UNIVERSITAT DE BARCELONA

Structural and functional analysis of natural protein-based inhibitors and their protease targets

Soraia Inês dos Reis Mendes

UNIVERSITAT DE BARCELONA

Departament de Bioquímica i Biologia Molecular
Facultat de Farmàcia I Ciències de L'alimentació

CONSELL SUPERIOR D'INVESTIGACIONS CIENTÍFIQUES
INSTITUT DE BIOLOGIA MOLECULAR DE BARCELONA

Departament de Biologia Estructural
Proteolysis Laboratory

Structural and functional analysis of natural protein-based inhibitors and their protease targets

Soraia Inês dos Reis Mendes
2022

UNIVERSITAT DE BARCELONA

FACULTAT DE FARMÀCIA I CIÈNCIES DE L'ALIMENTACIÓ

PhD programme in Biotechnology 2017-2022

Structural and functional analysis of natural protein-based inhibitors and their protease targets

Memòria presentada per Soraia Inês dos Reis Mendes per optar al títol de doctora per la universitat de Barcelona

Prof. Xavier Gomis-Rüth (Director)

Dr. Ulrich Eckhard (Co-Director)

Soraia Mendes

Prof. Josefa Badia Palacín (Academic Tutor)

Soraia Inês dos Reis Mendes
2022

To all those people that walked with me on this journey,

Valeu a pena? Tudo vale a pena
Se a alma não é pequena.
Quem quer passar além do Bojador
Tem que passar além da dor.
Deus ao mar o perigo e o abismo deu,
Mas nele é que espelhou o céu.

Fernando Pessoa

Acknowledgements

The elaboration of this thesis is the highlight of a five-years work and owns much to those that walked with me along this journey. I am therefore very grateful to all who supported me directly or indirectly. My many thanks to,

Professor Xavier Gomis-Rüth, my thesis supervisor, for all the support, for challenging me and for always believe in my work.

Ulrich, my thesis co-supervisor and friend, for all the nice words, advice, endless lessons and for the open and generous mentoring.

the Cri-3 Family, the family that welcomed me when I left my home. Among all, I specially thank to Theodoros Goulas for the mentoring at the beginning of my PhD studies; to Arturo and Cuppi for all the support and good conversations; to Tibi and Miguel Ângel, for being a loving and safe haven, and for the Spanish lessons. To Laura Mariño and Laura del Amo, for all the amazing moments we shared, for making me fell always welcome and for "españolizar me".

Inés, Andrea, Alex, Archadius, Didi, Danyhang, and Anna Rita for all the support and for giving colour to all the incredible moments we spend together.

À Cathy, à Sofia, ao Pedro e à Mafalda, amigos que mesmo à distância se mantiveram sempre presentes.

Por fim, quero agradecer a toda a minha família, a minha maior força e fonte incondicional de apoio. Um especial obrigado aos meus pais pela educação, pelo colo, pelo exemplo e pela força; à minha irmã pela ligação cúmplice, por toda a disponibilidade e pela alegria que traz à minha vida; ao João, por me fazer uma pessoa mais feliz, autêntica e segura; ao Jokinha e à avó Irene, pela presença e motivação contantes ao longo destes anos.

Abstract

Proteases were initially associated with indiscriminate digestion and degradation of dietary proteins and damaged proteins. However, the involvement of these enzymes in the specific, complex, and fine-tuned cleavage of target proteins became evident as a growing number of studies identified proteases as major players in a multitude of physiological processes. It is by now also well known that deregulation of such key molecules' activity is on the basis of several pathological conditions. Regulation of protease activity is essential for homeostasis, being attained by diverse mechanisms such as control of gene expression, protein compartmentalization, production of latent pro-forms (zymogens) that require subsequent activation, and by interaction with protein and peptide inhibitors. All forms of life are provisioned with a diverse repertoire of proteases and protease inhibitors working in a concerted and synergistic manner. Importantly, several proteases and their antagonistic protein or peptide inhibitors sparked researchers' interest due to their potential as therapeutic targets.

The present thesis unravels new information regarding protease and protease inhibitor structures and their mechanisms of action. To this end, this thesis compiles results obtained in three independent projects that addressed this topic focusing on distinct sets of proteases and inhibitors.

In the first project, we assessed the involvement of human reversion-inducing cysteine-rich protein with Kazal motifs (RECK) in embryogenesis and tumour suppression by investigating its involvement in the regulation of matrix metalloproteinases (MMPs). MMPs are proteases mainly involved in the degradation of protein constituents of the extracellular matrix (ECM) and are reported to be inhibited by three major protein inhibitors: α2-macroglobulin (α2M), tissue inhibitors of metalloproteinases (TIMPs) and RECK. Notably, abrogation or decrease of RECK levels is embryonically lethal for mice and causes increased tumour invasiveness and metastasis in a plethora of human cancers, with both phenotypes evincing profound disturbance of vascular structures. We developed bacterial and eucaryotic expression systems for the production of RECK variants and established an exhaustive purification protocol. The inhibitory activity of RECK variants towards MMP-2, MMP-7, MMP-9, and MMP-14 catalytic domain were assessed using fluorogenic peptides and natural protein substrates. Contrary to the published literature, no significant MMP

inhibition was detected, suggesting that RECK is not a direct inhibitor of MMPs activity (Mendes *et al.*, 2020).

In the second project, we aimed to develop a specific and potent inhibitor of aureolysin, a protease and key virulence factor secreted by the human pathogen *Staphylococcus aureus*. At present, we are witnessing the uncontrolled increase of antibiotic-resistant strains accountable for mild to life-threatening infections. *S. aureus*, one of those alarming pathogens, is equipped with several secreted proteases responsible for overcoming host defence molecules, one of which being aureolysin. Currently, there is no known specific inhibitor of aureolysin, but such molecules could be a valuable complement for antibiotic therapy. The insect metallopeptidase inhibitor (IMPI) is a unique low-molecular-weight defensive molecule produced by the greater wax moth *Galleria mellonella,* which effectively and potently inhibits thermolysins from some pathogenic bacteria. In the thesis we evaluated IMPI inhibition of aureolysin activity *in vitro* and explored its mechanism of action through analysis of the crystallographic structure of the complex. We conclude that the IMPI inhibition mechanism implies its proteolytic cleavage by the protease in the complex, an atypical mechanism of metallopeptidases inhibition. Furthermore, in an attempt to obtain a more effective or specific aureolysin inhibitor, we designed a set of twelve mutants displaying single or multiple point mutations in the reactive-centre loop. The best aureolysin inhibition was achieved by the $I^{57}F$ mutant, with a calculated inhibition constant (Ki) of 346 nM. Taking into account the lack of thermolysin-like peptidases, a family which includes aureolysin, in animals, the work presented here provides extra input for the development of therapeutic proteins and peptide-based inhibitors based on IMPI and its inactivation mechanism for the treatment of infections caused by antibiotic resistant pathogens (Mendes *et al.*, 2022).

In the third project, a highly collaborative work was developed in order to explore the unique inhibitory mechanism of human plasma α2M (hα2M). Hα2M is a glycosylated high-molecular weight homotetrameric protein of approximately 720 kDa, which is present in high concentrations in human plasma. Hα2M performs several functions including the transport of signalling molecules and inhibition of peptidases. Its unique mechanism of action, known as the "Venus fly-trap" mechanism, allows the non-specific inactivation of up to two protease molecules independently of their catalytic type. In this project we obtained eight α2M structures by cryo electron microscopy (cryo-EM) (of both native and induced states) demonstrating that the flexible native protein displays an open conformation that changes to a closed compact structure upon induction by the cleaving and consequently

entrapped peptidase, exposing the receptor binding domain and ultimately leading to the removal of the complex from circulation (Luque *et al.*, 2022). Sparked by remarks during the review process of this paper, we then compared the function and biophysical properties of hα2M purified from fresh plasma with that of thawed frozen plasma through a range of experiments, providing light on this for the very first time, and demonstrating, that both hα2M preparations are indistinguishable (Mendes *et al.*, in preparation).

Abstract (Spanish)

Las proteasas han estado asociadas desde los inicios de la bioquímica a la digestión indiscriminada y degradación de proteínas nutricionales y/o defectuosas. Sin embargo, la participación de estos enzimas en la escisión específica, compleja y precisa de sustratos proteicos diana concretos se hizo evidente a medida que incrementaban el número de estudios. A la par, varios de estos estudios evidenciaron el papel de las proteasas como actrices principales en multitud de procesos fisiológicos. Actualmente se sabe que la desregulación de la actividad de estas moléculas clave es consecuencia de varias condiciones patológicas. La regulación de la actividad de las proteasas es esencial para la homeostasis y se logra mediante diversos mecanismos, como el control de la expresión génica, la compartimentación de proteínas, la producción de proformas latentes (zimógenos) que requieren una activación posterior y la interacción con inhibidores de proteínas y péptidos. Todas las formas de vida cuentan con un repertorio diverso de proteasas e inhibidores de estas que funcionan de manera concertada y sinérgica. Es importante destacar que varias proteasas y sus inhibidores de péptidos o proteínas antagonistas han venido despertando el interés de investigadores debido a su gran potencial como dianas terapéuticas.

La presente tesis presenta nueva información sobre las estructuras de proteasas e inhibidores de proteasas y sus mecanismos de acción. Esta tesis recopila los resultados obtenidos en tres proyectos que abordan este tema centrándose en distintas proteasas e inhibidores.

