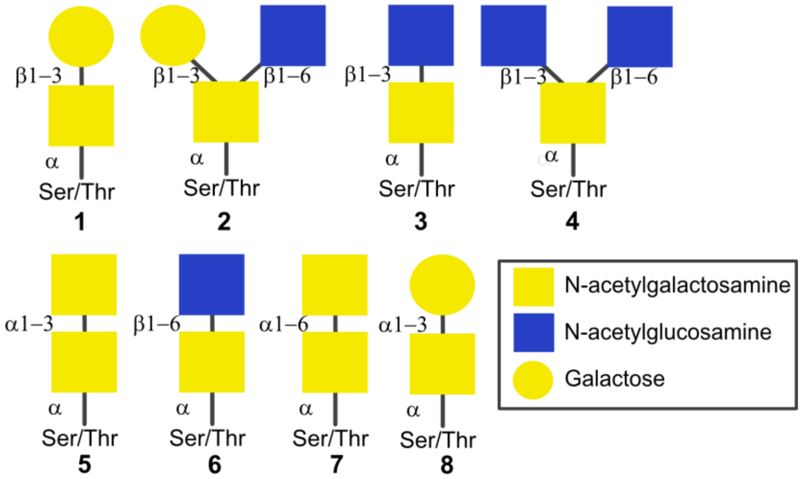


Figure 3: The structures of the eight core mammalian O-linked mucin glycans.[21]

However, there are a small number of other forms of O-linked glycosylation that do not contain the mucin linkage.[1] For example, the cysteine rich EGF domains of the transmembrane signalling protein Notch can be modified by a β-linked glucose residue (to the serine of C-X-S-X-A/P-C) or an α-linked fucose residue (to the serine or threonine of C-X-X-X-X-S/T-C).[27] These sugar residues are further glycosylated to influence Notch ligand binding and function.[27]

GPI-linked glycosylation

GPI-linked glycosylation only occurs on approximately 1 % of all eukaryotic proteins and involves the attachment of a pre-formed glycolipid, glycophosphatidlylinositol (GPI, Figure 4), to the C-terminus of a soluble protein.[28] This anchors the protein to the extracellular surface of the plasma membrane where it can act as a cell surface receptor or cell adhesion molecule.[28,29] Example mammalian GPI anchored proteins include the neural cell adhesion molecule (NCAM), which mediates cell-cell interactions in the brain, and alkaline phosphatase (APase), which catalyses dephosphorylation.[29]

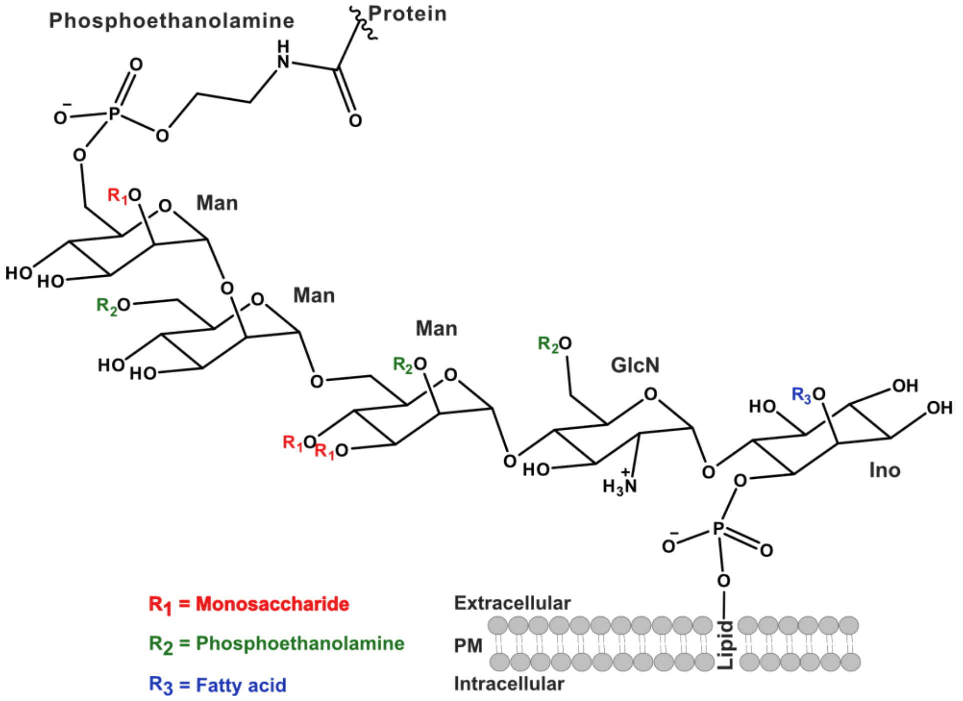


Figure 4: The structure of the glycophosphatidlylinositol (GPI) anchor. A phosphoethanolamine linker attaches a soluble protein to a pentasaccharide unit of three mannose residues, glucosamine and inositol (Man$_3$GlcNIno) that is in turn attached to a lipid in the plasma membrane. The modifications to the pentasaccharide unit and the lipid structure are tissue dependent.[29]

C-linked glycosylation

C-linked glycosylation involves the glycosylation of tryptophan residues with an α-mannose residue.[1] Uniquely, a carbon-carbon bond is formed between C1 of mannose and C2 of the first tryptophan in the consensus sequence W-X-X-W' (whereby X is any amino acid) (Figure 5A).[30] The presence of C-mannosyltransferases in mammals, birds, amphibians and fish suggests this is a widespread post-translational modification.[30] One of the most well studied examples of human C-mannosylation is that of thrombospondin, a large homotrimeric protein that, upon secretion, facilitates platelet aggregation resulting in wound healing, inflammation, tumour growth and metastasis.[30] Thrombospondin type 1 repeats containing the above consensus sequence and thus C-mannosylation are also found within netrin receptors and the ADAMTS family of secreted proteases, with C-mannosylation being critical for protein secretion.[31,32]

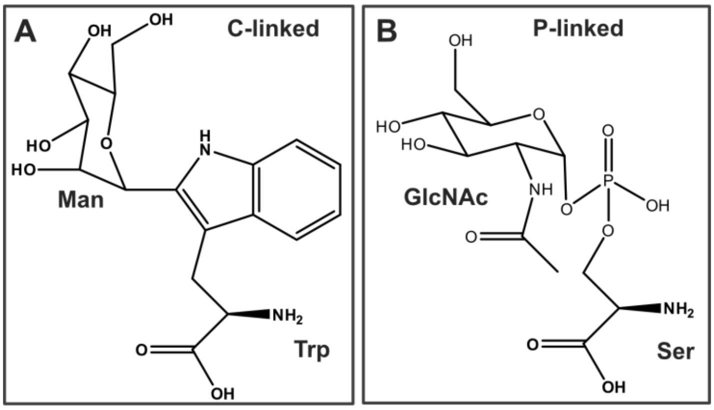


Figure 5: Examples of C-linked and P-linked protein glycosylation. A: C-linked glycosylation whereby tryptophan is modified with an α-D-mannose residue. **B:** P-linked glycosylation whereby a phosphoserine residue is modified with an α-D-N-acetylglucosamine (GlcNAc) residue.

P-linked glycosylation

A minor class of glycosylation, P-linked glycosylation is characterised by glycosylation of phosphorylated serine residues and has been observed in a number of unicellular parasites.[7,33] For example, the cysteine proteases (CP) proteinase I, CP4 and CP5 of the amoeba *Dictyostelium discoideum* have all displayed glycosylation with the addition of GlcNAc to phosphoserine residues in three sequence motifs: polyS, SGSQ and SGSG.[34,35] However the impact/ function of this modification is unknown.

Another example of phosphoglycosylation occurs on secreted acid phosphatase found in the protozoan parasite *Leishmania mexicana*.[36] Phosphorylated serine residues (of serine rich sequences) are mannosylated, creating a mannose α1-PO_4-Ser linkage (Figure 5B).[36] The resulting P-linked glycan which is rich in mannose and galactose could be cleaved from the protein by Peptide N-glycosidase F (PNGase F),[36] similarly to N-linked and O-linked glycans.

S-linked glycosylation

With only a handful of examples, the final class of glycosylation, S-linked, is the least common. The GlcNAc transferase employed in O-linked glycosylation may also attach a GlcNAc residue to cysteine (Figure 6).[37] The first observation of S-linked glycosylation arose from mass spectrometry of inter-α-inhibitor, a serine protease inhibitor that was found to be modified with two hexose sugars at C26.[38] S-linked glycosylation has since been observed on other intracellular proteins in mammals and on prokaryotic bacteriocins (antimicrobial polypeptides). [37,39]

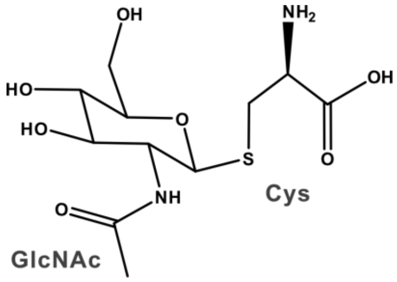


Figure 6: The structure of S-linked glycosylation. A single GlcNAc residue is covalently attached to a cysteine residue.

1.2. Lectins

Lectins are a class of proteins that selectively bind saccharides.[40] The term lectin does not include sugar-specific antibodies, sugar-transport proteins or enzymes.[40] Lectins were first discovered in 1888 by Russian scientist Stillmark, who identified the toxin Ricin from the caster bean plant *Ricinus communis*.[41] Today we know that plants, particularly legumes, are rich in lectins, with these proteins accounting for 5-10 % total protein content of legume seeds.[41] Lectins are, however, ubiquitous and found across all domains of life. The first animal lectins were identified in 1902 by American scientists Flexner and Noguchi, who observed agglutination of blood cells by snake venom.[42] It wasn't until 1980, however, that the first animal lectin, thrombolectin from *Bothrops atrox* (common lancehead snake), was isolated in pure form.[42]

Animal lectins are a large and diverse protein family that play a critical role in innate immunity, cell adhesion, protein trafficking and quality control.[43] Initially lectins were classified into five groups according to ligand specificity - for example, Galactose/*N*-Acetylgalactosamine (GalNAc) (61 % of animal lectins), Glucose/Mannose (14 %), *N*-Acetylglucosamine (GlcNAc) (12 %), Fucose (7 %) and sialic acid (6 %).[41] In 1988 Drickamer introduced a single letter naming system whereby lectins are instead classified based upon primary structure similarity (Table 1).[40]

Family	Example	Ligands	Cellular location	Function	Ref
C-type	Conglutinin	Various	Extracellular	Innate immunity	44,45
F-type	X-epilectin	Fucose	Extracellular	Innate immunity	46,47
H-type	HPA	GalNAc, galactose	Secreted	Innate immunity	48,49
I-type	CD22	Sialic acid	PM	Cell adhesion	50
L-type	CNX/CRT	Various	ER	Protein folding	51,52
M-type	OS-9	Mannose	ER	ERAD	46,53
P-type	CI-MPR	M6P	Secretory pathway	Protein trafficking	54
R-type	ppGalNAcTs	Various	Golgi, PM	Protein trafficking	46,55
S-type	Galectins	β-galactoside	Cytosol	Apoptosis	40
T-type	Tachylectins	GlcNAc, GalNAc	Hemocyte granules	Innate immunity	56
X-type	Interlectins	Furanoses	PM	Innate immunity	57
Pentraxins	CRP	Galactose	Cytosol	Innate immunity	58
CD 18	CR3	β-glucans	Plasma membrane	Cell adhesion	42,59

Table 1: Classification of animal lectins. Abbreviations used include: HPA - helix pomatia agglutinin, GalNAc - N-acetyl galactosamine, PM - plasma membrane, CNX/ CRT - calnexin/ calreticulin, ER - endoplasmic reticulum, ERAD - ER associated degradation, CI-MPR - Cation-independent mannose 6-phosphate receptor, M6P - mannose 6-phosphate, ppGalNAcTs - UDP-N-Acetylgalactosaminyltransferases, CRP - C-reactive protein.

Of lectins that have been structurally characterised (and deposited in the Glyco 3D database as of 2017) animal lectins represent ~33 % of structures, with plant lectins representing 24 %, followed by bacterial lectins (20%), viral lectins (14 %) and fungal lectins (9 %). The majority of these lectin structures exhibit predominantly β-strand secondary structure with approximately 50 % forming a β-sandwich fold (Figure 7A).[60] Other common folds include β-prisms (in plant lectins only), β-trefoils (in all kingdoms) and β-propellers (in animal, bacterial and fungal lectins) (Figure 7).[60]

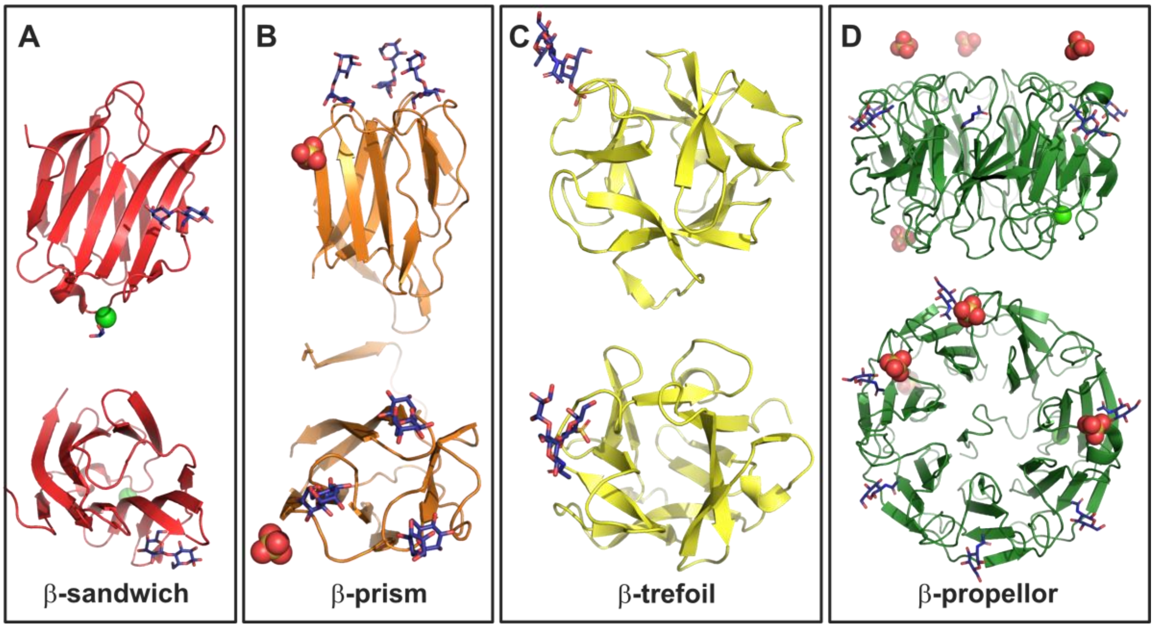


Figure 7: Structures of common lectin folds. A: β-sandwich fold of human galectin-3 which recognises β-galactoside sugars such as LacNAc. **B:** β-prism fold of the algal mannose lectin Griffithsin. **C:** β-trefoil fold of mouse mannose receptor, a C-type lectin. **D:** β-propeller fold of the *Psathyrella velutina* fungal lectin with six GlcNAc molecules bound. For each structure the protein is shown from the side (top) and top (bottom), glycans are shown as blue sticks, sulphate ions as red spheres and chloride ions as green spheres. PDB 2NN8, 2HYQ, 1FWV and 2C4D respectively.[61–64]

Different classes of lectins may contain similar carbohydrate binding domains (CBD). For example, the M-type and the P-type lectins, which recognise mannose and mannose 6-phosphate (M6P) respectively, both contain mannose 6-phosphate receptor homology (MRH) domains. Structural characterisation of the M-type and P-type lectins reveals a common flattened β-sandwich topology and a conserved 'QREY' motif that forms hydrogen bonds (H-bonds) to the sugar hydroxyl groups.[54,65–67]

The ER-resident, M-type lectins OS-9, XTP3-B and Glucosidase II (GII) play a critical role in glycoprotein quality control.[66] OS-9 and XTP3-B (otherwise known as Erlectin), which contain one or two MRH domains respectively, recognise the terminal α1-6 mannose residue on the C-arm of N-linked glycan precursors containing 5-7 mannose residues (and two GlcNAc residues) of misfolded glycoproteins.[24] This facilitates recruitment of the ubiquitin ligase SEL1L, targeting misfolded glycoproteins for ER associated degradation (ERAD).[24]

A soluble heterodimer, GII is composed of a catalytic α domain and regulatory β domain.[65] The catalytic domain recognises and removes the terminal glucose residue of the N-linked glycans $GlcNAc_2Man_9Glu_2$ and $GlcNAc_2Man_9Glu$, allowing release of the glycoprotein from the chaperones and L-type lectins CNX and CRT.[68] Meanwhile, the regulatory β domain contains an MRH domain that recognises a mannose residue in the B or C arm of the N-linked glycan.[65]

1.3. P-type lectins

Function of P-type lectins

Of the different classes of animal lectins, P-type lectins are unique in their ability to bind phosphorylated sugars, namely mannose 6-phosphate (M6P).[54] The P-type lectins recognise the phosphorylated terminal mannose residues of N-linked glycoproteins. The two members of the P-type lectin family, the cation-dependent mannose 6-phosphate receptor (CD-MPR) and the cation-independent mannose 6-phosphate receptor (CI-MPR), play a key role in intracellular protein trafficking and lysosome biogenesis. Together these P-type lectins are responsible for the trafficking of approximately 50 soluble lysosomal acid hydrolases from the trans Golgi network (TGN) to the late endosomes (LE) (Figure 8) and thus are critical for correct lysosome functioning.[54]

Traditionally referred to as the recycling centre of the cell, the lysosomes are responsible for the breakdown of proteins, lipids and polysaccharides.[69] However, through recruitment of mechanistic target of rapamycin complex 1 (mTORC1), the lysosomes also plays a role in metabolic signalling.[69] Recruitment of mTORC1, a phosphatidylinositol-3-kinase like Ser/Thr protein kinase, to the lysosomes allows it to regulate cell growth in response to nutrient levels.[69]

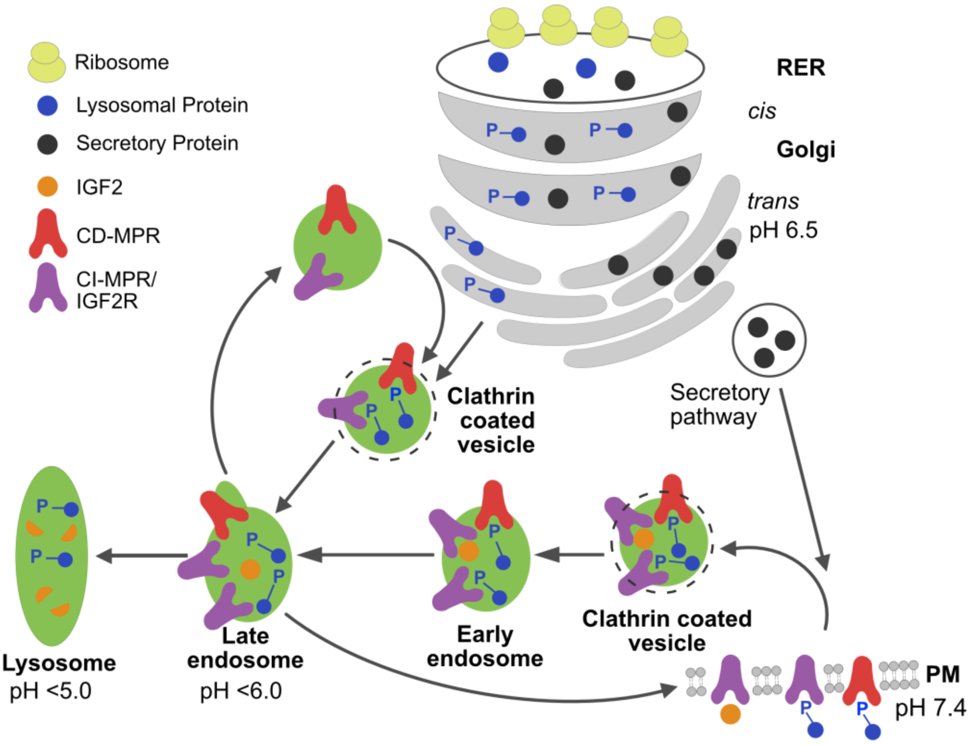


Figure 8: Localisation of the P-type lectins.[54,70] The CD-MPR (red) and CI-MPR (purple) transport M6P tagged lysosomal enzymes (blue) from the trans golgi network (TGN) to late endosomes (LE), which mature into the lysosomes. The CD-MPR and CI-MPR unload their cargo in the late endosomes and recycle to the TGN. When present in the plasma membrane (PM), the CD-MPR and CI-MPR may also import extracellular M6P tagged proteins. The CI-MPR can also bind the peptide hormone IGF2 (orange), facilitating its transport to the late endosomes for subsequent lysosomal degradation.

Although the dominant trafficker of lysosomal enzymes, the P-type lectins are not the only proteins involved in lysosome biogenesis. Sphingolipid activator proteins (SAPs) required for the degradation of glycosphingolipids from the plasma membrane are transported to the lysosome by the transmembrane glycoprotein sortilin independent of carbohydrate recognition.[71] Similarly, the transmembrane glycoprotein lysosomal integral membrane protein type 2 (LIMP2) has been demonstrated to transport β-glucocerebrosidase, a lysosomal enzyme that also catalyses the breakdown of sphingolipids, to the lysosome.[72,73] LIMP2 is not a lectin and this interaction between the proteins is instead mediated by a conserved coiled-coil of LIMP2.[72] Furthermore, trafficking lysosomal membrane proteins is also independent of the P-type lectins and carbohydrate recognition. Instead endocytosis sorting signals (e.g. YXXØ or

DEXXXLLI, whereby X is any amino acid and Ø a bulky, hydrophobic amino acid) are found at the C-terminus of the trafficked transmembrane lysosomal protein.[74]

As demonstrated in Figure 8, the CI-MPR is distinct from the CD-MPR in its ability to bind an array of ligands at neutral pH including non-glycosylated ligands such as insulin-like growth factor-2 (IGF2) and cellular repressor of E1A-stimulated genes (CREG) (see Section 1.5).[75–77]

Generation of the M6P recognition marker

Generation of the M6P recognition marker requires the concerted action of two key enzymes (Figure 9A). The first of these, GlcNAc-1-phosphotransferase, present in the Golgi apparatus, attaches an N-acetylglucosamine (GlcNAc) 1-phosphate residue onto the terminal mannose residue of a high mannose N-linked glycan precursor (Figure 9A).[78,79] GlcNAc phosphotransferase is a hetero-hexameric protein composed of the subunits $\alpha_2\beta_2\gamma_2$.[80] The catalytic α and β subunits may recognise 2-3 lysine residues ~34 Å apart that form a critical recognition patch for GlcNAc-phosphotransferase.[80] Meanwhile, the regulatory γ subunits enhance recruitment to the N-linked glycan through recognition of mannose residues by an MRH domain.[81]

The action of GlcNAc-phosphotransferase creates an M6P-GlcNAc di-ester (Figure 9A).[78] In the TGN a second enzyme, α-*N*-acetylglucosaminidase (uncovering enzyme (UCE)), continues the processing by removing the GlcNAc to expose the M6P mono-ester (Figure 9A).[82,83]

The N-glycan precursor can by phosphorylated at several mannose residues (Figure 9B). Furthermore, the actions of GlcNAc-1-phosphotransferase and UCE can produce mono and di-phosphorylated glycoproteins. 85 % of mono-phosphorylated glycoproteins, which contain 6-8 mannose residues, are phosphorylated on their α1-6 branch (B and C arms).[84] Di-phosphorylated glycoproteins, which contain no more than nine mannose residues, may be phosphorylated on both the α1-6 and α1-3 branch (A, B and C arms).[84]

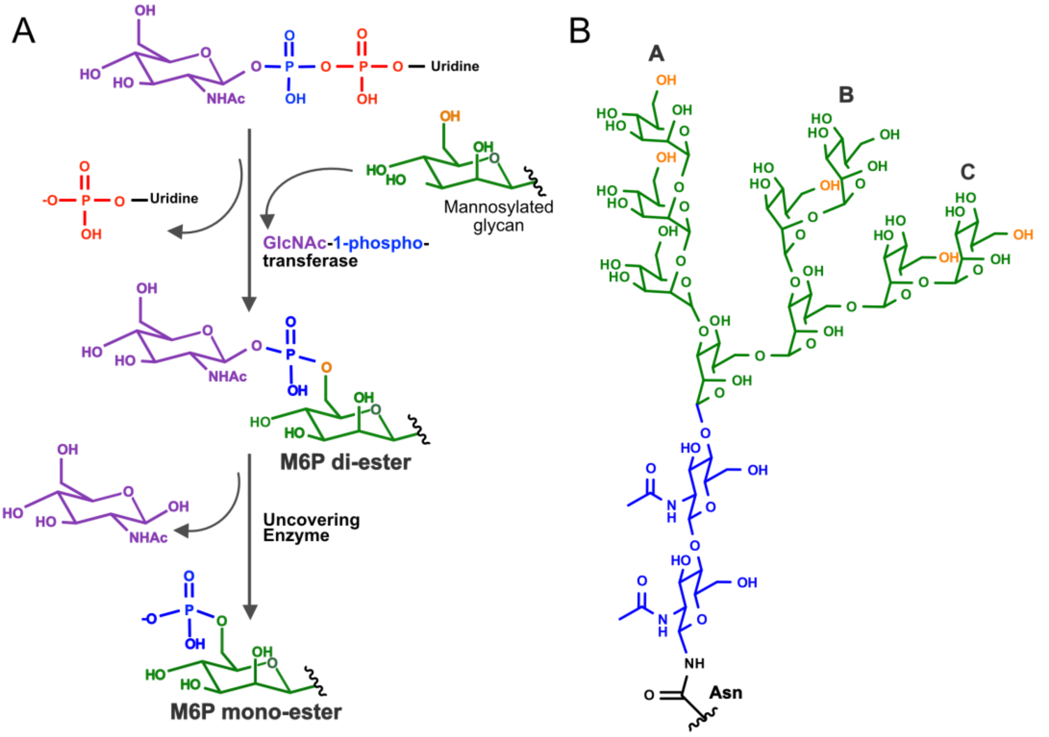


Figure 9: Generation of the mannose 6-phosphate (M6P) recognition marker. A: Generation of the M6P recognition marker occurs in two enzymatic steps. Firstly, GlcNAc-1-phosphotransferase attaches a GlcNAc-1-phosphate residue to the terminal mannose residue of an N-linked glycan on the lysosomal enzyme. This creates an M6P di-ester. Uncovering enzyme (UCE) removes the terminal GlcNAc to expose M6P mono-ester. **B:** The structure of the N-linked glycan precursor following transfer onto asparagine and removal of glucose residues. GlcNAc residues are coloured blue, mannose residues green and hydroxyl groups of mannose residues that can be phosphorylated are coloured orange.

1.4. CD-MPR structure and function

The CD-MPR recognises M6P mono-esters.[54] While it constitutively cycles between the TGN and lysosome, the CD-MPR functions to transport M6P tagged lysosomal enzymes to the late endosome (Figure 8).[85,86]

The CD-MPR, which is encoded by an 8 exon, 12 kb gene located on chromosome 12p, is a ~46 kDa, type I transmembrane glycoprotein, consisting of a cytoplasmic, carboxy-terminal domain (67 residues), a transmembrane domain (25 residues) and an extracellular amino-terminal domain (159 residues).[54] Crystal structures of bovine CD-MPR extracellular region have been determined in the absence of ligand, with the monosaccharide M6P bound (K_D 8 μM) and with the oligosaccharides trimannosyl phosphate and pentamannosyl phosphate bound (PDB 1KEO, 2RL8, 2RL9 and 1C39 respectively).[87–90] Each structure revealed a flattened nine stranded β-sandwich (β-strands A-I) with an N-terminal α-helix.[90] Four loops (termed the AB, CD, FG and HI loops) connect these β-strands at the top of the sandwich to form the ligand binding site.[90] There are four disulphide bonds between C6-C52 at the N-terminus, C106-C141 at the top of the FG and HI loops respectively, and C119-C153 at the base of βG and βI respectively (Figure 10A).

In each structure the CD-MPR formed a non-covalent homodimer (Figure 10B), with the βE-I surfaces of both monomers packing against one another. Chemical cross-linking, size exclusion chromatography and sucrose density centrifugation have demonstrated that the extracellular region of the CD-MPR is capable of dimerising in solution.[91] Further studies have found the CD-MPR to exist in monomeric, dimeric and tetrameric form in solution and dimeric form in membranes.[92] The oligomeric state of the CD-MPR appears to be concentration, pH and temperature sensitive, with lower concentration, lower pH and higher temperature each promoting CD-MPR monomerisation.[93]

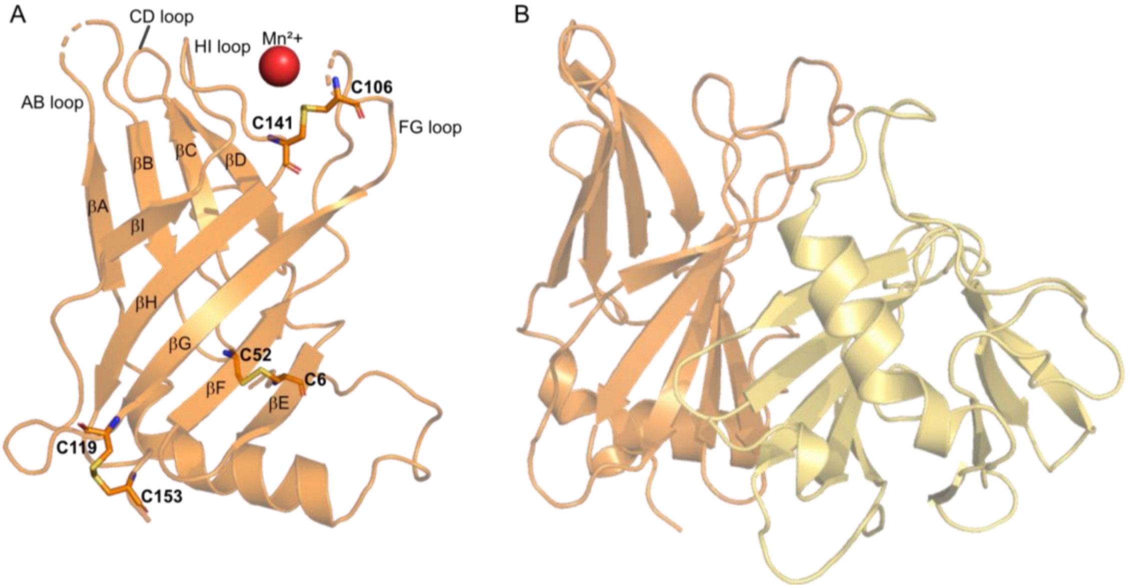


Figure 10: The structure of P-type lectin CD-MPR extracellular region. A: The crystal structure of bovine CD-MPR in the absence of ligand (PDB 1KEO) reveals that each monomer forms a nine stranded β-sandwich (βA-I) connected by flexible loops and three disulphide bonds (sticks). Mn^{2+}, not visible in the crystal structure, is modelled in (red sphere). **B:** The arrangement of the CD-MPR homodimer.

In comparison to other lectins, the CD-MPR has a large, deep binding site that buries ~65 % of bound M6P and both the M6P and penultimate mannose residue of bound pentamannosyl phosphate.[12, 24] X-ray crystallography reveals that the M6P residue is positioned in the same orientation in the structures of CD-MPR bound with M6P monosaccharide, trimannosyl phosphate and pentamannosyl phosphate (PDB 2RL8, 2RL9 and 1C39 respectively).

Residues Q66 (βC), R111 (βF), E133 (βH) and Y143 (βI) form the conserved 'QREY' motif characteristic of MRH domains.[54,67] Q66, Y143 and E133 are within hydrogen bonding distance of the 2' and 3' OH of M6P.[87] E133 may also form hydrogen bonds (H-bonds) to the 4' OH.[87] Additional binding site residues include Y45 (βB), R135 (HI loop) and H105 (FG loop). Y45 forms H-bonds to the 1'OH of M6P when free as a monosaccharide and when oxidised in an O-glycosidic bond.[87] R135 may form H-bonds with the 4'OH of M6P.[87] Lastly, H105 forms electrostatic interactions with the phosphate group of M6P (Figure 11A).[87]

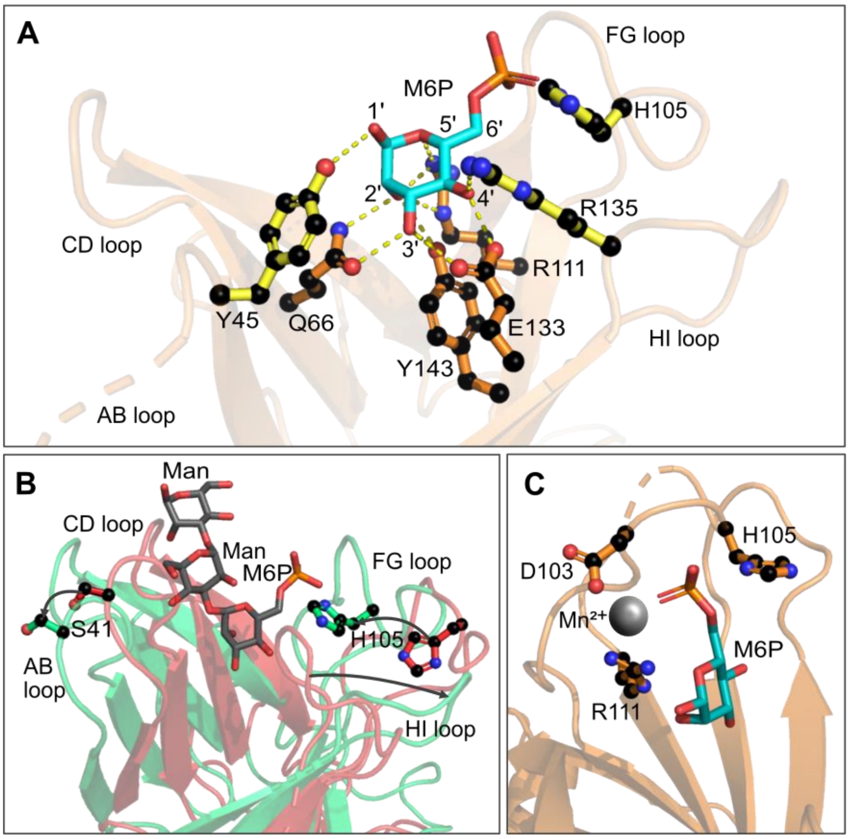


Figure 11: M6P binding by the CD-MPR extracellular region. A: The crystal structure of the extracellular region of bovine CD-MPR with M6P bound (PDB 2RL8) reveals the placement of the conserved 'QREY' motif (Q66, R111, E133, Y143, shown as orange balls-and-sticks). Additional M6P binding residues (Y45, R135, H105) are shown as yellow balls-and-sticks. M6P is shown as cyan sticks and possible H-bonds are shown as yellow dashed lines. **B:** Superimposition of the crystal structures of CD-MPR unbound (red) and with pentamannosyl phosphate bound (green, PDB 1KEO and 1C39 respectively) reveals large conformational changes upon ligand binding. Particularly, H105 on the FG loop and S41 on the AB loop move 2.9 Å and 4.1 Å respectively. **C:** The divalent cation Mn^{2+} binds at the edge of the M6P binding site and shields bound M6P from the acidic residue D103 on the FG loop (PDB 2RL8).

Upon ligand binding, the CD-MPR loops undergoes a large conformational change (Figure 11B).[90] S41 of the AB loop is displaced 4.1 Å and H105 on the FG loop 2.9 Å when pentamannosyl phosphate is bound.[90] Additionally, a flap comprising residues E134-C141 of the HI loop and E133 and F142 of βH and βI respectively, moves towards the binding pocket in the absence of ligand.[90]

Residues R111 (βF), D103 and H105 (FG loop) co-ordinate the Mn^{2+} cation (Figure 11C).[87] It is proposed that, when present, the Mn^{2+} cation will shield the phosphate group of M6P from the anion D103.[54] However, despite its name, the human CD-MPR is not strictly dependent upon the presence of the divalent cation Mn^{2+} in the binding site.[94] The CD-MPR retained the ability to bind phosphomannan-sepharose following stripping of the metal ion by EDTA.[94]

1.5. CI-MPR structure and function

The CI-MPR, which is encoded by a 48 exon, 136 kb gene located on human chromosome 6q, is a large (~300 kDa), type I transmembrane glycoprotein consisting of a cytoplasmic carboxyl-terminal domain (167 residues), a transmembrane domain (23 residues) and an extracellular amino-terminal domain (2269 residues).[87, 95-96]

Although the intracellular region of the CI-MPR lacks kinase activity, suggesting it is unable to signal, it does have several ligands critical for receptor endocytosis and trafficking (Table 2).[97] For packaging into clathrin coated vesicles that are secreted to the plasma membrane, adaptor protein-1 (AP1) interacts with the sequence TEWLI in the C-terminal region of the CI-MPR.[98, 74] In a similar manner, AP2 recognises the internalisation sequence YKYSKV of the CI-MPR C-terminus, facilitating receptor endocytosis from the plasma membrane to early endosome (EE).[54] The CI-MPR is also transported directly from the TGN to EE in clathrin-coated vesicles by Golgi-localized γ-ear-containing ADP-ribosylation factor binding protein (GGA).[54] The VHS domain of GGA recognises the di-leucine motif and central aspartic acid residue in the C-terminal TGN sorting signal sequence FHDDSDEDLL.[99]

Following cargo release in the late endosome, the CI-MPR is recycled back to the TGN by retrograde transport machinery, including TIP47, AP1 and PACS1.[54] TIP47 (Tail interacting protein of 47 kDa) recognises residues 48-75 of the CI-MPR C-terminal region and recruits Rab9 GTPase to facilitate retrograde transport.[54] Meanwhile, PACS1 (Phosphofurin acidic cluster sorting) recognises the di-leucine motif DDS (of the TGN sorting signal sequence FHDDSDEDLL) and recruits AP1.[54]

Interestingly, receptor trafficking is independent of extracellular ligand binding.[98] Instead, CI-MPR transport may be regulated by phosphorylation of these transport related recognition motifs in the intracellular region of the receptor.[54]

Intracellular ligand	Recognition site	Trafficking	Ref
AP1	TEWLI	TGN-PM	74
AP2	YKYSKV	PM-EE	54
GGA	FHDDSDEDLL	TGN-EE	99
TIP47	residues 48-75	LE-TGN	54
PACS1	DDS	LE-TGN	54
Retromer	WLM	LE-TGN	100

Table 2: Ligands that recognise the intracellular region of the CI-MPR to facilitate receptor trafficking. Abbreviations: TGN-trans golgi network, PM-plasma membrane, EE-early endosome, LE-late endosome.

The extracellular region of the CI-MPR is comprised of 15 homologous domains (each 124-192 amino acids) that have between 14-28 % sequence identity to the extracellular domain of the CD-MPR.[96] Figure 12 shows a comparison between the tertiary structure of the CD-MPR extracellular region and CI-MPR domain 11, the first CI-MPR domain to be structurally characterised. Similarly to the CD-MPR extracellular domain, each of the 15 CI-MPR extracellular domains forms a flattened, nine stranded, β-sandwich structure (Figure 12) with β-strands A-D and E-I forming two crossed β-sheets.[87,101] Each CI-MPR extracellular domain contains three or four conserved disulphide bonds.

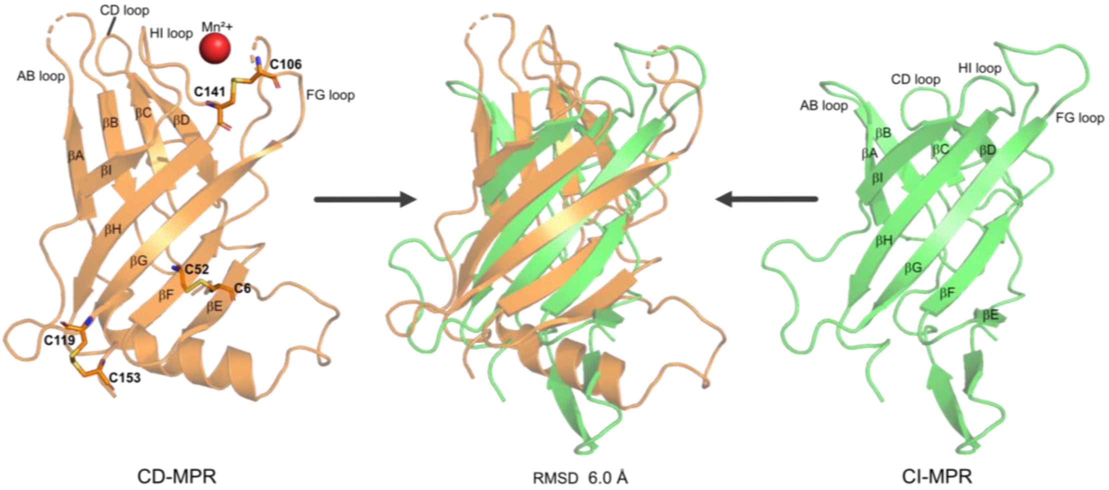


Figure 12: Structures of the P-type lectins. The extracellular regions of the CD-MPR (orange) and CI-MPR (green) exhibit high structural similarity - superimposition of the CD-MPR extracellular region and domain 11 of the CI-MPR gives an RMSD value of 6.0 Å over all backbone atoms (PDB 1KE0 and 1GP0 respectively).

The CI-MPR is functionally distinct to the CD-MPR, firstly, in its ability to bind both M6P mono-esters and di-esters, and, secondly, in its affinity for non-glycosylated ligands (Table 3).[54] In fact, the CI-MPR is remarkable in its ability to bind three distinct classes of ligand: protein (e.g. IGF2), small molecules (e.g. M6P) and lipids (e.g. retinoic acid).

Extracellular ligand	Consequence of binding	Ref
Lysosomal enzymes	Intracellular trafficking and lysosome biogenesis.	54
Granzyme B	Cytotoxic T cell mediated apoptosis.	102
Latent TGFβ	TGFβ activation.	103,104
LIF	Endocytosis and lysosomal degradation.	105
Prorenin	Activation to renin.	106
Proliferin	Angiogenesis.	105,107
CD26	T cell activation.	108
Thyroglobulin	Endocytosis.	109
EGFR	Endocytosis and lysosomal degradation	110
Herpes simplex viral glycoprotein D	Viral entry to cells.	111
CREG	Endocytosis and lysosomal degradation.	77
IGF2	Endocytosis and lysosomal degradation.	112
Retinoic acid	Suppression of cell growth and induction of apoptosis.	113
uPAR	Activation of TGFβ and plasminogen.	114,115
Plasminogen	Plasminogen activation, facilitating cell migration.	114,115
CD45	Regulation of T cell signalling.	115,116
Lck	T cell signalling.	115,116
Heparanase	Extracellular matrix degradation.	117

Table 3: Example known ligands of the CI-MPR extracellular region. M6P dependent ligands are highlighted in green, M6P independent ligands in blue and ligands that utilise M6P dependent and independent mechanisms in orange. Abbreviations: TGFβ -Transforming Growth Factor β, LIF-Leukaemia inhibitory factor, EGFR-Epidermal growth factor receptor, CREG-Cellular repressor of E1A-stimulated genes, IGF2-Insulin-like growth factor-2, uPAR-Urokinase-type plasminogen activator receptor.

IGF2 binding

The most well characterised CI-MPR ligand interaction is that with the peptide hormone insulin-like growth factor-2 (IGF2). Although expression of IGF2, which is encoded by a nine exon, 30 kb gene located on human chromosome 11p, is highest during embryogenesis, IGF2 is also present in adult liver and brain tissues.[118,119] IGF2, which has several post-translational isoforms, is produced as a prepro-peptide (180 residues).[118] The N-terminal signal sequence is cleaved to produce pro-IGF2 (156 residues, also termed IGF2^{156}).[118] Pro-IGF2 is O-link glycosylated, which signals further cleavage by pro-hormone convertase 4 (PC4) to produce mature IGF2^{67} (67 residues).[120,121] Incomplete processing of pro-IGF2 results in big IGF2 isoforms containing 104 and 87 residues (IGF2^{104} and IGF2^{87} respectively).[120,121] These big IGF2 isoforms make up approximately 10-20 % of IGF2 in circulation.[118]

At the plasma membrane mature IGF2^{67} signals cellular growth and proliferation through binding the receptor tyrosine kinases IGF1R, IR-A (an alternatively spliced isoform of the insulin receptor) or the hybrid receptor IGF1R-IR-A (Figure 13).[122] Through binding to IGF1R, IGF2 also plays a role in metabolic signaling by regulating glucose uptake.[123] Circulating IGF2 is sequestered by a family of six IGF binding proteins (IGFBPs).[123] IGF2 is also sequestered in circulation and at the plasma membrane by the CI-MPR, which is also known as the insulin-like growth factor 2 receptor (IGF2R).[124] The CI-MPR/ IGF2R facilitates IGF2 endocytosis and subsequent lysosomal degradation.[124] Early CI-MPR/ IGF2R knock-out experiments in mice resulted in an increase of IGF2, embryos that were 35-40 % larger than average due to hyperplasia (an increase in cell number) and death at birth.[125,126]

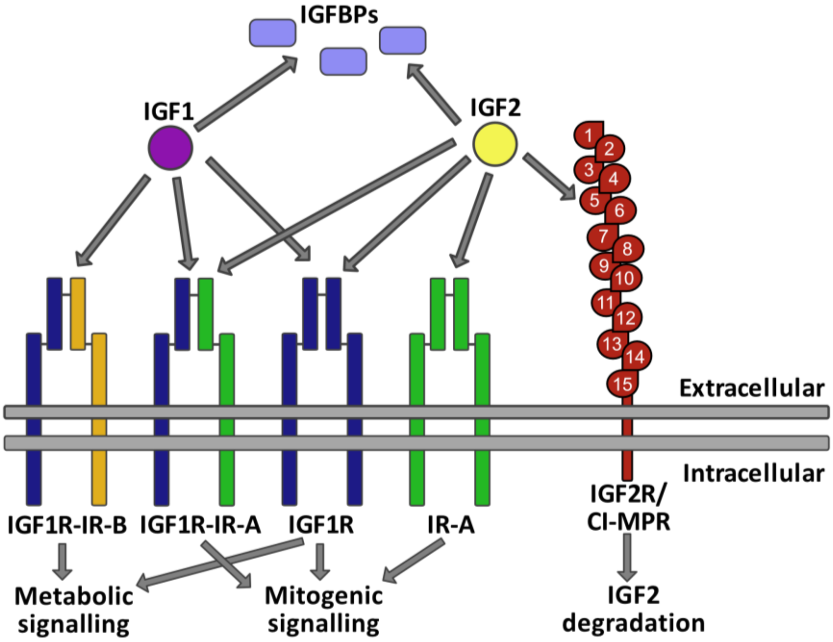


Figure 13: The Insulin-like growth factor (IGF) network. At the plasma membrane the IGFs signal growth via binding the receptor tyrosine kinases IGF1R, IGF1R-IR-A and, in the case of IGF2, IR-A. The IGFs signal metabolically through binding IGF1R and, in the case of IGF1, IGF1R-IR-B. The IGFs are sequestered by circulating IGF binding proteins (IGFBPs) in serum. IGF2 is also sequestered by the IGF2R/ CI-MPR, which facilitates IGF2 endocytosis and subsequent lysosomal degradation.[118] The extracellular region of the IGF2R/ CI-MPR consists of 15 domains, with domain 1 (D1) at the N-terminus, furthest from the plasma membrane and D15 closest to the plasma membrane. IGF2 is bound at D11.

IGF2 binds the CI-MPR/ IGF2R at extracellular domain 11 (D11). D11 contains a hydrophobic pocket formed from residues of the AB, CD, and FG loops (Figure 14A) that binds IGF2 with high-affinity (K_D 40-60 nM).[112,127] A combination of mutagenesis, crystallography and NMR studies has identified the interactions at the D11-IGF2 interface.[75,101,124] A group of 9 hydrophobic D11 residues interact with IGF2: V1574 (CD loop), L1626 (HI loop) and L1636 (HI loop) form a three-pronged interaction with IGF2, F1567 (CD loop), L1629 (HI loop) and Y1542 (AB loop) interact with IGF2 F19, Y1606 (FG loop) and I1572 (CD loop) interact with IGF2 T16, and K1631 (HI loop) interacts with IGF2 L53 (Figure 14A).[124] Electrostatic interactions are also involved in IGF2 binding, with the binding interface of IGF2 having a net negative charge and that of IGF2R a net positive charge.[101]

Residues of D13, which is the only domain to contain a fibronectin type II (FNII) insert (Figure 15), stabilise the D11 AB loop.[101] R1931 of the D13 FNII insert forms electrostatic interactions with E1553 of D11 AB loop (Figure 14D). Deletion of the FNII insert results in a 10-fold decrease in the rate of association (k_A).[101]

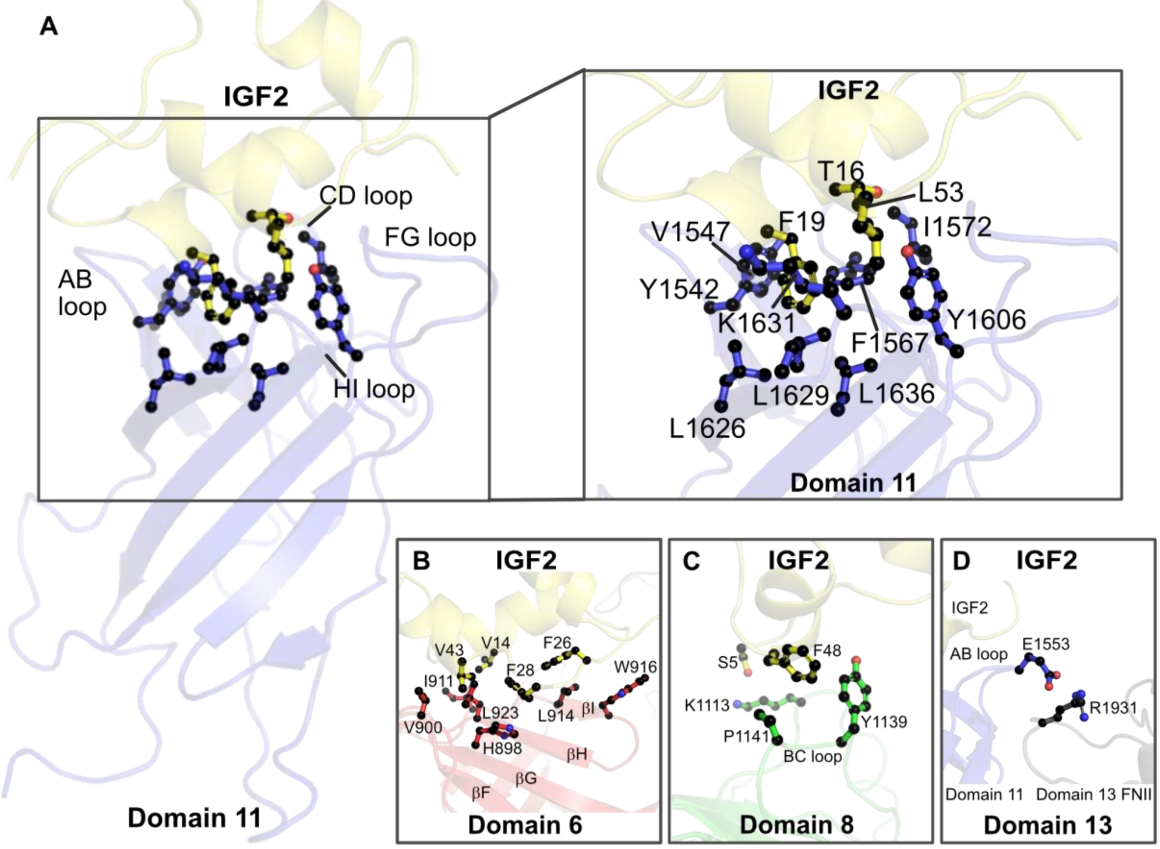


Figure 14: The extracellular region of the CI-MPR binds the growth hormone IGF2, facilitating its internalisation and subsequent degradation in the lysosome. A: The loops of D11 form the core IGF2 binding site. Interacting residues (T16, F19 and L53 of IGF2 and Y1542, V1547, F1567, I1572, Y1606, L1626, L1629, K1631, L1636 of D11) are shown as balls-and-sticks (PDB 2L29). **B:** The cryo-EM Structure of bovine D4-14 in complex with IGF2 reveals that residues on the βF-I surface of D6 (H898, V900, I911, L914, W916, L923) also interact with IGF2 (V14, F26, F28, V43) (PDB 6UM2). **C:** K1113 at the N-terminus of D8 and T1139 and P1141 of D8 BC loop interact with S5 and F48 of IGF2 (PDB 6UM2). **D:** A fibronectin type II insert in D13 stabilises the AB loop of D11, increasing its affinity for IGF2. A salt bridge forms between E1553 (D11) and R1931 (D13) (PDB 2V5P).

Recently, a low-resolution (4.3 Å) cryo-EM structure of D4-14 plus IGF2 (PDB 6UM2) has revealed that D6 and D8 also interact with bound IGF2.[128] The βI-F surface of D6 interacts with helices α1 and α2 of IGF2 (Figure 14B), while K1113 N-terminal to βA and Y1139 and P1141 both on the BC loop of D8 interact with α2 of IGF2 (Figure 14C). These interactions were confirmed by mutagenesis, with IGF2 mutations F28A, V43D (which both interact with D6) and F48D (which interacts with D8) each reducing CI-MPR-IGF2 binding 20-30 %.[128] In comparison, mutation of IGF2 residue L53, which interacts with D11, reduced binding by 60 %.[128]

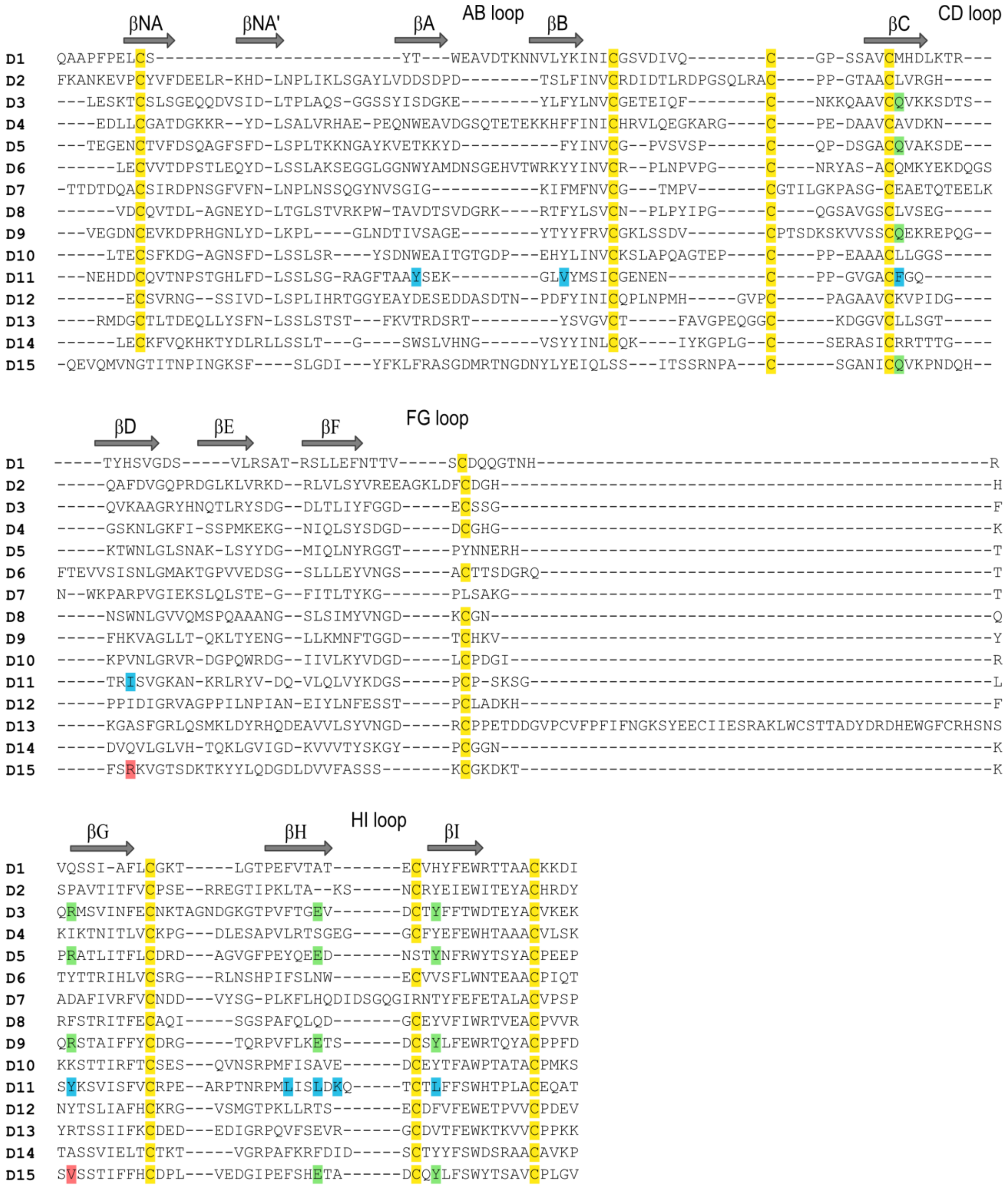
Despite the high CI-MPR amino acid sequence identity between species and animal classes (for example, *Bos taurus*, *Gallus gallus* and *Mus musculus* have 90 %, 60 % and 89 % sequence identity to *Homo sapiens* CI-MPR respectively), only the CI-MPR/ IGF2R of mammals (monotremes, marsupials and placentals) is capable of IGF2 binding.[97,124] The CI-MPR/ IGF2R of reptiles and birds does not bind IGF2.[124] Comparison of the IGF2 binding sites of D11 of different species reveals that the volume of the binding pocket increases and changes from a net negative charge to a net positive charge as IGF2 binding ability is acquired.[124] An exon splicing event whereby exon 34, which encodes the CD loop of D11, was mutated, is responsible for the IGF2 binding ability in D11.[124] Additionally, sequence alignment reveals that the residues I911, L914 and W916 of bovine D6 and Y1139, I1140 of bovine D8 which interact with IGF2 are conserved in mammals but are not found in birds and fish, which lack IGF2 binding capability (Appendix Figure 2).

Carbohydrate binding

While the CI-MPR and the CD-MPR both possess the ability to bind M6P, the CI-MPR exhibits more efficient binding to M6P-tagged lysosomal enzymes *in vitro* and *in vivo*.[82] This may be attributed to the presence of multiple MRH domains in the CI-MPR: D3, D5, D9 and D15 (Table 4). These MRH domains differ in their glycan specificity, affinity and dependency on neighbouring domains. Whilst the core of these domains retains the nine stranded β-sandwich structure shared with CI-MPR D11, the binding site residues vary. Sequence analysis, mutagenesis studies and structural biology have identified key sugar binding residues in these domains.[67,88,129–131] Most notably, four residues are conserved in the same orientations across each of these domains: Q, R, E, Y (Figure 15).

MRH domain	Ligands	Ref
CI-MPR D3	M6P monoesters, M6P methyl esters and mannose 6-sulphate.	132
CI-MPR D5	M6P di-esters.	131
CI-MPR D9	M6P mono-esters.	103
CI-MPR D15	M6P mono-esters and di-esters.	130
CD-MPR	M6P mono-esters.	133
Glucosidase II	Mannose.	65
OS-9	Mannose.	66
XTP3-B/Erlectin	Mannose.	134
GlcNAc-1-phosphotransferase	Mannose.	81

Table 4: Example lectins containing MRH domains and their ligands



```
D1   QAAPFPELCS-------------------------YT---WEAVDTKNNVLYKINICGSVDIVQ--------C----GP-SSAVCMHDLKTR---
D2   FKANKEVPCYVFDEELR-KHD-LNPLIKLSGAYLVDDSDPD--------TSLFINVCRDIDTLRDPGSQLRAC----PP-GTAACLVRGH-----
D3   ---LESKTCSLSGEQQDVSID-LTPLAQS-GGSSYISDGKE--------YLFYLNVCGETEIQF--------C----NKKQAAVCQVKKSDTS--
D4   ----EDLLCGATDGKKR--YD-LSALVRHAE-PEQNWEAVDGSQTETEKKHFFINICHRVLQEGKARG----C----PE-DAAVCAVDKN-----
D5   ---TEGENCTVFDSQAGFSFD-LSPLTKKNGAYKVETKKYD----------FYINVCG--PVSVSP------C----QP-DSGACQVAKSDE---
D6   ------LECVVTDPSTLEQYD-LSSLAKSEGGLGGNWYAMDNSGEHVTWRKYYINVCR--PLNPVPG-----C----NRYAS-ACQMKYEKDQGS
D7   -TTDTDQACSIRDPNSGFVFN-LNPLNSSQGYNVSGIG-----------KIFMFNVCG---TMPV-------CGTILGKPASG-CEAETQTEELK
D8   ------VDCQVTDL-AGNEYD-LTGLSTVRKPW-TAVDTSVDGRK----RTFYLSVCN--PLPYIPG-----C----QGSAVGSCLVSEG-----
D9   ---VEGDNCEVKDPRHGNLYD-LKPL----GLNDTIVSAGE--------YTYYFRVCGKLSSDV--------CPTSDKSKVVSSCQEKREPQG--
D10  -----LTECSFKDG-AGNSFD-LSSLSR----YSDNWEAITGTGDP---EHYLINVCKSLAPQAGTEP----C----PP-EAAACLLGGS-----
D11  ---NEHDDCQVTNPSTGHLFD-LSSLSG-RAGFTAAYSEK---------GLVYMSICGENEN----------C----PP-GVGACFGQ-------
D12  -------ECSVRNG--SSIVD-LSPLIHRTGGYEAYDESEDDASDTN--PDFYINICQPLNPMH-----GVPC-----PAGAAVCKVPIDG----
D13  ----RMDGCTLTDEQLLYSFN-LSSLSTST---FKVTRDSRT---------YSVGVCT-----FAVGPEQGGC------KDGGVCLLSGT-----
D14  ------LECKFVQKHKTYDLRLLSSLT---G----SWSLVHNG------VSYYINLCQK----IYKGPLG--C-----SERASICRRTTTG----
D15  -QEVQMVNGTITNPINGKSF----SLGDI---YFKLFRASGDMRTNGDNYLYEIQLSS---ITSSRNPA---C------SGANICQVKPNDQH--

D1   -----TYHSVGDS-----VLRSAT-RSLLEFNTTV-----SCDQQGTNH----------------------------------------------R
D2   -----QAFDVGQPRDGLKLVRKD--RLVLSYVREEAGKLDFCDGH--------------------------------------------------H
D3   -----QVKAAGRYHNQTLRYSDG--DLTLIYFGGD-----ECSSG--------------------------------------------------F
D4   -----GSKNLGKFI-SSPMKEKG--NIQLSYSDGD-----DCGHG--------------------------------------------------K
D5   -----KTWNLGLSNAK-LSYYDG--MIQLNYRGGT-----PYNNERH------------------------------------------------T
D6   FTEVVSISNLGMAKTGPVVEDSG--SLLLEYVNGS-----ACTTSDGRQ----------------------------------------------T
D7   N--WKPARPVGIEKSLQLSTE-G--FITLTYKG-------PLSAKG-------------------------------------------------T
D8   -----NSWNLGVVQMSPQAAANG--SLSIMYVNGD-----KCGN---------------------------------------------------Q
D9   -----FHKVAGLLT-QKLTYENG--LLKMNFTGGD-----TCHKV--------------------------------------------------Y
D10  -----KPVNLGRVR-DGPQWRDG--IIVLKYVDGD-----LCPDGI-------------------------------------------------R
D11  -----TRISVGKAN-KRLRYV-DQ-VLQLVYKDGS-----PCP-SKSG-----------------------------------------------L
D12  -----PPIDIGRVAGPPILNPIAN-EIYLNFESST-----PCLADKH------------------------------------------------F
D13  -----KGASFGRLQSMKLDYRHQDEAVVLSYVNGD-----RCPPETDDGVPCVFPFIFNGKSYEECIIESRAKLWCSTTADYDRDHEWGFCRHSNS
D14  -----DVQVLGLVH-TQKLGVIGD-KVVVTYSKGY-----PCGGN--------------------------------------------------K
D15  -----FSRKVGTSDKTKYYLQDGDLDVVFASSS-------KCGKDKT------------------------------------------------K

D1   VQSSI-AFLCGKT-----LGTPEFVTAT-------ECVHYFEWRTTAACKKDI
D2   SPAVTITFVCPSE--RREGTIPKLTA--KS-----NCRYEIEWITEYACHRDY
D3   QRMSVINFECNKTAGNDGKGTPVFTGEV-------DCTYFFTWDTEYACVKEK
D4   KIKTNITLVCKPG---DLESAPVLRTSGEG-----GCFYEFEWHTAAACVLSK
D5   PRATLITFLCDRD---AGVGFPEYQEED-------NSTYNFRWYTSYACPEEP
D6   TYTTRIHLVCSRG---RLNSHPIFSLNW-------ECVVSFLWNTEAACPIQT
D7   ADAFIVRFVCNDD---VYSG-PLKFLHQDIDSGQGIRNTYFEFETALACVPSP
D8   RFSTRITFECAQI-----SGSPAFQLQD-------GCEYVFIWRTVEACPVVR
D9   QRSTAIFFYCDRG-----TQRPVFLKETS------DCSYLFEWRTQYACPPFD
D10  KKSTTIRFTCSES---QVNSRPMFISAVE------DCEYTFAWPTATACPMKS
D11  SYKSVISFVCRPE--ARPTNRPMLISLDKQ-----TCTLFFSWHTPLACEQAT
D12  NYTSLIAFHCKRG---VSMGTPKLLRTS-------ECDFVFEWETPVVCPDEV
D13  YRTSSIIFKCDED---EDIGRPQVFSEVR------GCDVTFEWKTKVVCPPKK
D14  TASSVIELTCTKT-----VGRPAFKRFDID-----SCTYYFSWDSRAACAVKP
D15  SVSSTIFFHCDPL---VEDGIPEFSHETA------DCQYLFSWYTSAVCPLGV
```

Figure 15: Structure-based sequence alignment of human CI-MPR extracellular domains. The conserved 'QREY' residues of MRH domains (D3, D5, D9 and D15) are coloured green. In D15 the conserved arginine residue is replaced by R2170 and V2205 (coloured red). IGF binding residues of D11 are coloured blue. Cysteines are coloured yellow. Each domain contains four disulphide bonds, except for D5 and D7 which both contain three disulphide bonds. The position of the β-strands and binding loops are labelled.

The CI-MPR, which contains four MRH domains (in D3, D5, D9 and D15), exhibits a similar binding affinity for the lysosomal enzyme acid alpha-glucosidase (GAA) tagged with M6P mono-ester and GlcNAc-M6P di-ester with K_D values of 4.5 ± 0.7 nM and 51 ± 1 nM respectively.[130] Glycan micro-array analysis reveals that the CI-MPR does not interact with non-phosphorylated high mannose glycans.[82] Kinetic studies have shown that the CI-MPR has a higher affinity for oligosaccharide glycans over monosaccharides. For example, the CI-MPR binds M6P monosaccharide, pentamannose phosphate and the glycoprotein β-galactosidase with binding affinities (K_D) of 7 μM, 6 μM and 0.02 μM respectively.[135]

Domain 3

The structure of bovine CI-MPR D1-3 with M6P bound has been determined by X-ray crystallography (PDB 1SYO/1SZ0).[136] Each domain formed the conserved β-sandwich structure observed for D11. Similarly, to the CD-MPR structure, the crystal structure of the N-terminal region of the CI-MPR (D1-3) revealed a homodimer.[136] However, while dimerisation of D1-3 has been observed in crystals belonging to two different space groups (orthorhombic $P2_12_12_1$ and monoclinic $P2_1$), the oligomeric state of D1-3 has not been demonstrated in solution.[103,136] Thus, it is not known if dimerisation is relevant in solution or simply a crystal artefact.

Bovine residues Q348 (βC), R391 (βG), E416 (βH) and Y421 (βI) form the conserved 'QREY' motif characteristic of MRH domains (Figure 16). Q348, Y421 and E416 are within hydrogen bonding distance of the 2' and 3' OH of M6P.[136] E416 may also form H-bonds to the 4' OH.[136] Y324 (βB) may form H-bonds to the 1' OH of M6P.[136] Three bridging waters in the binding site further contribute to this network. Water 1 (W1) interacts with a phosphate oxygen, S386 and S387 (both on the FG loop).[136] The second and third water molecules (W2 and W3) form a hydrogen bond network - W2 forms H-bonds to W1 and W3, whilst W3 forms H-bonds to W2, S387 and D418 (βH) (Figure 16).[136]

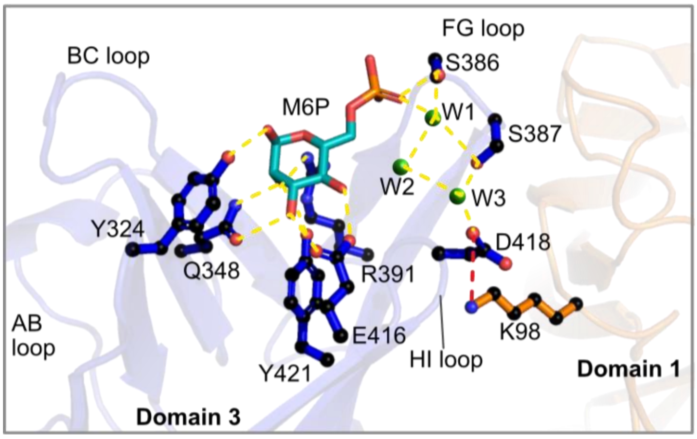


Figure 16: The M6P binding site of bovine CI-MPR D3. The crystal structure of bovine D1-3 M6P bound (PDB 1SYO/1SZ0) reveals the orientation of key M6P binding residues Q348, R391, E416, Y421, S386, S387 and Y324. H-bonds between these residues (balls-and-sticks) and M6P (cyan sticks) are shown as yellow lines. The salt bridge formed between D418 on the HI loop of D3 and K98 at the base of D1 is shown as a red line. The three bridging water molecules (W1, W2 and W3) are shown as green spheres.

Comparable to IGF2s interaction with D11 and D13, the binding of M6P to D3 is enhanced when expressed in a multi-domain construct of D1-3.[82] For example, D1-3 exhibited a binding affinity of 0.5 ± 0.1 nM to M6P-tagged β-glucuronidase, while D3 alone gave a value of ~500 nM.[67,137] The crystal structure of bovine D1-3 shows that these neighbouring domains stabilise the D3 binding loops, rather than interacting directly with the ligand.[82] For example, K98 in the GH loop at the base of D1 forms a salt bridge with D418 in the HI loop of D3.[136]

Recently the first structure of human CI-MPR D3 has been determined within the multi-domain construct D1-5 (PDB 6P8I and 6V02).[138] These are also the first structures of the D3 M6P binding site unoccupied by a ligand. Superimposition of human D1-3 at pH 5.5 and 7.0 with bovine D1-3 plus M6P at pH 6.35 (PDB 6P8I, 6V02 and 1SZ0 respectively) gave RMSD values of 3.8 Å and 3.5 Å over backbone atoms. The absence of ligand alters the domain arrangement of D1-3, with βE-I of D3 rearranging to pack against βE-I of D1.

D3 (within the bovine D1-3 construct) exhibits high-affinity binding (K_D 1.0 ± 0.3 nM) for the M6P mono-ester of β-glucuronidase.[132] However, D3 is also capable of binding small methyl-M6P di-esters and mannose 6-sulphates found in the amoeba *Dictyostelium discoideum*.[132] The CD loop of D3 is lacking in large, bulky residues. Most notably H1285 of D9 and N104 and H105 of the CD-MPR, which are both specific for M6P monoesters, are replaced in D3 by S386 and S387.[136] The bulky residues in the D9 and CD-MPR FG loops may occlude the binding site, preventing larger di-esters from binding.[136] D3 also has a shorter FG loop than that of D9 and the CD-MPR.[136] This further prevents residues of the FG loop from occluding the binding pocket.[136]

Domain 5

Glycan microarray analysis revealed that D5 preferentially bound GlcNAc-M6P di-esters.[82] When expressed in isolation, D5 exhibited a 4-fold higher affinity for M6P di-esters over M6P mono-esters when studied by SPR (K_D 18 μM versus 72 μM respectively).[139] This observation was supported by NMR titrations that found D5 bound M6P monosaccharide with a K_D of 20 mM and methyl-M6P-GlcNAc with a K_D of 1 mM.[131]

Similarly to D3, D5 exhibits higher affinity binding in the presence of neighbouring domains. D5 demonstrated a significantly higher affinity for GAA di-ester when expressed as a D1-5 penta-domain construct (K_D 60 nM) versus in isolation (K_D 10 μM).[82,138]

The structure of bovine D5 has been determined in isolation by solution NMR in the absence of ligand; with M6P monosaccharide (mono-ester) and with methyl-M6P-GlcNAc (di-ester) bound.[131] D5 does not undergo a conformational change upon ligand binding.[131] Bovine residues Q644 (βC), R687 (FG loop), E709 (FG loop) and Y714 (HI loop) form the conserved 'QREY' motif characteristic of MRH domains (Figure 17). Q644, R687 and Y714 are within hydrogen bonding distance of the 2' OH of M6P. Additionally, Y714 may also form H-bonds to the 3' OH and E416 to the 4' OH of M6P. S712 (HI loop) may H-bond to the 4' OH of the GlcNAc residue of M6P di-ester, while N680 (FG loop) may form H-bonds with the phosphate group and amide at the 2' position of GlcNAc (Figure 17).

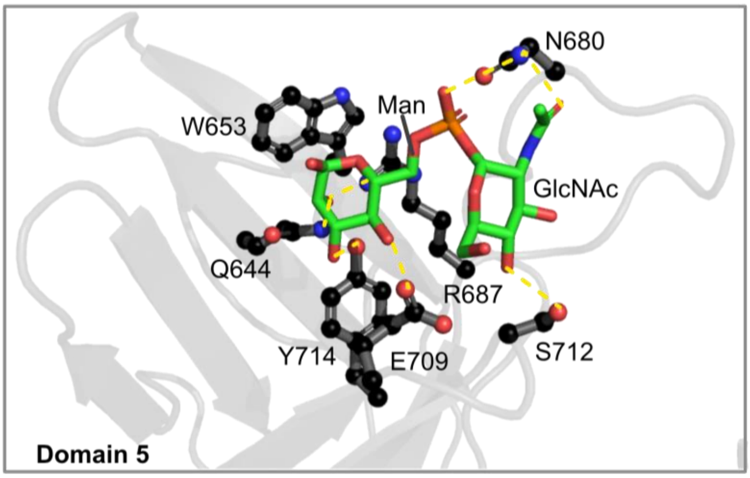


Figure 17: The M6P binding site of bovine CI-MPR D5. NMR solution structure of bovine D5 (PDB 2KVB) with GlcNAc-M6P (green sticks) docked. Sugar binding residues Q644, R687, E709, Y714, N680 and S712 are shown as balls-and-sticks. W653 interacts with R687 to stabilise the binding site. H-bonds are shown as yellow lines.

D5 has a larger, more open binding pocket than D3 and can thus accommodate the bulkier M6P-GlcNAc di-ester. Furthermore, D5 contains only three disulphide bonds (Figure 15), lacking a stabilising disulphide bond at the edge of the M6P binding site between β-strands C and D.[131] D5 instead contains W653 on the CD loop, which interacts with R687 to stabilise the binding pocket (Figure 17).[131] W653 may also interact with the protons at C1 and C5 of the mannose ring.[131] Also not present within D3, Y679 in the FG loop of D5 may interact with the GlcNAc methyl group.[131]

Domain 9

Glycan micro-array analysis and SPR studies have shown that D9, which contains the conserved 'QREY' motif, is specific for M6P monoesters: K_D 95 ± 12 nM for GAA mono-ester and K_D 1.4 ± 1 μM for GAA di-ester.[130,139,82] Glycan micro-array analysis also reveals that D9 recognises mono- and di-phosphorylated high mannose glycans with similar affinity.[82]

To date, D9 is the only CI-MPR domain to exhibit high affinity sugar binding independent of neighbouring domains.[103,130] For example, D9 alone bound M6P-tagged β-glucuronidase with a binding affinity of 0.3 ± 0.1 nM, while D7-9 bound with a K_D value of 0.5 ± 0.1 nM.[137]

Domain 15

As for D9, there is currently no structure of D15. However, sequence alignments (Figure 15) reveal that D15 contains only three (Q2160, E2227, Y2233) of the four conserved sugar binding residues ('QREY') found in other MRH domains, with R2170 and V2205 replacing the conserved arginine residue.[130] A model of D15 reveals that R2170 on βD is positioned similarly to W653 of D5, suggesting a possible role in GlcNAc binding.[130]

Carbohydrate recognition by bovine D15 has been studied in the context of D14-15, D7-15 and, briefly, D1-15. Bovine D15 (of the D14-15 construct) binds M6P mono-esters and di-esters with similar affinity: K_{D} 13 ± 3 μM and 17 ± 7 μM respectively.[130] This interaction was inhibited by addition of M6P but not G6P (glucose 6-phosphate).[130] Similarly to D3 and D5, M6P binding by D15 is enhanced by the presence of neighbouring domains.[130]

Other CI-MPR ligands

As illustrated in Tables 2 and 3, the CI-MPR binds a range of ligands in an M6P dependent or independent manner. The majority of these M6P dependent interactions occur intracellularly at the TGN (pH 6.5) between CI-MPR and lysosomal acid hydrolases. However, the CI-MPR extracellular region can also recognise M6P-tagged proteins at the plasma membrane, which is near neutral pH (pH 7.4). For example, the CI-MPR binds the M6P-tagged glycan of the inactive propreprotein transforming growth factor β (TGFβ), which is present in the extracellular matrix.[104] This interaction appears to be required for TGFβ activation by plasmin and urokinase-type plasminogen activator receptor (uPAR), which also binds the CI-MPR (discussed below).[104] Similarly, M6P-tagged glycans of the inactive proenzyme prorenin are recognised by CI-MPR at the plasma membrane, facilitating prorenin internalisation and subsequent activation to renin.[106]

Only a few ligands exhibit binding through both M6P dependent and independent mechanisms -one example being the secreted, dimeric, glycoprotein cellular repressor of E1A-stimulated genes (CREG).[76] N-link glycosylated CREG has been demonstrated to bind CI-MPR D7-10 through M6P recognition by D9, while deglycosylated CREG binds D11-13 independent of

M6P.[41,42] The interaction between CI-MPR and CREG has proven essential for CREG mediated inhibition of cell growth.[141]

The observation that CREG recognises two sites on the CI-MPR raises questions of whether the CI-MPR is capable of binding multiple ligands simultaneously. uPAR and plasminogen binding has been localised to the N-terminal half of D1 (residues 1-73).[114] This corresponds to the first β sheet (βA-D) and βE, creating a binding surface on D1 that is positioned on the opposite surface to the M6P binding site of D3.[103] This should allow both M6P tagged proteins (such as latent TGFβ) and the M6P independent uPAR or plasminogen to bind at the same time.[103,142] However, addition of β-glucuronidase (an M6P-tagged lysosomal enzyme) has been shown to reduce the binding of uPAR, likely through steric hindrance of the binding site.[142] A similar observation was observed upon addition of IGF2.[142] In the absence of other ligands, uPAR binds bovine CI-MPR with a K_D of 8.9 ± 1.9 μM.[142] Interestingly, however, a soluble, truncated form of uPAR (suPAR) binds the CI-MPR in an M6P dependent manner with co-immunoprecipitation experiments localizing suPAR binding to domains 1-3 and 7-9.[143]

A final example of a ligand that binds the CI-MPR in an M6P independent manner, is retinoic acid (RA) (Figure 18). A lipid metabolite of vitamin A, RA plays a critical role in immunity, cell growth and neuronal development, [144,145] and binds the CI-MPR with high affinity (K_D 2.5 ± 0.3 nM) independent of M6P.[113] RA binding does not inhibit the CI-MPRs ability to bind M6P tagged ligands nor IGF2. In fact, RA and M6P act in a co-operative manner with photolabeling experiments demonstrating an increase in RA binding affinity (to K_D 1.2 ± 0.4 nM) following addition of M6P.[113] Furthermore, the addition of RA to neonatal rat cardio myocytes resulted in a 30-50 % increase in IGF2 internalisation and an intracellular accumulation of M6P-tagged acid phosphatases in the lysosome.[113] Although not yet fully determined, RA has been proposed to bind the C-terminal 40 kDa of the CI-MPR which would encompass the intracellular region of the receptor.[113]

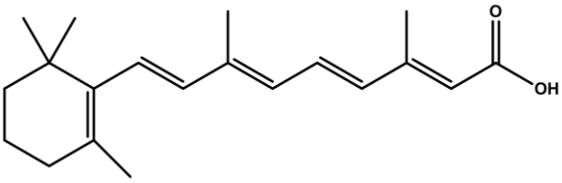


Figure 18: The structure of retinoic acid, a lipid metabolite of vitamin A, that is bound by the CI-MPR with nanomolar affinity.

CI-MPR in human disease

The P-type lectins were first discovered during investigations into the lysosomal storage disorder mucolipidosis II, also known as I-cell disease.[54] In I-cell disease a deficiency in catalytic subunits of GlcNAc-1-phosphotransferase results in missorting and secretion of lysosomal hydrolases.[98] Lysosomes are thus non-functional and form dense inclusion bodies of undigested material.[98] The same phenotype has been observed in transgenic mice lacking in both P-type lectins.[54] Meanwhile, mutations in UCE, the uncovering enzyme responsible for M6P mono-ester formation, have been implicated in persistent stuttering.[146] Other lysosomal storage disorders including Niemann-Pick disease and Krabbe disease, are due to deficiencies in lysosomal enzymes sphingomyelinase and galactocerebrosidase respectively.[147]

Thus, the P-type lectins are implicated in a number of lysosomal storage disorders and neurodegenerative diseases. In mice models of Parkinson's disease, the CI-MPR was missorted, resulting in a reduction of functional cathepsin D - a lysosomal protease that degrades α-synuclein aggregates.[148] This is supported by the observation that in human brain tissue of early stage Parkinson's patients, the CI-MPR was found to be down-regulated, resulting in reduced cathepsin D trafficking and thus impaired lysosomal clearance of the toxic α-synuclein aggregates.[148] Samples of human Alzheimer's brain tissue also shows mis-regulated CI-MPR expression and localisation, disrupting the trafficking of cathepsins B and L that are involved in the lysosomal breakdown of β-amyloid plaques.[149]

The significance of the CI-MPR/ IGF2R and IGF2 in human health and disease is reflected in the tight regulation of their expression. In viviparous mammals the genes encoding both IGF2 and IGF2R are imprinted - a form of epigenetic modification involving DNA methylation and histone acetylation to ensure transcription occurs from a single allele only.[150] During embryogenesis, *IGF2* and *IGF2R* are reciprocally imprinted, with *IGF2R* being maternally expressed and *IGF2* paternally expressed.[151] Imprinting of *IGF2R* leaves the receptor function susceptible to mutation.[118, 151] Loss of loss of function mutations in *IGF2R* are common in epithelial based cancers.[122] For example 60 % of early stage hepatocellular tumours and 30 % of early stage breast tumours exhibited *IGF2R* loss of heterozygosity.[96] To date nine cancer related missense mutations have been observed in the extracellular region of the CI-MPR.[67] These include: C1262S (BC loop) and G1296R (βD) in D9, Q1445H (DE loop), G1449V (DE loop) and G1464E (FG loop) in D10 and G1564R (βC), I1572T (BC loop), A1618T (GH loop) and G1619R (GH loop) in D11.[67,152–155] Loss of IGF2R activity leads to protein mis-sorting and decreased clearance of mannosylated proteinases such as cathepsins, which are involved in extracellular matrix degradation and thus aid tumour invasion.[156] The CI-MPR is, therefore, a tumour suppressor protein, with overexpression shown to contribute to increased apoptosis in tumours.[157]

Loss of IGF2 imprinting – by hypomethylation of the paternal imprinting control region 1 (ICR1) downstream of *IGF2*,[158] and subsequent bi-allelic *IGF2* expression is responsible for approximately 40 % of Silver-Russell Syndrome (SRS) cases.[159, 160] SRS is a rare (1 in 30,00-1 in 100,000) but severe growth retardation disorder.[159,161] Conversely, *IGF2* gene silencing by hypermethylation of ICR1 can cause the foetal over-growth disorder Beckwith-Weidemann syndrome, in which patients have a pre-disposition to tumours.[159,162] Thus correct regulation of IGF2 by IGF2R is important during embryogenesis. IGF2R knock-out mice displayed severe developmental abnormalities and died at birth.[115] Loss of IGF2R results in increased bio-availability of IGF2, which promotes cell growth, proliferation and angiogenesis, while suppressing apoptosis.[118] Increased big IGF2 isoforms also contribute to the syndrome non-islet cell tumour hypoglycaemia (NICTH).[118]

While the majority of the CI-MPR population is located intracellularly – constitutively cycling between the TGN and endosomes, approximately 10 % is present in the plasma membrane. In the plasma membranes of endothelial cells, the extracellular region of the CI-MPR can be cleaved at the C-terminus of D15 by the protease tumour necrosis factor α-converting enzyme (TACE/ ADAM17).[163] This results in soluble sCI-MPR (D1-15) that circulates in the blood.[163] sCI-MPR can form heterodimers with membrane bound CI-MPR, promoting receptor shedding from the plasma membrane.[164] Increased sCI-MPR has been observed in patients with breast cancer, liver cirrhosis and morbid obesity.[102,163,165] Similar to the IGFBPs, sCI-MPR binds and sequesters IGF2 to modulate organ size.[166] sCI-MPR also binds plasminogen, inhibiting its interaction with and activation by uPAR and thus down-regulating cell migration and angiogenesis.[163] In this manner sCI-MPR acts as a tumour suppressor protein.[163] This is supported by the observation that sCI-MPR halts the progression of intestinal adenoma in mice.[157,167] Additionally, sCI-MPR can sequester extracellular ligands including Granzyme B, CD26, CD45 and Lck, which are all involved in the regulation of cytotoxic T cells.[102]

The CI-MPR, or components derived from it, might therefore be potential therapeutics. Of the eleven enzyme replacement therapies (ERT) approved for the treatment of lysosomal storage disorders, nine rely upon the CI-MPRs ability to traffic M6P tagged proteins.[138] For example, recombinant human acid α-glucosidase (rhGAA) is an ERT to treat the lysosomal storage disorder Pompe disease.[168] *In vitro* and in mice models, Cheng *et al.* have demonstrated that the efficiency of this ERT can be enhanced by increasing rhGAA targeting (through the addition of M6P tagged oligosaccharides to rhGAA) to the CI-MPR and thus increasing rhGAA endocytosis.[168] Similarly, Bali *et al.* improved the efficiency of rhGAA ERT in mice by up-regulating CI-MPR.[169]

More recently, in 2016 Frago *et al.* developed the first high-affinity IGF2 specific ligand trap based upon D11 of the CI-MPR.[127] This has proven successful in inhibiting *in vivo* IGF2 signalling in hypoglycaemic mice.[127]

CI-MPR structural knowledge

Table 5 and Figure 19 below show the CI-MPR structures determined to at the outset of this work. There were no high-resolution structures of the sugar binding domains, D9 and D15 or D6-8 and D10. Neither are there any structures of the P-type lectins in complex with M6P-tagged glycoproteins. Of the 15 CI-MPR extracellular domains, the IGF2 binding domain, D11, is the most well studied, with high-resolution structures of human, chicken, echidna and opossum D11 being determined to delineate the evolution of IGF2 binding.[124] Bovine forms of the sugar binding domains, D3 and D5, have been expressed in *Trichoplusia ni* 5B1–4 (Tn-5B1-4) insect cells and *P. pastoris* yeast respectively.[131,136] Difficulties in obtaining soluble protein domains stems from the presence of 3-4 disulphide bonds in each domain, which require refolding after bacterial expression. This makes bacterial expression and purification of folded multi-domain constructs challenging.