En el primer proyecto, revisamos la participación de la proteína “reversion-inducing cysteine-rich protein with Kazal motifs” (RECK) en la embriogénesis, la supresión de tumores y su implicación en la regulación de las metaloproteinasas de matriz (MMP). Las MMPs son proteasas involucradas principalmente en la degradación de los componentes proteicos de la matriz extracelular (ECM) y se ha descrito que son inhibidas por tres tipos de proteínas: la α2-macroglobulina (α2M), los inhibidores tisulares de metaloproteinasas (TIMP) y RECK. En particular, la eliminación o disminución de los niveles de RECK provoca la muerte de embriones de ratón y también una mayor invasión tumoral y metástasis en una plétora de cánceres humanos. Ambos fenotipos están asociados con la alteración profunda de estructuras vasculares. En el marco de esta tesis desarrollamos sistemas de expresión bacterianos y eucariotas para la producción de variantes de RECK y establecimos un protocolo de purificación exhaustivo. Se evaluó la actividad inhibidora de las variantes

de RECK frente a MMP-2, MMP-7, MMP-9 y al dominio catalítico de MMP-14 utilizando péptidos fluorigénicos y proteínas naturales como sustratos. Durante los estudios realizados, no se detectó inhibición significativa de las MMP, lo que claramente indica que RECK no es un inhibidor directo de la actividad de MMPs (Mendes *et al.*, 2020).

En el segundo proyecto, nuestro objetivo fue desarrollar un inhibidor específico y potente de la aureolisina, una proteasa secretada por el patógeno humano *Staphylococcus aureus*. En la actualidad, se está observando el aumento descontrolado de cepas resistentes a los antibióticos responsables de infecciones leves o potencialmente mortales. *S. aureus*, uno de esos patógenos, está equipado con varias proteasas secretadas que anulan los mecanismos de defensa del huésped, una de las cuales es la aureolisina. No se conoce ningún inhibidor específico de la aureolisina, que sin duda sería un valioso complemento para la terapia con antibióticos. El inhibidor de metalopeptidasas de insectos (IMPI) es una molécula defensiva única de bajo peso molecular producida por la polilla de la cera *Galleria mellonella*, que inhibe eficaz y potentemente las termolisinas de algunas bacterias patógenas, una familia de metalopeptidasas a la que pertenece la aureolisina. En la tesis evaluamos la inhibición de la actividad de la aureolisina por parte de IMPI *in vitro* y también exploramos su mecanismo de acción a través del análisis de la estructura cristalográfica del complejo. Concluimos que el mecanismo de inhibición implica la escisión proteolítica por parte de la proteasa del inhibidor dentro del complejo, un mecanismo de inhibición atípico de metalopeptidasas. Asimismo, en un intento por obtener un inhibidor de aureolisina más eficaz o específico, diseñamos un conjunto de doce mutantes diferentes con mutaciones puntuales únicas o múltiples en el bucle del centro reactivo. La mejor inhibición de aureolisina se logró con el mutante $I^{57}F$, con una constante de inhibición (Ki) de 346 nM. Teniendo en cuenta la ausencia de termolisinas en animales, el trabajo aquí presentado podría sentar las bases para el desarrollo de proteínas y péptidos terapéuticos basados en IMPI para el tratamiento de infecciones causadas por patógenos resistentes a los antibióticos (Mendes *et al.*, 2022).

En el tercer proyecto, se desarrolló un trabajo colaborativo para explorar el mecanismo inhibidor único de la α2M de plasma humano (hα2M). La hα2M es una proteína homotetramérica glicosilada de alto peso molecular (720 kDa) presente en altas concentraciones en plasma humano. Realiza varias funciones, incluidos el transporte de moléculas de señalización y la inhibición de peptidasas. Su mecanismo de acción, conocido como “trampa Venus”, permite la inactivación inespecífica de hasta dos moléculas de proteasa, independientemente del tipo catalítico. En este proyecto, obtuvimos ocho

estructuras de α2M mediante criomicroscopía electrónica (cryo-EM), tanto de estados nativos como inducidos, que demuestran que la proteína nativa es muy flexible y muestra una conformación abierta que cambia a una cerrada tras la inducción por la peptidasa atrapada, exponiendo el dominio de unión al receptor (Luque *et al.*, 2022). Impulsados por los comentarios durante la revisión del manuscrito, procedimos posteriormente a comparar las propiedades biofísicas y funcionales de hα2M purificada a partir de plasma fresco no congelado con proteína obtenida a partir de plasma congelado (Mendes *et al.*, en preparación).

Table of contents

Index of figures and tables

Abbreviations and acronyms

$(h\alpha_2M)_4$: tetrameric human $h\alpha_2M$

α1I3: α1-inhibitor-3

α2M: alpha 2-macroglobulin

3D: three-dimensional

a.u.: asymmetric unit

AC: affinity chromatography

ADAM: a disintegrin and metalloproteinase

ADAMTS: a disintegrin and metalloproteinase with thrombospondin motifs

AEBSF: 4-[2-aminoethyl]benzenesulfonyl fluoride

AF: AlphaFold

AMPs: antimicrobial peptides and proteins

APMA: 4-aminophenylmercuric acetate

Aur: aureolysin

BRD: bait-region domain

BSA: bovine serum albumin

CD: catalytic domain

CHO: chinese hamster ovary

CP: cytoplasmic

CPAMD8: PZP-like α2M domain-containing 8

CRR: Cysteine Rich Regions

Cryo-EM: cryo–electron microscopy

CSD: C-terminal subdomain

CTS: C- terminal segment

CUB domain: complement protein subcomponents C1r/C1s, urchin embryonic growth factor, and bone morphogenetic protein domain

DTNB: 5,5'-dithiobis-2-nitrobenzoic acid

ECAM: *Escherichia coli* α_2-macroglobulin

ECM: extracellular matrix

EDTA: ethylenediamine tetraacetic acid

EGF: epidermal growth factor

FFP: fresh frozen plasma

FGF23: fibroblast growth factor 23

FN: fibronectin

FSC: Fourier shell correlation

FTP: fungalysin-thermolysin-pro-peptide

GPCR: G-protein-coupled receptors

GPI: glycosyl-phosphatidyl inositol

HEK: human embryonic kidney

HER-2/neu: human epidermal growth factor receptor 2

His_6-tag: hexahistidine-tag

HLA: human leukocyte antigen

HRP: horseradish peroxidase

IEC: ion exchange chromatography

IMAC: immobilised-metal affinity chromatography (

IMPI: insect metalloproteinase inhibitor

IPTG: isopropyl-β-D-1-thiogalactopyranoside

ISPIs: inducible serine proteinase inhibitors

IUBMB: International Union of Biochemistry and Molecular Biology

KL: Kazal-like

KN- cysteine knot

LB: lysogeny broth / Luria-Bertani

MA: methylamine

MALDI-TOF: *matrix-assisted laser desorption/ionization - time-of-flight*

MG domain: macroglobulin-like domain

miRNA: microRNAs

MMPs: matrix metallopeptidases

MPs: metalloproteases

MRSA: methicillin-resistant *S. aureus*

MT-MMP: membrane-type MMP

MW: molecular weight

N-TES: N-terminal region of human testican 3

Ni-NTA: affinity chromatography

NSD: N-terminal subdomain

o-Phe: o-phenanthroline

OD: optical density

PBS-T: PBS-Tween

PBS: phosphate buffered saline

PCR: polymerase chain reaction

PD: prodomain

PDB: protein data bank

PEG: polyethylene glycol

PEPs: prolyl endopeptidases

PEX: hemopexin-like domain

pFN: plasma fibronectin

pLDDT: per-residue confidence score

PMF: peptide-mass fingerprinting

PMSF: phenylmethanesulfonyl fluoride

PPC: proprotein convertase

PRR: Proline- Rich Regions

PTM: post-translational modification

PZP: pregnancy zone protein

RAS: rat sarcoma

RBD: receptor binding domain

RCL: reactive-centre loop

RECK: reversion-inducing-cysteine-rich protein with Kazal motifs

RMSD: root mean square deviation

S2: Schneider 2 cells

ScpA: staphopain A

SDS-PAGE: sodium dodecyl sulfate polyacrylamide gel electrophoresis

SDS: sodium dodecyl sulphate

SEC-MALLS: multi-angle laser light scattering

SEC: size exclusion chromatography

SNP: single nucleotide polymorphisms

SP: signal peptide

SP1: specificity protein 1

Spl: serine protease-like proteins

SspA: serine protease V8

SspB: staphopain B

TCEP: tris(2-carboxyethyl)phosphine

TED: thioester domain

TEPs: thioester-containing proteins

TEV: tobacco etch virus

TGAT: trio-related transforming gene in ATL tumour cells

TIL: trypsin inhibitor-like

TIMPs: tissue inhibitors of metalloproteinases

TLP-ste: TLP from *Bacillus stearothermophilus*

TLPs: thermolysin-like proteases

TM: transmembrane

TP: therapeutic proteins and peptides

UP: UniProt

WT: wild type

XMMP: MMP from *Xenopus laevis*

List of amino acids with respective one and three letter codes ("Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983," 1984)

Amino acid name	One letter code	Three letter code
Alanine	A	Ala
Arginine	R	Arg
Asparagine	N	Asn
Aspartic acid	D	Asp
Cysteine	C	Cys
Glutamine	Q	Gln
Glutamic acid	E	Glu
Glycine	G	Gly
Histidine	H	His
Isoleucine	I	Ile
Leucine	L	Leu
Lysine	K	Lys
Methionine	M	Met
Phenylalanine	F	Phe
Proline	P	Pro
Serine	S	Ser
Threonine	T	Thr
Tryptophan	W	Trp
Tyrosine	Y	Tyr
Valine	V	Val
Any amino acid	X	Xaa
termination codon		TERM

Introduction

1. Proteolytic enzymes

1.1. Brief historical contextualization

The Human Genome Project revealed that, contrary to expectations, only approximately 2 % of the human genome correspond to protein- or RNA-coding genes, amounting to a total of ~20,000 genes (Moraes & Góes, 2016; Nurk *et al.*, 2022)). The awareness that about 50% of the human coding genes present high similarity with those from other organisms and that genome sizes might be identical despite of the organism's complexity, dethroned DNA as key molecule responsible for complexity and variability. New significance was therefore awarded to proteins, the "one gene, one protein" dogma was ousted, and the pivotal role undertaken by proteins on the complexity and variability of all living organisms unravelled. For instance, one single gene could potentially give rise to 100 different protein variants with slightly different functions (lately reviewed by Ezkurdia *et al.*, 2014; Moraes & Góes, 2016; Ponomarenko *et al.*, 2016).

The term "protein", coined by Mulder in 1838 (Vickery, 1950), is a general and non-exquisite classification that encompasses a tremendous number of macromolecules. Enzymes are a large and relevant class of proteins, which speed up chemical reactions (Alberts *et al.*, 2002). According to the International Union of Biochemistry and Molecular Biology (IUBMB), these catalysts are grouped into seven classes (oxidoreductases (EC 1), transferases (EC 2), hydrolases (EC 3), lyases (EC 4), isomerases (EC 5), ligases (EC 6) and translocases (EC 7)) according to the chemical reaction they exert (Jeske *et al.*, 2019; McDonald *et al.*, 2009). Depicting 2% of the human coding genes, peptidases are one of the largest human enzyme classes with 588 putative representatives known to date (Pérez-Silva *et al.*, 2016; Quesada *et al.*, 2009; Rawlings, 2020). Peptidases (EC 3.4), also called proteases, proteinases, or proteolytic enzymes in an interchangeable manner, are hydrolases responsible for cleavage of peptide bonds in peptide or protein substrates (Barrett & Rawlings, 2007).

The historical development of scientific breakthroughs leading to enzyme and peptidase knowledge is very intriguing (**Figure 1**). Although the discovery of the first protease, pepsin, by William Beaumont's and Theodor Schwann took place already in 1836 (Cushing, 1935), the term "enzyme" was only coined almost 40 years later by Wilhelm Kühne (Gutfreund, 1976). More interestingly, the association of catalytic activity with

protein molecules was only established in 1926 when James Sumner achieved urease crystallisation (Simoni *et al.*, 2002).

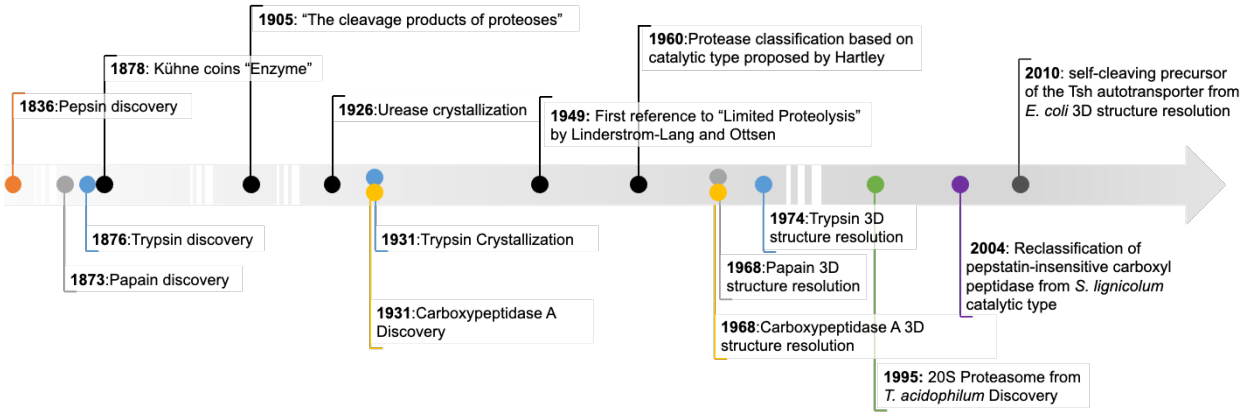


Figure 1: Historic timeline of proteolytic enzyme research. Chronological representation of protease history milestones, including the dates of discovery, crystallisation and three-dimensional (3D) structure resolution of the first known representatives of aspartic (pepsin), cysteine (papain), serine (trypsin), and metallopeptidases (carboxypeptidases A) along with the recently discovered threonine (20S Proteasome from T. acidophilum) and glutamic (pepstatin-insensitive carboxyl peptidase from S. lignicolum) peptidases and asparagine lyases (self-cleaving precursor of the Tsh autotransporter from E. coli). Urease (EC 3.5) is included as it was the very first enzyme reported to be crystalized.

1.2. The catalytic site

The active site of a protease is the region where the binding of the substrate and subsequent hydrolysis of its peptide bonds occurs. Protease-substrate binding is established by hydrogen bond interactions between the protease and the substrate peptide backbone and through hydrophobic and electrostatic contacts of substrate amino-acid side chains (P, P' residues) with protease substrate pockets (S, S' pockets) (**Figure 2**) (Deu *et al.*, 2012; Klein *et al.*, 2018). In agreement with the nomenclature established by Schechter & Berger (1967), the substrate residues and their cognate protease substrate pockets upstream (P, S) and downstream (P', S') of the scissile bond are numbered according to their relative position to the cleavage site. The permissiveness of a protease to bind to multiple or limited amino acid residues on the substrate determines its specificity (Schauperl *et al.*, 2015). On top of that, various proteases possess exosites, interaction surfaces without catalytic activity, that are crucial for regulation of specificity and catalytic efficiency of physiological substrates (Bock *et al.*, 2007).

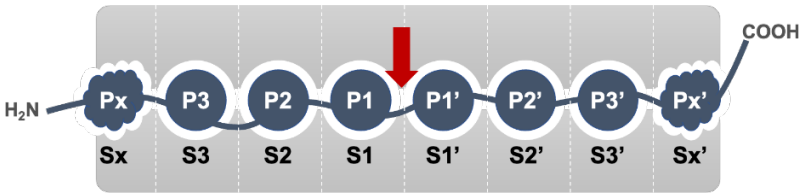


Figure 2: Schematic representation of a putative enzyme-substrate complex and its nomenclature. The enzyme active site (in grey) comprises a variable number of subsites (Sx to S1 and S1' to Sx') which accommodate the substrate amino-acid chains (likewise, Px to P1 and P1' to Px'), with cleavage occurring between the P1 and P1' residue. Non-primed and primed positions are upstream and downstream of the scissile bond (indicated by the red arrow), respectively.

1.3. Classification of proteases

A high number of peptidases was discovered since the report "The cleavage products of proteoses" in the Journal of Biological Chemistry in 1905 (Levene), leading to more than one million amino-acid sequences currently associated with peptidases in MEROPS (Rawlings *et al.*, 2018). Initially, peptidases were divided into endopeptidases, responsible for cleavage of internal peptide bonds, and exopeptidases, acting near or at the ends of polypeptide chains ─ either at the amino (aminopeptidases) or carboxyl (carboxypeptidases) terminus of the substrate (Barrett, 1994). While not being common, some proteolytic enzymes might function as both endo- and exopeptidases (Turk, 2006).

A more systematic classification of proteolytic enzymes is based on their catalytic mechanism **(Figure 3)**. Seven catalytic types are known to date, allowing separation into serine, cysteine, aspartic, threonine, metallo- and glutamic peptidases (Hartley, 1960; Kataoka *et al.*, 2005; Rawlings & Barrett, 1993; Seemüller *et al.*, 1995) and asparagine lyases (Rawlings *et al.*, 2011). It is important to emphasise that asparagine lyases do not fall within the definition of peptidases and hydrolases as they do not utilize a water molecule during peptide bond hydrolysis, these class members are nevertheless proteolytic enzymes. And while cysteine, serine and threonine peptidases catalyse the peptide cleavage through nucleophilic attack of their homonymous active site side chain residue and subsequent hydrolysis, the aspartic, glutamic and metallopeptidases draw on an activated water molecule as a nucleophile. In the case of asparagine lyases an asparagine residue is the nucleophile responsible for self-cleavage (Rawlings *et al.*, 2011).

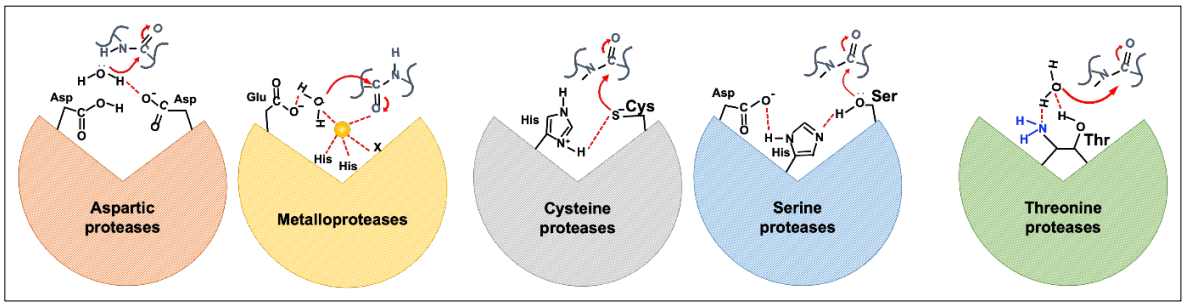


Figure 3: Simplified representation of the catalysis mechanism for each human protease class. Serine, cysteine, and threonine residues are responsible for the nucleophilic attack of the cognate protease class while in aspartic and metalloproteases this role is exerted by an activated water molecule. Metalloproteases comprise a divalent ion (yellow sphere) in their active site. Glutamic peptidases present an active site similar to aspartic peptidases but with a catalytic dyad of two glutamates or one glutamate and one glutamine. In threonine peptidases the α-amine of the protein N-terminus (shown in blue) complements the active-site dyad. Adapted from (L. Wang *et al.*, 2021).