Domain(s)	Method	Organism	Expression host	PDB
D1-3	X-ray diffraction	Bovine	Tn-5B1-4	1SYO, 1SZ0, 1Q25
D5	NMR	Bovine	*P. pastoris*	2KVA, 2KVB
D11	X-ray diffraction	Human	*E. coli*	1GP0, 1GP3
D11	NMR	Chicken, echidna, opossum	*E. coli*	2L21, 2LLA, 2L2G
D11-IGF2	NMR	Human	*E. coli*	2CNJ
D11-12	X-ray diffraction	Human	CHO	2V5N
D11-13-IGF2	X-ray diffraction	Human	CHO	2V5P
D11-14	X-ray diffraction	Human	CHO	2V5O

Table 5: CI-MPR extracellular domains solved at the outset of this work.

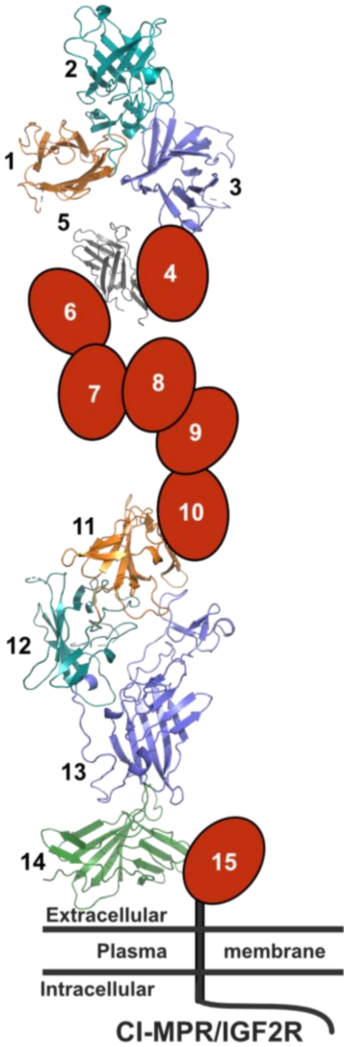


Figure 19: Structural characterisation of CI-MPR extracellular region. The extracellular region of the CI-MPR consists of 15 homologous domains. At the outset of this work the structures of bovine D1-3 had been determined with M6P or a mannosylated glycan in the D3 binding site (PDB 1SYO/1SZ0 and 1Q25 respectively). The structure of bovine D5 had been determined alone and in complex with methyl-M6P-GlcNAc by solution NMR (PDB 2KVA, 2KVB). The structure of human D11 had been determined alone and within multi-domain constructs D11-12, D11-14 and D11-13 plus IGF2 (PDB 1GP0/1GP3, 2V5N, 2V5O and 2V5P respectively). Domains coloured red had not been structurally characterised.

At the outset of this work, the largest fragment of the CI-MPR solved was a preliminary crystal structure of a 160 kDa homodimer of D7-11 (Figure 20). This was expressed, purified, and a single hit was crystallised by Hans Hoppe and Karl Harlos at the University of Oxford after numerous optimisations and trials. The crystals in question were, however, partially radiation damaged and it was suggested that D11 interacts weakly with the remaining D7-10 which may also have reduced the quality of crystals and diffraction data obtained. Nevertheless, a tentative model (Figure 20) was produced by Airlie McCoy (University of Cambridge) using a series of homology models generated for D7-10, the crystal structure of D11 (PDB 1GP0) and application and development of state-of-the-art molecular replacement approaches. However, a final, publishable, refined structure proved to be elusive.

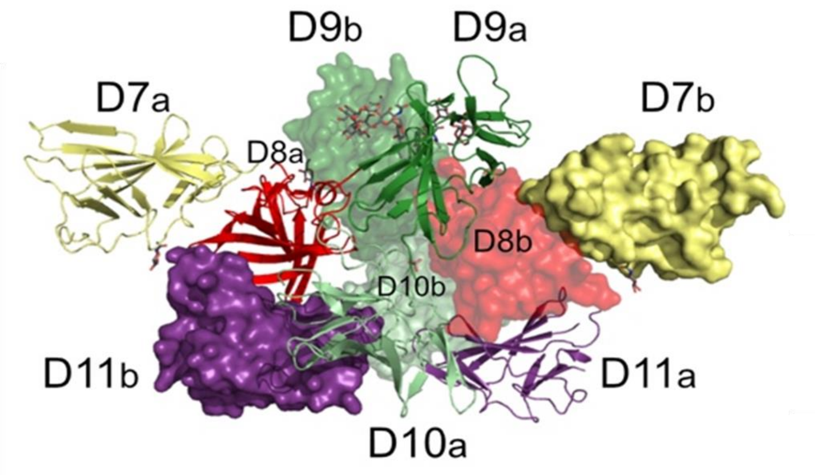


Figure 20: The low resolution (3.8 Å) crystal structure of the D7-11 homodimer. One monomer is in ribbon format, the other in surface representation.

At the start of this project there was no structure of the full extracellular region of the CI-MPR. An initial model for the CI-MPR extracellular region was proposed by Olson *et al.* in 2004 and comprised five tri-domain units each containing an MRH domain (D1-3, D4-6, D7-9, D10-12, D13-15).[103] This was consistent with the compact crystal structure of D1-3, D1-3 and D4-6 proteolytic stability and homology modelling the association of the remaining domains (D7-9, D10-12, D13-15).[103,170] This model was later modified by Brown *et al.* in 2008 to include the D11-14 crystal structure.[101] While maintaining the three N-terminal tri-domains D1-9, dimerisation interfaces at domains 3, 5, 9, 12 and 15 were introduced.[101]

Recently, however, cryo-EM has been performed on CI-MPR extracted from bovine liver. Wang *et al.* have determined the structures of bovine D1-14 at pH 4.5 (3.5 Å) and D4-14 plus IGF2 at pH 7.4 (4.3 Å) (PDB 6UM1 and 6UM2 respectively).[128] Although these structures are monomeric, CI-MPR oligomerisation remains poorly understood (see section 4.5). In the context of full length CI-MPR, both monomeric and dimeric forms of sCI-MPR purified from bovine livers have been observed by native gel electrophoresis.[171]

1.6. Project aims

Despite structural characterisation of several CI-MPR domains (Table 5), the structure of the selective, high-affinity and independent M6P binding domain D9 has, however, remained elusive. The primary aim of this work is to determine the structure of D9 and characterise its interaction with M6P. Previous work by Dr Chris Williams (University of Bristol) demonstrated that soluble D9 could not be obtained in isolation. Thus, the work presented in this thesis focuses on D9 within multi-domain constructs. This includes attempts to improve the low-resolution D7-11 crystal structure using two previously uncharacterised CI-MPR domains, D7 and D8, (chapter 2) and generation of novel multi-domain constructs that contain D9 (chapter 3).

This work also aims to study the domain arrangement, oligomerisation and effects of ligand binding in the context of the full extracellular region of the human CI-MPR, i.e. domains 1-15 (chapter 4). As a monomer the extracellular region is 250 kDa with 19 N-linked glycosylation sites and 59 disulphide bonds. Thus, D1-15 is a complex and ambitious target but ideal for characterisation by cryo-EM.

In parallel to this, a second project (chapter 5) aims to engineer a synthetic lectin based upon D11 of the CI-MPR. This work builds upon preliminary work by Dr Chris Williams (University of Bristol) and tests whether a scaffold evolved to bind a ‘large’ macromolecule (e.g. IGF2) can be engineered to bind a small ligand (i.e. mannose 6-phosphate or perhaps even glucose).

2. Domains 7 and 8

2.1. Introduction and aims

Currently D11 is the most well characterised domain of the CI-MPR owing to its role in regulating IGF2 bioavailability. Due to the presence of eight cysteines that form four disulphide bonds, D11 forms insoluble inclusion bodies when expressed bacterially. However, a robust *in vitro* refolding protocol for D11 expression in *E. coli* was developed by Brown *et al.* in 2002 (based upon Gao *et al.* 1998).[112,172] Using variations of this protocol, high-resolution X-ray or NMR solution structures have been determined for human, echidna, chicken and opossum D11 (PDB 1GP0, 2LLA, 2L21, 2L2G)[101,124] as well as D11 variants engineered for high-affinity ligand binding.[127]

The aim of the work in this chapter was to apply this refolding method to the expression of two previously uncharacterised domains – domain 7 (D7) and domain 8 (D8). When this project was initiated the only structural information available for D7 and D8 was from a tentative model of the 160 kDa homodimer of D7-11 collected to 3.8 Å. High-resolution crystal structures of domains 7, 8, 9 or 10, which have not been structurally characterised elsewhere, may assist in the phasing and refinement of this D7-11 crystal structure.

2.2. Bacterial expression of D7 and D8

Synthetic genes encoding D7 and D8 (GeneArt, ThermoFisher Scientific, Appendix Section 8.1.) were individually subcloned into the expression vector pET28a (Novagen) by traditional restriction cloning methods (using NdeI, XhoI and T4 ligase). Following confirmation by Sanger sequencing (Genewiz), constructs were expressed in *E. coli* BL21 (DE3, Novagen). D11 was expressed from pET28a in parallel as a control. Due to the reducing environment of the bacterial cytosol and the presence of three disulphide bonds in D7 and four in each of D8 and D11, these domains form insoluble inclusion bodies – dense aggregates of misfolded recombinant protein.[173]

The inclusion bodies were purified and resolubilised in 8 M urea. Denatured protein was refolded following the rapid dilution protocol established by Brown *et al.* for D11.[112] Briefly, denatured protein is diluted into refold buffer containing Tris pH 8 (a pH buffer), EDTA (a metal chelator that suppresses metal-catalysed cysteine oxidation), L-arginine (a chaotropic aggregation suppressor), cysteamine and cystamine (a redox pair that facilitates correct disulphide bond formation).[173,174] After 24 hrs recombinant protein was purified by size exclusion chromatography (SEC) and analysed by SDS-PAGE.

Figure 19 shows the successful application of this method to the refolding of D11. Folding was assessed by gel filtration and 1D ^{1}H-NMR. When purified by SEC (using a preparative 26/60 S75 column), natively folded, monomeric D11 elutes between 160-180 mL (Figure 21A). The minor peak at 100-120 mL seen in the purification of all three recombinant proteins (D11, D7 and D8) corresponds to high molecular weight aggregates. Protein aggregation results from non-native, intermolecular hydrophobic interactions due to protein mis-folding.[175] Purification of D11 was confirmed by SDS-PAGE and mass spectrometry. 1D ^{1}H-NMR confirmed native folding of the protein (Figure 21B). The upfield chemical shifts in the region 0.5 to -0.5 ppm correspond to methyl groups that have been shifted away from common methyl chemical shifts (e.g. ~0.85ppm for valine and leucine methyl groups).[176] This is due to the ring current effects of nearby aromatic residues and typically only occurs when a protein is folded.[176]

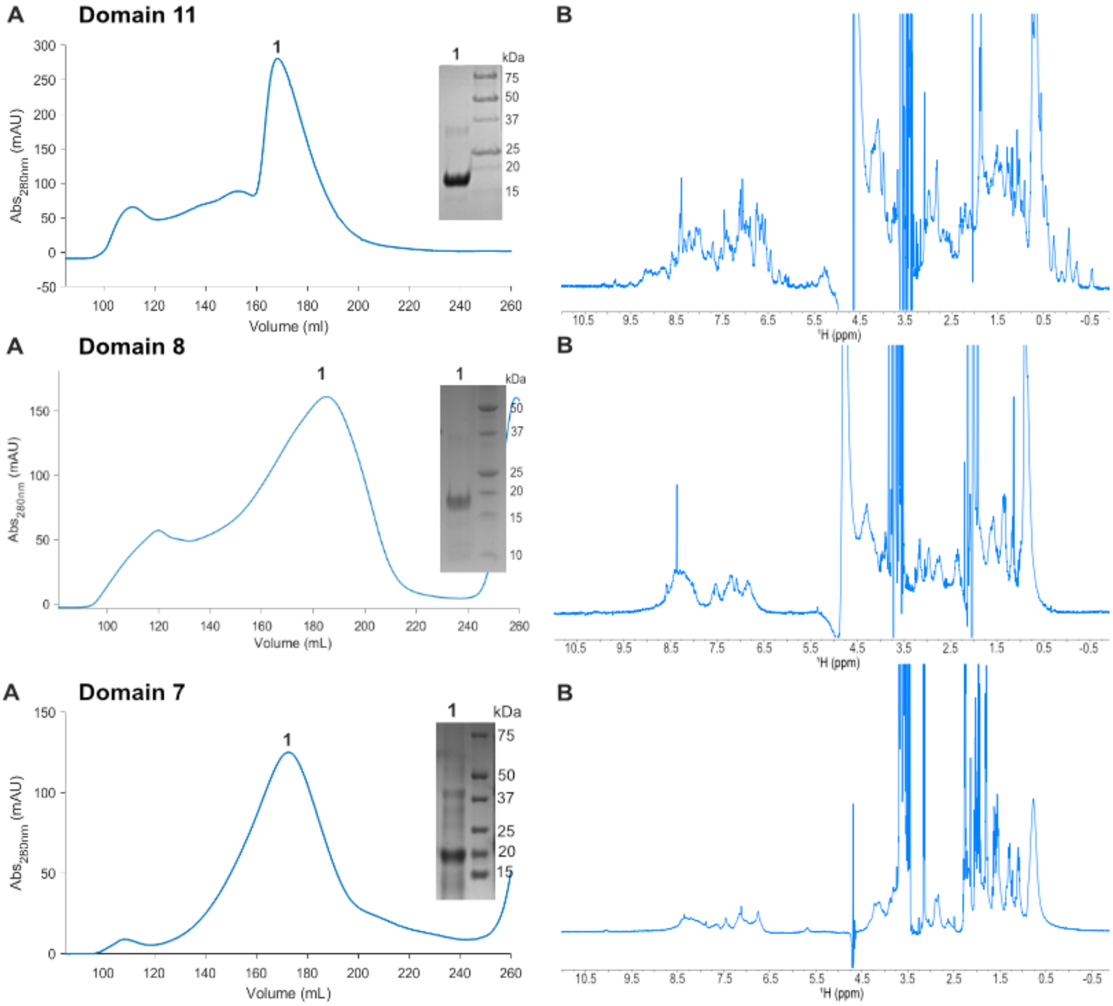


Figure 21: Bacterial expression and *in vitro* refolding of D11 (top), D8 (middle) and D7 (bottom). A: SEC trace of *in vitro* refold and SDS-PAGE analysis showing elution of purified recombinant protein. Expected molecular weights: D11 19.8 kDa, D8 17.5 kDa and D7 18.6 kDa. **B:** 1D ^{1}H-NMR of sample 1 from *in vitro* refolds confirming that only D11 has been successfully refolded.

However, use of this protocol did not yield natively folded D7 nor D8. Whilst refolded D7 and D8 eluted from the SEC column within the expected calibrated mass range (Figure 21A), the absorbance peak was broad suggesting the presence of multiple species/ states. 1D ^{1}H-NMR confirmed that the major peak fractions of D7 and D8 were unfolded (Figure 21B). An extensive screen of refold conditions was performed in attempt to optimise the *in vitro* refolding protocol. Changes were made to the sample preparation, refold buffer and to the refold method (Table 6). However, despite an extensive screen of refold conditions (Table 6) in attempt to optimise the *in vitro* refolding protocol and screening by 1D ^{1}H NMR, natively folded D7 or D8 could not be obtained following expression in *E. coli* BL21 (DE3).

Condition	Change
1. Protein concentration	1-0.01 mg/ mL
2. Protein construct	His_6 tagged, untagged
3. Reducing agent (DTT) concentration	0-100 mM
4. Denaturant	8 M urea, 6 M guanidinium hydrochloride
5. pH	pH 7-9
6. Arginine concentration	0-2 M
7. Redox pair	Cystamine-cysteamine, glutathione-reduced glutathione
8. Redox pair ratio	1:1 -10:1 (reduced: oxidised)
9. Additives	Glutamate, sucrose, ETDA, sodium chloride, potassium sulphate
10. Denaturing IMAC	
11. On column (IMAC) refold	
12. Dialysis	

Table 6: D7 and D8 *in vitro* refold conditions screened. Changes were made to sample preparation (green), to refold buffer (blue) and to refold method (red) individually and in combination.

2.3. Insect cell expression

Alternate approaches have proved successful in expressing single CI-MPR domains bacterially. For example, Olson *et al.* demonstrated successful expression of folded bovine D5 in *E. coli* BL21 (DE3).[177] D5, which contains three disulphide bonds, was expressed as an insoluble fusion protein with the small ubiquitin-like modifier (SUMO) protein at the N-terminus – a modification previously demonstrated to enhance protein solubility.[177,178] Recombinant SUMO-D5 was refolded *in vitro* using the rapid dilution method, purified and the SUMO tag cleaved.[177] A well dispersed 2D ^{1}H-^{15}N HSQC and crystals diffracting to 1.8 Å demonstrated native folding of bacterially expressed D5 suitable for structural characterisation.[177]

However, with the aspiration to structurally characterise larger, multi-domain CI-MPR constructs, the decision was made to use a eukaryotic expression system, which due to their compartmentalisation, are intrinsically capable of disulphide bond formation. Therefore, insect cells were chosen as expression hosts for the CI-MPR domains due to their post-translational modifications (notably disulphide bond formation and N-linked glycosylation), relatively high yields, and safety – the baculoviruses produced are unable to replicate within mammalian cells.[179]

Insect cell expression of D11

D11 has already been well characterised structurally and functionally through *E. coli* expression, *in vitro* refolding, X-ray and NMR studies, IGF2 binding assays and mutagenesis.[79,101,112,124] However, being only approx. 15 kDa in size, containing eight cysteines that form four disulphide bonds and having no *N*-linked glycosylation sites, D11 was used initially to demonstrate that insect cells were suitable expression hosts for single domains of the CI-MPR.

In the baculoviral expression vector system (BEVS), insect cells are infected with recombinant baculovirus containing the gene of interest on a baculoviral artificial chromosome (BAC or bacmid).[180] Here the MultiBac system developed by Professor Imre Berger, predominantly for the expression of multi-subunit complexes but also with the advantage of streamlined standard operating protocols, was used.[181] Table 7 below summarises the major modifications that have been made to the 130 kb double stranded DNA genome of the *Autographa californica* nuclear polyhedrosis virus (AcNPV) to produce the MultiBac BAC.[181] Although the AcNPV has a broad range of insect cell hosts, it cannot replicate within mammalian cells.[179]

Gene	Function	Modification	Result
polh	Very late gene encoding the viral coat protein polyhedron.	Removed	Recombinant protein can be placed under control of *polh* promoter.
p10	Very late gene encoding p10 – function unknown.	Removed	Recombinant protein can be placed under control of *p10* promoter.
v-cath	Encodes cathepsin - a cysteine protease involved in host cell degradation.[182]	Removed	Improves stability of secreted recombinant proteins.[183]
chiA	Encodes chitinase – degrades chitin and processes v-cath.[182]	Removed	Improves stability of secreted recombinant proteins.[183]
YFP	Encodes Yellow fluorescent protein.	Added	Facilitates monitoring of recombinant protein expression.[179,184]

Table 7: The major modifications that have been made to the AcNPV genome in the MultiBac system for recombinant protein expression.

A synthetic gene encoding D11, codon optimised for expression in *Spodoptera frugiperda* and with an N-terminal RPTPμ signal sequence for recombinant protein secretion and a C-terminal hexaHistidine tag for purification, was synthesised and sub-cloned (by GeneArt, ThermoFisher Scientific) into the transfer vector pFastBac, creating D11WT-pFastBac (Appendix Figure 1).

This places the D11 gene under the control of the very late baculoviral promoter *polh* (polyhedrin promoter) and within two Tn7 transposition sites that are required for bacmid formation.

The D11 gene was inserted into the baculoviral DNA from the pFastBac transfer vector by Tn7 transposition – the Tn7 enzyme being provided by a helper plasmid present within EMBacY DH10 *E. coli* cells (kindly provided by the Berger group, School of Biochemistry) (Figure 22).[179,185] Colonies containing recombinant bacmid were selected by blue-white screening. White colonies were simultaneously streaked onto fresh lysogeny broth (LB) agar plates and added to 3 mL LB containing the relevant antibiotics. Both plates and cultures were incubated overnight at 37 ºC before bacmid DNA was isolated by alkaline lysis from LB cultures of confirmed white colonies. Agarose gel electrophoresis (Figure 22) confirmed the presence of bacmid DNA.

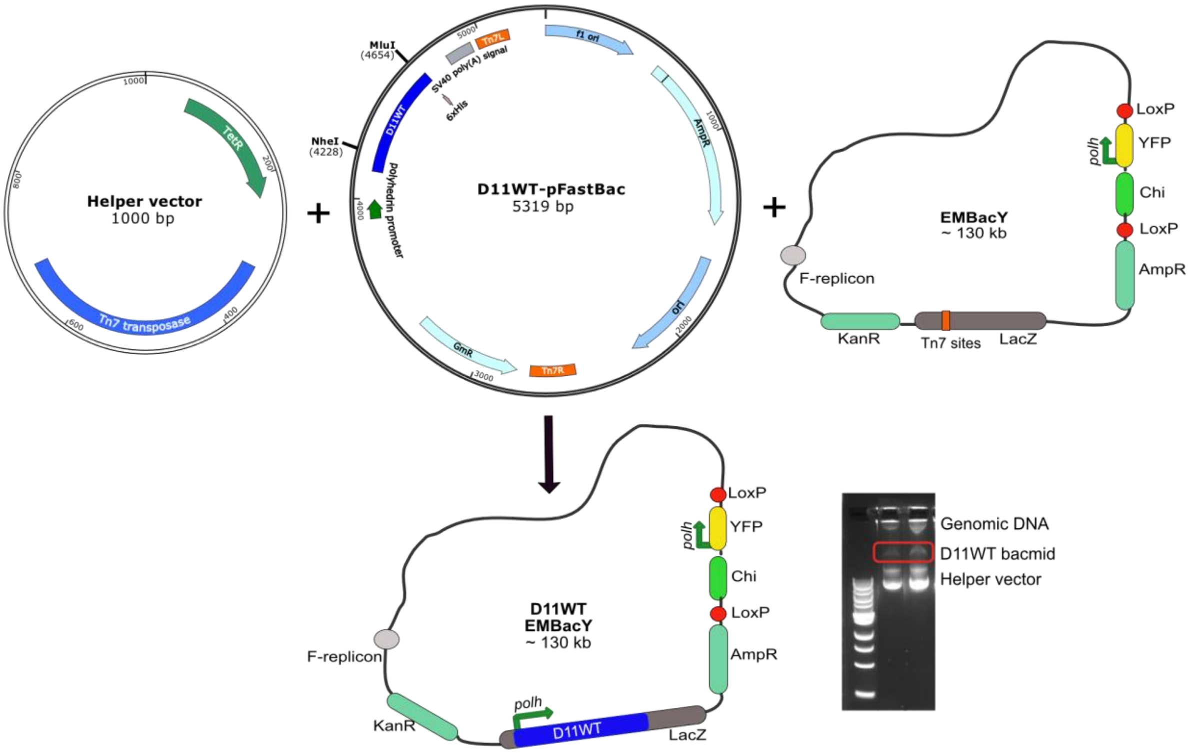


Figure 22: Generation of D11 bacmid DNA. EMBacY DH10 *E. coli* cells were transformed with D11WT-pFastBac transfer plasmid. Tn7 transposase, encoded on a helper plasmid in the *E. coli* cells, facilitated the integration of the D11WT gene into the EMBacY bacmid. Successful integration disrupted the LacZα gene of EMBacY giving rise to white colonies. Following selection and growth of white colonies, bacmid DNA was isolated and confirmed by 0.8 % agarose gel electrophoresis.

Bacmid DNA was transfected into an ovarian cell line from the fall army worm (*Spodoptera frugiperda* cell line 21, Sf21) via lipid transfection. Viral titre V_0 was harvested after 48 hours incubation at 27 ºC and amplified in fresh Sf21 shaking cultures. Viral titre $V_{1,}$ which was used to infect subsequent cultures, was harvested 24 hours after proliferation arrest. Cell proliferation was monitored daily with samples containing 1x10^{6} cells collected every 24 hours.

One of the advantages of the MultiBac system, is the addition of the YFP gene (as indicated in Table 7 above) to facilitate determination of the optimum harvest time.[181] YFP and the gene of interest (D11 here) are both under the control of the *polh* promoter (Figure 22, D11WT EMBacY).[181] It follows, therefore, that if YFP is expressed, the protein of interest should also be expressed. YFP emission is measured in soluble cell fractions prepared from 1x10^{6} cells, making this a quantitative analysis (Figure 23A). Cultures should be harvested when the YFP emission peaks or plateaus` as at this point protein expression has generally reached a maximum. Beyond this time point YFP emission decreases as YFP degradation, cell death and lysis increases.

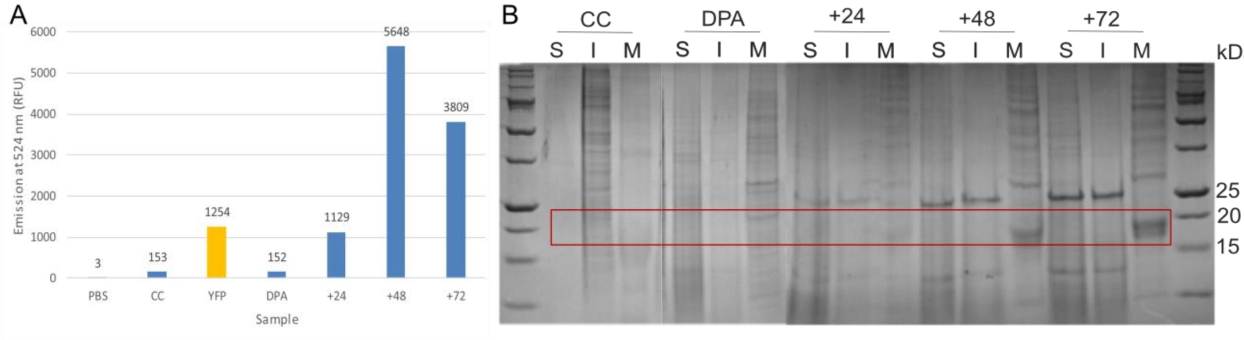


Figure 23: Expression of D11 in insect cells. A: YFP emission at 524 nm peaks. **B:** SDS-PAGE of D11 soluble cell fractions (S), insoluble (I) and media (M). Expression and secretion of D11 (within the red box, expected molecular weight 18.6 kDa) is first visible 48 hours after proliferation arrest. YFP (expected molecular weight 28.1 kDa) is also visible in the soluble and insoluble cell fractions. Abbreviations: PBS-phosphate buffered saline, CC-uninfected cell control, DPA-day of proliferation arrest, +24 -24 hours after proliferation arrest.

Based upon the YFP emission data and SDS-PAGE above (Figure 23), D11 was harvested 72 hours after proliferation arrest (DPA+72). Due to the presence of an N-terminal signal peptide, D11 is secreted into the culture medium (Figure 23B), which was clarified by centrifugation. The pH of the culture media containing recombinant D11 was adjusted from pH 6.5 to 5.5 by dilution with sodium acetate buffer pH 5.0, keeping below the isoelectric point of D11 (8.2). This buffered media was filtered and loaded onto an SP ion exchange (IEX) column and D11 eluted over a gradient of sodium chloride (Figure 24A). D11 was isolated to greater than 90 % purity and at a yield of 46 mg/ L.

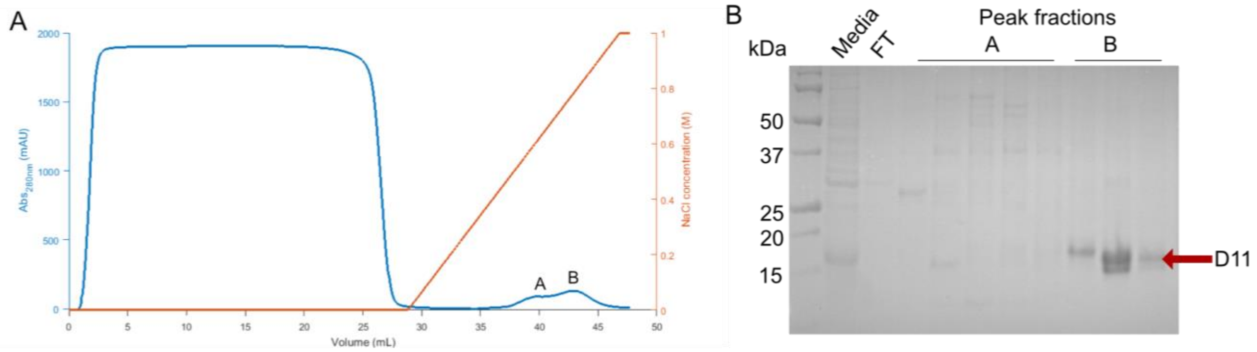
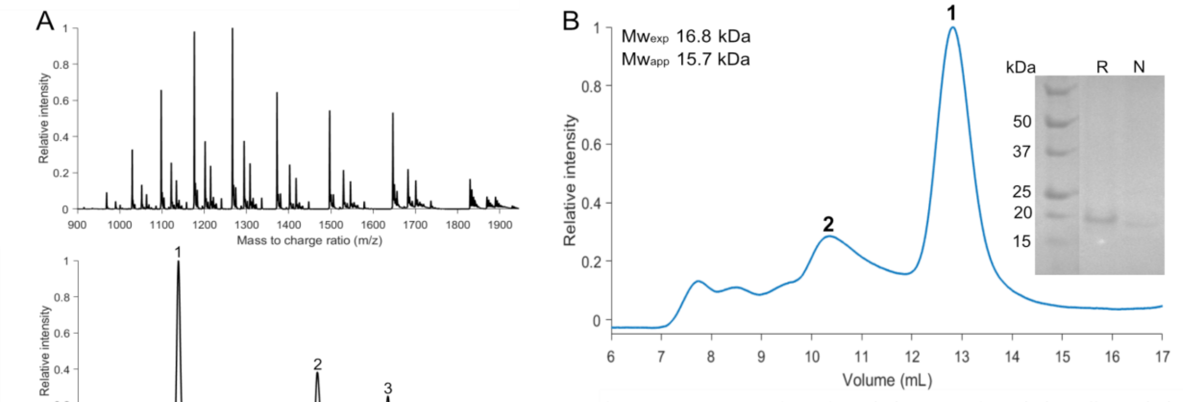


Figure 24: Purification of D11 from insect culture media. A: IEX chromatogram of D11 harvest from cell culture medium. Protein was eluted from the SP column over a gradient of sodium chloride. **B:** SDS-PAGE analysis of IEX purification confirmed the presence of purified D11 in peak B.

Electrospray ionisation-mass spectrometry (ESI-MS) (Figure 25A) confirmed expression and purification of D11 (expected mass: 16817.9 Da, observed mass: 16817.2 Da). However, ESI-MS revealed the sample to be heterogeneous with multiple charge state envelopes observed. The major species corresponds to a truncated version in which the four N-terminal residues (ETGA) of the signal sequence were absent (expected mass: 16548.9 Da, observed mass: 16548.5 Da). There is also a minor species at 16999.5 Da (182 Da greater than the expected mass) that has not been identified but may correspond to a combination of phosphorylation of D11 at a serine, threonine or tyrosine and a phosphate adduct (expected mass addition of +178 Da). There are no N-linked glycosylation sites in D11.



Number	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
1	Truncated D11	16458.5	16458.9	-0.4
2	Full length D11	16817.2	16817.9	-0.7
3	Unidentified	16999.5	16817.9	+182.3

Figure 25: Initial characterisation of insect expressed D11. A: ESI-MS of D11 reveals 3 species: full length D11 (2), N-terminally truncated D11 (1) and an unidentified species (3) 182 Da larger than expected. Small quantities of D11 sodium adducts (black crosses), potassium adducts (blue) and phosphate adducts (green) were also observed. **B:** Analytical SEC of D11 reveals the major species (1) to be monomeric (Mw$_{exp}$ 16.8 kDa, Mw$_{app}$ 15.7 kDa). SDS-PAGE of the major species under reducing (R) and non-reducing (N) conditions suggests disulphide bond formation.

Analytical SEC confirmed that the majority of D11 (1) was monomeric (Mw$_{exp}$16.8 kDa, Mw$_{app}$ 15.7 kDa), with a small peak (2) that may correspond to a D11 dimer (Mw$_{exp}$ 33.6 kDa,

Mw_{app} 40.5kDa) (Figure 25B and Appendix Figure 3). The major species was analysed further by denaturing and native SDS-PAGE (Figure 25B inset). Unlike in denaturing SDS-PAGE, where protein samples are denatured and reduced by boiling with SDS and β-mercaptoethanol, in native SDS-PAGE samples are not denatured, allowing separation of different oligomeric states.[186] Under native conditions, whereby the absence of reducing agent maintains the cysteine residues as oxidised disulphide bonds, protein samples migrate slightly faster and give rise to discrete bands. The slight band shift between D11 under denaturing, reducing conditions (R) and non-denaturing, non-reducing conditions (N) (Figure 25B inset) suggests the presence of disulphide bonds and is indicative of protein folding.

Native protein folding was confirmed by 1D ^{1}H-NMR (Figure 26A). The upfield chemical shifts in the region 0.5 to -0.5 ppm for methyl protons and sharp dispersed signals across the full range of protein signals, suggested that the protein is folded.[176] Figure 24 shows a comparison of 1D ^{1}H-NMR spectra of D11 expressed in insect cells (produced here) and bacterial cells (Dr Chris Williams, University of Bristol). Shifts in the upfield region are identical between the two spectra, suggesting that D11 expressed in insect cells is adopting the same fold as that expressed bacterially.

X-ray crystallography has been routinely used to structurally characterise multi-domain CI-MPR constructs (for example, D1-3, D11-12 and D11-14, PDB 1SYO/1SZ0, 2V5N and 2V5O respectively), single D11 constructs (PDB 1GP0 and 1GP3) and D11 mutants engineered for high-affinity IGF2 binding (for example D11 AB5 RHH, PDB 5IEI). Thus, X-ray crystallography was employed here to demonstrate that D11 expressed using the baculoviral expression vector system adopts the native β-sandwich fold previously observed by bacterial expression.

D11 (140 amino acids, N1511-T1651) in 0.025 M tris pH 7, 0.15 M sodium chloride crystallised from a solution of 30 % PEG 4000, 0.1 M Tris-HCl pH 8.5, 0.2 M sodium acetate for crystal structure determination. The construct crystallised in space group $P2_12_12_1$ with 1 molecule in the asymmetric unit. The structure of D11 was determined to 2.0 Å (Rwork and Rfree values of 18.3 and 22.4 % respectively) with 100 % of the backbone dihedral angles in allowed and favoured regions of the Ramachandran plot (Figure 26C, Appendix Table 4), by molecular replacement using the crystal structure of bacterially expressed D11 (PDB 1GP0) as a search model. There was no electron density for N1511-D1515 at the N terminus or Q1649-T1651 at the C terminus of D11.

Insect expressed D11 forms the same core β-sandwich topology composed of nine anti-parallel β-strands as previously observed in the structures of D1-3, D5, D11 and D11-14.[101,131,136] Superimposition of insect expressed D11 (this work) and bacterially expressed D11 (PDB 1GP0) crystal structures gives an RMSD value of 0.2 Å over backbone atoms (Figure 26C).

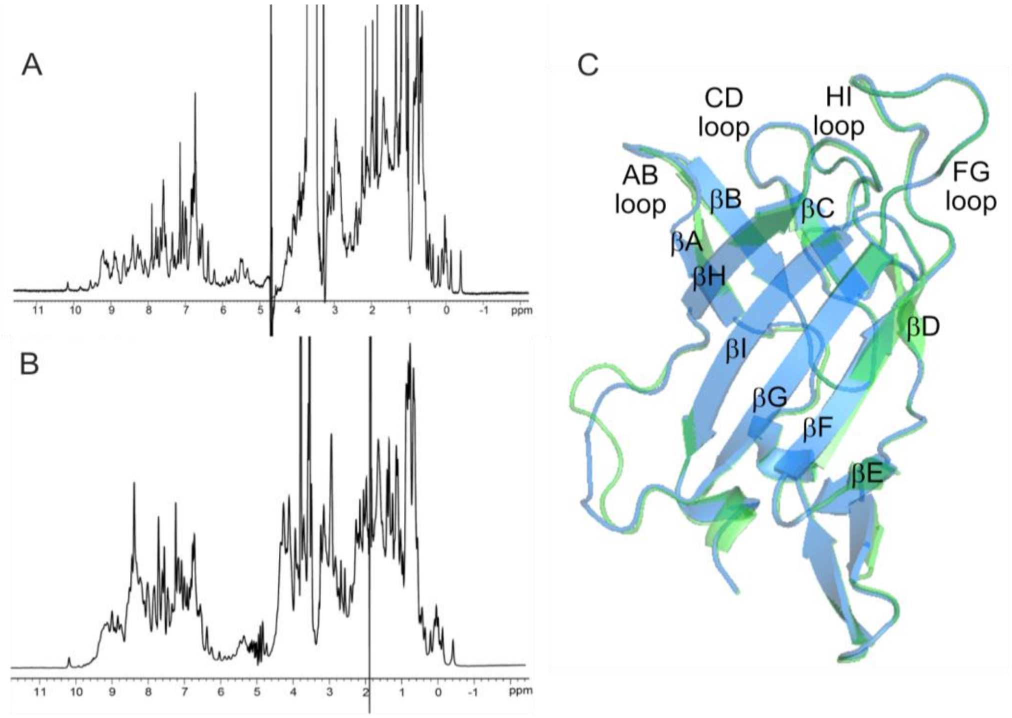


Figure 26: Structural characterisation of D11. A: 1D-^{1}H NMR spectra of D11 expressed in insect cells confirms native protein folding. **B:** 1D-^{1}H NMR spectra of D11 expressed bacterially (Dr Chris Williams, University of Bristol) has the same pattern of upfield peaks suggesting the same protein fold. **C:** Superimposition of the insect expressed D11 (blue) and bacterially expressed D11 (green, PDB 1GP0) crystal structures gives an RMSD value of 0.2 Å over backbone atoms.

Insect cell expression of D7 and D8

Expression and purification of D7 and D8

Following successful expression and purification of D11 in insect cells, this approach was applied to D7 and D8 for structural characterisation. DNA encoding D7 (465 bp) and D8 (420 bp) was amplified by PCR from a synthetic gene encoding D1-15 (GeneArt, ThermoFisher Scientific) and subcloned, using restriction endonucleases EcoRI and HindIII, into the modified pFastBac transfer vector (Appendix Figure 1). (pFastBac was again modified to contain the N-terminal RPTPμ signal sequence for recombinant protein secretion and a C-terminal hexaHistidine tag). EMBacY DH10 *E. coli* were transformed individually with D7pFastBac or D8pFastBac and recombinant D7 and D8 protein was expressed in Sf21 insect cells using the protocols previously established for the expression of D11WT.

The optimal harvest time for each protein was determined by SDS-PAGE analysis and YFP emission as 96 hrs after proliferation arrest. As for D11, D7 and D8 were secreted into the culture media that was harvested by centrifugation. However, due to lower isoelectric points (6.0 for D7 and 6.3 for D8 versus 8.2 for D11), the pH of the culture media was adjusted from pH 6.5 to pH 8.0 by the addition of an equal volume of tris buffer pH 9.0. This increase in pH simultaneously improved the affinity of His_6 tagged D7 and D8 to the Ni^{2+} IMAC resin and diluted chelating agents present in the serum free media that strip Ni^{2+} IMAC resin.[187] Above pH 7.0, components of the serum free culture media formed a white precipitate that was removed by centrifugation and filtration.[187] Buffered culture media was then loaded directly onto an Ni-NTA column and recombinant protein eluted with an imidazole gradient (Figure 27).

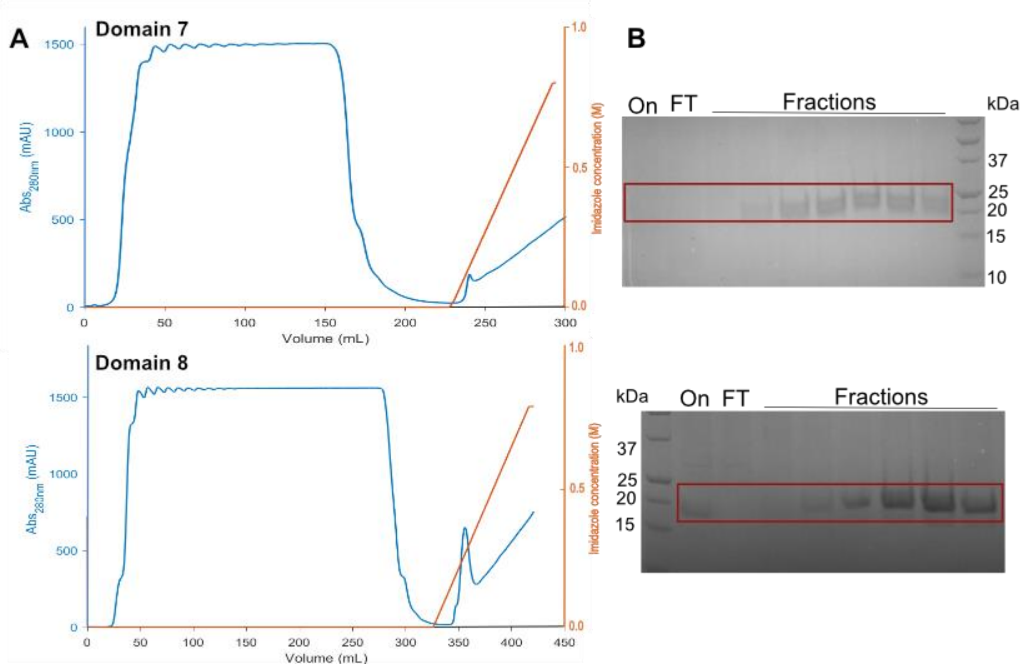


Figure 27: Purification of D7 and D8 from insect culture media. A: Ni^{2+} IMAC of D7 (top) and D8 (bottom) harvested from cell culture medium. Protein was eluted from the Ni-NTA resin over a gradient of imidazole. **B:** SDS-PAGE analysis of Ni^{2+} IMAC confirmed the presence of purified D7 and D8 (bands in the red box, expected molecular weights of 18.2 kDa and 16.6 kDa respectively). 'On' corresponds to media loaded onto the Ni^{2+} column, 'FT' to the flow through and 'Fractions' to the peak fractions following addition of imidazole.

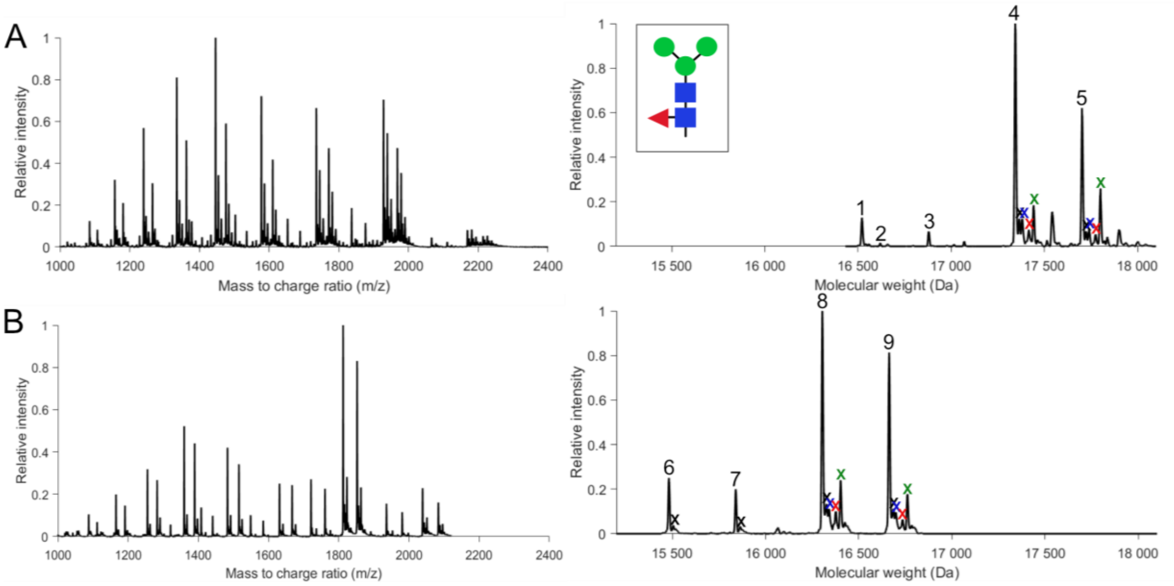
D7 and D8 were isolated to greater than 90 % purity and with a yield of 6 mg/ L and 10-15 mg/ L respectively. The yields were much lower than that observed for D11 (46 mg/ L), most likely due to the presence of one predicted glycosylation site in D8 and two in D7 that mass spectrometry has shown to be glycosylated. The yields of D7 and D8 are in line with literature values. For example, Farrell *et al.* reported a yield of 10 mg/ L for a non-glycosylated, secreted protein (BTF) from Hi5 insect cells and 10-20 mg/ L for a glycosylated, secreted protein (GM-CSF) from Bm5 insect cells.[188]

Biophysical characterisation of D7 and D8

SDS-PAGE analysis and mass spectrometry confirmed successful expression and purification of both D7 and D8. D8 contains one predicted N-linked glycosylation site: N1163 and was characterised first for simplicity. The major species in the D8 mass spectrum (Figure 28A) was derived from truncated D8, termed D8t, that is missing four residues (ETGA) of the N-terminal signal sequence. Mass spectrometry revealed that both D8 and D8t were glycosylated with the same core fucosylated paucimannosidic N-linked glycan (D8 expected molecular mass: 17699.7 Da, observed: 17701.0 Da, D8t expected molecular mass: 17341.3 Da, observed: 17343.0 Da, Figure 28A). This post-translational modification is typical of insect cells (Figure 29).[189]

Despite sample desalting and preparation by methanol-chloroform extraction, several smaller peaks were observed corresponding to common salt adducts (sodium adducts (+23 Da), potassium adducts (+39 Da), imidazole adducts (+68 Da) and phosphate adducts (+98 Da)). This was not improved by using C4 ziptips or dialysis into ammonium acetate. Small peaks were also seen in the deconvoluted spectra for full length non-glycosylated D8 and truncated and full length glycosylated D8 species minus the His_6 tag.

Although N-linked glycans can be removed using anhydrous hydrazine, this has the undesirable effect of cleaving peptide bonds.[190] Thus, for de-glycosylation under native conditions enzymatic methods are often employed. The N-linked glycan on D8 was removed by incubation with Peptide N-glycosidase F (PNGaseF), which cleaves the β-glycosidic linkage between GlcNAc and asparagine, converting asparagine into aspartic acid.[191] (Note PNGaseF cannot cleave N-linked glycans containing an α1-3 fucose residue but α1-3 fucosylation does not occur in the *Spodoptera frugiperda* cell line 21 used here (Figure 29)).[192–194] Following treatment with PNGaseF, D8 was completely de-glycosylated, with the two major species corresponding to de-glycosylated truncated and full length D8 (de-glycosylated D8t expected molecular mass: 16303.3 Da, observed: 16304.9 Da and de-glycosylated D8 expected molecular mass: 16661.7 Da, observed: 16663.0 Da) (Figure 28B).



Spectra	Number	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
D8	1	Truncated D8 minus His6 tag with 1 fucosylated glycan	16520.0	16518.5	+1.5
	2	Full length D8 not glycosylated	16658.0	16660.7	-2.7
	3	Full length D8 minus His6 tag with 1 fucosylated glycan	16879.0	16876.8	+1.2
	4	Truncated D8 with 1 fucosylated glycan	17343.0	17341.3	+1.7
	5	Full length D8 with 1 fucosylated glycan	17701.0	17699.7	+1.3
D8 PNGaseF	6	Truncated D8 minus His6 tag	15481.6	15480.5	+1.1
	7	Full length D8 minus His6 tag	15839.7	15838.8	+0.9
	8	Deglycosylated truncated D8	16304.9	16303.3	+1.6
	9	Deglycosylated full length D8	16663.0	16661.7	+1.3

🟢 Mannose
🟦 GlcNAc
◀ Fucose

Figure 28: ESI-MS of D8. A: Mass spectrometry revealed that D8 was glycosylated with a core fucosylated N-linked glycan. The major species (4) corresponded to truncated D8 in which four N-terminal residues (ETGA) of the signal sequence had been cleaved. **B:** Following treatment with PNGaseF, D8 was de-glycosylated. Species 6 and 7 have lost their His_6 tag. In both spectra, black crosses corresponded to sodium adducts (+23 Da), blue crosses to potassium adducts (+39 Da), red crosses to imidazole adducts (+68 Da) and green crosses to phosphate adducts (+98 Da).

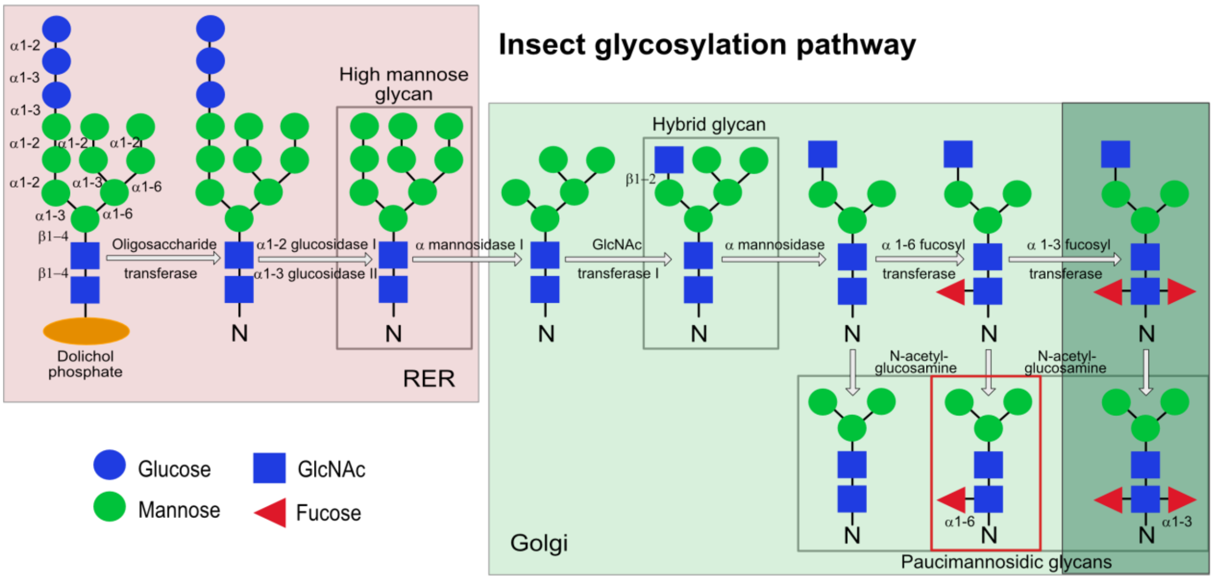
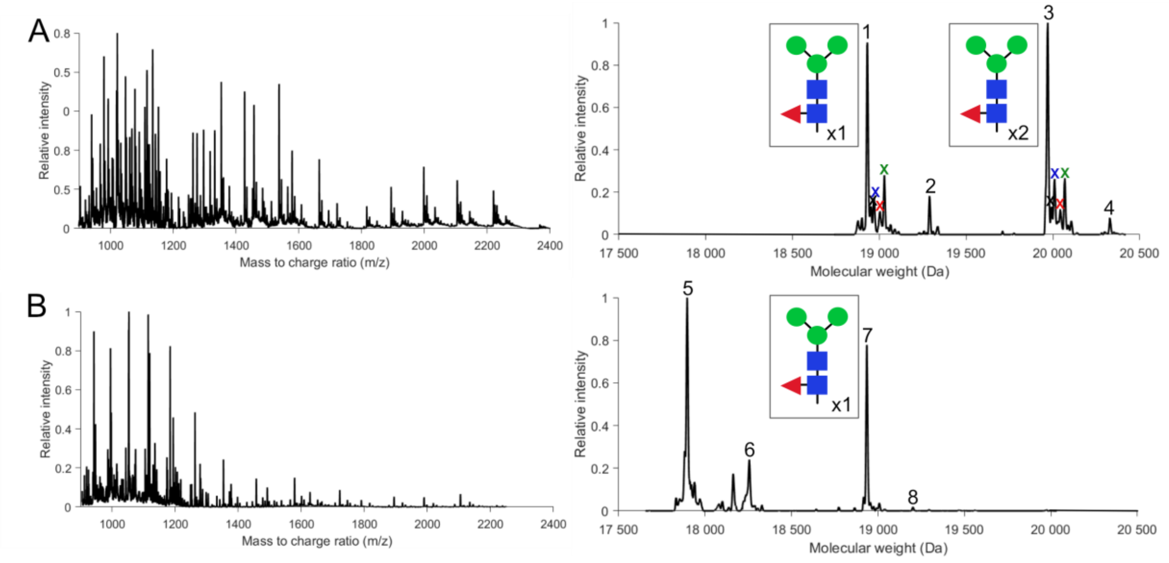


Figure 29: The N-linked glycosylation pathway in insect cells. In the rough endoplasmic reticulum (RER), a precursor N-linked glycan is transferred from the lipid carrier dolichol phosphate to an asparagine residue in the NST sequon (N-X-S/T). The glycan is trimmed and processed in the RER and golgi to give rise to the three classes of insect N-linked glycan: high mannose, hybrid and paucimannosidic. The pathway highlighted in red and pale green occurs in the *Spodoptera frugiperda* cell line (including Sf21 used here).[16,23,189] These Sf21 cells are incapable of α1-3 fucosylation (dark green box).[192,193] The paucimannosidic glycan in the red box was observed on both D7 and D8.

Mass spectrometry revealed that D7 was modified with the same core fucosylated paucimannosidic glycan as D8 (Figures 28A and 30A). The same N-terminal truncation observed for D11 and D8 was also detected for D7 and is termed D7t. D7 contains two predicted N-linked glycosylation sites: N951 and N957, which gave rise to singly and doubly glycosylated species (singly glycosylated D7t expected molecular mass: 18928.1 Da, observed: 18931.0 Da, doubly glycosylated D7t expected molecular mass: 19967.1 Da, observed: 19970.0 Da, D7 singly glycosylated expected molecular mass: 19286.1 Da, observed: 19289.0 Da, D7 doubly glycosylated expected molecular mass: 20325.5 Da, observed: 20327.0 Da). However, whilst ~60 % of D8 was isolated as the truncated form, a greater proportion (~80 %) of D7 was truncated.

PNGaseF de-glycosylation of D7 (Figure 30B) yielded a major species of non-glycosylated D7t (expected molecular mass: 17887.1 Da, observed: 17896.0 Da). A smaller peak was observed for D7t modified with a single fucosylated N-linked glycan (expected molecular mass: 18926.5 Da, observed: 18935.0 Da). The same modifications were also observed with full length D7 (non-glycosylated D7 expected molecular mass: 18248.6 Da, observed: 18255.0 Da, singly glycosylated D7 expected molecular mass: 19288.0 Da, observed: 19294.0 Da). Incomplete de-glycosylation may be due to sub-optimal conditions or steric hindrance of one of the glycosylation sites. N957 is predicted to be exposed on the surface of βA and N951 on the flexible N-terminus/ linker region between D6-D7.



Spectra	Number	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
D7	1	Truncated D7 with 1 fucosylated glycan	18931.0	18928.1	+2.9
	2	Full length D7 with 1 fucosylated glycan	19289.0	19286.5	+2.5
	3	Truncated D7 with 2 fucosylated glycans	19970.0	19967.1	+2.9
	4	Full length D7 with 2 fucosylated glycans	20327.0	20325.5	+1.5
D7 PNGaseF	5	Truncated D7	17896.0	17887.1	+8.9
	6	Full length D7	18255.0	18248.6	+6.4
	7	Truncated D7 with 1 fucosylated glycan	18935.0	18926.5	+8.5
	8	Full length D7 with 1 fucosylated glycan	19294.0	19288.0	+6.0

● Mannose
■ GlcNAc
◀ Fucose

Figure 30: ESI-MS of D7. A: Mass spectrometry revealed that D7 was glycosylated with either one (species 2) or two (species 4) core fucosylated N-linked glycans. The same N-terminal truncation seen in D8 is visible for D7 (species 1 and 3). **B:** Mass spectrum of D7 following treatment with PNGaseF. Truncated and full length D7 with no glycans (species 5 and 6) were visible. However, deconvoluted spectra also contains a peak corresponding to truncated D7 with a single fucosylated glycan (species 7). In both spectra, black crosses corresponded to sodium adducts (+23 Da), blue crosses to potassium adducts (+39 Da), red crosses to imidazole adducts (+68 Da) and green crosses to phosphate adducts (+98 Da).

D7 and D8 were both determined to be monomeric in solution by analytical SEC with an expected and observed monomeric molecular weight of 19.9 kDa and 21.3 kDa respectively for D7 and 17.3 kDa (Mw_{exp}) versus 15.5 kDa (Mw_{app}) for D8 (Figure 31A and Appendix Figure 3). Peak fractions were analysed by SDS-PAGE under denaturing, reducing (R) and non-denaturing, non-reducing (N) conditions (Figure 31A). The slight band shift under native conditions suggested a more compact, disulphide bonded structure consistent with a globular protein fold. Native protein folding of glycosylated D7 and D8 was confirmed by 1D-^{1}H NMR acquired at 700 MHz (Figure 31B) and both species yielded sharp, well resolved and dispersed chemical shift envelopes.

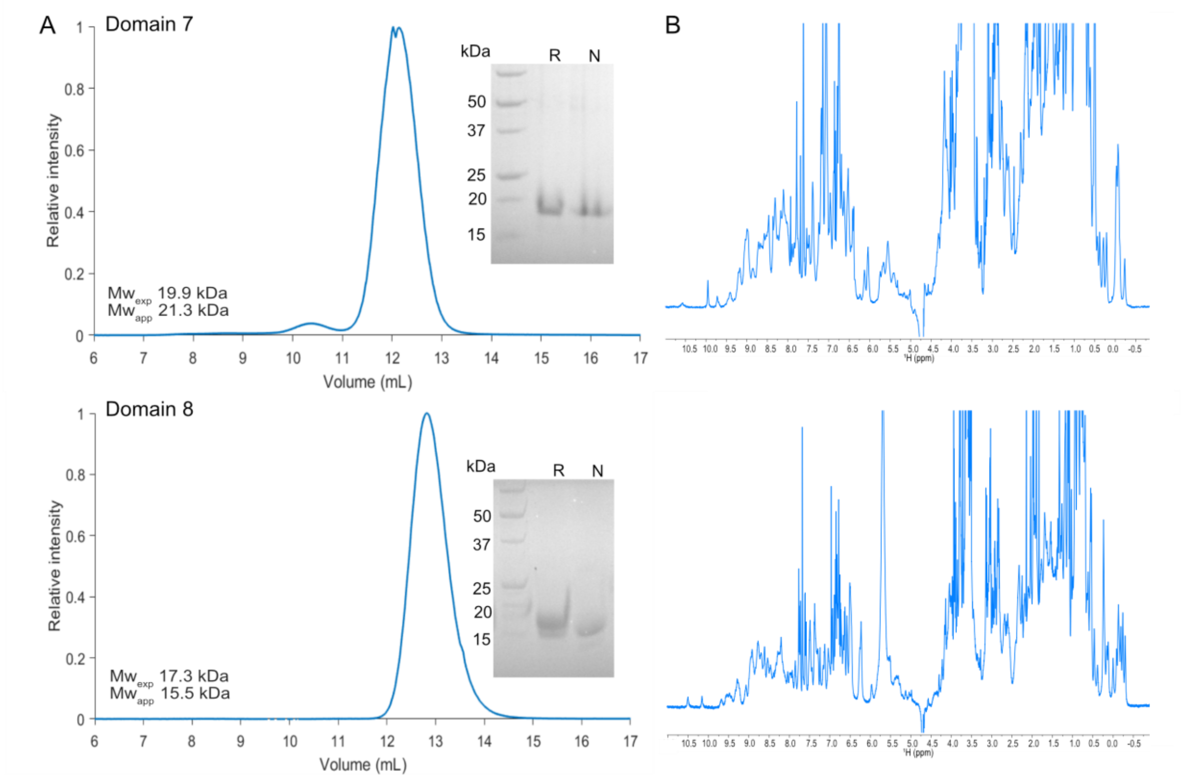


Figure 31: Biophysical characterisation of D7 and D8. A: analytical SEC chromatogram of D7 (top) and D8 (bottom) revealed their monomeric status. SDS-PAGE analysis of the peak under reducing (R) and non-reducing (N) conditions suggested protein folding. **B:** 1D ^{1}H-NMR spectra of glycosylated D7 (top) and D8 (bottom) suggested native protein folding.

Structural characterisation of D7 and D8

D7 and D8 have not been structurally characterised at high-resolution. If high-resolution structures of D7 and D8 could be obtained these would provide structural information on two new domains of the CI-MPR as well as potentially assisting in the phasing and refinement of larger fragments of the CI-MPR, including a dataset for D7-11. Thus, D7 and D8 structure determination was pursued by X-ray crystallography.

Domain 8:

Sparse matrix crystallisation screens were set up with varying concentrations of glycosylated D8. After approximately 6 months small birefringent crystals were seen from a solution of 1.5 M ammonium sulphate, 0.1 M Tris pH 8.5 and 12 % glycerol. These were looped, cryo-cooled and diffraction data collected at Diamond Light Source. D8 (V1082-R1221, 139 amino acids) had crystallised in space group $P12_11$ with two molecules in the asymmetric unit. The diffraction data was poor quality with high anisotropy and possible twinning. Nonetheless, by molecular replacement methods, Dr Chris Williams (University of Bristol) determined the structure of D8 to 2.56 Å resolution (Rwork and Rfree values of 22.3 % and 25.1 % respectively) with 94.9 % of the backbone dihedral angles in favoured regions of the

Ramachandran plot (Appendix Table 4). Electron density was observed for V1082-V1220 of D8 and S1081 from the N-terminal signal sequence. Although ESI-MS revealed D8 to be modified by the addition of an N-linked glycan at N1164 (Figure 28A), no electron density for the glycan was observed. This is likely due to the flexibility of the glycan and the positioning of N1164 towards a solvent channel. Alternatively, the glycan may have been lost with time, allowing eventual crystallisation.

The structure of D8 (Figure 32A) revealed a core β-sandwich structure as previously observed for D1-3, D5 and D11-14 of the CI-MPR.[101,131,136] Two antiparallel β-sheets (βA-D and βE-I) form a flattened, nine-stranded β-sandwich linked by five loop regions (termed AB, CD, FG and HI) which vary in length between four and nine amino acids. The β-sandwich is stabilised by four disulphide bonds: C1084-1125 between the N-terminus and the loop between βB and βC, C1142-C1134 between βC and the preceding loop region, C1177-C1204 between the FG and HI loop and C1190-C1207 at the base of βG and βI (Figure 32A).

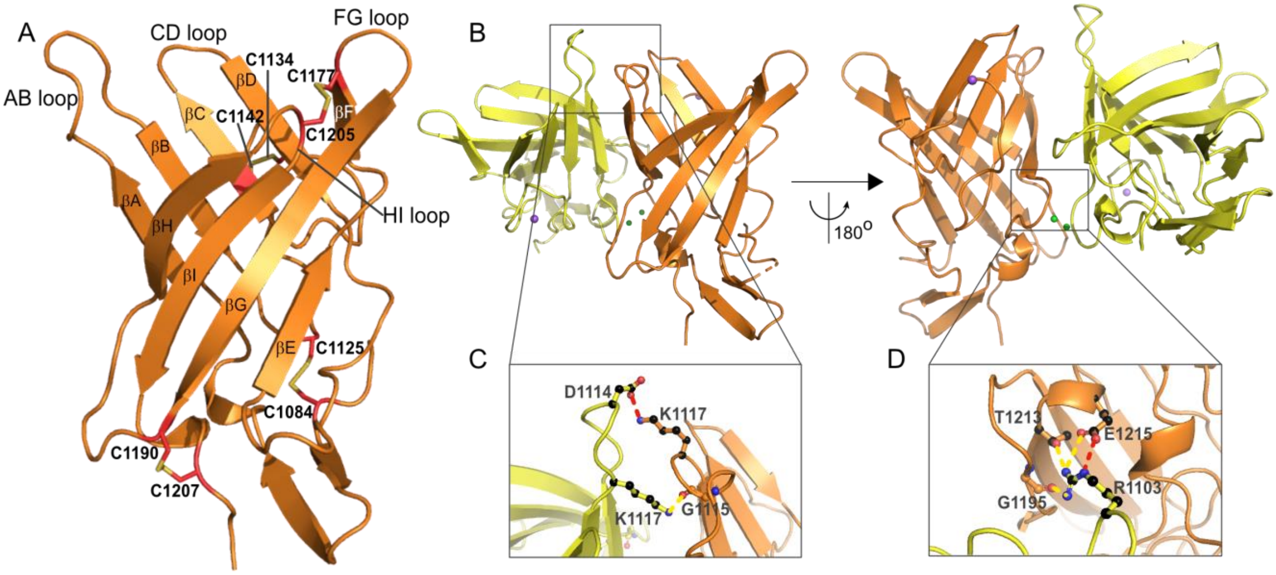


Figure 32: The crystal structure of human CI-MPR D8. A: D8 forms a nine-stranded β-sandwich (βA-I) stabilised by four disulphide bonds (red sticks). **B:** There are two copies of D8 per asymmetric unit. **C**: The D8 homodimer is stabilised by a salt bridge between D1114-K1117 and H-bond between K1117-G1115 in the AB loops. **D:** At the base of D8, dimerisation is stabilised by a salt bridge between R1103-E1215 and H-bonds between R1103-G1195 and R1103-T1213. Salt bridges are shown as red dashed lines and H-bonds as yellow dashed lines. An identical interaction site is seen for R1103 of molecule B.

The D8 crystal structure contains two copies per ASU (Figure 32B) with molecule A having a buried surface area of 682.4 Å^2 (9.2 %) and molecule B 662.9 Å^2 (8.9 %) (as determined by PISA, Krissinel *et al.*).[195] βA-B of molecule A packs against βA-B of molecule B. Specifically, a series of three hydrogen bonds (K1117-G1115, R1103-G1195 and R1103-T1213) and two salt bridges (D1114-K1117 and R1103-E215) form at the bases of the β-sandwich and the AB loops.

D8 has no obvious ligand binding site, indeed there is no known ligand for D8. Analysis of the AB, CD, FG and HI loops of D8 reveals that residue Y1207 on βI is in the same position as Y421 of D3, Y714 of D5, Y1351 of D9 and Y2233 of D15 that are all part of the conserved 'QREY' motif in these MRH domains. However, the remaining residues of this motif are not present in D8, where **L**1143, **F**1182, **Q**1202 replace **QRE**Y residues respectively (Figure 33A), suggesting that D8 is unlikely to possess any hitherto unreported weak M6P binding. Further analysis of the M6P binding sites of D3, D5, D9 and the loops at the top of D8 demonstrate that, D8 does not form a pocket with a charge distribution observed in the MRH domains (Figure 33B). Similarly, the region adjacent to the AB, CD, FG and HI loops of D8 does not form a hydrophobic patch as observed in the D11 IGF2 binding site (Figure 33B).

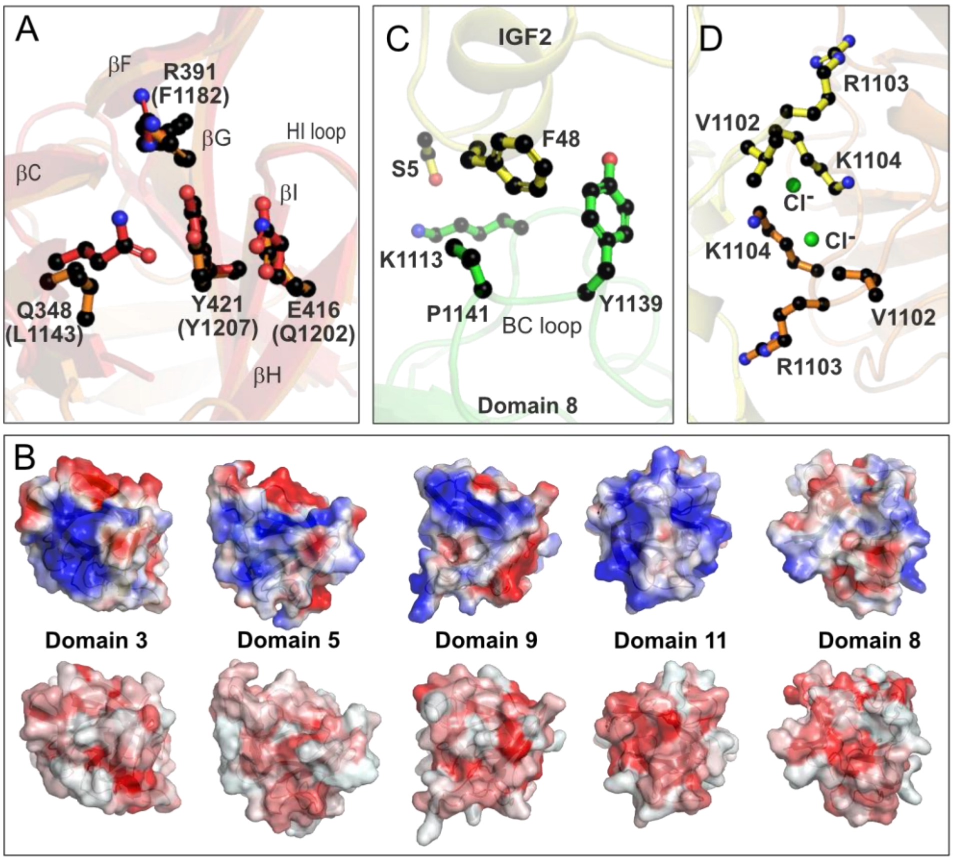


Figure 33: The potential D8 binding site. A: Superimposition of D3 (red) and D8 (orange). Residues of the 'QREY' M6P binding motif of D3 (Q348, R391, E416, Y421) are labelled and shown as red balls-and-sticks. The analogous residues of D8 (L1143, F1182, Q1202, Y1207) are labelled in brackets and shown as orange balls-and-sticks. **B:** Comparison of the binding sites (viewed looking down onto the AB, CD, FG and HI loops) of D3, D5, D9, D11 and D8 (PDB 1SYO, 2KVB, 6Z32, 1GP0 and 6Z31). Top: positively charged residues coloured blue and negatively charged residues red (range +2 to -2) as determined using the APBS software.[196] Bottom: hydrophobic residues coloured red according to the normalised hydrophobicity scale.[197] **C:** K1113 at the N-terminus of D8 and T1139 and P1141 of D8 BC loop interact with S5 and F48 of IGF2 (PDB 6UM2). **D:** The crystal structure of single human D8 also contains two chloride ions (green spheres) that are bound by the positively charged side chain of K1104 and the backbone amide groups of V1102 and R1103 of each D8 molecule.

However, the recent low-resolution (4.3 Å) cryo-EM structure of bovine D4-14 in complex with IGF2 at pH 7.4 (PDB 6UM2) reveals that D8 may play a role in IGF2 binding.[128] K1113 N-terminal to βA and Y1139 and P1141 both on the BC loop of D8 interact with α2 of IGF2 (Figure 33C). These interactions were confirmed by mutagenesis, with IGF2 mutant F48D resulting in a 20-30 % reduction in CI-MPR-IGF2 binding.[128]

The crystal structure of D8 also contains two chloride ions (confirmed by CheckMyMetal validation server [198]) bound at the base of the β-sandwich (Figure 33D). Although, chloride is not present in the crystallisation reservoir, D8 was purified into Tris pH 8 buffer containing 150 mM sodium chloride. Analysis of protein structures in the PDB shows it is not uncommon to find two chloride ions in structures containing 2 copies of identical sequence in a single ASU.[199] On average, such chloride ions have a total exposed surface area of ~29 Å^2 (22 % of their total surface area).[199] The chloride ions in the D8 crystal structure have exposed surface areas of 25.1 Å^2 (20.2 %) and 24.4 Å^2 (19.5 %).[195] The chloride ions are bound by the positively charged side chain of K1104 and the backbone amide groups of V1102 and R1103 (Figure 33D). This is in line with the observation that positively charged amino acids (R, H, K) and polar amino acids (N, Q, S, T) dominate in protein-chloride interactions.[199]

Domain 7:

Four commercial sparse matrix crystallisation screens (Morpheus, JCSG Plus, PACT Premier and Structure Screen I and II (Molecular Dimensions)) spanning 384 discrete crystallisation conditions were set up with D7 at a range of concentrations at 4 °C and 25 °C. However, birefringent protein crystals failed to form. This was attributed to the presence of the two predicted N-linked glycosylation sites at N951 and N957 that mass spectrometry revealed to be occupied by fucosylated paucimannosidic glycans (Figure 30A).

Crystallisation of glycosylated proteins is a recognised challenge with only ~ 10 % of all PDB entries being glycosylated (as of 2007).[25] The flexibility and heterogeneity of N-linked glycans hampers crystallisation and therefore the most common strategy for crystallisation of glycosylated proteins is simply to remove them. This was demonstrated by Fan *et al.* who found that although glycosylated follicle-stimulating hormone (FSH) crystallised, it diffracted to only 9 Å.[200] Meanwhile, FSH partially de-glycosylated with endoglycosidases F2 and F3 diffracted to 2.9 Å (PDB 1XWD).[200]

Mass spectrometry has shown that D7 can be partially de-glycosylated by PNGaseF (Figure 30B). However, complete removal of the N-linked glycan can result in protein aggregation.[25,201] Indeed, fully de-glycosylated, soluble D7 could not be obtained using PNGaseF when the reaction was performed under native conditions. A second enzyme, endo β-N-acetylglucosaminidase H (EndoH), which cleaves between the two GlcNAc residues of the chitobiose core,[25,202] was also tested but again failed to produce soluble, de-glycosylated D7 suitable for crystallisation. Co-crystallisation of D7 and EndoH also failed to produce crystals.

2.4. Conclusions

As expected, due to the presence of disulphide bonds and N-linked glycosylation sites, D7 and D8 expressed insolubly in bacteria, forming dense inclusion bodies of protein. Despite an extensive screen of conditions, the D11 *in vitro* refold protocol could not be adapted for D7 or D8.

With this in mind and the aspiration to structurally characterise larger, multi-domain fragments of the CI-MPR, eukaryotic expression protocols were established. Using the well characterised D11 as a control protein, insect cells proved to be suitable expression hosts capable of expressing natively folded, single CI-MPR domains for biophysical characterisation (that includes mass spectrometry, analytical SEC and 1D ^{1}H-NMR) and structural characterisation (by X-ray crystallography).

Although, D7 failed to crystallise (likely due to the presence of two N-linked glycans), D8, which contains only one N-linked glycosylation site, crystallised and its structure was determined to 2.5 Å resolution. This is the first high-resolution structure of human D8 and can hopefully be used in the determination of larger fragments of the CI-MPR such as D7-11.

3. Structural characterisation of D9

3.1. <u>Introduction and aims</u>

With a protocol established for the expression of single CI-MPR domains (chapter 2), the focus turned to the selective, high-affinity and independent M6P binding domain D9. To date, there exists no high-resolution structure of human D9, which has not been expressed in isolation. D5, D11, D15 and now D7 and D8 have each been expressed individually: D5 and D11 in both bacteria and yeast,[112,124,131,177] D15 in bacteria [130] and D7 and D8 in insect cells (chapter 2). However, despite extensive previous work in the Crump group, soluble folded D9 has not been obtained. *E. coli* expression of D9 from four species (*Homo sapiens, Bos taurus, Lemur catta* and *Ornithorhynchus anatinus)* yielded inclusion bodies that require refolding. A range of *in vitro* refolding protocols yielded soluble protein but in no case was a well dispersed, native-like ^{1}H or ^{1}H-^{15}N HSQC NMR spectrum observed that would indicate correctly refolded protein. Mammalian expression of glycosylated D9 in HEK293 cells did eventually produce soluble, disulphide bridged, glycosylated protein but again this yielded poor ^{1}H-NMR spectra suggesting this construct was not suitable for structural studies. *In vitro* de-glycosylation of the protein failed to improve the NMR spectra.[70]

There are conflicting reports by others regarding the success in expressing the sugar binding domains D3 and D9 in isolation. For example, attempts by Dahms *et al.* have similarly failed to obtain D3 and D9 in isolation.[67] However, Chavez *et al.* and Hancock *et al.* have expressed bovine D3 and D9 alone in *Pichia pastoris*, demonstrated the specificity of D9 for M6P mono-esters and determined the binding affinity of D9 as 75 ±11 nM for GAA mono-ester and 0.3 ±0.1 nM for M6P-tagged β-glucuronidase.[137,139]

Based on the observation that a D9 protein could not be isolated alone for structural characterisation, this work aimed at expressing a stable, D9 containing, multi-domain construct. Due to their large size and the presence of three to four disulphide bonds per domain, eukaryotic expression hosts are required for the expression of multi-domain CI-MPR constructs. The most common eukaryotic expression hosts are the mammalian cell lines human embryonic kidney (HEK) cells and Chinese hamster ovary (CHO) cells, the insect cell lines *Spodoptera frugiperda* (Sf21) and *Trichoplusia ni* 5B1-4 (High five) cells, and the yeast strain *Pichia pastoris*.[203] Olson *et al.* have expressed the three N-terminal domains of the CI-MPR extracellular region, D1-3, in High five insect cells and determined the structure by X-ray crystallography (PDB 1SYO/ 1SZ0 and 1Q25).[103,136] Similarly, Brown *et al.* have recombinantly expressed several D11 containing multi-domain constructs including D11-12, D11-14 and D11-13 in complex with IGF2 for X-ray crystallography (PDB 2V5N, 2V5O, 2V5P) using the mammalian CHO cell line.[101] However, due to availability and the establishment of a protocol for expression of single CI-MPR domains in insect cells (chapter 2), Sf21 insect cells were employed here.

A range of di-, tri- and larger multi-domain constructs encompassing D9 were designed and sub-cloned in parallel for expression in Sf21 cells (Table 8). The most successful constructs (D9-10 and D7-10) are discussed here.

Construct	**PCR amplified?**	**Ligated into pFastBac?**	**Correct sequence?**	**Bacmid?**	**Virus?**	**Protein?**
D8-9	🟩	🟩				
D8-10	🟩	🟩				
D8-11	🟩	🟩				
D9-10	🟩	🟩	🟩	🟩	🟩	🟩
D9-11	🟩	🟩				
D6-10	🟩	🟩				
D6-11	🟩	🟩				
D7-10	🟩	🟩	🟩	🟩	🟩	🟩
D7-12	🟩	🟩				
D7-14	🟩					

Table 8: The progress of D9 containing multi-domain constructs. Constructs were PCR amplified from a gene encoding human CI-MPR D1-15 and ligated into a pFastBac vector modified with an N-terminal signal sequence and C-terminal His$_6$ tag. Ligation was confirmed by sanger sequencing (Genewiz) (column 3). D9-10 and D7-10 were the first to be successfully sub-cloned and recombinant bacmid generated. Transfection of Sf21 insect cells with recombinant bacmid resulted in recombinant baculovirus and recombinant D9-10 and D7-10 protein.

3.2. Domains 9-10

Expression and purification of D9-10

DNA encoding a D9-10 (852 bp) di-domain construct (Appendix Section 8.1**.**) was amplified by PCR from a synthetic gene encoding D1-15 (GeneArt, ThermoFisher Scientific) and subcloned, using restriction endonucleases EcoRI and HindIII, into a modified pFastBac transfer vector with an N-terminal RPTPμ signal sequence and C-terminal His_6 tag. EMBacY DH10 *E. coli* were transformed with D9-10pFastBac and D9-10 protein was expressed in Sf21 insect cells using the protocols established in chapter 2 for single domain constructs.

The optimal harvest time was determined by SDS-PAGE analysis and YFP emission to be 48 hours after the day of proliferation arrest (Figure 34A, 34B). However, SDS-PAGE revealed a substantial amount of D9-10 protein was retained in the insect cells (in both the soluble and insoluble cell fractions, Figure 34A). This is likely due to the presence of 16 cysteines that form 8 disulphide bonds. As for previous constructs, D9-10 was harvested from culture media by Ni^{2+} IMAC (Figure 34C, 34D) and was isolated to greater than 90 % purity and with a yield of 7-13 mg/ L (varying with batch).

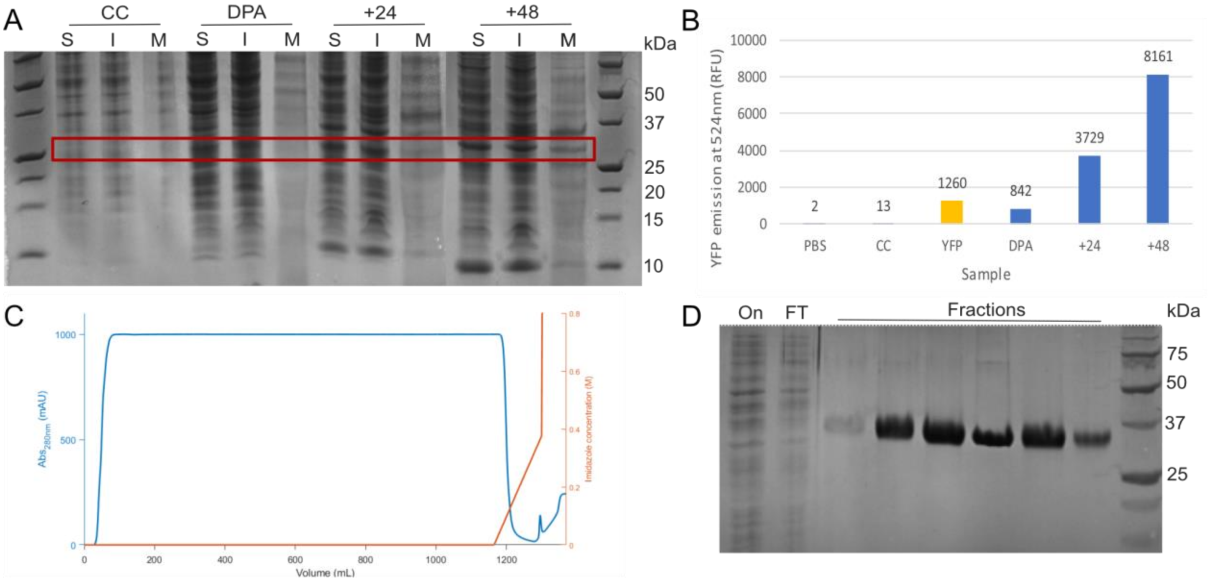


Figure 34: Expression and purification of D9-10. A: SDS-PAGE analysis of D9-10 expression showing D9-10 is secreted into culture media. D9-10 (within the red box) has an expected molecular weight of 33.6 kDa. **B:** YFP emissions confirming YFP expression. **C:** Ni^{2+} IMAC of D9-10 harvest from the culture media. **D:** SDS-PAGE analysis of IMAC purification. 'On' corresponds to media loaded onto the column, 'FT' to the flow through and 'Fractions' to the peak fractions eluted following imidazole addition.

Characterisation of D9-10

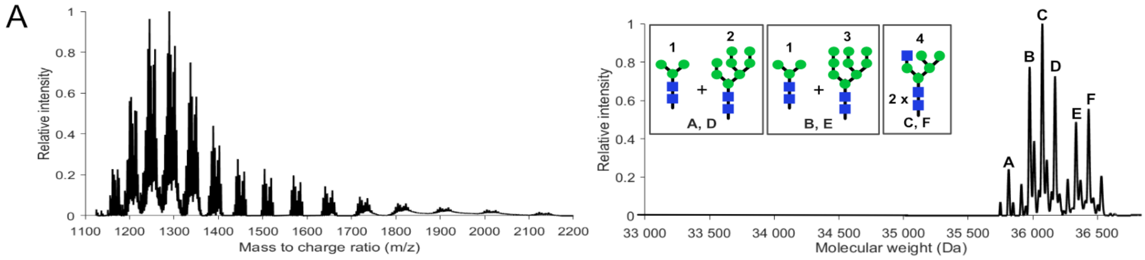
Expression and purification of D9-10 was confirmed by SDS-PAGE (Figure 34D) and mass spectrometry (Figure 35). The ESI-MS m/z envelope (Figure 35A) revealed the presence of multiple species with masses exceeding full length D9-10 (33,588.0 Da). This is consistent with modification of both predicted N-linked glycosylation sites of D9 (N1246 and N1312) with a mixture of typical insect cell derived glycans (Figure 36).[16,23,189]

With a mass of 36,172.1 Da, peak D was close to the expected mass of D9-10 plus $GlcNAc_4Man_{11}$ (36,168.6 Da) (Figure 35A). Similarly, peak E (36,333.9 Da) was close to the expected mass of D9-10 plus $GlcNAc_4Man_{12}$ (36,330.6 Da). N-linked glycosylation is a highly heterogenous modification with incomplete mannosidase processing giving rise to intermediate glycans with a range of mannose residues. Although the exact glycan structures are unknown, glycans 1-3 (Figure 35A) represent possible configurations corresponding to modification of D9 at both sites with a low (1) and high (2 and 3) mannose glycan. The higher mass peak F (36,431.9 Da) was close to the expected mass of D9-10 plus $GlcNAc_6Man_{10}$ (36,428.7 Da). This may correspond to dual modification with hybrid glycan 4 (Figure 35A), a typical insect cell derived, N-linked glycan (Figure 36).