Categorization of proteolytic enzymes based on their catalytic type is the foundation of their hierarchical organization into databases such as MEROPS (Rawlings *et al.*, 2016). Proteases presenting similar amino acid sequences are clustered into families where amino acid sequence similarity might be observed at the whole protein level or within the sequence of the catalytic domain (Rawlings, 2013; Rawlings & Barrett, 1993). Members of a protease family presenting high similarity within important features, such as substrate binding or biological function, are then subdivided into so-called peptidase "species" (Barrett & Rawlings, 2007). The best characterised peptidase from such a peptidase species is called "*holotype"* (Rawlings, 2016; Rawlings *et al.*, 2016). Peptidase families are further clustered into "clans" when an evolutionary relation could be appreciated by the three-dimensional structure and by the linear order of the residues from the catalytic sites (Rawlings *et al.*, 2016). The complete repertoire of peptidases expressed by an organism or tissue is entitled *"degradome"* (Pérez-Silva *et al.*, 2016).

One particular organism does not necessarily have proteolytic enzymes of all catalytic types **(**Figure 4**)**. For instance, human and mouse degradomes comprise aspartic, cysteine, threonine, serine, and metalloproteases, the last two being the classes with more representatives (Quesada *et al.*, 2009). Glutamic peptidases are mainly found in fungi but also on archaea and bacteria (K. Jensen *et al.*, 2010; Sims *et al.*, 2004), just as asparagine lyases that were additionally detected on viruses (Rawlings *et al.*, 2011). Furthermore, aspartic peptidases are mainly encoded by fungi (Nguyen *et al.*, 2019). Moreover, from over the 270 protease families hitherto identified, none has homologues in species from every

phylum of organisms, except some who undertake essential housekeeping functions, such as methionyl aminopeptidase for N-terminal methionine excision and signal peptidases for the removal of targeting signals (Rawlings & Bateman, 2019).

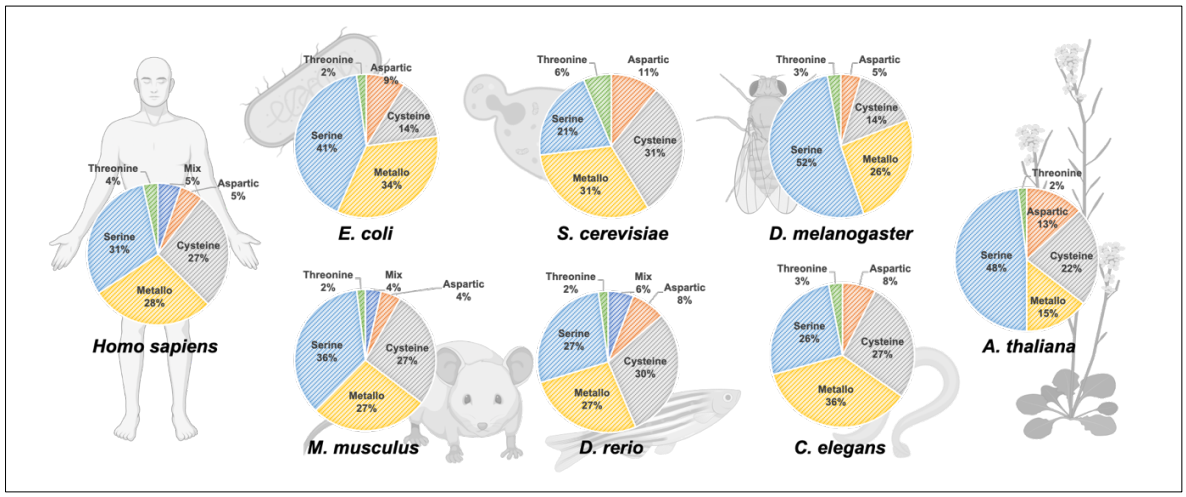
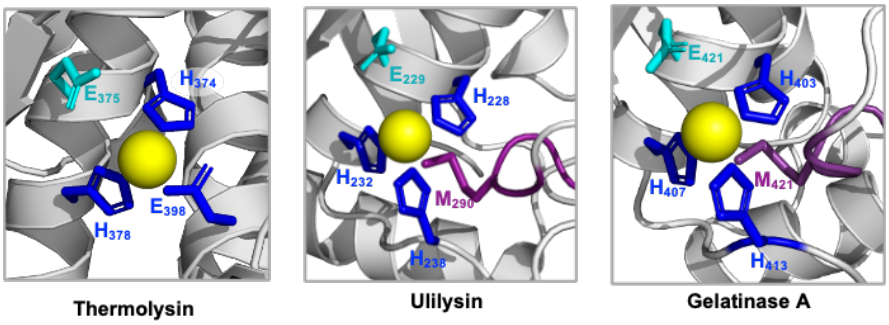


Figure 4: Graphical representation of peptidase classes constituting the degradomes of different model organisms. Serine, metallo-, cysteine, aspartic and threonine proteases are shown in light blue, yellow, grey, orange, and green, respectively. Mix (shown in dark blue) stands for peptidases with protein nucleophiles of mixed catalytic type. Data from (Rawlings, 2020).

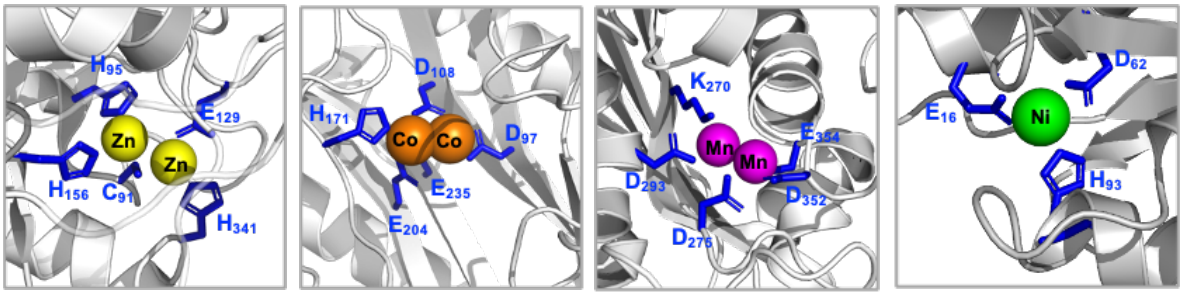
1.4. Metalloproteases

The metalloproteases (MPs) are one of the protease classes with more representatives in all forms of life, together with serine and cysteine proteases **(**Figure 4**)**. The diverse members of this class harbour at least one divalent metal ion, typically a zinc and less frequently cobalt, manganese or nickel, in their active site **(Figure 5** and Figure 6**).** The active site metal ion is coordinated by at least one water molecule, the ultimate responsible for the hydrolysis of the scissile peptide bond, and three amino acid side chains of peptidase residues. The most common metal binding residues are histidines, followed by glutamates, aspartates, and lysines (Cerdà-Costa & Gomis-Rüth, 2014; Klein *et al.*, 2018; Vallee & Auld, 1990). An additional residue, typically a glutamate, which acts as a general base/acid in the proteolytic mechanism of MPs, is also in the neighbourhood of the catalytic metal, required to activate the catalytic water (**Figure 5**).



Thermolysin **Ulilysin** **Gelatinase A**

Figure 5: Schematic representation of zinc metallopeptidases catalytic sites. The catalytic sites of the M4 holotype thermolysin (from *B. thermoproteolyticus*; UP P00800; PDB 1kei), the pappalysin-like protease ulilysin (from *M. acetivorans*; UP Q8TL28; PDB 2cki; Tallant *et al.*, 2006) and the human matrixin gelatinase A (UP P08253; PDB 3ayu; (Hashimoto *et al.*, 2011) are shown with the catalytic zinc ion displayed as a yellow sphere. Residues engaged in ion binding and the general base/acid residue are shown in dark and light blue, respectively. The Met turn motif and its methionine residue are highlighted in purple. All structural models were prepared in PyMOL (*The PyMOL Molecular Graphics System*, n.d.).



Peptidase M20 domain-containing protein 2 *Staphylococcus aureus*

Methionine aminopeptidase *Escherichia coli*

Cytosol aminopeptidase *Escherichia coli*

Hydrogenase 2 maturation protease *Escherichia coli*

Figure 6: Illustrative metallopeptidase catalytic sites harbouring one or two distinct divalent metal ions. *S. aureus* Peptidase M20 domain-containing protein 2 (UP A0A0H3JAV7; PDB 3ram; Botelho *et al.*, 2011) and *E. coli* methionine aminopeptidase (UP.P0AE18; PDB 1mat; Roderick & Matthews, 1993) and cytosol aminopeptidase (UP P68767; PDB 1gyt; Colloms, 2004; Strater, 1999) catalytic sites comprise two zinc, cobalt, and manganese ions (yellow, orange, and pink spheres), respectively. *E. coli* Hydrogenase 2 maturation protease (UP P37182; PDB 1cfz; Fritsche *et al.*, 1999) catalytic site contains one nickel ion (green sphere). Peptidase residues coordinating the catalytic metal ions are represented in blue. All structural models were prepared in PyMOL (*The PyMOL Molecular Graphics System*, n.d.).

Metallopeptidases are systematically organized and classified according to their characteristic metal binding motifs and additional structural features (**Figure 7**). They are allocated into two subclasses, the mononuclear peptidases with a single catalytic metal ion, and the dimetalate peptidases displaying two catalytic ions in their active site. These subclasses are then further subdivided into tribes (Cerdà-Costa & Gomis-Rüth, 2014).

For example, zinc metalloproteases, which comprise mostly mononuclear peptidases, gather the zincin, inverzincin and αβα-exopeptidase tribes. Members of the *zincin* peptidase tribe are characterised by the short zinc binding motif **HExxH** while *inverzincins* present a consensus sequence with identical residues in inverted order, **H**xx**EH**. In turn, zincins split into two main clans, the *gluzincins*, whose third zinc ligand is a glutamic acid (E), and the *metzincins*, which presents a histidine as the third zinc binding residue and a characteristic methionine-containing 1,4-β-turn (Met-turn) responsible for their designation **(Figure 5 and Figure 7)** (Bode *et al.*, 1993; Fushimi *et al.*, 1999; Gomis-Rüth *et al.*, 2012; Hooper, 1994). Finally, peptidase clans are formed by families, which represent the most specific aggregating units.

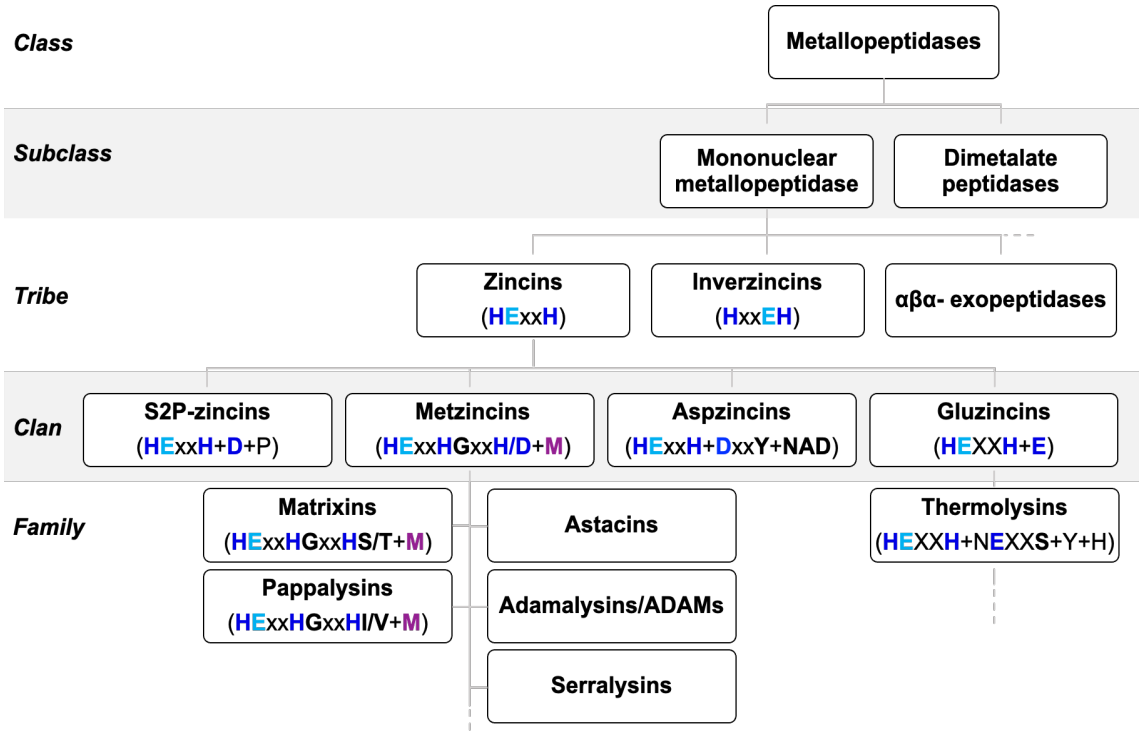


Figure 7: Diagrammatic representation of zinc metallopeptidases classification. The suggested tree is structure-based and comprises five hierarchical levels: class, subclass, tribe, clan, and family. Consensus motifs comprising residues responsible for zinc ion binding and catalysis (residues acting as a general base/acid) are shown in bold and dark and light blue, respectively. The methionine residue from the Met-turn consensus sequence of metzincins is highlighted in bold purple letter. Adapted from (Cerdà-Costa & Gomis-Rüth, 2014).

1.4.1. The MMP family

Matrix metallopeptidases (MMPs), interchangeably called matrixins, were initially identified sixty years ago as collagenolytic agents during amphibian metamorphosis (Gross & Lapiere, 1962). They constitute an important family of the metzincin clan of MPs (M10A subfamily of the MEROPS database) and are found in all kingdoms of life, being primarily engaged in the degradation of extracellular-matrix (ECM) proteins, as their name implies. The ECM is a non-cellular three-dimensional network formed by an agglomerate of various macromolecules. Fibrous-forming proteins, namely collagens, elastin, laminin and fibronectin (FN), together with proteoglycans and many other glycoproteins, like tenascin and vitronectin, are the main constituents of the ECM (Laronha *et al.*, 2020; Theocharis *et al.*, 2016).

The members of this family are chiefly found in vertebrates and were denoted with MMP numbers following their order of discovery in addition to their individual trivial names (**Table 1**) (Nagase *et al.*, 2006). In humans, the MMP family encompasses 23 members, which are expressed in different tissues and recognize and cleave different substrates (**Table 1** and **Table 2**) (Cui *et al.*, 2017). Their general architecture is well conserved comprising a signal peptide (SP) followed by a pro-peptide for zymogenic latency, a catalytic domain, a linker (hinge) region and a hemopexin-like domain (PEX) (**Figure 8**) (Pulkoski-Gross, 2015). The SP is a typical feature of secreted proteins, which is removed during secretion/export from the cell and is responsible for ensuring proper protein trafficking. In the case of MMPs, this entails translocation to the extracellular space (Nagase *et al.*, 2006; Pulkoski-Gross, 2015).

MMPs are expressed as latent zymogens, which are activated thought the removal of their pro-peptide (or pro-domain) after secretion. The pro-domain is composed of three α-helices (α1–α3) lending stability to this domain and of flexible connecting loops, which are highly prone to proteolytic cleavage (**Figure 9A**) within a protease susceptible "bait region" located between α1 and α2 (Jozic *et al.*, 2005; Nagase *et al.*, 2006). This inhibitory domain encloses a highly conserved PRCGXPDV motif containing a cysteine residue which is responsible for the for the "cysteine-switch" mechanism of inhibition (van Wart & Birkedal-Hansen, 1990). The thiol group (-SH) of the cysteine residue (C) binds to the catalytic zinc ion precluding the binding of the active-site water molecule responsible for the nucleophilic attack onto the scissile-bond carbonyl group (van Wart & Birkedal-Hansen, 1990). The

adjacent arginine (R) and aspartate (D) residues form a salt bridge that is essential for the stability of the cysteine-zinc interaction (Galazka *et al.*, 1996; Suzuki *et al.*, 1990).

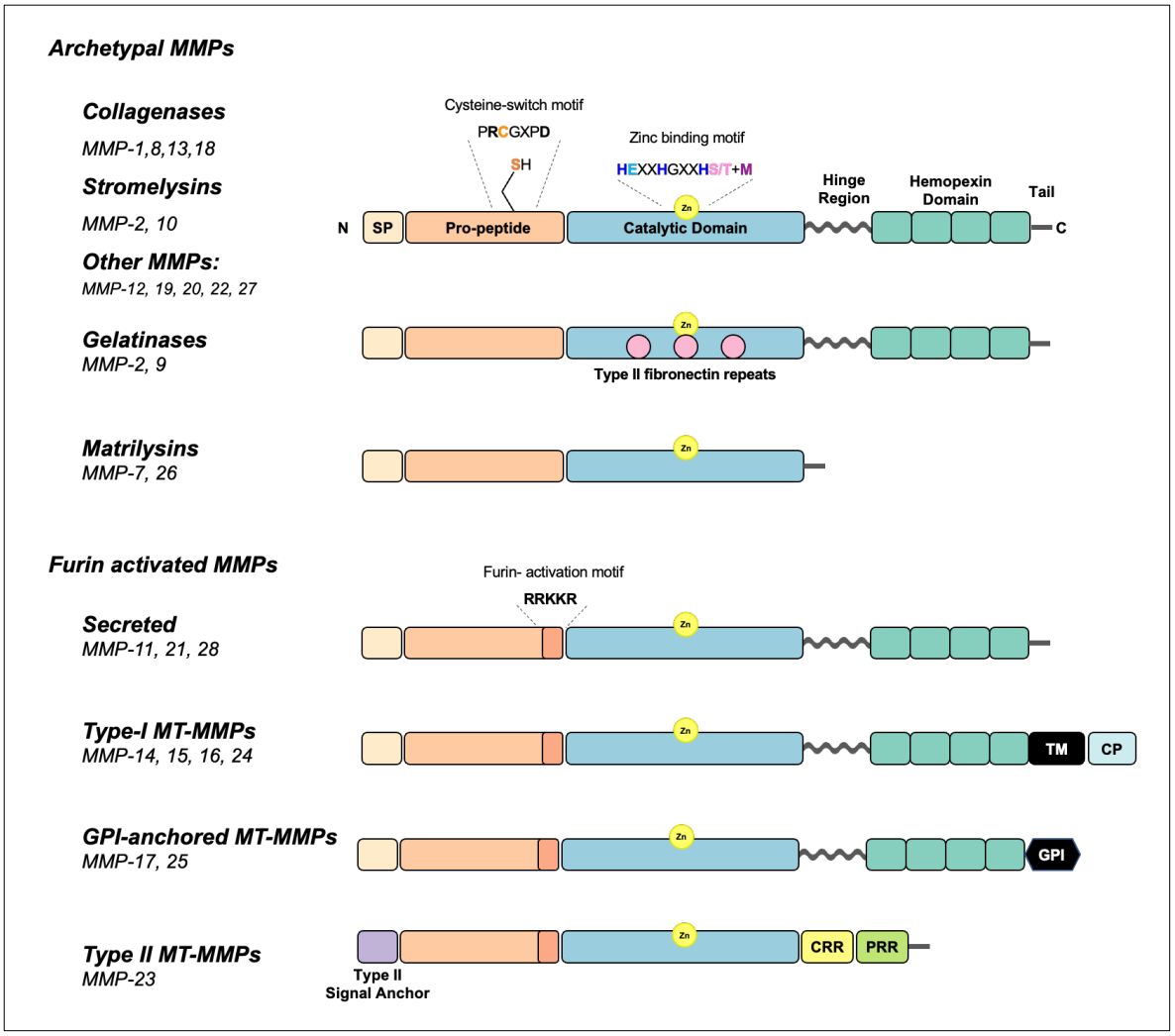


Figure 8: Schematic representation of the matrixins general architecture. The signal peptide (SP), pro-peptide, catalytic and hemopexin-like domains linked by the hinge region are conserved in all MMPs except in matrilysins which lack both hinge and hemopexin domains. The additional fibronectin-like, furin recognition, transmembrane (TM) and cytoplasmic (CP) domains, the alternative Type-II Signal Anchor, and the Cysteine- and Proline- Rich Regions (CRR and PRR, respectively), which are characteristic for some MMPs, are also represented. The cysteine-switch, zinc-binding and the furin activation consensus motifs protrude from the respective domains and their key residues are highlighted in bold and coloured letters. Adapted from (Cui et al., 2017; Parks et al., 2004; Pulkoski-Gross, 2015).