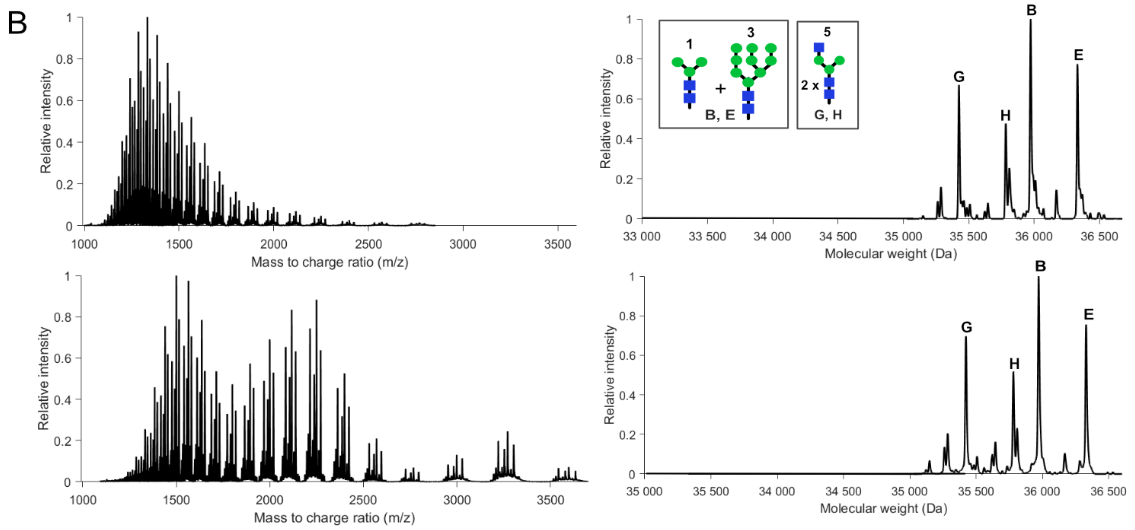
The same glycan modifications (peaks A, B and C) were also observed on a second species attributed to an N-terminal truncation of D9-10 in which four amino acids (ETGA) from the N-terminal signal sequence had been lost resulting in 358.3 Da mass difference. This N-terminal truncation was previously observed for D11, D7 and D8 (chapter 2). Small peaks corresponding to salt adducts were not removed by dialysis into 0.1 % formic acid or 100 mM ammonium acetate. Batch to batch preparations showed slight variability in the distribution of glycans (Figure 35B).

Under native conditions the m/z envelope has a more bimodal distribution and contains several smaller charge states corresponding to higher molecular weight species (Figure 35B). However, the deconvoluted spectra is unchanged, suggesting D9-10 may exist as a dimer in the gas-phase.

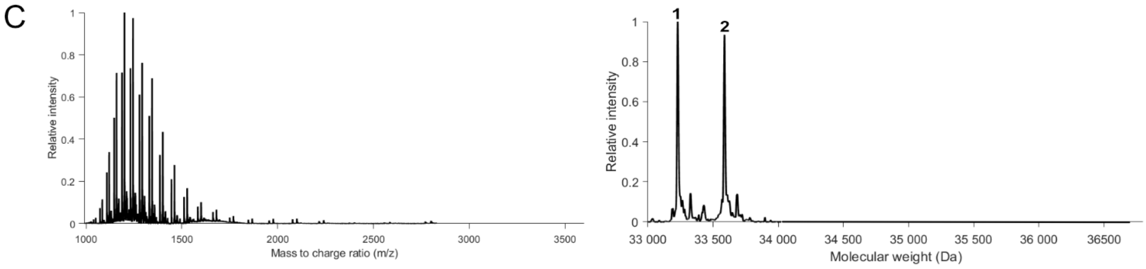
Treatment of D9-10 with PNGaseF confirmed complete loss of glycosylation leaving truncated and full length D9-10 de-glycosylated at both sites (truncated (1) expected molecular weight: 33,231.4 Da, observed: 33,231.0 Da, full length (2) expected molecular weight: 33,589.7 Da, observed: 33,589.0 Da) (Figure 35C).



Label	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
A	Truncated D9-10 with glycans 1 and 2	35813.8	35810.3	+3.5
B	Truncated D9-10 with glycans 1 and 3	35975.6	35972.3	+3.3
C	Truncated D9-10 with two glycan 4	36073.6	36070.4	+3.2
D	Full length D9-10 with glycans 1 and 2	36172.1	36168.6	+3.5
E	Full length D9-10 with glycans 1 and 3	36333.9	36330.6	+3.3
F	Full length D9-10 with two glycan 4	36431.9	36428.7	+3.2



Spectra	Label	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
D9-10 denatured	B	Truncated D9-10 with glycans 1 and 3	35972.0	35972.3	-0.3
	E	Full length D9-10 with glycans 1 and 3	36331.0	36330.6	+0.4
	G	Truncated D9-10 with two glycan 5	35454.0	35422.2	+1.8
	H	Full length D9-10 with two glycan 5	35782.0	35780.5	+1.5
D9-10 native	B	Truncated D9-10 with glycans 1 and 3	35981.0	35972.3	+8.3
	E	Full length D9-10 with glycans 1 and 3	36333.0	36330.6	+3.0
	G	Truncated D9-10 with two glycan 5	35420.0	35422.2	-2.2
	H	Full length D9-10 with two glycan 5	35779.0	35780.5	-1.5



Label	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
1	Truncated D9-10 deglycosylated	33231.0	33231.4	-0.4
2	Full length D9-10 deglycosylated	33589.0	33589.7	-0.7

Figure 35: Mass spectrometry of recombinant human D9-10 from insect cells. A: ESI-MS of D9-10 used in X-ray crystallography reveals the glycosylation status of the two N-linked glycosylation sites of D9 (N1246 and N1312). Species D and species E, with masses 36,172.1 Da and 36,333.9 Da were close to the expected masses of 36,168.6 Da and 36,330.6 Da consistent with the presence of a low (1) and high (2 and 3) mannose glycan typical of on-pathway insect glycosylation processing.[16,23,189] Similarly, the higher mass species F (36,431.9 Da) closely matched the expected mass for dual modification with hybrid glycan 4. The same glycan modifications (A, B and C) were also observed on a second species attributed to an N-terminal truncation of D9-10 in which four amino acids (ETGA) from the N-terminal signal sequence had been lost resulting in 358.3 Da mass difference. **B**: ESI-MS analysis of D9-10 batch used in SAXS studies shows the slight variability in glycosylation patterns. Higher m/z peaks under native conditions (bottom) suggest D9-10 may exist as a dimer in the gas phase. **C:** ESI-MS analysis of D9-10 de-glycosylated by PNGaseF.

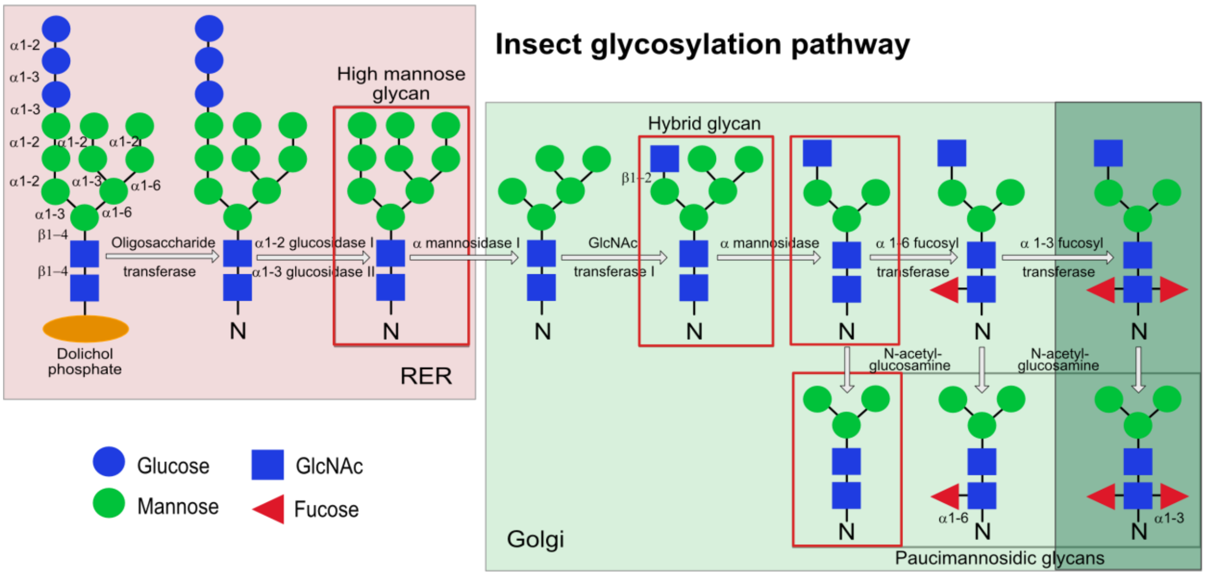


Figure 36: The N-linked glycosylation pathway in insect cells. In the rough endoplasmic reticulum (RER), a precursor N-linked glycan is transferred from the lipid carrier dolichol phosphate to an asparagine residue in the NST sequon (N-X-S/T). The glycan is trimmed and processed in the RER and golgi to give rise to the three classes of insect N-linked glycan: high mannose, hybrid and paucimannosidic. The pathway highlighted in red and pale green occurs in *Spodoptera frugiperda* cell line (including Sf21 used here).[16,23,189] These cells are incapable of α1-3 fucosylation (dark green box).[192,193] The glycans in the red boxes are observed on D9-10.

Native protein folding was assessed by SDS-PAGE and 1D ^{1}H-NMR (Figure 37). SDS-PAGE analysis of D9-10 run under denatured and reduced (R) versus non-denatured and non-reduced (N) conditions confirmed a band shift of ~3-4 kDa (Figure 37). The lower mass under native conditions strongly suggested the presence of intact disulphide bridges and the formation of a compact protein fold. Protein folding was further confirmed by 1D-^{1}H NMR analysis at 700 MHz. A 90 μM solution of D9-10 dissolved in 25 mM Tris pH 7.5, 150 mM NaCl revealed a well dispersed chemical shift profile (Figure 37). The amide region contained numerous chemical shifts ~8.5 ppm that would be characteristic of the presence of β-strands. In addition, the presence of upfield chemical shifts between 0.5 and -0.5 ppm most likely corresponds to methyl groups experiencing the ring current shifts from aromatic residues, again indicative of the formation of a hydrophobic core of the protein. The amide chemical shift region (10-6 ppm) is not as well defined when compared to that of individual domains but is judged to be good for a glycosylated protein with a molecular weight of 35-36 kDa.

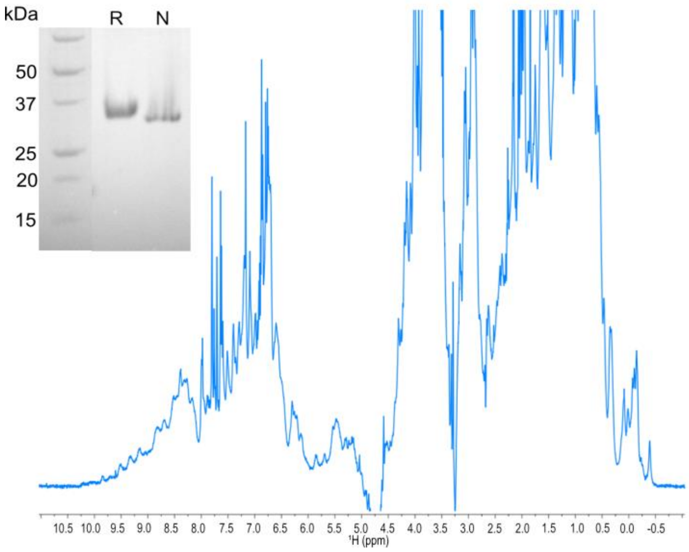


Figure 37: SDS-PAGE and ^{1}H-1D NMR to assess D9-10 folding. The slight band shift between D9-10 under reducing (R) and non-reducing (N) conditions suggests a compact structure stabilised by disulphide bonds. Protein folding was confirmed by 1D ^{1}H-NMR collected at 700 MHz with 90 μM D9-10 dissolved in 25 mM Tris pH 7.5, 150 mM NaCl.

Crystallisation and structure determination of D9-10

Sparse matrix crystallisation screens were set up with glycosylated D9-10. After ~7 weeks crystals were observed in PACT Premier condition A6: 0.1 M SPG (succinic acid, sodium dihydrogen phosphate monohydrate, glycine) pH 9, 25 % PEG 1500. These were looped, dipped in 25 % glycerol and cryo-cooled in liquid nitrogen. Diffraction data was collected on beamline I04 at Diamond Light Source. The construct crystallised in space group $P2_12_12_1$ with 1 molecule in the asymmetric unit. The structure of D9-10 was determined to 1.5 Å (Rwork and Rfree values of 19.9 % and 22.8 % respectively) with 98.6 % of the backbone dihedral angles in allowed regions of the Ramachandran plot (Appendix Table 5) by molecular replacement using homology models of D9 and D10.

The structure of D9-10 (Figure 36A) is formed from two well-defined domains of similar size, comprised of D9 residues D1225-D1365 and D10 residues L1366-K1509 and is the first reported to encompass the specific M6P binding site of D9. There was no electron density for amino acids V1222-G1224 at the N-terminus of D9 or S1510 at the C-terminus of D10. Electron density was observed for one N-linked glycan on N1312 of D9 (Figure 38) which is positioned through association with D9 of a symmetry related partner (Figure 38B). Modelling $GlcNAc_2Man_4$ yielded the best fit to the electron density (Figure 38B). The second glycan predicted to be linked to N1246, which is positioned next to a solvent channel, was not observed.

Both D9 and D10 form a core β-sandwich structure stabilised with four disulphide bonds as previously observed for D1-3, D5 and D11-14 of the CI-MPR.[101,131,136] Two antiparallel β-sheets (βA-D and βE–I) form a flattened, nine-stranded β-sandwich linked by four loop regions (termed AB, CD, FG and HI) which vary in length between four and seven amino acids. D9 and D10 each contain four disulphide bonds. These are: C1227-C1262 (at the N-terminus), C1270-C1282 (in the BC loop and βC), C1319-C1349 (in the FG and HI loops), C1333-C1361 (in βG and the linker region) in D9 and C1369-C1408, C1420-C1427, C1461-C1494, C1476-C1506 in D10.

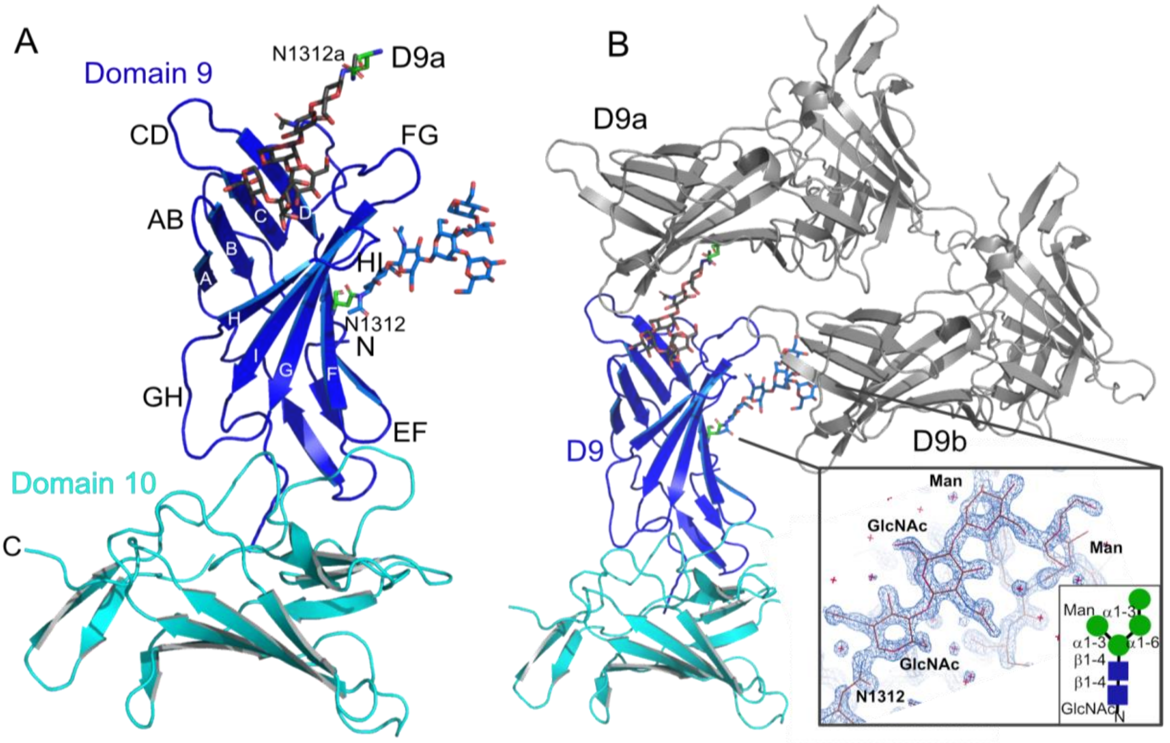


Figure 38: Crystal structure of human CI-MPR D9-10. A: The structure of D9-10 at 1.5 Å resolution. Each domain forms a nine-stranded flattened β–sandwich stabilised by four disulphide bonds. N1312 of D9 was glycosylated with $GlcNAc_2Man_4$. D9 (blue) and D10 (cyan) are orientated at ~90° to one another. **B:** D9 (blue) binds an N-linked $GlcNAc_2Man_4$ glycan of a neighbouring D9 molecule (D9a, grey). Meanwhile the glycan of D9 (blue) is bound by a third D9 molecule (D9b, grey). The inset shows the structure and fit to electron density (2Fo-Fc map at 1.16 electrons per $Å^3$ (2.8σ)) of the $GlcNAc_2Man_4$ glycan at N1312 of D9.

D9 and D10 are arranged at approximately 90° to one another (Figure 38A) with the βA-D surface of D10 interacting with the loops at the base of D9. The interface between D9 and D10 buries surface areas of ~600 $Å^2$ (8 %) and 625 $Å^2$ (8 %) respectively (as calculated by PISA[195]). The electron density observed in this region was well defined. The interface is stabilised by a combination of hydrophobic interactions, hydrogen bonding and the disulphide bridge between C1333-C1361 at the base of D9 βG and the linker region (Figure 39A). H1234 in the β-hairpin at the N-terminus of D9 sits within a pocket formed by the proline rich BC loop of D10 (Figure 39A) and is sandwiched between P1422 on D10 and P1362 on D9 with separations consistent with the formation of CH-π interactions.[204,205] R1233 also appears to pack against P1419 yielding a CH-π interaction. Y1387 of D10 also interacts with F1364 and R1335 to stabilise the opposing side of the interface. Hydrogen bonded residues include residues at the base of D9 and residues in the BC loop of D10 (N1306-G1416, R1356-S1388, R1233-G1416/Q1414, D1365-S1410) (Figure 39B).[70]

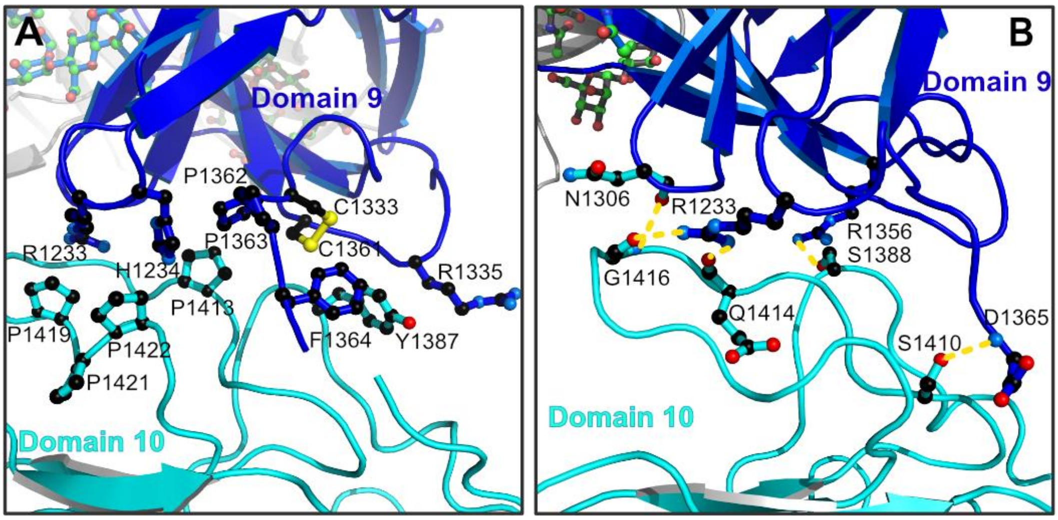


Figure 39: The interface between D9 and D10. A: H1234 in the β-hairpin at the N-terminus of D9 sits within a pocket formed by the proline rich BC loop of D10 and is sandwiched between P1422 on D10 and P1362 on D9 with separations consistent with the formation of CH-π interactions.[204,205] R1233 (left) also appears to pack against P1419 yielding a CH-π interaction. Y1387 (right) of D10 also interacts with F1364 and R1335 of D9 to stabilise the opposing side of the interface. **B:** H-bonds (yellow lines) form between the base of D9 and residues in the BC loop of D10 (N1306-G1416, R1356-S1388, R1233-G1416/Q1414, D1365-S1410).

D9-10 oligomeric state

Having observed a glycan bridging D9s in the D9-10 crystal structure, the oligomeric state of D9-10 was investigated further by native mass spectrometry, analytical SEC and SAXS. As mentioned, under native conditions the m/z envelope of D9-10 contained several smaller charge states corresponding to higher molecular weight species (Figure 35B). The deconvoluted spectra was unchanged, suggesting D9-10 may exist as a dimer in the gas-phase.

Analytical SEC of D9-10 at pH 7.5 gave an apparent molecular weight of ~64 kDa, close to the glycosylated, dimeric molecular weight of ~72 kDa (0.9 times the dimeric molecular weight, Figure 40, Appendix Figure 5). The broadness of the peak may be attributed to the two flexible N-linked glycans of D9 and the possible exchange between monomeric and dimeric forms. Upon addition of 100-fold excess M6P, D9-10 eluted only slightly later, giving an apparent molecular weight of ~57 kDa (0.8 times the dimeric molecular weight) (Figure 40A). De-glycosylated D9-10 (using PNGaseF) eluted even later and corresponded to ~39 kDa (0.6

times the dimeric molecular weight), which closely matches the expected de-glycosylated molecular weight of monomeric D9-10 (~33 kDa) (Figure 40A). Similarly to the glycosylated sample, there was little change upon addition of 100-fold excess M6P to de-glycosylated D9-10, which corresponded to 43 kDa (0.7 times the dimeric molecular weight, Figure 40A).

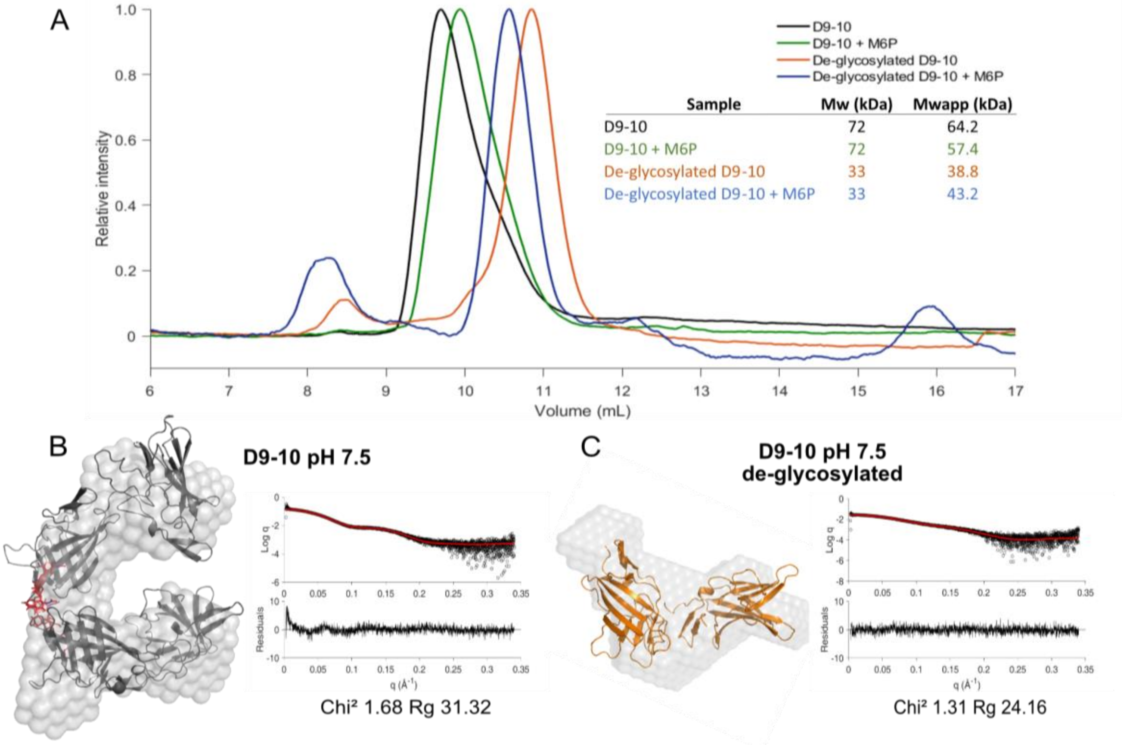


Figure 40: D9-10 is dimeric in solution. A: Analytical SEC of D9-10 at pH 7.5 glycosylated (black) and de-glycosylated (orange). Upon incubation with 100-fold excess M6P (green) D9-10 elutes later. **B:** SEC-SAXS of glycosylated D9-10 at pH 7.5. Glycans are shown in red. **C:** SEC-SAXS of D9-10 at pH 7.5 de-glycosylated by PNGaseF. For each SEC-SAXS result: *Ab initio* DAMMIN bead density shape envelope of D9-10 is represented as a surface (light grey) overlaid with a 1-state-optimised model. For each model, the *I*(q) vs q plot for a 1-state optimised model (red line) against the experimental scattering data (black) is shown with the residual difference plot underneath.

Small angle X-ray scattering (SAXS) was employed to further study D9-10 in solution and confirmed a folded flat entity with a calculated molecular mass associated with a dimer (apparent molecular mass 70.5 kDa (as calculated by SAXSMoW 2.0) expected dimeric molecular mass 70.8-71.6 kDa, Appendix Table 6). *Ab initio* shape envelope modelling of D9-10 revealed a dimer with symmetric cross-binding of the two N-linked glycan moieties (Figure 40B). Fitting required flexibility within the linker region between D9 and D10 (residues 1358-1364) and flexibility of the N-linked glycan at N1312 but not the principle D9-10 interface. A χ^2 value of 1.68 for a 1 state model of D9-10 indicates that this model is a good fit to the experimental data (Figure 40B).[70]

De-glycosylation of D9-10 by PNGaseF resulted in monomerisation with an apparent molecular mass of 33.7 kDa (expected monomeric molecular mass 33.2-33.6 kDa, as calculated by SAXSMoW 2.0, Appendix Table 6). Furthermore, the scattering data could be modelled to a shape envelope consistent with the D9-10 crystal structure (χ^2 value of 1.31, Figure 40C, Appendix Table 6).

D9 binding site and comparison to other MRH domains

D9 adopts the conserved MRH fold previously observed for CI-MPR D3 and D5, and the CD-MPR.[90,101,131] Sequence alignments and mutagenesis studies have shown that these MRH domains contain a four residue 'QREY' binding site motif present as Q1283, R1325, E1345 and Y1351 in human D9 (Figure 41A).[67] Q1283 on βC forms a hydrogen bond (H-bond) with the 2'OH of the terminal mannose residue, while E1345 and Y1351 on βH and βI respectively can H-bond to the 3'OH and 4'OH. From the FG loop, the charged guanidinium group of R1325 interacts with the 6'OH of the terminal mannose residue. H1320, previously identified as an essential residue for M6P binding,[132] also faces into the binding site of D9 and may engage the 6'OH of the same mannose residue (Figure 41A). The N-linked glycan branches, with the α1-3 branch of the glycan sitting inside the M6P binding site. Y1255 at the top of βB may form H-bonds with the 4'OH of this branched mannose residue and 1'OH of the terminal mannose.

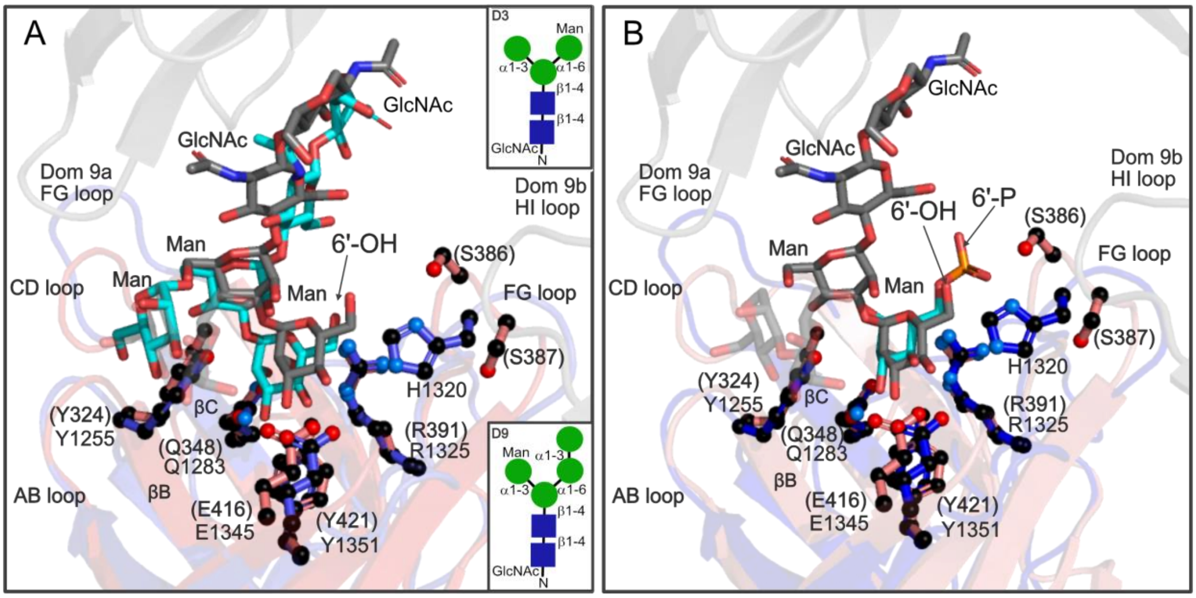


Figure 41: Comparison of the binding sites of CI-MPR MRH D3 and D9. A: The structure of the N-linked glycan (grey sticks) observed in the D9-10 crystal structure (blue cartoon) is superimposed with the structure of the N-linked glycan (cyan sticks) observed bound to bovine D3 (red cartoon, PDB 1Q25). Sugar binding residues of D9 are labelled and shown as blue balls-and-sticks. Homologous residues in D3 are shown as red balls-and-sticks and labelled in brackets. The insets show the structures of the N-linked glycans present at N76 of D3 (top) and N1312 of D9 (bottom). **B:** The structure of the N-linked glycan (grey sticks) observed in the D9-10 crystal structure (blue cartoon) is superimposed with the structure of M6P (cyan sticks) observed bound to bovine D3 (red cartoon, PDB 1SZ0). H1320 in the FG loop of D9, which restricts M6P di-ester binding, is replaced by S386 and S387 in the FG loop of D3.

As a P-type lectin, D9 is not expected to bind mannose. Indeed, glycan microarray analyses have shown no interaction between the CI-MPR and glycans with a terminal mannose residue.[82,133] However, crystal structures of bovine D1-3 also revealed occupancy of the M6P binding site either by M6P or a non-phosphorylated branched glycan that varied with protein preparation (PDB 1Q25 and 1SZ0/1SYO respectively).[103,136] Interactions were observed between the conserved D3 'QREY' motif and three terminal mannose residues of a branched $GlcNAc_2Man_3$ N-linked glycan on N76 of a neighbouring D3 (Figure 41A).[103] One mannose residue (of the α1-3 branch) sits within the M6P binding site interacting with the M6P binding residues Q348, R391, E416, Y421. Meanwhile, the mannose residue of the α1-6 branch sits outside the M6P binding site. Y324 on βB forms hydrogen bonds with the 1'OH of the terminal α1-3 mannose, 4'OH of the branching mannose residue and the 6'OH of the α1-6 branched mannose residue. The M6P binding pocket of D3 is shallow, with only 16 % of the three mannose residues being solvent-inaccessible.[103]

A comparison between D9-10 and D1-3 structures reveals a common orientation of either the glycan (Figure 41A) or M6P (Figure 41B) and amino acid residues shown to be essential for M6P binding.[67] H1320 in D9 replaces S386 and S387 residues in the equivalent position of D3 that lie either side of the histidine side chain and, assuming a similar orientation of M6P, H1320 would form a direct favourable charge-charge interaction with the phosphate group (Figure 41B). The orientation of the histidine sidechain in D9 is homologous to the equivalent histidine residue (H105) from the bovine CD-MPR high-resolution structure that binds the phosphate of M6P (PDB 2RL8)[206] and again there is close structural conservation of residues in the M6P binding site (Figure 42A).

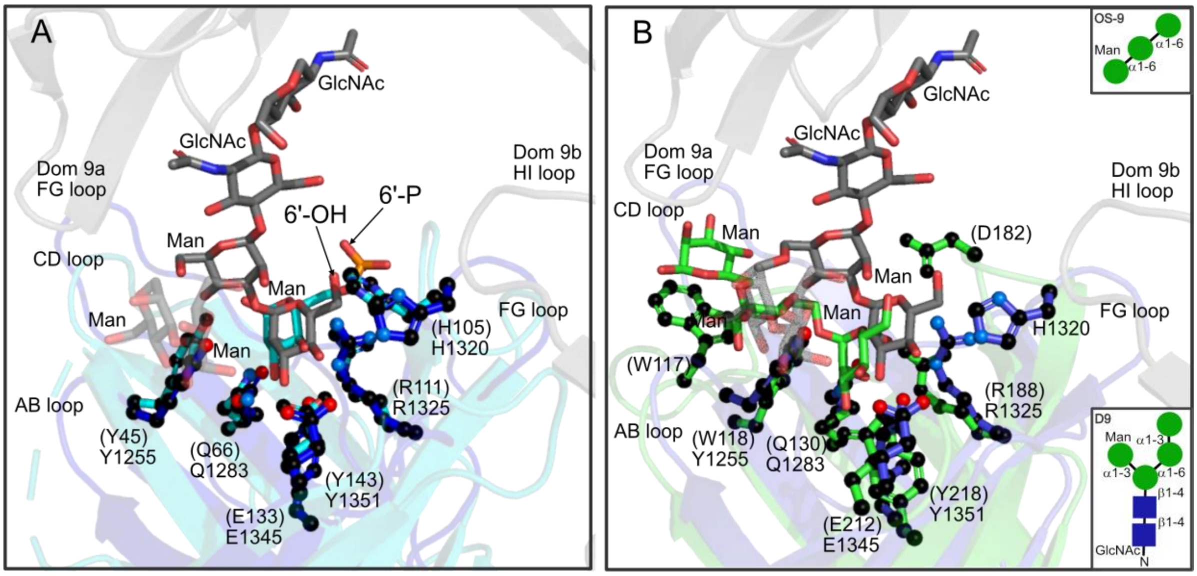


Figure 42: Comparison of the binding sites of the MRH domains of CI-MPR D9, CD-MPR and OS-9. A: The structure of the N-linked glycan (grey sticks) observed in the D9-10 crystal structure (blue cartoon) is superimposed with the structure of the M6P (cyan sticks) observed bound to bovine CD-MPR (cyan cartoon, PDB 2RL8). Sugar binding residues of D9 are labelled and shown as blue balls-and-sticks. Homologous residues in the CD-MPR binding site are shown as cyan balls-and-sticks and labelled in brackets. The critical H1320 (H105) adopts a similar juxtaposition to the 6'OH or 6'-P in both structures underlying its critical role in binding M6P mono-esters. **B:** The structure of the N-linked glycan (grey sticks) observed in the D9-10 crystal structure (blue cartoon) is superimposed with the structure of Man$_3$ (green sticks) observed bound to OS-9 (green cartoon, PDB 3AIH). H1320 in the FG loop of D9, which restricts M6P di-ester binding, is replaced by D182 in the FG loop of OS-9. Y1255 of βB of D9 is replaced by a di-tryptophan motif (W117 W118) in OS-9. The insets show the structures of the N-linked glycan bound by D9 (bottom) and OS-9 (top).

Both D9 and D3 also show close structural homology to OS-9 (RMSD values of 2.4 Å and 2.1 Å respectively over backbone atoms), a lectin that recognises two α1-6 linked mannose residues on the C-arm of high-mannose type N-linked glycans on ER-associated degradation (ERAD) substrates.[207] Q1283, R1325, E1345 and Y1351 of D9 all have conserved counterparts in OS-9 that form interactions with the hydroxyl groups of the bound mannose residues (Figure 42B). The exceptions are Y1255 of βB which is substituted as part of a crucial di-tryptophan glycan binding motif and H1320 of the FG loop which is substituted with D182 that binds the 6'OH of Man(B) and may prevent M6P binding.[66] The switch therefore to a polar residue (D3 S386/S387) or positively charged residue (D9 H1320) appears to be associated with the presence of the negatively charged M6P moiety.[70]

The interaction between the FG and HI loops defines an important region of the sugar binding pocket across the MRH domains. In D9, H1320 lies adjacent to the disulphide bridge (C1319-C1349) that connects the FG and HI loops and together these residues occlude one side of the binding pocket, preventing GlcNAc-M6P di-ester binding (Figure 43A). This region of the binding pocket was proposed to be important for the recognition of M6P di-esters by bovine

D5.[131] The FG and HI loops of D5 are not connected by a disulphide bridge and lack bulky residues, creating an open binding pocket that can easily accommodate the GlcNAc residue of the M6P di-ester (Figure 43B). Attempting to model in an equivalent di-ester into the D9 structure creates clear steric clashes (Figure 43A). The D9 binding site closely resembles the deep narrow pocket that is formed by the surface of the CD-MPR M6P specific binding pocket (Figure 43D) which is defined by an almost identical packing of H105, R111 and an equivalent disulphide bridge connecting the FG and HI loops. Although lacking the bulkier histidine residue, the disulphide bridge also partially occludes the binding pocket in D3 (Figure 43C). Taken together, occlusion of this region of the MRH binding pocket appears to be a common mechanism for ensuring specificity for M6P mono-esters.[70]

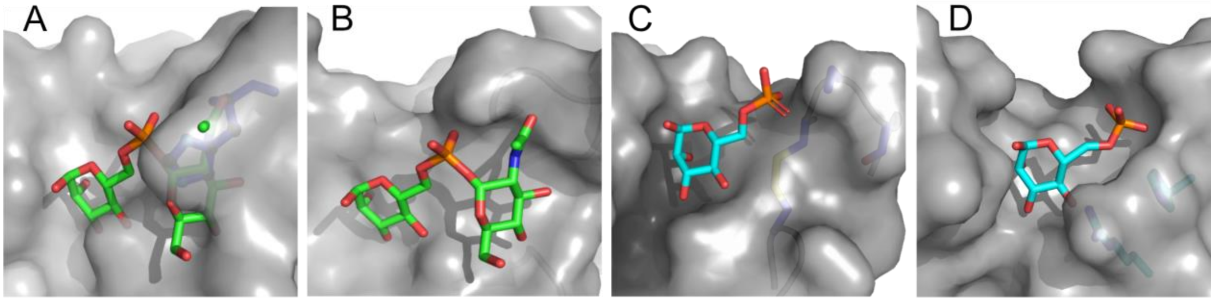


Figure 43: Comparison of the binding pockets of the P-type lectin MRH domains. A: The binding surface of D9 (grey) with the modelled GlcNAc-M6P di-ester (green sticks). H1320 (blue sticks) on the FG loop occludes di-ester binding. **B:** The binding surface of bovine D5 (grey, PDB 2KVB) with a modelled GlcNAc-M6P di-ester (green sticks) bound in the more open binding pocket. D5 lacks a disulphide bridge between the FG and HI loops and a bulky histidine on the FG loop. **C:** The binding surface of bovine D3 (grey) with bound M6P mono-ester (cyan, PDB 1SZ0). S386 and S387 of the FG loop and the disulphide bridge between C385 and C419 of the FG and HI loops respectively are shown as sticks. **D:** The binding surface of the CD-MPR (grey) with bound M6P mono-ester (cyan, PDB 2RL8). H105 and R111 are shown as sticks.

M6P binding

Having observed occupancy of the D9 carbohydrate binding site by a mannosylated glycan in the crystal structure and solution (Figure 38, Figure 40B), attention turned to demonstrating that D9 of the D9-10 construct has retained M6P binding ability. However, despite co-crystallisation of D9-10 with M6P and soaking of D9-10 crystals in an excess of M6P, a structure of M6P bound D9-10 could not be determined. Therefore, SAXS was performed on D9-10 in the absence and presence of M6P.

As mentioned above (Figure 40), glycosylated D9-10 is dimeric in solution through the bridging N-linked glycan of N1312. Enzymatic de-glycosylation of D9-10 resulted in monomerisation (Figure 40C, Figure 44C). However, the shape parameters (R_g and D_{max}) and shape envelope associated with de-glycosylated D9-10 were largely unaffected by the addition of a 100-fold excess of M6P, with the structure again fitting to the D9-10 crystal structure (χ^2 value of 1.15) (Figure 44D, Appendix Table 6, Appendix Figure 8). Thus, if D9 has bound M6P, there are no significant structural changes to the arrangement of D9-10.

The shape parameters (R_g and D_{max}) and shape envelope associated with glycosylated D9-10 were similarly unaffected by the addition of a 100-fold excess of M6P, with the structure remaining dimeric and again fitting to the D9-10 dimer arrangement of D7-11 (χ^2 value of 1.75) (Figure 44B, Appendix Table 6, Appendix Figure 8). This suggests a relatively tight interaction between D9 and the mannosylated N-linked glycan, that could not be disrupted by addition of M6P monosaccharide at pH 7.5.

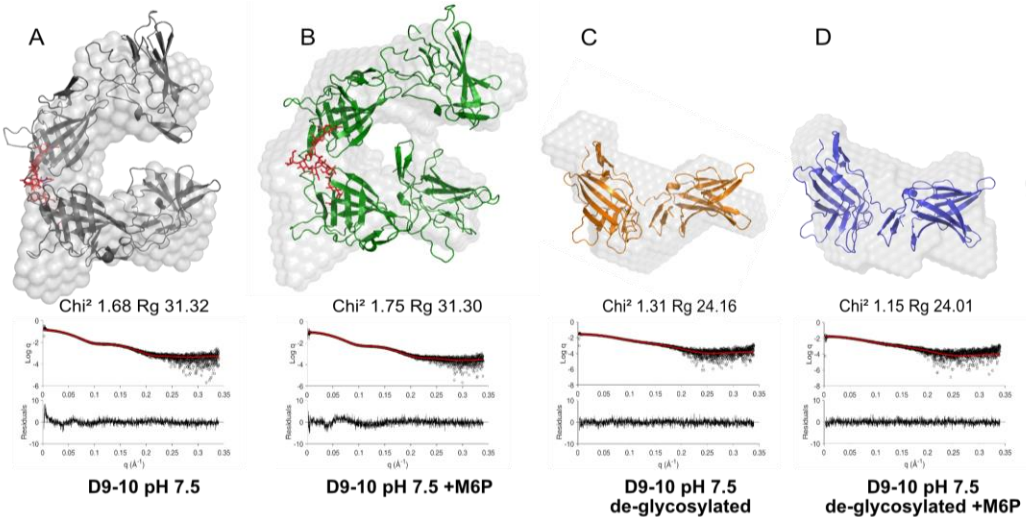


Figure 44: SEC-SAXS of D9-10 at pH 7.5. For each result: *Ab initio* DAMMIN bead density shape envelope of D9-10 represented as a surface (light grey) overlaid with a 1-state-optimised model. For each model, the *I*(q) vs q plot for a 1-state optimised model (red line) against the experimental scattering data (black) with the residual difference plot shown underneath. For each structure, glycans are shown in red with the χ^2 value of the fit and R_g value of the model shown. **A:** D9-10 pH7.5. **B:** D9-10 pH7.5 plus 100-fold excess M6P. **C:** D9-10 pH 7.5 de-glycosylated. **D:** D9-10 pH 7.5 de-glycosylated plus 100-fold excess M6P.

The CI-MPR has been demonstrated to have a higher affinity for M6P when presented within an N-linked glycoprotein versus M6P monosaccharide. For example, using SPR, Olson *et al.* found that D1-15 binds M6P monosaccharide with micromolar affinity (K_D 7 μM), while Chavez *et al.* found that D1-15 binds M6P mono-ester glycoprotein with nanomolar affinity (K_D 4.5 nM).[130,139] The interaction between D9 and M6P monosaccharide has not been studied. However, Chavez *et al.* and Hancock *et al.* have demonstrated by SPR that bovine D9 expressed alone in *Pichia pastoris* bound the M6P mono-ester tagged glycoproteins GAA and β-glucuronidase with nanomolar affinity (75 ±11 nM and 0.3 ±0.1 nM respectively).[137,139] Thus, for comparison to the literature and for physiological relevance, an M6P-tagged glycoprotein is desirable. Chavez *et al.* describe a complex protocol for obtaining GAA tagged with either M6P mono-ester or GlcNAc-M6P di-ester using the mammalian cell line CHO-K1 treated with the N-linked glycosylation inhibitor kifunensine, mammalian expressed GlcNAc phosphotransferase, UCE and a phosphatase spPAP.[139] However, these methods were not available to us. Thus, in the absence of an M6P-tagged glycoprotein, 1D ligand observed NMR experiments were employed with M6P monosaccharide and D9-10.

Ligand observed NMR is often used in fragment-based drug discovery to identify hit compounds as it requires very little protein and is capable of screening multiple compounds simultaneously.[208] Two ligand observed NMR methods were employed here: saturation transfer difference (STD) spectroscopy and water-ligand observed via gradient spectroscopy (WaterLOGSY).

In STD (developed by Mayer and Meyer in the late 1990s [209,210]), two spectra are collected. In the first spectrum aliphatic methyl groups of the protein are irradiated.[211] Cross-relaxation spin diffusion ensures that this saturation is spread throughout the protein.[209–211] Upon protein binding, this spin polarisation is transferred to protons of the ligand, giving rise to a positive signal.[209–211] The second spectrum collected is a reference spectrum whereby an empty spectral region is irradiated.[209–211] The reference spectrum is subtracted from the STD spectrum giving rise to an STD difference spectrum that shows only signals corresponding to the protons of ligands that interact with the protein.[209–211]

In the second method used here, WaterLOGSY (developed by Dalvit *et al.* [212,213]), the bulk water is irradiated. Through intermolecular NOE and chemical exchange this magnetisation is transferred to the protein and then to bound ligand.[214] However, the ligand may also be magnetised through interaction directly with the bulk water. Thus, similarly to STD, a reference spectrum must also be collected in the absence of protein.[214] Ligands that interact directly with the bulk water have positive NOE, giving rise to a negative peak, while ligands that interact with the protein have negative NOE, giving a positive peak in the WaterLOGSY spectrum.[208,210]

Figure 45 shows the STD and WaterLOGSY experiments of D9-10 and the control ligands mannose, glucose and glucose 6-phosphate (G6P) which are not known to be bound by the CI-MPR.[82,133] For each ligand, the WaterLOGSY contains only a single positive peak at ~4.7 ppm that corresponds to water. Protons of the sugar display negative peaks suggesting that, as expected, there is no interaction between D9-10 and mannose (green panel), glucose (red panel) or G6P (blue panel). This is supported by a lack of peaks in the STD spectra (Figure 45). Similar results were obtained for de-glycosylated D9-10 with mannose (Figure 45D, 45E).

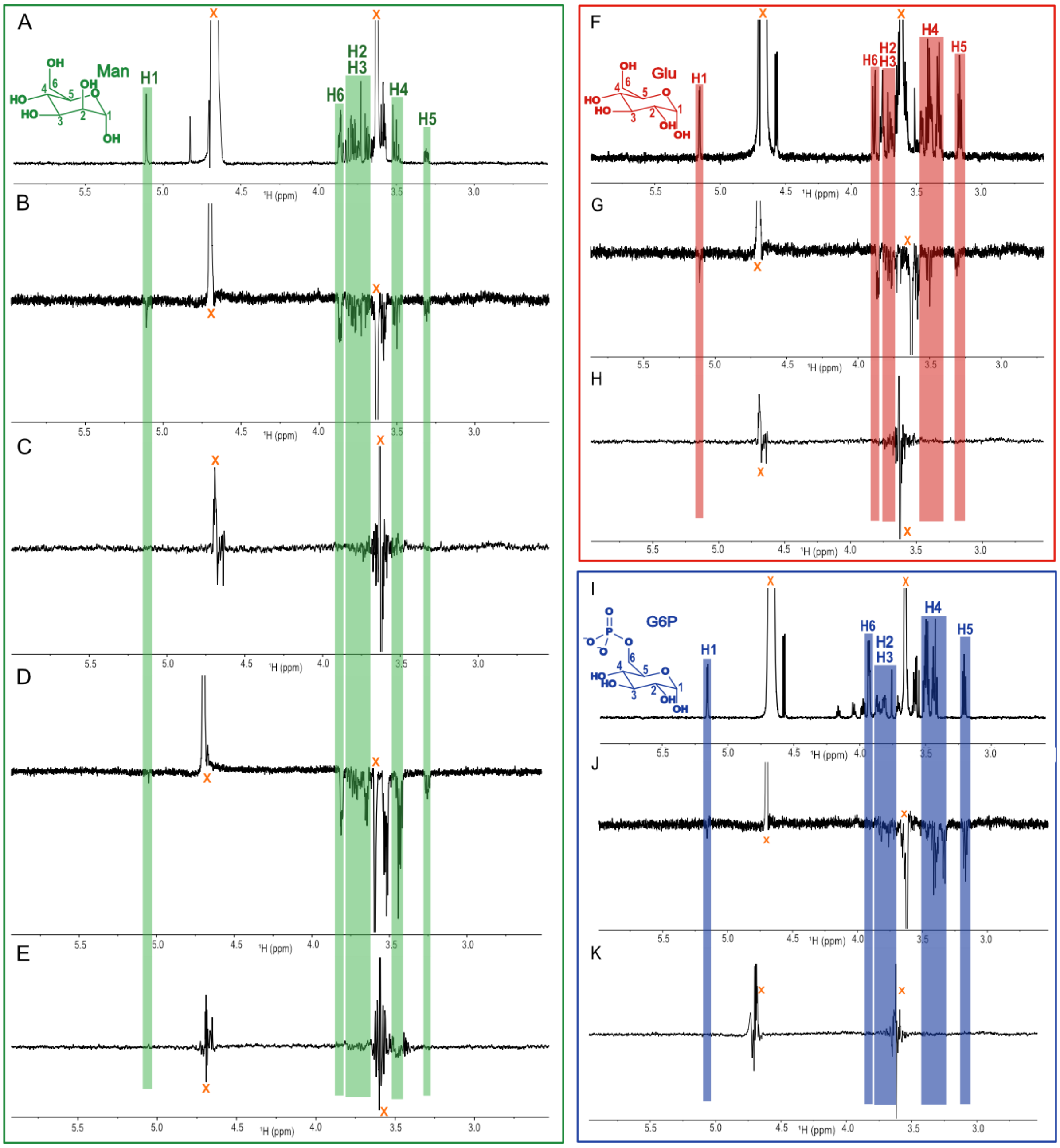


Figure 45: Control saturation transfer NMR experiments. A: CPMG spectrum of mannose. **B:** WaterLOGSY spectrum of glycosylated D9-10 plus mannose. **C:** The STD difference spectrum of glycosylated D9-10 plus mannose. **D:** WaterLOGSY spectrum of de-glycosylated D9-10 plus mannose. **E:** The STD difference spectrum of de-glycosylated D9-10 plus mannose. **F:** CPMG spectrum of glucose**. G:** WaterLOGSY spectrum of glycosylated D9-10 plus glucose. **H:** The STD difference spectrum of glycosylated D9-10 plus glucose. **I:** CPMG spectrum of glucose 6-phosphate (G6P). **J:** WaterLOGSY spectrum of glycosylated D9-10 plus G6P. **K:** The STD difference spectrum of glycosylated D9-10 plus G6P. Antiphase peaks corresponding to sugar protons (~3.4-4ppm and 5.1 ppm) in the WaterLOGSY and absence of positive peaks in the STD difference spectra suggest that D9-10 does not bind mannose, glucose or G6P. Peaks marked with an orange cross correspond to water (~4.7 ppm) and tris (~3.6 ppm). STD spectra were collected with an on-resonance frequency of 0.58 ppm and off-resonance frequency of -28 ppm. All spectra were collected at 700 MHz by Dr Chris Williams with 20 μM D9-10 and 2 mM sugar in 25 mM Tris, 150 mM NaCl pH 7.4, 60 % D_2O.

Saturation transfer experiments were also performed on de-glycosylated D9-10 in the presence of M6P (Figure 47C, 47D). Except for a very small positive peak at ~5.1 ppm corresponding to the anomeric proton, the proton peaks of M6P are negative in the WaterLOGSY (Figure 47C). While the WaterLOGSY result is largely inconclusive, the STD shows possible binding. The anomeric proton, along with protons H2-4, gives a weak positive signal in the STD spectrum (Figure 47D). The crystal structure of D9-10 shows that of the conserved sugar binding residues, Q1283 on βC, E1345 on βH, Y1351 on βI are positioned to interact with the hydroxyl groups at positions 2, 3 and 4 respectively (Figure 46).

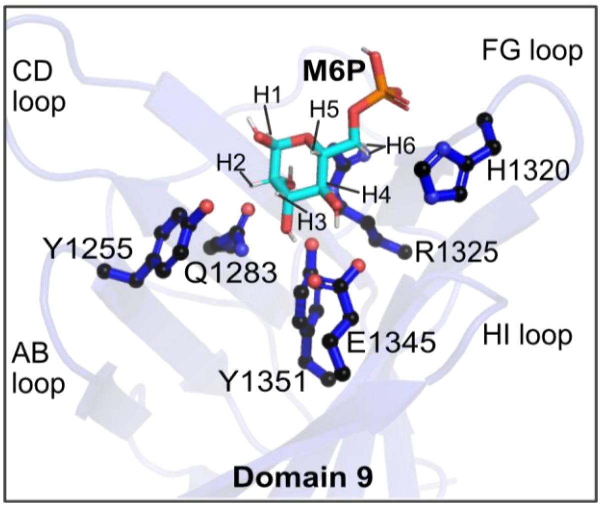


Figure 46: The binding site of D9 (blue) with M6P docked (cyan sticks). Key sugar binding residues of D9 are shown as balls-and-sticks (blue). Y1255 on βB, Q1283 on βC, E1345 on βH, Y1351 on βI and R1325 and H1320 on the FG loop are positioned to interact with the hydroxyl groups at positions 1, 2, 3, 4 and 6 respectively.

Ambiguity may arise from the fact that ligand observed NMR is only suitable for a narrow range of binding affinities.[214] WaterLOGSY and STD are only capable of detecting ligands that bind with millimolar to micromolar affinity (K_D mM-μM) due to their slow exchange with the protein (slow k_{off}).[214] Thus, de-glycosylated D9-10 may be binding M6P monosaccharide too tightly or too weakly to observe here.

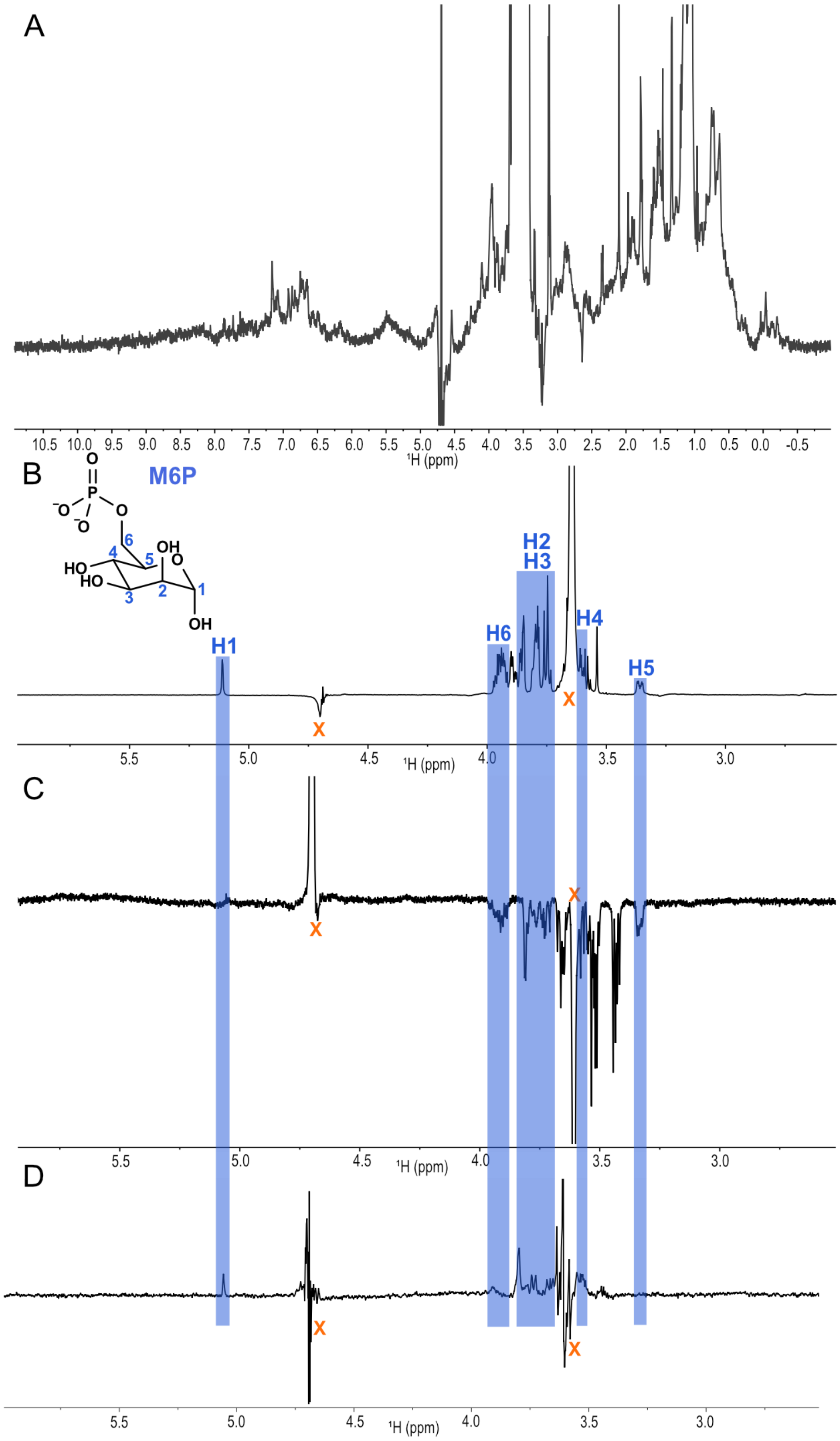


Figure 47: NMR experiments to study de-glycosylated D9-10 M6P binding ability. A: 1D ^{1}H-NMR spectrum of 20 µM de-glycosylated D9-10 plus M6P confirming that D9-10 is folded after de-glycosylation. **B:** 1D ^{1}H-NMR spectrum of M6P. **C:** WaterLOGSY of de-glycosylated D9-10 plus M6P. **D:** STD difference spectrum of de-glycosylated D9-10 plus M6P with an on-resonance frequency of 0.58 ppm and off-resonance frequency of -28 ppm. Shifts corresponding to protons of M6P are highlighted in blue. Peaks marked with an orange cross correspond to water (~4.7 ppm) and tris (~3.6 ppm). All spectra were collected at 700 MHz by Dr Chris Williams with 20 µM de-glycosylated D9-10 and 2 mM M6P in 25 mM Tris, 150 mM NaCl pH 7.4 in 60 % D_2O.

Nonetheless, identical WaterLOGSY and STD experiments were also performed on glycosylated D9-10 in the presence of M6P (Figure 48). Positive WaterLOGSY signals were observed for the anomeric proton of M6P along with H2, H3 and H6 (Figure 48C). However, H4 has a negative peak and H5 has no observable peak. In a control WaterLOGSY experiment of M6P in the absence of protein, the protons of M6P, which are in fast exchange with the protons of water, do not give rise to positive peaks, indicating these WaterLOGSY signals observed in the presence of protein are not false positives. Furthermore, the STD spectrum closely matches the WaterLOGSY spectrum, with clear peaks for the anomeric, H2 and H3 protons. This suggests some interaction between D9-10 and M6P. The crystal structure of D9-10 shows that Y1255 on βB, Q1283 on βC, E1345 on βH, Y1351 on βI and R1325 and H1320 on the FG loop are positioned to interact with the hydroxyl groups at positions 1, 2, 3, 4 and 6 respectively (Figure 46).

To determine the binding affinity of M6P under these conditions, a WaterLOGSY titration experiment was performed. However, the low concentrations of M6P required (50 μM to obtain a 1:5 ratio of protein: ligand) could not be detected over a reasonable time frame using the micro-cryoprobe equipped 700 MHz spectrometer. Furthermore, Huang *et al.* have demonstrated that, due to the ligand re-binding the protein, determination of binding affinity by WaterLOGSY is affected by protein concentration, with higher K_D values observed when using higher protein concentrations.[215]

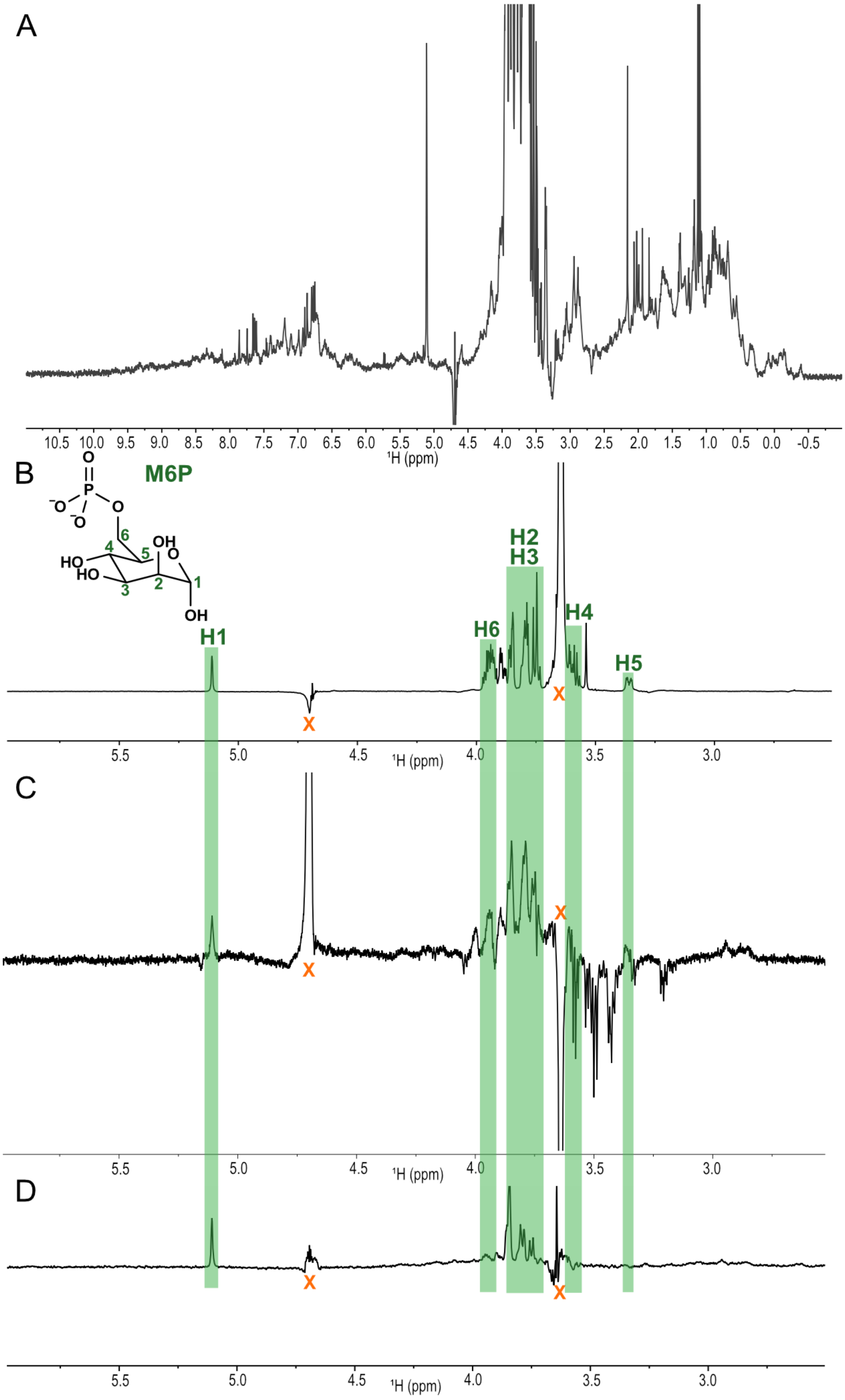


Figure 48: Saturation transfer NMR experiments of D9-10 plus M6P. A: 1D ^{1}H-NMR spectrum of D9-10 plus M6P confirmed that glycosylated D9-10 was folded. **B:** 1D ^{1}H-NMR spectrum of M6P. **C:** WaterLOGSY of D9-10 plus M6P. **D:** STD difference spectrum of D9-10 plus M6P with an on-resonance frequency of 0.58 ppm and off-resonance frequency of -28 ppm. Shifts corresponding to protons of M6P are highlighted in green. Peaks marked with an orange cross correspond to water (~4.7 ppm) and tris (~3.6 ppm). All spectra were collected at 700 MHz by Dr Chris Williams with 20 μM D9-10 and 2 mM M6P in 25 mM Tris, 150 mM NaCl pH 7.4 in 60 % D_2O.

Thus, alternate methods to determine the binding/ inhibition constant were explored. Preliminary isothermal titration calorimetry (ITC) experiments were performed using glycosylated D9-10 and an excess of M6P. However, the ligand, M6P, is an acidic monosodium salt, which makes exact buffer matching challenging and thus obscures the signal of M6P interacting with D9-10 even following buffer subtraction. Furthermore, due to the N-linked glycan, this would be a competitive experiment requiring subtraction of a non-competitive control experiment that would consist of glycosylated D9-10 in the syringe being titrated into glycosylated D9-10 in the sample cell. However, the ratio of ligand to protein established in preliminary experiments with M6P would require 2.5 mM glycosylated D9-10, which is not feasible.

3.3. Domains 7-10

Having succeeded in expressing a D9 containing di-domain construct in insect cells and determining the first high-resolution structure of the elusive sugar binding domain D9, attention turned to larger multi-domain CI-MPR constructs. In the low-resolution D7-11 crystal structure, D11 interacts weakly with the remaining D7-10, which may have reduced the quality of crystals and diffraction data obtained. Thus, the D9-10 construct was extended to encompass D7-10, a ~66 kDa tetra-domain construct containing 5 N-linked glycosylation sites and 15 disulphide bonds.

Two D7-10 constructs were expressed in parallel. The first contained a C-terminal His_6 tag. This construct was generated by former student Ryan Nicholls (University of Bristol), who sub-cloned and performed small-scale expression tests of CI-MPR multi-domain constructs in both mammalian and insect expression hosts at the Oxford Protein Production Facility (OPPF).

Meanwhile, in the second D7-10 construct the C-terminal His_6 tag was replaced by a Strep II affinity tag. A sequence of 8 amino acids (WSHPQFEL), the Strep II tag interacts with an engineered form of Streptavidin called StrepTactin with moderate affinity (K_D ~1μM).[216] The Strep II tag is, therefore, routinely used in the purification and detection of recombinant proteins.[216]

Expression and purification of D7-10

D7-10His:

The D7-10His construct (Appendix Section 8.1.) contains the same N-terminal RPTPμ signal sequence and C-terminal His_6 tag as the D11, D7, D8 and D9-10 constructs. A frozen baculoviral infected insect cell stock (BIIC) from Ryan Nicholls (University of Bristol) was thawed and used to infect fresh Sf21 cultures. Since the EMBacY MultiBac bacmid had not been used, protein expression could not be monitored by measuring YFP emission. Instead the optimal harvest time of D7-10His was determined by SDS-PAGE analysis to be 48 hours after the day of proliferation arrest (Figure 49A). However, only a very faint band for D7-10 was visible in samples of the culture media. SDS-PAGE and western blot (Figure 49) showed that, while some D7-10 has been secreted into the media, a substantial proportion has been retained intracellularly. This may be due to the increased size of the construct (~65 kDa versus the ~16-19 kDa D8 and D7 constructs), the increased number of disulphide bonds (fifteen versus four and three for D8 and D7 respectively) and the increase in N-linked glycosylation sites (five versus one and two for D8 and D7 respectively).

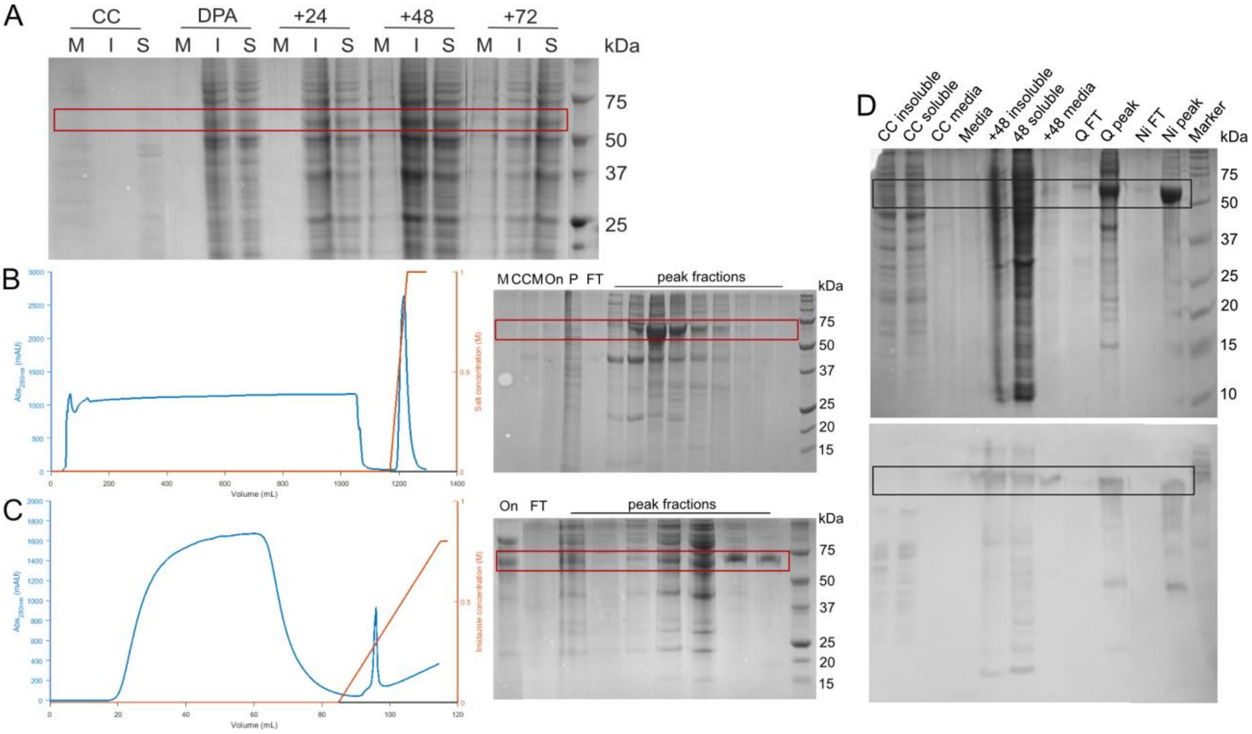


Figure 49: Expression and purification of D7-10His. A: SDS-PAGE analysis of D7-10His expression in Sf21 insect cells. Samples of media (M), insoluble (I) and soluble (S) cell fractions were collected every 24 hours from the day after proliferation arrest (DPA). D7-10His (within the red box) has an expected molecular weight of ~65 kDa. **B:** D7-10His was harvested from culture media on DPA +48 by IEX chromatography using Q sepharose resin. **C:** Harvested D7-10His was further purified by Ni^{2+} IMAC. **D:** SDS-PAGE (top) and western blot (bottom) following expression and purification of D7-10His via IEX chromatography and Ni^{2+} IMAC. Abbreviations: M – media, I – insoluble cell fraction, S – soluble cell fraction, DPA – day after proliferation arrest, CC – uninfected cell control, FT – flow through, Q – Q sepharose IEX.

D7-10His was harvested from Sf21 culture media by ion-exchange (IEX) chromatography. D7-10His has a net negative charge at physiological pH (pI of 5.8) so anion exchange chromatography was performed using a Q sepharose^{TM} column – whereby the resin is conjugated to positively charged quaternary ammonium ions and the protein of interest eluted off over a salt gradient (Figure 49B). Although this concentrated D7-10, SDS-PAGE analysis shows that IEX was not sufficient to purify D7-10His, with multiple higher and lower molecular weight species co-eluting across the peak (Figure 49B). D7-10His was therefore further purified by affinity chromatography by Ni^{2+} IMAC (Figure 49C).

Although eluted D7-10His initially appeared pure by SDS-PAGE (Figure 49C), western blot analysis (Figure 49D) and subsequent biophysical experiments (below) revealed the recurring presence a lower molecular weight impurity. The detection of this species by western blot probing the C-terminal His_6 tag suggests that this must be a degradation product of D7-11 (Figure 49D). An N-terminal truncation of D7-11 would retain the C-terminal His_6 tag. Thus, this species likely corresponds to loss of the N-terminal D7 (~16 kDa unglycosylated and ~19 kDa glycosylated), giving a molecular weight of ~47.8 kDa unglycosylated (loss of D7 has also resulted in loss of two glycosylation sites, lowering the glycosylated molecular weight and heterogeneity) that is consistent with SDS-PAGE analysis. While the cause of this truncation is unknown, cleavage by a protease in the culture medium or improper processing of the signal sequence by the signal peptidase may be responsible.

D7-10Strep:

A synthetic gene encoding D7-10 with an N-terminal RPTPμ signal sequence and a C-terminal Strep II tag (D7-10Strep, Appendix Section 8.1., GeneArt, ThermoFisher Scientific) was sub-cloned from the synthetic vector pMK-rq into the transfer vector pFL (kindly provided by the Berger group, University of Bristol) using the restriction endonucleases BamHI and HindIII. EMBacY DH10 *E. coli* were transformed with D7-10Strep-pFL and D7-10 protein was expressed in Sf21 insect cells using the protocols established in chapter 2 for single domain constructs.

The optimal harvest time was determined by YFP emission, SDS-PAGE and western blot to be 96 hours after the day of proliferation arrest (Figure 50). As observed in the expression of D9-10, SDS-PAGE and western blot reveal that some D7-10 protein is retained in the insect cells (in both the soluble and insoluble cell fractions, Figure 50B). This is likely due to the presence of 30 cysteines that form 15 disulphide bonds. However, the majority of D7-10strep is secreted. D7-10strep from samples of culture media was not visible by SDS-PAGE unless concentrated 10-fold or western blot analysis performed probing for the Strep II tag (Figure 50B).

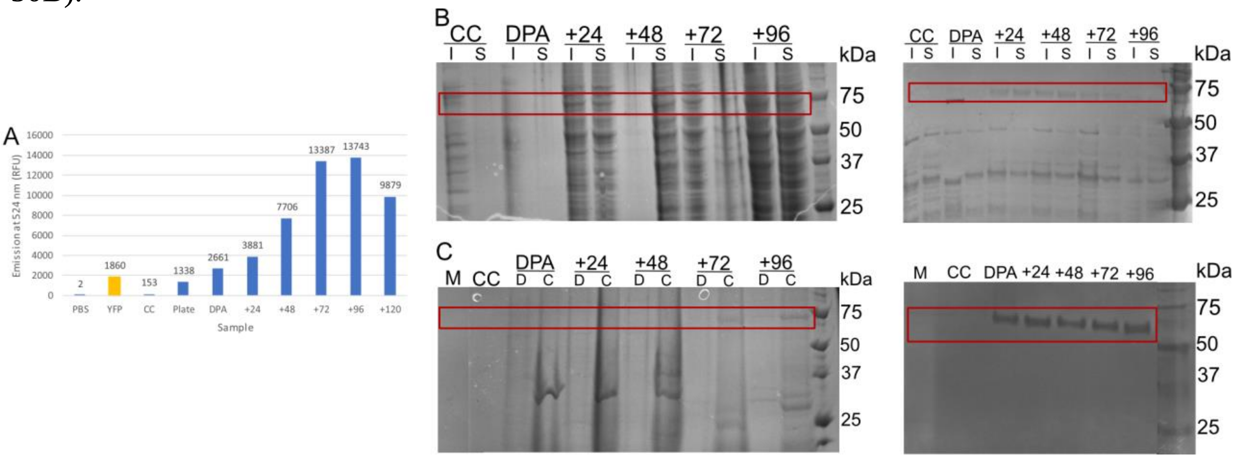


Figure 50: Expression of D7-10Strep from Sf21 insect cells. A: YFP emission peaks 96 hours after the day of proliferation arrest (DPA +96). **B:** SDS-PAGE (left) and western blot (right, developed using Streptactin-AP) of samples of insoluble (I) and soluble (S) cell fractions taken every 24 hours from the day of proliferation arrest (DPA). **C:** SDS-PAGE (left) and western blot (right) of samples of media dilute (D) or concentrated 10-fold (C) collected every 24 hours from the day of proliferation arrest. D7-10Strep (within the red box) has an expected molecular weight of ~66 kDa. Abbreviations: CC-uninfected cell control, I-insoluble, S-soluble, M-media only, DPA-day of proliferation arrest, D-dilute, C-concentrated 10-fold.

Although D7-10Strep contains a C-terminal Strep II tag for affinity purification, addition of culture medium directly to a 5 mL StrepTrap^{TM} was not practical on the scale required. Therefore, as for D7-10His, IEX chromatography was performed first as a crude purification and concentration step. Owing to D7-10Strep's net negative charge at physiological pH (pI of 5.2) anion exchange chromatography was performed (Figure 51A). SDS-PAGE and western blot analysis were used to select fractions containing D7-10strep. These were pooled and applied to the pre-equilibrated StrepTactin column and eluted off over a gradient of desthiobiotin (Figure 51B).

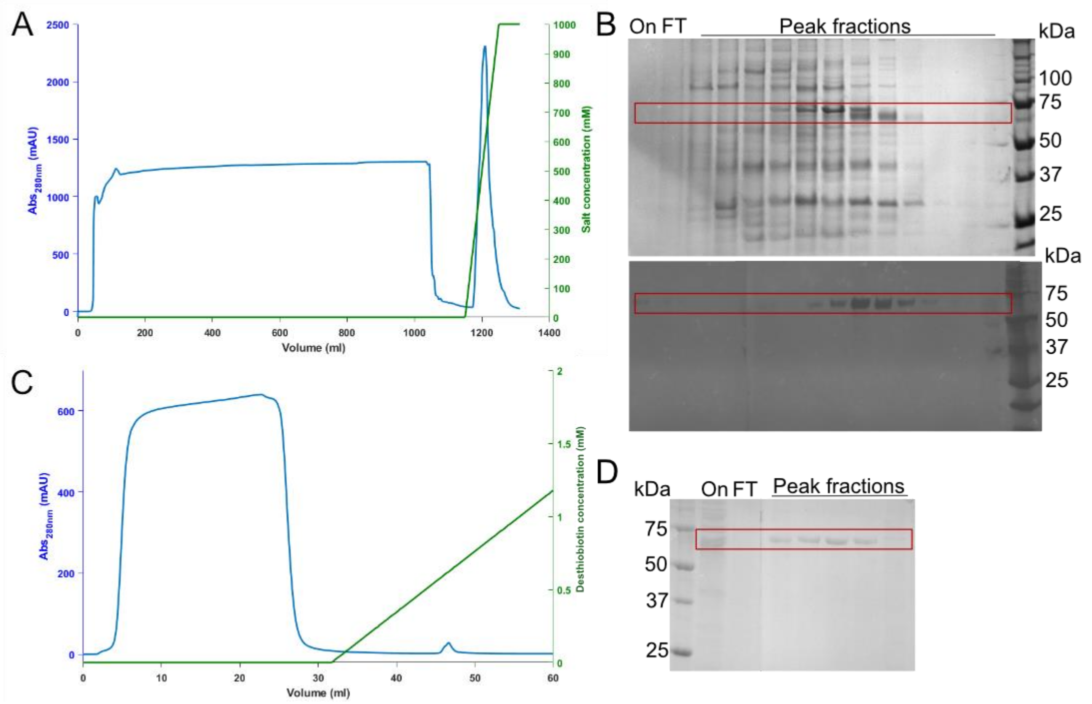


Figure 51: Purification of D7-10Strep. A: D7-10Strep was purified from culture medium by IEX chromatography. D7-10strep was eluted over a gradient of salt (NaCl). **B:** SDS-PAGE (top) and western blot (bottom) analysis of IEX chromatography. D7-10Strep (within the red box) has an expected molecular weight of ~66 kDa. **C:** IEX peak fractions were then purified by affinity chromatography with a 5 mL StrepTrap^{TM} column. D7-10Strep was eluted over a gradient of desthiobiotin. **D:** SDS-PAGE analysis confirmed isolation of pure D7-10strep. 'On' corresponds to media loaded onto the column, 'FT' to the flow through and 'Fractions' to the peak fractions eluted following desthiobiotin addition.

However, even following optimisation of the purification protocol, the final yield of D7-10Strep was very low, approximately 0.5 mg/ L. Due to a much greater yield of D7-10His (8.6 mg/L) all subsequent protein characterisation was performed with D7-10His only.

<u>Characterisation of D7-10</u>

ESI-MS was performed with D7-10 (Figure 52). However, as seen by the broadness of the m/z peaks, D7-10 was very heterogeneous. This can be attributed to the five predicted N-linked glycosylation sites (N951 and N957 in D7, N1163 in D8 and N1246 and N1312 in D9) giving an expected molecular weight range of 65.9-71.1 kDa (65.9 kDa being unglycosylated and 71.1 kDa being D7-11 plus five fucosylated paucimannosidic N-linked glycans). When deconvoluted, the ESI-MS gave a molecular weight range of 66.3-66.9 kDa. SDS-PAGE revealed that attempts to de-glycosylate D7-10 under denaturing conditions using EndoH and PNGaseF (individually) were unsuccessful (Appendix Figure 9).

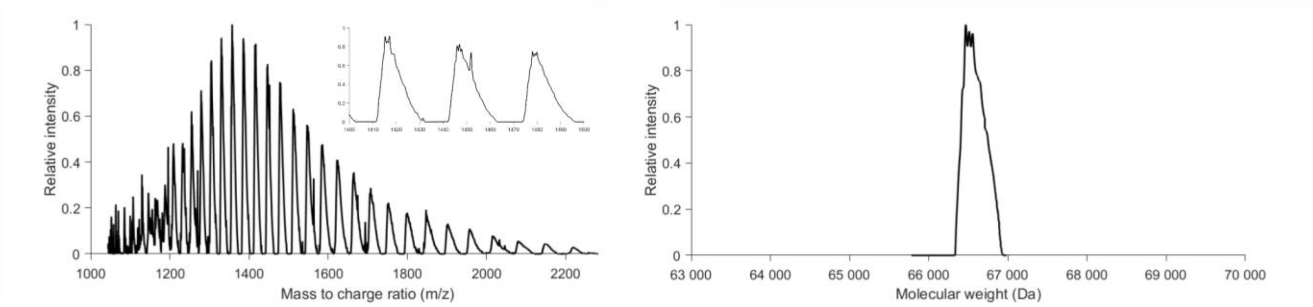


Figure 52: ESI-MS of D7-10. Left: raw m/z spectrum of D7-10. Inset zooms in on m/z region 1400-1500 to demonstrate the broadness of each peak. **Right:** deconvoluted spectra for D7-10 is again very broad, with an observed molecular weight range of 66.3-66.9 kDa and expected molecular weight range of 65.9-71.1 kDa, due to the presence of five N-linked glycosylation sites.

The only evidence that suggests D7-10 is folded is SDS-PAGE analysis (Figure 53A). D7-10His ran faster under non-denaturing, non-reducing conditions suggesting a more compact form stabilised by disulphide bonds - of which D7-10 has fifteen.

When analysed by analytical SEC (Figure 53A), D7-10 eluted as a single, broad peak. The peak maxima (14.2 mL) corresponded to an apparent molecular weight of 63.4 kDa (Appendix Table 7). With a calculated molecular weight of 65.9-71.1 kDa this suggests D7-10 is monomeric in solution. However, the peak is very broad and spanned 11.3-15.6 mL, which corresponds to 380-26.5 kDa. It is possible, therefore, that a smaller population of D7-10 exists as a dimer. There was no change in elution volume following the addition of mannose or M6P.

Size exclusion chromatography coupled to multi-angle light scattering (SEC-MALS) was used to further characterise the oligomeric state of D7-10. D7-10 eluted as a broad peak with an apparent molar mass of 51.0 kDa ± 38 % (corresponding to 31.6-70.4 kDa) suggesting D7-10 was monomeric in solution (Figure 53B). There was no significant change in elution profile or calculated molar mass following incubation with 10-fold excess M6P (73.1 kDa ± 4.5 %, corresponding to 69.8-76.4 kDa, Figure 53B). D7-10 did not elute later following treatment with EndoH and PNGaseF under native conditions (individually). This is in-line with the observation, by SDS-PAGE, that D7-10 was not enzymatically de-glycosylated under denaturing conditions (Appedix Figure 9).

Sedimentation equilibrium analytical ultracentrifugation (SE-AUC) was also performed (with the help of Dr Guto Rhys, University of Bristol) to determine the oligomeric state of D7-10 (Figure 53C). SE-AUC, which reports molecular mass and oligomeric state, uses low centrifugal forces to create an equilibrium between sedimentation flux and diffusional flux.[217] The sedimentation coefficient (s) can be defined as below (Equation 1) whereby M_b is the buoyant molar mass and *f* the frictional coefficient. The buoyant molar mass (M_b) is calculated using the mass of the protein (M_p) and partial specific volume of the protein, $\bar{v}_p$ (Equation 2).

$$s = \frac{M_b}{f}$$

Equation 1: Calculation of the sedimentation coefficient (s) for AUC analysis. M_b is the buoyant molar mass and f the frictional coefficient.

$$M_b = M_p(1 - \bar{v}\rho)$$

Equation 2: Calculation of the buoyant molar mass (M_b). M_b depends upon the mass of the protein (M_p) and partial specific volume of the protein ($\bar{v}_p$).