The catalytic domain of MMPs is very well conserved, as highlighted by the good superimposition of the various structures. It shows a flattened ellipsoidal shape, with the active-site cleft traversing the domain in extended horizontal arrangement, which enables binding of a peptide substrate from left to right according the so-called “standard orientation” (Nagase *et al.*, 2006). The

catalytic domain is constituted by three α-helices (α1–α3) and a five-stranded twisted β-sheet (β1–β5) **(Figure 9B)**. The loops connecting β-strands coordinate non-catalytic ions, specifically a second zinc ion and up to three calcium ions, which confer structural stability (Bode *et al.*, 1999; Tallant *et al.*, 2010)

The active-site cleft separates the catalytic domain in two asymmetric parts, the larger N-terminal subdomain ("upper" subdomain) and the smaller C-terminal subdomain ("lower" subdomain). The first half of a zinc-binding consensus sequence (**HE**xx**H**xx**G**xx**H**), typical of MMPs, is part the helix α2, which ends in a turn allowed by the glycine (**G**) residue. This turn is fundamental to allow the contact of the third histidine residue with the catalytic zinc ion. Downstream in the sequence is another characteristic structural feature, the Met-turn loop, which is found in MMPs and other members of the metzincin clan (**Figure 7**) (Nagase *et al.*, 2006).

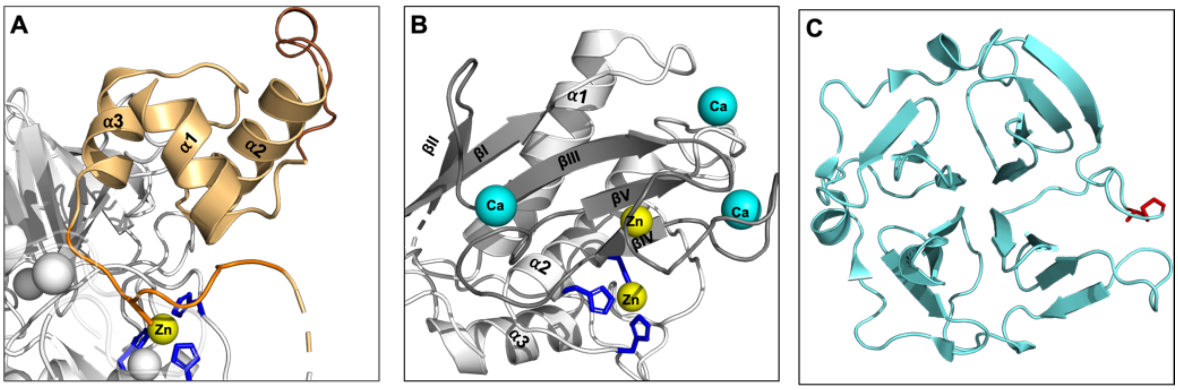


Figure 9: Schematic representation of MMPs pro-peptide (A), catalytic (B) and hemopexin-like (C) domains. (**A**) Human MMP-1 pro-peptide structure (shown in light orange; UP P03956; PDB 1su3; Jozic *et al.*, 2005) comprises three α-helices (α1–α3) connected by flexible loops. Pro-peptide key features as the MMP bait region and the cysteine switch motif are show in brown and dark orange, respectively. (**B**) MMP-1 catalytic domain structure (in grey) is constituted by three α-helices (α1–α3), five β-strands (β1–β5) and the connecting loops harbouring zinc and calcium ions (yellow and cyan spheres, respectively). MMP-1 residues coordinating the catalytic zinc ions are shown in dark blue. (**C**) MMP-1 hemopexin-like domain presents a four-bladed β-propeller structure (light blue) bordered by two cysteine residues forming a disulphide bridge (show in red). All structural models were prepared in PyMOL (*The PyMOL Molecular Graphics System*, n.d.).

Variations to the catalytic domain structure previously described are embodied by the membrane-type MMPs (MT-MMPs) catalytic domains, which harbour an insertion of eight extra residues in the loop connecting two strands of the β-sheet, known as the MT-loop, and by the gelatinases MMP-2 and MMP-9 catalytic domains, which accommodate three fibronectin-like repeats (Maskos, 2005; Nagase *et al.*, 2006; Tallant *et al.*, 2010).

A hinge region links the catalytic and hemopexin domains. The length of the linker varies in the different MMPs, ranging from 15 to 65 amino acids. This proline-rich region interacts with both catalytic and hemopexin-like domains assuring proper catalysis and specificity (Fasciglione *et al.*, 2012; Iyer *et al.*, 2006; Knäuper *et al.*, 1997).

The hemopexin-like domain (PEX) presents a characteristic four-bladed β-propeller structure entrenched by a disulphide bond formed between cysteine residues located at the N- and C-termini of the PEX domain (**Figure 9C**). This domain is involved in substrate recognition and specificity, being crucial for collagenolytic activity (Iyer *et al.*, 2006; Overall, 2002).

Moreover, MMPs might contain additional distinctive features such as the furin-recognition motif located between the pro- and catalytic domains, which is found in several representatives, or the exclusive vitronectin-like domain present in the MMP from *Xenopus laevis* (XMMP), which is flanked by the pro-peptide and the furin recognition domain. Surprisingly, this additional domain has no equivalent in the human ortholog, *viz*. MMP-21 (Ahokas *et al.*, 2002; Pei & Weiss, 1995).

Furthermore, certain MMPs are tethered to cell membranes, either by a transmembrane (TM) domain followed by a cytoplasmic (CP) domain or by a glycosyl-phosphatidyl inositol (GPI) anchor (**Figure 8**) (Itoh, 2015; Sohail *et al.*, 2008). An exception to this is MMP-23, which has a type-II transmembrane domain. MMP-23 is unique in that it lacks the signal peptide and contains a shorter pro-peptide, in which the key cysteine residue is found within an ALCLLPA motif. In addition, it encompasses a cysteine-array region (CRR) and a proline-rich immunoglobulin-like domain (PRR) downstream of the catalytic domain (Pei *et al.*, 2000; Velasco *et al.*, 1999).

Table 1: Members of the MMP Family: Nomenclature and Tissue Distribution. Table adapted from (Cui et al., 2017; Laronha & Caldeira, 2020).

MMP Designation	Common name	Other name(s)	IUBMB Enzyme Nomenclature	Production
1	Collagenase-1	Interstitial collagenase Fibroblast collagenase	EC 3.4.24.7	Cells: fibroblasts, keratinocytes, endothelial cells, macrophages, hepatocytes, chondrocytes, platelets, and osteoblasts.
2	Gelatinase A	72-kD gelatinase 72-kD type IV collagenase	EC 3.4.24.24	Cells: dermal fibroblasts, keratinocytes, endothelial cells, chondrocytes, osteoblasts, leukocytes, platelets, and monocytes
3	Stromelysin-1	Transin-1	EC 3.4.24.17	Cells: fibroblasts and platelets
7	Matrilysin	PUMP	EC 3.4.24.23	Cells: epithelia cells Organs: mammalian glands, liver, pancreas, prostate, and skin
8	Collagenase-2	Neutrophil collagenase PMNL collagenase	EC 3.4.24.34	Cells: chondrocytes, endothelial cells, activated macrophages, neutrophils, and smooth muscle cells
9	Gelatinase B	92-kD gelatinase macrophage gelatinase	EC 3.4.24.35	Cells: neutrophils, macrophages, polymorphonuclear leucocytes, osteoblasts, epithelial cells, fibroblasts, dendritic cells, granulocytes, T-cells, and keratinocytes
10	Stromelysin-2	Transin-2 Protoglycanase 2	EC 3.4.24.22	Cells: keratinocytes, macrophages, and epithelium
11	Stromelysin-3			Cells: fibroblasts Organs: uterus, placenta, and mammary glands
12	Metalloelastase	Macrophage metalloelastase	EC 3.4.24.65	Cells: chondrocytes, macrophages and other stromal cells, osteoblasts, fibroblasts. Organs: placenta
13	Collagenase-3	-	-	Connective tissue (cartilage and developing bone) Cells: macrophages and epithelial and neuronal cells
14	MT1-MMP	Membrane-type MMP	EC 3.4.24.80	Cells: fibroblasts, platelets, and osteoblasts
15	MT2-MMP	-	-	Organs: placenta, heart, and brain
16	MT3-MMP	-	-	Cells: cardiomyocytes progenitor cells, leucocytes Organs: lungs, placenta, kidney, ovaries, intestine, prostate, spleen, heart, and skeletal muscle
17	MT4-MMP	-	-	Cells: leucocytes Organs: brain, colon, ovaries, and testicles
19	RASI-1	Stromelysin-4	-	Cells: leucocytes Organs: colon, intestine, ovary, testis, prostate, thymus, spleen, pancreas, kidney, skeletal muscle, liver, lung, placenta, brain, and heart
20	Enamelysin	-	-	Organs: dental tissue (enamel)
21	-	Xenopus-MMP	-	Cells: leucocytes, macrophages, fibroblasts, basal and squamous cell Organs: ovary, kidney, lung, placenta, intestine, neuroectoderm, skin and brain.
23	Cysteine array (CA)-MMP	-	-	Organs: ovary, testicles, and prostate
24	MT5-MMP	-	-	Cells: leucocytes Organs: brain, kidney, pancreas, and lung
25	Leukolysin	MT6-MMP	-	Cells: leucocytes and cancer tissue Organs: testicles, kidney, and skeletal muscle
26	Endometase	Matrilysin-2	-	Cancer cells of epithelial origin
27	-	-	-	Cells: B-lymphocytes Organs: testicles, intestine, lung, and skin.
28	Epilysin	-	-	Cells: basal keratinocytes Organs: epidermis. High levels- testis. Low levels- lungs heart, intestine, colon, placenta, and brain.

Table 2: Members of the MMP Family and their substrates. Adapted from (Cui *et al.*, 2017; Laronha & Caldeira, 2020).

MMP Designation	Collagen Substrate	Non-collagen ECM Substrates	Other Substrates and Targets
1	I, II, III, VII, VIII, X and XI, gelatin	Aggrecan, fibronectin, nidogen, perlecan, proteoglycan link protein, serpins, tenascin, versican, vitronectin	Casein, IGF-BP-3 and -5, IL-1β, L-selectin myelin basic protein, ovostatin, pro-TNF-α SDF-1, α1-antichymotrypsin, α1-antitrypsin
2	I, II, III, IV, V, VII, X and XI, gelatin	Aggrecan, elastin, fibronectin, laminin, nidogen, proteoglycan link protein, tenascin, versican, vitronectin	FGF-R1, IGF-BP-3, and -5, IL-1β, myelin basic protein, pro-TNF-α, TGF-β, **Active MMP-9 and -13**
3	I, II, III, IV, V, IX, X and XI, gelatin	Aggrecan, decorin, elastin, fibronectin, laminin, nidogen, perlecan, proteoglycan link protein, proteoglycans, tenascin, versican, vitronectin	antithrombin III, casein, E-cadherin, fibrinogen, IGF-BP-3, L-selectin, myelin basic protein, osteonectin, ovostatin, pro-HB-EGF, pro-TNF-α, pro-IL-1β, SDF-1, α1-antichymotrypsin, α1-antitrypsin, **pro-MMP-1, -8, and -9,**
7	IV and X, gelatin	Aggrecan, elastin, fibronectin, laminin, nidogen, proteoglycan link proteins, proteoglycans, tenascin, vitronectin	Casein, decorin, defensin, E-cadherin, Fas–ligand, myeline, plasminogen, pro-TNF-α, syndecan-1 transferrin β4 integrin **pro-MMP-2, -7, and -8**.
8	I, II, III, V, VII, VIII and X, gelatin	Aggrecan, elastin, fibronectin, laminin, nidogen	Ovostatin, α2-Antiplasmin, **pro-MMP-8**
9	IV, V, VII, X, XI, and XIV, gelatin	Aggrecan, decorin, elastin, fibronectin, laminin, nidogen, proteoglycan link protein, versican	Casein, CXCL5, IL-1β, IL-8, IL2-R, myelin basic protein, plasminogen, pro-TNF-α, SDF-1, TGF-β
10	III, IV, V, IX and X, gelatin	Aggrecan, elastin, fibronectin, laminin, nidogen, proteoglycans	Casein, fibrilin-10, **pro-MMP-1, -8, and -10**
11	-	Aggrecan, laminin, fibronectin	IGF-BP-1, α1-antitrypsin
12	I, IV and V, gelatin	Aggrecan, elastin, fibronectin, laminin, nidogen, proteoglycans, vitronectin	Casein, fibrinogen, myelin, osteonectin, α1-antitripsin
13	I, II, III, IV, IX, X and XIV, gelatin	Aggrecan, fibronectin, laminin, tenascin	Casein, osteonectin, plasminogen, SDF-1, **pro-MMP-9 and -13,**
14	I, II and III, gelatin	Aggrecan, elastin, fibrin, fibronectin, laminin, nidogen, perlecan, tenascin, vitronectin	CD44, pro-TNF-α, SDF-1, tissue transglutaminase, α1-antitrypsin, α2-macroglobulin, αvβ3 integrin, **pro-MMP-2 and -13**
15	I, gelatin	Aggrecan, fibronectin, laminin, nidogen, perlecan, vitronectin, tenascin	Tissue transglutaminase, **pro-MMP-2 and -13**
16	I and III, gelatin	Aggrecan, fibronectin, laminin, perlecan, vitronectin	casein, **pro-MMP-2 and -13**
17	Gelatin	Fibrin	Fibrinogen
19	I and IV, gelatin	Aggrecan, fibronectin, laminin, nidogen, tenascin-C isoform	Casein
21			α1-antitrypsin
20	V	Aggrecan, amelogenin cartilage oligomeric protein	-
23	Gelatin		-
24	Gelatin	Chondroitin sulphate, dermatin sulphate Fibrin, fibronectin, N-cadherin	**Pro-MMP-2 and -13**
25	IV, gelatin	Fibrin, fibronectin, proteoglycans	α1-antitrypsin, **pro-MMP-2**
26	IV, gelatin,	Fibrin, fibronectin, vitronectin,	Casein, fibrinogen, IGFBP-1, α1-antitrypsin, α2-macroglobulin, **pro-MMP-2**
27	Gelatin	-	-
28	-	-	Casein

MMPs are typically classified into (true) collagenases, gelatinases, stromelysins, matrilysins, membrane-type MMPs (MT-MMPs), and other MMPs according to their substrates or organization of their structural domains (**Figure 8**) (**Table 3**) (Cui *et al.*, 2017). Collagenases and gelatinases are peptidases, which primarily cleave native or denatured

collagens (gelatins), respectively (Allan *et al.*, 1995; Chung *et al.*, 2004; Holmbeck & Birkedal-Hansen, 2013), while stromelysins and matrilysins present a broad substrate specificity against protein constituents of the ECM and cell surface proteins (**Table 2**). The last two differ in structural features: matrilysins are the smallest MMPs since they are constituted only by the pro- and catalytic domains and lack both the hinge region and the PEX domain, while stromelysins have the same domain arrangement as collagenases (**Figure 8**) (Piskór *et al.*, 2020).

Table 3: MMPs Classification

MMP Family	MMP Designation
Collagenases	MMP-1, -8, -13, -18
Gelatinases	MMP-2, -9
Stromelysins	MMP-3, -10, -11
Matrilysins	MMP-7, -26
Membrane-type MMPs	MMP-14, -15, -16, -17, -24, -25
Other MMPs	MMP-12, -19, -20, -23, -28

An alternative classification, based on structural similarities and function, has been proposed in which MMPs are grouped into minimal domain, simple hemopexin, gelatin binding, furin-activated, vitronectin-activated, transmembrane, GPI-anchored and type II transmembrane MMPs (Caley *et al.*, 2015).

The members of the matrixin family are responsible not only for the turnover and degradation of ECM components as initially described, but also for the proteolytic cleavage of several non-matrix substrates, like cytokines, chemokines, receptors, and others (**Table 2**). Therefore, rigorous regulation of MMPs activity is fundamental, and it takes place at four levels: gene expression, compartmentalization, temporal and spatial control of pre-form activation, and enzyme inhibition by their specific endogenous inhibitors, the tissue inhibitors of metalloproteinases (TIMPs; Laronha & Caldeira, 2020). Moreover, MMPs are also inhibited by endogenous non-specific inhibitors like α2-macroglobulin (α2M) and, based on some studies, by the reversion-inducing-cysteine-rich protein with Kazal motifs (RECK; Laronha *et al.*, 2020).

In 1990, van Wart & Birkedal-Hansen proposed the still accepted “cysteine switch” mechanism for MMPs activation (van Wart & Birkedal-Hansen, 1990). During this activation process, the MMP pro-domain is removed through disruption of the interaction between its conserved cysteine residue thiol group with the catalytic zinc ion. This ligand

position is then taken over by a water molecule, complementing the proteinaceous zinc binding residues of the catalytic domain. The thiol–zinc ion interaction can be disrupted by (i) direct proteolytic cleavage of the pro-domain, (ii) allosteric disruption of the zymogen provoked by chaotropic agents or surfactants like sodium dodecyl sulphate (SDS), or by (iii) reduction of the free thiol by chemical agents such as disulphide compounds, sulfhydryl alkylating agents, oxidants, and heavy metal ions such as organomercurial compounds. Allosteric disruption or sulfhydryl reduction activation further leads to subsequent autolytic cleavage of the pro-domain (Ra & Parks, 2007; Springman *et al.*, 1990; Yamamoto *et al.*, 2015).

In vivo, latent pro-MMPs are mainly activated by proteolytic cleavage. One third of MMPs comprise a RxKR or RRKR motif recognized and cleaved by proprotein convertases (PPC) (**Figure 8**). Furin, a subtilisin-like serine protease and the prototypic PPC, is localized in the *trans*-Golgi network and therefore is accountable for MMPs intracellular activation before secretion. All the other MMPs are secreted as zymogens requiring posterior activation. *In vitro* studies reveal that pro-MMPs might be also activated by distinct serine proteases like plasmin, trypsin, chymase and elastase, or even by other MMPs. Nevertheless, the *in vivo* activation mechanism of this proteases remains unclear in many cases (Loffek *et al.*, 2011; Ra & Parks, 2007). Pro-MMP-2 activation by active MMP-14 at the cell surface is probably the best described pro-MMP activation mechanism exerted by other active MMP members. Notably, this activation mechanism requires the concerted participation of TIMP-2 and MMP-14, which form an activation complex with a 1:1:1 stoichiometric ratio on the cell surface (Itoh, 2001; Z. Wang *et al.*, 2000).