The presence of glycans, however, will affect the partial specific volume ($\bar{v}$) of gylcoproteins.[218] The mass of the carbohydrate (M_c) can be estimated as the total mass of the protein (M_p) minus the mass of the amino acids (M_A).[219] The partial specific volume of glycoproteins can therefore be calculated as below (Equation 3).[219] Mass spectrometry has shown the sample to be highly heterogeneous, containing multiple glycosylated forms. Subsequently an average mass and partial specific volume were estimated for the carbohydrate content and amino acid composition ($\bar{v}$ values calculated using SEDNTERP).[220]

$$\bar{v} = \frac{1}{M_p}(M_A\bar{v}_A + M_C\bar{v}_C)$$

Equation 3: Calculation of the total partial specific volume ($\bar{v}$) of glycoproteins. M_p is the mass of the total protein, M_A the mass of the amino acid component, M_c the mass of the carbohydrate component, $\bar{v}_A$ the partial specific volume of the amnio acid component and $\bar{v}_C$ the partial specific volume of the carbohydrate component.[219]

Due to the presence of two species on the SDS-PAGE, results were fitted to a two-component model (Ultrascan II).[221] This gave molecular masses of 64,354 Da (95 % confidence intervals: +1090 Da -800 Da, which corresponds to 63,554-65,446 Da) and 43,029 Da (95 % confidence intervals: +7470 Da -6180 Da, which corresponds to 36,849-50,499 Da). The former corresponds well to monomeric D7-10 (expected mass 65.9-71.1 kDa) and the later to truncated D7-10 (ie. D8-10, 47.8-51.1 kDa).

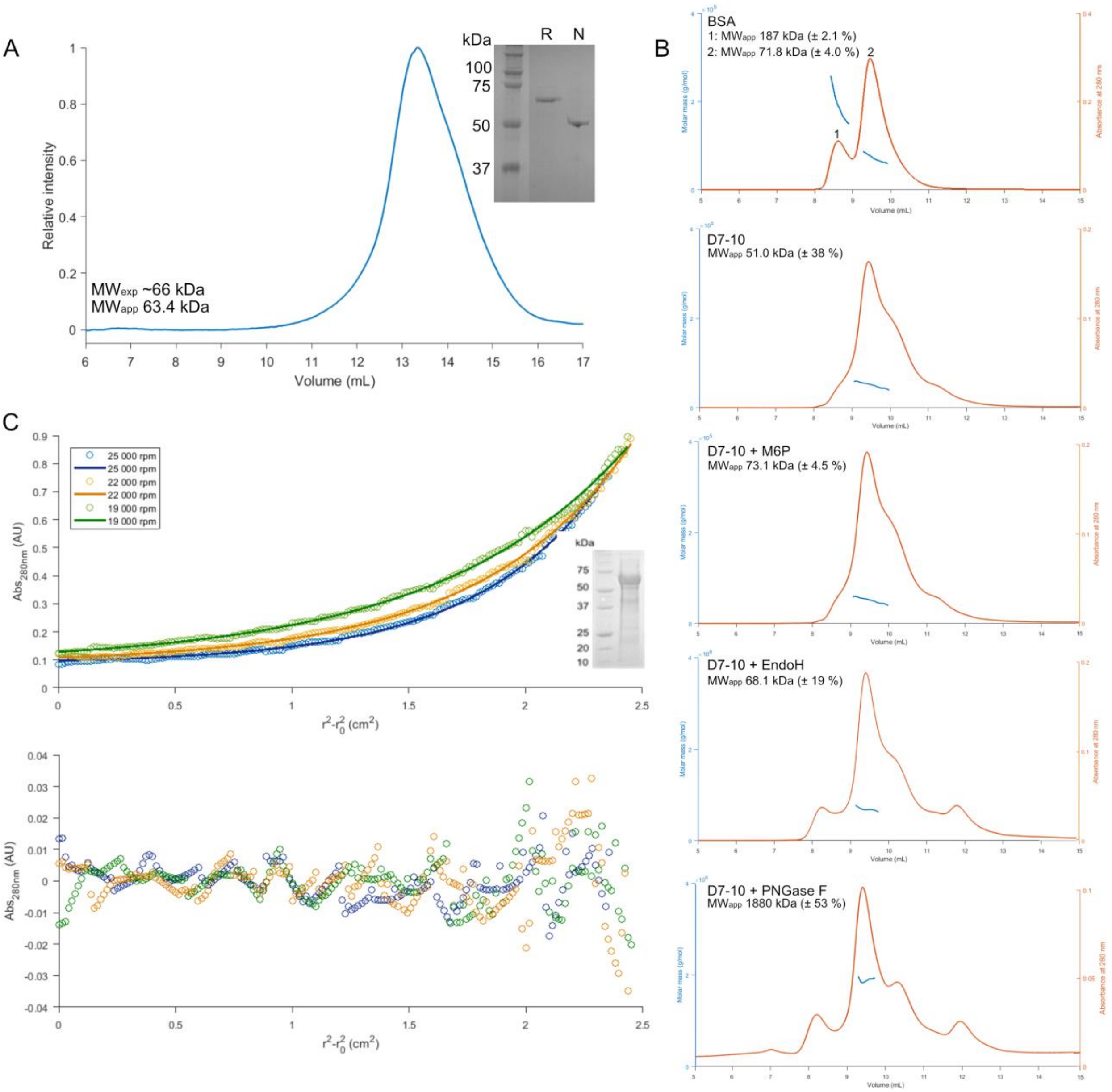


Figure 53: Determination of D7-10 folding and oligomeric state. A: Analytical SEC and SDS-PAGE analysis of the peak fraction under reducing (R) and native (N) conditions suggested D7-10 was folded and monomeric. **B:** SEC-MALS of D7-10 confirmed the monomeric status of D7-10 (51.0 kDa ± 38 %) that was unchanged by the presence of M6P (73.1 kDa ± 4.5 %). D7-10 was not successfully de-glycosylated by EndoH or PNGaseF. **C:** Experimental SE-AUC data (circles) and fits (lines) to a two-component model (top) along with residuals between the experimental and filled data points (bottom) reveal the presence of two species of molecular weights 64,354 Da and 43,029 Da.

Negative stains of D7-10, which were prepared and imaged with the help of Professor Christiane Berger-Schaffitzel (University of Bristol), showed the sample to be uniform (Figure 54). However, structure determination by Cryo-EM was not pursued due to the relatively small size of the protein (~65 kDa) and sample impurity.

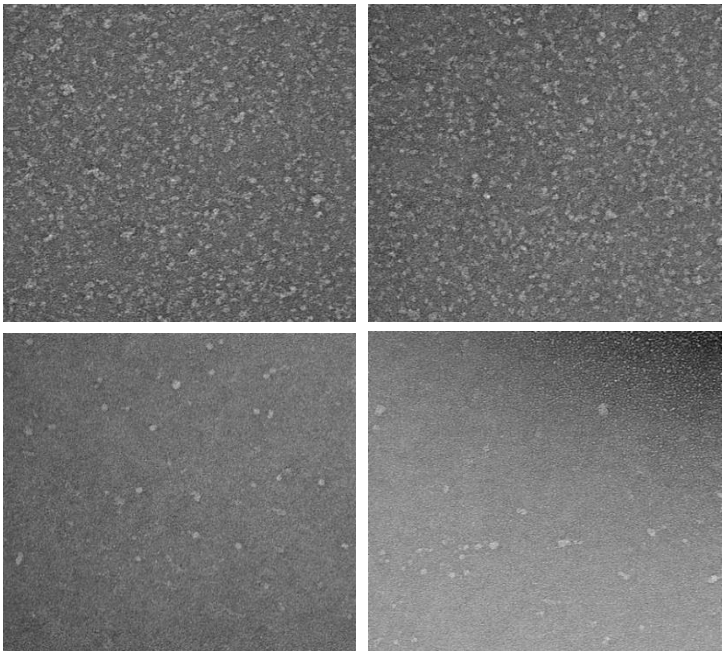


Figure 54: Transmission electron micrographs of D7-10 negatively stained with uranyl acetate. 100 k x magnification. Taken with an FEI Tecnai 20 200 kV (Wolfson Bioimaging facility) Top: 50 μg/mL. Bottom: 5 μg/mL.

Instead, in attempt to determine the structure of D7-10, sparse matrix crystallisation screens were prepared and after 3 days very small, birefringent crystals were observed in Structure screen I+II condition E8 (0.2 M ammonium phosphate monobasic, 0.1 M Tris pH 8.5, 50 % v/v MPD) (Appendix Figure 10). These were looped, cryo-cooled and taken to Diamond Light Source for data collection. However, the observed diffraction pattern (Appendix Figure 10) suggests these crystals were salt. Unfortunately, further attempts to crystallise D7-10 with longer crystallisation times, varying protein concentrations, with and without EndoH and with and without M6P failed to produce diffracting crystals. Similarly, SAXS was also unsuccessful due to protein aggregation. Thus, structural characterisation of D7-10 was not pursued further.

3.4. Domains 7-11

Although, a bovine construct comprised of D7-11 has previously been shown to be stable to proteolytic digestion [170] and expressible as a soluble protein in High five insect cells that maintains M6P binding (K_D = 0.5 nM) as well as an intact IGF2 binding site (D11)[132], there is no structural information for this central region of the human CI-MPR.

Collaborators led by Professor Bass Hassan at the University of Oxford expressed human D7-11 (and other multi-domain constructs containing D9: D8-9, D9-10 and D9-11) in HEK293T cells and assayed for M6P-binding by surface plasmon resonance (SPR) using the known CI-MPR ligand Leukaemia inhibitory factor (LIF). The binding affinities (K_D) determined by SPR of each construct were between 60-80 nM and in line with literature values for a single, bovine D9 construct binding M6P mono-ester glycoprotein GAA (K_D 95 ± 12 nM).[139] The binding affinity of D9 did not change significantly between constructs, confirming the unique property of D9 that M6P mono-ester binding affinity is independent of the presence of neighbouring domains.[67,70,82,130]

In parallel expression experiments, human D7-11 was expressed in mammalian HEK293S cells by Hans Hoppe (University of Oxford) for structure determination. HEK293S cells lack N-acetyl-glucosaminyl transferase I (GnTI) and therefore secrete protein containing shorter N-linked $GlcNAc_2Man_5$ glycans.[222] D7-11 crystallised from a solution of 0.1 M MES, 1.6 M $MgSO_4$, 10 mM M6P at pH 6.5 and, in 2013, X-ray diffraction data was collected at Diamond Light Source by Karl Harlos (University of Oxford).[70]

D7-11 crystallised in space group $P4_12_12$ with two molecules in the asymmetric unit. However, D7-11 was partially radiation damaged and diffracted to low-resolution. Thus, D7-11 was a challenging structure to solve that required Airlie McCoy (University of Cambridge) to phase using a series of homology models generated for D7-10 and the crystal structure of D11 (PDB 1GP0). Unfortunately, the preliminary D7-11 structure was not suitable for publication.

However, the recent determination of high-resolution structures of human D8 (2.5 Å, chapter 2) and human D9-10 (1.5 Å) has assisted in the phasing and refining of D7-11. Adding restraints from these three previously uncharacterised domains allowed the structure of D7-11 to be determined to 3.5 Å (Rwork and Rfree values of 26.1 % and 30.0 % respectively) with 99.6 % of the backbone dihedral angles in the allowed and favoured regions of the Ramachandran plot (Appendix Table 5, Appendix Figure 4). There was no electron density for A933 at the N-terminus of D7 or E1647-T1651 at the C-terminus of D11. D7-11 contains five predicted N-linked glycosylation sites: N951 and N957 on D7, N1163 on D8 and N1246 and N1312 on D9. Electron density was observed for a single GlcNAc residue at N951 of both chains, N1246 of chain B and $GlcNAc_2Man_4$ at N1312 of both chains.[70]

D7-11 crystal structure

The structure of D7-11 (Figure 55) comprises residues C934 to C1646 of the human CI-MPR and provides the first structure of human D7. The construct forms two 80 kDa chains that associate to form an intertwined 160 kDa homodimer. Each chain is comprised of five well-defined domains of similar size, comprised of D7 (C934-P1081, 147 amino acids), D8 (V1082-R1221, 139 amino acids), D9 (V1222-D1365, 143 amino acids), D10 (L1366-S1510, 144 amino acids) and D11 (N1511-C1644, 143 amino acids). Both chains wrap around each other primarily through associations between D8-10. Each domain has the same core β-sandwich topology composed of nine anti-parallel β-strands. The IGF2 binding domain, D11, is flexible and disordered in the complex probably due to the lack of neighbouring domains. In the D7-11 structure, D11 associates with D10 via the βE-βI surface. This is the same region of D11 observed to pack against D12 in the crystal structures of D11-12 and D11-14 (PDB 2V5N, 2V5O)[101] and may therefore be an artefact arising from the burial of this hydrophobic face of D11, possibly due to the absence of D12. This packing may be sub-optimal and may explain the poor electron density observed for this domain and a degree of conformational averaging.[70]

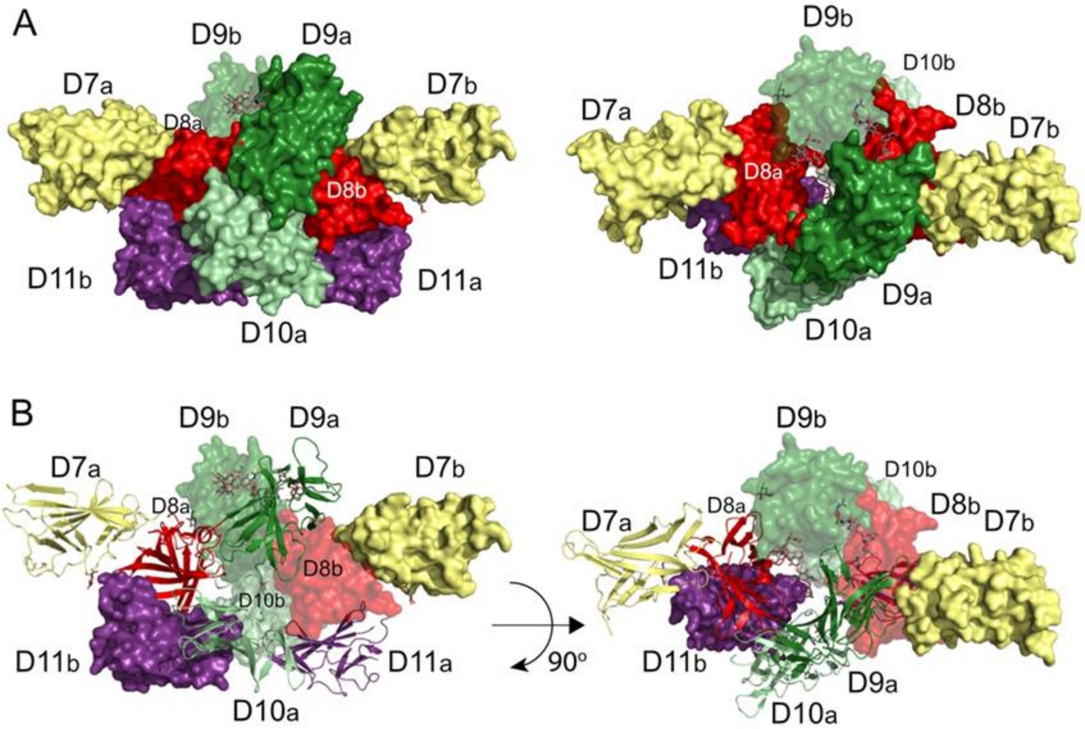


Figure 55: Structure of human CI-MPR D7-11 homodimer.[70] **A**: Structure of human CI-MPR D7-11 at 3.5 Å resolution. Both monomers are shown in surface representation with the two chains labelled a or b and coloured yellow (D7, C934-P1081), red (D8, V1082-R1221), dark green (D9, V1222-D1365), light green (D10, L1366-S1510) and purple (D11, N1511-C1644). The structure on the right is rotated by 90° and the glycans extending into the opposing D9 M6P binding sites are shown as sticks. **B**: One monomer is shown in cartoon format, the other in surface representation and two orientations are shown.

The dimeric interface of D7-11 is formed from predominantly hydrophobic contacts with a total buried surface area of approximately 19500 Å^2 and a solvent accessible area of 66400 Å^2 (as calculated by PISA[195]). The binding site loops of D9 are solvent exposed and, as observed in the D9-10 structure, a mannose residue from the modelled portion of the N1312-linked glycan of D9 extends into the D9 M6P binding site of the other chain. However, unlike in the D9-10 crystal structure, in the D7-11 crystal structure this is a symmetric interaction with both N-linked glycans binding to opposing D9 binding sites in the dimer (Figure 56). The terminal mannose residues are oriented to bring the 6'OH within ~ 3 and 6 Å of H1320 in the two sites and the two interactions bring the FG loops of neighbouring D9 molecules into close proximity. The conserved 'QREY' residues of the D9 M6P binding site interact with the terminal mannose residue supplemented by additional interactions to the preceding sugar (i.e. 3'OH to Y1255 at the N-terminus of βB) (Figure 56). Interestingly, the M6P binding site of D9 and the IGF2 binding site of D11 point away from one another on opposing surfaces, suggesting that it may be possible for an M6P-tagged glycoprotein and IGF2 to bind simultaneously.

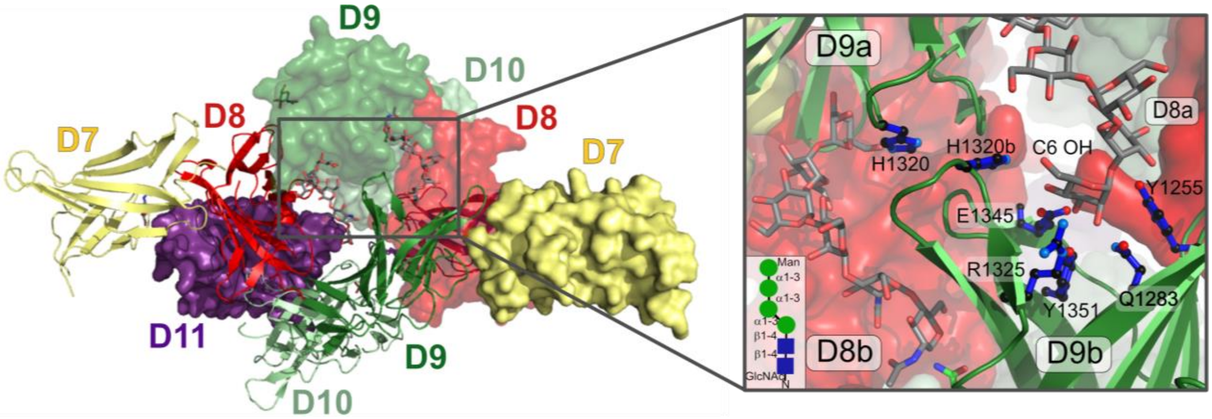


Figure 56: The bridging glycan of D7-11.[70] The structure of the D7-11 homodimer with one chain shown as cartoon and one as surface render. The inset shows the interface of the D9 structures with glycans (grey sticks) extending into the opposing D9 M6P binding site. Conserved residues ('QREY') are shown (blue balls-and-sticks) along with H1320 and Y1255 which binds a mannose group adjacent to the terminal mannose residue.

Superimposing the single human D8 crystal structure and D8 of the human D7-11 structure gives an RMSD value of 0.51 Å over backbone atoms (Appendix Figure 4). D8 and D10 form the most extensive inter-domain contacts and the core of the D7-11 dimer. Interestingly however, there are no homo-dimeric contacts (i.e. 8 to 8b and 10 to 10b) for either of these domains and instead the D9 dimer forms a capstone that binds a ring of domains formed from 8/10-8b/10b. D8-10 (of D7-11) forms a compact tri-domain structure comparable to the crystal structure of D1-3 (PDB 1SYO/ 1SZ0, 1Q25) (Figure 57).[70] This extensive packing around D8 suggests the role of this domain is purely structural.

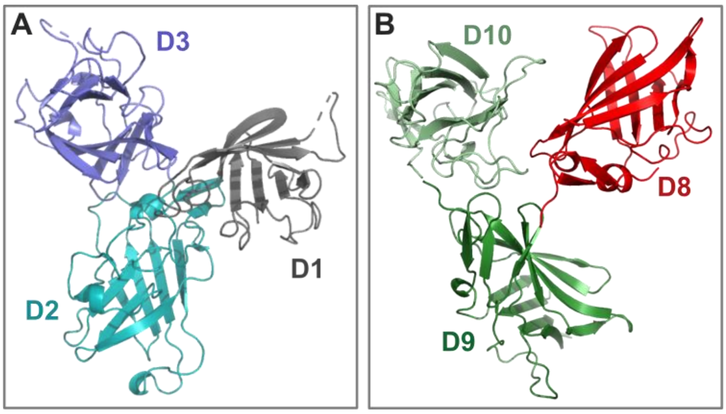


Figure 57: Compact tri-domain structures of D1-3 and D8-10. A: Crystal structure of D1-3 (PDB 1SZ0). **B:** Crystal structure of D8-10 of the human D7-11 crystal structure (PDB 6Z32).

The D7-11 crystal structure exposes the first structure of human D7. D7 is the most solvent exposed domain of D7-11, although this structure lacks neighbouring D6. The N-terminal βN and βN' strands, the GH loop and the C-terminal loop of βI from D7 all contact D8 of the same chain (i.e. D7a to D8a). S1602 of the HI loop of D11b from the opposing chain of the dimer is also brought into close proximity to the N-terminus of D7a βA.

There is no known ligand of D7. However, as pointed out by Brown *et al.* 2002, all the known functional domains to date are odd numbered domains.[112] For example, D1 binds uPAR, D3, D5, D9 and D15 bind M6P-tagged proteins while D11 and D13 bind IGF2.[75,101,103,130–132] Indeed, D7 is the last odd numbered domain with no known binding partner. In the D7-11 structure, the AB, CD, FG, and HI loops of D7 point away from the dimeric hub and enclose a slightly positively charged pocket (Figure 58B). D7 has a short AB loop, extended CD and HI loops and lacks one disulphide bridge (cf C1598-C1634) that connects the FG loop N-terminal of βI. This is substituted for on βI by R1065 that along with residues K963 (AB loop), K1000, R1003 (CD loop) and K1030 (FG loop) forms a cluster of positively charged residues. However, compared to D3, D9 and D11 which have more extended AB loops but shorter CD loops, the D7 loops do not form an obvious binding groove.[70]

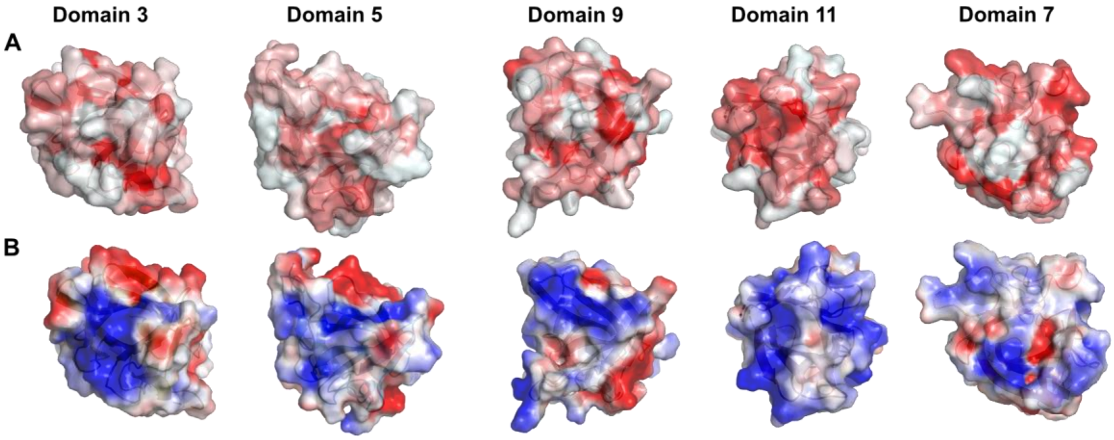


Figure 58: Comparison of the known CI-MPR binding sites and the loops of D7. A: Comparison of the binding sites (viewed from the top) of D3, D5, D9, D11 and D7 (PDB 1SYO, 2KVB and D7-11 structure). Each domain is shown with surface render with hydrophobic residues coloured red according to the normalised hydrophobicity scale.[197] **B:** Electrostatic surface potential of the same domains in identical orientation coloured by charge with positively charged residues blue and negatively charged residues red (range +2 to -2) as determined using the APBS PyMOL plug-in.[196]

Interdomain orientation and interfaces

Recent work by Wang *et al.* 2020 has determined the cryo-EM structure of bovine CI-MPR D1-14 to 3.5 Å at the physiological pH of the lysosome, pH 4.5 (PDB 6UM1). However, the structure of bovine CI-MPR D1-14 could not be obtained at neutral pH in the absence of ligand (likely due to flexibility).[128] Only upon incubation with IGF2 could a low-resolution (4.3 Å) structure of the bovine CI-MPR D4-14 be obtained at neutral pH, with D1-3 omitted due to flexibility (PDB 6UM2).[128]

At pH 4.5, the extracellular region of the bovine CI-MPR forms a compact, helical structure.[128] Seven sub-groups (D1+D3, D2+D5, D4+D7, D6+D9, D8+D11, D10+D13 and D12+D15) form with the βE-I surface of each domain packing against the βE-I surface of its partner.[128] A similar interaction is observed in crystal structures of bovine CD-MPR at pH 6.5 (PDB 2RL8, 2RL9, 1KEO), whereby the βE-I surfaces of each monomer pack against one another and the HI loop of one monomer interacts with the N-terminus of the other.

The cryo-EM structure of bovine CI-MPR at pH 4.5 reveals that the βE-I surfaces of D6 and D9 pack against one another (in a similar manner to CD-MPR dimerisation).[128] However, this interaction is shown to be disrupted in the cryo-EM structure of bovine D4-14 in complex with IGF2 at pH 7.4 (PDB 6UM2).[128] The βE-I surface of D9 is similarly exposed in the crystal structure of human D7-11 in the absence of IGF2 but near neutral pH (pH 6.5). It is from this βE-I surface that, in the crystal structures of human D9-10 and D7-11, the N-linked glycan projects from N1312 into the neighbouring D9 binding site. Superimposing a single chain of human D7-11 at pH 6.5 and bovine D7-11 at pH 7.4 (PDB 6UM2) gives an RMSD value of 4.9 Å (PyMOL cealign structure-based alignment Figure 59). Thus, identical packing around D9 can be derived from interaction with one of two distinct ligands, IGF2 or glycan, or, more likely, by neutral pH conditions. Under these conditions, D9 forms protein-protein interactions with the βA-D surface of D10 and, in the D7-11 crystal structure, with the BC loop at the base of D8b.

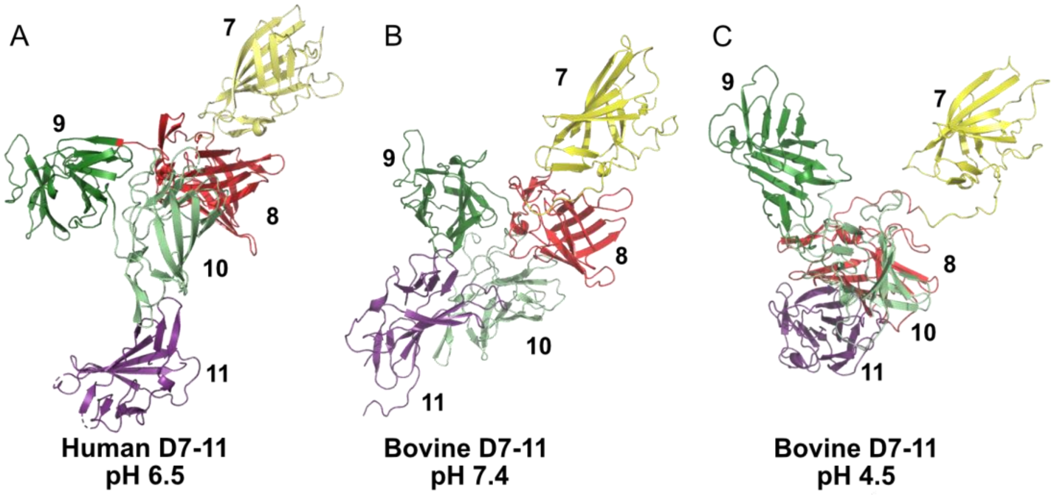


Figure 59: Comparison of human and bovine D7-11 structures. A: A single chain from the crystal structure of human D7-11 at pH 6.5 (PDB 6Z30). **B:** D7-11 from the cryo-EM structure of bovine D4-14 in complex with IGF2 at pH 7.4 (PDB 6UM2). Superimposition of human and bovine D7-11 at near neutral pH shows the same domain orientation and gives an RMSD of 4.9 Å. **C:** D7-11 from the cryo-EM structure of bovine D1-14 at pH 4.5 forms a more compact arrangement (PDB 6UM1). Superimposition of human D7-11 at pH 6.5 and bovine D7-11 at pH 4.5 gives an RMSD value of 7.9 Å. (PyMOL cealign structure-based alignments)

A comparison of the relative orientations of D9-10 in isolation and D9-10 from the D7-11 crystal structure revealed a surprisingly close similarity despite the presence of additional domains. *Ab initio* shape envelope modelling of D9-10 SAXS data was consistent with D9-10 of the D7-11 crystal structure, with symmetric cross-binding of the two N-linked glycan moieties and no contact between D10s (Figure 44). The *Ab initio* shape envelopes also confirm that the dimerisation of D9-10 visible in the crystal structures of D9-10 and D7-11 is driven by the bridging N-linked glycan at N1312 of D9 and not by protein-protein interactions (Figure 40).

Further analysis, superimposing a single copy of D9 and D10 from the D9-10 and D7-11 crystal structures, at pH 9.0 and pH 6.5 respectively, yielded an RMSD value of 3.1 Å (over backbone atoms), consistent with this close alignment. Moreover, the inter-domain interactions also appear to be well conserved, with the CH-π interaction between histidine and proline residues at the D9-10 interface also evident in the D7-11 crystal structure (Figure 60A). It appears therefore that at neutral/ high pH D9-10 forms a rigid body and D10 may play a key role in stabilising the fold of D9.

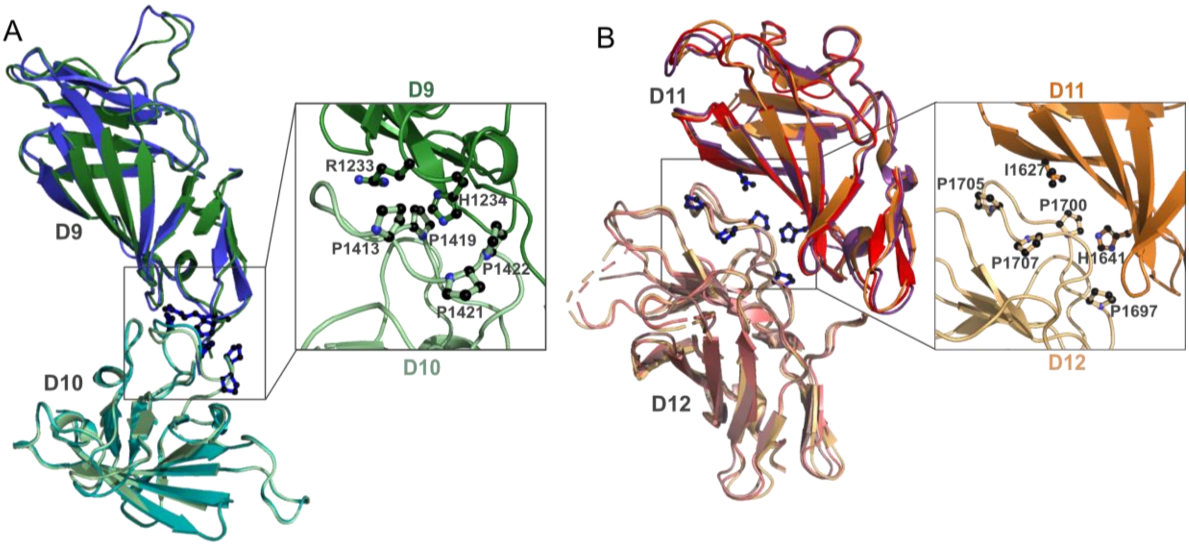


Figure 60: Identification of rigid di-domain structures stabilised by CH-π interactions.[70] **A:** D9 and D10 are in the same orientation in the D9-10 structure (blue, pH 9) and D7-11 structure (green, pH 6.5). Inset shows the His-Pro interaction at the interface of D9 and D10 of the D7-11 structure involving H1234 of D9 and P1413, P1419, P1421 and P1422 of D10. **B:** D11 and D12 also form a rigid unit that is conserved in the crystal structures of human D11-12 (purple), D11-13 (red) and D11-14 (orange) (all at pH 7.5, PDB 2V5N, 2V5P and 2V5O respectively). A similar His-Pro interaction to that in D9-10 is seen at the D11-12 interface involving H1641 of D11 and P1697, P1700, P1705 and P1707 of D12.

The crystal structure of human D11-12 at neutral pH (pH 7.5, PDB 2V5N) revealed the hydrophobic patch of D11 formed from βE-βI packs against the BC loop of D12.[112] This interface includes similar interactions to D9-10, with H1641 of D11 interacting with P1697 and P1700 of the proline rich BC loop of D12 as well as several other hydrophobic interactions (Figure 60B). However, H1641 resides on βI of D11 rather than the loop between βNA and βNA' observed in D9-10 leading to a different association and side-on packing of the D11-12 crossed β-sheets. This interaction was maintained in two further crystal structures of D11-14 and D11-13 bound to IGF2 (PDB 2V5O and 2V5P respectively) suggesting a degree of rigidity between at least D11-12 (Figure 60B).[70]

Analysis of the full extracellular region of the CI-MPR reveals that these His-Pro interactions observed between D9-10 and D11-12 are unique when compared across D1-14. These His-Pro interactions are also observed in the recent, cryo-EM structure of bovine D4-14 at pH 7.4 in complex with IGF2 (PDB 6UM2) with the equivalent residues H1243 of D9; P1422, P1428, P1430 and P1431 of D10; H1641 of D11 and P1697, P1700, P1705 and P1707 of D12.

Additionally, these His-Pro interactions at the D9-10 and D11-12 interfaces are conserved across species. Sequence analysis reveals that residues H1234, P1413, P1421 of human D9 and

H1641, P1697, P1707 of human D10 are present in placentals (e.g. Human and Rodent), marsupials (e.g. Kangaroo), monotremes (e.g. Echidna and Opossum), birds (e.g. Chicken) and fish (e.g. Zebrafish) with P1419, P1422 and P1700 also present in human, bovine and murine CI-MPR (Appendix Figure 11).

The role of pH

The CI-MPR binds its cargo at near neutral pH. For example, IGF2 is bound by D11 at the plasma membrane and in circulation by sCI-MPR at pH 7.4, while the MRH domains D3, D5, D9 and D15 bind M6P-tagged proteins at the plasma membrane and TGN (pH 6.5).[54] Although the pH profiles of individual domains is broad (spanning approximately pH 6.0-7.5), more detailed studies have defined the pH optimum of D3 as pH 6.9 and D9 as pH 6.4-6.5.[132,137] In the low pH environment of the late endosome (pH <6.0), the CI-MPR releases its ligands and is recycled to the TGN and PM. However, the precise mechanism of cargo release at low pH is poorly understood. Is cargo release simply the result of protonation of critical binding site residues in the acidic environment of the late endosome? Or are there conformational changes to the receptor structure that drive ligand dissociation?

Olson *et al.* have proposed that cargo dissociation by the CD-MPR is facilitated by protonation of key binding site residues, particularly E133 of the conserved ‘QREY’ motif.[90] It follows therefore that protonation of E416 of D3 and E1345 of D9 similarly play a role in ligand dissociation from the CI-MPR.

Olson *et al* also observed large conformational changes to the binding loops of the CD-MPR upon ligand binding.[90] Superimposition of bovine CD-MPR unbound and with M6P bound (both at pH 6.5, PDB 1KEO and 2RL8 respectively) gives an RMSD value of 0.47 Å over backbone atoms but a HI loop movement of 16.1 Å (Figure 61A). Similarly, upon binding pentamannosyl phosphate (PDB 1C39) the HI loop is displaced by 15.8 Å. Furthermore, S41 of the AB loop is displaced 4.1 Å and H105 on the FG loop 2.9 Å when pentamannosyl phosphate is bound.[90] However, superimposition of the crystal structures of bovine CD-MPR at pH 6.5 and pH 4.8 (both unbound, PDB 1KEO and 2RL7 respectively) gives an RMSD value of 0.27 Å over backbone atoms and no change to the positioning of the four binding loops.

Meanwhile, superimposition of bovine D5 at pH 6.5 and pH 4.5 (PDB 6UM2 and 6UM1 respectively, RMSD value of 1.0 Å over backbone atoms), demonstrates a HI loop movement of ~7.0 Å (Figure 61B). A similar movement is seen in D9, with superimposition of bovine D9 at pH 6.5 and pH 4.5 (PDB 6UM2 and 6UM1 respectively, RMSD value of 1.1 Å over backbone atoms), resulting in a HI loop moving 6.5 Å (Figure 61C). These domains have no ligands bound suggesting that this movement of the HI loop is pH dependent and may act to occlude the binding site at low pH.

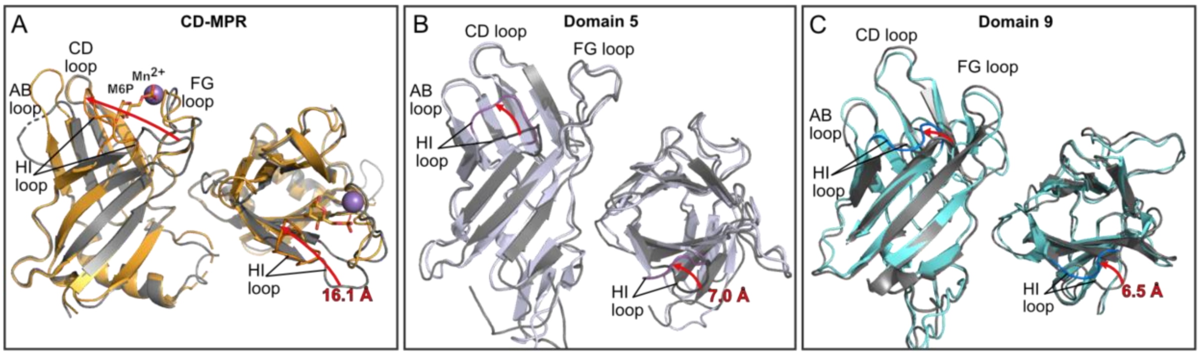


Figure 61: Effect of pH on CI-MPR binding domains. A: Superimposition of bovine CD-MPR unbound (grey) and with M6P bound (orange) both at pH 6.5 (PDB 1KEO and 2RL8 respectively) gives an RMSD value of 0.47 Å over backbone atoms. The HI loop moves 16.1 Å. **B:** Superimposition of bovine D5 at pH 6.5 (grey) and pH 4.5 (purple) (PDB 6UM2 and 6UM1 respectively), viewed from the side and top of the binding site, gives an RMSD value of 1.0 Å over backbone atoms. The HI loop moves 7.0 Å. **C:** Superimposition of bovine D9 at pH 6.5 (grey) and pH 4.5 (blue) gives an RMSD value of 1.1 Å over backbone atoms. The HI loop moves 6.5 Å.

Analysis of the bovine cryo-EM structures reveals further pH sensitive regions of the CI-MPR, particularly the His-Pro interactions at the interfaces of D9-10 and D11-12. At low pH, the histidine residues of D9 and D11 (H1234/ H1243 human/ bovine D9 and H1641/ H1650 human/ bovine D11) are protonated and the domains reorient so that the histidine residues no longer sit within the pocket of proline residues and are therefore unable to form CH-π interactions (Figure 62A, 62C). Superimposing D9-10 of the human D7-11 crystal structure at pH 6.5 with D9-10 of the bovine D1-14 cryo-EM structure at pH 4.5 (PDB 6UM1) (Figure 62B) demonstrates the domain rearrangement, with an RMSD value of 12.3 Å over backbone atoms (versus an RMSD value of 2.7 Å over backbone atoms between D9-10 of human D7-11 at pH 6.5 and D9-10 of bovine D4-14 at pH 7.4, PDB 6UM2). Similarly, superimposition of the human D11-12 structure at pH 7.5 (PDB 2V5N) with D11-12 of the bovine D1-14 structure at pH 4.5 (PDB 6UM1) demonstrates the rotation of D12 relative to D11 (Figure 62D).

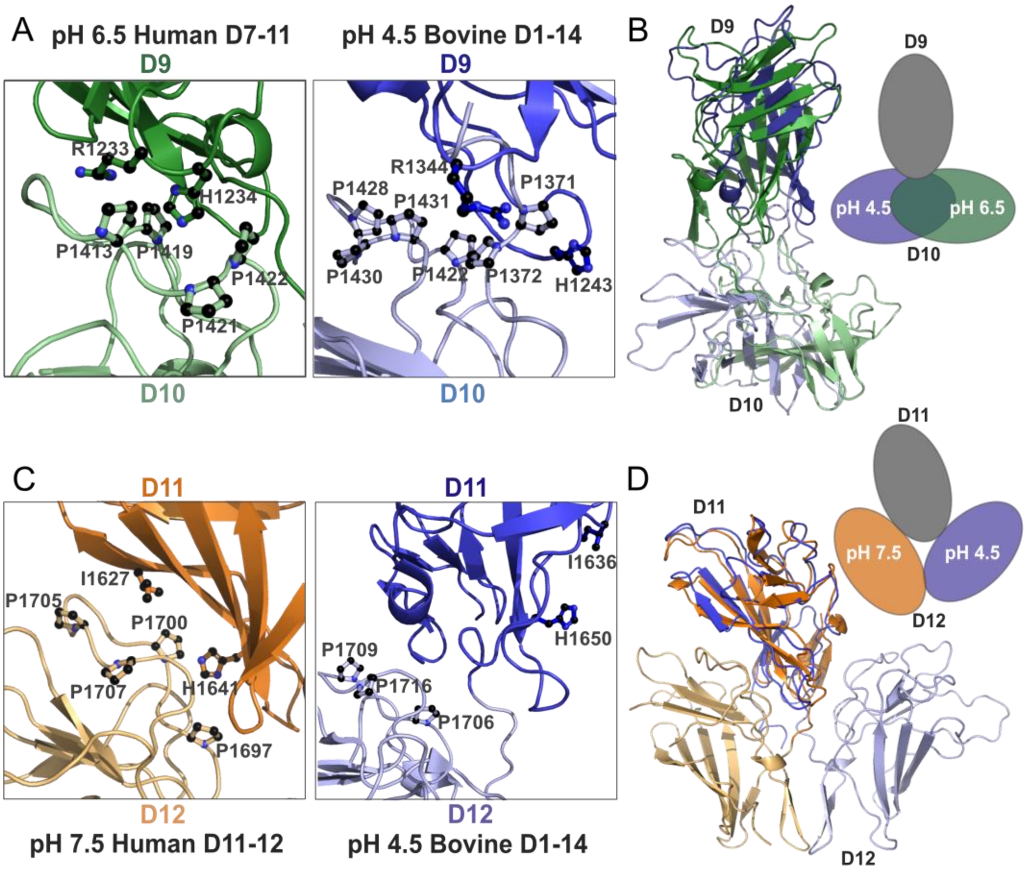


Figure 62: The influence of pH on the structure of di-domain units D9-10 and D11-12. A: The His-Pro interaction at the interface of D9-10 of the human D7-11 crystal structure (PDB 6Z32) at pH 6.5 involves H1234 of D9 and P1413, P1419, P1421 and P1422 of D10. The equivalent residues (H1243, P1422, P1428, P1430 and P1431) are observed in the cryo-EM structure of bovine D1-14 at pH 4.5 (PDB 6UM1). However, at low pH the histidine residue sits outside the proline pocket. **B:** Superimposing D9-10 at pH 6.5 with D9-10 at pH 4.5 reveals the altered domain arrangement. **C:** A similar His-Pro interaction is seen at the interface of D11 and D12 in the human D11-12 crystal structure at pH 7.5 (PDB 2V5N), involving H1641 of D11 and P1697, P1700, P1705 and P1707 of D12. The equivalent residues (H1650, P1706, P11709 and P1716) are observed in the cryo-EM structure of bovine D1-14 at pH 4.5. However, again, at low pH this interaction is disrupted. **D:** Superimposing D11-12 at pH 7.5 with D11-12 at pH 4.5 demonstrates the domain rearrangement.

The same conformational change to the D9-10 and D11-12 di-domains is also seen upon comparison of bovine D1-14 at pH 4.5 and bovine D4-14 at pH 7.5 in complex with IGF2 (PDB 6UM1 and 6UM2). However, the observation of this structural rearrangement in the absence of IGF2 (as in the ligand free structures of D9-10, D7-11, D11-12 and D11-14, PDB 6Z30, 6Z31, 2V5N and 2V5O respectively) confirms that this structural rearrangement is not the result of IGF2 binding but a pH induced change. Could these pH sensitive His-Pro interactions at the interfaces of D9-10 and D11-12 represent two possible hinge points for the collapse of the receptor to a more compact structure at low pH?

To further explore the effects of pH, analytical SEC and SAXS were performed with D9-10. During analytical SEC at pH 7.5, D9-10 elutes as a single broad peak (Figure 63A). The broadness of the peak, which may be attributed to the two flexible N-linked glycans of D9 and possible exchange between monomeric and dimeric forms, makes molecular weight determination inaccurate. However, at 64 kDa, the apparent molecular weight is near the calculated dimeric molecular weight of 72 kDa.

Before further gel filtration, D9-10 was checked by non-native ESI-MS at pH 7.5 and 5.5 (Appendix Figure 7) and resulted in the same mass, ruling out any unexpected degradation. When the pH was lowered from pH 7.5 to pH 5.5, D9-10 eluted later and sharper, suggesting a more homogeneous, compact structure, with an apparent molecular weight of 57 kDa (Figure 63A). This is consistent with recent SEC experiments performed by Olson *et al.*, which found that, upon lowering the pH from 6.5 to 4.5, the stokes radius of human D1-5, D7-15 and D1-15 decreased.[138] This indicates formation of a more compact structure at acidic pH.

Hancock *et al.* and Marron-Terada *et al.* have demonstrated that D9 does not undergo acid-dependent cargo dissociation as efficiently as D3.[132,137] For example, up to 90 % of bovine D7-9 and D7-11 was retained on a pentamannosyl phosphate column at pH 4.6.[132] Therefore, analytical SEC was repeated at lower pH in attempt to disrupt the N-linked glycan bridge of D9 and dissociate D9-10 to a monomer. However, when the pH was lowered to pH 4.0 the elution volume decreased suggesting aggregation.

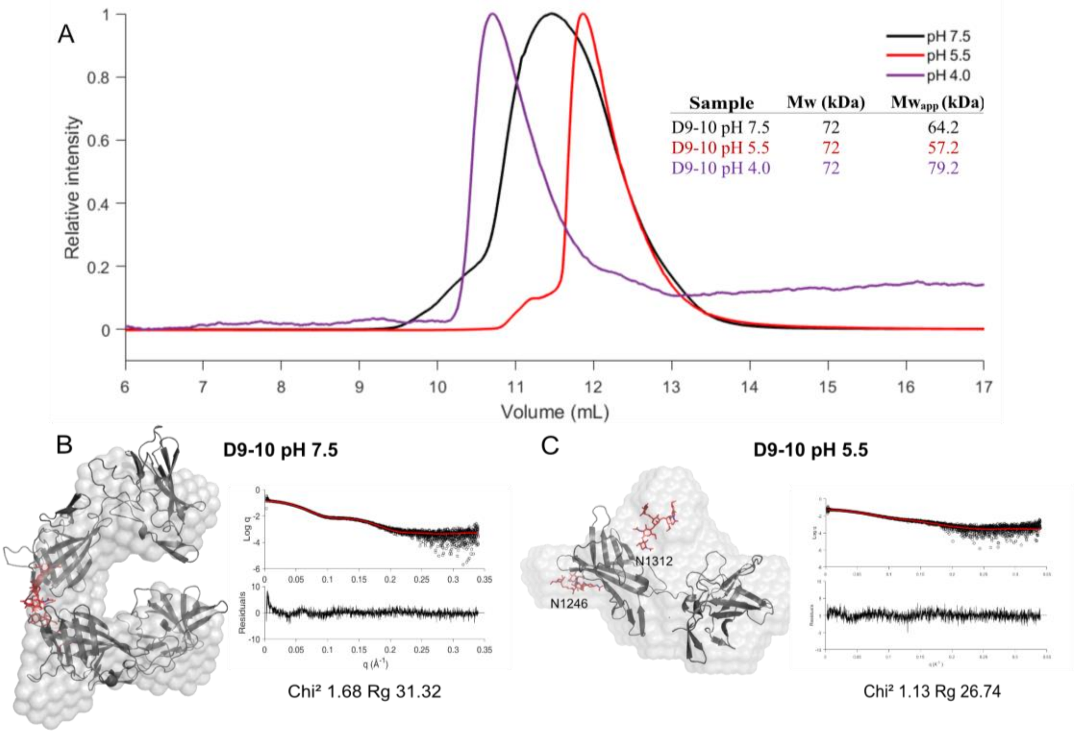


Figure 63: Analytical SEC and SEC-SAXS of D9-10 at pH 7.5 and pH 5.5. A: Analytical SEC of D9-10 at pH 7.5 (black), pH 5.5 (red) and pH 4.0 (purple). **B:** SEC-SAXS of D9-10 at pH 7.5. **C:** SEC-SAXS of D9-10 at pH 5.5. For each SAXS result: *Ab initio* DAMMIN bead density shape envelope of D9-10 represented as a surface (light grey) overlaid with a 1-state-optimised model. For each model, the *I*(q) vs q plot for a 1-state optimised model (red line) against the experimental scattering data (black) with the residual difference plot shown underneath. The χ^2 value of the fit and R_g value of the model are shown. Glycans are shown in red.

The SEC data at pH 5.5 suggests a change in quaternary structure of D9-10, but uncertainties associated with SEC mean monomer or monomer/dimer equilibrium cannot be ruled out. As described earlier, SAXS analysis of glycosylated D9-10 at pH 7.5 revealed the two domains form a folded flat entity with a calculated molecular weight associated with a dimer (apparent molecular mass 70.5 kDa (as calculated by SAXSMoW 2.0), expected dimeric molecular mass 70.8-71.6 kDa, Appendix Table 6). *Ab initio* shape envelope modelling of D9-10 was consistent with the structure of D9-10 in the same conformation as the D7-11 crystal structure, with symmetric cross-binding of the two N-glycan moieties at N1312 of D9 (Figure 63B). A χ^2 value of 1.68 for a 1 state model of D9-10 indicated that these models were a good fit to the experimental data.

SAXS of glycosylated D9-10 at pH 5.5 gave an apparent molecular mass of 44.7 kDa (as calculated by SAXSMoW), which is between the expected monomeric and dimeric molecular mass (1.3 times monomeric and 0.75 times dimeric molecular mass, Appendix Table 6). Similarly, the radius of gyration (R_g) is intermediate of the monomeric and dimeric values obtained at pH 7.5 (R_g ~29 Å at pH 5.5 versus ~25 and ~32 Å for pH 7.5 monomer and dimer, Appendix Table 6). This would suggest that, at pH 5.5, there is exchange between the monomer and dimer.

Multi-state modelling with the D9-10 dimeric model that fits the pH 7.5 scattering data gave a χ^2 value of 31.7 for a one-state system and 16.7 for a two-state system when modelled to the pH 5.5 data indicating an incorrect fit. Similarly, modelling the recent bovine cryo-EM structures of D9-10 at pH 4.5 and pH 7.4 (PDB 6UM1 and 6UM2) gave χ^2 values of 7.37 and 7.57 respectively, again indicating an incorrect fit to the scattering data. However, a good fit was found between the scattering data and monomeric D9-10 glycosylated at both sites (with glycan 4, $GlcNAc_2Man_3GlcNAc$, Figure 33B) (χ^2 value of 1.13, Figure 63C, Appendix Table 6).

Although further work/ repeats are required to confidently define the oligomeric state of D9-10 at low pH, the SAXS data presented here (Figure 63B, 63C) suggests that low pH may disrupt the bridging D9-glycan interaction resulting in exchange between dimeric and monomeric forms. This would be consistent with CI-MPR cargo release in the acidic environment of the late endosome (pH <6.0).

3.5. Conclusions

This chapter describes the successful application of methods developed in chapter 2 for the expression and purification of single CI-MPR domains in insect cells to multi-domain constructs encompassing the elusive sugar binding domain, D9.

A di-domain construct of D9-10 was expressed and secreted from Sf21 insect cells with greater than 90 % purity and with a yield of 7-13 mg/ L, allowing structure determination by X-ray crystallography. This was the first high-resolution (1.5 Å) structure of D9, which binds M6P mono-esters specifically with high affinity and independently of neighbouring domains. D9-10 contains two predicted N-linked glycosylation sites that were shown by mass spectrometry to be modified with glycans typical of insect cells. The D9-10 crystal structure and SAXS data reveal that one of these N-linked glycans forms a bridge between the two D9s, driving dimerisation of the rigid D9-10 unit.

Two tetra-domain constructs of D7-10 were also expressed in Sf21 insect cells. However, these constructs were not suitable for structural characterisation due to low yield, N-terminal truncation and the presence of five heterogeneous, glycans that could not be enzymatically removed. This impaired further biophysical characterisation and structure determination by X-ray crystallography.

However, the high-resolution crystal structures of human D9-10 and D8 (chapter 2) have improved the refinement of the human D7-11 crystal structure for publication. As a 160 kDa penta-domain homodimer, this is the largest fragment of the human CI-MPR determined to date and comprises the core of the CI-MPR extracellular region. Despite the addition of neighbouring domains, D9-10 of D7-11 exhibits the same domain orientation and interface as observed in the structure of D9-10 alone. Both structures reveal dimerisation through a bridging N-linked glycan.

4. The full extracellular region of CI-MPR

4.1. Introduction and aims

To date, high-resolution X-ray or NMR solution structures have been determined for bovine D1-3 (PDB 1SZ0, 1SYO, 1Q25[103,136]), bovine D5 (PDB 2KVA, 2KVB[131]), human D1-5 (PDB 6P8I, 6V02), human, echidna, chicken and opossum D11 (PDB 1GP0, 2LLA, 2L21, 2L2G[101,124]), human D11-12 (PDB 2V5N[101]) and human D11-14 (PDB 2V50[101]) as well as D11 variants engineered for high-affinity ligand binding (Table 9).[127] However, there exists no structure for the full extracellular region of the human CI-MPR.

Domain(s)	Method	Organism	Expression host	PDB
D1-3	X-ray diffraction	Bovine	Tn-5B1-4	1SYO, 1SZ0, 1Q25
D5	NMR	Bovine	*P. pastoris*	2KVA
D1-5	X-ray diffraction	Human	Sf9	6P8I, 6V02
D8	X-ray diffraction	Human	Sf21	6Z31
D9-10	X-ray diffraction	Human	Sf21	6Z30
D7-11	X-ray diffraction	Human	HEK293T	6Z32
D11	X-ray diffraction	Human	*E. coli*	1GP0, 1GP3
D11	NMR	Chicken, echidna, opossum	*E. coli*	2L21, 2LLA, 2L2G
D11-IGF2	NMR	Human	*E. coli*	2CNJ
D11-12	X-ray diffraction	Human	CHO	2V5N
D11-13-IGF2	X-ray diffraction	Human	CHO	2V5P
D11-14	X-ray diffraction	Human	CHO	2V5O

Table 9: Recombinant CI-MPR extracellular domains whose structures have been solved to date. Determination of the structures of human D8, 9-10 and D7-11 are described in chapters 2 and 3.

The first model for the CI-MPR extracellular region was proposed by Olson *et al.* in 2004 and comprised five tri-domain units (D1-3, D4-6, D7-9, D10-12 and D13-15) each containing a functional binding domain (D3, D5, D9, D11 and D15) (Figure 64A).[103] This tri-domain arrangement is supported by the observation that D1-3 and D4-6 form proteolytically stable units upon treatment of bovine extracted CI-MPR with the protease subtilisin.[103,170] This model is also consistent with the compact crystal structures of bovine D1-3 with M6P or mannosylated glycan bound (PDB 1SYO/ 1SZ0 and 1Q25 respectively).[136] Homology models of the remaining, uncharacterised domains (D7-9, D10-12, D13-15) were used.[103]

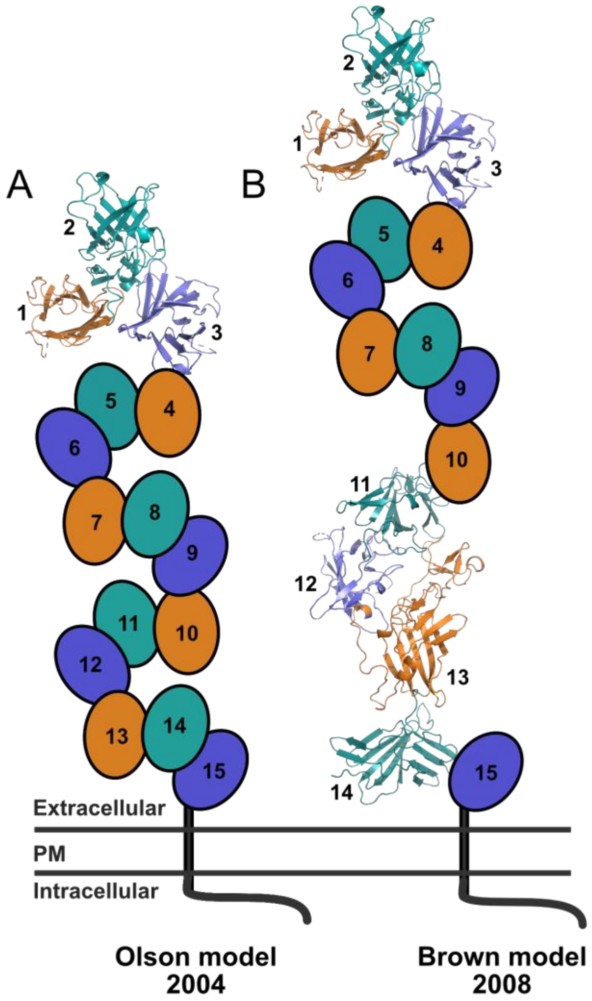


Figure 64: Early models of the CI-MPR extracellular region. A: The first model was proposed by Olson *et al.* in 2004 and consisted of five compact tri-domain units based upon the crystal structure of bovine D1-3 (PDB 1SYO/1SZ0).[103] **B:** This model was updated in 2008 by Brown *et al.* to incorporate the crystal structure of human D11-14 (PDB 2V5O).[101]

This model was later modified in 2008 by Brown *et al.* (Figure 64B) to include the crystal structure of human D11-14 (PDB 2V5O), while maintaining the three N-terminal tri-domains D1-9.[101] Brown *et al.* also proposed a dimeric model of the extracellular region in which domains 3, 5, 9, 12 and 15 formed homodimeric contacts. However, while dimerisation of D1-3 has been observed in crystals belonging to two different space groups (orthorhombic $P2_12_12_1$ and monoclinic $P2_1$), the oligomeric state of D1-3 has not been demonstrated in solution.[103,136]

Until recently, there has been very little work performed on the full CI-MPR. In 1989 Tong *et al.* extracted CI-MPR from bovine livers and determined, by radio-labelled assays, its binding affinity (K_D) for M6P monosaccharide, pentamannosyl phosphate and the glycoprotein β-galactosidase to be 7 μM, 6 μM and 0.02 μM respectively.[135] More recently, Olson *et al.* 2015 have used SPR to determine the binding affinity of the extracellular region (D1-15) of bovine

CI-MPR for the lysosomal enzyme GAA tagged with M6P mono-ester and di-ester (K_D 4.5 ±0.7 nM and 51 ±1 nM respectively).[130] Olson *et al.* were the first to recombinantly express bovine D1-15 of the CI-MPR and did so using Sf9 insect cells.[130] In 2019 Hughes *et al.* expressed human CI-MPR D1-15 using the mammalian HEK293T cell line and determined the binding affinity of human D1-15 for IGF2 (0.14 nM) and M6P-tagged β-glucuronidase (52.3 nM).[223] Most recently, Dwyer *et al.* 2020, have recombinantly expressed the extracellular region of human CI-MPR using the mammalian (human fibrosarcoma) cell line HT1080.[224]

Having demonstrated that Sf21 insect cells are capable of expressing soluble, folded single and multi-domain CI-MPR constructs (chapters 2 and 3), human D1-15 was expressed here in insect cells with the goal of structure elucidation by cryo-EM.

4.2. Expression and purification of human D1-15

DH10 EMBacY *E. coli* cells were transformed with a synthetic gene encoding human D1-15 of the CI-MPR with an N-terminal RPTPμ signal sequence and C-terminal His_6 tag (Appendix Section 8.1., GeneArt, Thermo Fischer). D1-15 EMBacY bacmids were purified from white colonies and used to transfect Sf21 insect cells as described previously for D11, D7, D8, D9-10 and D7-10 (chapters 2, 3 and 6). Recombinant virus was harvested, amplified and used to infect subsequent cultures.

The optimal harvest time was determined by SDS-PAGE analysis and YFP expression levels (Figure 65A, 65B). Due to the presence of an N-terminal signal sequence, soluble D1-15 should be secreted from the Sf21 cells. SDS-PAGE analysis revealed a band at ~250 kDa corresponding to D1-15 (expected monomeric, unglycosylated molecular weight 248.6 kDa) in the insoluble cell fraction that increases in intensity over time (Figure 65A). There was no visible band on the SDS-PAGE for secreted D1-15 (Figure 65A). Nonetheless, the culture media was harvested by centrifugation on DPA +72 before loading onto an Ni^{2+} IMAC column. A small absorbance peak (190 mAU) was observed upon application of a gradient of imidazole (Figure 65C). The peak species was confirmed by SDS-PAGE to be D1-15 (Figure 65D). The slight band-shift observed under non-denaturing, non-reducing (N) conditions versus denaturing, reducing conditions (R) (Figure 65D and Figure 66) is indicative of a more compact, disulphide bonded structure, suggesting that D1-15 is folded. However, due to the large size of D1-15 (250 kDa monomeric molecular weight), this could not be confirmed by

1D ^{1}H NMR. Its large size and heterogeneous glycans also prevented acquisition of an accurate mass of D1-15 by ESI-MS.

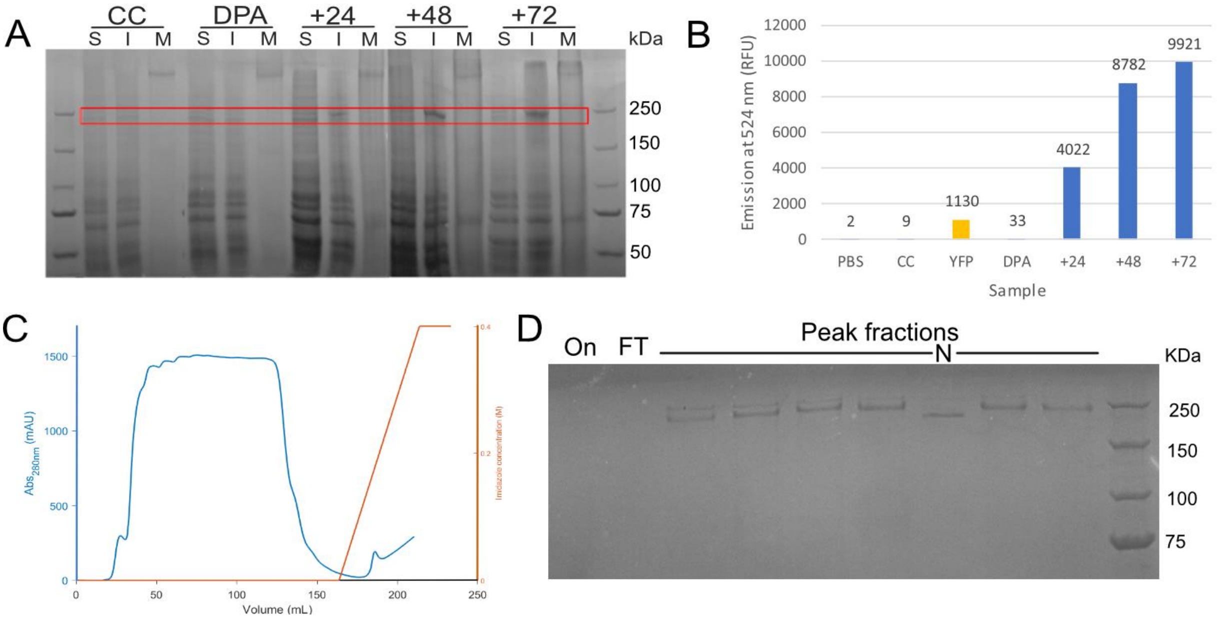


Figure 65: Expression and purification of D1-15 from insect cells. A: 5 % acrylamide SDS-PAGE analysis of D1-15 expression in Sf21 insect cells. Samples of media (M), insoluble cell fraction (I) and soluble cell fraction (S) were collected every 24 hours from the day after proliferation arrest (DPA) to determine optimal harvest time. Samples were also collected from an uninfected cell control (CC). D1-15 (within the red box) has an expected molecular weight of ~250 kDa. **B:** YFP emission was also measured every 24 hours to determine harvest time. **C:** D1-15 was harvested from culture media by Ni^{2+} IMAC. **D:** 5 % acrylamide SDS-PAGE confirmed purification of D1-15 from the culture media. All peak fractions except that labelled N (non-denatured, non-reduced) were denatured by boiling and reduced by DTT. 'On' corresponds to media loaded onto the column, 'FT' to the flow through and 'Fractions' to the peak fractions eluted following imidazole addition.

4.3. <u>Initial characterisation of human D1-15</u>

Analytical SEC resulted in two small peaks (Figure 66). The first peak elutes at the column void volume (~7 mL) and corresponds to aggregated D1-15. However, the second peak (~10.9 mL) has an apparent molecular weight of 506 kDa (Appendix Table 8). This closely matches the expected dimeric mass of D1-15, which could range between an unglycosylated molecular mass of 497 kDa and 517 kDa, 526 kDa and 536 kDa assuming 50 %, 75 % and 100 % N-linked glycosylation with high-mannose glycans typical of on-pathway insect cell glycosylation.

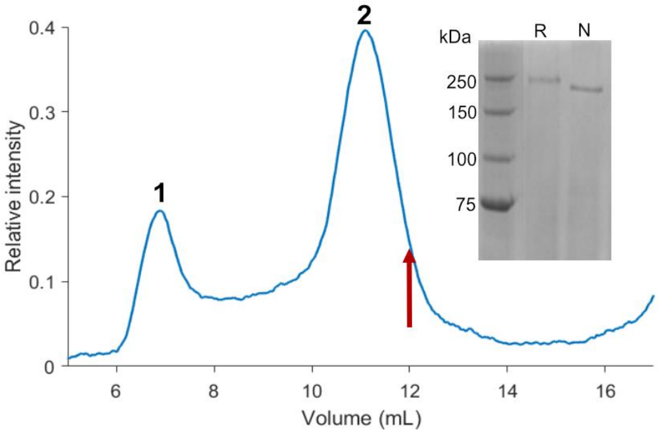


Figure 66: Analytical SEC of D1-15 reveals the major species (2) to be dimeric. The first peak (1, ~7.0 mL) corresponds to aggregated D1-15. The second peak (2, ~10.9 mL) corresponds to D1-15 with an apparent molecular weight of 506 kDa and expected dimeric, unglycosylated molecular weight of 497 kDa. The red arrow marks the expected elution volume (~12.0 mL) of monomeric D1-15 (248 kDa). SDS-PAGE analysis of the major species (2) under denatured, reduced conditions (R) and non-denatured, non-reduced conditions (N).

Negative stains of D1-15 were prepared and imaged with the help of Professor Christiane Berger-Schaffitzel and Dr Sathish Yadav (University of Bristol). With a resolution limit of ~15 Å, negative staining is often used as a screening step before protein structure determination by single-particle cryo-electron microscopy (cryo-EM).[225] During negative staining the protein sample is adsorbed onto a carbon coated copper grid, stained with the heavy metal solution uranyl acetate or uranyl formate and imaged by transmission electron microscopy (TEM).[226]

Negative stain images were collected for D1-15 in the absence of any ligand, with 10-fold excess M6P monosaccharide and with 1.5-fold excess of IGF2 (Figure 67). The protein molecules exclude the uranyl acetate stain and thus appear with white contrast.[226,227] The negative stain of D1-15 alone (120 μM) at pH 7.5 (Figure 67A) revealed a uniform sample suitable for cryo-EM. However, a more concentrated sample would be required to facilitate particle picking for classification and cryo-EM data collection. In samples of D1-15 plus IGF2 at pH 7.5 (120 μM D1-15 and 180 μM IGF2) large clumps of aggregated protein are observed (Figure 67B). This is due to the pH sensitivity of IGF2. Thus, formation of a D1-15-IGF2 complex at neutral pH may require mixing/ complex formation at very dilute concentrations before concentration and imaging.

D1-15 plus a 10-fold excess of M6P monosaccharide at pH 7.5 resulted in high quality negative stain micrographs (Figure 67C). Visual comparison to micrographs of D1-15 alone suggest the receptor forms a more compact structure in the presence of M6P and shows some heterogeneity. 2D classification was performed to understand if this heterogeneity might have arisen from either different oligomeric states, different conformations or simply different orientations of D1-15 on the grids.[226] Approximately 280 negative stain micrographs were collected of D1-15 plus M6P and ~15,000 particles picked for 2D classification (Figure 67D). Using the software package Scipion [228] images of D1-15 particles were picked, aligned, sorted into 36 classes and averaged.[226] The resulting 2D classifications (Figure 67D) show an elongated form of the receptor and a more compact form at different orientations. Several of the images demonstrated a clear multi-domain structure.

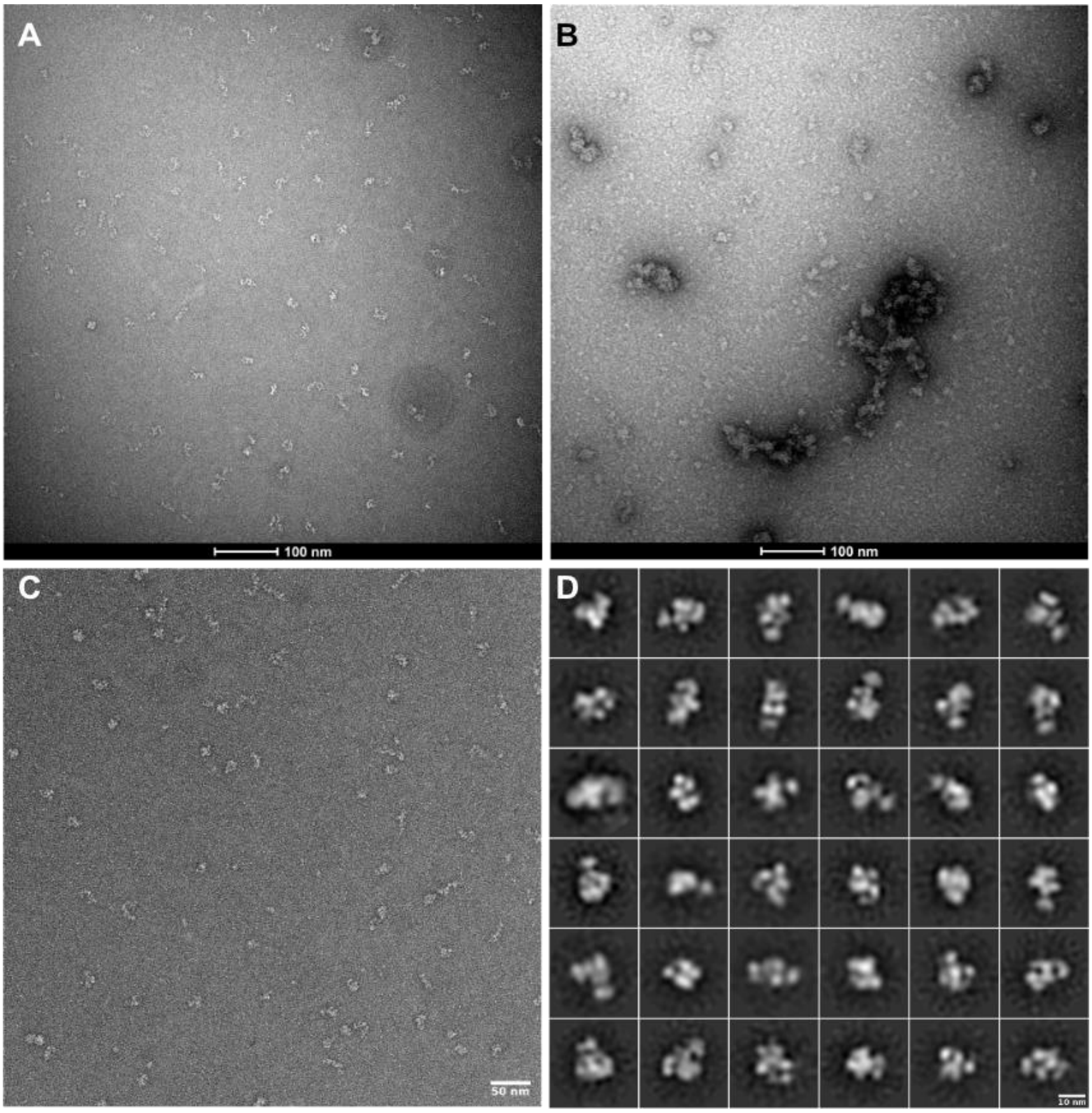


Figure 67: Transmission electron micrographs of D1-15 negatively stained with uranyl acetate. A: 120 μM of D1-15 alone at pH 7.5. **B:** 120 μM of D1-15 with 180 μM IGF2 at pH 7.5. IGF2, which is pH sensitive, has aggregated. **C:** 120 μM of D1-15 with 1200 μM of M6P monosaccharide at pH 7.5. All micrographs were collected at 100 k x magnification on an FEI Tecnai 20 200 kV microscope (Wolfson Bioimaging Facility, University of Bristol) with the help of Professor Christiane Berger-Schaffitzel and Dr Sathish Yadav.

4.4. The structure of bovine D1-15

Recent work by Wang *et al.* 2020 has determined the cryo-EM structure of bovine CI-MPR D1-14 to 3.5 Å at the physiological pH of the lysosome, pH 4.5 (Figure 68A, PDB 6UM1). At this pH, the extracellular region of the CI-MPR forms a compact, helical structure.[128] Seven sub-groups (D1+D3, D2+D5, D4+D7, D6+D9, D8+D11, D10+D13 and D12+D15) form with the βE-I surface of each domain packing against the βE-I surface of its partner.[128]

However, the structure of bovine CI-MPR D1-14 could not be obtained at neutral pH (pH 7.4 as found at the plasma membrane) in the absence of ligand.[128] Only upon incubation with IGF2 could a low-resolution (4.3 Å) structure of the bovine CI-MPR be obtained at neutral pH (Figure 68B, PDB 6UM2).[128] The result is an elongated structure of D4-14, with D1-3 and D15 being omitted due to flexibility.[128] In the structure of D4-14 the domains have rearranged and the seven sub-group interactions have been disrupted.

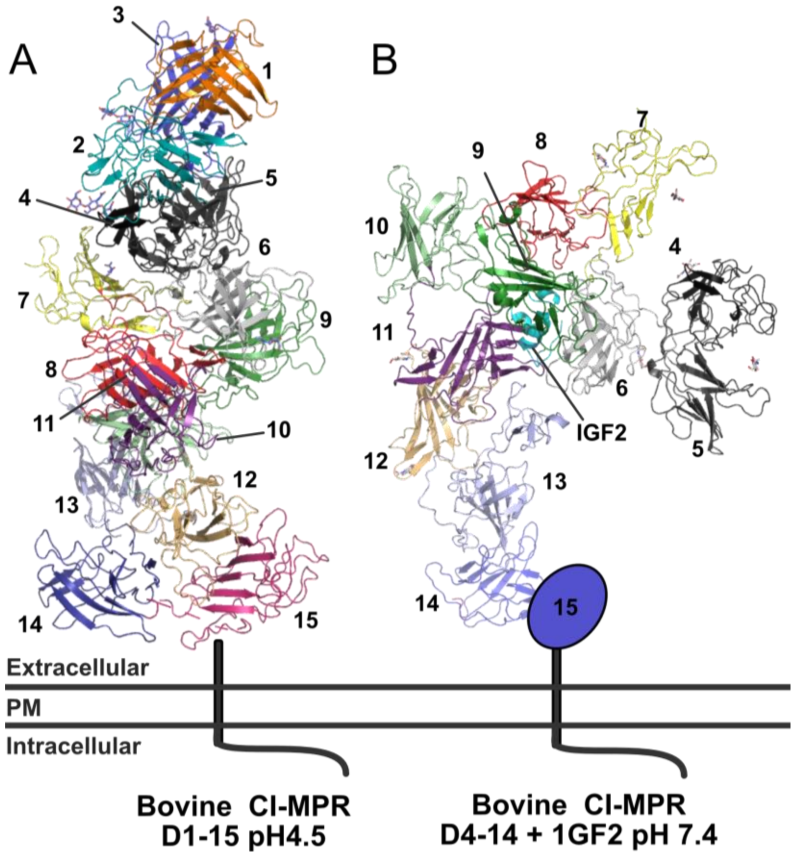


Figure 68: Cryo-EM structures of bovine CI-MPR extracellular region. A: The cryo-EM structure of bovine CI-MPR D1-15 (PDB 6UM1) at pH 4.5 demonstrates a compact, helical structure. **B:** The cryo-EM Structure of bovine CI-MPR D4-14 (PDB 6UM2) in complex with IGF2 (cyan) at pH 7.4. Structures of D1-3 and D15 could not be determined due to flexibility.

4.5. An updated model of D1-15

Combining the crystal structure of D7-11 with the previously characterised domains (D1-3, D5, D11-14) produces a model for the CI-MPR extracellular region (D1-14) that is different to previous proposals (Figure 69A). D4-6 is assumed to adopt a stable tri-domain which, alongside D1-3, maintains the two tri-domain model at the N-terminus of the CI-MPR. D7-10 and D11-14 do not conform to the tri-domain pattern but instead form two sequential tetra-domain units. Therefore, the C-terminal half can be modelled from the crystal structures of D7-11 and D11-14 using the alignment of D11 in the D11-14 structure. D15 was not modelled as it is anchored to the cell membrane, lacks any structural data and is susceptible to cleavage by TACE protease to release soluble CI-MPR (sCI-MPR).[224] Therefore, it is currently not clear how D15 interfaces with the remainder of the receptor.

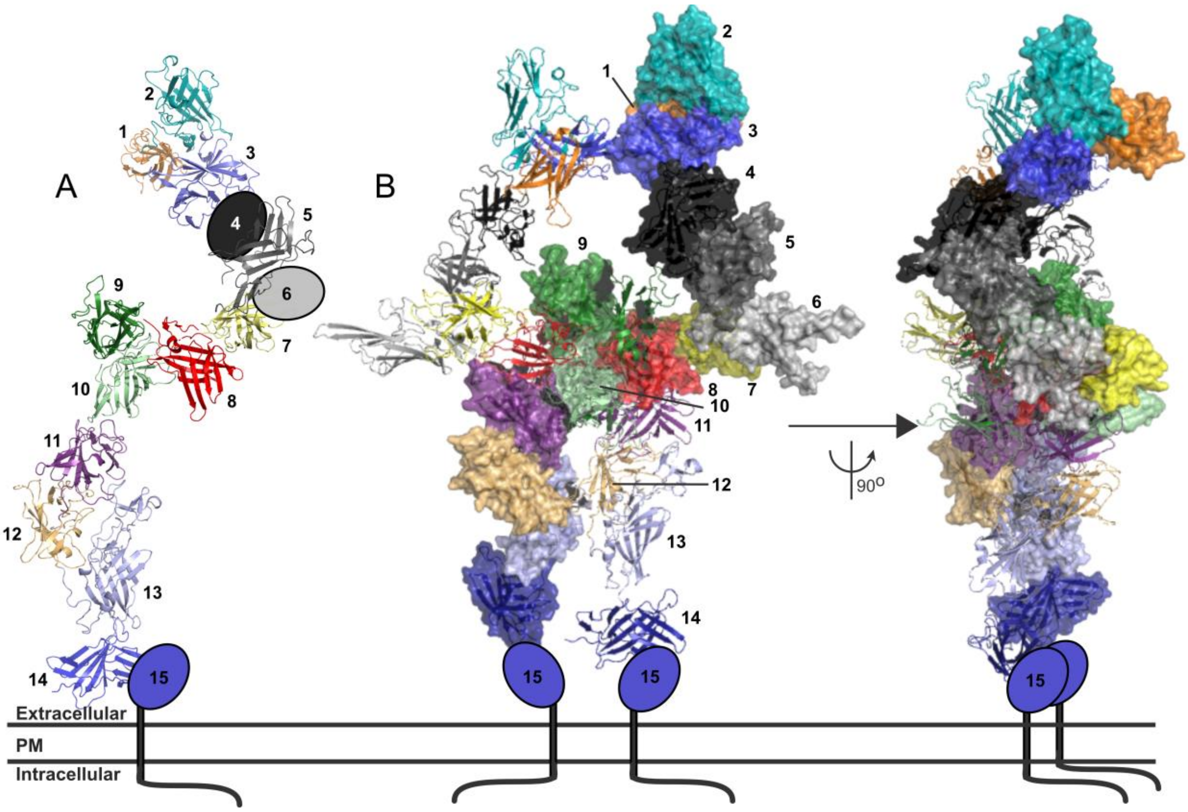


Figure 69: An updated model of the CI-MPR extracellular region.[70] **A:** A single chain of our updated model of D1-15 is based upon the crystal structures of bovine D1-3, human D7-11, human D11-14 and solution NMR structure of bovine D5 (PDB 1SZ0, 6Z32, 2V5O, 2KVA respectively). **B:** The dimeric form of our updated model. Chain A (cartoon) crosses over chain B (surface render) in the centre. Only the structures of D4, D6 and D15 have not yet been determined to high-resolution. The updated model was built by former PhD student Ryan Nicholls.

The physiological oligomeric state of the CI-MPR at neutral pH is poorly understood, with both monomeric and dimeric forms of CI-MPR purified from bovine livers observed by native gel electrophoresis.[171] When studied by gel filtration and sucrose gradient centrifugation, bovine extracted CI-MPR exists predominantly as a monomer.[171,229] Similarly, recombinant bovine D7-9 and D7-11 were determined monomeric by cross-linking, size exclusion chromatography and sucrose density gradient centrifugation.[132] However, the purification method used in each of these studies involved pentamannosyl phosphate affinity chromatography and might be selective for monomers as CI-MPR dimers mediated by N-linked glycosylation (as seen for bovine D1-3, human D9-10 and human D7-11) may not be retained on the column. Similarly, mammalian expressed human CI-MPR D1-15 (purified by similar affinity chromatography methods) was also determined monomeric by SEC-MALS.[70,224]

On the other hand, truncated human CI-MPR constructs and fusion proteins of human CI-MPR extracellular domains with epidermal growth factor receptor are capable of forming dimers in the plasma membrane of mammalian cells (HEK293T).[171,230] Furthermore, dimers of bovine CI-MPR observed in the membranes of mouse cells were found to be cross-linked by the multi-valent M6P-tagged ligand β-glucuronidase, which resulted in an increased rate of β-glucuronidase internalisation.[70,229]

The updated model maintains the crystal structures of D1-3, D7-11 and D11-14 in their dimeric forms, creating an elongated structure with approximate dimensions of 260 by 130 by 92 Å assembled around the D7-11 core (Figure 69B). In this updated model the two CI-MPR chains intersect and cross at D8-10, resulting in D7 sitting above D14 of the opposing chain (Figure 69B). D3, D5, D9 and D11 maintain solvent accessibility for (known) ligand binding. The role of D12 in receptor dimerisation has been proposed,[101,171,231] but the mechanism of dimerisation in terms of protein-protein interactions has not been explicitly proven. The assembly of the D7-10 hub and the rigidity of the D11-12 interface suggests minimal interactions between D12s but does permit extensive inter-chain contacts between D12 and D13. Similarly the positioning of the two D7s seems to preclude the incorporation of D5-D5 contacts suggesting D5 may not act as a dimerisation point or D7-D8 contact may be flexible.[131] The D4-6 structure remains unknown but could conceivably mark a further point at which the chains intertwine and cross and may be more compact than represented here.[70]

Despite the crowding caused by this coiled dimerisation, the updated model remains compatible with simultaneous multi-ligand binding. For example, the M6P binding site of D3 is positioned on the opposite face of the D1-3 tri-domain to the uPAR and plasminogen binding sites at the N-terminus of D1.[103,114,142] This should allow both M6P tagged proteins and the M6P independent ligand uPAR or plasminogen to bind simultaneously.[103,142] However, addition of β-glucuronidase has been shown to reduce the binding of uPAR.[142] A similar decrease in uPAR binding was observed upon addition of IGF2.[142] However, IGF2 binding at D11 is unlikely to affect the structure of D1, as comparison of the crystal structures of D11-13 in complex with IGF2 and D11-14 unbound (PDB 2V5P and 2V5O respectively) show no structural rearrangement upon IGF2 binding.[101]

Similarly D3, D5, D9 and D11 are distributed so that IGF2 and M6P tagged proteins may bind simultaneously.[101,103] For example, as in the previous models of the extracellular region,[101,103] the binding sites of D9 and D3 are on opposing surfaces of the receptor and therefore binding of IGF2 should not occlude D9 and D3. However, Kiess *et al.* found that addition of IGF2 reduced β-galactosidase binding by ~75 %.[232]

There have also been multiple reports that bi-phosphorylated oligosaccharides exhibit higher affinity binding to the CI-MPR than mono-phosphorylated ligands [135,233] Tong *et al.* reported high affinity CI-MPR binding of a glycan containing two M6P mono-esters, which would have a maximum distance of 30 Å between the two M6P residues.[135] To satisfy this, Olson *et al.* proposed that the extracellular region of the receptor exists in a dynamic state with the hinge regions between the two N-terminal tri-domains (D1-3 and D4-6) flexing to bring the M6P mono-ester sites of D3 and D9 to within 45-85 Å.[103] Our model, however, proposes a dimer in which the bi-phosphorylated M6P oligosaccharide would bind two D9s, a distance of ~22 Å. The two D9 binding sites are also in close enough proximity to bind the bi-phosphorylated molecular rulers developed by Fei *et al.*, which had a distance of 16-26 Å between the two M6P residues.[234] The distance between D3 and D9 of the same chain (intramolecular) and the opposing chain (intermolecular) in our model is ~58 Å.

This updated model of D1-14 has since been validated by the low-resolution cryo-EM structures of bovine extracted CI-MPR (PDB 6UM1 and 6UM2) (Figure 68).[128] Superimposing D4-14 of chain A of our model with the structure of bovine D4-14 at pH 7.4 gives an RMSD value of 29.9 Å over backbone atoms and similar dimensions (204 by 112 by 63 Å versus 230 by 170 by 70 Å for bovine D4-14). Furthermore, superimposing D7-11 of one chain of the dimeric core with corresponding domains of the extended D4-14-IGF2 complex (PDB 6UM2) revealed a close inter-domain arrangement (Figure 70), with an RMSD value of 4.9 Å (PyMOL cealign structure-based alignment). This is versus 7.9 Å when superimposed with bovine D7-11 at pH 4.5. Thus, identical packing around D9 can be derived from interaction with one of two distinct ligands, IGF2 or glycan, or, more likely, by neutral pH conditions.[70]

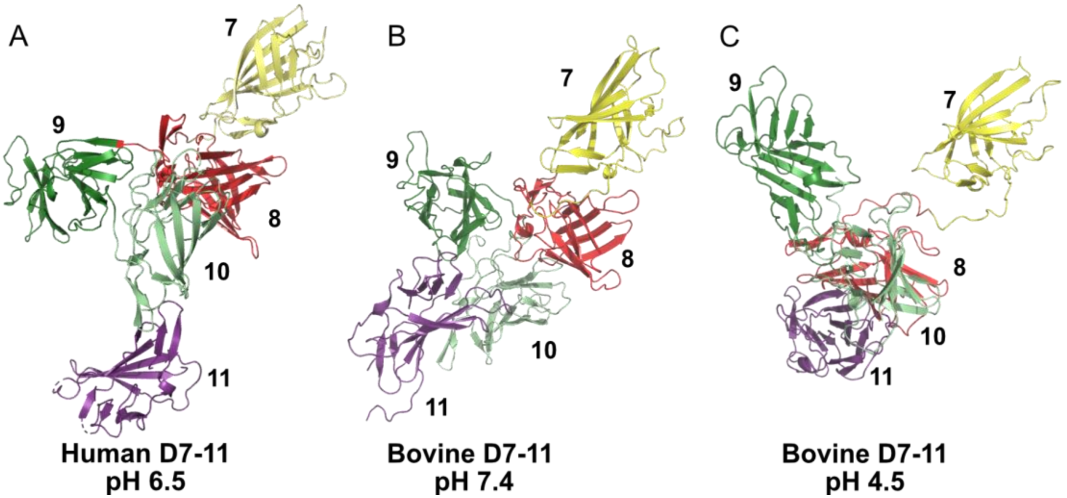


Figure 70: Comparison of human and bovine D7-11 structures.[70] **A:** A single chain from the crystal structure of human D7-11 at pH 6.5 (PDB 6Z30). **B:** D7-11 from the cryo-EM structure of bovine D4-14 in complex with IGF2 at pH 7.4 (PDB 6UM2). Superimposition of human and bovine D7-11 at near neutral pH shows the same domain orientation and gives an RMSD of 4.9 Å. **C:** D7-11 from the cryo-EM structure of bovine D1-14 at pH 4.5 forms a more compact arrangement (PDB 6UM1). Superimposition of human D7-11 at pH 6.5 and bovine D7-11 at pH 4.5 gives an RMSD value of 7.9 Å. PyMOL cealign structure-based alignments.

Our updated model of D1-14 is further supported by the recent crystal structures of human D1-5. Superimposition of human D1-5 at pH 7.0 (PDB 6V02) and the corresponding domains of our model gives an RMSD value of 7.2 Å over backbone atoms (versus an RMSD value of 13.6 Å when superimposed with D1-5 at pH 5.5, PDB 6P8I). Similarly, identical packing was observed in the structure of human D1-5 at pH 5.5 and the corresponding domains of bovine D1-14 at pH 4.5 (PDB 6P8I and 6UM1 respectively) (Figure 71). Superimposition of these structures, which share the same compact domain arrangement, gives an RMSD value of 1.4 Å, over backbone atoms revealing a rigid N-terminal unit formed at low pH in both bovine and human CI-MPR.

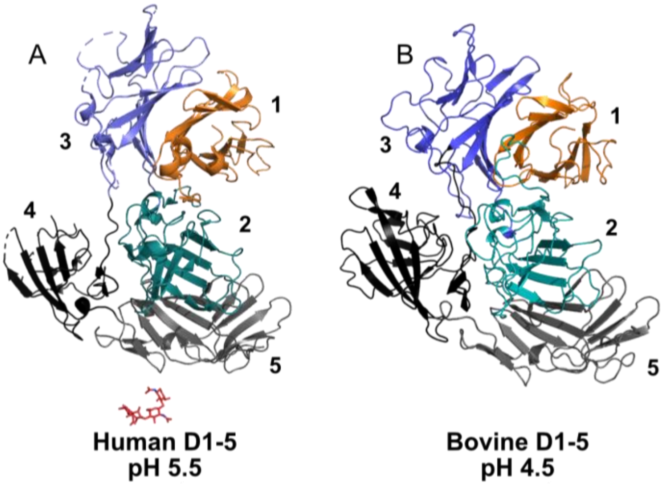


Figure 71: Comparison of human and bovine D1-5 structures at acidic pH. A: The crystal structure of human D1-5 at pH 5.5 (PDB 6P8I). The N-linked glycan (GlcNAc$_2$Man$_2$) at N591 of D5 is shown as red sticks. **B:** The cryo-EM structure of bovine D1-5 at pH 4.5 (PDB 6UM1) forms the same compact domain arrangement. Superimposition gives an RMSD value of 1.4 Å over backbone atoms.

4.6. Conclusions

With a monomeric molecular weight of ~250 kDa and containing 19 predicted N-linked glycosylation sites and 59 disulphide bonds, the full extracellular region (D1-15) of the CI-MPR is a complex and ambitious target ideal for structure determination by cryo-EM. This chapter describes the successful expression of D1-15 of the human CI-MPR in insect cells using the protocols developed for single and multi-domain constructs (chapters 2 and 3). Although the expression protocol requires optimisation to increase the ratio of secreted to insoluble D1-15, preliminary characterisation by analytical SEC and negative stain EM has been performed.

Determination of the structure of human D7-11 (chapter 3) has allowed an updated model of the CI-MPR extracellular region to be proposed that was validated by the recent, low-resolution cryo-EM structure of bovine D4-14. Supported by the observation that insect expressed, human D1-15 is dimeric in solution, the updated model of D1-15 maintains the crystal structures of D1-3, D7-10 and D11-14 in their dimeric forms. Dimerisation through a bridging N-linked glycan has now been observed in insect expressed bovine D1-3, insect expressed human D9-10 and mammalian expressed human D7-11.[103] This work opens up fresh avenues for investigation into the ground state of the CI-MPR at the cell surface, how receptor dimerisation is regulated and how the D9 M6P binding site is unmasked to allow ligand binding.

5. Engineering a synthetic lectin

5.1. Introduction and aims

Background

In 2016 Frago *et al.* developed the first high-affinity IGF2 specific ligand trap by fusing high-affinity D11 mutants with the Fc region of human IgG1.[127] This has proven successful in inhibiting *in vivo* IGF2 signalling in hypoglycaemic mice and may be a suitable therapeutic for CI-MPR deficient cancers.

Therefore, is it possible to make a similar CI-MPR-based trap for M6P-tagged proteins? Could an M6P trap be administered to stratified cancer patients alongside the IGF2 trap? Or could an M6P trap be used to treat lysosomal storage disorders, where it may prove useful in conjunction with enzyme replacement therapies (ERT). For example, Cheng *et al.* have demonstrated (*in vitro* and in mice models) that endocytosis of recombinant human acid α-glucosidase (rhGAA), which is a recognised ERT for the treatment of the lysosomal storage disorder Pompe disease, is enhanced by tagging rhGAA with M6P-tagged glycans.[168] Furthermore, Bali *et al.* have improved the efficiency of rhGAA ERT in mice by up-regulating CI-MPR.[169]

Additionally, could an M6P trap be adapted to bind other carbohydrate ligands? Recognition of glucose, for example, could facilitate glucose monitoring and diabetes treatments (through glucose-responsive insulin).[235] Evidently, a selective synthetic lectin has huge potential as a therapeutic, drug carrier, diagnostic or research tool. Thus, we sought to take some preliminary steps to engineer a synthetic lectin based upon the MRH domains of the P-type lectins.

Example synthetic lectins

There are two broad approaches to creating a synthetic lectin: 1) employing organic chemistry methods and 2) engineering biological molecules. Each of these methods can be classified further. For example, the first of these approaches exploits the interaction between boronic acids and sugars to form boronate esters and these so-called boronolectins were initially developed to bind glucose.[236,237] However, boronolectins may have limited application as boronic acids exhibit preferential binding to furanose monosaccharides over pyranoses.[238] Nonetheless, boronolectins that are capable of recognising the pyranose polysaccharides heparin and sialyl Lewis X (SLe^X) have been developed.[236] A lectin specific for SLe^X, which is an O-linked glycan over-expressed in epithelial based cancers, may be of value in diagnostics.[239] However, the bis-borane SLe^X lectin developed by Levonis *et al.* recognises the anomeric carboxylic acid group and glycerol tail rather that the hexose ring itself.[240]

The second approach to creating a synthetic lectin is to synthesise a biomimetic compound. This has successfully been applied to β-D-glucose, which, with all equatorial hydroxyl groups, can be simplified as two apolar surfaces with polar sides.[6] This may slot into a 'temple' structure of two apolar, aromatic surfaces that form CH-π interactions to the ring protons, separated by polar spacers that form hydrogen bonds to the axial hydroxyl groups.[6] Using this approach, Davis *et al.* have created a synthetic lectin that binds glucose with high affinity (K_D ~18 mM) and specificity over other monosaccharides (~50:1 versus galactose).[241] More recently, Davis *et al.* have updated their design to consist of a single apolar, aromatic surface and polar arches that can hydrogen bond to both axial and equatorial hydroxyl groups.[242] This yielded binding affinities of K_D 320 μM, 555 μM and 960 μM for D-mannosamine, D-galactosamine and D-glucosamine respectively.[242] Another design update to further increase versatility of these biomimetic synthetic lectins is the 'asymmetric temple', whereby one of the apolar surfaces is smaller than the other and linked by apolar spacers on one side and a variable region on the other side.[243] This yielded binding affinities of K_D 0.25 M to D-mannose but 3.8 mM to D-cellobiose, a disaccharide of D-glucose.[243]

Development of biological synthetic lectins has also taken two approaches, starting with either a nucleic acid scaffold or amino acid scaffold.[60] Referred to as nucleic acid antibodies, aptamers are single-stranded oligonucleotides (20-100 nucleotides) that can bind a range of targets including carbohydrates, metal ions, small molecules, peptides/ proteins, viruses, bacteria and cells.[60,244] Aptamers are generated by systemic evolution of ligands by exponential enrichment (SELEX) that involves multiple rounds of PCR amplification and binding selection.[244] Both RNA and DNA aptamers have been developed for carbohydrate binding. For example, Jeong *et al.* have developed an RNA aptamer (of 110 nucleotides) that binds SLe^X with very high affinity (K_D 0.085 nM).[245] This aptamer can inhibit, *in vitro*, the interaction between the cell surface adhesion proteins E- and P-selectin and SLe^X-tagged transmembrane proteins overexpressed on the surface of HL60 (human promyelocytic leukaemia) cells.[245] A DNA SLe^X aptamer had a similar effect, *in vitro*, of inhibiting the cell adhesion interactions of HepG2 (hepatocellular carcinoma) cells.[246]

An alternative approach to creating a biological lectin utilises peptides/ proteins as a scaffold. The simplest and most common method of creating a proteinogenic synthetic lectin is to re-engineer an existing lectin. This involves cycles of mutagenesis (either site-directed mutagenesis, random mutagenesis by error-prone PCR or domain swapping) followed by screening/ selection (typically by display methods or glycan microarrays).[247] For example, Yabe *et al.* used error-prone PCR and ribosome display to re-engineer EW29Ch, an R-type lectin with two carbohydrate binding domains (CBD) that bind galactose.[248] Wild-type EW29Ch does not bind sialic acid but the re-engineered construct exhibited micromolar binding to α2-6-Sialylated N-linked glycans.[248] A similar approach was taken to re-engineer EW29Ch to bind 6-sulpho-galactose, a component of O-linked glycans, with millimolar affinity.[249] Further site-directed mutagenesis revealed that a basic residue (lysine, arginine or histidine) is required at position 20 in the α CBD to alter its specificity.[249]

Engineering a proteinogenic synthetic lectin by site-directed mutagenesis requires an understanding of protein-carbohydrate interactions. Analysis of bound lectin structures in the PDB (by Hudson *et al.*) reveal that aliphatic residues (A, P, V, M, C, L, I) are disfavoured in carbohydrate binding sites, while aromatic residues (W, Y, F, H) are favoured due to the formation of CH-π interactions between the aromatic side chain and sugar C-H.[204,250] These interactions are influenced by the electronics of the sugar C-H bond and the aromatic side chain and therefore vary with the monosaccharide (e.g. Glc versus Gal), its stereochemistry (e.g. α versus β) and its position relative to the protein side chains.[250]

Our approach

Ideally CI-MPR MRH D9 would be used as an M6P trap and synthetic lectin scaffold due to its high affinity binding, specificity for M6P mono-esters and independence from neighbouring domains.[82] However, as previous work by the Crump lab has found, D9 is insoluble in isolation. Therefore, we are utilising D11 of the CI-MPR as a scaffold due to its high stability and robust *E. coli* expression and refolding protocols previously established within the Crump lab.[112,124]

The MRH domains of the P-type lectins and D11 of the CI-MPR have been well studied and there now exist high-resolution structures for each of these domains (Figure 72).[101,103,124,131,136] Structural characterisation, sequence alignments and mutagenesis studies have identified key carbohydrate binding residues in these MRH domains (Figure 72A-C).[67]

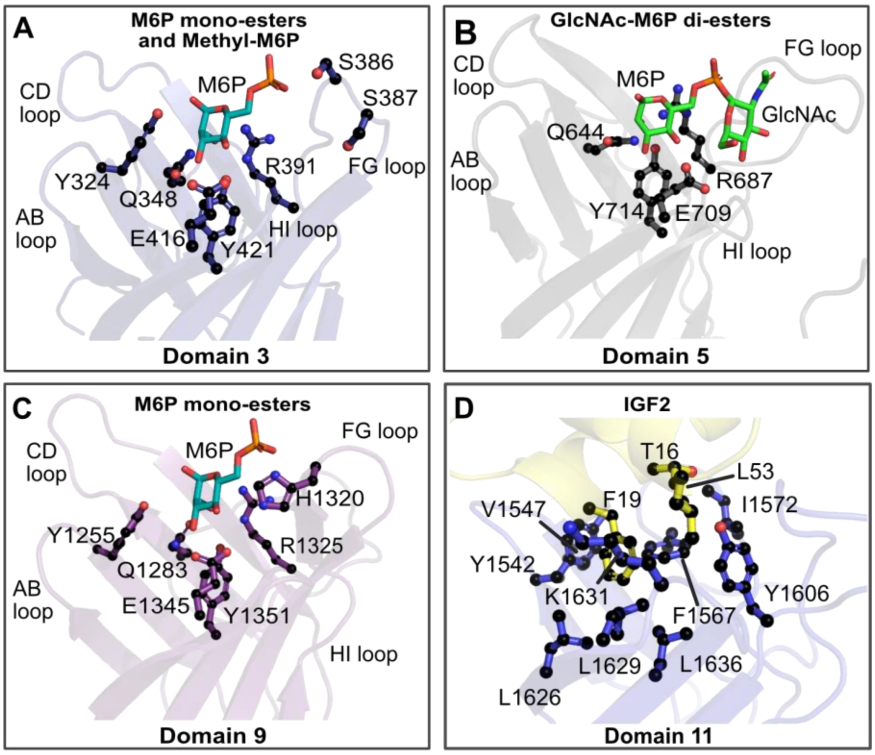


Figure 72: The binding sites of CI-MPR MRH domains D3, D5 and D9 and CI-MPR IGF2 binding domain D11. A: The binding site of bovine D3 (blue, PDB 1SYO) with M6P bound (cyan sticks). The conserved 'QREY' motif (Q348, R391, E416, Y421) and the additional sugar binding residues Y324, S386 and S387 are shown as balls-and-sticks. **B:** The binding site of bovine D5 (grey, PDB 2KVB) with M6P-GlcNAc di-ester modelled into the binding site (green sticks). The conserved 'QREY' motif (Q644, R687, E709, Y714) is shown as balls-and-sticks. **C:** The binding site of human D9 (purple, PDB 63Z0) with M6P docked (cyan sticks). Again, the conserved 'QREY' motif (Q1283, R1325, E1345, Y1351) and H1320 of the FG loop are shown as balls-and-sticks. **D:** The binding site of human D11 (blue) with the peptide IGF2 bound (yellow, PDB 2L29). Interacting residues (T16, F19 and L53 of IGF2 and Y1542, V1547, F1567, I1572, Y1606, L1626, L1629, K1631, L1636 of D11) are shown as balls-and-sticks.

Thus, we have taken a structure-based approach to engineer CI-MPR D11 from a hydrophobic, protein binding domain to a positively charged, carbohydrate binding domain (i.e. a synthetic lectin). Initial attempts grafted the binding site loops of D9 straight onto D11. However, this mutant failed to fold (consistent with the observation from generating the IGF2 trap that too many mutations disrupts D11 folding). Therefore, an iterative approach was taken (Figure 73), with point mutations generated by site-directed mutagenesis, mutants expressed and re-folded *in vitro*, before screening for M6P binding by 2D ^{1}H-^{15}N HSQC NMR.

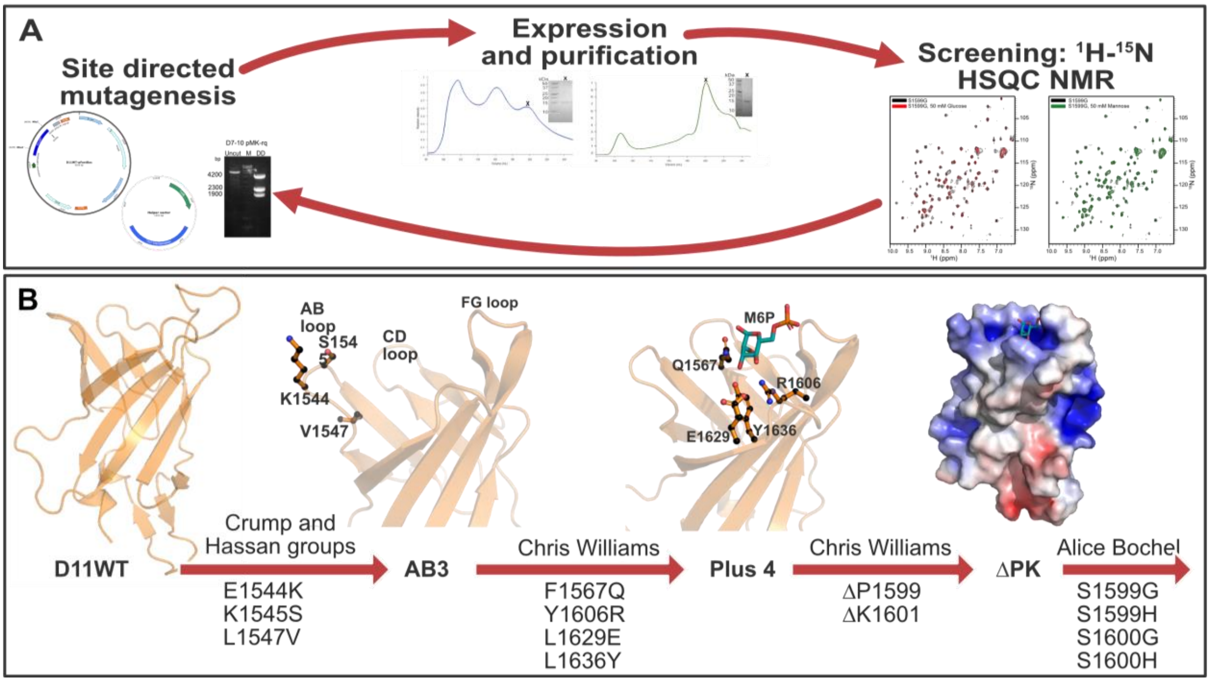


Figure 73: Engineering a biological synthetic lectin. A: Overview of our iterative, structure-based approach. Mutants are generated by site directed mutagenesis, expressed in *E. coli*, refolded *in vitro* and purified by SEC. Screening for sugar binding ability is performed by 2D ^{1}H-^{15}N HSQC NMR. Results obtained inform the next cycle. **B:** Evolution of our D11 construct. Three mutations in the AB loop of D11WT gave rise to AB3, a stable IGF2 binding domain. The conserved sugar binding 'QREY' motif was incorporated into this construct, generating the 'Plus 4' construct. This was mutated further by truncation of the FG loop to form D11 ΔPK. The latest round of mutations focuses on two residues of the FG loop S1599 and S1600. Crystal structures are shown for D11WT, AB3, Plus 4 and ΔPK (PDB 1GP0, 2L29 and unpublished (Dr Chris Williams (University of Bristol))). Mutated residues are shown as balls-and-sticks and M6P (cyan sticks) is docked into the structures of Plus 4 and ΔPK. The electrostatic surface is shown for ΔPK (blue positively charged, range +5 to -5 as determined using the APBS software).[196]

The AB loop of D11 has been extensively mutated and a construct termed AB3 containing three mutations (E1544K, K1545S, L1547V) exhibited a 3-4 fold increase in IGF2 binding affinity.[124] AB3 was chosen for further mutation due to an established expression and *in vitro* refolding protocol and excellent stability (AB3 was observed by ^{1}H NMR to have retained native folding for at least a year when stored at 4 ^{o}C). AB3 was mutated by Dr Chris Williams (University of Bristol) to form 'Plus 4', a construct that knocked out the IGF2 binding residues F1567, Y1606, L1629 and L1636 which contribute to the hydrophobic surface of D11 and would otherwise form hydrophobic interactions with F19 and T16 of IGF2 (Figure 72D). The 'Plus 4' construct simultaneously introduces the conserved 'QREY' motif observed in MRH domains (Figure 72A-C). As illustrated in the high-resolution crystal structure of human D9 (1.5 Å resolution, PDB 6Z30), Q1283 on βC forms a hydrogen bond with the 2'OH of mannose, while E1345 and Y1351 on βH and βI respectively can hydrogen bond to the 3'OH and 4'OH (Figure 73A). From the base of the FG loop, the charged guanidinium group of R1325 interacts with the 6’OH of mannose group. Docking M6P into the D9 crystal structure demonstrates the same orientation of the mannose ring and thus same interactions to the ‘QREY’ motif (Figure 74A).

Although only the 2-epimer of M6P (Figure 74B), glucose 6-phosphate (G6P) binds the P-type lectins with ~10,000 fold lower affinity.[54,89] G6P and 2-deoxyG6P bind the CI-MPR with similar affinity to one another.[89] This suggests that movement of the 2’OH from the axial position in mannose (and M6P) to the equatorial position in glucose (and G6P) (Figure 74B) does not result in steric clash but instead disrupts hydrogen bonds between the 2’OH and binding site residues.[89] Specifically, glutamine and arginine of the conserved ‘QREY’ motif (Q1283 and R1325 of D9, Figure 74A) are positioned to hydrogen bond the 2’OH of mannose/ M6P.

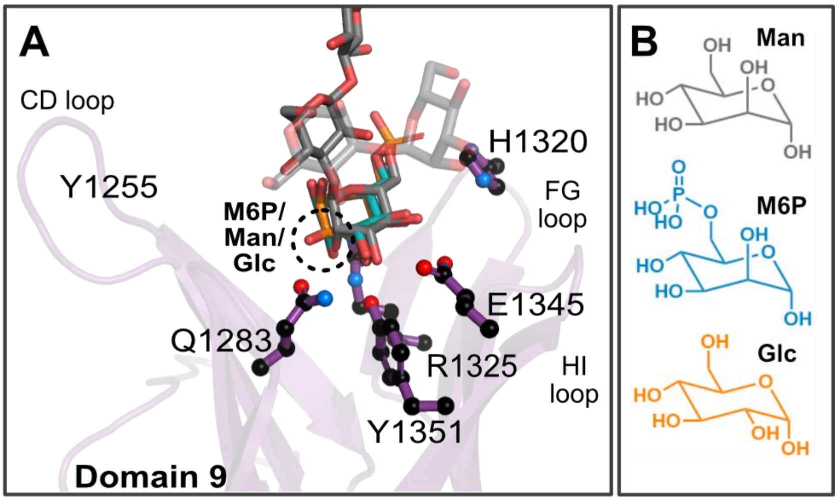


Figure 74: MRH domains are selective for M6P over G6P. A: Residues of the conserved ‘QREY’ (Q1283, R1325, E1345, Y1351) motif in the binding site of D9 are shown as purple balls-and-sticks. The mannosylated N-linked glycan bound by D9 in the crystal structure of D9-10 (PDB 6Z30) is shown as grey sticks. M6P (cyan sticks) and glucose (Glc, orange sticks) are docked into the crystal structure. Mannose and glucose are epimers, with only the 2’OH (circled) differing. **B:** The structures of α-D-mannose (grey), M6P (blue) and glucose (orange).

The ‘Plus 4’ construct was later mutated further by Dr Chris Williams (University of Bristol) to form ‘ΔPK’. P1599 and K1601 both in the FG loop were knocked out to shorten this loop, thereby making it similar to D3 and D9. This chapter summarises the progress made to date and describes the latest round of mutations made to the FG loop.

5.2. AB3

Before designing further mutants, the D11 AB3 construct, which contains the mutations E1544K, K1545S and L1547V in the AB loop and exhibits high affinity IGF2 binding,[124] was studied as a control protein to confirm that the D11 scaffold is incapable of M6P binding.

The AB3 gene was expressed in *E. coli* BL21 (DE3, Novagen). Due to the reducing environment of the bacterial cytosol and the presence of eight cysteines that form four disulphide bonds, AB3 forms insoluble inclusion bodies – dense aggregates of misfolded recombinant protein.[173] However, a robust *in vitro* refolding protocol for D11 expression in *E. coli* was developed by Brown *et al.* in 2002 (based upon Gao *et al.* 1998).[112,172]

The AB3 inclusion bodies were purified and resolubilised in 8 M urea. Denatured protein was refolded by rapid dilution as described in chapter 2.[112] After 24 hrs recombinant protein was purified by SEC and analysed by SDS-PAGE (Figure 75). Folding was assessed by gel filtration and 1D ^{1}H-NMR (Figure 75A, 75C).

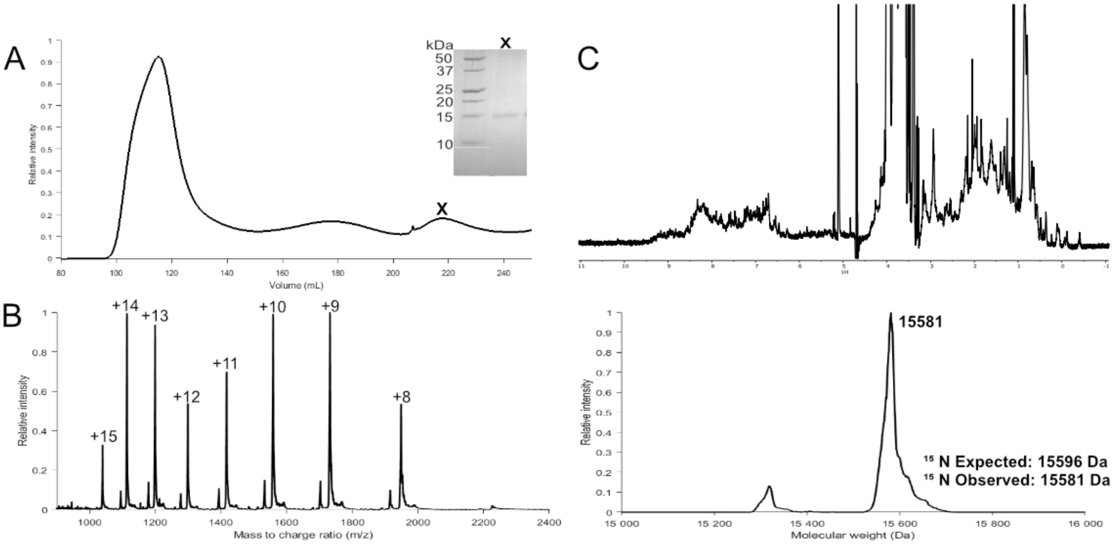


Figure 75: Purification and initial characterisation of D11 AB3. A: SEC trace of AB3 *in vitro* refold. The peak at 160-180 mL marked X corresponded to natively folded AB3. This was analysed by SDS-PAGE (inset). **B:** ESI-MS of AB3 confirmed purification of AB3 and 94 % incorporation of ^{15}N with an observed molecular mass of 15,581 Da versus an expected molecular mass of 15,596 Da assuming 100 % ^{15}N incorporation and reduction of cysteine residues. **C:** 1D ^{1}H-NMR spectrum of AB3 confirmed native protein folding.

When purified by SEC, natively folded, monomeric AB3 eluted between 160-180 mL (Figure 75A). Expression and purification of AB3 was confirmed by SDS-PAGE and mass spectrometry. As described previously, 1D ^{1}H-NMR confirmed native folding of the protein (Figure 75C). Protein folding was also confirmed by the presence of well-dispersed chemical shifts in the 2D ^{1}H-^{15}N HSQC (heteronuclear single quantum coherence) NMR spectra (Figure

76). The spectrum was assigned in CCPN analysis [251] by comparison to a previously assigned spectrum of AB3 at pH 4.0 (supplied by Dr Chris Williams (University of Bristol)).

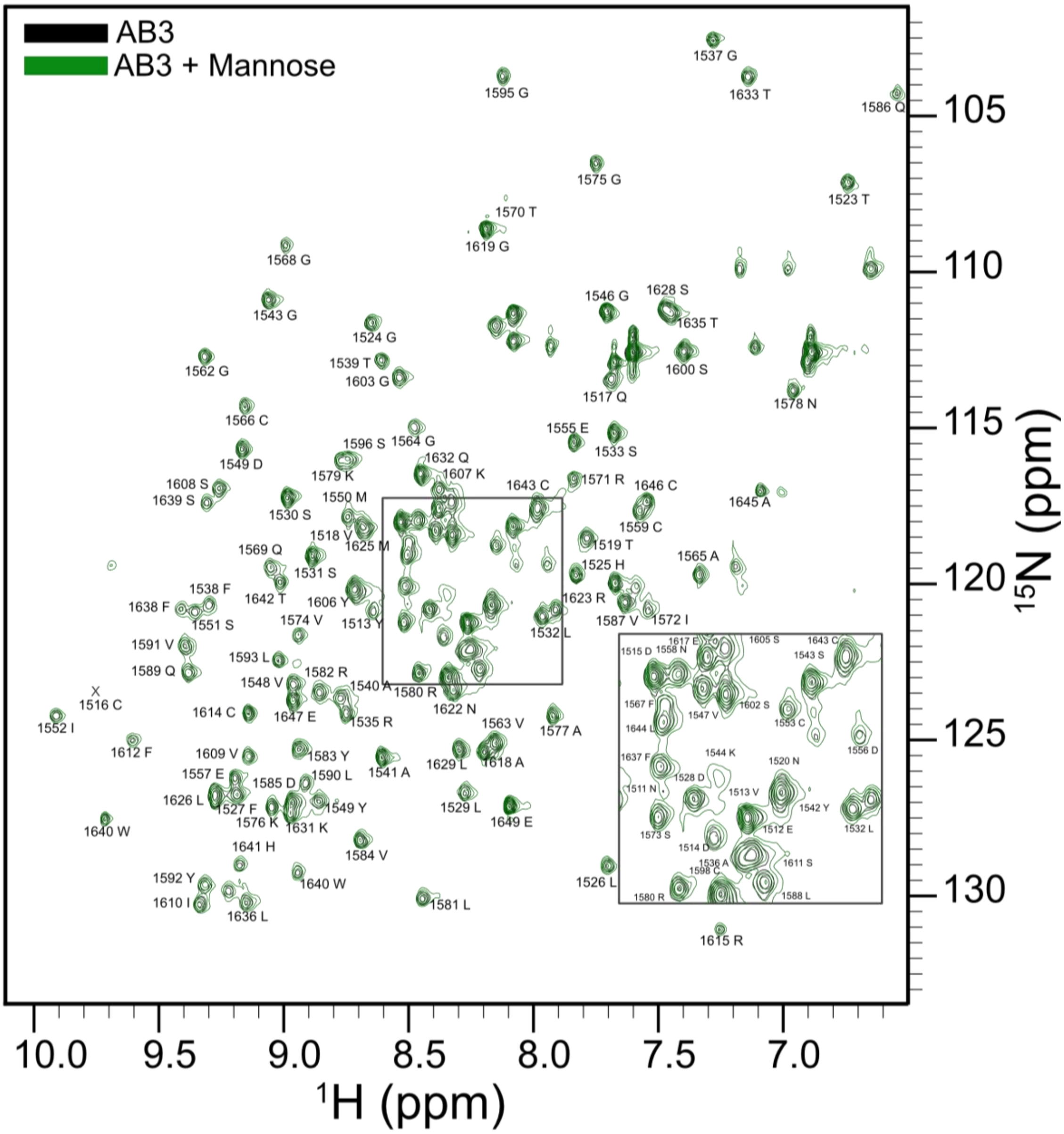


Figure 76: 2D ^{1}H-^{15}N HSQC NMR spectra of AB3 alone (black) and with ~100-fold excess of mannose (green). Spectra were collected at 700 MHz by Dr Chris Williams (University of Bristol) with 120 μM AB3 and 10 mM mannose in 25 mM Bis-Tris, 150 mM NaCl pH 6.5 in 10 % D_2O.

A protein-observed NMR method, HSQC experiments are frequently employed during the study of protein-ligand interactions and during fragment based-drug discovery.[252] HSQC experiments have relatively fast acquisition times and the required ^{15}N labelled protein is relatively easy to obtain by *E. coli* culture.[214] ^{1}H-^{15}N HSQC spectra are recorded with increasing concentrations of ligand.[253] If a residue interacts with the ligand its shielding is altered and the position/ intensity of its chemical shift will change.[253,211] These chemical shift perturbations (CSPs) allow identification of binding site residues and calculation of binding affinity. Olson *et al.* determined the NMR structure of bovine D5 (using 3D NMR experiments) before using ^{1}H-^{15}N HSQC spectra to locate the carbohydrate binding site (by mapping CSPs) and to determine binding affinities to M6P and methyl-M6P-GlcNAc (by HSQC titrations).[131]

Figure 76 also shows the HSQC spectrum of AB3 following incubation with approximately 100-fold excess of mannose. As expected, there were minimal changes with small chemical shift perturbations (CSPs) (Figure 77A), which were calculated according to Equation 4 below. The ten greatest CSPs (which are N1520, K1579, D1630, D1514, K1544, I1572, V1518, N1556, G1603 and E1647) have been mapped to the structure of AB3 (PDB 2L29) (Figure 77B). Their distribution throughout AB3 confirms the absence of a mannose binding pocket. Furthermore, the observed CSPs were very small, with only three residues giving a CSP above 0.1 ppm and a maximum CSP of 0.26 ppm (for N1520) (Figure 77A). In comparison, titration of bovine D5 with M6P and methylM6P-GlcNAc by Olson *et al.* resulted in multiple CSPs between 0.5-0.8 ppm and clusters of CSPs that mapped to the binding loops.[131]

$$\Delta\delta = \sqrt{(5 \times \Delta\delta_{H})^{2} + \Delta\delta_{N}{}^{2}}$$

Equation 4: Calculation of chemical shift perturbation (Δδ) whereby $\Delta\delta_{H}$ and $\Delta\delta_{N}$ correspond to perturbations of ^{1}H and ^{15}N chemical shifts.

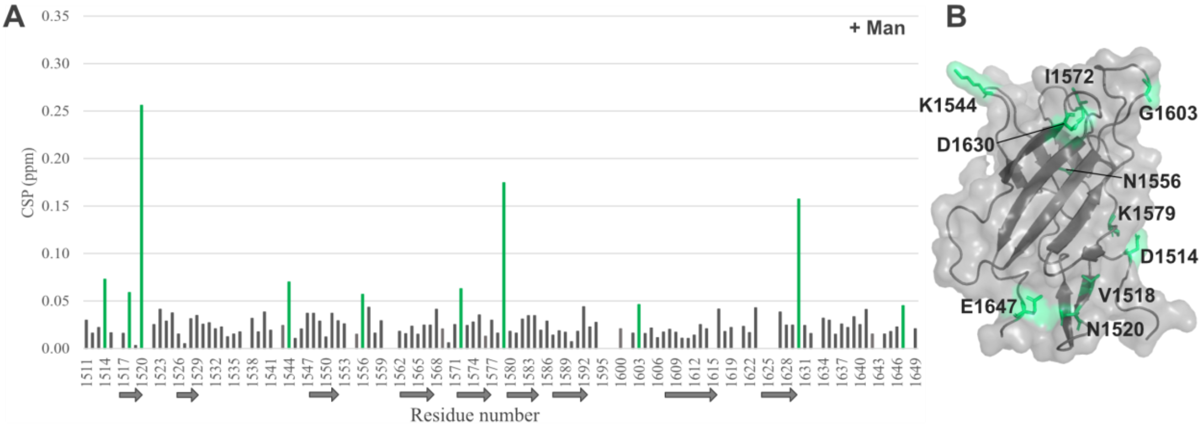


Figure 77: CSP of AB3 following addition of ~100-fold excess of mannose. A: CSPs of AB3 residues following addition of 10 mM mannose. The ten greatest CSPs are coloured green. β-strands are represented by arrows underneath the residue number. **B:** The ten residues with the greatest CSP values are shown as green sticks on the grey surface of AB3 (PDB 2L29).

The 2D ^{1}H-^{15}N HSQC was repeated in the presence of 100-fold excess of M6P (Figure 78A). Comparison to the ^{1}H-^{15}N HSQC of AB3 alone revealed small changes. Furthermore, some of these chemical shifts tracked with increasing concentration of M6P (10 and 20 mM M6P) suggesting an interaction between AB3 and M6P. However, upon CSP calculation and mapping to the structure of AB3 (Figure 78B, 78C), it became evident that there is no specific

M6P binding site on AB3. The ten residues that gave the largest CSPs (E1617, F1567, E1647, N1520, T1519, H1525, H1513, W1640, V1613 and A1618) were spread throughout AB3 and did not form a single M6P binding site (Figure 78C). Only one of these residues also exhibited a significant CSP upon addition of mannose, N1520 at the N-terminus. Interestingly, F1567, which has the second greatest CSP, is a key IGF2 binding residue.

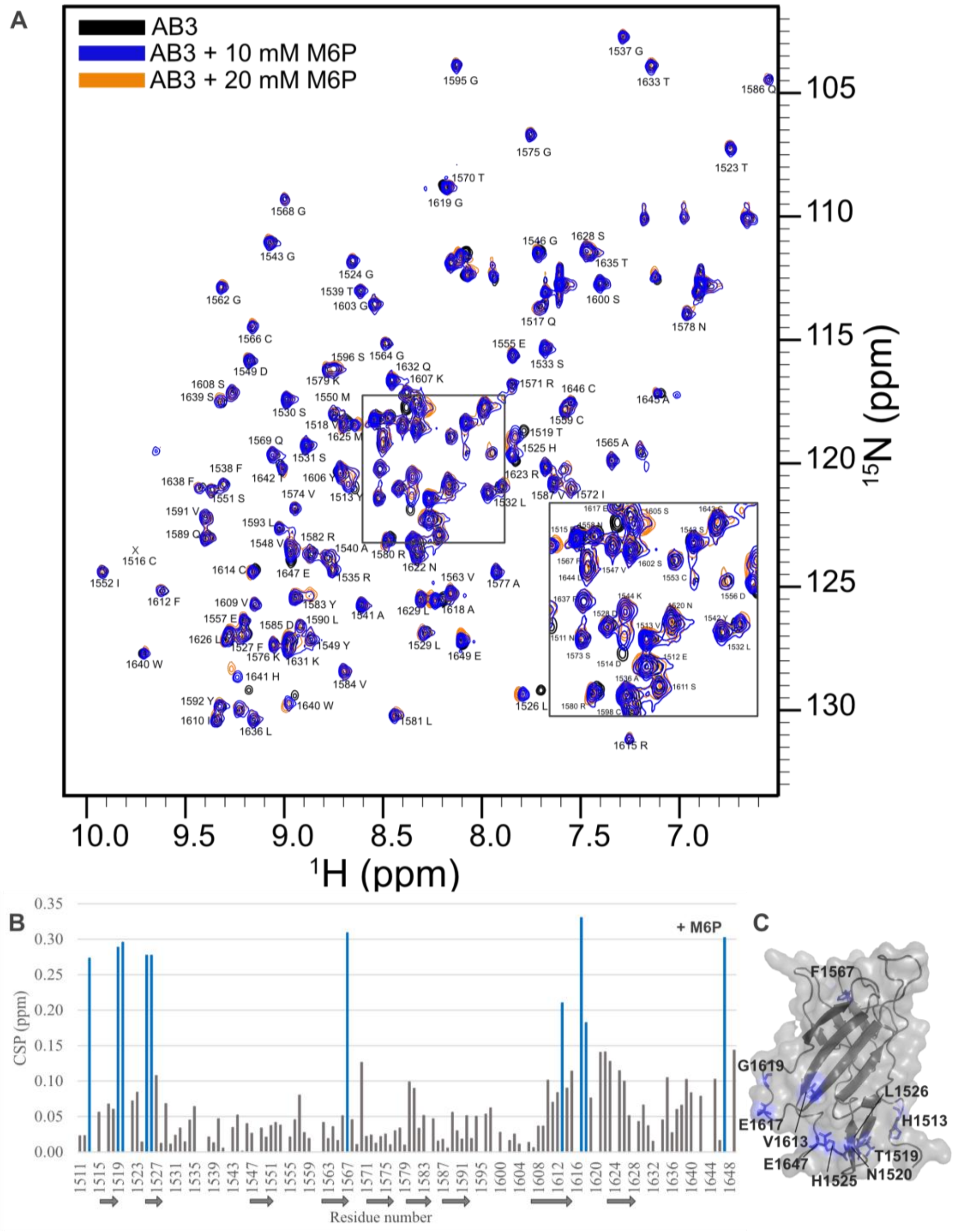


Figure 78: Analysis of AB3 binding M6P. A: 2D ^{1}H-^{15}N HSQC NMR spectra of AB3 alone (black), with 10 mM M6P (blue) and 20 mM M6P (orange). **B:** CSP of AB3 following addition of 10 mM M6P. The ten greatest CSPs are coloured blue. **C:** The ten residues with the greatest CSP values are shown as blue sticks on the grey surface of the crystal structure of AB3 (PDB 2L29). Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 120 µM D11 AB3 in 25 mM Bis-Tris, 150 mM NaCl pH 6.5 in 10 % D_2O.

M6P monosodium salt is acidic. For example, creating a stock solution of 100 mM M6P in 25 mM Tris lowered the pH from 8.0 to 5.0. Subsequent dilution to 2.5 mM M6P with 25 mM Tris lowered the pH from 8.0 to 7.1, a reduction of 0.9 pH units. A pH calibration curve (Appendix Figure 12) was determined by measuring the ^{1}H chemical shift of Tris at a range of pHs. From this the pH of AB3 was calculated to be reduced from pH 6.7 to pH 6.1 upon addition of 10 mM M6P, a decrease of 0.6 pH units (compared to a decrease of 0.4 pH units upon addition of 10 mM mannose, Appendix Table 16).

To determine if the small CSPs observed upon M6P addition were due to sugar-protein interactions or simply a result of the pH change, ^{1}H-^{15}N HSQC spectra of AB3 were collected at a range of pHs (pH 4, 5, 6, 7 and 8, Figure 79A). In each of the spectra the peaks were dispersed demonstrating that AB3 was folded and had not aggregated. CSPs were calculated for spectra collected at pH 7 and 6 (Figure 79B) and the ten residues giving the greatest CSPs (L1526, H1525, E1647, C1516, S1531, M1625, Q1586, C1553, L1581, S1543) mapped to the structure of AB3 (Figure 79C). With values ranging between 0.56-2.6 ppm, these residues demonstrated greater CSPs than observed upon addition of mannose or M6P (likely due to a larger pH change). Three residues, L1526, H1525 and E1647 that gave significant CSPs upon reduction of the pH from pH 7 to 6 also exhibited large CSPs following addition of M6P suggesting these were pH dependent changes and not the result of sugar binding.

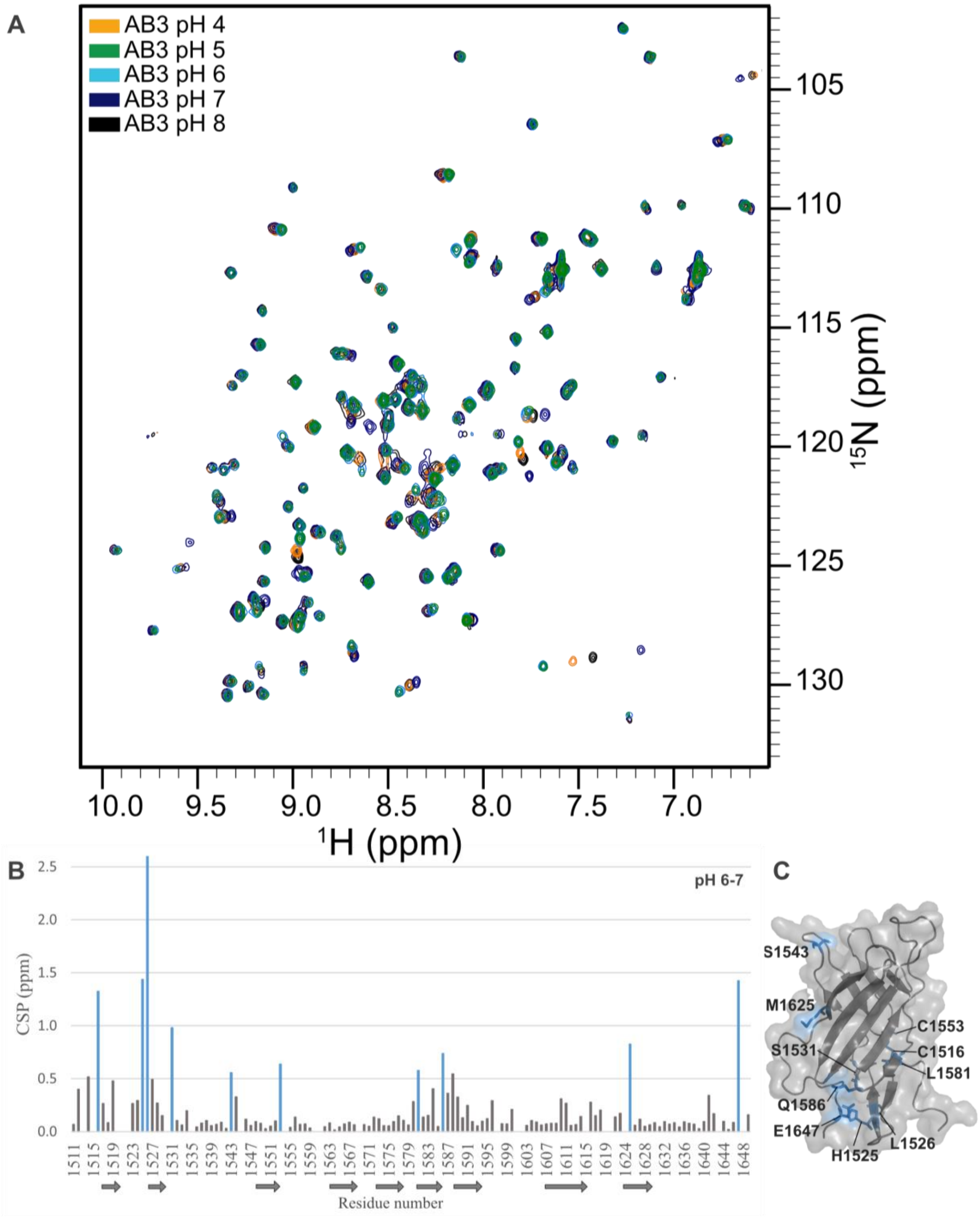
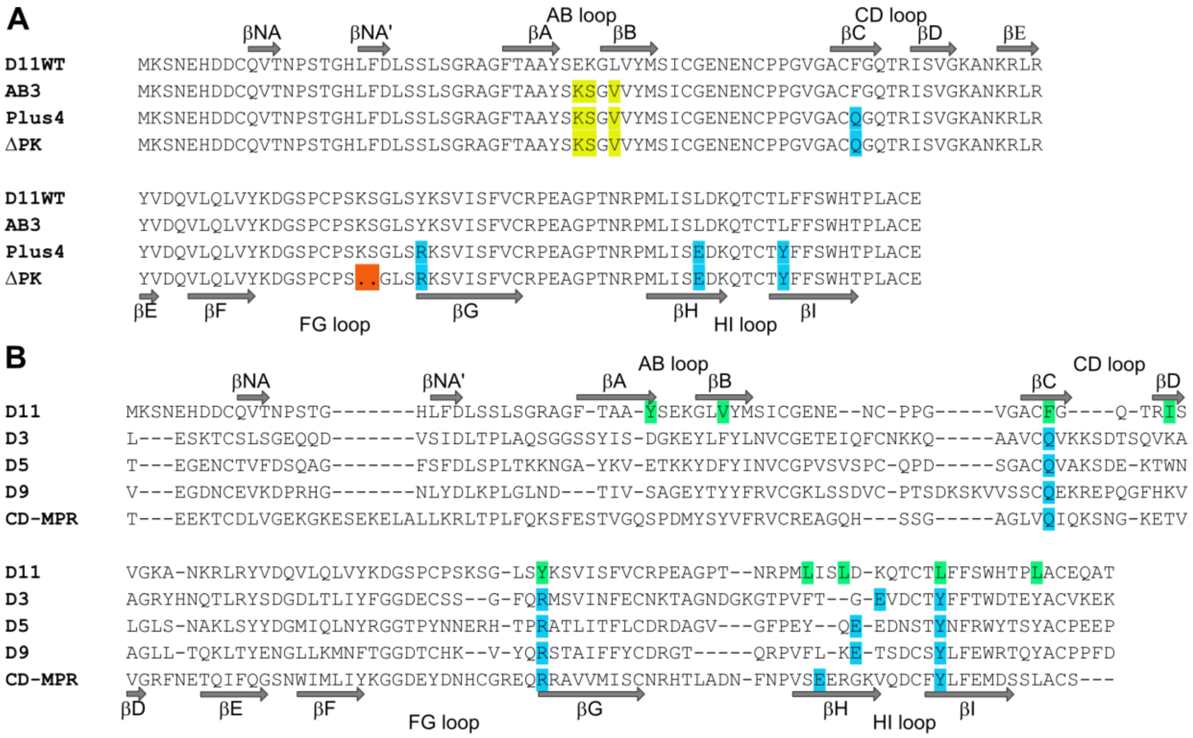


Figure 79: The effect of pH on analysis of AB3. A: 2D ^{1}H-^{15}N HSQC NMR spectra of AB3 at pH 4 (cyan), 5 (light blue), 6 (blue), 7 (navy) and 8 (black). **B:** CSP of AB3 at pH 6 and 7. The ten greatest CSPs are coloured blue. **C:** The ten residues with the greatest CSP values are shown as blue sticks on the grey surface of the crystal structure of AB3 (PDB 2L29). Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 120 μM D11 AB3 in 25 mM Bis-Tris, 150 mM NaCl in 10 % D_2O.

5.3. D11 ΔPK mutant

Having demonstrated that the control protein AB3 is incapable of M6P binding, attention turned to the D11 mutant ΔPK (designed by Dr Chris Williams (University of Bristol)), in which, following incorporation of the 'QREY' motif conserved in MRH domains, two residues of the FG loop (P1599 and K1601) have been deleted (Figure 80A). This was based upon the observation that the carbohydrate binding sites of CI-MPR D3 and D9 have FG loops composed of 11 residues compared to 13 residues in the FG loop of D11WT (and AB3) (Figure 80B). Furthermore, the FG loop of the CD-MPR is 13 residues in length. X-ray crystallography has demonstrated that, in the absence of ligand, the CD-MPR FG loop occludes the binding site.[254]



```
A
D11WT   MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSEKGLVYMSICGENENCPPGVGACFGQTRISVGKANKRLR
AB3     MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACFGQTRISVGKANKRLR
Plus4   MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKRLR
ΔPK     MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKRLR

D11WT   YVDQVLQLVYKDGSPCPSKSGLSYKSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACE
AB3     YVDQVLQLVYKDGSPCPSKSGLSYKSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACE
Plus4   YVDQVLQLVYKDGSPCPSKSGLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACE
ΔPK     YVDQVLQLVYKDGSPCPS..GLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACE

B
D11     MKSNEHDDCQVTNPSTG-------HLFDLSSLSGRAGF-TAA-YSEKGLVYMSICGENE--NC-PPG-----VGACFG----Q-TRIS
D3      L---ESKTCSLSGEQQD-------VSIDLTPLAQSGGSSYIS-DGKEYLFYLNVCGETEIQFCNKKQ-----AAVCQVKKSDTSQVKA
D5      T---EGENCTVFDSQAG-------FSFDLSPLTKKNGA-YKV-ETKKYDFYINVCGPVSVSPC-QPD-----SGACQVAKSDE-KTWN
D9      V---EGDNCEVKDPRHG-------NLYDLKPLGLND---TIV-SAGEYTYYFRVCGKLSSDVC-PTSDKSKVVSSCQEKREPQGFHKV
CD-MPR  T---EEKTCDLVGEKGKESEKELALLKRLTPLFQKSFESTVGQSPDMYSYVFRVCREAGQH---SSG-----AGLVQIQKSNG-KETV

D11     VGKA-NKRLRYVDQVLQLVYKDGSPCPSKSG-LSYKSVISFVCRPEAGPT--NRPMLISLD-KQTCTLFFSWHTPLACEQAT
D3      AGRYHNQTLRYSDGDLTLIYFGGDECSS--G-FQRMSVINFECNKTAGNDGKGTPVFT--G-EVDCTYFFTWDTEYACVKEK
D5      LGLS-NAKLSYYDGMIQLNYRGGTPYNNERH-TPRATLITFLCDRDAGV---GFPEY--QE-EDNSTYNFRWYTSYACPEEP
D9      AGLL-TQKLTYENGLLKMNFTGGDTCHK--V-YQRSTAIFFYCDRGT-----QRPVFL-KE-TSDCSYLFEWRTQYACPPFD
CD-MPR  VGRFNETQIFQGSNWIMLIYKGGDEYDNHCGREQRRAVVMISCNRHTLADN-FNPVSEERGKVQDCFYLFEMDSSLACS---
```

Figure 80: Sequence alignment of D11 mutants and CI-MPR domains. A: Alignment of D11 mutants. **B:** Alignment of human D11 (WT) against the MRH domains of human CI-MPR D3, D5, D9 and bovine CD-MPR. The three mutations in AB3 (E1544K, K1545S, L1547V) are coloured yellow, the conserved 'QREY' carbohydrate binding motif blue, the deletion of P1599 and K1601 in the FG loop orange and IGF2 binding residues of D11 (Y1542, V1547, F1567, I1572, Y1606, L1626, L1629, K1631, L1636) green. The location of β-strands is shown as arrows.

The crystal structure of D11 mutant ΔPK was previously determined by Dr Chris Williams (University of Bristol) to 2.4 Å resolution (Rwork and Rfree values of 24.0 % and 27.1 % respectively) with 96.8 % of the backbone dihedral angles in allowed regions of the Ramachandran plot (Appendix Table 17).

Superimposition of the crystal structure of ΔPK with the structures of bovine D3 and human D9 gives RMSD values of 2.4 Å and 1.0 Å respectively over backbone atoms (Figure 81A). The crystal structure of ΔPK confirms incorporation of the conserved 'QREY' motif. Furthermore these residues (Q1567, R1604, E1627, Y1634 of ΔPK) are in the same position on βC, βG, βH and βI respectively as in D3 and D9 and should, therefore, be capable of forming H-bonds to M6P (Figure 81A). Although Q1567 at the top of βC and R1604 near the top of βG are in slightly different conformations to their counterparts in D3 and D9 (Q348, R391 and Q1283, R1325 respectively) they are still within H-bonding distance to a docked M6P residue (Figure 81A). The ΔPK construct also contains a deletion of P1599 and K1601, truncating the FG loop from 13 residues found in D11 to 11 residues as found in D3 and D9 (Figure 81B).

The IGF2 binding site on D11WT consists of a hydrophobic pocket, surrounded by a ring of positive charge (Figure 81C).[101] The carbohydrate binding sites of D3 and D9 are less hydrophobic and have a positively charged pocket (Figure 81C). A similar surface is observed in the crystal structure of ΔPK (Figure 81C).

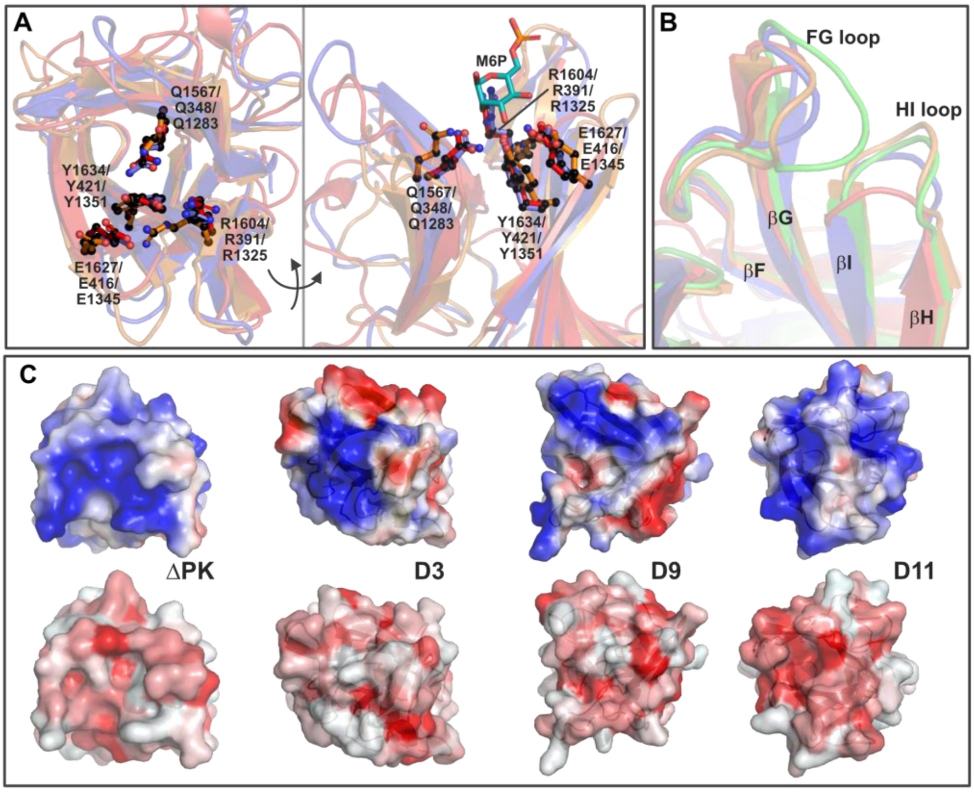


Figure 81: The crystal structure of ΔPK. A: The crystal structure of ΔPK (orange, Dr Chris Williams (University of Bristol)) superimposed with the structures of D3 (red) and D9 (blue) (PDB 1SYO and 6Z30 respectively). Residues of the conserved 'QREY' motif (Q1567, R1604, E1627, Y1634 in ΔPK, Q348, R391, E416, Y421 in D3 and Q1283, R1325, E1345, Y1351 in D9) are shown as balls-and-sticks. The orientation on the right demonstrates the positioning of these side chains relative to a docked M6P residue (cyan sticks). **B:** The FG loop of ΔPK was truncated to be the same length as that of D3 and D9, compared to that of D11WT (green, PDB 1GP0). **C:** The binding site of ΔPK (from the top) compared to that of D3, D9 and D11. Electrostatic surfaces (top, blue positively charged, range +2 to -2 as determined using the APBS software)[196] and hydrophobic surfaces (bottom) are shown.

However, despite co-crystallisation with M6P and soaking, an M6P-bound structure of D11 ∆PK could not be obtained. D11 ∆PK was, therefore, expressed as insoluble inclusion bodies in *E. coli* BL21 (DE3, Novagen) and refolded *in vitro* for characterisation by HSQC NMR. Successful expression, purification and refolding was confirmed by SEC, SDS-PAGE, MS and NMR (Figure 82).

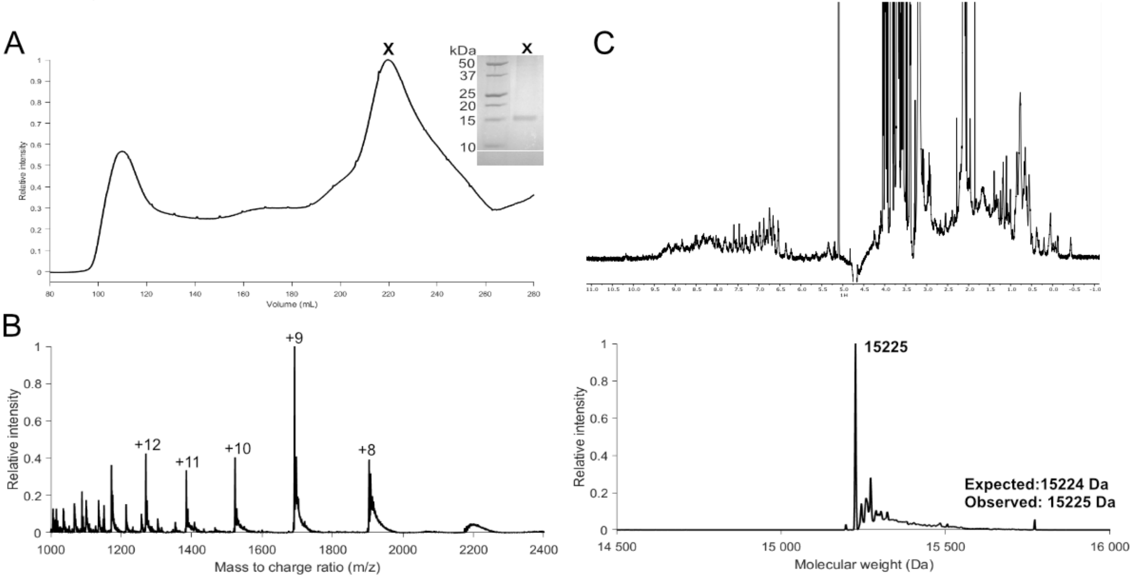


Figure 82: Purification and initial characterisation of D11 ∆PK mutant. A: SEC trace of ∆PK i*n vitro* refold. The peak at 160-180 mL marked X corresponded to natively folded ∆PK. This was analysed by SDS-PAGE (inset). **B:** ESI-MS confirmed purification of ∆PK with an observed molecular mass of 15,225 Da versus an expected molecular mass of 15,224 Da assuming reduction of cysteine residues. **C:** 1D 1H-NMR spectrum of ∆PK confirmed ^{15}N incorporation and protein folding.

^{1}H-^{15}N HSQC NMR experiments with ∆PK demonstrated no significant changes upon addition of glucose, with a maximum CSP of 0.37 ppm (F1635) (Figure 83, Appendix Table 14). Similarly, upon addition of M6P, only small changes were observed, with a maximum CSP of 0.27 ppm (H1525) and a distribution of CSPs throughout the protein (Figure 83E). Several residues that shifted upon addition of M6P also shifted upon addition of glucose (for example, H1525, F1635, T1519, R1580). Furthermore, residues H1525 and E1645 of ∆PK that exhibited two of the greatest CSPs upon addition of M6P (0.27 ppm and 0.23 ppm respectively) were previously demonstrated to be pH sensitive (Appendix Table 14).

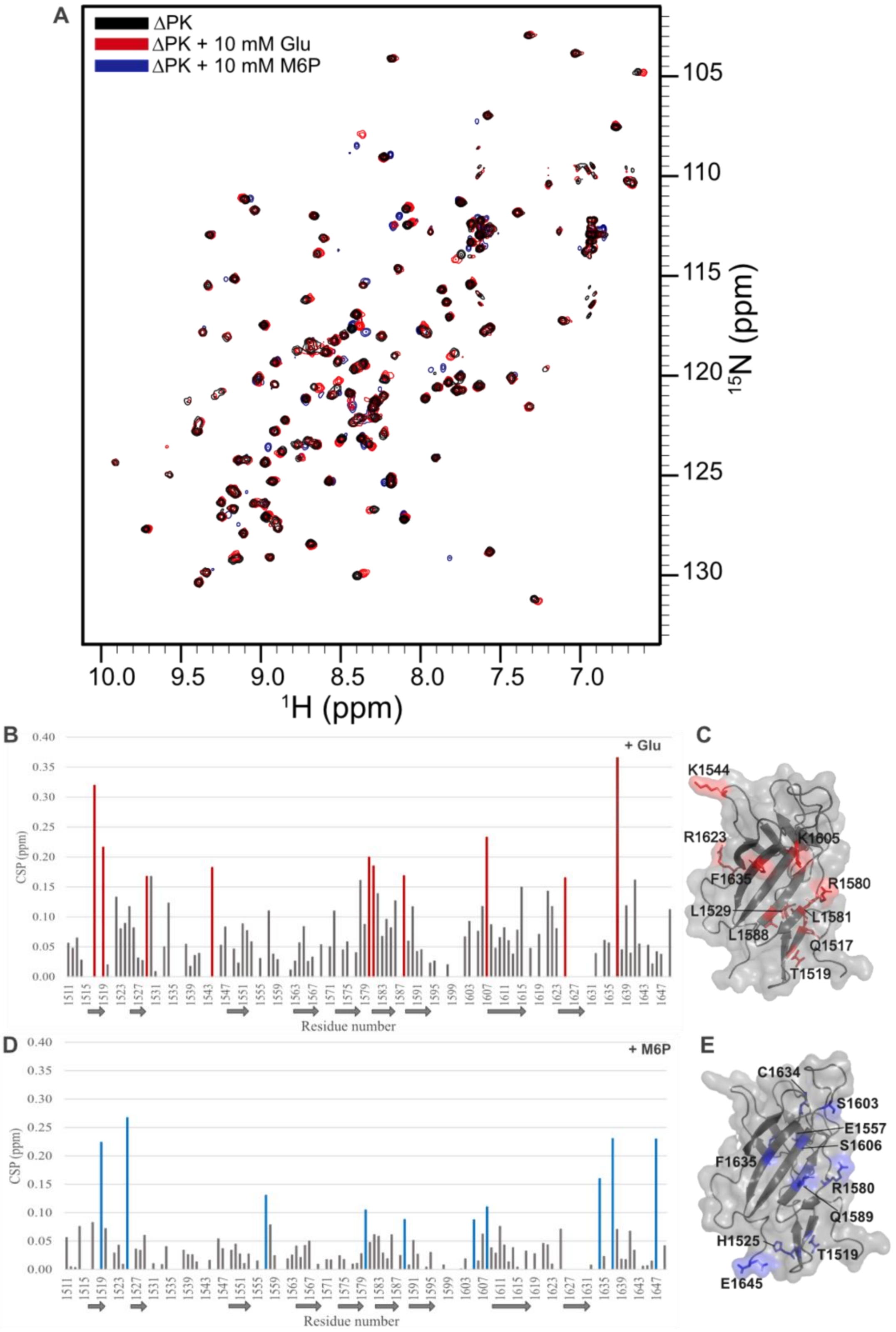


Figure 83: Analysis of D11 mutant ΔPK. A: 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK alone (black) and with 10 mM glucose (red) and with 10 mM M6P (blue). **B:** CSP of D11 ΔPK following addition of 10 mM glucose. The ten greatest CSPs are coloured red. **C:** The ten residues with the greatest CSP values are shown as red sticks on the grey surface of the crystal structure of ΔPK (Dr Chris Williams (University of Bristol)). **D:** CSP of D11 ΔPK following addition of 10 mM M6P. The ten greatest CSPs are coloured blue. **E:** The ten residues with the greatest CSP values are shown as blue sticks on the grey surface of the crystal structure of ΔPK. Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 50 µM ΔPK in 25 mM Bis-Tris, 150 mM NaCl pH 8.5 in 10 % D_2O.

5.4. FG loop mutants

A further round of site-directed mutagenesis was performed on D11 mutant ΔPK based upon the crystal structure of ΔPK (Figure 81, Figure 84). Truncation of the FG loop in the ΔPK construct has re-positioned two serine residues (S1599 and S1600) in the FG loop (Figure 84A). These serine residues are found in the same position in the FG loop of D3 (S386 and S387, Figure 84B) and have previously been demonstrated by mutagenesis and affinity chromatography to be essential sugar binding residues (mutation of S386 reduced D3 ligand binding, while mutation of S387 completely eliminated D3 ligand binding).[67] However, the surface representation of ΔPK with M6P docked demonstrates that the FG loop currently occludes this part of the binding site (Figure 84C). Specifically, S1599 on the FG loop clashes with the phosphate group of M6P. Therefore, ΔPK mutants were designed whereby S1599 and S1600 were mutated to the smallest residue glycine (S1599G also termed ΔPK1G, and S1600G or ΔPK2G). Although partially responsible for restricting this side of the binding pocket, the disulphide bridge formed between C385-C419 of ΔPK was left unmutated as it is conserved in D3 and D9 and likely critical to native protein folding (Figure 84C).

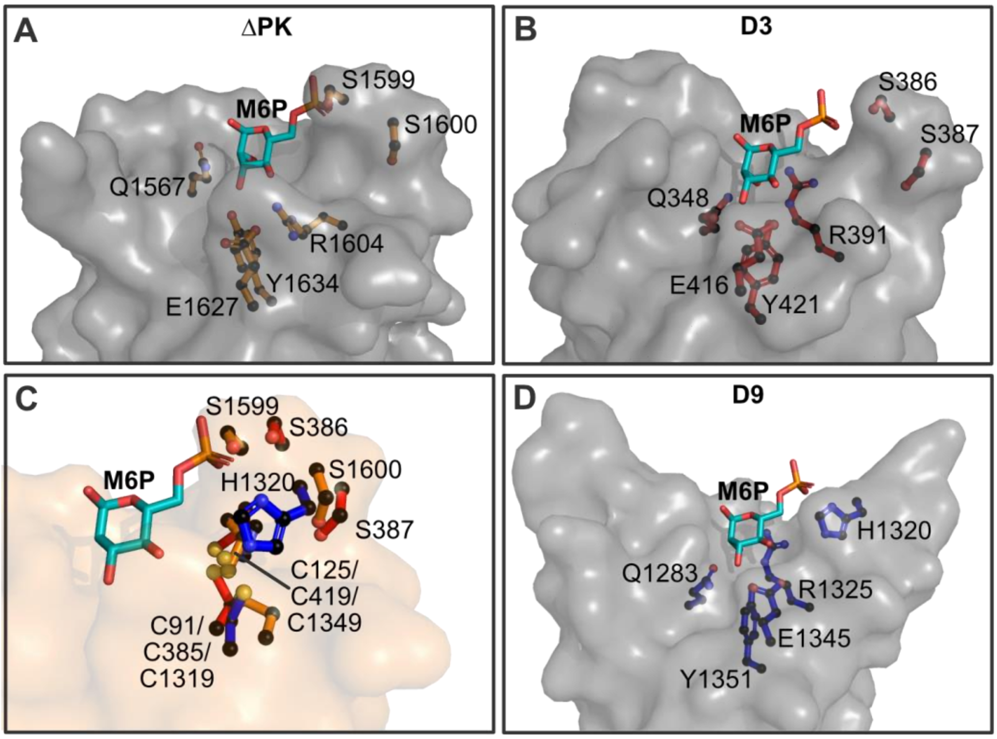


Figure 84: Comparison of the carbohydrate binding pocket of D3, D9 and ΔPK. A: The structure of ΔPK binding surface (grey) with M6P docked (cyan sticks). The conserved 'QREY' motif (Q1567, R1604, E1627, Y1634), S1599 and S1600 in the FG loop are shown as balls-and-sticks. **B:** The structure of bovine D3 binding surface (grey) with M6P bound (cyan sticks, PDB 1SYO). The conserved 'QREY' motif (Q348, R391, E416, Y421), S386 and S387 in the FG loop are shown as balls-and-sticks. **C:** Superimposition of ΔPK, D3 and D9 reveals a conserved disulphide bond between C91/C385/C1319 in the FG loop (of ΔPK/D3/D9) and C125/C419/ C1349 of βI. S1599 and S1600 of ΔPK are replaced by S386 and S387 in D3 and H1320 in D9. **D:** The structure of human D9 binding surface (grey) with M6P docked (cyan sticks, PDB 6Z30). The conserved 'QREY' motif (Q1328, R1325, E1345, Y1351) and H1320 in the FG loop are shown as balls-and-sticks.

Although glycan microarray analysis demonstrated no interaction between mannose and CI-MPR domains, crystal structures have been determined of D1-3 and D9-10 occupied with a mannosylated N-linked glycan (PDB 1Q25 and 6Z30 respectively).[82,103] Comparison to the structure of D1-3 with M6P bound (PDB 1SYO) reveals that the conserved 'QREY' motif interacts with the hydroxyl groups of mannosylated glycan and M6P in the same manner. This 'QREY' motif is observed in a handful of other MRH lectins. For example, ER-resident lectins GIIβ, OS-9 and XTP3-B each contain an MRH domain that possesses mannose binding ability and is incapable of M6P binding.[65,66,134] Therefore, residues outside this 'QREY' motif are responsible for the P-type lectins preferential binding of M6P over mannose.[82] The crystal structures of GIIβ and OS-9 reveal that their FG loops contain either bulky residues (W409 and N410 in GIIβ) that occlude this part of the binding pocket or residues (D182 and L183 in OS-9) that interact with the 6'OH (PDB 4XQM and 3AIH respectively).[65,66] In D3, this is replaced by S386 and S387 which interact with the phosphate oxygen directly and through a bridging water molecule.[136] Meanwhile, D9 contains a histidine residue, H1320, in the FG loop that forms a direct, favourable charge-charge interaction with the phosphate group of M6P (Figure 84D). Therefore, a further two FG loop mutants were designed whereby S1599 and S1600 are replaced by a histidine residue (S1599H also termed ΔPK1H, and S1600H or ΔPK2H).

These four new D11 mutations (S1599G, S1600G, S1599H and S1600H) were incorporated individually into the FG loop of D11 ΔPK by site-directed mutagenesis. As for earlier D11 constructs, the FG loop mutants were expressed as insoluble inclusion bodies in *E. coli* BL21 (DE3, Novagen) and refolded *in vitro*. Successful expression, purification and refolding was confirmed by SEC, SDS-PAGE, MS and NMR (Appendix Figure 13-14).

D11 mutants ΔPK1G and ΔPK1H (Figure 85) were crystallised by the hanging drop vapour diffusion method. Screening conditions were based on the crystallisation conditions of D11WT (PDB 1GP0) and ΔPK. D11 ΔPK1G crystals were observed in a solution of 15 % PEG 4K, 0.1 M sodium cacodylate pH 5.2, 0.1 M sodium acetate. The construct crystallised in space group C121 with 9 molecules in the asymmetric unit. The structure of D11 ΔPK1G was determined to 2.8 Å (Rwork and Rfree values of 26.3 % and 30.6 % respectively) with 97.9 % of the backbone dihedral angles in allowed and favoured regions of the Ramachandran plot (Figure 85A, Appendix Table 17) by molecular replacement using the crystal structure of D11WT (PDB 1GP0) as the search model. There was no electron density for M1508-D1514 at the N terminus or E1647 at the C terminus.

In parallel, D11 ΔPK1H crystals were observed in a solution 18 % PEG 4K, 0.1 M sodium cacodylate pH 5, 0.05 M sodium acetate. The construct crystallised in space group P12$_1$1 with 6 molecules in the asymmetric unit. The structure of D11 ΔPK1H was determined to 2.5 Å (Rwork and Rfree values of 25.3 % and 28.0 % respectively) with 98.4 % of the backbone dihedral angles in allowed and favoured regions of the Ramachandran plot (Figure 85B, Appendix Table 17) by molecular replacement using the crystal structure of D11WT (PDB 1GP0) as the search model. There was no electron density for M1508-D1514 at the N terminus or P1646-E1647 at the C terminus.

The crystal structures of D11 ΔPK1G and ΔPK1H (Figure 85) exhibit the core β-sandwich topology observed for D11WT with RMSD values (over backbone atoms) of 2.2 Å, 1.0 Å and 0.56 Å to D3, D9, D11WT respectively for ΔPK1G and 2.3 Å, 0.97 Å and 0.25 Å to D3, D9, D11WT respectively for ΔPK1H. The presence of both mutations (serine and histidine) is clearly visible in the corresponding electron density maps (Figure 85). However, as for ΔPK, despite co-crystallisation and soaking with M6P, an M6P-bound structure could not be obtained.

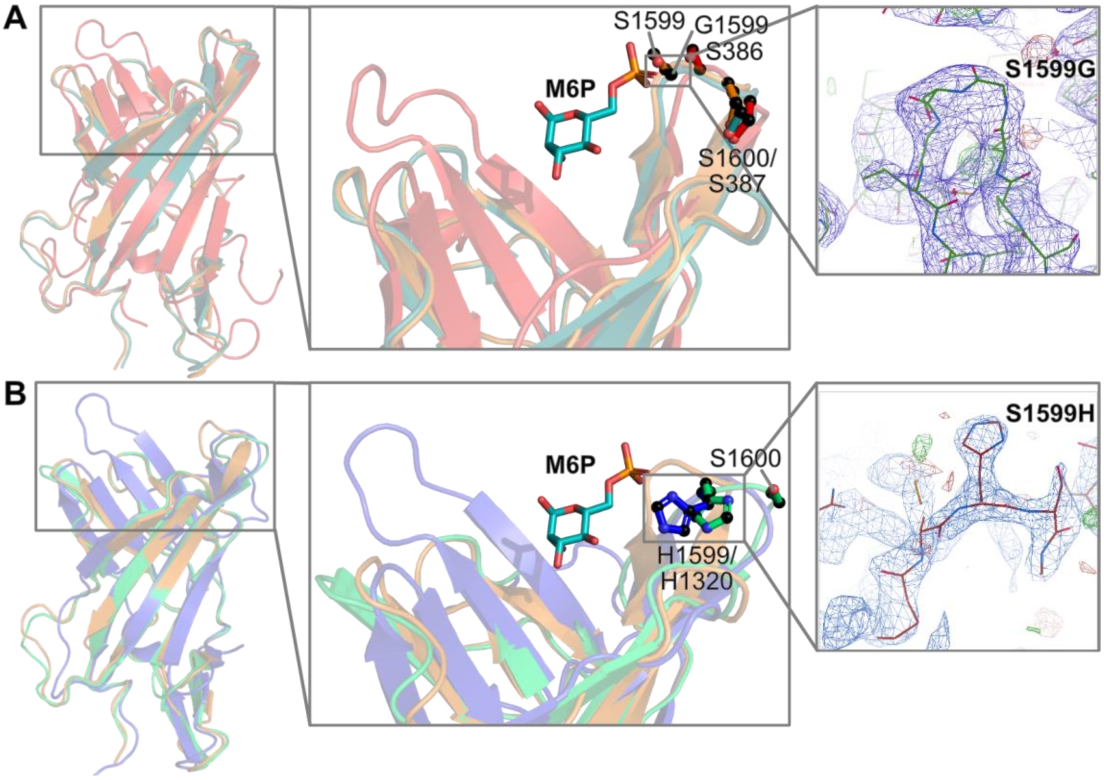


Figure 85: Crystal structures of D11 ΔPK1G and ΔPK1H. A: Superimposition of D11 ΔPK1G (teal), ΔPK (orange, Dr Chris Williams (University of Bristol)) and D3 (red, PDB 1SYO). Zoom in on the binding site shows carbohydrate binding residues of the FG loop: S386 and S387 (red) in D3, S1599 and S1600 (orange) in ΔPK and G1599 and S1600 in the ΔPK1G. M6P (cyan sticks) docked. Inset shows the electron density for G1599 in ΔPK1G. **B:** Superimposition of D11 ΔPK1H (green), ΔPK (orange, Dr Chris Williams (University of Bristol)) and D9 (blue, PDB 6Z30). Zoom in on the binding site shows the carbohydrate binding residue H1320 (blue ball-and-stick) in the FG loop of D9 and H1599 (green ball-and-stick) in the ΔPK1H. M6P (cyan sticks) docked. Inset shows the electron density for H1599 in ΔPK1H. For each structure the 2Fo-Fc map is shown at 1.8σ.

As for AB3 and ΔPK, ^{1}H-^{15}N HSQC was used to assess sugar binding capability of the D11 FG loop mutants. Due to changes in the protein sequence and pH variation, ^{1}H-^{15}N HSQC of D11 ΔPK1G were only 76-82 % assigned (Appendix Table 11). Very few changes were observed upon addition of excess mannose or glucose to D11 ΔPK1G (Figure 86). Analysis of the CSPs revealed a maximum CSP of 0.15ppm (for R1613 and V1548) in the presence of mannose and 0.14 ppm (for Q1630 and H1525) in the presence of glucose (Appendix Table 15, Figure 87A-B).

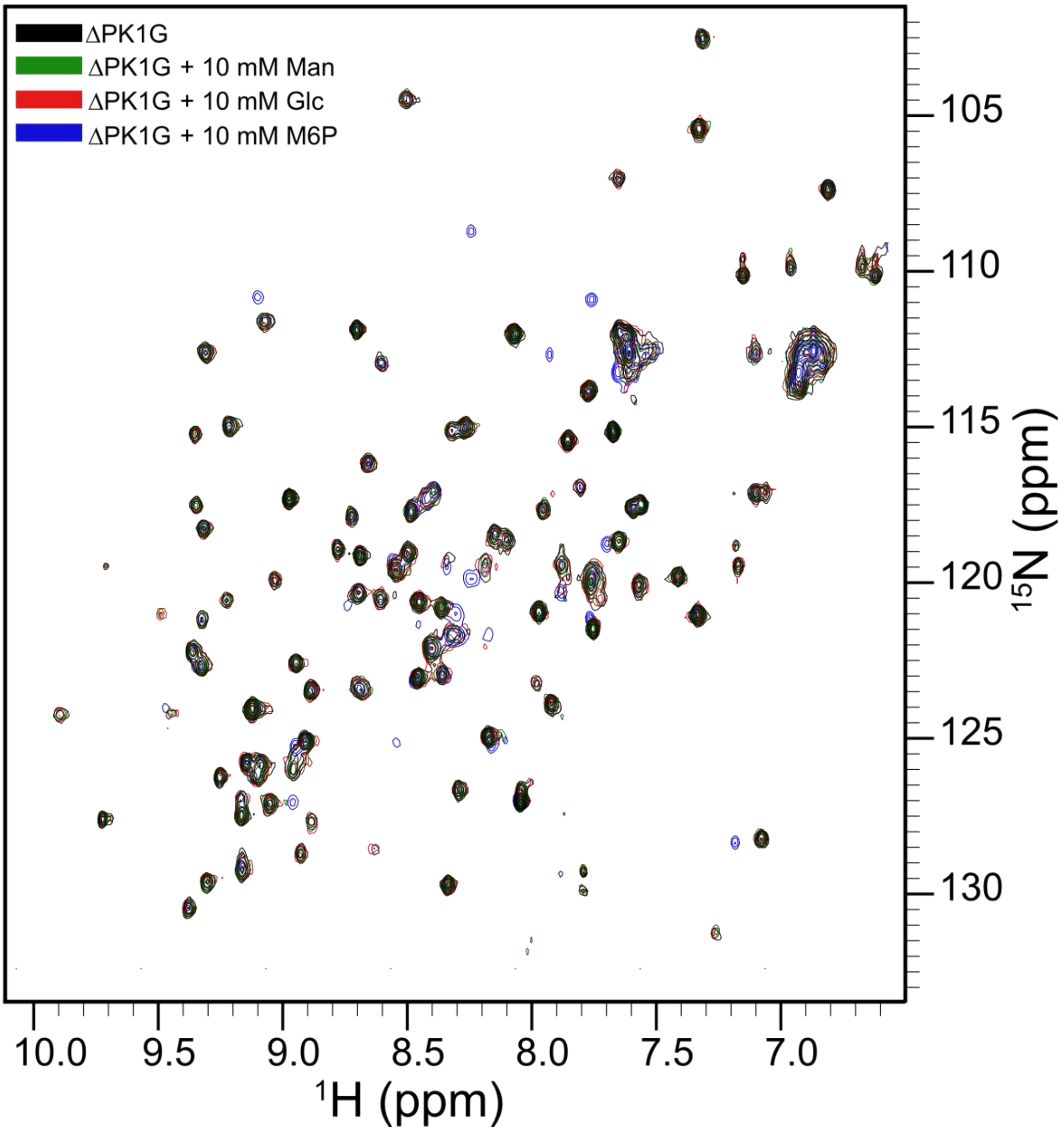


Figure 86: Analysis of D11 mutant ΔPK1G. 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK1G alone (black), with 10 mM mannose (green), with 10 mM glucose (red) and 10 mM M6P (blue). Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 50 μM D11 ΔPK1G and 10 mM sugar in 25 mM Tris pH 8.5, 150 mM NaCl in 10 % D_2O.

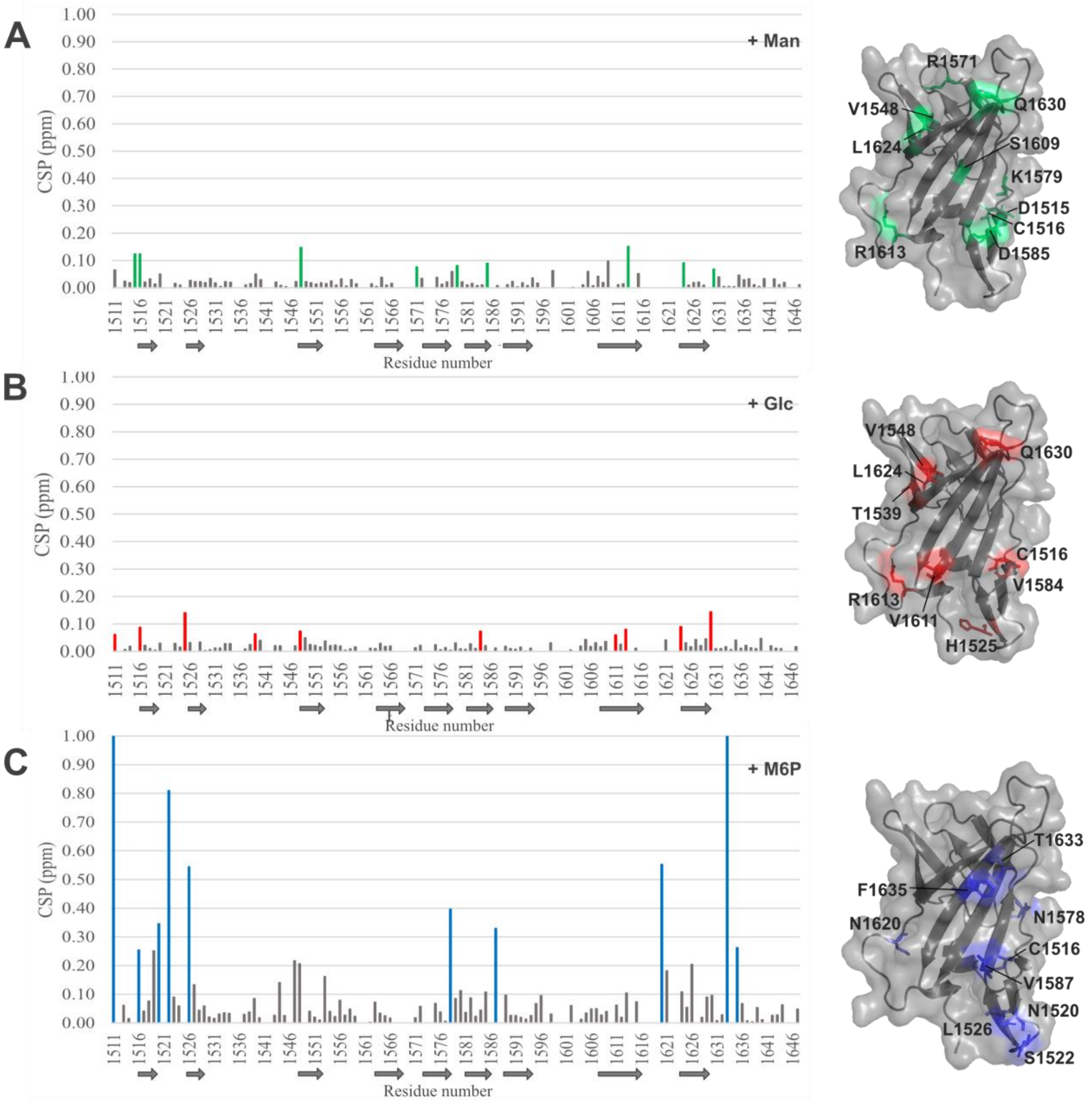


Figure 87: CSP values of D11 ΔPK1G following addition of mannose, glucose and M6P. A: CSP of D11 ΔPK1G following addition of 10 mM mannose. The ten greatest CSPs are coloured green. β-strands are represented by arrows underneath the residue number. The ten residues with the greatest CSP values are shown as green sticks on the grey surface of ΔPK1G crystal structure. **B:** CSP of D11 ΔPK1G following addition of 10 mM glucose. The ten greatest CSPs are coloured red. The ten residues with the greatest CSP values are shown as red sticks on the grey surface of ΔPK1G crystal structure. **C:** CSP of D11 ΔPK1G following addition of 10 mM M6P. The ten greatest CSPs are coloured blue. The ten residues with the greatest CSP values are shown as blue sticks on the grey surface of ΔPK1G crystal structure. N1511 has been omitted as it was not visible in the crystal structure.

However, upon addition of excess M6P, several changes were observed in the ^{1}H-^{15}N HSQC spectrum (Figure 86). Addition of 10 mM M6P gave large CSPs, with the ten greatest CSPs ranging from 0.25-1.06 ppm (Figure 87C, Appendix Table 15). However, mapping these CSPs to the crystal structure of D11 ΔPK1G did not reveal a distinct M6P binding site. Instead there is a cluster of sensitive residues at the N-terminus (C1516, N1520, S1522, L1526) away from the binding loops (Figure 87C). T1633 and F1635 both on βH gave CSP values of 1.03 ppm and 0.26 ppm (Appendix Table 15). However, these two residues protrude from the β-strand into the solvent rather than into the core of the β-sandwich and are thus not in a position to interact with bound ligand (Figure 87C).

The effect of M6P was studied further by titration with a range of M6P concentrations (5-50 mM, i.e. 100-1000-fold excess) (Figure 88). The ^{1}H-^{15}N HSQC revealed that some peaks clearly tracked with an increased concentration of M6P, while other peaks only appeared at high M6P concentrations (Figure 88B).

Addition of 50 mM M6P resulted in even larger CSPs, with the ten greatest values between 0.39-3.32 ppm (Appendix Table 15). Only three of these residues (L1526, V1587 and the N-terminal residue N1511) previously exhibited large CSPs upon addition of 10 mM M6P (Appendix Table 15). Three residues L1526, C1553 and C1516 that gave large CSP values following addition of M6P were demonstrated in earlier experiments with AB3 to be pH sensitive (Appendix Table 14). Of the remaining residues giving large CSPs upon titration with 50 mM M6P three were in the binding loops of ΔPK1G (Figure 89). These were S1545, one of the AB loop residues mutated to form AB3, S1596 in the FG loop and Q1630 in the HI loop (Figure 89).

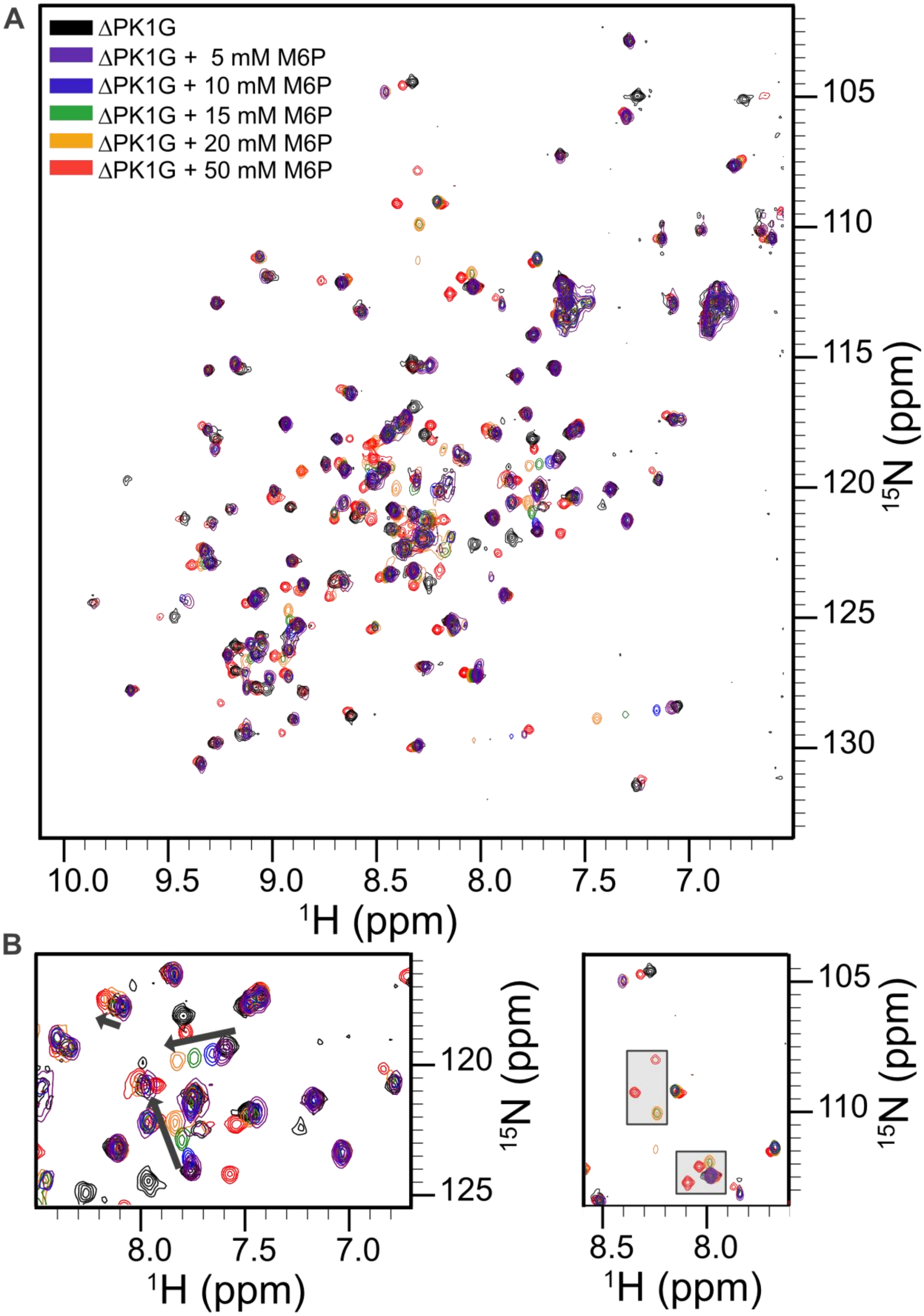


Figure 88: Analysis of D11 mutant ΔPK1G and M6P. A: 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK1G with increasing concentrations of M6P: 0 mM M6P black, 5 mM M6P purple, 10 mM M6P blue, 15 mM M6P green, 20 mM M6P orange, 50 mM M6P red. **B:** Examples of chemical shifts tracking with increasing M6P concentration (left, arrows) and new peaks appearing at high ligand concentrations (right, grey boxes). Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 50 μM D11 ΔPK1G in 25 mM Tris pH 8.5, 150 mM NaCl in 10 % D_2O.

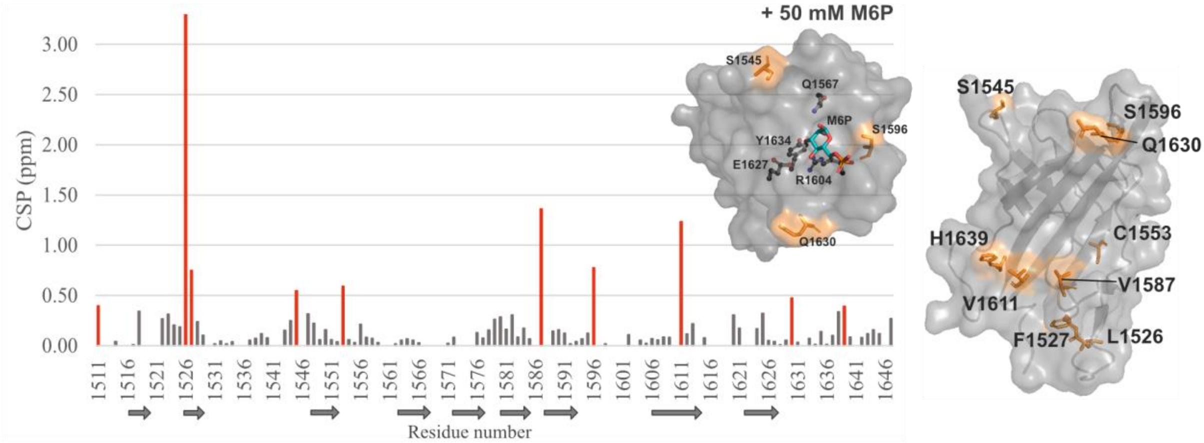


Figure 89: CSP values of D11 ΔPK1G following addition of 50 mM M6P. CSP of D11 ΔPK1G following addition of 50 mM M6P. The ten greatest CSPs are coloured orange. β-strands are represented by arrows underneath the residue number. The ten residues with the greatest CSP values are shown as orange sticks on the grey surface of ΔPK1G crystal structure viewed from the side (right) and from the top (left). The view from the top has M6P docked into the binding site and the conserved 'QREY' motif shown as balls-and-sticks. N1511 has been omitted as it was not visible in the crystal structure.

^{1}H-^{15}N HSQC were also collected of FG loop mutant ΔPK1H in the presence of 50 mM M6P (Figure 90A). CSPs were of the same magnitude to those observed for ΔPK1G (Figure 89), ranging from 0.41-3.67 ppm (Figure 90B, Appendix Table 15). Three residues, L1526, S1545 and N1511, gave high CSPs in both D11 ΔPK1H and ΔPK1G plus 50 mM M6P (Appendix Table 15). Similarly to D11 ΔPK1G, three ΔPK1H residues that gave relatively large CSPs upon addition of M6P, L1526, C1516 and Q1586, were demonstrated to be pH sensitive in AB3 (Appendix Table 14). The remaining D11 ΔPK1H residues giving large CSPs formed a small cluster near the base of D11. The only binding loop residues with significant CSPs were S1545 in the AB loop and T1631 in the HI loop (Figure 90C). However, both of these residues appeared to be positioned too far away to bind M6P.

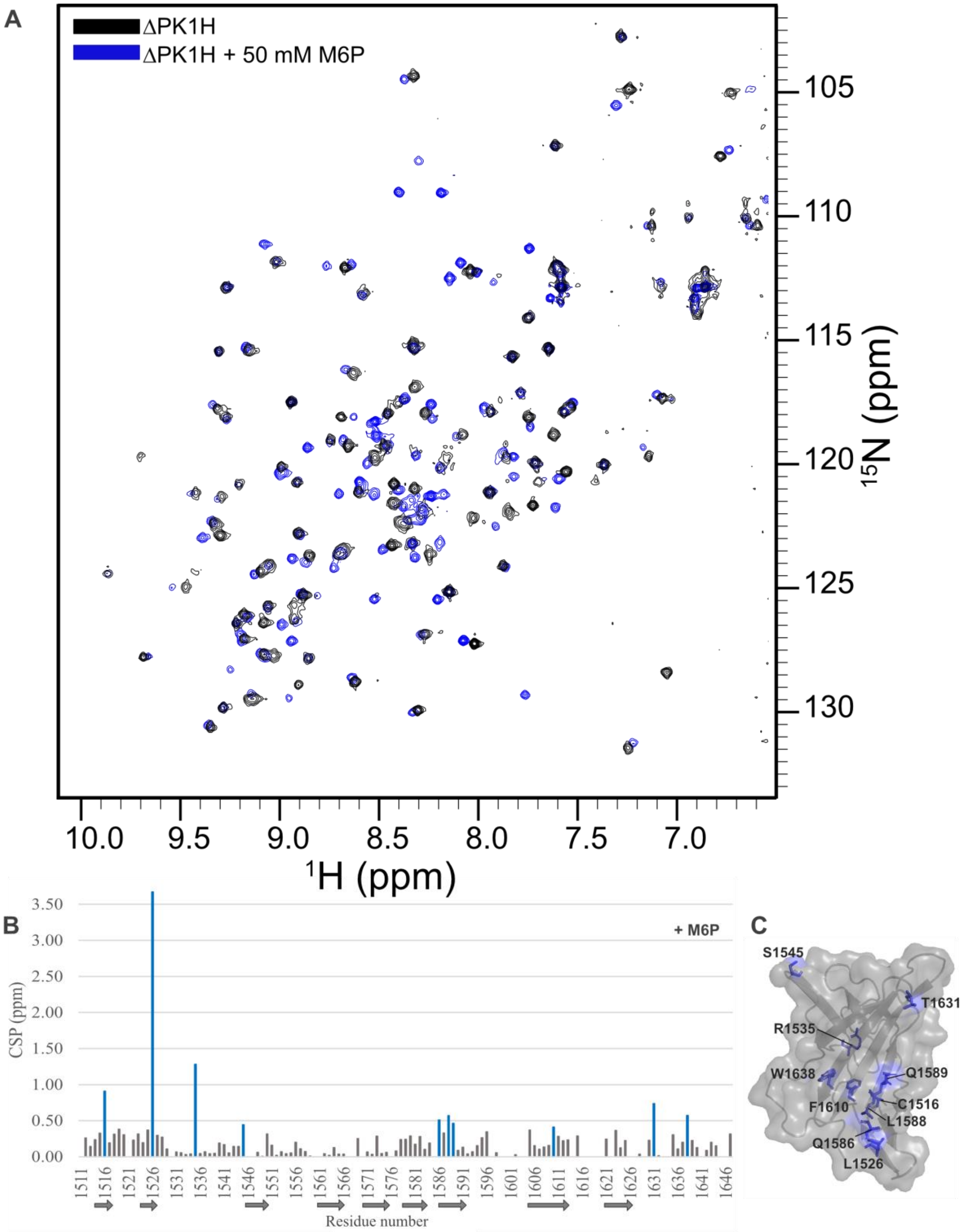


Figure 90: Analysis of D11 mutant ΔPK1H and M6P. A: 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK1H in the absence (black) and presence of 50 mM M6P (blue). **B:** CSPs of D11 ΔPK1H following addition of 50 mM M6P. The ten greatest CSPs are coloured blue. β-strands are represented by arrows underneath the residue number. **C:** The ten residues with the greatest CSP values are shown as blue sticks on the grey surface of ΔPK1H crystal structure. Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 50 µM D11 ΔPK1H in 25 mM Tris pH 8.5, 150 mM NaCl in 10 % D$_2$O.

The above ^{1}H-^{15}N HSQC experiments involving D11 FG loop mutants ΔPK1G and ΔPK1H were performed at pH 8.5 (above the proteins pI 7.5). However, Marron-Terada *et al.* have demonstrated by affinity chromatography that D3 exhibits optimal M6P binding at pH 6.9, while D9 displays optimal binding at pH 6.4-6.5.[132,137] This is in-line with the observation that the CI-MPR binds M6P-tagged glycoproteins at the TGN (pH 6.5) and PM (pH 7.4) and releases its cargo in the late endosome (pH <6.0).[54] Therefore these ^{1}H-^{15}N HSQC experiments were performed at lower pH, pH 6.5.

Addition of 10mM mannose (100-fold excess) to D11 ΔPK1H at pH 6.5 did not result in significant CSPs, with the 10 greatest CSPs distributed throughout the protein and of the same magnitude of those observed for the control protein AB3 in the presence of mannose and M6P (CSPs 0.19-0.44 ppm, Figure 91B, Appendix Table 15). Meanwhile, addition of 10 mM M6P resulted in slightly larger CSPs ranging from 0.17-0.91 ppm (Figure 91D, Appendix Table 15). However, these remained smaller than the CSPs observed for D11 ΔPK1G and ΔPK1H at pH 8.5 (Appendix Table 15).

When mapped to the crystal structure of ΔPK1H, the ten residues with the greatest CSPs did not form a binding pocket but were again distributed throughout the protein (Figure 91E). Only two of these residues were in the binding loops: K1544 in the AB loop and G1568 in the CD loop (Figure 91E). Furthermore, three of these residues (L1526, E1645, H1525) have been previously demonstrated to be pH sensitive (Appendix Table 14). As observed for AB3, analysis of the bis-Tris peak revealed a pH change of 0.6 pH units (pH 6.9 to 6.3) following addition of 10 mM M6P (Appendix Figure 12, Appendix Table 16, there was no pH change following addition of 10 mM mannose).

D11 FG loop mutant ΔPK2H exhibited very similar behaviour at pH 6.5 (Figure 92). A similar CSP range was observed upon addition of 10 mM M6P (0.40-1.19 ppm) with several residues of ΔPK1H also shifting in ΔPK2H (L1526, H1526, E1645, V1611, K1544) (Appendix Table 15). However, no discrete M6P binding site was observed (Figure 92E) and pH sensitive residues (L1526, H1525, E1645, C1553, L1581) again exhibited some of the greatest CSPs (Appendix Table 14).

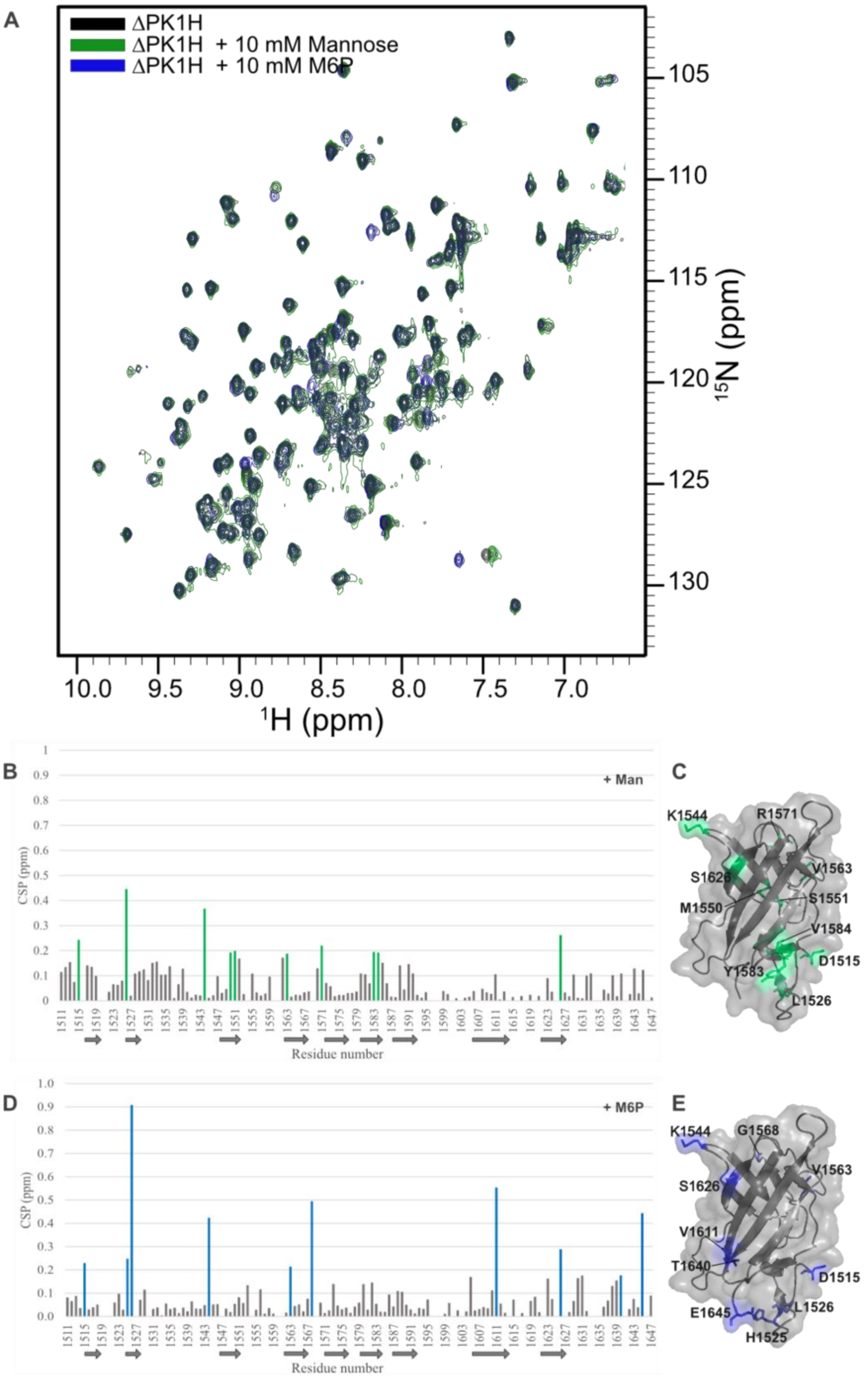


Figure 91: Analysis of D11 mutant ΔPK1H binding M6P at pH 6.5. A: 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK1H alone (black), with 10 mM mannose (green) and 10 mM M6P (blue). **B:** CSP of D11 ΔPK1H following addition of 10 mM mannose. The ten greatest CSPs are coloured green. β-strands are represented by arrows underneath the residue number. **C:** The ten residues with the greatest CSP values are shown as green sticks on the grey surface of ΔPK1H crystal structure. **D**: CSP of D11 ΔPK1H following addition of 10 mM M6P. The ten greatest CSPs are coloured blue. **E:** The ten residues with the greatest CSP values are shown as green sticks on the grey surface of ΔPK1H crystal structure. Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 100 μM D11 ΔPK1H in 25 mM Bis-Tris, 150 mM NaCl pH 6.5 in 10 % D$_2$O.

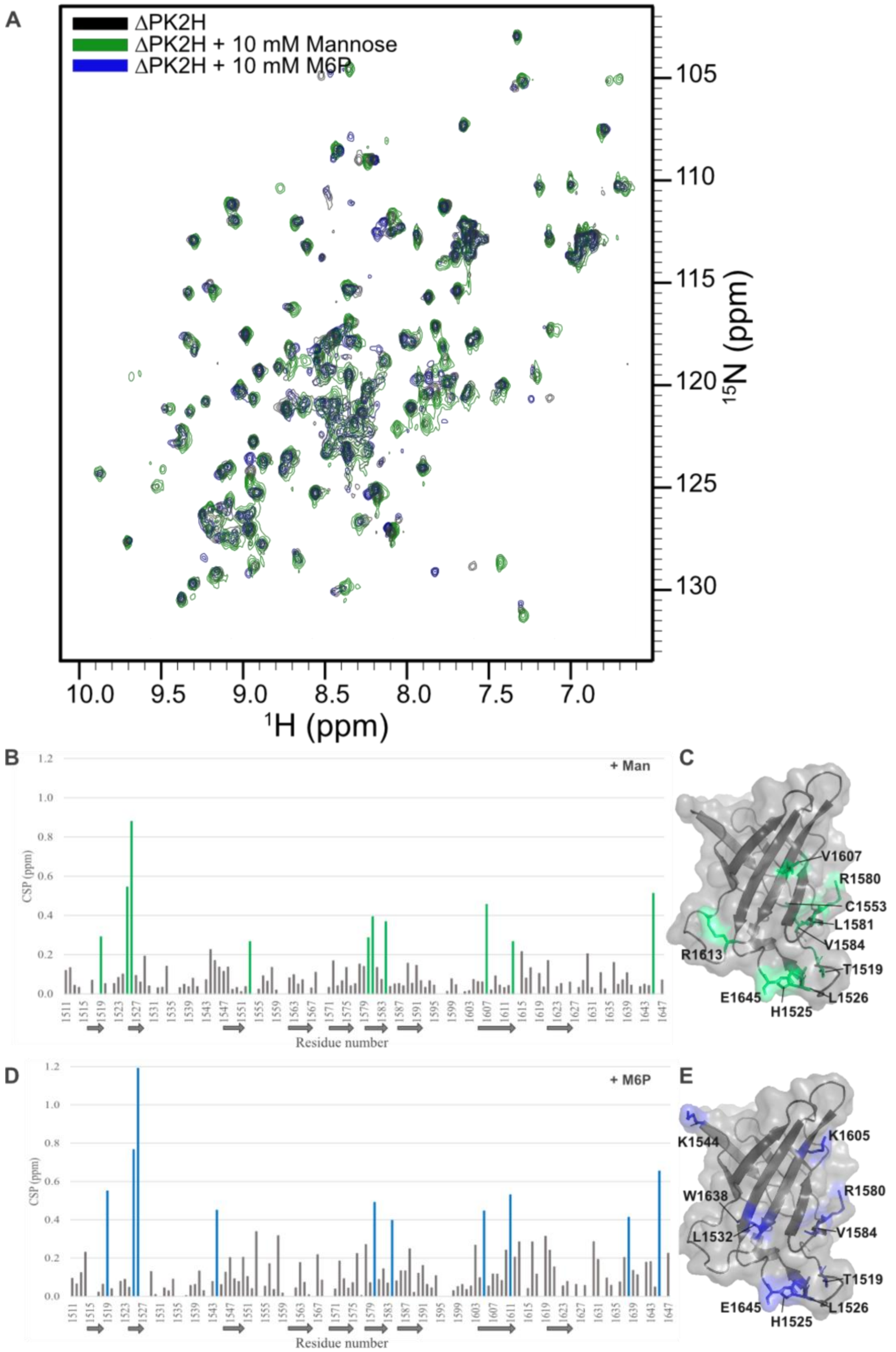


Figure 92: Analysis of D11 mutant ΔPK2H binding M6P at pH 6.5. A: 2D ^{1}H-^{15}N HSQC NMR spectra of D11 ΔPK2H alone (black), with 10 mM mannose (green) and 10 mM M6P (blue). **B:** CSP of D11 ΔPK2H following addition of 10 mM mannose. The ten greatest CSPs are coloured green. β-strands are represented by arrows underneath the residue number. **C:** The ten residues with the greatest CSP values are shown as green sticks on the grey surface of ΔPK2H crystal structure. **D**: CSP of D11 ΔPK2H following addition of 10 mM M6P. The ten greatest CSPs are coloured blue. **E:** The ten residues with the greatest CSP values are shown as green sticks on the grey surface of ΔPK2H crystal structure. Each spectrum was collected at 700 MHz by Dr Chris Williams (University of Bristol) with 100 μM D11 ΔPK2H in 25 mM Bis-Tris, 150 mM NaCl pH 6.5 in 10 % D_2O.

5.5. Conclusions

This chapter describes our structure-based approach to engineer a synthetic lectin using CI-MPR D11 as a scaffold. D11 AB3 has undergone iterative rounds of site directed mutagenesis, expression/ purification, and screening for M6P binding. The first round of mutagenesis incorporated the conserved 'QREY' motif (Q1567, R1604, E1627, Y1634) present in MRH domains. X-ray crystallography revealed that these residues are orientated in the same position as those present within CI-MPR D3 and D9 and should thus be capable of forming hydrogen bonds to M6P. The second round of mutagenesis truncated the FG loop by two residues, ensuring this loop is now the same length as that of M6P binding domains D3 and D9. In the final round of mutagenesis, FG loop mutants ΔPK1G, ΔPK1H, ΔPK2G and ΔPK2H (S1599G, S1599H, S1600G and S1600H respectively) were created to resemble the FG loops of D3 (which contains a di-serine motif, S387 and S386) and D9 (which contains a histidine residue, H1320).

Each D11 FG loop mutant tested (ΔPK1G, ΔPK1H and ΔPK2H) exhibited greater CSPs upon addition of M6P than the earlier construct ΔPK, the negative control protein D11 AB3 and the negative control sugars mannose and glucose. However, affected residues did not cluster around the loops at the top of D11 (the intended binding site) and did not form a distinct binding site elsewhere on the protein surface. The assignment of D11 mutant spectra was based upon a ^{1}H-^{15}N HSQC spectra of AB3 at lower pH (pH 4 versus pH 8.5 or 6.5). Thus, ambiguity arose from changes to buffer composition and pH and from the protein itself, resulting in spectra that are only 76-90 % assigned (Appendix Tables 9-13). In particular the chemical shifts of residues that have been mutated are unknown. These residues have been incorporated to bind M6P and thus it is impossible to know if their chemical shifts are influenced by the addition of M6P. Currently, it is unclear to what extent the observed chemical shift changes can be attributed to sugar-protein interactions and how much is a result of pH reduction upon addition of the acidic M6P monosodium salt. Thus, further work is required to fully characterise these interactions (chapter 6).

6. Future work

6.1. Characterisation of the CI-MPR extracellular region

This work describes the modular approach to structurally characterise the extracellular region of the human CI-MPR, with a particular focus on the elusive, high-affinity, M6P binding domain 9. Work began on single domain constructs (D7, D8) before progressing to di-domain constructs (D9-10), larger tetra-domain constructs (D7-10) and, finally, the full extracellular region (D1-15).

The two single domain constructs, D7 and D8 were previously uncharacterised and as such have no known ligands. Having obtained soluble D7 and D8 by expression in Sf21 insect cells, a pull-down assay should be performed to identify any binding partners. The His_6 tagged domains may be immobilised on Ni-NTA resin, incubated with human serum and any resulting complexes characterised by mass spectrometry.

The structure of D8 was determined to 2.5 Å by X-ray crystallography. However, D7 failed to crystallise. This is likely due to modification with two flexible and heterogeneous N-linked glycans. N-linked glycosylation is sequence specific, with the asparagine residue on the NST sequon (N-X-S/T) being glycosylated.[9] Site-directed mutagenesis was performed to knockout the two predicted glycosylation sites (N921D and N957D) individually and in combination. In an attempt to cause least disruption to secondary and tertiary structure, asparagine was mutated to aspartic acid as is the case following enzymatic de-glycosylation with PNGase F.[191] Work was ongoing to express these constructs in insect cells.

An alternative approach is to express D7 in the presence of glycosylation inhibitors. This has been successfully applied by Dias *et al.* to the chemokine binding protein Evasin-1.[255] Following expression in insect cells, glycosylated Evasin-1, which contains three predicted N-linked glycosylation sites, crystallised and diffracted to 2.7 Å (PDB 3FPT).[255] Dias *et al.* were unable to obtain diffracting crystals of Evasin-1 following enzymatic de-glycosylation by EndoH or PNGaseF.[255] However, addition of the antibiotic Tunicamycin, which inhibits the first step of the N-linked glycosylation pathway (transfer of GlcNAc-1-phosphate to the dolichol lipid anchor),[256] to the insect cell culture medium resulted in soluble, non-glycosylated Evasin-1 that diffracted to 1.6 Å (PDB 3FPR).[255]

Although also modified by two N-linked glycans, a construct encoding D9-10 did crystallise and the structure was determined to 1.5 Å. The M6P binding site of D9 was occupied by a mannose residue from an N-linked glycan of a neighbouring protein. The same mechanism of dimerisation (through a bridging N-linked glycan at D9) was also observed in the crystal structure of D7-11. Further characterisation, such as SEC and SAXS, should be performed on D9-10 and D7-11 at varying pHs and ligand occupancies (M6P-tagged glycoprotein and IGF2) to investigate the regulation of dimerisation and the mechanisms of cargo dissociation. Competition assays should also be performed to determine the affinity between D9 and the mannosylated N-linked glycan.

Following identification of two conserved His-Pro pockets at the interfaces of D9-10 and D11-12 that may form critical pH-dependent hinge points in the CI-MPR extracellular region, knock-out mutation of H1234 of D9 in the D9-10 construct should also be performed to determine the role of this interaction in domain arrangement and receptor structure.

The above experiments to study receptor oligomerisation, ligand occupancy and domain arrangement/ structure should also be performed on the full extracellular region of human CI-MPR. Although, D1-15 was expressed here in Sf21 insect cells, a more concentrated sample of D1-15 alone and in complex with IGF2 is required for preliminary comparison to the 2D classification of D1-15 with M6P and for full structure determination by cryo-EM. This necessitates a more robust expression protocol. Attempts to affect this (by screening harvest times, recombinant bacmids, viral stocks and insect cell lines), were on-going. Work was also on-going to sub-clone and express truncated forms of the extracellular region based upon the recent observation that recombinant human CI-MPR extracellular domain expressed in the mammalian cell line HT1080 was truncated at F1182 of D15.[224]

Nonetheless, the work presented here opens up many new avenues of investigation by cryo-EM and super-resolution light microscopy methods into the domain arrangement, oligomeric state and trafficking of human CI-MPR and the influence of ligand binding.

6.2. Engineering a synthetic lectin

In a parallel project, a structure-based approach was taken to engineer human CI-MPR D11 from a hydrophobic, protein binding domain to a positively charged, carbohydrate binding domain (i.e. to engineer a synthetic lectin). A further round of protein engineering should be performed. Additional structure-based mutations may be suggested from comparison of the crystal structures of D11 FG loop mutants, D3 and D9. For example, the mutations S1599G and S1600H should be combined so that the FG loop resembles that of D9. G1546 in the AB loop should also be mutated to tyrosine as is found in D3 (Y324) and D9 (Y1255). Crystal structures suggest this tyrosine residue may be capable of forming hydrogen bonds to the 2’OH and 3’OH, and CH-π interactions with the 2’C-H.

Following this, a higher throughput method for designing and generating D11 mutants may be desired. In designing the IGF2 trap, Frago *et al.* used a combination of two parallel approaches.[127] The first involved site-directed mutagenesis, *E. coli* expression and screening by NMR, as was employed here.[127] The second approach involved random mutagenesis, expression in *P. pastoris* and screening by yeast surface display FACS (fluorescence activated cell cytometry).[127] All mutants were then validated by SPR.[127] This combined approach allowed screening of libraries of mutant D11s. However, yeast surface display relies upon a fluorescently labelled substrate for selection by FACS.[257] Thus, to select a synthetic lectin for M6P by yeast surface display would require either M6P monosaccharide labelled with a fluorophore (which would likely interfere with any binding event) or an M6P-tagged glycoprotein that can be detected by a fluorescently labelled antibody or streptavidin.

Alternatively, computational methods such as *in silico* molecular docking using the Bristol University Docking Engine (BUDE) may be used to design and screen mutants.[258] Although BUDE has recently been used to identify hotspot residues in protein-protein interactions and characterise the protein-protein interactions in the clathrin coat, it was originally designed to screen libraries of small molecule drug candidates.[258–260] BUDE may therefore be used here to simulate M6P (and other sugars) binding to D11 mutants. Alternative programs such as RosettaScripts and AutoDock Vina could also be employed for *in silico* docking studies.[261,262]

D11 was chosen as the starting construct as a robust *E. coli* expression and *in vitro* refolding protocol already exists, allowing soluble D11 to be obtained in isolation. Soluble D3 and D9, which are both specific for M6P mono-esters, cannot be obtained in isolation. However, having determined the structure of D9 within the di-domain construct D9-10 (PDB 6Z30), it may now be possible to mutate D9 residues at the interface with D10 to improve the solubility of D9 alone.

The D11 mutants engineered to date were analysed by 2D ^{1}H-^{15}N HSQC NMR experiments. ^{1}H-^{15}N HSQC titrations (a protein observed NMR method) are commonly employed during the screening of protein-protein and protein-small molecule interactions as they are direct, reliable, relatively quick and easy to perform.[211] Olson *et al.* have demonstrated the power of ^{1}H-^{15}N HSQC titrations in identifying the binding site residues of bovine D5 and determining its binding affinities for M6P and methylM6P-GlcNAc.[131] Fully assigned HSQC spectra of each D11 FG loop mutant are therefore required. This can be obtained by acquiring 3D NMR spectra of double-labelled (^{15}N, ^{13}C) protein samples.

However, other screening methods should also be employed. Barile *et al.* describe using protein observed 1D ^{1}H-NMR to monitor protein-protein interactions.[211] Chemical shifts may be monitored in uncrowded spectral regions such as below 0.7 ppm, which corresponds to methyl groups, and above 10 ppm, which corresponds to tryptophan side chains.[211] Chemical shift and intensity changes suggest an interaction.[211] However, this is assuming that the binding site contains either methyl groups with chemical shifts below 0.7 ppm or tryptophan residues.[211] Furthermore, there is no information on the location of these interactions on the protein surface.

1D ^{31}P-NMR is an alternative 1D NMR method relevant here due to the phosphate group of M6P. However, further work is required to optimise protein and ligand concentrations to observe a reasonable phosphorous signal in a realistic timeframe.

Alternative ligand observed NMR methods that are frequently used during fragment-based drug discovery, for example STD and WaterLOGSY, may also be employed here to identify an interaction between M6P and D11 mutants. As discussed in chapter 3, these saturation transfer methods are capable of detecting ligands that bind with millimolar to micromolar affinity (K_D mM-µM) due to their slow exchange with the protein (slow k_{off}).[214] We are not expecting to have engineered a high affinity (nM) sugar binder at this point. Therefore, STD and WaterLOGSY are ideal experiments to identify an interaction between D11 mutants and M6P, mannose and glucose. However, again, these experiments will not identify interacting residues.

Non-NMR techniques may also be employed to detect protein-carbohydrate interactions. Glycan microarrays, whereby libraries of glycan structures are immobilised, are a common tool used to characterise lectin binding.[263–265] Song *et al.* and Bohnsack *et al.* describe the use of glycan microarrays to determine the specificities of CI-MPR MRH domains for M6P mono-

or di-esters.[82,133] Another powerful technique to study protein-carbohydrate interactions is ITC. However, as discussed in chapter 3 with D9-10, due to the acidity of M6P monosodium salt, ITC with this ligand requires optimisation. Alternatively, an M6P-tagged glycoprotein, may be used.

Based upon the ^{15}N-^{1}H HSQC experiments performed to date, it is possible that we have engineered sugar binding ability into D11. However, if our D11 mutants are binding M6P monosaccharide it is likely a very weak interaction (possibly still too weak to detect). Thus, further work is required to fully characterise and optimise this.

7. Materials and Methods

7.1. Sources of materials

Unless otherwise stated, reagents were purchased from Sigma Aldrich, Thermo Fisher Scientific and Merck Millipore. Competent *E. coli* cells -NEB 5α and Novagen BL21 (DE3) - were purchased from New England Biolabs and Merck Millipore respectively. All synthetic genes were synthesised by GeneArt Thermo Fisher Scientific. Primers (Appendix Section 8.1.) were purchased from IDT Ltd. PVDF membranes and antibodies for western blotting were purchased from Bio-Rad. The following kits were used: GenElute plasmid miniprep kit (Sigma Aldrich), Qiaquick gel extraction kit (Qiagen) and PureLink PCR purification kit (Thermo Fisher Scientific).

7.2. Techniques

The methods employed here build upon those described in the first- and second-year reports (A. Bochel). Recombinant protein expression in *E. coli*, *in vitro* refolding and crystallisation was based on methods described by Brown *et al.* 2002 and Williams *et al.* 2012.[112,124] Recombinant protein expression in insect cells and general insect cell handling was based upon the standard operating procedures at the Eukaryotic Expression Facility, University of Bristol (Professor Imre Berger) and methods described by Bieniossek *et al.* 2008.[266,184] All X-ray diffraction data and SAXS data was collected on beamlines I03, I04, I24 and B21 at Diamond Light Source.

Sterile technique

All sterile work was performed beside a flame or in a biohood. Media, pipette tips and Eppendorf's were sterilised by autoclaving at 121 °C, 151 psi for 15 minutes. Antibiotics and solutions were sterilised by filtration through a 0.22 µm filter.

Media	**Ingredients**
LB	10 g/L tryptone, 10 g/L NaCl, 5 g/L yeast extract
LB agar	LB, 15 g/L agar
2YT	20 g/L tryptone, 10 g/L yeast extract
5052 solution	25 % w/v glucose, 25 % v/v glycerol
M9 minimal media	3 g/L KH_2PO_4, 1.8 g/L Na_2HPO_4, 0.5 g/L NaCl
Trace metal elements	20 mM $CaCl_2$, 2 mM $CoCl_2.6H_2O$, 2 mM $CuCl_2.2H_2O$, 60 mM H_3BO_3.HCl, 10 mM $MnCl_2.4H_2O$, 2 mM $Na_2MoO_4.2H_2O$, 2 mM $Na_2SeO_3.5H_2O$, 2 mM $NiCl_2.6H_2O$, 2 mM $ZnSO_4.7H_2O$

Table 10: Media used for *E. coli* culture.

Buffer	Procedure	Composition
50x TAE buffer	Agarose gel electrophoresis	242 g Trizma base, 57 mL 100 % glacial acetic acid, 100 mL of 0.5 M EDTA, up to 1 L dH_2O
Resuspension buffer	Resuspension of *E. coli* pellets	50 mM Tris pH 8, 0.5 M NaCl, 0.05 % Triton X-100, 1 mM EDTA, 10% glycerol, 1 mM benzamidine, 5 mM β-mercaptoethanol
Buffer 2	Inclusion body isolation	50 mM Tris pH 8, 0.5 M NaCl, 1 % Triton X-100, 1 mM EDTA, 10 % glycerol, 1 mM benzamidine, 5 mM β-mercaptoethanol, 2 M urea
Buffer 3	Inclusion body isolation	50 mM Tris pH 8, 0.1 M NaCl, 1 mM EDTA, 1 mM benzamidine, 2 M urea
Denaturing buffer	Protein denaturation/ solubilisation	0.1 M Tris pH 8, 0.1 M NaCl, 8 M urea
Refold buffer	Protein refolding	0.1 M Tris pH 8.5, 1 mM EDTA, 1 M L-Arginine, 4 μM benzamidine, 3.7 mM cystamine, 6.5 mM cysteamine
Gel filtration buffer	Gel filtration chromatography	25 mM Tris pH 7.5-8.5, 150 mM NaCl
10 % separating gel	SDS-PAGE	2.5 mL 40 % acrylamide, 3.3 mL gel buffer, 2.8 mL H_20, 50 μL 10 % AMPS, 20 μL TEMED
4 % stacking gel	SDS-PAGE	0.5 mL 40 % acrylamide, 1.25 mL gel buffer, 3.3 mL H_20, 50 μL 10 % AMPS, 20 μL TEMED
Gel buffer	SDS-PAGE	3 M Tris, 10 mM SDS, pH 8.45
Denaturing gel loading dye	SDS-PAGE	780 mM Tris, 142 mM β-mercaptoethanol, 35 mM SDS, 10 % glycerol, 140 mM bromophenol blue, pH 6.8
Native gel loading dye	SDS-PAGE	780 mM Tris, 10 % glycerol, 140 mM bromophenol blue, pH 6.8
Anode buffer	SDS-PAGE	20 mM Tris, pH 8.45
Cathode buffer	SDS-PAGE	100 mM Tris, 100 mM Tricine, 3.5 mM SDS, pH 8.25
Phosphate buffered saline (PBS)	Monitoring protein expression	137 mM NaCl, 2.7 mM g KCl, 10 mM Na_2HPO_4, 1.8 mM KH_2PO_4 pH 7.4
Blocking solution	Western blot	50 mM Tris HCl, 140 mM NaCl pH 7.4, 0.1 % Tween20, 3 % BSA
Tris buffered saline (TBS)	Western blot	50 mM Tris HCl, 140 mM NaCl pH 7.4
TBST	Western blot	50 mM Tris HCl, 140 mM NaCl pH 7.4, 0.1 % Tween20
SP column buffer A	Cation exchange chromatography	50 mM Sodium acetate, pH 5.5
SP column buffer B	Cation exchange chromatography	50 mM Sodium acetate, pH 5.5, 1 M NaCl
IMAC column buffer A	Ni-NTA IMAC	25 mM Tris pH 8.5, 150 mM NaCl
IMAC column buffer B	Ni-NTA IMAC	25 mM Tris pH 8.5, 150 mM NaCl, 800 mM Imidazole
Q column buffer A	Anion exchange chromatography	25 mM Tris pH 7.5
Q column buffer B	Anion exchange chromatography	25 mM Tris pH 7.5, 1 M NaCl
Strep column buffer A	Streptactin affinity chromatography	100 mM Tris pH 8, 1 mM EDTA, 150 mM NaCl
Strep column buffer B	Streptactin affinity chromatography	100 mM Tris pH 8, 1 mM EDTA, 150 mM NaCl, 2.5 mM Desthiobiotin

Table 11: Details of buffers used.

7.3. Methods

1. Agarose gel electrophoresis

Agarose gel solutions were made by dissolving 0.5-1 g agarose in 100 mL 1x TAE buffer. 2 μL Midori Green advance dye (NIPPON Genetics) was added to molten agarose. 5 μL of DNA sample was mixed with 1 μL 5x loading buffer and loaded onto the 0.5-1 % agarose gel. Gels were run in 10x TAE buffer at 200 mV, 100 mA for 20 mins and then visualised under UV light.

2. Cell transformations

1 μL of plasmid DNA was incubated with 50 μL NEB 5α or BL21 Novagen *E. coli* cells on ice for 30 minutes. The cells were transformed by heatshock: 30 seconds at 42 ^{o}C, and then returned to ice for 5 minutes before the addition of 200 μL sterile LB medium and 1-hour incubation at 37 ^{o}C, 220 rpm. 100 μL of culture was then spread onto LB agar plates containing the relevant antibiotic and incubated overnight at 37 ^{o}C before storage at 4 ^{o}C.

3. Isolation of plasmid DNA

Overnight cultures of 10 mL sterile LB, 100 μg/mL antibiotic and a single bacterial colony containing plasmid DNA (prepared as above) were incubated overnight at 37 ^{o}C, 220 rpm. Plasmid DNA was extracted from these cultures using a miniprep kit or alkaline lysis precipitation. Successful isolation of plasmid DNA was confirmed by 1 % agarose gel electrophoresis and the concentration determined by measuring the absorbance at 260 nm (A260) using a MicroVolume DeNovix spectrophotometer.

4. PCR

The following Master mix was made (Table 12) using the KOD hot start DNA Polymerase kit (Sigma-Aldrich). Where necessary up to 5 % DMSO was added. Table 13 shows the standard PCR cycle. Annealing temperatures were screened based upon primer melting temperatures (Appendix Section 8.1. for primer sequences) and extension time was adjusted based upon length of desired PCR product. Following PCR, samples were analysed by gel electrophoresis (see above) and the PCR products were then purified using the PureLink PCR purification kit.

Component	**Volume (μL)**	**Final concentration**
KOD DNA polymerase	1	0.01 units/mL
10X KOD hot start buffer	5	1X
2 mM dNTPs	5	0.2 mM
25 mM $MgSO_4$	3	1.5 mM
Template DNA	1	-
5' Primer	1.5	0.3 mM
3' Primer	1.5	0.3 mM
dH_20	32	-

Table 12: Components of master mix for PCR using KOD polymerase. Where kappa polymerase was used the above concentrations of template DNA and primers were added to a 2x ready-mix.

Step	Temperature (ºC)	Time
Hot start	95	10 mins
Denaturation	95	2 mins
Annealing	Gradient	30 sec
Extension	70	30 sec /0.5 kb
Final extension	70	10 mins
Storage	4	Infinite

Table 13: Standard PCR Cycle. A cycle of denaturation, annealing and extension was repeated 35 times.

5. Site-directed mutagenesis

PCR reactions were set up as per Table 12 above (Appendix Section 8.1. for primer sequences). The annealing temperature and extension time of the PCR cycle in Table 13 were adjusted to reflect the primers used and amplification of the whole plasmid respectively. Again, results were analysed by gel electrophoresis.

Successful reactions were pooled and subject to DpnI digestion - 10 % (v/v) cutsmart buffer and 10 % (v/v) DpnI enzyme were added and samples incubated at 37 ºC for 1 hour before clean up using the PureLink PCR purification kit.

6. Restriction enzyme digests

Per 1 μg template DNA 1 μL restriction enzyme was added along with 5 μL 10x fast digest buffer. The total reaction volume was increased to 50 μL by addition of water. Reactions were incubated at 37 ºC for 1-4 hours before analysis by gel electrophoresis. Digested plasmid samples were excised from the gel and purified using the QIAquick gel extraction kit. Digested PCR products were purified using the PureLink PCR purification kit.

7. Ligation

Ligation reactions were performed using a 1:1 - 1:3 molar ratio of vector: insert DNA with a total DNA concentration of 50-100 ng. The DNA was mixed with 2 μL 10x ligase buffer, 1 μL T4 DNA ligase and the volume increased to 20 μL with water before incubation at room temperature for 10 mins. 5 μL of ligation reaction was added to 50 μL NEB 5α *E. coli* cells. The cell transformation and isolation of plasmid DNA were performed as above. All ligation reactions were confirmed by Sanger sequencing performed by Genewiz.

8. Recombinant protein expression in *E. coli*

A single colony of BL21 (DE3) *E. coli* containing the recombinant plasmid prepared as above was added to 100 mL LB containing 100 μg/mL kanamycin and incubated at 37 ºC, 220 rpm for 16 hours. 2 mL of this pre-culture was used to inoculate 200 mL 2YT media supplemented with 100 μg/mL kanamycin, 5 mL 5052 solution and 1 mM $MgSO_4$. Cultures were incubated at 37 ºC, 220 rpm until they reached an optical density at 600 nm (OD600) of 0.8. The temperature was then lowered to 25 ºC and protein expression induced by addition of 0.5 mM IPTG. After 16 hours, cultures were harvested by centrifugation at 6000 rpm (F10S-6X500 rotor, RC-6 centrifuge) for 10 mins. Cell pellets were resuspended in 50 ml resuspension buffer and inclusion bodies isolated as described below or stored at -20 °C for future analysis.

For isotopic labelling cultures were grown in 2YT media as described above. At OD600 1-2 cells were centrifuged at 4000 rpm, 10 minutes (F10S-6X500 rotor, RC-6 centrifuge). Cell pellets were resuspended and washed in 50 mL pre-aerated M9 media (37 °C, 200 rpm) before a second centrifugation at 4000 rpm, 10 minutes. Cells were resuspended in a fresh 50 mL sample of pre-aerated M9 media and split between 10 500 mL flasks each containing 100 mL pre-aerated M9 media supplemented with 1 g/L $^{15}NH_4Cl$, 6 g/L glucose, 100 μL trace metal solution, 1 mM of $MgSO_4$ and 100 μg/mL kanamycin. Cultures were incubated at 37 ^{o}C, 30 minutes prior to induction with 0.5 mM IPTG. Following overnight incubation at 25 ^{o}C, cultures were harvested and resuspended as above.

9. Inclusion body isolation

Resuspended cell pellets were sonicated (5 sec on, 10 sec off, 80 % amplitude) with a sonicator probe, pelleted (8600 x g, 20 mins) and resuspended in ~ 30 mL buffer 2. This cycle of sonication and centrifugation was repeated with buffer 3. Following centrifugation in buffer 3, supernatant was discarded, and the pellet resuspended in a minimal volume of denaturing buffer supplemented with 10 mM DTT and 6 M NaOH to obtain a final protein concentration of 20-50 mg/ mL. Denatured protein was either stored at -20 ^{o}C or underwent the refolding process described below.

10. Refolding and purification

Denatured protein was diluted 3-fold into fresh denaturing buffer reduced with 100 μL 1M DTT and 100 μL 10 mM EDTA on ice for 1 hour. Denatured protein was then added dropwise to fresh refolding buffer at 4 ^{o}C with rapid mixing. Refold solutions were left at 4 ^{o}C, with gentle stirring, for 24-48 hours.

The resulting protein samples were centrifuged (15 minutes, 8600 x g) to remove precipitate and concentrated using an AMI UHP62 stirred pressure cell (Advantec) at 50-75 psi under nitrogen. The protein was then purified by size exclusion chromatography (SEC) using a Highload S26/60 Superdex 75 prep grade column (GE Healthcare) pre-equilibrated with gel filtration buffer, and a Fast Purification Liquid Chromatography (FPLC) system (GE Healthcare).

11. Recombinant protein expression in insect cells

Bacmid generation and isolation

1 μg plasmid DNA was added to 100 μL DH10EmbacY cells (a gift from the Berger group) and incubated on ice for 30 mins. Following the addition of 400 μL LB, samples were incubated at 37 ^{o}C, 220 rpm for 16 hours. Samples were then spread onto agar plates containing 50 μg/ mL kanamycin, 10 μg/ mL tetracycline, 10 μg/ mL gentamycin, 1 mM IPTG and 100 μg/ mL X-Gal, in a dilution series (1:1, 1:100, 1:1000, 1:10000) and incubated at 37 ^{o}C for a further 16-24 hours.

White colonies containing the recombinant bacmid DNA were then picked and simultaneously streaked onto a fresh LB agar plate (containing the above antibiotics, IPTG and X-Gal) and added to 3 mL of LB containing 50 μg/ mL kanamycin, 10 μg/ mL tetracycline, 10 μg/ mL gentamycin. These cultures were incubated overnight at 37 °C, 220 rpm.

Recombinant bacmid DNA was extracted from confirmed white colonies by alkaline lysis miniprep. Briefly, resuspended cell pellets were lysed and neutralised using the GenElute plasmid miniprep kit (Sigma). The DNA was then precipitated with isopropanol and washed twice with 70 % ethanol. Under a sterile hood the ethanol was removed and the bacmid gently resuspended in 30 μL sterile water. Successful bacmid isolation was confirmed by 0.8 % agarose gel electrophoresis.

Baculovirus generation and amplification

6 well plates were set up as below (Figure 93). 0.5 mL of *Spodoptera frugiperda* cell line 21 (Sf21) insect cells at a density of 0.7-1.0x10^6 cells/mL were added to all wells (except for well M) along with 2.5 mL serum free media (SF900II, Gibco or ESF921, Expression Systems). To wells labelled M (media only) 3 mL media was added.

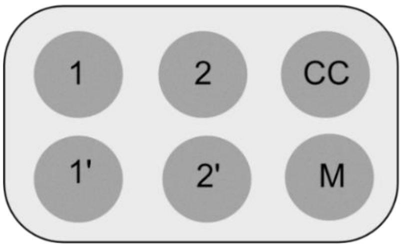


Figure 93: Schematic overview of 6 well plates used in Sf21 culture and virus production. Wells labelled 1-2 correspond to two different bacmid clones, of which 1' and 2' are duplicates of 1 and 2. CC is a cell control of uninfected cells, whilst M is a media only control.

200 μL media was added to each bacmid clone. For x number of bacmid clones x0 μL XtremeGene Transfection Reagent (Sigma) was added to x00 μL medium. 100 μL of this transfection mixture was then added to each bacmid clone and 150 μL of each bacmid-XtremeGene cocktail was added to each of the two dedicated wells (e.g. 1 and 1'). After 48 hours incubating at 27 °C, the supernatant, which contains the V_0 viral titre, was removed from each clone and stored in foil at 4 °C. The supernatant of each infected well was replaced with 3 mL fresh medium and the plates incubated at 27 °C for a further 48 hours before the supernatant was removed and the plate stored at -20 °C for protein expression tests (see 'Monitoring protein expression' below).

To amplify the virus, 3 mL V_0 was added to 25 mL cultures of 0.5x10^6 cells/mL in 250 mL Erlenmeyer flasks. These were incubated at 27 °C, 100 rpm, for approximately 60 hours, with cells counted every 24 hours. On the day after proliferation arrest (DPA) samples containing 1x10^6 cells were taken to monitor protein expression (see below). Samples containing 1x10^6 cells were collected 24 hours later (DPA + 24). Viral titre V_1 was also collected on DPA + 24; cultures were centrifuged (800 rpm, 3 minutes) and the supernatant, which contains V_1, removed and stored in foil at 4 °C. The cell pellet was gently resuspended in 50 mL fresh medium and returned to the shaker flask. Samples containing 1x10^6 cells were taken every 24 hours until cultures were harvested. All cell samples were stored at 4 °C until analysis.

Monitoring protein expression

Each sample containing 1x10^6 cells was first centrifuged for 2 mins at 13 000 rpm before processing as below. Samples in 6 well plates were first resuspended in 500 μL 1xPBS buffer (duplicate wells combined) before above centrifugation.

A. Monitoring YFP

Cell pellets from the above centrifugation were resuspended in 500 μL 1xPBS and sonicated for 5 seconds. A sample of this insoluble cell fraction (I), was taken before centrifugation (12 000 rpm, 10 minutes) and the supernatant, the soluble cell fraction (S), retained. 100 μL S samples were loaded into a FluoroNunc 96 well black plate and YFP emission measured using a Tecan plate reader (BrisSynBio biosuite). 1xPBS was used as a negative control and a YFP standard as a positive control.

B. Monitoring protein of interest

The above samples, alongside samples of the culture media (M) were analysed by SDS-PAGE and western blot to monitor expression and secretion of the protein of interest (see 'Protein detection below').

Large scale protein expression

400 mL cultures (2 L Erlenmeyer flasks) of Sf21 cells at a density of 0.5-1.0x10^6 cells/mL were infected with x mL of V_1 virus (as determined by small scale expression tests). Cultures were maintained at 0.5-1.0x10^6 cells/mL until the day of proliferation arrest (DPA) and samples containing 1.0x10^6 cells were taken daily until the media was harvested.

Baculoviral-infected insect cell (BIIC) stocks were created as follows. On the day of proliferation arrest (DPA) infected insect cells at density 1.0x10^6 cells/mL were gently centrifuged (800 rpm, 10 mins) and resuspended to a density of 1.0x10^7 cells/mL in a solution of 90 % medium (sterile filtered), 10 % DMSO (already sterile) and 10 g/L BSA (sterile filtered). 1 mL aliquots were frozen at -20 ^{o}C for 1 hour before long term storage at -80 ^{o}C. When required, a 1 mL aliquot of BIIC was thawed and diluted to 50 mL with medium before being split between two flasks of 400 mL Sf21 cells at density 1.0x10^6 cells/mL and cultured as above.

Protein purification

All recombinant proteins expressed here using insect cells were secreted into culture medium due to the presence of an N-terminal signal sequence. Culture media containing the protein of interest was harvested by centrifugation at 4000 rpm, 10 mins (F10S-6X500 rotor, RC-6 centrifuge).

For purification of **D11**, the media was adjusted to pH 5.5 by the addition of 0.5 M sodium acetate pH 5.5 before centrifugation at 6000 rpm 20 mins (F10S-6X500 rotor, RC-6 centrifuge). The media was then filtered under vacuum (0.45 μm pore size filters) and pumped onto a 1 mL SP XL column and 1 mL SP FF column in tandem (GE Healthcare) pre-equilibrated with SP column buffer A. The column was washed with SP column buffer A before elution over a gradient of SP column buffer B.

For purification of **D7**, **D8**, **D9-10** and **D1-15**, the media was adjusted to pH 8.3 by the addition of 0.5 M Tris pH 9 before centrifugation at 6000 rpm 20 mins (F10S-6X500 rotor, RC-6 centrifuge). The media was then filtered under vacuum (0.45 μm pore size filters) and pumped onto a 25 mL Ni-NTA column pre-equilibrated with 200 mM $NiSO_4$ and 5 column volumes (CV) of IMAC column buffer A. The column was washed with IMAC column buffer A before elution over a 10 CV gradient of IMAC column buffer B.

For purification of **D7-10**, the media was adjusted to pH 7.5 by the addition of 0.5 M Tris pH 9.0 before centrifugation at 6000 rpm 20 mins (F10S-6X500 rotor, RC-6 centrifuge). The media was then filtered under vacuum (0.45 μm pore size filters) and pumped onto a 25 mL Hi Load^{TM} 16/10 Q sepharose^{TM} high performance anion exchange column (GE healthcare) pre-equilibrated with Q column buffer A. The column was washed with Q column buffer A before elution over a gradient of Q column buffer B. Fractions containing D7-10Strep were purified further by affinity chromatography using a 5 mL StrepTrap^{TM} column (GE healthcare) pre-equilibrated with Strep column buffer A. D7-10Strep was eluted over a gradient of Strep column buffer B. The StrepTrap^{TM} column was regenerated with 1 mM HABA. Fractions from IEX containing D7-10His were purified further by affinity chromatography using a 5 mL HisTrap^{TM} column (GE healthcare) pre-equilibrated with IMAC column buffer A. D7-10His was eluted over a gradient of IMAC column buffer B.

Recombinant proteins were concentrated using Amicon Ultra-15 centrifugal filter units (Sigma) and a sample analysed by analytical SEC using either a Superdex 75 10/200 column or Superdex 200 increase 10/200 GL column (GE Healthcare).

Protein de-glycosylation

Small scale de-glycosylation tests were performed on D7, D8, D9-10 and D7-10 using EndoH and PNGaseF following the manufacturers protocol (NEB). Briefly, 1-20 μg of glycoprotein was denatured by boiling at 95 ^{o}C for 10 minutes with 1 μL 10X NEB denaturing buffer (0.5 % SDS, 40 mM DTT). 2 μL 10x NEB glyco buffer 2 (50 mM sodium phosphate pH 7.5), 2 μL 1x NP40 (1 % NP40) and 1 μL EndoH or PNGaseF were added and the reaction incubated at 37 ^{o}C. Under native conditions the denaturing buffer was replaced with non-denaturing buffer (25 mM Tris pH 7.5, 150 mM sodium chloride) and the samples were not boiled. De-glycosylation was monitored by SDS-PAGE and mass spectrometry.

De-glycosylation of D9-10 was scaled up such that 1500 μg D9-10 protein was incubated with 60 μL NP40, 60 μL NEB glycobuffer 2 and 16 μL PNGaseF (NEB) in a total volume of 600 μL for 16 hrs at 37 ^{o}C. De-glycosylation was again monitored by SDS-PAGE and mass spectrometry. De-glycosylated D9-10 was purified by analytical SEC using a Superdex 75 10/200 column (GE Healthcare).

13. Protein detection

SDS-PAGE and western blot

Samples were prepared by addition of 5x loading dye before boiling (95 ^{o}C, 10 mins). 10 % Tris-tricine acrylamide gels were run for ~45 mins at 150 volts before staining in Coomassie brilliant blue stain (30 mins), and de-staining in water.

For western blotting, samples were run on an SDS-PAGE gel as described above before transfer to a pvdf membrane (Bio-Rad) using a Turbo Transfer machine (Bio-Rad) (1.3 A, 7 mins). Membranes were then incubated in blocking solution for 1 hour at room temperature with shaking before rinsing with TBST buffer. Primary antibody (α-His tag, 1:4000 dilution) was added and membranes incubated for 1 hr at room temp with shaking. Following two five-minute washes with TBST, membranes were visualised by addition of BCIP/NBT (Bio-Rad).

Mass spectrometry

Samples were prepared for mass spectrometry by methanol chloroform extraction.[267] Briefly, 20 μL of protein was mixed with 60 μL methanol and 15 μL chloroform, vortexed and 45 μL distilled water added. After centrifugation (5 mins 9000 x g) the top, organic layer was removed, and the protein precipitated at the interface was resuspended in 45 μL fresh methanol. The protein was pelleted by centrifugation (5 mins 9000 x g). The supernatant was removed and the protein pellet dissolved in a solution of 50% acetonitrile, 0.01 % formic acid.

Data was collected on a Synapt G2-Si (Waters) electrospray ionisation–time of flight (ESI-TOF) mass spectrometer fitted with a Triverse Nanomate (Advion) spraying device in positive ion mode. Samples were sprayed using a capillary voltage of 1.5 kV, set up for resolution mode. Data was acquired over 500-3000 m/z for 10 minutes, before analysis using MassLynx 4.1 or MagTran.

Samples dialysed overnight into 100 mM Ammonium acetate using 7.5 kDa Slide-A-Lyzer mini dialysis caps (ThermoFisher Scientific) were analysed as above on the Synapt G2-Si (Waters) mass spectrometer under Ion Mobility mode, where the ion mobility cell was filled with N_2 at 0.5 mbar. The sample cone was adjusted to 80 V, bias voltage 35 V, trap collision energy 10 V and transfer collision energy 5 V.

14. Protein quantitation

Protein concentration was determined using a DeNovix micro-volume spectrophotometer to measure absorbance at 280 nm (A280) using the Beer-Lambert Law. Extinction coefficients were calculated from the amino acid sequence using ExPASy Protparam.[268]

15. Biophysical techniques

Analytical SEC

Samples of D11WT, D7, D8, D9-10, D7-10 and D1-15 were analysed by analytical SEC using either a Superdex 75 10/200 column or Superdex 200 increase 10/200 GL column (GE Healthcare) and gel filtration buffer.

The columns were calibrated with a combination of Beta amylase, ADH, Conalbumin, BSA, ovalbumin, Carbonic anhydrase, Cytochrome C and aprotinin. A calibration curve was generated using the equation below (Equation 5).

$$K_{av} = \frac{V_e - V_t}{V_t - V_o}$$

Equation 5: Calculation of the particle coefficient value, *K*$_{av}$, for analytical SEC analysis. V_e is the sample elution volume, V_t the total column volume and V_0 the column void volume.

SEC-MALS

The oligomeric state of D7-10 in solution was analysed by SEC-MALS using a Dawn Heleos II light scattering instrument (Wyatt) coupled to an OptilabrEX online refractive index detector (Wyatt). Protein samples (100 μl) were resolved using a Superdex 75 10/200 column (GE Healthcare) running at 0.4 ml/min in 25 mM Tris pH 7.5, 150 mM NaCl before passing through the light scattering and refractive index detectors. The molar mass was calculated using Wyatt's ASTRA software.

SEC-MALS data was collected for the control protein BSA (68 μM), D7-10 (30 μM), D7-10 (30 μM) plus 10-fold excess M6P and D7-10 following overnight incubation at 37 ^{o}C with EndoH and PNGase F (individually).

NMR

All data was collected using a Varian 600 MHz VNMR Spectrometer equipped with a triple-resonance 6.5 mm cryoprobe or Bruker Avance III 700 MHz NMR Spectrometer equipped with a 1.7 mm inverse triple-resonance microcryocoil probe. One-dimensional ^{1}H-NMR spectra and two-dimensional ^{1}H-^{15}N HSQC NMR spectra were acquired for D11 mutants. Spectra were processed with Topsin 3.6 (Bruker) or NMRPipe [269] and analysed using CcpNmr Analysis (version 2.4.2).[270]

One-dimensional ^{1}H-NMR spectra were also acquired for D7, D8 and D9-10. Additional NMR samples of D9-10 were prepared in 50 μL with 25 mM Tris, 150 NaCl pH 7.4 in 60 % D_2O (uncorrected for D_2O). Sugar (20 mM) was used with a final concentration of 2 mM compound. For the STD experiments, the standard Bruker stddiffesgp.3 pulse sequence was used with a saturation time of 7 s and a spectral width of 15.9 ppm with 256 scans. The on-resonance frequency was set to 0.58 ppm, while the off-resonance frequency was set to −28 ppm. Appropriate blank experiments, in the absence of protein or ligand, were performed to test the lack of direct saturation to the ligand protons. For the WATERLOGSY experiments, the standard Bruker ephogsygpno pulse sequence was used with relaxation delay of 1s and a mixing time of 1s with a spectral width of 15.9 ppm with 256 scans. Spectrometer operation was kindly performed by Dr Chris Williams (University of Bristol).

X-ray crystallography

Domain 11 and mutants:

D11 WT and mutants ΔPK1G and ΔPK1H were crystallised by the hanging drop vapour diffusion method using 24 well plates. Screening conditions were based on crystallisation conditions of D11 wild-type (PDB 1GP0) and ΔPK (Dr Chris Williams unpublished data). Crystals were observed in 30 % PEG 4K, 0.1 M Tris pH8.5, 0.2 M sodium acetate for D11WT, 15 % PEG 4K, 0.1 M sodium cacodylate pH 5.2, 0.1 M sodium acetate for ΔPK1G and 15 % PEG 4K, 0.1 M sodium cacodylate pH 5.0, 0.05 M sodium acetate for ΔPK1H. Suitable looking crystals were looped and cryo-cooled in liquid nitrogen, using 20-30 % glycerol as cryo-protectant.

Data was collected on beamlines I04 and I23 at Diamond Light Source (Didcot, UK). Diffraction images were processed and integrated using the Xia2 software package.[271] D11WT was solved using phaser with an existing D11 wild-type structure (PDB 1GP0) as a search model. Using these phases about 90 % of the model was built using autobuild.[272] The remaining residues were manually built in *Coot*[273] and the model was subjected to several rounds of refinement and model building using phenix refine[272] and *Coot.*[273] The final model had an *R*work of 18.3 % and an *R*free of 22.4 % to 2.0 Å resolution. All figures were made in PyMOL.[274]

D11 mutants ΔPK1G and ΔPK1H were solved using phaser with an existing structure of ΔPK as a search model. Again, using these phases about 90 % of the model was built using autobuild.[272] The remaining residues were manually built in *Coot* [273] and the model was subjected to several rounds of refinement and model building using phenix refine [272] and *Coot.*[273] ΔPK1H was refined with tNCS. The final model of ΔPK1G had an *R*work of 26.3 % and an *R*free of 30.6 % to 2.8 Å resolution, while the final model of ΔPK1H had an *R*work of 25.3 % and an *R*free of 28.0 % to 2.5 Å resolution. Figures were made in PyMOL.[274]

Domains 7 and 8:

Commercial sparse matrix crystallisation screens (Morpheus, Structure I and II, PACT Premier and JCSG plus, Molecular Dimensions) were set up with glycosylated D7 and D8 (His_6 tagged, 5-10 mg/mL) in Swissci 96-well plates (Molecular Dimensions) using Art Robbins Gryphon and Phoenix liquid handling robots (BrisSynBio biosuite). After ~6 months crystals of D8 were seen in Structure screen I+II condition E7 (Molecular Dimensions): 1.5 M Ammonium sulfate, 0.1 M Tris pH 8.5, 12 % glycerol. Suitable crystals were looped, dipped in 25 % glycerol and cryo-cooled in liquid nitrogen.

Diffraction data was collected on beamline I24 at Diamond Light Source (Didcot, UK). Diffraction images 1-958 and 1298-1800 were processed and integrated in the Xia2 software package.[271] D8 was solved using phaser with D10 from the D9-10 crystal structure as a search model. Using these phases about 90% of the model was automatically built using autobuild.[272] The remaining residues were manually built in *Coot* [273] and the model was subjected to several rounds of refinement and model building using phenix refine [272] and *Coot.*[273] The final model had an *R*work of 22.3 % and an *R*free of 25.1 % to 2.5 Å resolution. Figures were made in PyMOL.[274]

Domains 9-10:

Commercial sparse matrix crystallisation screens (Morpheus, Structure I and II, PACT Premier and JCSG plus, Molecular Dimensions) were set up with D9-10 (His$_6$ tagged, 4-8 mg/mL) in Swissci 96-well plates (Molecular Dimensions) using Art Robbins Gryphon and Phoenix liquid handling robots (BrisSynBio biosuite). After ~7 weeks crystals were seen in PACT Premier condition A6 (Molecular Dimensions): 0.1 M SPG (succinic acid, sodium dihydrogen phosphate monohydrate, glycine) pH 9, 25 % PEG 1500. Suitable crystals were looped, dipped in 25 % glycerol and cryo-cooled in liquid nitrogen.

Diffraction data was collected on beamline I04 at Diamond Light Source (Didcot, UK). Diffraction images were processed and integrated in the Xia2 software package.[271] D9-10 was solved using phaser with homology models of D9 and D10 as search models. Using these phases about 90% of the model was automatically built using autobuild.[272] The remaining residues were manually built in *Coot* [273] and the model was subjected to several rounds of refinement and model building using phenix refine [272] and *Coot*.[273] Glycans were modelled using the carbohydrate module in *Coot* [273] and validated with Privateer.[275] The final model had an *R*work of 19.9 % and an *R*free of 22.8 % to 1.5 Å resolution. Figures were made in PyMOL.[274]

Domains 7-11:

Note all protein expression, purification, crystallisation and data collection was performed by Hans Hoppe, Karl Harlos and Yvonne Jones (University of Oxford). Data processing and structure determination was performed by Airlie McCoy (University of Cambridge) and Chris Williams (University of Bristol). The structure was refined with Dr Chris Williams using crystal structures of D8 and D9-10 (this work).

Commercial sparse matrix crystallisation screens were set up with domains 7-11 (His$_6$ tagged, 3.5-7 mg/mL) in CrystalQuick 96-well plates (Greiner Bio-One). After ~20 hours crystals were observed in ProPlex condition H06 (Molecular Dimensions): 0.1 M MES pH 6.5, 1.6 M $MgSO_4$ with the addition of 10 mM M6P. These were looped, dipped in perfluoro-polyether-oil and cryo-cooled in liquid nitrogen. Many other crystal conditions were trialled with the inclusion and exclusion of mannose and M6P, but an improved resolution diffraction dataset was not achieved.

Diffraction data was collected on beamline I03 at Diamond Light Source (Didcot, UK). Diffraction images were processed and integrated by XDS[276] in the Xia2 software package.[271] Phaser[272] was used to serially place Rosetta models of D7-10 and the crystal structure of D11 (PDB 1GP0). The model was built through iterative rounds of manual building in *Coot* [273] and refinement in REFMAC5.[277] *B*-factor sharpening was used to enhance the low-resolution maps. Once the model was built, external structural restraints generated with ProSMART [278] using high resolution models for domains 8, 9, 10 and 11 (1.4-2.5 Å) were added to the REFMAC5 refinement.[279] Glycans were modelled using the carbohydrate module in *Coot* [273] and validated with Privateer.[275] The final model had an *R*work of 26.1 % and an *R*free of 30.0 % to 3.5 Å resolution. Figures were made in PyMOL.[274]

Negative stains

Grids of D7-10 (50 μg/mL) and D1-15 (30 μg/mL) were prepared, stained with uranyl acetate and imaged with the help of Dr Sathish Yadavik and Professor Christiane Berger-Schafittzel. Grids were imaged at room temperature using an FEI Tecnai 20 200 kV twin lens transmission electron microscope (Wolfson Bioimaging Facility, University of Bristol). 281 micrographs of D1-15 plus 10-fold excess of M6P were taken. Approximately 15,000 particles were picked and 2D classification performed in Scipion.[228]

Analytical ultracentrifugation

Sedimentation equilibrium-analytical ultracentrifugation (SE-AUC) experiments on D7-10 were carried out with the help of Dr Guto Rhys (University of Bristol) at 20 ^{o}C in a Beckman-Optima XL-I analytical ultracentrifuge. 400 μL of purified protein at OD_{280} = 0.4 was added to the sample channel, whilst the reference channel contained the matching buffer: 25 mM tris pH 7.5, 150 mM NaCl. SE-AUC was performed using an AN50 rotor with an Epon 6 channel centre piece at 7000, 10 000, 13 000, 16 000, 19 000, 22 000 and 25 000 rpm.

Partial specific volumes of the amino acid component ($\bar{v}_A$) and carbohydrate content ($\bar{v}_c$) of D7-10 were estimated using SEDNTERP [220]. Ultrascan II [221] was used to fit the data to a two-component model and calculate molecular weight distributions.

Small angle X-ray scattering

In-line SEC-SAXS of D9-10 was collected with the help of Dr Ash Winter at B21 Diamond Light Source using an Agilent 1200 HPLC and 2.4 mL Superdex S200 column (GE Healthcare). 50 μL of D9-10 (180 μM) was loaded onto the S200 column in running buffer (25 mM Tris, 150 mM sodium chloride) at pH 7.5. Frames were collected at 3 seconds per frame at 25 °C and X-ray scattering was recorded (Pilatus 2M detector) at a fixed camera length of 4.014 m, at 12.4 keV. Angular q range data were collected between 0.0025- 0.34 Å^{-1}. Data reduction, buffer subtraction and modelling of the radius of gyration (R_g), the maximum particle dimension (D_{max}) and the pair distribution function (P(r)) were determined using ScÅtter 3.1r.[280] *Ab initio* bead density shape envelope models for each dataset were generated by programs within the ATSAS 2.7,2 package.[281] DAMMIF[282] averaging over twenty three independent runs using the program DAMAVER,[283] before a single DAMMIN[284] refinement run. *Ab initio* bead density shape envelope models were superimposed to three dimensional structures of proteins using SUPCOMB.[285] FoXS and MultiFoXS [286,287] was used to model flexible regions and quantitatively compare the calculated X-ray scattering of three-dimensional models with the experimental scattering profile of each protein. SEC-SAXS data was also collected for D9-10 (270 μM) in the presence of 100-fold excess (27 mM) M6P and de-glycosylated D9-10 (160 μM) with and without M6P (16 mM).

8. Appendix

8.1. Protein constructs

Amino acid sequences

D7 (orange) and D8 (green) were expressed in *E. coli* in the pET28a vector (black) with an N-terminal hexa-histidine tag (red). D7, D8, D9-10 (blue), D7-10His and D1-15 were expressed in Sf21 insect cells with a C-terminal His$_6$ tag (red). D7-10Strep was expressed in Sf21 insect cells with a C-terminal Strep II tag (purple) and flanked with TEV cleavage sites (pink). Residual residues of the N-terminal RPTPμ signal sequence (GILPSPGMPALLSLVSLLSVLLMGCVAETGAS) required for secretion from insect cells are coloured black. Cysteines are underlined and predicted N-linked glycosylation sites bold. All molecular weights were calculated assuming cysteines are reduced and glycosylation sites not modified. Domain boundaries are coloured according to Brown et al.[112] D11 mutants were expressed in *E. coli* in the pET26b vector (black) with no purification tag.

Domain 7 (*E. coli* expression)			
	MGSSHHHHHHSSGLVPRGSHMMQACSIRDPNSGFVFNLNPLNSSQGYNVSGIGKIFMFNV	969	D7
970	CGTMPVCGTILGKPASGCEAETQTEELKNWKPARPVGIEKSLQLSTEGFITLTYKGPLSA	1029	D7
1010	KGTADAFIVRFVCNDDVYSGPLKFLHQDIDSGQGIRNTYFEFETALACVPSP*	1081	D7
Molecular weight 18,614 Da	**pI 6.7**		

Domain 8 (*E. coli* expression)			
	MGSSHHHHHHSSGLVPRGSHMMVDCQVTDLAGNEYDLTGLSTVRKPWTAVDTSVDGRKRTF	1220	D8
1221	YLSVCNPLPYIPGCQGSAVGSCLVSEGNSWNLGVVQMSPQAAANGSLSIMYVNGDKCGNQR	1181	D8
1182	FSTRFTIECAQISGSPAFQLQDGCEYVFIWRTVEACPVVR*		
Molecular weight 17,561 Da	**pI 6.6**		

Domain 11 (Sf21 expression)			
1511	ETGASNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSEKGLVYMSICGENENCPPGVGA	1575	D11
1576	CFGQTRISVGKANKRLRYVDQVLQLVYKDGSPCPSKSGLSYKSVVISFVCRPEARPTNRP	1626	D11
1627	MLISLDKQTCTLFFSWHTPLACEQATTRHHHHHH*		
Molecular weight 16,818 Da	**pI 8.2**	**N-linked glycosylation sites: 0**	

Domain 7 (Sf21 expression)			
	ETGASTTDTDQACSIRDPNSGFVFNLNPL**N**SSQGY**N**VSGIGKIFMFNVCGTMPVCGTILG	979	D7
980	KPASGCEAETQTEELKNWKPARPVGIEKSLQLSTEGFITLTYKGPLSAKGTADAFIVRFV	1040	D7
1041	CNDDVYSGPLKFLHQDIDSGQGIRNTYFEFETALACVPSPTRHHHHHH*	1081	D7
Molecular weight 18,248 Da	**pI 6.0**	**N-linked glycosylation sites: 2**	

Domain 8 (Sf21 expression)			
	ETGASVDCQVTDLAGNEYDLTGLSTVRKPWTAVDTSVDGRKRTFYLSVCNPLPYIPGCQG	1134	D8
1135	SAVGSCLVSEGNSWNLGVVQMSPQAAA**N**GSLSIMYVNGDKCGNQRFSTRITFECAQISGS	1195	D8
1196	PAFQLQDGCEYVFIWRTVEACPVVRTRHHHHHH*		
Molecular weight 16,661 Da	**pI 6.3**	**N-linked glycosylation sites: 1**	

Domains 9-10 (Sf21 expression)

```
                                ETGASVEGDNCEVKDPRHGNLYDLKPLGLNDT 1248 D9
1249 IVSAGEYTYYFRVCGKLSSDVCPTSDKSKVVSSCQEKREPQGFHKVAGLLTQKLTYENGL 1308 D9
1309 LKMNFTGGDTCHKVYQRSTAIFFYCDRGTQRPVFLKETSDCSYLFEWRTQYACPPFDLTE 1368 D10
1369 CSFKDGAGNSFDLSSLSRYSDNWEAITGTGDPEHYLINVCKSLAPQAGTEPCPPEAAACL 1428 D10
1429 LGGSKPVNLGRVRDGPQWRDGIIVLKYVDGDLCPDGIRKKSTTIRFTCSESQVNSRPMFI 1488 D10
1489 SAVEDCEYTFAWPTATACPMKSTRHHHHHH*
```

Molecular weight 33,588 Da	**pI 6.2**	**N-linked glycosylation sites: 2**

Domains 7-10Strep (Sf21 expression)

```
     ETGENLYFQGACSIRDPNSGFVFNLNPLNSSQGYNVSGIGKIFMFNVCGTMPVCGTILGK  982 D7
 983 PASGCEAETQTEELKNWKPARPVGIEKSLQLSTEGFITLTYKGPLSAKGTADAFIVRFVC 1043 D7
1044 NDDVYSGPLKFLHQDIDSGQGIRNTYFEFETALACVPSPVDCQVTDLAGNEYDLTGLSTV 1104 D8
1105 RKPWTAVDTSVDGRKRTFYLSVCNPLPYIPGCQGSAVGSCLVSEGNSWNLGVVQMSPQAA 1165 D8
1166 ANGSLSIMYVNGDKCGNQRFSTRITFECAQISGSPAFQLQDGCEYVFIWRTVEACPVVRV 1226 D9
1227 EGDNCEVKDPRHGNLYDLKPLGLNDTIVSAGEYTYYFRVCGKLSSDVCPTSDKSKVVSSC 1287 D9
1288 QEKREPQGFHKVAGLLTQKLTYENGLLKMNFTGGDTCHKVYQRSTAIFFYCDRGTQRPVF 1348 D9
1349 LKETSDCSYLFEWRTQYACPPFDLTECSFKDGAGNSFDLSSLSRYSDNWEAITGTGDPEH 1409 D10
1410 YLINVCKSLAPQAGTEPCPPEAAACLLGGSKPVNLGRVRDGPQWRDGIIVLKYVDGDLCP 1470 D10
1471 DGIRKKSTTIRFTCSESQVNSRPMFISAVEDCEYTFAWPTATACPENLYFQGWSHPQFEK*
```

Molecular weight 65,906 Da	**pI 5.2**	**N-linked glycosylation sites: 5**

Domains 7-10His (Sf21 expression)

```
     ETGACSIRDPNSGFVFNLNPLNSSQGYNVSGIGKIFMFNVCGTMPVCGTILGKPASGCEA  989 D7
990  ETQTEELKNWKPARPVGIEKSLQLSTEGFITLTYKGPLSAKGTADAFIVRFVCNDDVYSG 1049 D7
1050 PLKFLHQDIDSGQGIRNTYFEFETALACVPSPVDCQVTDLAGNEYDLTGLSTVRKPWTAV 1109 D7
1110 DTSVDGRKRTFYLSVCNPLPYIPGCQGSAVGSCLVSEGNSWNLGVVQMSPQAAANGSLSI 1169 D8
1170 MYVNGDKCGNQRFSTRITFECAQISGSPAFQLQDGCEYVFIWRTVEACPVVRVEGDNCEV 1229 D8
1230 KDPRHGNLYDLKPLGLNDTIVSAGEYTYYFRVCGKLSSDVCPTSDKSKVVSSCQEKREPQ 1289 D8
1290 GFHKVAGLLTQKLTYENGLLKMNFTGGDTCHKVYQRSTAIFFYCDRGTQRPVFLKETSDC 1349 D9
1350 SYLFEWRTQYACPPFDLTECSFKDGAGNSFDLSSLSRYSDNWEAITGTGDPEHYLINVCK 1409 D9
1410 SLAPQAGTEPCPPEAAACLLGGSKPVNLGRVRDGPQWRDGIIVLKYVDGDLCPDGIRKKS 1469 D10
1470 TTIRFTCSESQVNSRPMFISAVEDCEYTFAWPTATACPMKHHHHHH*
```

Molecular weight 64,113 Da	**pI 5.8**	**N-linked glycosylation sites: 5**

Domains 1-15 sequence (Sf21 expression)

```
     ETGASAAPFPELCSYTWEAVDTKNNVLYKINICGSVDIVQCGPSSAVCMHDLKTRTYHSV    96 D1
97   GDSVLRSATRSLLEFNTTVSCDQQGTNHRVQSSIAFLCGKTLGTPEFVTATECVHYFEWR   157 D1
158  TTAACKKDIFKANKEVPCYVFDEELRKHDLNPLIKLSGAYLVDDSDPDTSLFINVCRDID   218 D2
219  TLRDPGSQLRACPPGTAACLVRGHQAFDVGQPRDGLKLVRKDRLVLSYVREEAGKLDFCD   279 D2
280  GHSPAVTITFVCPSERREGTIPKLTAKSNCRYEIEWITEYACHRDYLESKTCSLSGEQQD   340 D3
341  VSIDLTPLAQSGGSSYISDGKEYLFYLNVCGETEIQFCNKKQAAVCQVKKSDTSQVKAAG   401 D3
402  RYHNQTLRYSDGDLTLIYFGGDECSSGFQRMSVINFECNKTAGNDGKGTPVFTGEVDCTY   462 D3
463  FFTWDTEYACVKEKEDLLCGATDGKKRYDLSALVRHAEPEQNWEAVDGSQTETEKKHFFI   523 D4
524  NICHRVLQEGKARGCPEDAAVCAVDKNGSKNLGKFISSPMKEKGNIQLSYSDGDDCGHGK   584 D4
585  KIKTNITLVCKPGDLESAPVLRTSGEGGCFYEFEWHTAAACVLSKTEGENCTVFDSQAGF   645 D5
645  SFDLSPLTKKNGAYKVETKKYDFYINVCGPVSVSPCQPDSGACQVAKSDEKTWNLGLSNA   706 D5
707  KLSYYDGMIQLNYRGGTPYNNERHTPRATLITFLCDRDAGVGFPEYQEEDNSTYNFRWYT   767 D5
768  SYACPEEPLECVVTDPSTLEQYDLSSLAKSEGGLGGNWYAMDNSGEHVTWRKYYINVCRP   828 D6
829  LNPVPGCNRYASACQMKYEKDQGSFTEVVSISNLGMAKTGPVVEDSGSLLLEYVNGSACT   889 D6
890  TSDGRQTTYTTRIHLVCSRGRLNSHPIFSLNWECVVSFLWNTEAACPIQTTTDTDQACSI   950 D7
951  RDPNSGFVFNLNPLNSSQGYNVSGIGKIFMFNVCGTMPVCGTILGKPASGCEAETQTEEL  1011 D7
1012 KNWKPARPVGIEKSLQLSTEGFITLTYKGPLSAKGTADAFIVRFVCNDDVYSGPLKFLHQ  1072 D7
1073 DIDSGQGIRNTYFEFETALACVPSPVDCQVTDLAGNEYDLTGLSTVRKPWTAVDTSVDGR  1133 D8
1133 KRTFYLSVCNPLPYIPGCQGSAVGSCLVSEGNSWNLGVVQMSPQAAANGSLSIMYVNGDK  1194 D8
1195 CGNQRFSTRITFECAQISGSPAFQLQDGCEYVFIWRTVEACPVVRVEGDNCEVKDPRHGN  1255 D9
1256 LYDLKPLGLNDTIVSAGEYTYYFRVCGKLSSDVCPTSDKSKVVSSCQEKREPQGFHKVAG  1316 D9
1317 LLTQKLTYENGLLKMNFTGGDTCHKVYQRSTAIFFYCDRGTQRPVFLKETSDCSYLFEWR  1377 D9
1378 TQYACPPFDLTECSFKDGAGNSFDLSSLSRYSDNWEAITGTGDPEHYLINVCKSLAPQAG  1438 D10
1439 TEPCPPEAAACLLGGSKPVNLGRVRDGPQWRDGIIVLKYVDGDLCPDGIRKKSTTIRFTC  1499 D10
1500 SESQVNSRPMFISAVEDCEYTFAWPTATACPMKSNEHDDCQVTNPSTGHLFDLSSLSGRA  1560 D11
1561 GFTAAYSEKGLVYMSICGENENCPPGVGACFGQTRISVGKANKRLRYVDQVLQLVYKDGS  1621 D11
1622 PCPSKSGLSYKSVISFVCRPEARPTNRPMLISLDKQTCTLFFSWHTPLACEQATECSVRN  1682 D12
1683 GSSIVDLSPLIHRTGGYEAYDESEDDASDTNPDFYINICQPLNPMHGVPCPAGAAVCKVP  1743 D12
1744 IDGPPIDIGRVAGPPILNPIANEIYLNFESSTPCLADKHFNYTSLIAFHCKRGVSMGTPK  1804 D12
1805 LLRTSECDFVFEWETPVVCPDEVRMDGCTLTDEQLLYSFNLSSLSTSTFKVTRDSRTYSV  1865 D13
1866 GVCTFAVGPEQGGCKDGGVCLLSGTKGASFGRLQSMKLDYRHQDEAVVLSYVNGDRCPPE  1926 D13
1927 TDDGVPCVFPFIFNGKSYEECIIESRAKLWCSTTADYDRDHEWGFCRHSNSYRTSSIIFK  1987 D13
1987 CDEDEDIGRPQVFSEVRGCDVTFEWKTKVVCPPKKLECKFVQKHKTYDLRLLSSLTGSWS  2048 D14
2049 LVHNGVSYYINLCQKIYKGPLGCSERASICRRTTTGDVQVLGLVHTQKLGVIGDKVVVTY  2109 D14
2110 SKGYPCGGNKTASSVIELTCTKTVGRPAFKRFDIDSCTYYFSWDSRAACAVKPQEVQMVN  2170 D15
2171 GTITNPINGKSFSLGDIYFKLFRASGDMRTNGDNYLYEIQLSSITSSRNPACSGANICQV  2231 D15
2232 KPNDQHFSRKVGTSDKTKYYLQDGDLDVVFASSSKCGKDKTKSVSSTIFFHCDPLVEDGI  2292 D15
2293 PEFSHETADCQYLFSWYTSAVCPLGVTRHHHHHH*
```

Molecular weight 248,588 Da	pI 5.7	N-linked glycosylation sites: 19

Domain 11 AB3 sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACFGQTRISVGKANKR LRYVDQVLQLVYKDGSPCPSKSGLSYKSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACEPE	
Molecular weight 15,224 Da	**pI 7.5**

Domain 11 ΔPK sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACFGQTRISVGKANKR LRYVDQVLQLVYKDGSPCSSGLSYKSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACEPE	
Molecular weight 15,224 Da	**pI 7.5**

Domain 11 ΔPK1G sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKR LRYVDQVLQLVYKDGSPCGSGLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACEPE	
Molecular weight 15,194 Da	**pI 7.5**

Domain 11 ΔPK1H sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKR LRYVDQVLQLVYKDGSPCHSGLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACEPE	
Molecular weight 15,274 Da	**pI 7.5**

Domain 11 ΔPK2G sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKR LRYVDQVLQLVYKDGSPCSGGLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACEPE	
Molecular weight 15,194 Da	**pI 7.5**

Domain 11 ΔPK2H sequence (*E. coli* expression)	
MKSNEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAYSKSGVVYMSICGENENCPPGVGACQGQTRISVGKANKR LRYVDQVLQLVYKDGSPCSHGLSRKSVISFVCRPEAGPTNRPMLISEDKQTCTYFFSWHTPLACEPE	
Molecular weight 15,274 Da	**pI 7.5**

Primer sequences

Primer	Sequence
D7F	aac gta **gct agc** ACC ACC GAC ACT GAC CAG GCT
D7R	ttg cat **acg cgt** CGG GCT CGG CAC GCA
D8F	aac gta **gct agc** GTG GAC TGC CAA GTT ACT GAC CTG
D8R	ttg cat **acg cgt** ACG CAC GAC AGG GCA AGC
D9F	aac gta **gct agc** GTT GAA GGC GAC AAC TGC GAA GTG
D10R	ttg cat **acg cgt** GGA CTT CAT AGG GCA AGC AGT AGC

Appendix Table 1: Primers for sub-cloning domains into pFastBac for expression in insect cells (chapters 2 and 3). The adaptor region containing the restriction site (in bold, NheI in forwards primers (F), MluI in reverse primers (R)) is in lower case.

Mutation	Forwards primer	Reverse primer
D8 N1163D	CAG GCT GCC GCT **GAT** GGT TCC CTG TCC	GGA CAG GGA ACC **ATC** AGC GGC AGC CTG
D7 N951D	CTG AAT CCA CTG **GAC** TCC AGC CAG GGT	ACC CTG GCT GGA **GTC** CAG TGG ATT CAG
D7 N957D	AGC CAG GGT TAC **GAC** GTG TCC GGT ATC	GAT ACC GGA CAC **GTC** GTA ACC CTG GCT

Appendix Table 2: Primers for site directed mutagenesis of D7 and D8 (chapter 2). Asparagine residues of predicted N-linked glycosylation sites were mutated to aspartate. Mutated codon in bold.

Mutation	Forwards primer	Reverse primer
ΔPK1G	TCC CCT TGT **GGC** TCC GGC CTG	CAG GCC GGA **GCC** ACA AGG GGA
ΔPK1H	GGG TCC CCT TGT **CAC** TCC GGC CTG AGC	GCT CAG GCC GGA **GTG** ACA AGG GGA CCC
ΔPK2G	TCC CCT TGT AGT **GGA** GGC CTG AGC AGA	TCT GCT CAG GCC **TCC** ACT ACA AGG GGA
ΔPK2H	TCC CCT TGT AGT **CAC** GGC CTG AGC AGA	TCT GCT CAG GCC **GTG** ACT ACA AGG GGA

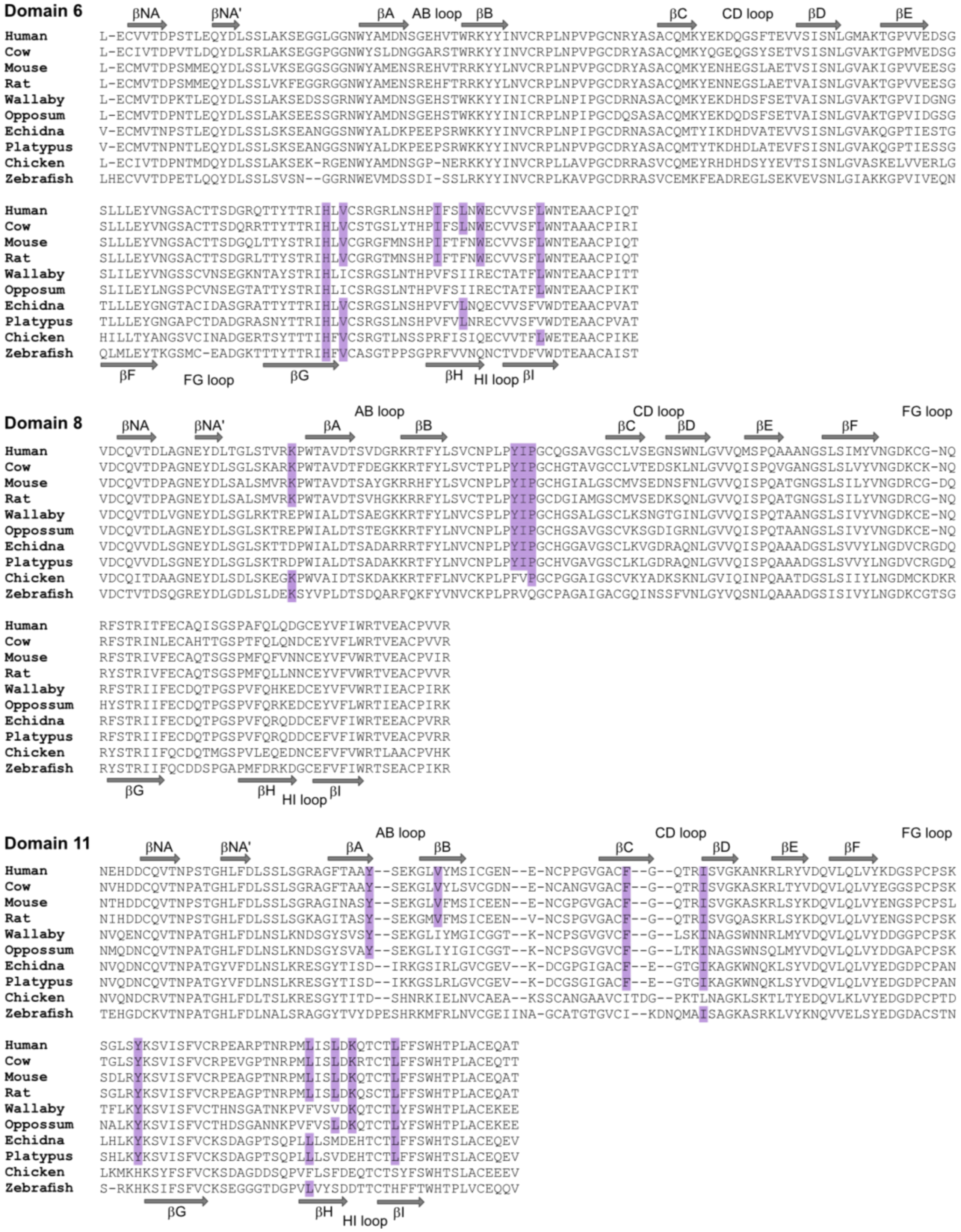
Appendix Table 3: Primers for site directed mutagenesis of D11 (chapter 5). Mutated codon in bold.

pFastBac vector design

ATGGGGATCCTTCCCAGCCCTGGGATGCCTGCGCTGCTCTCCCTCGTGAGCCTTCTCTCCGTGCTGCTGACGGGT
TGCGTAGCTGAAACCGGT**GCTAGC**AACGAGCACGACGACTGTCAAGTGACTAACCCCTCTACCGGTCACCTGTTC
GATCTGTCCTCACTGTCCGGTCGTGCTGGTTTCACCGCTGCCTACTCTGAGAAGGGCCTCGTGTACATGTCCATC
TGCGGAGAGAACGAAAACTGCCCTCCAGGTGTCGGTGCTTGCTTCGGACAGACCCGTATCTCCGTCGGAAAGGCT
AACAAGCGTCTGCGTTACGTGGACCAGGTGCTGCAGCTGGTGTACAAGGACGGATCCCCTTGTCCTTCCAAGTCC
GGCCTGTCCTACAAGTCCGTCATCTCCTTCGTTTGCCGTCCTGAGGCTCGCCCTACCAACCGTCCAATGCTGATC
TCCCTCGACAAGCAGACCTGTACTCTGTTCTTCTCCTGGCACACCCCACTGGCCTGCGAGCAAGCTACT**ACGCGT**
CACCATCATCACCATCAT**TAA**

Appendix Figure 1: Nucleotide sequence of D11WT synthetic gene showing modification of the pFastBac transfer vector. The start and stop codons are shown in bold, MluI (5’) and NheI (3’) restriction sites underlined blue, the RPTPμ signal sequence in green and the hexa-histidine tag in red. DNA encoding D11WT is shown here in black as an example. DNA encoding D7, D8, D9-10 were individually sub-cloned into this pFastBac using MluI and NheI.

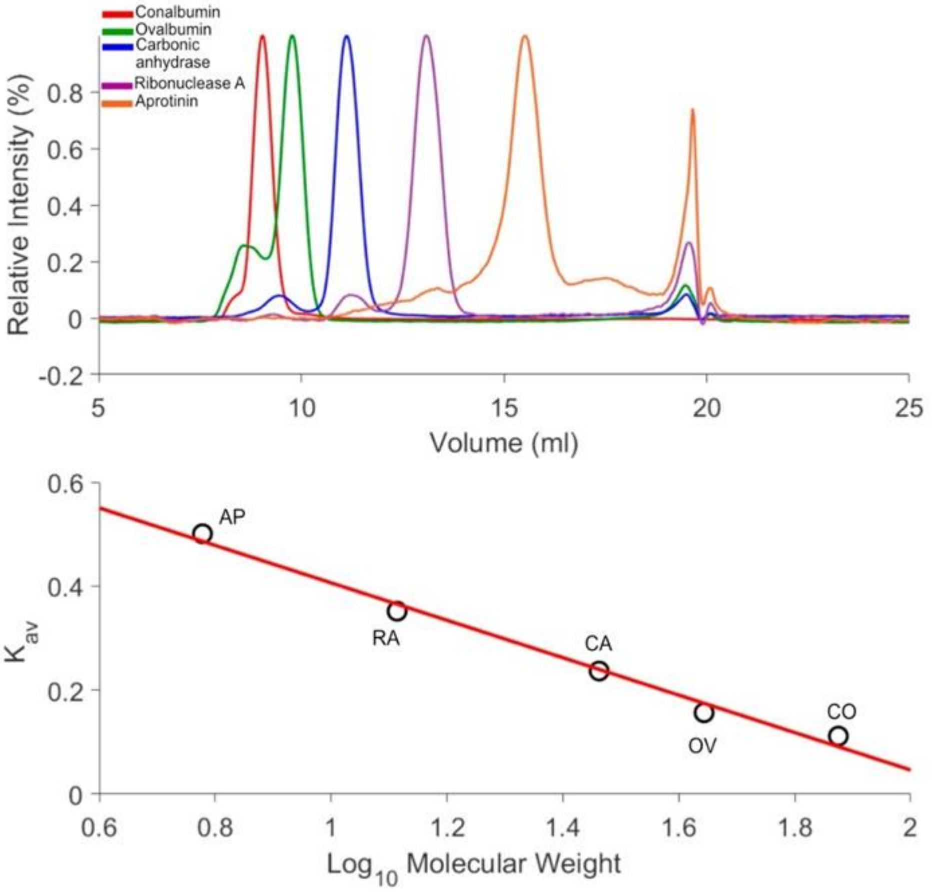
8.2. Chapter 1: Introduction



```
Domain 6
Human      L-ECVVTDPSTLEQYDLSSLAKSEGGLGGNWYAMDNSGEHVTWRKYYINVCRPLNPVPGCNRYASACQMKYEKDQGSFTEVVSISNLGMAKTGPVVEDSG
Cow        L-ECIVTDPVTLDQYDLSRLAKSEGGPGGNWYSLDNGGARSTWRKYYINVCRPLNPVPGCDRYASACQMKYQGEQGSYSETVSISNLGVAKTGPMVEDSG
Mouse      L-ECMVTDPSMMEQYDLSSLVKSEGGSGGNWYAMENSREHVTRRKYYLNVCRPLNPVPGCDRYASACQMKYENHEGSLAETVSISNLGVAKIGPVVEESG
Rat        L-ECMVTDPSMMEQYDLSSLVKFEGGRGGNWYAMENSREHFTRRKYYLNVCRPLNPVPGCDRYASACQMKYENNEGSLAETVAISNLGVAKTGPVVEESG
Wallaby    L-ECMVTDPKTLEQYDLSSLAKSEDSSGRNWYAMDNSGEHSTWKKYYINICRPLNPIPGCDRNASACQMKYEKDHDSFSETVAISNLGVAKTGPVIDGNG
Opposum    L-ECMVTDPNTLEQYDLSSLAKSEESSGRNWYAMDNSGEHSTWKKYYINICRPLNPIPGCDQSASACQMKYEKDQDSFSETVAISNLGVAKTGPVIDGSG
Echidna    V-ECMVTNPSTLEQYDLSSLSKSEANGGSNWYALDKPEEPSRWKKYYINVCRPLNPIPGCDRNASACQMTYIKDHDVATEVVSISNLGVAKQGPTIESTG
Platypus   V-ECMVTNPNTLEQYDLSSLSKSEANGGSNWYALDKPEEPSRWKKYYINVCRPLNPVPGCDRYASACQMTYTKDHDLATEVFSISNLGVAKQGPTIESSG
Chicken    L-ECIVTDPNTMDQYDLSSLAKSEK-RGENWYAMDNSGP-NERKKYYINVCRPLLAVPGCDRRASVCQMEYRHDHDSYYEVTSISNLGVASKELVVERLG
Zebrafish  LHECVVTDPETLQQYDLSSLSVSN--GGRNWEVMDSSDI-SSLRKYYINVCRPLKAVPGCDRRASVCEMKFEADREGLSEKVEVSNLGIAKKGPVIVEQN

Human      SLLLEYVNGSACTTSDGRQTTYTTRIHLVCSRGRLNSHPIFSLNWECVVSFLWNTEAACPIQT
Cow        SLLLEYVNGSACTTSDQRRTTYTTRIHLVCSTGSLYTHPIFSLNWECVVSFLWNTAAACPIRI
Mouse      SLLLEYVNGSACTTSDGQLTTYSTRIHLVCGRGFMNSHPIFTFNWECVVSFLWNTEAACPIQT
Rat        SLLLEYVNGSACTTSDGRLTTYSTRIHLVCGRGTMNSHPIFTFNWECVVSFLWNTEAACPIQT
Wallaby    SLILEYVNGSSCVNSEGKNTAYSTRIHLICSRGSLNTHPVFSIIRECTATFLWNTEAACPITT
Opposum    SLILEYLNGSPCVNSEGTATTYSTRIHLICSRGSLNTHPVFSIIRECTATFLWNTEAACPIKT
Echidna    TLLLEYGNGTACIDASGRATTYTTRIHLVCSRGSLNSHPVFVLNQECVVSFVWDTEAACPVAT
Platypus   TLLLEYGNGAPCTDADGRASNYTTRIHLVCSRGSLNSHPVFVLNRECVVSFVWDTEAACPVAT
Chicken    HILLTYANGSVCINADGERTSYTTTIHFVCSRGTLNSSPRFISIQECVVTFLWETEAACPIKE
Zebrafish  QLMLEYTKGSMC-EADGKTTTYTTRIHFVCASGTPPSGPRFVVNQNCTVDFVWDTEAACAIST

Domain 8
Human      VDCQVTDLAGNEYDLTGLSTVRKPWTAVDTSVDGRKRTFYLSVCNPLPYIPGCQGSAVGSCLVSEGNSWNLGVVQMSPQAAANGSLSIMYVNGDKCG-NQ
Cow        VDCQVTDPAGNEYDLSGLSKARKPWTAVDTFDEGKKRTFYLSVCTPLPYIPGCHGTAVGCCLVTEDSKLNLGVVQISPQVGANGSLSLVYVNGDKCK-NQ
Mouse      VDCQVTDPAGNEYDLSALSMVRKPWTAVDTSAYGKRRHFYLSVCNPLPYIPGCHGIALGSCMVSEDNSFNLGVVQISPQATGNGSLSILYVNGDRCG-DQ
Rat        VDCQVTDPAGNEYDLSALSMVRKPWTAVDTSVHGKKRRFYLSVCTPLPYIPGCDGIAMGSCMVSEDKSQNLGVVQISPQATGNGSLSILYVNGDRCG-NQ
Wallaby    VDCQVTDLVGNEYDLSGLRKTREPWIALDTSAEGKKRTFYLNVCSPLPYIPGCHGSALGSCLKSNGTGINLGVVQISPQTAANGSLSIVYVNGDKCE-NQ
Oppossum   VDCQVTDLAGNEYDLSGLSKTREPWIALDTSTEGKKRTFYLNVCNPLPYIPGCHGSAVGSCVKSGDIGRNLGVVQISPQTAANGSLSIVYVNGDKCE-NQ
Echidna    VDCQVVDLSGNEYDLSGLSKTTDPWIALDTSADARRRTFYLNVCNPLPYIPGCHGGAVGSCLKVGDRAQNLGVVQISPQAAADGSLSVVYLNGDVCRGDQ
Platypus   VDCQVVDLSGNEYDLSGLSKTRDPWIALDTSADAKKRTFYLNVCNPLPYIPGCHVGAVGSCLKLGDRAQNLGVVQISPQAAADGSLSVVYLNGDVCRGDQ
Chicken    VDCQITDAAGNEYDLSDLSKEGKPWVAIDTSKDAKKRTFFLNVCKPLPFVPGCPGGAIGSCVKYADKSKNLGVIQINPQAATDGSLSIIYLNGDMCKDKR
Zebrafish  VDCTVTDSQGREYDLGDLSLDEKSYVPLDTSDQARFQKFYVNVCKPLPRVQGCPAGAIGACGQINSSFVNLGYVQSNLQAAADGSISIVYLNGDKCGTSG

Human      RFSTRITFECAQISGSPAFQLQDGCEYVFIWRTVEACPVVR
Cow        RFSTRINLECAHTTGSPTFQLQNDCEYVFLWRTVEACPVVR
Mouse      RFSTRIVFECAQTSGSPMFQFVNNCEYVFVWRTVEACPVIR
Rat        RYSTRIVFECAQTSGSPMFQLLNNCEYVFVWRTVEACPVVR
Wallaby    RFSTRIIFECDQTPGSPVFQHKEDCEYVFVWRTIEACPIRK
Oppossum   HYSTRIIFECDQTPGSPVFQRKEDCEYVFLWRTIEACPIRK
Echidna    RFSTRIIFECDQTPGSPVFQRQDDCEFVFIWRTEEACPVRR
Platypus   RFSTRIIFECDQTPGSPVFQRQDDCEFVFIWRTVEACPVRR
Chicken    RYSTRIIFQCDQTMGSPVLEQEDNCEFVFVWRTLAACPVHK
Zebrafish  RYSTRIIFQCDDSPGAPMFDRKDGCEFVFIWRTSEACPIKR

Domain 11
Human      NEHDDCQVTNPSTGHLFDLSSLSGRAGFTAAY--SEKGLVYMSICGEN--E-NCPPGVGACF--G--QTRISVGKANKRLRYVDQVLQLVYKDGSPCPSK
Cow        NVHDDCQVTNPATGHLFDLSSLSGRAGFTAAY--SEKGLVYLSVCGDN--E-NCANGVGACF--G--QTRISVGKASKRLTYVDQVLQLVYEGGSPCPSK
Mouse      NTHDDCQVTNPSTGHLFDLSSLSGRAGINASY--SEKGLVFMSICEEN--E-NCGPGVGACF--G--QTRISVGKASKRLSYKDQVLQLVYENGSPCPSL
Rat        NIHDDCQVTNPSTGHLFDLSSLSGKAGITASY--SEKGMVFMSICEEN--V-NCSPGVGACF--G--QTRISVGQASKRLSYKDQVLQLVYENGSPCPSK
Wallaby    NVQENCQVTNPATGHLFDLNSLKNDSGYSVSY--SEKGLIYMGICGGT--K-NCPSGVGVCF--G--LSKINAGSWNNRLMYVDQVLQLVYDDGGPCPSK
Oppossum   NMQDNCQVTNPATGHLFDLNSLKNDSGYSVAY--SEKGLIYIGICGGT--K-NCPSGVGVCF--G--LTKINAGSWNSQLMYVDQVLQLVYDDGAPCPSK
Echidna    NVQDNCQVTNPATGYVFDLNSLKRESGYTISD--IRKGSIRLGVCGEV--K-DCGPGIGACF--E--GTGIKAGKWNQKLSYVDQVLQLVYEDGDPCPAN
Platypus   NVQDNCQVTNPATGYVFDLNSLKRESGYTISD--IKKGSLRLGVCGEV--K-DCGSGIGACF--E--GTGIKAGKWNQKLSYVDQVLQLVYEDGDPCPAN
Chicken    NVQNDCRVTNPATGHLFDLTSLKRESGYTITD--SHNRKIELNVCAEA--KSSCANGAAVCITDG--PKTLNAGKLSKTLTYEDQVLKLVYEDGDPCPTD
Zebrafish  TEHGDCKVTNPATGHLFDLNALSRAGGYTVYDPESHRKMFRLNVCGEIINA-GCATGTGVCI--KDNQMAISAGKASRKLVYKNQVVELSYEDGDACSTN

Human      SGLSYKSVISFVCRPEARPTNRPMLISLDKQTCTLFFSWHTPLACEQAT
Cow        TGLSYKSVISFVCRPEVGPTNRPMLISLDKRTCTLFFSWHTPLACEQTT
Mouse      SDLRYKSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACEQAT
Rat        SGLRYKSVISFVCRPEAGPTNRPMLISLDKQSCTLFFSWHTPLACEQAT
Wallaby    TFLKYKSVISFVCTHNSGATNKPVFVSVDKQTCTLYFSWHTPLACEKEE
Oppossum   NALKYKSVISFVCTHDSGANNKPVFVSLDKQTCTLYFSWHTPLACEKEE
Echidna    LHLKYKSVISFVCKSDAGPTSQPLLLSMDEHTCTLFFSWHTSLACEQEV
Platypus   SHLKYKSVISFVCKSDAGPTSQPLLLSVDEHTCTLFFSWHTSLACEQEV
Chicken    LKMKHKSYFSFVCKSDAGDDSQPVFLSFDEQTCTSYFSWHTSLACEEEV
Zebrafish  S-RKHKSIFSFVCKSEGGGTDGPVLVYSDDTTCTHFFTWHTPLVCEQQV
```

Appendix Figure 2: Sequence alignment of CI-MPR D6, D8 and D11 from placental mammals (human, cow, mouse, rat), marsupial mammals (wallaby), monotreme mammals (opossum, echidna), platypus, birds (Chicken) and fish (zebrafish). Residues that interact with IGF2 are highlighted purple.

8.3. Chapter 2: Domains 7 and 8



Sample	Mw (kDa)	Mw_{app} (kDa)	$Log_{10}Mw$	V_e (ml)	K_{av}
Conalbumin	75	65.9	1.87	9.02	0.11
Ovalbumin	44	49.4	1.64	9.78	0.16
Carbonic anhydrase	29	29.6	1.46	11.13	0.24
Ribonuclease A	13	14.2	1.11	13.07	0.35
Aprotinin	6	5.5	0.78	15.58	0.50
D11 Peak 1	33.6	40.5	1.53	10.3	0.19
D11 Peak 2	16.8	15.7	1.23	12.8	0.34
D7	19.9	21.3	1.30	12.0	0.29
D8	17.3	15.5	1.24	12.8	0.34

Appendix Figure 3: Analytical SEC of D11, D7 and D8 using a Superdex 75 10/300 column. Top: SEC chromatograms of calibrants: Aprotinin (AP), Ribonuclease A (RA), Carbonic Anhydrase (CA), Ovalbumin (OV) and Conalbumin (CO). **Middle:** Calibration curve of the calibrants used to determine the Mw_{app} for analysed proteins. **Bottom:** Summary of results for each calibrant, D11, D7 and D8 (red). D11 has a monomeric molecular weight of ~16.8 kDa (peak 2) and dimeric molecular weight of ~33.6 kDa (peak 1). D7 has a glycosylated monomeric molecular weight of ~20 kDa, while D8 has a glycosylated, monomeric molecular weight of ~17 kDa. The SEC column was calibrated by Dr Ash Winter (University of Bristol).

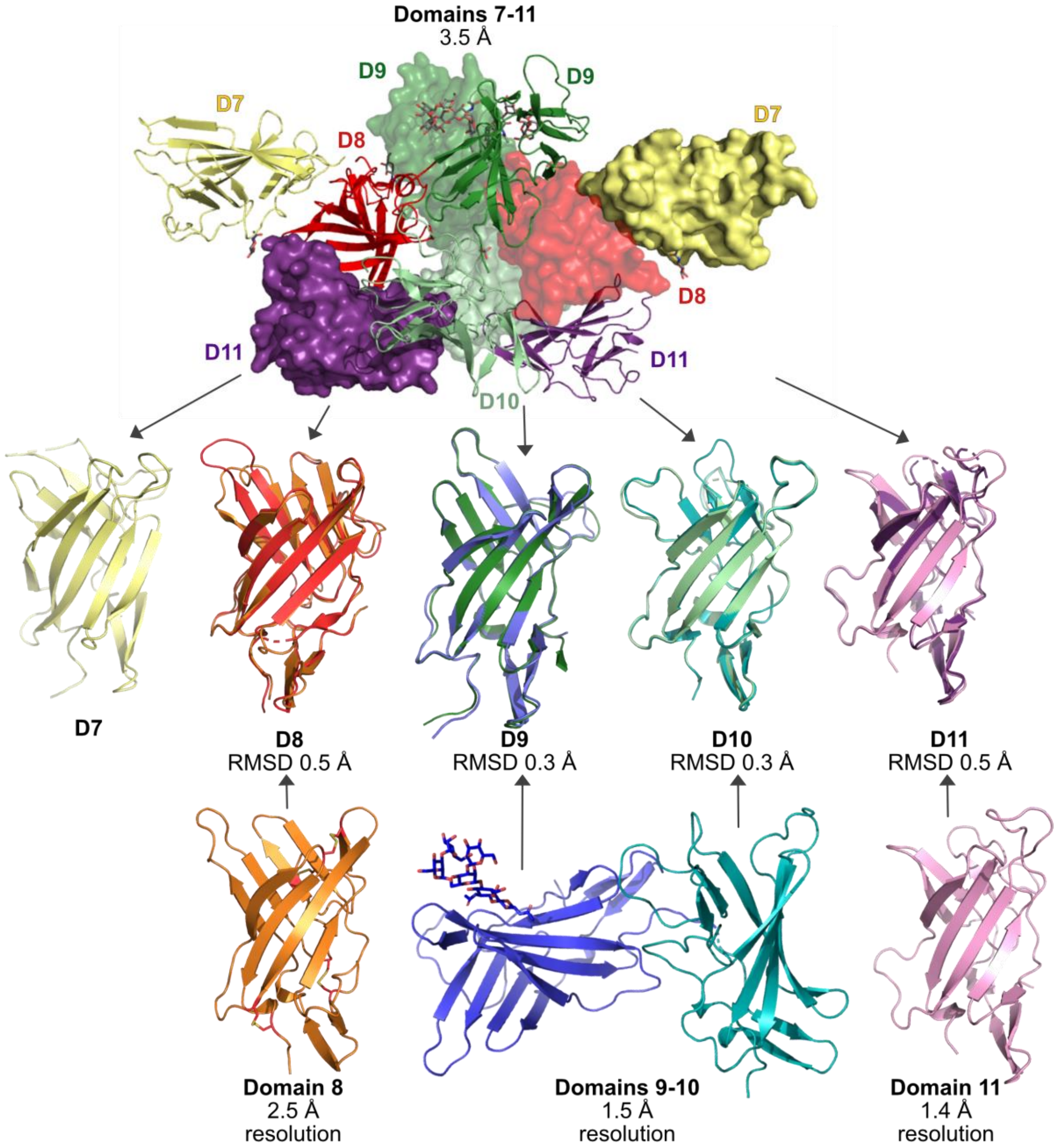
Construct	**Domain 11**	**Domain 8**
Data collection		
PDB accession code	N/A	**6Z31**
Space group	P 2_1 2_1 2_1	P12$_1$1
Unit cell		
a, b, c (Å)	31.5, 49.9, 82.9	65.6, 45.6, 70.3
α, β, γ (°)	90, 90, 90	90, 116, 90
X-ray wavelength (Å)	0.97	0.97
Resolution range (Å)	40.73-2.00 (2.07-2.00)	36.02-2.56 (2.65-2.56)
Total reflections	18262 (1794)	56943 (5415)
Unique reflections	9145 (897)	12121 (1170)
Multiplicity	2.0 (2.0)	4.7 (4.6)
Completeness (%)	98.2 (100)	98.98 (95.82)
R_{meas}	0.05 (0.29)	0.21 (0.71)
Mean I/σ (I)	9.6 (3.1)	5.59 (1.97)
Wilson B-factor (Å^2)	24.41	40.52
CC ½	0.997 (0.871)	0.975 (0.686)
Refinement		
Reflections used in refinement	9144 (897)	12116 (1169)
R_{work} (%)	18.3 (22.9)	22.3 (27.1)
R_{free} (%)	22.4 (29.2)	25.1 (38.4)
Root mean squared deviation		
Bond lengths (Å)	0.004	0.004
Bond angles (°)	1.01	0.70
Ramachandran plot (%)		
Favoured	98.5	94.9
Allowed	1.5	5.1
Outliers	0.0	0.0
Average B-factor (Å^2)	33.61	30.62
Protein	33.12	30.70
Ligand/ glycan	74.81	41.31

Appendix Table 4: Data collection and refinement statistics for D11 and D8 crystal structures. Values in parentheses are for the outer resolution shell.

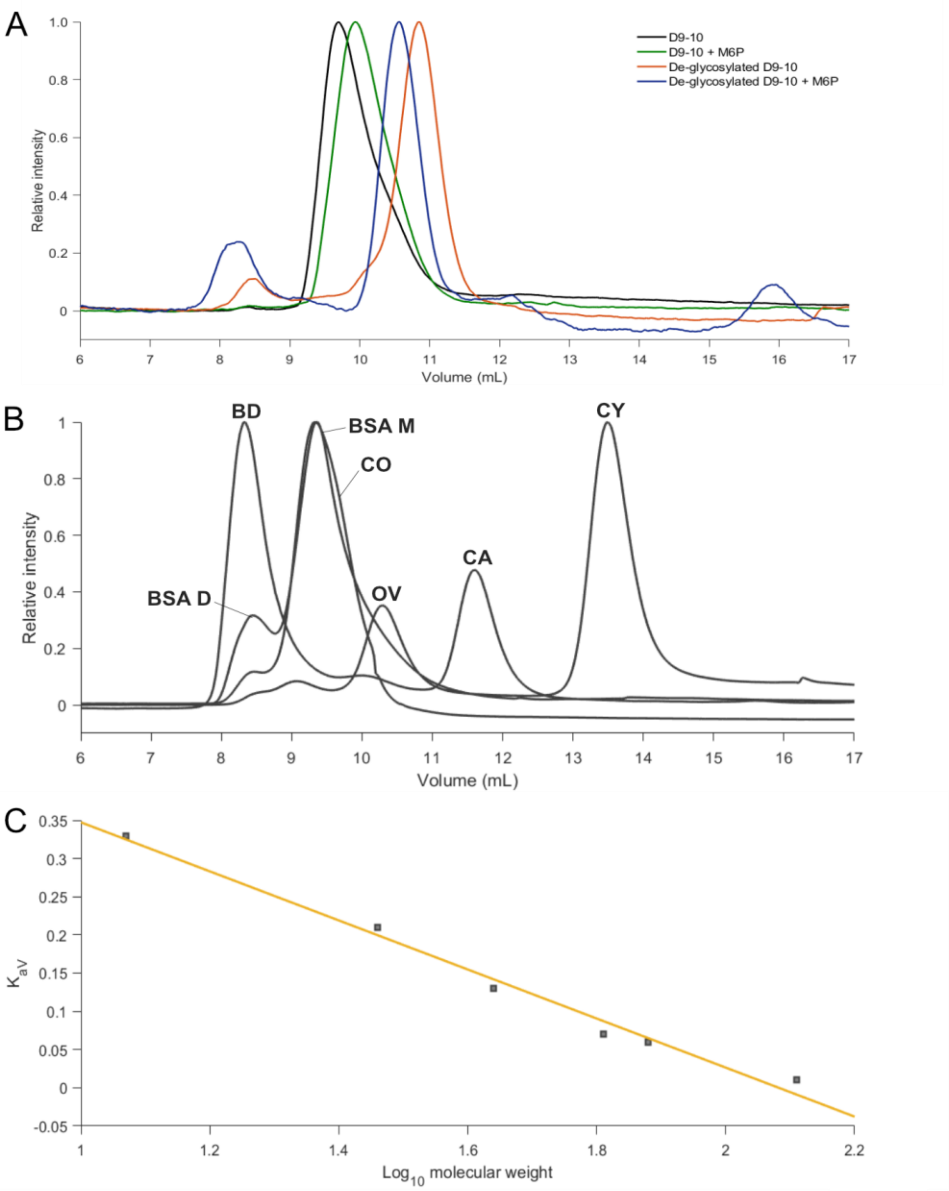
8.4. Chapter 3: Structural characterisation of D9

Construct	**Domains 9-10**	**Domains 7-11**
Data collection		
PDB accession code	6Z30	6Z32
Space group	P 2_1 2_1 2_1	P 4_1 2_1 2
Unit cell		
a, b, c (Å)	40.4, 55.8, 32.7	139.2 139.2 234.7
α, β, γ (°)	90, 90, 90	90, 90, 90
X-ray wavelength (Å)	0.98	1.06
Resolution range (Å)	51.46-1.50 (1.55-1.50)	89.72-3.47 (3.59-3.47)
Total reflections	630776 (64,630)	438,999 (43,737)
Unique reflections	48,880 (4805)	30,224 (2944)
Multiplicity	12.9 (13.5)	14.5 (14.9)
Completeness (%)	99.9 (99.9)	98.6 (98.4)
R_{meas}	0.15 (2.4)	0.21 (5.06)
Mean I/σ (I)	11.12 (1.24)	6.03 (0.84)
Wilson B-factor (Å^2)	21.15	157.79
CC ½	0.999 (0.488)	0.997 (0.127)
Refinement		
Reflections used in refinement	48877 (4805)	30215 (2943)
R_{work} (%)	19.9 (30.02)	26.1 (42.1)
R_{free} (%)	22.8 (32.0)	30.0 (43.2)
Root mean squared deviation		
Bond lengths (Å)	0.005	0.014
Bond angles (°)	0.84	1.97
Ramachandran plot (%)		
Favoured	98.6	94.5
Allowed	1.4	5.1
Outliers	0.0	0.4
Average B-factor (Å^2)	27.43	197.43
Protein	25.78	197.85
Ligand/ glycan	22.68	174.54

Appendix Table 5: Data collection and refinement statistics for D9-10 and D7-11 crystal structures. Values in parentheses are for the outer resolution shell.



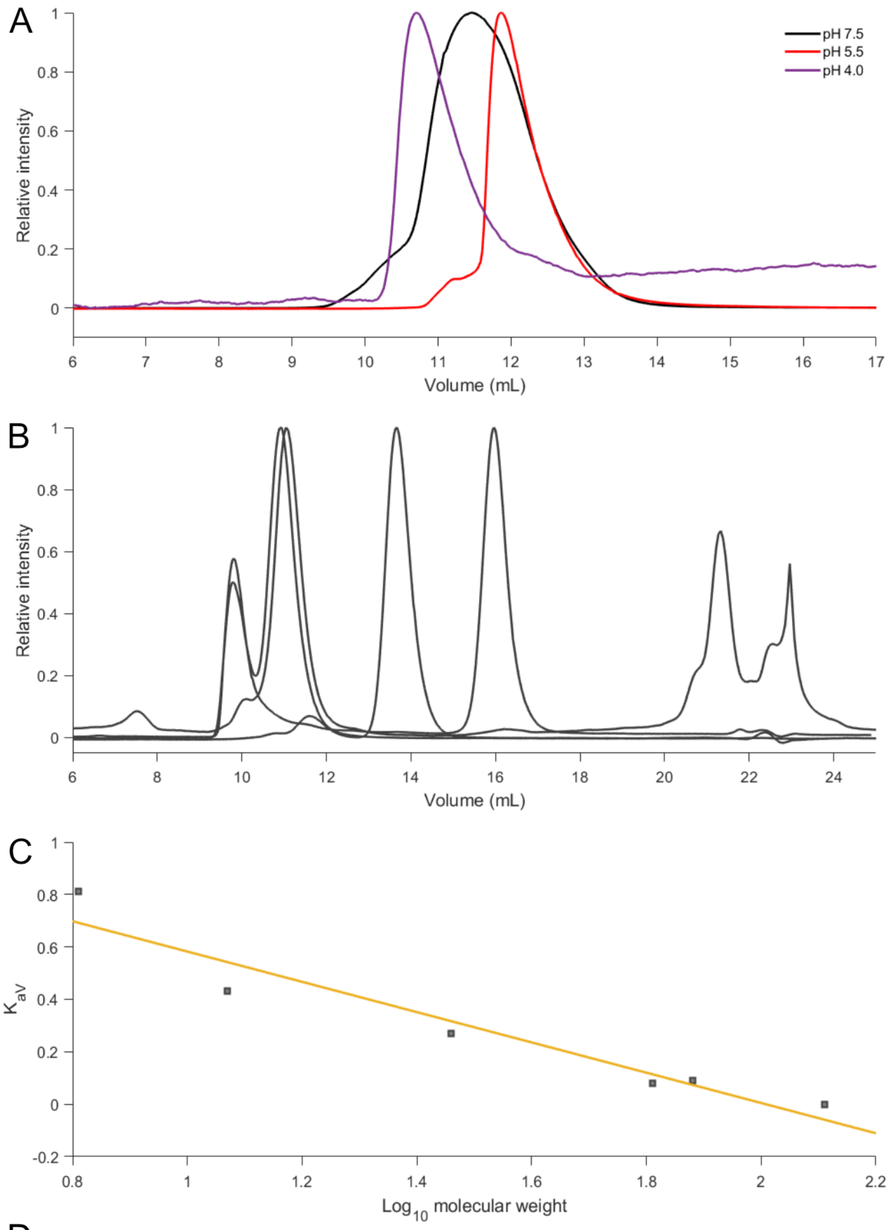
Appendix Figure 4: Crystal structures used in the phasing and refinement of human D7-11 crystal structure. D7-11 was phased by Airlie McCoy (University of Cambridge) using homology models of D7, D8, D9, D10 and the high-resolution (1.4 Å) crystal structure of D11 (PDB 1GP0). Chris Williams (University of Bristol) refined the D7-11 structure using the high-resolution crystal structures of D8 (2.5 Å) and D9-10 (1.5 Å). Superimposition of individual domains of the D7-11 structure with those of the high-resolution reveals close structural similarities.



D

Sample	Mw (kDa)	Mw$_{app}$ (kDa)	Log$_{10}$Mw	V$_e$ (ml)	K$_{av}$
Blue dextran	2 000 000	-	6.30	8.32	0.00
BSA dimer	130	112.7	2.11	8.43	0.01
Conalbumin	75	75.6	1.88	9.32	0.06
BSA monomer	65	74.4	1.81	9.36	0.07
Ovalbumin	44	49.4	1.64	10.28	0.13
Carbonic anhydrase	29	27.3	1.46	11.61	0.21
Cytochrome C	12	11.7	1.07	13.50	0.33
D9-10	72	64.2	1.86	9.69	0.09
D9-10 + M6P	72	57.4	1.86	9.94	0.1
De-glycosylated D9-10	33	38.8	1.52	10.82	0.16
De-glycosylated D9-10 + M6P	33	43.2	1.52	10.58	0.14

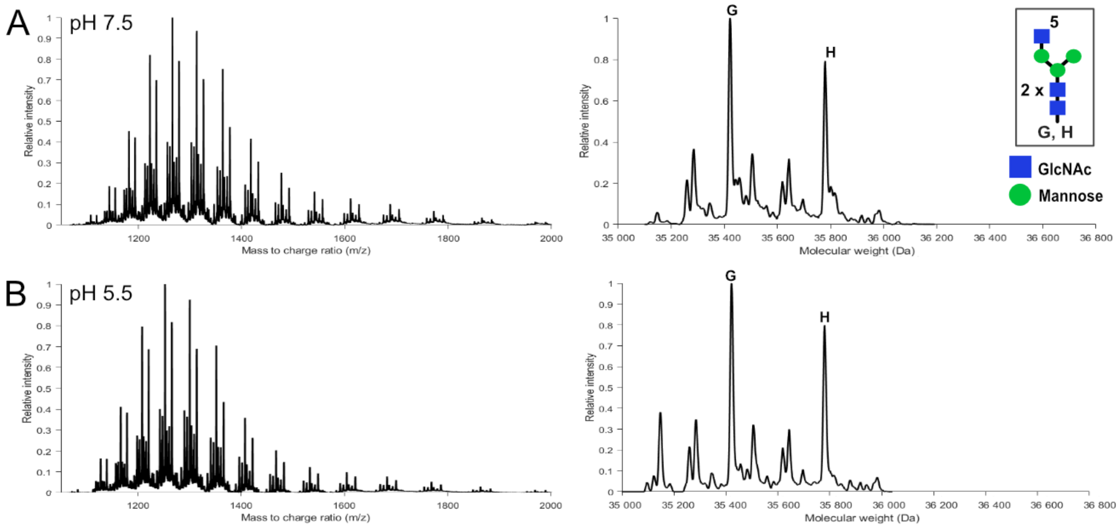
Appendix Figure 5: analytical SEC of D9-10 using a Superdex 75 10/300 column. A: chromatograms of D9-10 with (green) and without M6P (black) and de-glycosylated D9-10 with (blue) and without (orange) M6P. **B:** chromatogram of calibrants: Blue dextran (BD), BSA dimer (BSA D), Conalbumin (CO), BSA monomer (BSA M), Ovalbumin (OV), Carbonic anhydrase (CA), Cytochrome C (CY). **C:** Calibration curve of the calibrants used to determine the Mw$_{app}$ for analysed proteins. **D:** Summary of results for each calibrant and D9-10. D9-10 has a glycosylated monomeric molecular weight of ~36 kDa, glycosylated, dimeric molecular weight of ~72 kDa and de-glycosylated monomeric molecular weight of ~33 kDa.



D

Sample	Mw (kDa)	Mw_{app} (kDa)	$Log_{10}Mw$	V_e (ml)	K_{av}
Blue dextran	2000	-	6.30	9.81	0.00
BSA dimer	130	101.2	2.11	9.80	0.00
Conalbumin	75	71.5	1.88	11.06	0.09
BSA monomer	65	74.7	1.81	10.90	0.08
Carbonic anhydrase	29	35.1	1.46	13.64	0.27
Cytochrome C	12	18.5	1.07	15.96	0.43
Aprotinin	6	1.2	0.81	21.32	0.81
D9-10 pH 7.5	72	64.2	1.86	11.45	0.12
D9-10 pH 5.5	72	57.2	1.86	11.87	0.15
D9-10 pH 4.0	72	79.2	1.52	10.69	0.06

Appendix Figure 6: analytical SEC of D9-10 using a Superdex 75 10/300 column. A: chromatograms of D9-10 at pH 7.5 (black), pH 5.5 (red) and pH 4.0 (purple). **B:** chromatogram of calibrants: Blue dextran (BD), BSA dimer (BSA D), Conalbumin (CO), BSA monomer (BSA M), Carbonic anhydrase (CA), Cytochrome C (CY) and Aprotinin (AP). **C:** Calibration curve of the calibrants used to determine the Mw_{app} for analysed proteins. **D:** Summary of results for each calibrant and D9-10. D9-10 has a glycosylated monomeric molecular weight of ~36 kDa, glycosylated, dimeric molecular weight of ~72 kDa and de-glycosylated monomeric molecular weight of ~33 kDa.



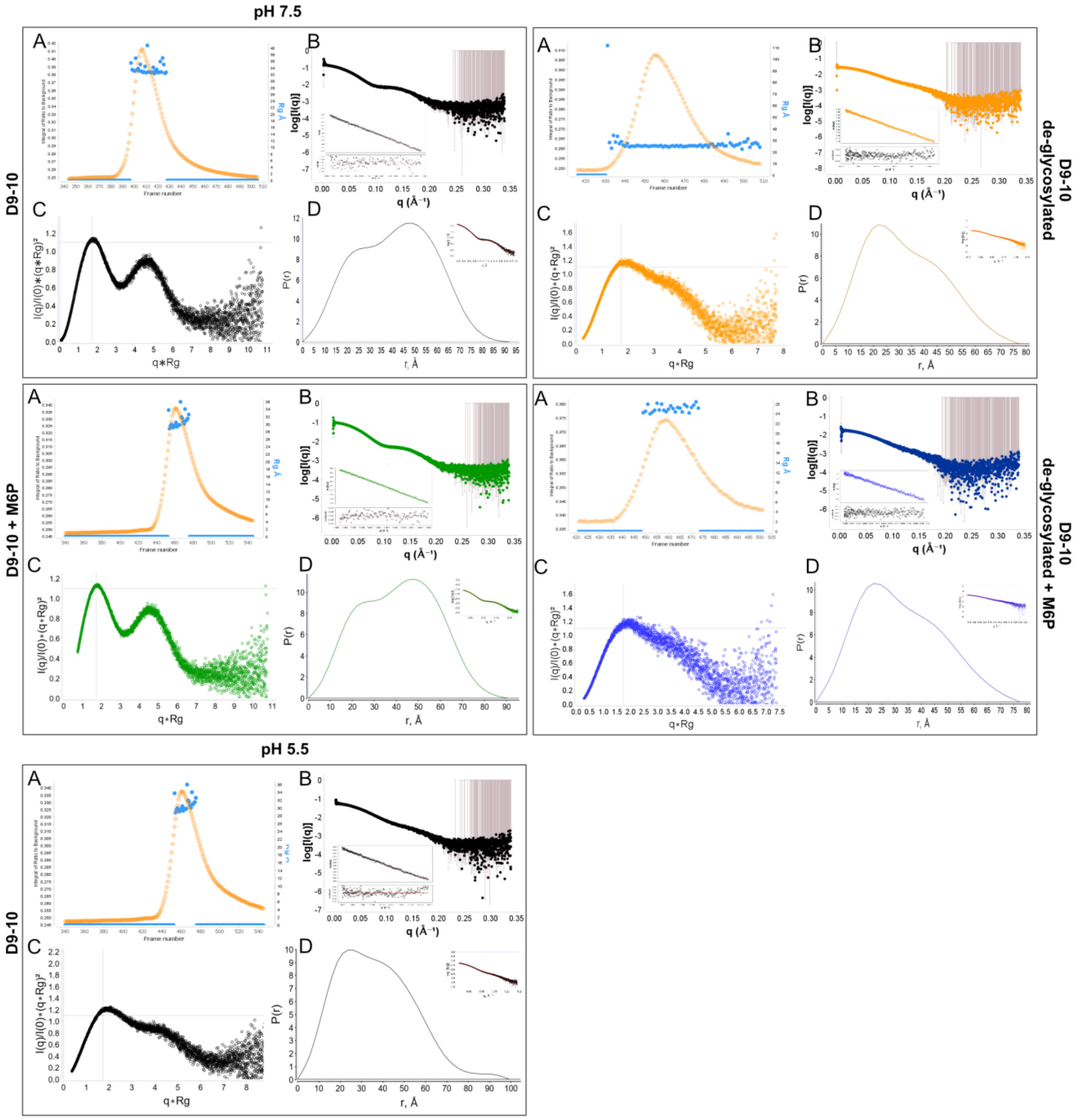
Spectra	Label	Species	Observed mass (Da)	Expected mass (Da)	Difference (Da)
D9-10	G	Truncated D9-10 with two glycan 5	35423.0	35422.2	+0.8
pH 7.5	H	Full length D9-10 with two glycan 5	35781.0	35780.5	+0.5
D9-10	G	Truncated D9-10 with two glycan 5	35424.0	35422.2	+1.8
pH 5.5	H	Full length D9-10 with two glycan 5	35782.0	35780.5	+1.5

Appendix Figure 7: ESI-MS of D9-10 at different pH. A: D9-10 at pH 7.5 is glycosylated at both sites by glycan 5, $GlcNAc_2Man_3GlcNAc$. **B:** The molecular mass of D9-10 at pH 5.5 is unchanged.

8. **Appendix**

Data collection parameters					
Sample	**D9-10 pH 7.5**	**D9-10 pH 7.5 + M6P**	**D9-10 pH 7.5 de-glycosylated**	**D9-10 pH 7.5 de-glycosylated +M6P**	**D9-10 pH 5.5**
SASBDB code	**SASDH79**	**SASDH59**	**SASDH69**	**SASDJ23**	**N/A**
Instrument	SEC-SAXS at B21 Diamond Light Source				
SEC Column	Superdex 200 Increase 3.2/300				
Loading concentration (mg ml^{-1})	6.5				
Injection volume (μl)	45				
Flow rate (ml min^{-1})	0.075				
Average C in combined data frames (mg ml^{-1})	1.65				
Solvent	25 mM Tris-HCl, 150 mM NaCl				
Data analysis					
Guinier analysis:					
I(0) (cm^{-1})	1.37E-1± 1.4E-4	9.72E-2± 1.3E-4	2.74E-2 ±2.3E-4	1.74E-2 ±2.9E-5	4.96E-2 ±9.2E-5
Rg (Å)	32.79±0.19	32.51±0.19	25.21±0.13	25.05±0.21	29.05±0.3
q_{min} (Å^{-1})	0.015	0.022	0.011	0.012	0.009
qR_g max (q_{min} max)	1.25 (0.034)	1.27 (0.034)	1.29 (0.040)	1.29 (0.041)	1.29 (0.034
Coefficient of Correlation	0.9980	0.9986	0.9961	0.9898	0.9893
P(r) analysis:					
I(0) (cm^{-1})	1.31E-1 ±1.5E-3	9.29E-2 ±1.3E-3	2.76E-2 ±1.8E-4	1.79E-2 ±1.6E-4	4.83E-2 ±4.0E-4
Rg (Å)	31.38±0.20	31.16±0.18	25.23±0.11	25.32±0.13	28.65±0.22
d_{max} (Å)	92.0	91.0	77.0	77.5	93.5
q range (Å^{-1})	0.0026-0.34	0.0026-0.34	0.0026-0.34	0.0026-0.34	0.0090-0.21
χ^2	0.77	0.92	1.18	1.06	1.03
Porod volume (Å^{-3})	135,067	127,660	68,881	71,840	112,600
(ratio V_P/calculated M)	(2.01)	(1.87)	(2.05)	(2.14)	(3.35)
Resolution (from SASRES) (Å)	39 ± 3	37 ± 3	30 ±2	34 ±3	33 ±3
DATCLASS designation	Flat	Flat	Flat	Compact	Flat
Molecular mass determination (ratio to predicted value)					
Partial specific volume ($\bar{v}$, cm^{3}g^{-1})	0.726				
Particle contrast from sequence and solvent constituents $\Delta\bar{\rho}$ ($P_{\text{protein}} - P_{\text{solvent}}$; 10^{10} cm^{-2})	3.049 (12.462 – 9.413)				
Mr from sequence (KDa)	33.6	33.6	33.6	33.6	33.6
Mr form I(0) Reciprocal (KDa)	83.6 (1.24)	78.9 (1.17)	48.0 (1.43)	41.8 (1.24)	67.1 (1.99)
Mr from I(0) Real (KDa)	64.0 (0.95)	62.0 (0.92)	47.7 (1.42)	43.3 (1.28)	65.3 (1.94)
Mr from Porod volume (Vporod/1.7) (KDa)	79.5 (1.18)	75.1 (1.12)	40.7 (1.21)	42.3 (1.26)	66.2 (1.97)
Mr from SAXSMoW (KDa)	70.5 (1.05)	68.0 (1.01)	33.7 (1.01)	34.8 (1.04)	44.7 (1.33)
Shape model fitting results					
DAMMIF (default parameters, 23 calculations)					
q range for fitting (Å^{-1})	0.0026-0.233	0.0026-0.233	0.0026-0.233	0.0026-0.239	0.0026-0.233
Symmetry, anisotropy assumptions	P1, none	P1, none	P1, none	P1, none	P1, none
NSD (standard deviation)	0.766 (0.021)	0.612 (0.039)	0.548 (0.019)	0.597 (0.018)	0.621 (0.043)
DAMMIN (default parameters, slow run)					
q range for fitting (Å^{-1})	0.0026-0.233	0.0026-0.233	0.0026-0.233	0.0026-0.233	0.0026-0.233
Symmetry, anisotropy assumptions	P1, none	P1, none	P1, none	P1, none	P1, none
χ^2	1.937	2.158	1.504	1.320	1.252
Atomistic Modelling					
Model (PDB code)	**6Z32**	**6Z32**	**6Z30**	**6Z30**	**6Z30**
q range for all modelling	0.0026-0.34	0.0026-0.34	0.0026-0.34	0.0026-0.34	0.0026-0.34
FoXS:					
χ^2	3.56	9.90	2.13	1.35	3.08
Predicted R_g (Å)	31.21	31.21	23.06	23.06	23.54
MultiFoXS multistate models:					
(10,000 iterations)					
Flexible residues	1358-1364	1358-1364	1358-1364	1358-1364	1358-1364
No of states	1	1	1	1	1
χ^2	1.68	1.75	1.31	1.15	1.13
c_1, c_2	1.03, 0.72	1.02, 1.53	0.99, 2.92	0.99, 2.74	1.01, 4.00
R_g values of each state (Å)	31.32	31.30	24.15	24.01	26.74

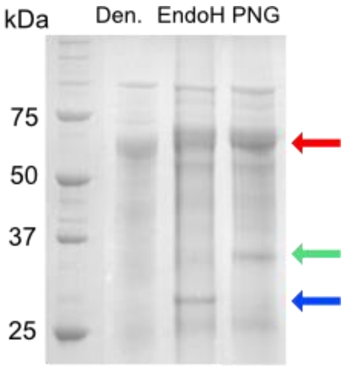
Appendix Table 6: Data collection parameters and statistics obtained from SEC-SAXS experiments of D9-10 at pH7.5 in the absence of M6P, presence of 100-fold excess M6P, de-glycosylated in the absence of M6P and presence of M6P and glycosylated D9-10 at pH 5.5. For glycosylated samples D9-10 from the D7-11 crystal structure, which was singly glycosylated, was used as a starting model. For de-glycosylated samples D9-10 from the D9-10 crystal structure, with glycans removed, was used as the starting model. For modelling the pH 5.5 data D9-10 from the D9-10 crystal structure was used, with the second glycan at N1246 being built in. All glycan residues were also specified as flexible during multistate modelling.



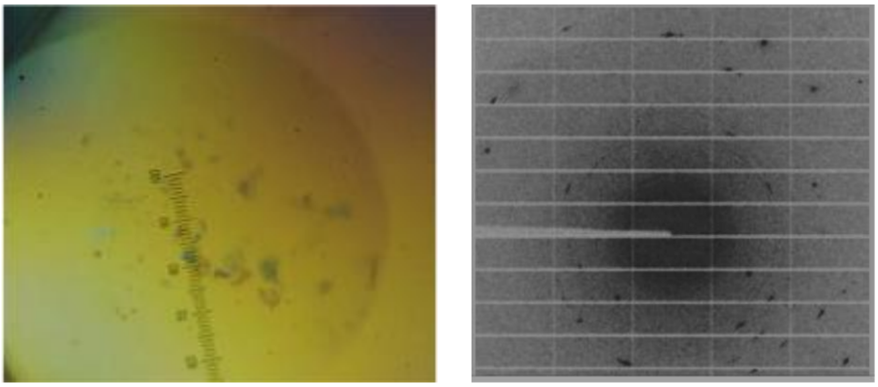
Appendix Figure 8: SEC-SAXS of D9-10. For each result: **A:** SEC trace of D9-10 showing the intensity of scattering as a ratio to the buffer background vs frame number. The Rg across the peak is highlighted in blue. **B**: I(q) vs q as log-linear plots with errors shown, the inset displays the guinier fit and residuals for data at q*R_g < 1.3. **C:** Dimensionless Kratky plot data shown in panel B. The crosshair (1.10, 1.73) is the expected peak maxima for a folded globular protein. The two peaks observed indicate that the species is a two-domain structure. The definition of the second peak is reduced when the sample is deglycosylated. **D:** Pair-distance, P(r) distribution function showing the P(r) vs r profile. The inset shows the fit of the P(r) distribution profile (red line) to the raw scattering data used in the P(r) vs r analyses. SEC-SAXS of D9-10 at pH 7.5 in the absence of M6P (black), presence of 100-fold excess of M6P (green), de-glycosylated in the absence of M6P (orange) and presence of 100-fold excess of M6P (blue). Data was collected at pH 7.5 (left) and pH 5.5 (right). D9-10 de-glycosylated pH 5.5 gave poor scattering due to low concentration. Data for D9-10 de-glycosylated and at pH 5.5 was not collected.

Sample	Mw (kDa)	Mw_{app} (kDa)	$Log_{10}Mw$	V_e (ml)	K_{av}
Blue dextran	2000	-	3.73	7.07	0.00
Beta amylase	200	225.5	2.30	12.16	0.30
ADH	150	137.1	2.18	12.96	0.35
Conalbumin	75	72.7	1.88	13.98	0.41
BSA monomer	66	57.4	1.82	14.36	0.43
Ovalbumin	44	52.6	1.64	14.50	0.44
Carbonic anhydrase	29	27.2	1.46	15.56	0.50
Cytochrome C	12	12.7	1.09	16.79	0.57
D7-10His	~66	63.4	1.82	14.20	0.57

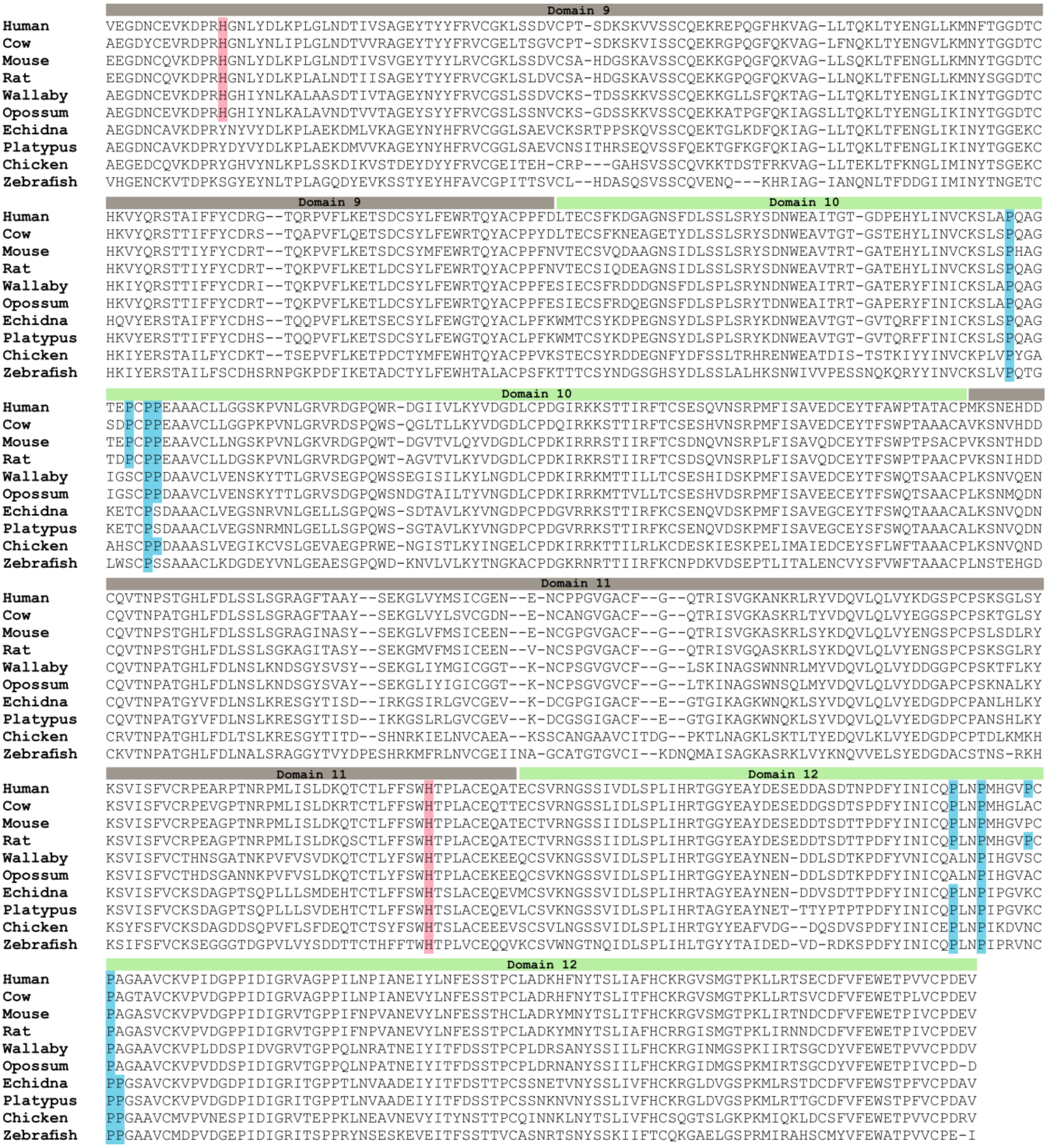
Appendix Table 7: Analytical SEC of D7-10 using a Superdex 200 10/300 column reveals D7-10 to be monomeric. A summary of results for each calibrant and glycosylated D7-10His (red). The SEC column was calibrated by Dr Ash Winter (University of Bristol).



Appendix Figure 9: D7-10 de-glycosylation attempts. Lane 1 contains denatured, fully glycosylated D7-10His (~65 kDa, red arrow). There is no shift to a lower molecular weight species following incubation with EndoH (29 kDa, blue arrow) or PNGaseF (36 kDa, green arrow).



Appendix Figure 10: Crystallography of D7-10His. Small birefringent crystals (left) were looped, cryo-cooled and analysed at Diamond Light Source. Unfortunately, their poor diffraction pattern (right) suggests salt crystals.



```
                                                        Domain 9
Human      VEGDNCEVKDPRHGNLYDLKPLGLNDTIVSAGEYTYYFRVCGKLSSDVCPT-SDKSKVVSSCQEKREPQGFHKVAG-LLTQKLTYENGLLKMNFTGGDTC
Cow        AEGDYCEVRDPRHGNLYNLIPLGLNDTVVRAGEYTYYFRVCGELTSGVCPT-SDKSKVISSCQEKRGPQGFQKVAG-LFNQKLTYENGVLKMNYTGGDTC
Mouse      EEGDNCQVKDPRHGNLYDLKPLGLNDTIVSVGEYTYYLRVCGKLSSDVCSA-HDGSKAVSSCQEKKGPQGFQKVAG-LLSQKLTFENGLLKMNYTGGDTC
Rat        EEGDNCQVKDPRHGNLYDLKPLALNDTIISAGEYTYYFRVCGKLSLDVCSA-HDGSKAVSSCQEKKGPQGFQKVAG-LLNQKLTFENGLLKMNYSGGDTC
Wallaby    AEGDNCEVKDPRHGHIYNLKALAASDTIVTAGEYNYYFRVCGSLSSDVCKS-TDSSKKVSSCQEKKGLLSFQKTAG-LLTQKLTYENGLIKINYTGGDTC
Opossum    AEGDNCEVKDPRHGHIYNLKALAVNDTVVTAGEYSYYFRVCGSLSSNVCKS-GDSSKKVSSCQEKKATPGFQKIAGSLLTQKLTYENGLIKINYTGGDTC
Echidna    AEGDNCAVKDPRYNYVYDLKPLAEKDMLVKAGEYNYHFRVCGGLSAEVCKSRTPPSKQVSSCQEKTGLKDFQKIAG-LLTQKLTFENGLIKINYTGGEKC
Platypus   AEGDNCAVKDPRYDYVYDLKPLAEKDMVVKAGEYNYHFRVCGGLSAEVCNSITHRSEQVSSFQEKTGFKGFQKIAG-LLTQKLTFENGLIKINYTGGEKC
Chicken    AEGEDCQVKDPRYGHVYNLKPLSSKDIKVSTDEYDYYFRVCGEITEH-CRP---GAHSVSSCQVKKTDSTFRKVAG-LLTEKLTFKNGLIMINYTSGEKC
Zebrafish  VHGENCKVTDPKSGYEYNLTPLAGQDYEVKSSTYEYHFAVCGPITTSVCL--HDASQSVSSCQVENQ---KHRIAG-IANQNLTFDDGIIMINYTNGETC

                      Domain 9                                       Domain 10
Human      HKVYQRSTAIFFYCDRG--TQRPVFLKETSDCSYLFEWRTQYACPPFDLTECSFKDGAGNSFDLSSLSRYSDNWEAITGT-GDPEHYLINVCKSLAPQAG
Cow        HKVYQRSTTIFFYCDRS--TQAPVFLQETSDCSYLFEWRTQYACPPYDLTECSFKNEAGETYDLSSLSRYSDNWEAVTGT-GSTEHYLINVCKSLSPQAG
Mouse      HKVYQRSTTIYFYCDRT--TQKPVFLKETSDCSYMFEWRTQYACPPFNVTECSVQDAAGNSIDLSSLSRYSDNWEAVTRT-GATEHYLINVCKSLSPHAG
Rat        HKVYQRSTTIYFYCDRT--TQKPVFLKETLDCSYLFEWRTQYACPPFNVTECSIQDEAGNSIDLSSLSRYSDNWEAVTRT-GATEHYLINVCKSLSPQAG
Wallaby    HKIYQRSTTIFFYCDRI--TQKPVFLKETLDCSYLFEWRTQYACPPFESIECSFRDDDGNSFDLSPLSRYNDNWEAITRT-GATERYFINICKSLAPQAG
Opossum    HKVYQRSTTIFFYCDRT--TQKPVFLKETLDCSYLFEWRTQYACPPFESIECSFRDQEGNSFDLSPLSRYTDNWEAITRT-GAPERYFINICKSLAPQAG
Echidna    HQVYERSTAIFFYCDHS--TQQPVFLKETSECSYLFEWGTQYACLPFKWMTCSYKDPEGNSYDLSPLSRYKDNWEAVTGT-GVTQRFFINICKSLSPQAG
Platypus   HKVYERSTTIFFYCDHS--TQQPVFLKETSDCSYLFEWGTQYACLPFKWMTCSYKDPEGNSYDLSPLSRYKDNWEAVTGT-GVTQRFFINICKSLSPQAG
Chicken    HKIYERSTAILFYCDKT--TSEPVFLKETPDCTYMFEWHTQYACPPVKSTECSYRDDEGNFYDFSSLTRHRENWEATDIS-TSTKIYYINVCKPLVPYGA
Zebrafish  HKIYERSTAILFSCDHSRNPGKPDFIKETADCTYLFEWHTALACPSFKTTTCSYNDGSGHSYDLSSLALHKSNWIVVPESSNQKQRYYINVCKSLVPQTG

                                                        Domain 10
Human      TEPCPPEAAACLLGGSKPVNLGRVRDGPQWR-DGIIVLKYVDGDLCPDGIRKKSTTIRFTCSESQVNSRPMFISAVEDCEYTFAWPTATACPMKSNEHDD
Cow        SDPCPPEAAVCLLGGPKPVNLGRVRDSPQWS-QGLTLLKYVDGDLCPDQIRKKSTTIRFTCSESHVNSRPMFISAVEDCEYTFSWPTAAACAVKSNVHDD
Mouse      TEPCPPEAAVCLLNGSKPVNLGKVRDGPQWT-DGVTVLQYVDGDLCPDKIRRRSTIIRFTCSDNQVNSRPLFISAVQDCEYTFSWPTPSACPVKSNTHDD
Rat        TDPCPPEAAVCLLDGSKPVNLGRVRDGPQWT-AGVTVLKYVDGDLCPDKIRKRSTIIRFTCSDSQVNSRPLFISAVQDCEYTFSWPTPAACPVKSNIHDD
Wallaby    IGSCPPDAAVCLVENSKYTTLGRVSEGPQWSSEGISILKYLNGDLCPDKIRRKMTTILLTCSESHIDSKPMFISAVEDCEYTFSWQTSAACPLKSNVQEN
Opossum    IGSCPPDAAVCLVENSKYTTLGRVSDGPQWSNDGTAILTYVNGDLCPDKIRRKMTTVLLTCSESHVDSRPMFISAVEECEYTFSWQTSAACPLKSNMQDN
Echidna    KETCPSDAAACLVEGSNRVNLGELLSGPQWS-SDTAVLKYVNGDPCPDGVRRKSTTIRFKCSENQVDSKPMFISAVEGCEYTFSWQTAAACALKSNVQDN
Platypus   KETCPSDAAACLVEGSNRMNLGELLSGPQWS-SGTAVLKYVNGDPCPDGVRRKSTTIRFKCSENQVDSKPMFISAVEGCEYSFSWQTAAACALKSNVQDN
Chicken    AHSCPPDAAASLVEGIKCVSLGEVAEGPRWE-NGISTLKYINGELCPDKIRRKTTILRLKCDESKIESKPELIMAIEDCEYSFLWFTAAACPLKSNVQND
Zebrafish  LWSCPSSAAACLKDGDEYVNLGEAESGPQWD-KNVLVLKYTNGKACPDGKRNRTTIIRFKCNPDKVDSEPTLITALENCVYSFVWFTAAACPLNSTEHGD

                                                        Domain 11
Human      CQVTNPSTGHLFDLSSLSGRAGFTAAY--SEKGLVYMSICGEN--E-NCPPGVGACF--G--QTRISVGKANKRLRYVDQVLQLVYKDGSPCPSKSGLSY
Cow        CQVTNPATGHLFDLSSLSGRAGFTAAY--SEKGLVYLSVCGDN--E-NCANGVGACF--G--QTRISVGKASKRLTYVDQVLQLVYEGGSPCPSKTGLSY
Mouse      CQVTNPSTGHLFDLSSLSGRAGINASY--SEKGLVFMSICEEN--E-NCGPGVGACF--G--QTRISVGKASKRLSYKDQVLQLVYENGSPCPSLSDLRY
Rat        CQVTNPSTGHLFDLSSLSGKAGITASY--SEKGMVFMSICEEN--V-NCSPGVGACF--G--QTRISVGQASKRLSYKDQVLQLVYENGSPCPSKSGLRY
Wallaby    CQVTNPATGHLFDLNSLKNDSGYSVSY--SEKGLIYMGICGGT--K-NCPSGVGVCF--G--LSKINAGSWNNRLMYVDQVLQLVYDDGGPCPSKTFLKY
Opossum    CQVTNPATGHLFDLNSLKNDSGYSVAY--SEKGLIYIGICGGT--K-NCPSGVGVCF--G--LTKINAGSWNSQLMYVDQVLQLVYDDGAPCPSKNALKY
Echidna    CQVTNPATGYVFDLNSLKRESGYTISD--IRKGSIRLGVCGEV--K-DCGPGIGACF--E--GTGIKAGKWNQKLSYVDQVLQLVYEDGDPCPANLHLKY
Platypus   CQVTNPATGYVFDLNSLKRESGYTISD--IKKGSLRLGVCGEV--K-DCGSGIGACF--E--GTGIKAGKWNQKLSYVDQVLQLVYEDGDPCPANSHLKY
Chicken    CRVTNPATGHLFDLTSLKRESGYTITD--SHNRKIELNVCAEA--KSSCANGAAVCITDG--PKTLNAGKLSKTLTYEDQVLKLVYEDGDPCPTDLKMKH
Zebrafish  CKVTNPATGHLFDLNALSRAGGYTVYDPESHRKMFRLNVCGEIINA-GCATGTGVCI--KDNQMAISAGKASRKLVYKNQVVELSYEDGDACSTNS-RKH

                       Domain 11                                           Domain 12
Human      KSVISFVCRPEARPTNRPMLISLDKQTCTLFFSWHTPLACEQATECSVRNGSSIVDLSPLIHRTGGYEAYDESEDDASDTNPDFYINICQPLNPMHGVPC
Cow        KSVISFVCRPEVGPTNRPMLISLDKRTCTLFFSWHTPLACEQTTECSVRNGSSLIDLSPLIHRTGGYEAYDESEDDGSDTSPDFYINICQPLNPMHGLAC
Mouse      KSVISFVCRPEAGPTNRPMLISLDKQTCTLFFSWHTPLACEQATECTVRNGSSIIDLSPLIHRTGGYEAYDESEDDTSDTTPDFYINICQPLNPMHGVPC
Rat        KSVISFVCRPEAGPTNRPMLISLDKQSCTLFFSWHTPLACEQATECTVRNGSSIIDLSPLIHRTGGYEAYDESEDDTSDTTPDFYINICQPLNPMHGVPC
Wallaby    KSVISFVCTHNSGATNKPVFVSVDKQTCTLYFSWHTPLACEKEEQCSVKNGSSVIDLSPLIHRTGGYEAYNEN-DDLSDTKPDFYVNICQALNPIHGVSC
Opossum    KSVISFVCTHDSGANNKPVFVSLDKQTCTLYFSWHTPLACEKEEQCSVKNGSSVIDLSPLIHRTGGYEAYNEN-DDLSDTKPDFYINICQALNPIHGVAC
Echidna    KSVISFVCKSDAGPTSQPLLLSMDEHTCTLFFSWHTSLACEQEVMCSVKNGSSVIDLSPLIHRTAGYEAYNEN-DDVSDTTPDFYINICQPLNPIPGVKC
Platypus   KSVISFVCKSDAGPTSQPLLLSVDEHTCTLFFSWHTSLACEQEVLCSVKNGSSVIDLSPLIHRTAGYEAYNET-TTYPTPTPDFYINICQPLNPIPGVKC
Chicken    KSYFSFVCKSDAGDDSQPVFLSFDEQTCTSYFSWHTSLACEEEVSCSVLNGSSVIDLSPLIHRTGYYEAFVDG--DQSDVSPDFYINICEPLNPIKDVNC
Zebrafish  KSIFSFVCKSEGGGTDGPVLVYSDDTTCTHFFTWHTPLVCEQQVKCSVWNGTNQIDLSPLIHLTGYYTAIDED-VD-RDKSPDFYINICQPLNPIPRVNC

                                                        Domain 12
Human      PAGAAVCKVPIDGPPIDIGRVAGPPILNPIANEIYLNFESSTPCLADKHFNYTSLIAFHCKRGVSMGTPKLLRTSECDFVFEWETPVVCPDEV
Cow        PAGTAVCKVPVDGPPIDIGRVAGPPILNPIANEVYLNFESSTPCLADRHFNYTSLITFHCKRGVSMGTPKLLRTSVCDFVFEWETPLVCPDEV
Mouse      PAGASVCKVPVDGPPIDIGRVTGPPIFNPVANEVYLNFESSTHCLADRYMNYTSLITFHCKRGVSMGTPKLIRTNDCDFVFEWETPIVCPDEV
Rat        PAGASVCKVPVDGPPIDIGRVTGPPIFNPVANEVYLNFESSTPCLADKYMNYTSLIAFHCRRGISMGTPKLIRNNDCDFVFEWETPIVCPDEV
Wallaby    PAGAAVCKVPLDDSPIDVGRVTGPPQLNRATNEIYITFDSSTPCPLDRSANYSSIILFHCKRGINMGSPKIIRTSGCDYVFEWETPVVCPDDV
Opossum    PAGAAVCKVPVDDSPIDIGRVTGPPQLNPATNEIYITFDSSTPCPLDRNANYSSIILFHCKRGIDMGSPKMIRTSGCDYVFEWETPIVCPD-D
Echidna    PPGSAVCKVPVDGDPIDIGRITGPPTLNVAADEIYITFDSTTPCSSNETVNYSSLIVFHCKRGLDVGSPKMLRSTDCDFVFEWSTPFVCPDAV
Platypus   PPGSAVCKVPVDGDPIDIGRITGPPTLNVAADEIYITFDSSTPCSSNKNVNYSSLIVFHCKRGLDVGSPKMLRTTGCDFVFEWSTPFVCPDAV
Chicken    PPGAAVCMVPVNESPIDIGRVTEPPKLNEAVNEVYITYNSTTPCQINNKLNYTSLIVFHCSQGTSLGKPKMIQKLDCSFVFEWETPVVCPDRV
Zebrafish  PPGAAVCMDPVDGEPIDIGRITSPPRYNSESKEVEITFSSTTVCASNRTSNYSSKIIFTCQKGAELGSPRMIRAHSCMYVFEWATPVVCPE-I
```

Appendix Figure 11: Sequence alignment of CI-MPR domains 9-12 from placental mammals (human, cow, mouse, rat), marsupial mammals (wallaby), monotreme mammals (opossum, echidna), platypus, birds (Chicken) and fish (zebrafish). Histidine residues in the His-Pro pocket are coloured red, while proline residues coloured blue.

8.5. Chapter 4: The full extracellular region of CI-MPR

Sample	Mw (kDa)	Mw$_{app}$ (kDa)	Log$_{10}$Mw	V$_{e}$ (ml)	K$_{av}$
Blue dextran	2000	-	3.73	7.07	0.00
Beta amylase	200	225.5	2.35	12.16	0.30
ADH	150	137.1	2.14	12.96	0.35
BSA monomer	66	57.4	1.76	13.76	0.40
Conalbumin	75	72.7	1.86	13.98	0.41
Ovalbumin	44	52.6	1.72	14.50	0.44
Carbonic anhydrase	29	27.2	1.43	15.46	0.50
Cytochrome C	12	12.7	1.10	16.79	0.57
D1-15	~250	506.2	2.70	10.86	0.22

Appendix Table 8: Analytical SEC of D1-15 using a Superdex 200 10/300 column reveals D1-15 to be dimeric. A summary of results for each calibrant and glycosylated D1-15 (red).

8.6. Chapter 5: Engineering a synthetic lectin

Spectra:	AB3	AB3 Man	AB3 M6P	AB3	AB3
pH:	6.5	6.5	6.5	6	7
Residues:	M1508	M1508	M1508	M1508	M1508
	K1509	K1509	K1509	K1509	K1509
	S1510	S1510	S1510	S1510	S1510
	P1521	C1516	D1514	P1521	H1513
	P1560	P1521	C1516	P1560	D1515
	P1561	G1537	P1521	P1561	N1520
	P1597	Y1542	A1536	P1597	P1521
	P1599	G1554	G1537	P1599	S1522
	P1616	P1560	Y1542	P1616	R1535
	P1620	P1561	V1548	P1620	**S1545**
	P1624	G1595	G1554	P1624	G1554
	P1643	S1596	P1560	P1643	P1560
	P1648	C1598	P1561	P1648	P1561
		P1597	P1597		Q1569
		P1599	P1599		P1597
		K1601	K1601		S1602
		R1615	S1605		G1603
		P1616	P1616		P1616
		P1620	P1620		P1620
		P1624	P1624		T1621
		M1625	S1628		P1624
		L1626	T1633		P1643
		T1633	H1641		P1648
		P1643	P1643		
		P1648	L1644		
			P1648		
Percentage unassigned (%):	0	9	9	0	8

Appendix Table 9: Residues of AB3 unassigned by ^{1}H-^{15}N HSQC. Unless otherwise stated 10 mM sugar was added to ~100 μM protein. The percentage of unassigned residues does not include proline residues (highlighted orange) and the three N-terminal residues M1508-S1510 (highlighted blue) that are not visible in any of the HSQC experiments. Residues that are mutated from D11WT are coloured red.

Spectra:	**ΔPK**	**ΔPK Glc**	**ΔPK M6P**
pH:	8.5	8.5	8.5
Residues:	M1508	M1508	M1508
	K1509	K1509	K1509
	S1510	S1510	S1510
	D1515	D1515	D1515
	D1516	D1516	D1516
	V1518	V1518	V1518
	P1521	P1521	P1521
	L1532	L1532	L1526
	R1535	R1535	L1532
	A1536	A1536	R1535
	G1537	G1537	A1536
	Y1542	Y1542	G1537
	S1543	S1543	Y1542
	V1548	V1548	S1543
	G1554	G1554	**S1545**
	N1556	N1556	V1548
	P1560	P1560	G1554
	P1561	P1561	N1556
	G1568	G1568	P1560
	T1570	T1570	P1561
	S1573	S1573	G1568
	K1576	K1576	T1570
	V1587	K1593	S1573
	K1593	S1596	K1576
	S1596	P1597	V1587
	P1597	L1602	K1593
	L1602	P1614	S1596
	P1614	A1616	P1597
	A1616	P1618	L1602
	P1618	P1622	P1614
	P1622	L1624	A1616
	L1624	I1625	P1618
	I1625	S1626	P1622
	S1626	**E1627**	L1624
	E1627	D1628	I1625
	D1628	**E1629**	S1626
	E1629	T1631	**E1627**
	T1631	**Y1634**	D1628
	Y1634	P1641	**E1629**
	P1641	P1646	T1631
	P1646		**Y1634**
			P1641
			P1646
Percentage unassigned (%):	21	20	22

Appendix Table 10: Residues of ΔPK unassigned by ^{1}H-^{15}N HSQC. Unless otherwise stated 10 mM sugar was added to ~100 µM protein. The percentage of unassigned residues does not include proline residues (highlighted orange) and the three N-terminal residues M1508-S1510 (highlighted blue) that are not visible in any of the HSQC experiments. Residues that are mutated from D11WT are coloured red.

Spectra:	ΔPK1G	ΔPK1G Man	ΔPK1G Glc	ΔPK1G 10 mM M6P	ΔPK1G 50 mM M6P
pH:	**8.5**	**8.5**	**8.5**	**8.5**	**8.5**
Residues:	M1508	M1508	M1508	M1508	M1508
	K1509	K1509	K1509	K1509	K1509
	S1510	S1510	S1510	S1510	S1510
	E1512	E1512	E1512	E1512	E1512
	P1521	P1521	D1515	D1515	H1513
	R1535	S1522	P1521	P1521	D1515
	A1536	H1525	S1522	H1525	C1516
	A1541	R1535	F1527	R1535	T1519
	Y1542	A1536	R1535	A1536	D1520
	G1546	A1541	A1536	A1541	P1521
	P1560	Y1542	A1541	Y1542	R1535
	P1561	G1546	Y1542	G1546	A1536
	Q1567	P1560	**S1545**	Y1549	A1541
	G1568	P1561	G1546	P1560	Y1542
	Q1569	**Q1567**	P1560	P1561	G1546
	T1570	G1568	P1561	**Q1567**	P1560
	S1573	Q1569	**Q1567**	G1568	P1561
	Q1586	T1570	G1568	Q1569	**Q1567**
	L1588	S1573	Q1569	T1570	G1568
	P1597	V1574	T1570	S1573	Q1569
	G1599	Q1586	S1573	V1574	T1570
	S1600	L1588	V1574	Q1586	S1573
	G1601	S1596	Q1586	L1588	V1574
	S1603	P1597	L1588	P1597	Q1586
	F1610	**G1599**	S1596	**G1599**	L1588
	P1614	S1600	P1597	S1600	P1597
	A1616	G1601	**G1599**	G1601	**G1599**
	G1617	S1603	S1600	S1603	S1600
	P1618	F1610	G1601	F1610	G1601
	E1619	P1614	S1603	P1614	S1603
	P1622	A1616	F1610	A1616	F1610
	T1623	G1617	P1614	G1617	P1614
	E1627	P1618	A1616	P1618	A1616
	Y1634	E1619	G1617	E1619	G1617
	P1641	N1620	P1618	P1622	P1618
	E1645	G1621	E1619	T1623	E1619
	P1646	P1622	N1620	**E1627**	T1623
		T1623	P1622	**Y1634**	**E1627**
		E1627	T1623	P1641	D1632
		Y1634	**E1627**	E1645	**Y1634**
		P1641	P1641	P1646	P1641
		E1645	**Y1634**		P1646
		P1646	E1645		
			P1646		
Percentage unassigned (%):	18	23	24	21	23

Appendix Table 11: Residues of D11 ΔPK1G unassigned by ^{1}H-^{15}N HSQC. Unless otherwise stated 10 mM sugar was added to ~100 μM protein. The percentage of unassigned residues does not include proline residues (highlighted orange). Residues that are mutated from D11WT are coloured red.

Spectra:	**ΔPK1H**	**ΔPK1H M6P**	**ΔPK1H**	**ΔPK1H Man**	**ΔPK1H M6P**
pH:	**8.5**	**8.5**	**6.5**	**6.5**	**6.5**
Residues:	M1508	M1508	M1508	M1508	M1508
	K1509	K1509	K1509	K1509	K1509
	S1510	S1510	S1510	S1510	S1510
	P1521	N1511	P1521	C1516	T1519
	G1546	P1521	G1554	P1521	N1520
	V1547	**V1547**	P1560	G1554	P1521
	P1650	P1560	P1561	P1560	F1527
	P1651	P1561	Q1569	P1561	**V1547**
	F1567	**F1567**	S1596	Q1569	P1560
	G1568	K1576	P1597	S1596	P1561
	T1570	D1585	C1598	P1597	Q1569
	K1576	P1597	**H1599**	C1598	S1596
	D1585	**H1599**	G1601	**H1599**	P1597
	P1597	S1600	V1607	G1601	C1598
	H1599	G1601	P1614	S1603	**H1599**
	S1600	S1603	A1616	V1607	G1601
	G1601	**R1604**	P1618	P1614	S1603
	S1603	P1614	P1622	A1616	P1614
	R1604	P1618	N1620	N1620	A1616
	P1614	E1619	I1625	P1618	P1618
	A1616	N1620	**E1627**	P1622	P1622
	G1617	P1622	**Y1634**	I1625	I1625
	P1618	S1626	F1635	**E1627**	**E1627**
	E1619	**E1627**	P1641	**Y1634**	T1633
	N1620	K1629	P1646	F1635	**Y1634**
	P1622	T1633		P1641	F1635
	S1626	**Y1634**		P1646	P1641
	E1627	P1641			P1646
	K1629	E1645			
	T1633	P1646			
	Y1634				
	P1641				
	E1645				
	P1646				
Percentage unassigned (%):	16	13	10	11	12

Appendix Table 12: Residues of D11 ΔPK1H unassigned by ^{1}H-^{15}N HSQC. Unless otherwise stated 10 mM sugar was added to ~100 μM protein. The percentage of unassigned residues does not include proline residues (highlighted orange). Residues that are mutated from D11WT are coloured red.

Spectra:	**ΔPK2H**	**ΔPK2H Man**	**ΔPK2H M6P**
pH:	**6.5**	**6.5**	**6.5**
Residues:	M1508	M1508	M1508
	K1509	K1509	K1509
	S1510	S1510	S1510
	D1515	D1515	D1515
	C1516	P1521	C1516
	P1521	R1535	P1521
	R1535	A1536	R1535
	A1536	Y1542	A1536
	Y1542	G1554	Y1542
	G1554	P1560	G1554
	P1560	P1561	P1560
	P1561	C1566	P1561
	C1566	**Q1567**	C1566
	Q1567	Q1569	**Q1567**
	T1570	T1570	E1569
	P1597	S1596	T1570
	H1600	P1597	S1596
	P1614	**H1600**	P1597
	P1618	G1601	**H1600**
	P1622	P1614	P1614
	I1625	P1618	P1618
	E1627	P1622	P1622
	Y1634	I1625	I1625
	F1635	**E1627**	**E1629**
	P1641	**Y1634**	**E1627**
	P1646	F1635	**Y1634**
		P1641	F1635
		P1646	P1641
			P1646
Percentage unassigned (%):	11	12	13

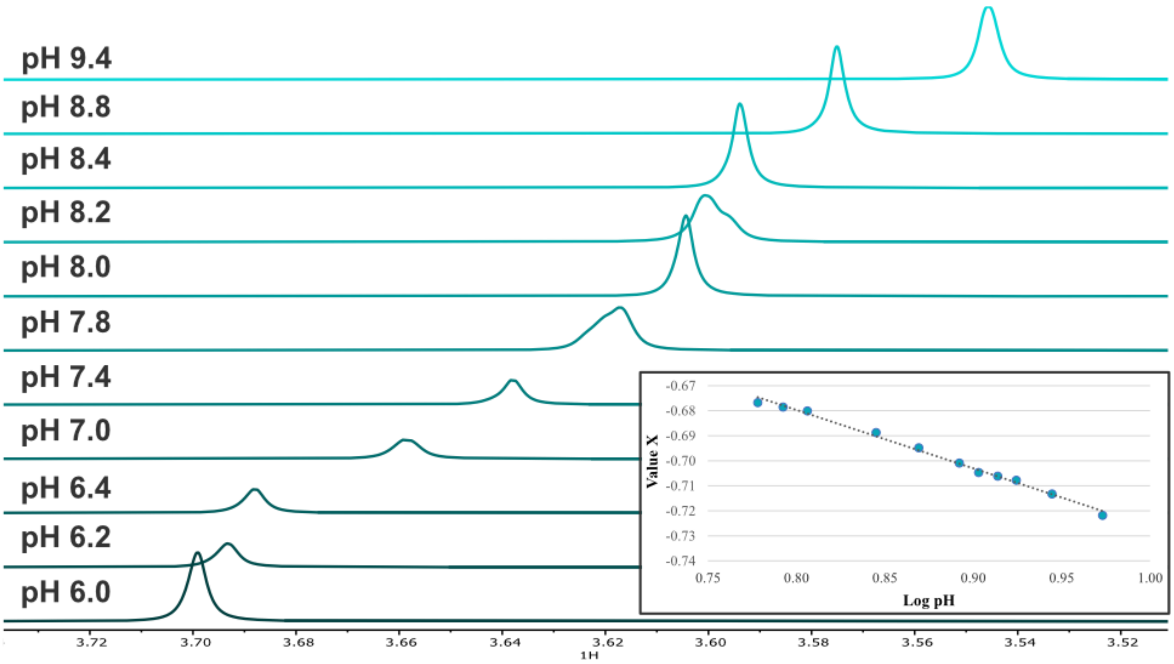
Appendix Table 13: Residues of D11 ΔPK1G unassigned by ^{1}H-^{15}N HSQC. Unless otherwise stated 10 mM sugar was added to ~100 μM protein. The percentage of unassigned residues does not include proline residues (highlighted orange). Residues that are mutated from D11WT are coloured red.

Spectra:	AB3	AB3	AB3	ΔPK	ΔPK
	Man	M6P	pH	Glc	M6P
pH:	6.5	6.5	6-7	8.5	8.5
Residues	N1520	E1617	L1526	F1635	H1525
(CSP,	(0.26)	(0.33)	(2.61)	(0.37)	(0.27)
ppm):	K1579	**F1567**	H1525	Q1517	F1635
	(0.17)	(0.31)	(1.44)	(0.32)	(0.23)
	D1630	E1647	E1647	K1605	E1645
	(0.16)	(0.30)	(1.42)	(0.23)	(0.23)
	D1514	N1520	C1516	T1519	T1519
	(0.07)	(0.30)	(1.32)	(0.22)	(0.22)
	K1544	T1519	S1531	R1580	C1632
	(0.07)	(0.29)	(0.98)	(0.20)	(0.16)
	I1572	H1525	M1625	L1581	E1557
	(0.06)	(0.28)	(0.83)	(0.19)	(0.13)
	V1518	L1526	Q1586	**K1544**	S1608
	(0.06)	(0.28)	(0.74)	(0.18)	(0.11)
	N1556	H1513	C1553	L1588	R1580
	(0.06)	(0.27)	(0.64)	(0.17)	(0.10)
	G1603	V1613	L1581	L1529	Q1589
	(0.05)	(0.21)	(0.58)	(0.17)	(0.09)
	E1647	A1618	S1543	M1623	S1605
	(0.05)	(0.18)	(0.56)	(0.17)	(0.09)

Appendix Table 14: Residues of D11 AB3 and ΔPK that gave the 10 largest CSPs upon addition of 100-fold excess mannose, M6P, glucose or change in pH (pH 6-7). Mutated residues are coloured red**.**

Spectra:	ΔPK1G	ΔPK1G	ΔPK1G	ΔPK1G	ΔPK1H	ΔPK1H	ΔPK1H	ΔPK2H	ΔPK2H
	Man	Glc	10 mM	50 mM	50 mM	Man	M6P	Man	M6P
			M6P	M6P	M6P				
pH:	8.5	8.5	8.5	8.5	8.5	6.5	6.5	6.5	6.5
Residues	R1613	Q1630	N1511	L1526	L1526	L1526	L1526	L1526	L1526
(CSP,	(0.15)	(0.14)	(1.06)	(3.32)	(3.67)	(0.44)	(0.91)	(0.88)	(1.19)
ppm):	V1548	H1525	T1633	V1587	R1535	**K1544**	V1611	H1525	H1525
	(0.15)	(0.14)	(1.03)	(1.36)	(1.28)	(0.37)	(0.55)	(0.55)	(0.77)
	D1515	L1624	S1522	V1611	C1516	S1626	G1568	E1645	E1645
	(0.12)	(0.09)	(0.81)	(1.23)	(0.91)	(0.26)	(0.49)	(0.51)	(0.65)
	C1516	C1516	N1620	S1596	T1631	D1515	E1645	V1607	T1519
	(0.12)	(0.09)	(0.55)	(0.78)	(0.74)	(0.24)	(0.44)	(0.46)	(0.55)
	S1609	R1613	L1526	F1527	W1638	R1571	**K1544**	L1581	V1611
	(0.10)	(0.08)	(0.54)	(0.75)	(0.57)	(0.22)	(0.42)	(0.39)	(0.53)
	L1624	V1548	N1578	C1553	L1588	S1551	S1626	V1584	R1580
	(0.09)	(0.07)	(0.40)	(0.59)	(0.57)	(0.20)	(0.29)	(0.37)	(0.49)
	D1585	V1584	N1520	**S1545**	Q1586	Y1583	H1525	T1519	**K1544**
	(0.09)	(0.07)	(0.35)	(0.55)	(0.51)	(0.19)	(0.25)	(0.29)	(0.45)
	K1579	T1569	V1587	Q1630	Q1589	M1550	D1515	R1580	K1605
	(0.08)	(0.06)	(0.33)	(0.47)	(0.46)	(0.19)	(0.23)	(0.29)	(0.45)
	R1571	N1511	F1635	N1511	**S1545**	V1584	V1563	C1553	W1638
	(0.08)	(0.06)	(0.26)	(0.40)	(0.44)	(0.19)	(0.211)	(0.27)	(0.41)
	Q1630	V1611	C1516	H1639	F1610	V1563	T1640	R1613	V1584
	(0.07)	(0.06)	(0.25)	(0.39)	(0.41)	(0.19)	(0.17)	(0.27)	(0.40)

Appendix Table 15: Residues of D11 FG loop mutants (ΔPK1G, ΔPK1H and ΔPK2H) that gave the 10 largest CSPs upon addition of mannose, glucose or M6P. Mutated residues are coloured red**.**

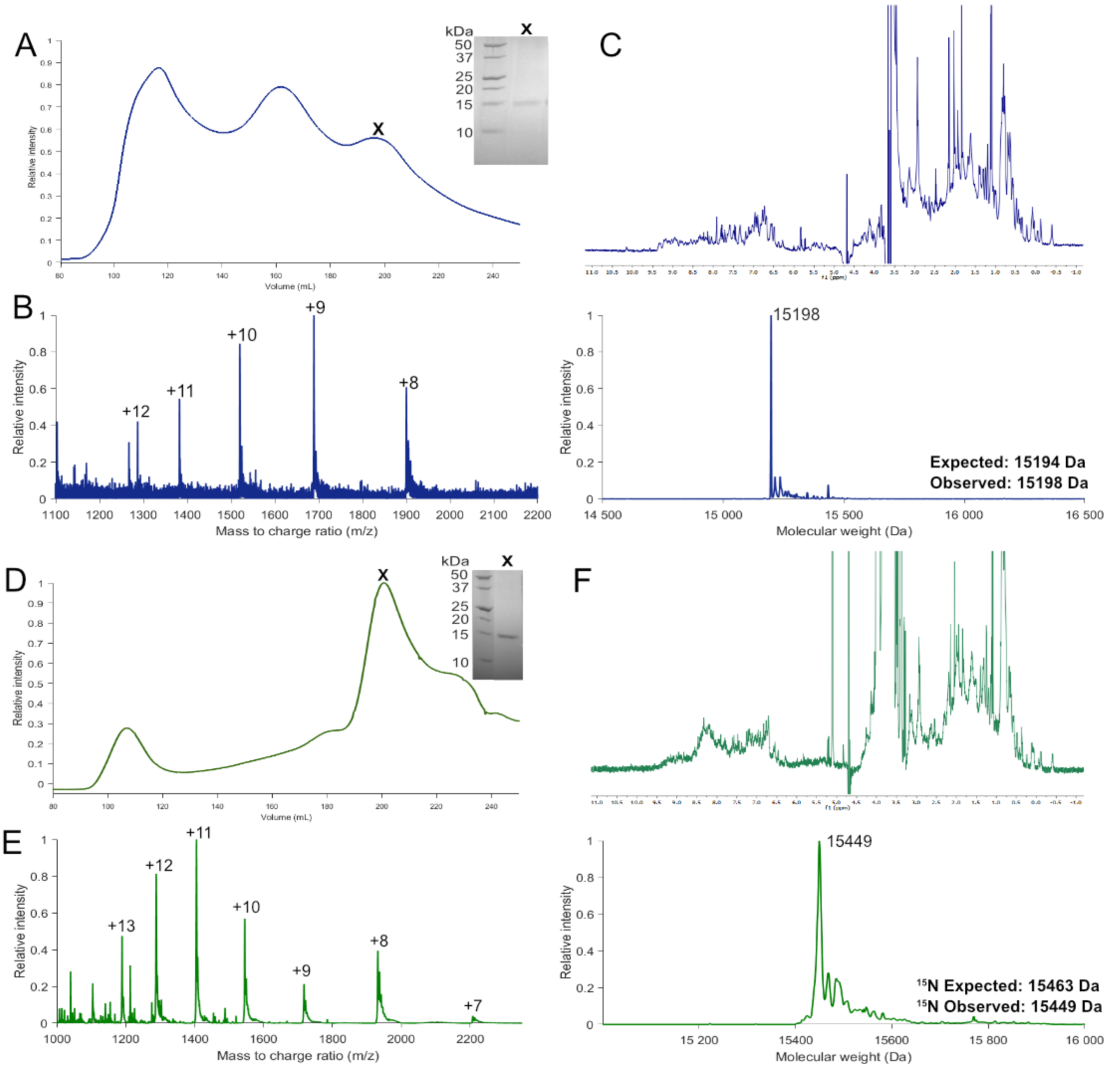


pH	Tris δ (ppm)	Log pH	Value X	Calc log pH	Calc pH	Difference
9.4	3.55	0.97	-0.72	0.98	9.56	-0.16
8.8	3.58	0.94	-0.71	0.94	8.79	0.01
8.4	3.59	0.92	-0.71	0.92	8.32	0.08
8.2	3.60	0.91	-0.71	0.91	8.19	0.01
8.0	3.60	0.90	-0.70	0.91	8.07	-0.07
7.8	3.62	0.89	-0.70	0.89	7.78	0.02
7.4	3.64	0.87	-0.69	0.86	7.32	0.08
7.0	3.66	0.85	-0.69	0.84	6.89	0.11
6.4	3.69	0.97	-0.72	0.98	9.56	-0.16
6.2	3.69	0.79	-0.68	0.79	6.24	-0.04
6.0	3.70	0.78	-0.68	0.79	6.13	-0.13

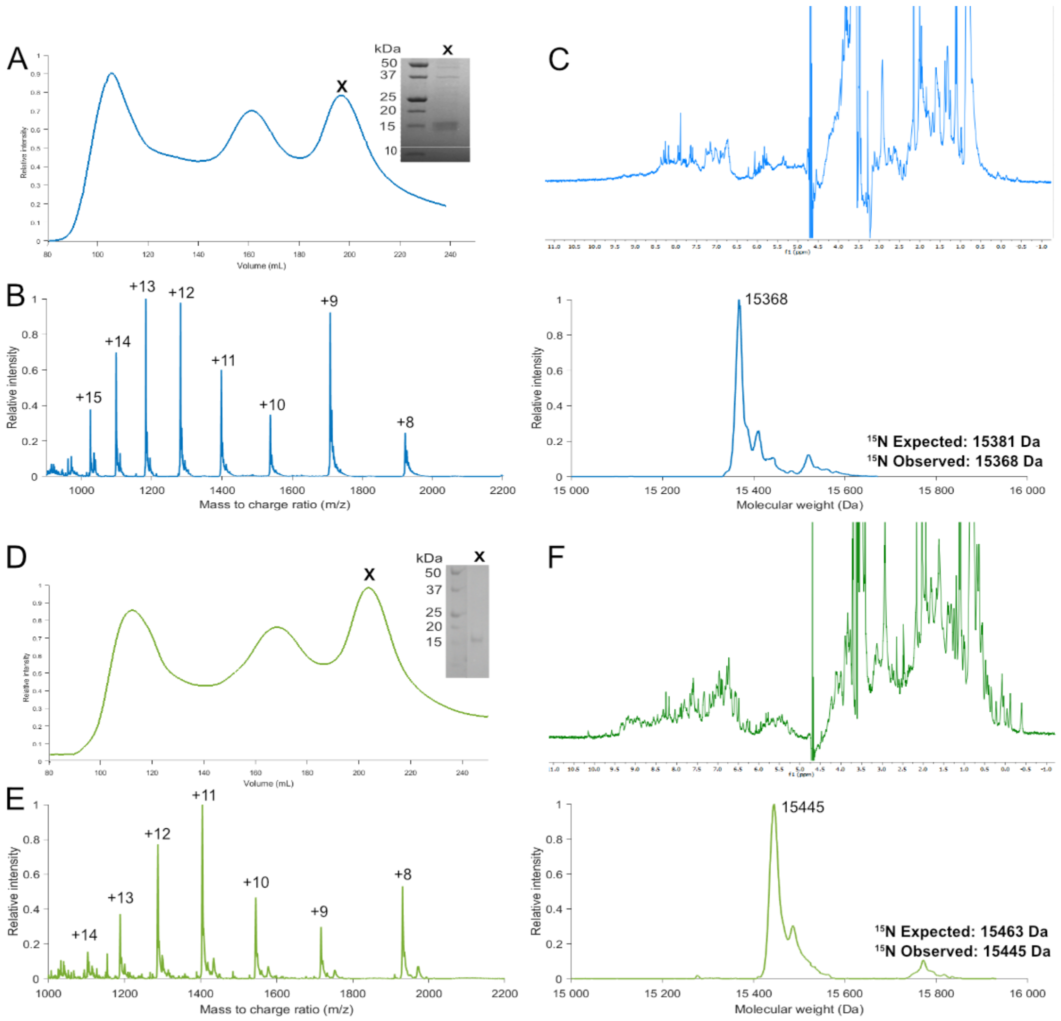
Appendix Figure 12: Generation of a pH calibration curve using the chemical shifts of Tris. Top: 1D 1H-NMR spectra of Tris at pH 6.0-9.4. Inset shows the calibration curve. Bottom: Table of values used to calculate pH. Value X was calculated using the equation: (δ-lowest pH)/ (highest pH-lowest pH) whereby the highest pH measured was 9.7 and the lowest pH 5.0. The difference between expected and calculated pH values are colour coded such that differences < 0.1 are green, >0.1 red.

Sample	Expected	Tris δ	Log	Value	Calc log	Calc	Difference	pH
	pH	(ppm)	pH	X	pH	pH		change
AB3	6.5	3.67	0.81	-0.69	0.82	6.67	0.17	NA
AB3 Man	6.5	3.69	0.81	-0.68	0.80	6.29	-0.21	-0.38
AB3 10 mM M6P	6.5	3.70	0.81	-0.68	0.79	6.11	-0.39	-0.56
AB3 20 mM M6P	6.5	3.70	0.81	-0.68	0.79	6.11	-0.39	-0.56
ΔPK	8.5	3.54	0.93	-0.72	0.99	9.73	+1.23	NA
ΔPK 10 mM Glc	8.5	3.54	0.93	-0.72	0.99	9.73	+1.23	0.00
ΔPK 10 mM M6P	8.5	3.56	0.93	-0.72	0.96	9.18	+0.68	-0.55
ΔPK1G	8.5	3.60	0.93	-0.71	0.91	8.17	-0.33	NA
ΔPK1G 10 mM Man	8.5	3.60	0.93	-0.71	0.91	8.17	-0.33	0.00
ΔPK1G 10 mM Glc	8.5	3.60	0.93	-0.71	0.91	8.15	-0.35	-0.02
ΔPK1G 10 mM M6P	8.5	3.63	0.93	-0.70	0.88	7.60	-0.90	-0.57
ΔPK1G 50 mM M6P	8.5	3.63	0.93	-0.70	0.88	7.60	-0.90	-0.57
ΔPK1H	8.5	3.56	0.93	-0.72	0.97	9.29	+0.79	NA
ΔPK1H 50 mM M6P	8.5	3.58	0.93	-0.71	0.94	8.74	+0.24	-0.56
ΔPK1H	6.7	3.66	0.81	-0.69	0.84	6.86	+0.17	NA
ΔPK1H 10 mM Man	6.7	3.66	0.81	-0.69	0.84	6.86	+0.17	0.00
ΔPK1H 10 mM M6P	6.7	3.69	0.81	-0.69	0.80	6.29	-0.41	-0.57
ΔPK2H	6.5	3.67	0.81	-0.69	0.82	6.67	+0.17	NA
ΔPK2H 10 mM Man	6.5	3.69	0.81	-0.68	0.80	6.29	-0.21	-0.38
ΔPK2H 10 mM M6P	6.5	3.70	0.81	-0.68	0.79	6.11	-0.39	-0.56

Appendix Table 16: Calculated pH values of AB3 NMR samples. The difference column is the difference between expected and calculated pH values and is colour coded such that differences < 0.5 are green, >0.5 orange and >1.0 red. The pH change column is the pH change upon addition of sugar and is coloured such that changes <0.5 are coloured blue and changes >0.5 are coloured purple.



Appendix Figure 13: Purification and characterisation of D11 FG loop mutants ΔPK1G and ΔPK1H. A: SEC of D11 ΔPK1G *in vitro* refold. The peak marked with a cross corresponds to folded, purified D11. **B:** ESI-MS of D11 ΔPK1G. The expected molecular weight, assuming cysteines reduced is 15,194 Da, the observed molecular weight is 15,198 Da. **C:** 1D ^{1}H-NMR of D11 ΔPK1G confirms the protein is folded. **D:** SEC of D11 ΔPK1H *in vitro* refold. The peak marked with a cross corresponds to folded, purified D11. **E:** ESI-MS of D11 ΔPK1H. The expected molecular weight, assuming cysteines reduced and 100 % 15N incorporation is 15,463 Da, observed molecular weight 15,449 Da. This corresponds to 93 % ^{15}N incorporation. **F:** 1D ^{1}H-NMR of D11 ΔPK1H confirms protein folding.



Appendix Figure 14: Purification and characterisation of D11 FG loop mutants ΔPK2G and ΔPK2H. A: SEC of D11 ΔPK2G in vitro refold. The peak marked with a cross corresponds to folded, purified D11. **B:** ESI-MS of D11 ΔPK2G. The expected molecular weight, assuming cysteines reduced and 100 % 15N incorporation is 15,381 Da, the observed molecular weight 15,368 Da. This corresponds to 93 % 15N incorporation. **C:** 1D ^{1}H-NMR of D11 ΔPK2G confirms the protein is folded. **D:** SEC of D11 ΔPK2H *in vitro* refold. The peak marked with a cross corresponds to folded, purified D11. **E:** ESI-MS of D11 ΔPK2H. The expected molecular weight, assuming cysteines reduced and 100 % ^{15}N incorporation is 15,463 Da, observed molecular weight 15,445 Da. This corresponds to 90 % ^{15}N incorporation. **F:** 1D ^{1}H-NMR of D11 ΔPK2H confirms protein folding.

Construct	**ΔPK**	**ΔPK1G**	**ΔPK1H**
Data collection			
Space group	C121	C121	$P12_11$
Unit cell			
a, b, c (Å)	150.2, 49.2, 214.9	144.4, 48.8, 214.0	48.3, 82.9, 93.8
α, β, γ (°)	90, 105, 90	90, 106, 90	90, 90, 90
X-ray wavelength (Å)	0.98	0.98	0.97
Resolution range (Å)	68.21-2.37 (2.46-2.37)	46.07-2.77 (2.87-2.77)	42.37-2.53 (2.62-2.53)
Total reflections	125063 (12398)	241364 (19278)	145012 (14222)
Unique reflections	62588 (6228)	37254 (3692)	25300 (2482)
Multiplicity	2.0 (2.0)	6.5 (5.2)	5.7 (5.7)
Completeness (%)	98.4 (94.6)	99.3 (98.5)	99.4 (98.7)
R_{meas}	0.11 (1.94)	0.20 (1.44)	0.13 (0.34)
Mean I/σ (I)	10.1 (0.9)	5.9 (1.2)	9.3 (5.8)
Wilson B-factor (Å^2)	44.51	57.62	23.40
CC ½	0.998 (0.488)	0.994 (0.733)	0.998 (0.989)
Refinement			
Reflections used in refinement	61567 (5897)	37104 (3672)	25204 (2478)
R_{work} (%)	24.0 (41.0)	26.3 (45.6)	25.3 (31.6)
R_{free} (%)	27.1 (43.5)	30.6 (56.8)	28.0 (31.8)
Root mean squared deviation			
Bond lengths (Å)	0.009	0.017	0.013
Bond angles (°)	1.34	2.60	1.86
Ramachandran plot (%)			
Favoured	96.8	93.5	93.5
Allowed	2.8	4.4	4.9
Outliers	0.4	2.1	1.6
Average B-factor (Å^2)	69.03	98.15	13.45
Protein	69.39	98.26	13.04
Ligand/ glycan	44.77	56.98	23.81

Appendix Table 17: Data collection and refinement statistics for D11 mutants ΔPK, ΔPK1G and ΔPK1H crystal structures. Values in parentheses are for the outer resolution shell. The structure of ΔPK (including crystallisation, data collection, phasing and refinement) was determined by Dr Chris Williams (Crump group).

9. References

1. Moremen, K. W., Tiemeyer, M. & Nairn, A. V. Vertebrate protein glycosylation: Diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.* **13**, 448–462 (2012).
2. Chang, I. J., He, M. & Lam, C. T. Congenital disorders of glycosylation. **6**, 1–13 (2018).
3. Schedin-Weiss, S., Winblad, B. & Tjernberg, L. O. The role of protein glycosylation in Alzheimer disease. *FEBS J.* **281**, 46–62 (2014).
4. Stowell, S. R., Ju, T. & Cummings, R. D. Protein Glycosylation in Cancer. (2015) doi:10.1146/annurev-pathol-012414-040438.
5. Laine, R. A. A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 × 1012 structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology* **4**, 759–767 (1994).
6. Davis, A. P. Synthetic lectins. *Org. Biomol. Chem.* **7**, 3629–3638 (2009).
7. Pandhal, J. & Wright, P. C. N-Linked glycoengineering for human therapeutic proteins in bacteria. *Biotechnol. Lett.* **32**, 1189–1198 (2010).
8. Zafar, S., Nasir, A. & Bokhari, H. B. Computational analysis reveals abundance of potential glycoproteins in Archaea, Bacteria and Eukarya. *Bioinformation* **6**, 352–355 (2011).
9. Welply, J. K., Shenbagamurthili, P., Lennarzlll, W. J. & Naiderp, F. Substrate Recognition by Oligosaccharyltransferase. *J. Biol. Chem.* **258**, 11856–11863 (1983).
10. Petrescu, A., Milac, A. & Stefana, M. Statistical analysis of the protein environment of N -glycosylation sites : implications for occupancy , structure , and folding. **14**, (2004).
11. Shakin-eshleman, S. H., Spitalnik, S. L. & Kasturi, L. The Amino Acid at the X Position of an Asn- X -Ser Sequon Is an Important Determinant of N -Linked Core-glycosylation Efficiency. *J. Biol. Chem.* **271**, 6363–6366 (1996).
12. Rao, R. S. P. & Wollenweber, B. Do N-glycoproteins have preference for specific sequons? *Bioinformation* **5**, 208–212 (2010).
13. Bañó-Polo, M., Baldin, F., Tamborero, S., Marti-Renom, M. A. & Mingarro, I. N-Glycosylation efficiency is determined by the distance to the C-terminus and the amino acid preceding an Asn-Ser-Thr sequon. *Protein Sci.* **20**, 179–186 (2011).
14. Helenius, A. & Aebi, M. Roles of N-linked Glycans in the Endoplasmic Reticulum. *Annu. Rev. Biochem.* **73**, 1019–1049 (2004).
15. Kornfeld, R. & Kornfeld, S. Assembly of Asparagine-linked oligosaccharides. *Annu. Rev. Biochem.* **54**, 631–664 (1985).
16. Aebi, M. N-linked protein glycosylation in the ER. *Biochim. Biophys. Acta* **1833**, 2430–2437 (2013).
17. Nilsson, I. Determination of the Distance between the Oligosaccharyltransferase Active Site and the Endoplasmic Reticulum Membrane. *J. Biol. Chem.* **268**, 5798–5801 (1993).
18. Stanley, P., Taniguchi, N. & Aebi, M. Chapter 9: N-Glycans. in *Essentials of Glycobiology* (2017).
19. Traving, C. & Schauer, R. Structure, function and metabolism of sialic acids. *Cell. Mol. Life Sci.* **54**, 1330–1349 (1998).
20. Varki, A., Schnaar, R. L. & Schauer, R. Chapter 15: Sialic acids and other Nonulosonic acids. in *Essentials of Glycobiology* (2017).
21. Nettleship, J. E. Structural Biology of Glycoproteins. in *Glycoproteins* 41–62 (2012).
22. Hebert, D. N., Zhang, J., Chen, W., Foellmer, B. & Helenius, A. The Number and Location of Glycans on Influenza Hemagglutinin Determine Folding and Association with Calnexin and Calreticulin. *J. Cell Biol.* **139**, 613–623 (1997).
23. Aebi, M., Bernasconi, R., Clerc, S. & Molinari, M. N-glycan structures: recognition and processing in the ER. *Trends Biochem. Sci.* **35**, 74–82 (2010).

24. Hosokawa, N., Kamiya, Y., Kamiya, D., Kato, K. & Nagata, K. Human OS-9 , a Lectin Required for Glycoprotein Endoplasmic Reticulum-associated Degradation, Recognizes Mannose-trimmed N-glycans. *J. Biol. Chem.* **284**, 17061–17068 (2009).
25. Chang, V. T. *et al.* Glycoprotein Structural Genomics: Solving the Glycosylation Problem. *Structure* **15**, 267–273 (2007).
26. Brockhausen, I., Schachter, H. & Stanley, P. Chapter 10: O-GalNAc glycans. in *Essentials of Glycobiology* (2017).
27. Rana, N. A. & Haltiwanger, R. S. Fringe benefits : Functional and structural impacts of O -glycosylation on the extracellular domain of Notch receptors. *Curr. Opin. Struct. Biol.* **21**, 583–589 (2011).
28. Orlean, P. & Menon, A. K. Thematic review series : Lipid Posttranslational Modifications GPI anchoring of protein in yeast and mammalian cells , or : how we learned to stop worrying and love glycophospholipids. *J. Lipid Res.* **48**, 993–1011 (2007).
29. Paulick, M. G. & Bertozzi, C. R. The glycosylphosphatidylinositol anchor: A complex membrane-anchoring structure for proteins. *Biochemistry* **47**, 6991–7000 (2008).
30. Hofsteenge, J. *et al.* C -Mannosylation and O -Fucosylation of the Thrombospondin Type 1 Module. *J. Biol. Chem.* **276**, 6485–6498 (2001).
31. Shcherbakova, A., Tiemann, B., Buettner, F. F. R. & Bakker, H. Distinct C-mannosylation of netrin receptor thrombospondin type 1 repeats by mammalian. *Proc. Natl. Acad. Sci.* **114**, 2574–2579 (2017).
32. Wang, L. W., Leonhard-melief, C., Haltiwanger, R. S. & Apte, S. S. Post-translational Modification of Thrombospondin Type-1 Repeats in ADAMTS-like 1/ Punctin-1 by C -Mannosylation of Tryptophan. *J. Biol. Chem.* **284**, 30004–30015 (2009).
33. Haynes, P. A. Phosphoglycosylation : a new structural class of glycosylation? *Glycobiology* **8**, 1–5 (1998).
34. Gustafsonl, G. L. & Milner, L. A. Occurrence of N-Acetylglucosamine-1-phosphate in Proteinase I from Dictyostelium discoideum. *J. Biol. Chem.* **255**, 7208–7210 (1980).
35. Souza, G. M., Hirai, J., Mehta, D. P. & Freeze, H. H. Identification of Two Novel Dictyostelium discoideum Cysteine Proteinases That Carry N -Acetylglucosamine-1-P Modification. *J. Biol. Chem.* **270**, 28938–28945 (1995).
36. Ilgs, T. *et al.* O- and N-Glycosylation of the Leishmania mexicana-secreted Acid Phosphatase. *J. Biol. Chem.* **269**, (1994).
37. Maynard, J. C., Burlingame, A. L. & Medzihradszky, K. F. Cysteine S-linked N-acetylglucosamine (S-GlcNAcylation), A New Post-translational Modification in Mammals. *Mol. Cell. proteomics* **15**, 3405–3411 (2016).
38. Olsen, E. H. N., Rahbek-nielsen, H., Thøgersen, I. B., Roepstorff, P. & Enghild, J. J. Posttranslational Modifications of Human Inter- R -Inhibitor : Identification of Glycans and Disulfide Bridges in Heavy Chains 1 and 2. *Biochemistry* **37**, 408–416 (1998).
39. Stepper, J. *et al.* Cysteine S -glycosylation , a new post-translational modification found in glycopeptide bacteriocins. *FEBS J.* **585**, 645–650 (2011).
40. Drickamer, K. Two Distinct Classes of Carbohydrate-recognition Domains in Animal Lectins. *J. Biol. Chem.* **263**, 9557–9560 (1988).
41. Hirabayashi, J., Tateno, H., Shikanai, T., Aoki-kinoshita, K. F. & Narimatsu, H. The Lectin Frontier Database (LfDB), and Data Generation Based on Frontal Affinity Chromatography. *Molecules* **20**, 951–973 (2015).
42. Kilpatrick, D. C. Animal lectins: A historical introduction and overview. *Biochim. Biophys. Acta - Gen. Subj.* **1572**, 187–197 (2002).
43. Taylor, M. E. & Drickamer, K. Mammalian sugar-binding receptors : known functions and unexplored roles. *FEBS J.* **286**, 1800–1814 (2019).

44. Vastaa, G. R. *et al.* Structural and functional diversity of the lectin repertoire in teleost fish: Relevance to innate and adaptive immunity. *Dev. Comp. Immunol.* **35**, 1388–1399 (2011).
45. Hansen, S. & Holmskov, U. Structural Aspects of Collectins and Receptors for Collectins. *Immunobiology* **199**, 165–189 (1998).
46. Kumar, K., Prakash Chandra, L., Sumanthi, J., Reddy, S. & Shekar, C. Biological role of lectins : A review. **4**, 20–25 (2012).
47. Massé, K., Baldwin, R., Barnett, M. W. & Jones, E. A. X-epilectin: A novel epidermal fucolectin regulated by BMP signalling. *Int. J. Dev. Biol.* **48**, 1119–1129 (2004).
48. Pietrzyk-Brzezinska, A. J. & Bujacz, A. H-type lectins – Structural characteristics and their applications in diagnostics, analytics and drug delivery. *Int. J. Biol. Macromol.* **152**, 735–747 (2020).
49. Sanchez, J. F. *et al.* Biochemical and structural analysis of Helix pomatia agglutinin: A hexameric lectin with a novel fold. *J. Biol. Chem.* **281**, 20171–20180 (2006).
50. Angata, T. & Linden, E. C. M. B. Der. I-type lectins. *Biochem. Biophys. Acta* **1572**, 294–316 (2002).
51. Cummings, R. D., Etzler, M. E. & Surolia, A. L-tpye Lectins (chapter 32). in *Essentials of Glycobiology* (Cold Spring Harbor Laboratory Press, 2017).
52. Kamiya, Y. *et al.* Molecular Basis of Sugar Recognition by the Human L-type. **283**, 1857–1861 (2008).
53. Benyair, R. *et al.* Mammalian ER mannosidase i resides in quality control vesicles, where it encounters its glycoprotein substrates. *Mol. Biol. Cell* **26**, 172–184 (2015).
54. Dahms, N. M. & Hancock, M. K. P-type lectins. *Biochim. Biophys. Acta* **1572**, 317–340 (2002).
55. Varki, A., Cummings, R. D. & Schnaar, R. Chapter 31: R-Type Lectins. in *Essentials of Glycobiology* (2017).
56. Inamori, K. I. *et al.* A newly identified horseshoe crab lectin with specificity for blood group A antigen recognizes specific O-antigens of bacterial lipopolysaccharides. *J. Biol. Chem.* **274**, 3272–3278 (1999).
57. Wesener, D. A. *et al.* Recognition of Microbial Glycans by Human Intelectin. *Nat. Struct. Mol. Biol.* **22**, 603–610 (2015).
58. Gabius, H. J. Animal lectins. *Eur. J. Biochem.* **243**, 543–576 (1997).
59. Yu Xia, Václav Vetvicka, Jun Yan, M. H. & Ross, T. M. and G. D. The β -Glucan-Binding Lectin Site of Mouse CR3 (CD11b/CD18) and Its Function in Generating a Primed State of the Receptor That Mediates Cytotoxic Activation in Response to iC3b-Opsonized Target Cells. *J. Immunol.* **162**, 2281–2290 (1999).
60. Arnaud, J., Audfray, A. & Imberty, A. Binding sugars: From natural lectins to synthetic receptors and engineered neolectins. *Chem. Soc. Rev.* **42**, 4798–4813 (2013).
61. Collins, P. M., Hidari, K. I. P. J. & Blanchard, H. Slow diffusion of lactose out of galectin-3 crystals monitored by X-ray crystallography: Possible implications for ligand-exchange protocols. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **63**, 415–419 (2007).
62. Ziolkowska, N. E. *et al.* Crystallographic, Thermodynamic, and Molecular Modeling Studies of the Mode of Binding of Oligosaccharides to the Potent Antiviral Protein Griffithsin. *PROTEINS Struct. Funct. Bioinforma.* **67**, 661–670 (2007).
63. Liu, Y., Misulovin, Z. & Bjorkman, P. J. The molecular mechanism of sulfated carbohydrate recognition by the cysteine-rich domain of mannose receptor. *J. Mol. Biol.* **305**, 481–490 (2001).
64. Cioci, G. *et al.* β-Propeller crystal structure of Psathyrella velutina lectin: An integrin-like fungal protein interacting with monosaccharides and calcium. *J. Mol. Biol.* **357**,

1575–1591 (2006).

65. Olson, L. J. *et al.* Crystal Structure and Functional Analyses of the Lectin Domain of Glucosidase II: Insights into Oligomannose Recognition. *Biochemistry* **54**, 4097–4111 (2015).
66. Satoh, T. *et al.* Structural Basis for Oligosaccharide Recognition of Misfolded Glycoproteins by OS-9 in ER-Associated Degradation. *Mol. Cell* **40**, 905–916 (2010).
67. Hancock, Haskins, D. J., Sun, G. & Dahms, N. M. Identification of residues essential for carbohydrate recognition by the insulin-like growth factor II/mannose 6-phosphate receptor. *J. Biol. Chem.* **277**, 11255–11264 (2002).
68. D'Alessio, C. & Dahms, N. M. Glucosidase II and MRH-Domain Containing Proteins in the Secretory Pathway. *Curr. Protein Pept. Sci.* **16**, 31–48 (2015).
69. Lamming, D. W. & Bar-Peled, L. Lysosome: The metabolic signaling hub. *Traffic* **20**, 27–38 (2019).
70. Bochel, A. J. *et al.* Structure of the Human Cation-Independent Mannose 6-Phosphate / IGF2 Receptor Domains 7 – 11 Uncovers the Mannose 6-Phosphate Binding Site of Domain 9. *Structure* 1–13 (2020) doi:10.1016/j.str.2020.08.002.
71. Lefrancois, S., Zeng, J., Hassan, A. J., Canuel, M. & Morales, C. R. The lysosomal trafficking of sphingolipid activator proteins (SAPs) is mediated by sortilin. *EMBO J.* **22**, 6430–6437 (2003).
72. Reczek, D. *et al.* LIMP-2 Is a Receptor for Lysosomal Mannose-6-Phosphate-Independent Targeting of β-Glucocerebrosidase. *Cell* **131**, 770–783 (2007).
73. Gegg, M. E. & Schapira, A. H. V. The role of glucocerebrosidase in Parkinson disease pathogenesis. *FEBS J.* **285**, 3591–3603 (2018).
74. Braulke, T. & Bonifacino, J. S. Sorting of lysosomal proteins. *Biochim. Biophys. Acta - Mol. Cell Res.* **1793**, 605–614 (2009).
75. Zaccheo, O. J. *et al.* Kinetics of Insulin-like Growth Factor II (IGF-II) Interaction with Domain 11 of the Human IGF-II/Mannose 6-phosphate Receptor: Function of CD and AB Loop Solvent-exposed Residues. *J. Mol. Biol.* **359**, 403–421 (2006).
76. Sacher, M. *et al.* The crystal structure of CREG, a secreted glycoprotein involved in cellular growth and differentiation. *Proc. Natl. Acad. Sci. USA* **102**, 18326–18331 (2005).
77. Ghobrial, G., Araujo, L., Jinwala, F., Li, S. & Lee, L. Y. The Structure and Biological Function of CREG. *Front. Cell Dev. Biol.* **6**, 1–7 (2018).
78. Bao, M., Elmendorf, B. J., Booth, J. L., Drake, R. R. & Canfield, W. M. Bovine UDP-N -acetylglucosamine : Lysosomal-enzyme N -Acetylglucosamine-1-phosphotransferase. *J. Biol. Chem.* **271**, 31446–31451 (1996).
79. Reitman, M. L. & Kornfeld, S. UDP-N-Acetylg1ucosamine : Glycoprotein phosphotransferase. *J. Biol. Chem.* **256**, 4275–4281 (1981).
80. Kudo, M. *et al.* The α- and β-subunits of the human UDP-N-acetylglucosamine: lysosomal enzyme phosphotransferase are encoded by a single cDNA. *J. Biol. Chem.* **280**, 36141–36149 (2005).
81. Qian, Y. *et al.* Functions of the a,b and y Subunits of UDP-GlcNAc : Lysosomal. *J. Biol. Chem.* **285**, 3360–3370 (2010).
82. Bohnsack, R. N. *et al.* Cation-independent mannose 6-phosphate receptor: A composite of distinct phosphomannosyl binding sites. *J. Biol. Chem.* **284**, 35215–35226 (2009).
83. Rohrer, J. & Kornfeld, R. Lysosomal hydrolase mannose 6-phosphate uncovering enzyme resides in the trans-Golgi network. *Mol. Biol. Cell* **12**, 1623–1631 (2001).
84. Lazzarino, D. A. & Gabels, A. Mannose Processing Is an Important Determinant in the Assembly of Phosphorylated High Mannose-type Oligosaccharides. *J. Biol. Chem.* **2134**, 5015–5023 (1989).

85. Chao, H. H., Waheed, A., Pohlmann, R., Hille, A. & von Figura, K. Mannose 6-phosphate receptor dependent secretion of lysosomal enzymes. *EMBO J.* **9**, 3507–3513 (1990).
86. Sleat, D. E. & Lobel, P. Ligand binding specificities of the two mannose 6-phosphate receptors. *J. Biol. Chem.* **272**, 731–738 (1997).
87. Roberts, D. L., Weix, D. J., Dahms, N. M. & Kim, J.-J. P. Molecular basis of lysosomal enzyme recognition: Three-dimensional structure of the cation-dependent mannose 6-phosphate receptor. *Cell* **93**, 639–648 (1998).
88. Olson, L. J., Zhang, J., Lee, Y. C., Dahms, N. M. & Kim, J. J. P. Structural basis for recognition of phosphorylated high mannose oligosaccharides by the cation-dependent mannose 6-phosphate receptor. *J. Biol. Chem.* **274**, 29889–29896 (1999).
89. Tong, P. Y. & Kornfeld, S. Ligand interactions of the cation-dependent mannose 6-phosphate receptor. Comparison with the cation-independent mannose 6-phosphate receptor. *J. Biol. Chem.* **264**, 7970–7975 (1989).
90. Olson, L. J., Zhang, J., Dahms, N. M. & Kim, J. P. Twists and turns of the cation-dependent mannose 6-phosphate receptor. Ligand-bound versus ligand-free receptor. *J. Biol. Chem.* **277**, 10156–10161 (2002).
91. Marron-terada, P. G., Bollinger, K. E. & Dahms, N. M. Characterization of Truncated and Glycosylation-Deficient Forms of the Cation-Dependent Mannose 6-Phosphate Receptor Expressed in. *Biochemistry* **37**, 17223–17229 (1998).
92. Li, M., Distler, J. J. & Jourdian, G. W. The Aggregation and Dissociation Properties of a Low Molecular Weight Mannose 6-Phosphate Receptor from Bovine Testis'. *Arch. Biochem. Biophys.* **283**, 150–157 (1990).
93. Waheed, A. & Von Figura, K. Rapid equilibrium between monomeric, dimeric and tetrameric forms of the 46-kDa mannose 6-phosphate receptor at 37oC. *Eur. J. Biochem.* **193**, 47–54 (1990).
94. Junghans, U., Waheed, A. & von Figura, K. The 'cation-dependent' mannose 6-phosphate receptor binds ligands in the absence of divalent cations. *FEBS Lett.* **237**, 81–84 (1988).
95. Killian, J. K. & Jirtle, R. L. Genomic structure of the human M6P/IGF2 receptor. *Mamm. Genome* **10**, 74–77 (1999).
96. Hassan, A. B. Keys to the hidden treasures of the mannose 6-phosphate/insulin-like growth factor 2 receptor. *Am. J. Pathol.* **162**, 3–6 (2003).
97. Lemamy, G. J., Ndeboko, B., Omouessi, S. T. & Mouecoucou, J. Mannose-6-Phosphate/ Insulin-Like Growth Factor 2 Receptor (M6P/IGF2-R) in Growth and Disease: A Review. in *Restricted Growth - Clinical, Genetic and Molecular Aspects* 147–162 (2016).
98. Van Meel, E. & Klumperman, J. TGN exit of the cation-independent mannose 6-phosphate receptor does not require acid hydrolase binding. *Cell. Logist.* **4**, e954441 (2014).
99. Misra, S., Puertollano, R., Kato, Y., Bonifacino, J. S. & Hurley, J. H. Structural basis for acidic-clusterdileucine sorting-signal recognition by VHS domains. *Nature* **415**, 933–937 (2002).
100. Seaman, M. N. J. Identification of a novel conserved sorting motif required for retromer-mediated endosome-to-TGN retrieval. *J. Cell Sci.* **120**, 2378–2389 (2007).
101. Brown, J. *et al.* Structure and functional analysis of the IGF-II/IGF2R interaction. *EMBO J.* **27**, 265–276 (2008).
102. Motyka, B. *et al.* Mannose 6-phosphate/insulin-like growth factor II receptor is a death receptor for granzyme B during cytotoxic T cell-induced apoptosis. *Cell* **103**, 491–500 (2000).

103. Olson, L. J., Yammani, R. D., Dahms, N. M. & Kim, J. P. Structure of uPAR, plasminogen, and sugar-binding sites of the 300 kDa mannose 6-phosphate receptor. *EMBO J.* **23**, 2019–28 (2004).
104. Dennis, P. A. & Rifkin, D. B. Cellular activation of latent transforming growth factor B requires binding to the cation-independent mannose 6-phosphate / insulin- like growth factor type II receptor. **88**, 580–584 (1991).
105. Blanchard, F. *et al.* The mannose 6-phosphate/insulin-like growth factor II receptor is a nanomolar affinity receptor for glycosylated human leukemia inhibitory factor. *J. Biol. Chem.* **273**, 20886–20893 (1998).
106. Van den Eijnden, M. M. E. D. *et al.* Prorenin accumulation and activation in human endothelial cells: Importance of mannose 6-phosphate receptors. *Arterioscler. Thromb. Vasc. Biol.* **21**, 911–916 (2001).
107. Lee, S.-J. & Nathans, D. Proliferin Secreted by Cultured Cells Binds to Mannose 6-Phosphate Receptors. *J. Biol. Chem.* **263**, 3521–3527 (1988).
108. Ikushima, H. *et al.* Internalization of CD26 by mannose 6-phosphate/insulin-like growth factor II receptor contributes to T cell activation. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 8439–44 (2000).
109. Herzog, V., Neumüller, W. & Holzmann, B. Thyroglobulin, the major and obligatory exportable protein of thyroid follicle cells, carries the lysosomal recognition marker mannose-6-phosphate. *EMBO J.* **6**, 555–560 (1987).
110. Todderud, G. & Carpenter, G. Presence of mannose phosphate on the epidermal growth factor receptor in A-431 cells. *J. Biol. Chem.* **263**, 17893–17896 (1988).
111. Brunetti, C. R., Dingwell, K. S., Wale, C., Graham, F. L. & Johnson, D. C. Herpes simplex virus gD and virions accumulate in endosomes by mannose 6-phosphate-dependent and -independent mechanisms. *J. Virol.* **72**, 3330–9 (1998).
112. Brown, J. *et al.* Structure of a functional IGF2R fragment determined from the anomalous scattering of sulfur. *EMBO J.* **21**, 1054–1062 (2002).
113. Kang, J. X., Li, Y. & Leaf, A. Mannose-6-phosphate/insulin-like growth factor-II receptor is a receptor for retinoic acid. *Proc. Natl. Acad. Sci.* **94**, 13671–6 (1997).
114. Leksa, V. *et al.* The N terminus of mannose 6-phosphate/insulin-like growth factor 2 receptor in regulation of fibrinolysis and cell migration. *J. Biol. Chem.* **277**, 40575–40582 (2002).
115. Leksa, V., Ilková, A., Vi, K. & Stockinger, H. Unravelling novel functions of the endosomal transporter mannose 6- phosphate / insulin-like growth factor receptor (CD222) in health and disease : An emerging regulator of the immune system. **190**, 194–200 (2017).
116. Pfisterer, K. *et al.* The Late Endosomal Transporter CD222 Directs the Spatial Distribution and Activity of Lck. *J. Immunol.* **193**, 2718–2732 (2014).
117. Wood, R. J. & Hulett, M. D. Cell surface-expressed cation-independent mannose 6-phosphate receptor (CD222) binds enzymatically active heparanase independently of mannose 6-phosphate to promote extracellular matrix degradation. *J. Biol. Chem.* **283**, 4165–4176 (2008).
118. Livingstone, C. IGF2 and cancer. *Endocr. Relat. Cancer* **20**, 321–339 (2013).
119. Engström, W. *et al.* Transcriptional regulation and biological significance of the insulin like growth factor II gene. *Cell prolif.* **31**, 173–189 (1998).
120. Greenall, S. A. *et al.* Biochemical characterization of individual human glycosylated pro-insulin-like growth factor (IGF)-II and big-IGF-II isoforms associated with cancer. *J. Biol. Chem.* **288**, 59–68 (2013).
121. Qiu, Q., Basak, A., Mbikay, M. & Tsang, B. K. Role of pro-IGF-II processing by proprotein convertase 4 in human placental development. **102**, (2005).

122. Foulstone, E. *et al.* Insulin-like growth factor ligands, receptors, and binding proteins in cancer. *J. Pathol.* **205**, 145–153 (2005).
123. Haywood, N. J., Slater, T. A., Matthews, C. J. & Wheatcroft, S. B. The insulin like growth factor and binding protein family: Novel therapeutic targets in obesity & diabetes. *Mol. Metab.* **19**, 86–96 (2019).
124. Williams, C. *et al.* An Exon Splice Enhancer Primes IGF2:IGF2R Binding Site Structure and Function Evolution. *Science (80-.).* **338**, 1209–1213 (2012).
125. Ludwig, T. *et al.* Mouse Mutants Lacking the Type 2 IGF Receptor (IGF2R) Are Rescued from Perinatal Lethality in Igf2 and Igf1r Null Backgrounds. *Dev. Biol.* **177**, 517–535 (1996).
126. Wang, Z., Fung, M. R., Barlow, D. P. & Wagner, E. F. Regulation of embryonic growth and lysosomal targeting by the imprinted. *Lett. to Nat.* **372**, 464–467 (1994).
127. Frago, S. *et al.* Functional evolution of IGF2:IGF2R domain 11 binding generates novel structural interactions and a specific IGF2 antagonist. *Proc. Natl. Acad. Sci.* **113**, E2766–E2775 (2016).
128. Wang, R., Qi, X., Schmiege, P., Coutavas, E. & Li, X. Marked structural rearrangement of mannose 6-phosphate/IGF2 receptor at different pH environments. *Sci. Adv.* **6**, eaaz1466 (2020).
129. Dahms, N. M., Rose, P. a, Molkentin, J. D., Zhang, Y. & Brzycki, M. a. The Bovine Mannose 6-Phosphate/Insulin-like Growth Factor II Receptor. **268**, 5457–5463 (1993).
130. Olson, L. J. *et al.* Identification of a fourth mannose 6-phosphate binding site in the cation-independent mannose 6-phosphate receptor. *Glycobiology* **25**, 591–606 (2015).
131. Olson *et al.* Structural basis for recognition of phosphodiester-containing lysosomal enzymes by the cation-independent mannose 6-phosphate recepbtor. *Proc. Natl. Acad. Sci.* **107**, 12493–12498 (2010).
132. Marron-Terada, P. G., Hancock, M. K., Haskins, D. J. & Dahms, N. M. Recognition of Dictyostelium discoideum lysosomal enzymes is conferred by the amino-terminal carbohydrate binding site of the insulin-like growth factor II/mannose 6-phosphate receptor. *Biochemistry* **39**, 2243–2253 (2000).
133. Song, X. *et al.* Glycan microarray analysis of P-type lectins reveals distinct phosphomannose glycan recognition. *J. Biol. Chem.* **284**, 35201–35214 (2009).
134. Christianson, J. C., Shaler, T. A., Tyler, R. E. & Kopito, R. R. OS-9 and GRP94 deliver mutant α 1-antitrypsin to the Hrd1 – SEL1L ubiquitin ligase complex for ERAD. *Nat. Cell Biol.* **10**, 272–282 (2008).
135. Tong, P. Y., Gregory, W. & Kornfeld, S. Ligand interactions of the cation-independent mannose 6-phosphate receptor. The stoichiometry of mannose 6-phosphate binding. *J. Biol. Chem.* **264**, 7962–7969 (1989).
136. Olson, L. J., Dahms, N. M. & Kim, J. P. The N-terminal Carbohydrate Recognition Site of the Cation-independent Mannose 6-Phosphate Receptor. *J. Biol. Chem.* **279**, 34000–34009 (2004).
137. Hancock, Yammani, R. D. & Dahms, N. M. Localization of the Carbohydrate Recognition Sites of the Insulin- like Growth Factor II / Mannose 6-Phosphate Receptor to Domains 3 and 9 of the Extracytoplasmic Region. *J. Biol. Chem.* **277**, 47205–47212 (2002).
138. Olson, L. J. *et al.* Allosteric regulation of lysosomal enzyme recognition by the cation-independent mannose 6-phosphate receptor. *Commun. Biol.* **3**, 1–15 (2020).
139. Chavez, C. A. *et al.* Domain 5 of the cation-independent mannose 6-phosphate receptor preferentially binds phosphodiesters (mannose 6-phosphate N-acetylglucosamine ester). *Biochemistry* **46**, 12604–12617 (2007).
140. Han, Y. *et al.* Glycosylation-independent binding to extracellular domains 11-13 of

mannose-6-phosphate/insulin-like growth factor-2 receptor mediates the effects of soluble CREG on the phenotypic modulation of vascular smooth muscle cells. *J. Mol. Cell. Cardiol.* **50**, 723–730 (2011).

141. Bacco, A. Di & Gill, G. The secreted glycoprotein CREG inhibits cell growth dependent on the mannose-6-phosphate / insulin-like growth factor II receptor. *Oncogene* **22**, 5436–5445 (2003).

142. Nykjær, A. *et al.* Receptor Targets the Urokinase Receptor to. *Cell* **141**, 815–828 (1998).

143. Kreiling, J. L. *et al.* Binding of Urokinase-type Plasminogen Activator Receptor (uPAR) to the Mannose 6-Phosphate / Insulin-like Growth Factor II Receptor. **278**, 20628–20637 (2003).

144. Oliveira, L. D. M. *et al.* Impact of Retinoic Acid on Immune Cells and Inflammatory Diseases. *Nat. Rev. Mol. Cell Biol.* **10**, 445–457 (2018).

145. Janesick, A., Cherie, S. & Blumberg, B. Retinoic acid signaling and neuronal differentiation. *Cell. Mol. Life Sci.* **72**, 1559–1576 (2015).

146. Lee, W. S., Kang, C., Drayna, D. & Kornfeld, S. Analysis of mannose 6-phosphate uncovering enzyme mutations associated with persistent Stuttering. *J. Biol. Chem.* **286**, 39786–39793 (2011).

147. Brady, R. O. Enzyme Replacement for Lysosomal Diseases. *Annu. Rev. Med.* **57**, 283–296 (2006).

148. Matrone, C. *et al.* Mannose 6-phosphate receptor is reduced in -synuclein overexpressing models of parkinsons disease. *PLoS One* **11**, 1–21 (2016).

149. Kar, S. *et al.* Cellular distribution of insulin-like growth factor-II/mannose-6-phosphate receptor in normal human brain and its alteration in Alzheimer's disease pathology. *Neurobiol. Aging* **27**, 199–210 (2006).

150. Killian JK, Nolan CM, Stewart N, Munday BL, Andersen NA, Nicol S, J. R. Monotreme IGF2 expression and ancestral origin of genomic imprinting. *J Exp Zool.* **291**, 205–12 (2001).

151. Williams, C. *et al.* Structural Insights into the Interaction of Insulin-like Growth Factor 2 with IGF2R Domain 11. *Structure* **15**, 1065–1078 (2007).

152. Byrd, J. C., Devi, G. R., De Souza, A. T., Jirtle, R. L. & MacDonald, R. G. Disruption of ligand binding to the insulin-like growth factor II/mannose 6-phosphate receptor by cancer-associated missense mutations. *J. Biol. Chem.* **274**, 24408–24416 (1999).

153. Souza, A. T. De *et al.* M6P/IGF2R gene is mutated in human hepatocellular carcinomas with loss of heterozygosity. *Nat. Genet.* **11**, 447–449 (1995).

154. Souza, A. T. De, Yamada, T., Mills, J. J. & Jirtle, R. L. Imprinted genes in liver carcinogenesis. *FASEB J.* **11**, 60–67 (1997).

155. Devi, G. R., Souza, A. T. De, Byrd, J. C., Jirtle, R. L. & Macdonald, R. G. Altered Ligand Binding by Insulin-like Growth Factor II / Mannose 6-Phosphate Receptors Bearing Missense Mutations in Human Cancers 1. *Cancer Res.* **59**, 4314–4319 (1999).

156. Probst, O. C. *et al.* The mannose 6-phosphate/insulin-like growth factor II receptor restricts the tumourigenicity and invasiveness of squamous cell carcinoma cells. *Int. J. Cancer* **124**, 2559–2567 (2009).

157. Harper, J. *et al.* Soluble IGF2 receptor rescues ApcMin/+ intestinal adenoma progression induced by Igf2 loss of imprinting. *Cancer Res.* **66**, 1940–1948 (2006).

158. Sasaki, H., Ishihara, K. & Kato, R. Mechanisms of Igj2 / H19 Imprinting : DNA Methylation , Chromatin and Long-Distance Gene Regulation. *J Biochem.* **127**, 711–715 (2000).

159. Abi Habib, W. *et al.* 11p15 ICR1 Partial Deletions Associated with IGF2/H19 DMR Hypomethylation and Silver-Russell Syndrome. *Hum. Mutat.* **38**, 105–111 (2017).

160. Iliev, D. & Binder, G. Silver-Russell syndrome. *Pediatriya* **51**, 29–31 (2011).

161. NIH Genetics Home Reference. https://ghr.nlm.nih.gov/condition/russell-silver-syndrome#statistics.
162. Weksberg, R., Shuman, C. & Beckwith, J. B. Beckwith-Wiedemann syndrome. *Eur. J. Hum. Genet.* **18**, 8–14 (2010).
163. Leksa, V. *et al.* Soluble M6P/IGF2R released by TACE controls angiogenesis via blocking plasminogen activation. *Circ. Res.* **108**, 676–685 (2011).
164. Kreiling, J. L. *et al.* Dominant-negative effect of truncated mannose 6-phosphate/insulin-like growth factor II receptor species in cancer. *FEBS J.* **279**, 2695–2713 (2012).
165. Jeyaratnaganthan, N. *et al.* Circulating levels of insulin-like growth factor-II/mannose-6-phosphate receptor in obesity and type 2 diabetes. *Growth Horm. IGF Res.* **20**, 185–191 (2010).
166. Zaina, S. & Squire, S. The soluble type 2 insulin-like growth factor (IGF2) receptor reduces organ size by IGF2-mediated and IGF2-independent mechanisms. *J. Biol. Chem.* **273**, 28610–28616 (1998).
167. Prince, S. N., Foulstone, E. J., Zaccheo, O. J., Williams, C. & Hassan, A. B. Functional evaluation of novel soluble insulin-like growth factor (IGF)-II-specific ligand traps based on modified domain 11 of the human IGF2 receptor. *Mol. Cancer Ther.* **6**, 607–17 (2007).
168. Zhu, Y. *et al.* Conjugation of mannose 6-phosphate-containing oligosaccharides to acid α-glucosidase improves the clearance of glycogen in Pompe mice. *J. Biol. Chem.* **279**, 50336–50341 (2004).
169. Koeberl, D. D. *et al.* Enhanced Efficacy of Enzyme Replacement Therapy in Pompe Disease Through Mannose-6-Phosphate Receptor Expression in Skeletal Muscle. *Mol. Genet. Metab.* **103**, 107–112 (2011).
170. Westlund, B., Dahms, N. M. & Kornfeld, S. The bovine mannose 6-phosphate/insulin-like growth factor II receptor. Localization of mannose 6-phosphate binding sites to domains 1-3 and 7-11 of the extracytoplasmic region. *J. Biol. Chem.* **266**, 23233–23239 (1991).
171. Byrd, J. C., Park, J. H. Y., Schaffer, B. S., Garmroudi, F. & MacDonald, R. G. Dimerization of the insulin-like growth factor II/mannose 6-phosphate receptor. *J. Biol. Chem.* **275**, 18647–18656 (2000).
172. Gao G *et al.* Assembly and crystallization of the complex between the human T cell coreceptor CD8 alpha-alpha dimer and HLA-A2. *Protein Sci.* **7**, 1245–1249 (1998).
173. Singh, A., Upadhyay, V. & Panda, A. K. Solubilization and refolding of inclusion body proteins. *Insoluble Proteins Methods Protoc.* **99**, 283–291 (2014).
174. Bajorunaite, E., Sereikaite, J. & Bumelis, V. A. L-arginine suppresses aggregation of recombinant growth hormones in refolding process from E. coli inclusion bodies. *Protein J.* **26**, 547–555 (2007).
175. Vallejo, L. F. & Rinas, U. Strategies for the recovery of active proteins through refolding of bacterial inclusion body proteins. *Microb. Cell Fact.* **3**, 11 (2004).
176. Wüthrich, K. *NMR of Proteins and Nucleic Acids.*
177. Olson, L. J. *et al.* Bacterial expression of the phosphodiester-binding site of the cation-independent mannose 6-phosphate receptor for crystallographic and NMR studies. *Protein Expr. Purif.* **111**, 91–97 (2015).
178. Marblestone, J. G. *et al.* Comparison of SUMO fusion technology with traditional gene fusion systems : Enhanced expression and solubility with SUMO. *Protein Sci.* **15**, 182–189 (2006).
179. Trowitzsch, S. *et al.* New baculovirus expression tools for recombinant protein complex production New baculovirus expression tools for recombinant protein complex

production. *J. Struct. Biol.* (2010) doi:10.1016/j.jsb.2010.02.010.

180. Van Oers, M. M., Pijlman, G. P. & Vlak, J. M. Thirty years of baculovirus-insect cell protein expression: From dark horse to mainstream technology. *J. Gen. Virol.* **96**, 6–23 (2015).
181. Berger, I., Fitzgerald, D. J. & Richmond, T. J. Baculovirus expression system for heterologous multiprotein complexes. *Nat. Biotechnol.* **22**, 1583–1587 (2004).
182. Hom, L. G. & Volkman, L. E. Autographa californica M nucleopolyhedrovirus chiA is required for processing of V-CATH. *Virology* **277**, 178–183 (2000).
183. Kaba, S. A., Salcedo, A. M., Wafula, P. O., Vlak, J. M. & Oers, M. M. Van. Development of a chitinase and v-cathepsin negative bacmid for improved integrity of secreted recombinant proteins. *J. Virol. Methods* **122**, 113–118 (2004).
184. Bieniossek, C., Richmond, T. J. & Berger, I. MultiBac: Multigene baculovirus-based eukaryotic protein complex production. *Curr. Protoc. Protein Sci.* **5**, (2008).
185. Luckow, V. a, Lee, S. C., Barry, G. F. & Olins, P. O. Efficient generation of infectious recombinant baculoviruses by site-specific transposon-mediated insertion of foreign genes into a baculovirus genome propagated in Escherichia coli. *J. Virol.* **67**, 4566–4579 (1993).
186. Nowakowski, A. B., Wobig, W. J. & Petering, D. Native SDS-PAGE: High Resolution Electrophoretic Separation of Proteins With Retention of Native Properties Including Bound Metal Ions. *Metallomics* **6**, 1068–1078 (2014).
187. Kollewe, C. & Vilcinskas, A. Production of recombinant proteins in insect cells. *Am. J. Biochem. Biotechnol.* **9**, 255–271 (2013).
188. Farrell, P. & Iatrou, K. Transfected insect cells in suspension culture rapidly yield moderate quantities of recombinant proteins in protein-free culture medium. *Protein Expr. Purif.* **36**, 177–185 (2004).
189. Shi, X. & Jarvis, D. L. Protein N-glycosylation in the baculovirus-insect cell system. *Curr. Drug Targets* **8**, 1116–25 (2007).
190. Morelle, W., Faid, V., Chirat, F. & Michalski, J. C. *Analysis of N-and O-linked glycans from glycoproteins using MALDI-TOF mass spectrometry*. *Methods in Molecular Biology* vol. 534 (2009).
191. Norris, G. E., Stillman, T. J., Anderson, B. F. & Baker, E. N. The three-dimensional structure of PNGase F , a glycosyl- asparaginase from Flavobacterium meningosepticum. *StructureStructure* **2**, 1049–1059 (1994).
192. Seismann, H. *et al.* Dissecting cross-reactivity in hymenoptera venom allergy by circumvention of. **47**, 799–808 (2010).
193. Mabashi-Asazuma, H., Kuo, C.-W., Khoo, K.-H. & Jarvis, D. L. A novel baculovirus vector for production of nonfucosylated recombinant glycoproteins in insect cells. *Glycobiology* **24**, 325–340 (2014).
194. Tretter, V., Altmann, F. & Marz, L. Peptide-N4- (N-acetyl- / 3-glucosaminyl) asparagine amidase F cannot release glycans with fucose attached a1 + 3 to the asparagine-linked N-acetylglucosamine residue. **652**, 647–652 (1991).
195. Krissinel, E. & Henrick, K. Inference of macromolecular assemblies from crystalline state. *J. Mol. Biol.* **372**, 774–797 (2007).
196. Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Sci.* **27**, 112–128 (2018).
197. Eisenberg, D., Schwarz, E., Komarony, M. & Wall, R. Amino acid scale: Normalized consensus hydrophobicity scale. *J. Mol. Biol.* **179**, 125–142 (1984).
198. Zheng, H. *et al.* CheckMyMetal : a macromolecular metal-binding validation tool research papers. *Acta Crystallogr. Sect. D Struct. Biol.* **73**, 223–233 (2017).
199. Carugo, O. Buried chloride stereochemistry in the Protein Data Bank. *BMC Struct. Biol.*

14, 1–7 (2014).

200. Hendrickson, W. A. & Fan, Q. R. Structure of human follicle-stimulating hormone in complex with its receptor. *Nature* **433**, 269–277 (2005).
201. Davis, S. J. *et al.* Ligand binding by the immunoglobulin superfamily recognition molecule CD2 is glycosylation-independent. *Journal of Biological Chemistry* vol. 270 369–375 (1995).
202. Maley, F., Trimble, R. B., Tarentino, A. L. & Plummer, T. H. Characterization of Glycoproteins and Their Associated Oligosaccharides through the Use of Endoglycosidases 1. **204**, 195–204 (1989).
203. Brondyk, W. H. *Chapter 11 Selecting an Appropriate Method for Expressing a Recombinant Protein*. *Methods in Enzymology* vol. 463 (Elsevier Inc., 2009).
204. Nishio, M., Umezawa, Y., Fantini, J., Weiss, M. S. & Chakrabarti, P. CH-π hydrogen bonds in biological macromolecules. *Phys. Chem. Chem. Phys.* **16**, 12648–12683 (2014).
205. Saha, R. P., Bhattacharyya, R. & Chakrabarti, P. Interaction Geometry Involving Planar Groups in Protein–Protein Interfaces. *PROTEINS Struct. Funct. Bioinforma.* **67**, 84–97 (2007).
206. Olson, L. J., Hindsgaul, O., Dahms, N. M. & Kim, J. P. Structural insights into the mechanism of pH-dependent ligand binding and release by the cation-dependent mannose 6-phosphate receptor. *J. Biol. Chem.* **283**, 10124–10134 (2008).
207. Satoh, T. *et al.* Structural basis for recognition of high mannose type glycoproteins by mammalian transport lectin VIP36. *J. Biol. Chem.* **282**, 28246–28255 (2007).
208. Raingeval, C. *et al.* 1D NMR WaterLOGSY as an efficient method for fragment-based lead discovery. *J. Enzyme Inhib. Med. Chem.* **34**, 1218–1225 (2019).
209. Mayer, M. & Meyer, B. Characterization of ligand binding by saturation transfer difference NMR spectroscopy. *Angew. Chemie - Int. Ed.* **38**, 1784–1788 (1999).
210. Mayer, M. & Meyer, B. Group epitope mapping by saturation transfer difference NMR to identify segments of a ligand in direct contact with a protein receptor. *J. Am. Chem. Soc.* **123**, 6108–6117 (2001).
211. Barile, E. & Pellecchia, M. NMR-based approaches for the identification and optimization of inhibitors of protein-protein interactions. *Chem. Rev.* **114**, 4749–4763 (2014).
212. Dalvit, C., Fogliatto, G., Stewart, A., Veronesi, M. & Stockman, B. WaterLOGSY as a Method for Primary NMR Screening: Practical Aspects and Range of Applicability. *J. Biomol. NMR* **4**, 349–359 (2001).
213. Dalvit, C., Pevarello, P., Tato, M., Veronesi, M. & Sundstrom, M. Identification of Compounds With Binding Affinity to Proteins via Magnetization Transfer From Bulk Water. *J. Biomol. NMR* **1**, 65–68 (2000).
214. Fielding, L. NMR methods for the determination of protein-ligand dissociation constants. *Prog. Nucl. Magn. Reson. Spectrosc.* **51**, 219–242 (2007).
215. Huang, R., Bonnichon, A., Claridge, T. D. W. & Leung, I. K. H. Protein-ligand binding affinity determination by the waterLOGSY method: An optimised approach considering ligand rebinding. *Sci. Rep.* **7**, 1–6 (2017).
216. Schmidt, T. G. M. & Skerra, A. The Strep-tag system for one-step purification and high-affinity detection or capturing of proteins. *Nat. Protoc.* **2**, 1528–1535 (2007).
217. Cole, J. L., Lary, J. W., Moody, T. & Laue, T. M. Analytical Ultracentrifugation: Sedimentation Velocity and Sedimentation Equilibrium. *Methods Cell Biol.* **84**, 143–179 (2008).
218. Lebowitz, J., Lewis, M. S. & Schuck, P. Modern analytical ultracentrifugation in protein science: A tutorial review. *Protein Sci.* **11**, 2067–2079 (2009).

219. Balbo, A. & Schuck, P. Analytical ultracentrifugation in the study of protein self-association and heterogeneous protein-protein interactions. *Protein-Protein Interact. Golemis ...* 253–277 (2005).
220. Laue, T. M., Shah, B. ., Ridgeway, T. . & Pelletier, S. . Analytical Ultracentrifugation in Biochemistry and Polymer Science. *R. Soc. Chem.* 90–125 (1992).
221. Demeler, B., Scott, D. ., Harding, S. E. & Rowe, A. J. Ultrascan: A comprehensive data analysis software package for analytical ultracentrifugation data. *R. Soc. Chem.* 210–229 (2005).
222. Büssow, K. Stable mammalian producer cell lines for structural biology. *Curr. Opin. Struct. Biol.* **32**, 81–90 (2015).
223. Hughes, J. *et al.* Maternal transmission of an Igf2r domain 11: IGF2 binding mutant allele (Igf2r I1565A) results in partial lethality, overgrowth and intestinal adenoma progression. *Sci. Rep.* **9**, 1–16 (2019).
224. Dwyer, B. *et al.* Expression, purification, and characterization of human mannose-6-phosphate receptor – Extra cellular domain from a stable cell line utilizing a small molecule biomimetic of the mannose-6-phosphate moiety. *Protein Expr. Purif.* **170**, 105589 (2020).
225. Takizawa, Y. *et al.* While the revolution will not be crystallized, biochemistry reigns supreme. *Protein Sci.* **26**, 69–81 (2017).
226. Ohi, M., Li, Y., Cheng, L. & Walz, T. Negative Staining and Image Classification - Powerful Tools in Modern Electron Microscopy. *Biol. Proced. Online* **6**, 23–34 (2004).
227. Booth, D. S., Avila-Sakar, A. & Cheng, Y. Visualizing proteins and macromolecular complexes by negative stain EM: from grid preparation to image acquisition. *J. Vis. Exp.* 1–8 (2011) doi:10.3791/3227.
228. Rosa-trevín, J. M. De *et al.* Scipion : A software framework toward integration , reproducibility and validation in 3D electron microscopy. **195**, 93–99 (2016).
229. York, S. J., Arneson, L. S., Gregory, W. T., Dahms, N. M. & Kornfeld, S. The Rate of Internalization of the Mannose 6-Phosphate / Insulin-like Growth Factor II Receptor Is Enhanced by Multivalent Ligand Binding. *J. Biol. Chem.* **274**, 1164–1171 (1999).
230. Byrd, J. C. & MacDonald, R. G. Mechanisms for high affinity mannose 6-phosphate ligand binding to the insulin-like growth factor II/mannose 6-phosphate receptor. Negative cooperativity and receptor oligomerization. *J. Biol. Chem.* **275**, 18638–18646 (2000).
231. Kreiling, J. L., Byrd, J. C. & MacDonald, R. G. Domain interactions of the mannose 6-phosphate/insulin-like growth factor II receptor. *J. Biol. Chem.* **280**, 21067–77 (2005).
232. Kiess, W. *et al.* Insulin-like Growth Factor-II (IGF-1I) Inhibits Both the Cellular the Binding of & Galactosidase to Uptake of b-Galactosidase and Purified IGF-II / Mannose 6-Phosphate Receptor. *J. Biol. Chem.* **264**, 4710–4714 (1989).
233. Distler, J. J., Guo, J. & Jourdians, G. W. The Binding Specificity of High and Low Molecular Weight Phosphomannosyl Receptors from Bovine Testes. *J. Biol. Chem.* **266**, 21687–21692 (1991).
234. Fei, X., Connelly, C. M., MacDonald, R. G. & Berkowitz, D. B. A set of phosphatase-inert ‘molecular rulers’ to probe for bivalent mannose 6-phosphate ligand-receptor interactions. *Bioorganic Med. Chem. Lett.* **18**, 3085–3089 (2008).
235. Zeng, Y., Wang, J., Gu, Z. & Gu, Z. Engineering glucose-responsive insulin. *Med. Drug Discov.* **3**, 1–5 (2019).
236. Houston, T. A. Developing high-affinity boron-based receptors for cell-surface carbohydrates. *ChemBioChem* **11**, 954–957 (2010).
237. James, T. D., Samankumara Sandanayake, K. R. A. & Shinkai, S. Chiral discrimination of monosaccharides using a fluorescent molecular sensor. *Nature* **374**, 345–347 (1995).

238. Eggert, H., Frederiksen, J., Morin, C. & Norrild, J. C. A new glucose-selective fluorescent bisboronic acid. First report of strong α-furanose complexation in aqueous solution at physiological pH. *J. Org. Chem.* **64**, 3846–3852 (1999).

239. Yang, W. *et al.* Diboronic acids as fluorescent probes for cells expressing sialyl Lewis X. *Bioorganic Med. Chem. Lett.* **12**, 2175–2177 (2002).

240. Levonis, S. M., Kiefel, M. J. & Houston, T. A. Boronolectin with divergent fluorescent response specific for free sialic acid. *Chem. Commun. R. Soc. Chem.* **7**, 2278–2280 (2009).

241. Ke, C., Destecroix, H., Crump, M. P. & Davis, A. P. A simple and accessible synthetic lectin for glucose recognition and sensing. *Nat. Chem.* **4**, 718–723 (2012).

242. Carter, T. S. *et al.* Platform Synthetic Lectins for Divalent Carbohydrate Recognition in Water. *Angew. Chemie - Int. Ed.* **55**, 9311–9315 (2016).

243. Mooibroek, T. J., Crump, M. P. & Davis, A. P. Synthesis and evaluation of a desymmetrised synthetic lectin: An approach to carbohydrate receptors with improved versatility. *Org. Biomol. Chem.* **14**, 1930–1933 (2016).

244. Zhou, J. & Rossi, J. Aptamers as targeted therapeutics: Current potential and challenges. *Nat. Rev. Drug Discov.* **16**, 181–202 (2017).

245. Jeong, S., Eom, T. Y., Kim, S. J., Lee, S. W. & Yu, J. In vitro selection of the RNA Aptamer against the Sialyl Lewis X and its inhibition of the cell adhesion. *Biochem. Biophys. Res. Commun.* **281**, 237–243 (2001).

246. Wang, X. kang *et al.* Inhibition of adhesion and metastasis of HepG2 hepatocellular carcinoma cells in vitro by DNA aptamer against sialyl Lewis X. *J. Huazhong Univ. Sci. Technol. - Med. Sci.* **37**, 343–347 (2017).

247. Hu, D., Tateno, H. & Hirabayashi, J. Lectin Engineering, a Molecular Evolutionary Approach to Expanding the Lectin Utilities. *Molecules* **20**, 7637–7656 (2015).

248. Yabe, R. *et al.* Tailoring a Novel Sialic Acid-Binding Lectin from a Ricin-B Chain-like Galactose-Binding Protein by Natural Evolution-Mimicry. *J. Biochem.* **141**, 389–399 (2007).

249. Hu, D., Tateno, H., Kuno, A., Yabe, R. & Hirabayashi, J. Directed evolution of lectins with sugar-binding specificity for 6-sulfo-galactose. *J. Biol. Chem.* **287**, 20313–20320 (2012).

250. Hudson, K. L. *et al.* Carbohydrate-Aromatic Interactions in Proteins. *J. Am. Chem. Soc.* **137**, 15152–15160 (2015).

251. Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, L. E. The CCPN Data Model for NMR Spetroscopy: Development of a Software Pipeline. *Proteins* **59**, 687–696 (2005).

252. Harner, M. J., Frank, A. O. & Fesik, S. W. Fragment-Based Drug Discovery Using NMR Spectroscopy. *J. Biomol. NMR* **56**, 65–75 (2014).

253. Bieri, M. *et al.* Macromolecular NMR spectroscopy for the non-spectroscopist: Beyond macromolecular solution structure determination. *FEBS J.* **278**, 704–715 (2011).

254. Zhang, J., Dahms, N. M. & Kim, J.-J. P. Twists and Turns of the Cation-dependent Mannose 6-Phosphate Receptor. *J. Biol. Chem.* **277**, 10156–10161 (2002).

255. Dias, J. M. *et al.* Structural basis of chemokine sequestration by a tick chemokine binding protein: The crystal structure of the complex between Evasin-1 and CCL3. *PLoS One* **4**, e8514 (2009).

256. Duksin, D. & Mahoney, W. C. Relationship of the structure and biological activity of the natural homologues of tunicamycin. *J. Biol. Chem.* **257**, 3105–3109 (1982).

257. Mei, M. *et al.* Application of modified yeast surface display technologies for non-Antibody protein engineering. *Microbiol. Res.* **196**, 118–128 (2017).

258. McIntosh-Smith, S., Price, J., Sessions, R. B. & Ibarra, A. A. High performance in silico

virtual drug screening on many-core processors. *Int. J. High Perform. Comput. Appl.* **29**, 119–134 (2015).

259. Ibarra, A. A. *et al.* Predicting and Experimentally Validating Hot-Spot Residues at Protein-Protein Interfaces. *ACS Chem. Biol.* **14**, 2252–2263 (2019).
260. Morris, K. L. *et al.* Cryo-EM of multiple cage architectures reveals a universal mode of clathrin self-assembly. *Nat. Struct. Mol. Biol.* **26**, 890–898 (2019).
261. Fleishman, S. J. *et al.* Rosettascripts: A scripting language interface to the Rosetta Macromolecular modeling suite. *PLoS One* **6**, 1–10 (2011).
262. Trott, O. & Olson, A. J. AutoDock Vina: Improving the Speed and Accuracy of Docking with a New Scoring Function, Efficient Optimization, and Multithreading OLEG. *J. Comput. Chem.* **31**, 455–461 (2010).
263. Drickamer, K. & Taylor, M. E. Glycan arrays for functional glycomics. *Genome Biol.* **3**, 10–13 (2002).
264. Fukui, S., Feizi, T., Galustian, C., Lawson, A. M. & Chai, W. Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. *Nat. Biotechnol.* **20**, 1011–1017 (2002).
265. Wang, D., Liu, S., Trummer, B. J., Deng, C. & Wang, A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat. Biotechnol.* **20**, 275–281 (2002).
266. Berger & Group. Eukaryotic Expression Facility at University of Bristol. 1–18 (2016).
267. Wessel, D. A Method for the Quantitative Recovery of Protein in Dilute Solution in the Presence of Detergents and Lipids. **143**, 141–143 (1984).
268. Gasteiger, E. *et al.* The Proteomics Protocols Handbook. *Proteomics Protoc. Handb.* 571–608 (2005) doi:10.1385/1592598900.
269. Delaglio, F. *et al.* NMRPipe : A multidimensional spectral processing system based on UNIX pipes. *J. Biomol. NMR* **6**, 277–293 (1995).
270. Vranken WF, Boucher W, Stevens TJ, Fogh RH, Pajon A, Llinas M, Ulrich EL, Markley JL, Ionides J, L. E. The CCPN data model for NMR spectroscopy: development of a software pipeline. *Proteins* **59**, 687–696 (2005).
271. Winter, G. Xia2: An expert system for macromolecular crystallography data reduction. *J. Appl. Crystallogr.* **43**, 186–190 (2010).
272. Adams, P. D. *et al.* PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 213–221 (2010).
273. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 486–501 (2010).
274. The PyMOL molecular graphics system, version 1.5.0.4 Schrodinger, LLC.
275. Winn, M. D. *et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **67**, 235–242 (2011).
276. Kabsch, W. XDS. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 125–132 (2010).
277. Vagin, A. & Teplyakov, A. Molecular replacement with MOLREP. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **66**, 22–25 (2010).
278. Nicholls, R. A., Long, F. & Murshudov, G. N. Low-resolution refinement tools in REFMAC5. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **68**, 404–417 (2012).
279. Vagin, A. A. *et al.* REFMAC5 dictionary: Organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**, 2184–2195 (2004).
280. Rambo, R. ScAtter. https://bl1231.als.lbl.gov/scatter/.
281. Petoukhov, M. V. *et al.* New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 (2012).

282. Franke, D. and Svergun, D. I. DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 (2009).

283. Volkov, V. & Svergun, D. I. Uniqueness of ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 (2003).

284. Svergun, D. I. Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.* **76**, 2879–2886 (1999).

285. Svergun, D. I. & Kozin, M. Automated matching of high- and low-resolution structural models. *J. Appl. Crystallogr.* **34**, 33–41 (2001).

286. Schneidman-Duhovny, D. Hammel, M., Tainer, J. A. & Sali, A. FoXS, FoXSDock and MultiFoXS: Single-state and multi-state structural modeling of proteins and their complexes based on SAXS profiles. *Nucleic Acids Res.* **44**, 424–429 (2016).

287. Schneidman-Duhovny, D. Hammel, M., Tainer, J. A. & Sali, A. Accurate SAXS profile computation and its assessment by contrast variation experiments. *Biophys. J.* **105**, 962–974 (2013).