1.4.2. The thermolysin family

Thermolysin (EC 3.4.24.27) was the first metalloendopeptidase whose X-ray crystal structure was solved (Matthews, 1988). Thermolysin is a neutral metallopeptidase of 34.6 kDa secreted by *Bacillus thermoproteolyticus*, a thermophilic Gram-positive bacterium (Titani *et al.*, 1972), and the prototypical member of the thermolysin family, also called thermolysin-like proteases (TLPs), which is included within family M4 of the MEROPS database. This is an important family of the gluzincin clan of zinc metallopeptidases and it comprises a large number of neutral metallopeptidases with similar amino-acid sequences, three-dimensional structures, catalytic mechanisms and, to some extent, substrate

specificities (de Kreij *et al.*, 2000; van den Burg & Eijsink, 2013). TLPs are zinc and calcium-dependent metallopeptidases expressed by several microorganisms and are often virulence factors implicated in severe bacterial infections accountable not only for degradation of several host proteins (**Table 4**), but also for the proteolytic activation of bacterial toxin precursors (Adekoya & Sylte, 2009).

Table 4: Thermolysin-like proteases (TLPs) secreted by pathogenic bacteria, their substrates, and pathological implications

TLP	Organism	Host Substrates	Pathological consequences	Reference(s)
λ-toxin	*Clostridium perfringens*	Casein, collagen, complement C3, fibrinogen, fibronectin, immunoglobulin A	Bacterial invasion, haemorrhagic oedema, increased vascular permeability, tissue destruction	(Jin *et al.*, 1996)
Coccolysin	*Enterococcus faecalis*	Casein, collagen, fibrinogen, gelatin, haemoglobin, endothelin-1	Food poisoning, intra-abdominal abscesses, secondary bacteraemia, and root canal, soft tissue, and urinary tract infections	(Mäkinen *et al.*, 1989; Makinen & Makinen, 1994)
M4 TLP	*Helicobacter pylori*	Gelatin, mucin	Gastric carcinoma, gastritis, peptic ulcer	(Smith *et al.*, 1994)
	Vibrio cholerae	Fibronectin, lactoferrin, ovomucin	severe vomiting and watery diarrhoea	(Booth *et al.*, 1983; Finkelstein & Hanne, 1982)
ProA	*Legionella*	α1-antitrypsin, CD4, Interleukin 2	Legionnaire's disease and pneumonia	(Conlan *et al.*, 1988; Mintz *et al.*, 1993; Scheithauer *et al.*, 2021)
Pseudolysin	*Pseudomonas aeruginosa*	α1-antitrypsin, casein, coagulation and complement factors, collagen, elastin, gelatin, immunoglobulins	chronic ulcers, corneal and lung infections, muscle damage, severe haemorrhages	(Heck *et al.*, 1986; Hobden, 2002; Kessler & Safrin, 2014; Komori *et al.*, 2001; Schmidtchen *et al.*, 2003; Wretlind & Pavlovskis, 1983; Yanagihara *et al.*, 2003)

TLPs, as well as other proteases with broad specificity such as the *Bacillus* subtilisins, may be dangerous for the secreting bacterium, so they are produced as partially unfolded pre-proenzymes that are only activated after secretion (Eijsink *et al.*, 2011). The pre-peptide, which acts as a signal peptide for transport, is removed during secretion (Eder & Fersht, 1995). The inactivating pro-peptide comprises two domains: the fungalysin-thermolysin-pro-peptide (FTP) domain, common to bacterial and fungal metallopeptidase families (M4 and M36 in MEROPS database), and the PepSy domain, exclusively conserved within the M4 family (Markaryan *et al.*, 1996; Tang *et al.*, 2003). This pro-domain behaves as an intramolecular chaperone assisting protease folding in the extracellular milieu. After completion of the folding process, it is removed by autoproteolytic maturation (Eijsink *et al.*, 2011; Gao *et al.*, 2010; Marie-Claire *et al.*, 1998).

The TLPs structures display two domains connected by a central α-helix, which comprises the residues of the central zinc binding motif (**HE**xx**H**) (**Figure 10**). The N- and

C-terminal domains are mainly constituted by β-sheets and α-helices, respectively. The thermolysin catalytic zinc ion is deeply buried within the active-site cleft, at the junction between the domains, which has been shown to display a more closed conformation once a ligand is bound (Holland *et al.*, 1992; van den Burg & Eijsink, 2013; Veltman *et al.*, 1998). *In vitro*, removal of the catalytic zinc ion from prototypical thermolysin generates an inactive apoenzyme whose activity can be restored to 100%, 200%, 60% or 10% of the native activity through addition of zinc, cobalt, iron or manganese ions, respectively. Interestingly, excess of zinc ions leads to thermolysin inactivation due to binding of an additional zinc ion to H_{231}, assuming the position usually occupied by the catalytic water molecule (H_2O^{231}) (Holland *et al.*, 1995; Holmquist & Vallee, 1974). Furthermore, thermolysin binds four calcium ions (Ca_1–Ca_4). Whereas Ca_1 and Ca_2 occupy a double binding site near the active site cleft, Ca_3 and Ca_4 are exposed in loops on the N- and C-terminal domains, respectively (**Figure 10**). Calcium ions are critical for enzyme stability, and thus itis not surprising that thermostable TLPs bind four calcium ions while less stable TLPs bind only two (Ca_1 and Ca_2). Studies on TLP from *Bacillus stearothermophilus* (TLP-ste) suggest that the calcium ions have an important regulatory role; the intracellular environment with low calcium concentrations favours protein instability and ensures partial protein unfolding while the higher availability of calcium ions in the extracellular space triggers complete folding and ensures protein stability (Eijsink *et al.*, 2011).

Just like in other gluzincins, the catalytic zinc of TLPs is coordinated by a water molecule and by the histidines and the downstream glutamate residue of the zinc-binding motif **H***E*xx**H**+**E,** in which the first glutamate (*E*) residue acts as a general base/acid for catalysis.

The main specificity site of archetypal thermolysin is the S1' pocket, which preferably accepts aliphatic amino acids like L, F, I, and V (**Figure 11**) (Adekoya & Sylte, 2009; Heinrikson, 1977; Ligné *et al.*, 1997). The TLPs aureolysin, coccolysin and λ-toxin present similar specificities, preferring the hydrophobic residues L, V, Y, I, F and A in S1' (de Kreij *et al.*, 2000; Drapeau, 1978; Mäkinen *et al.*, 1989; Makinen & Makinen, 1994). Pseudolysin and griselysin specificities are also similar to the prototype since they favour hydrophobic residues but they prefer aromatic over aliphatic residues (Kajiwara *et al.*, 1991; Tsuyuki *et al.*, 1991).

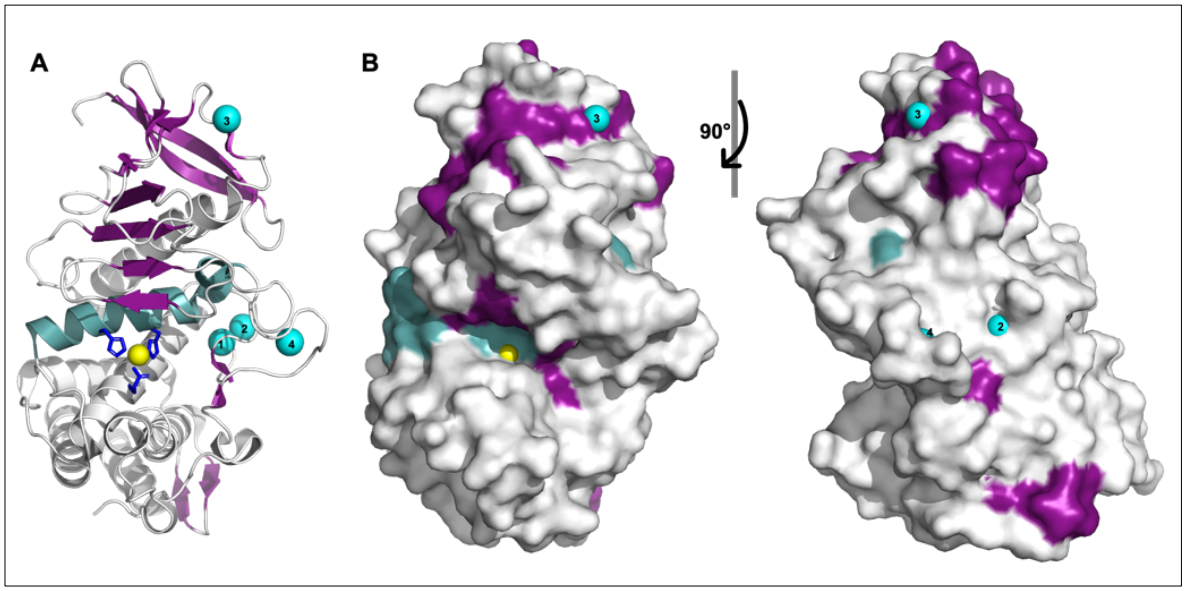


Figure 10: Schematic representation of thermolysins structure. (A) Structure of prototypical thermolysin (UP P00800; PDB 1kei). β-strands and α-helices are coloured in purple and grey, respectively. Central α-helix, connecting N- and C-terminal domains, is shown in light teal. Zinc ion and zinc binding residues are highlighted in yellow and dark blue, respectively. Calcium ions are numbered and shown in light blue. (B) Thermolysin surface plot in the (left) same orientation as (A) and after (right) clockwise 90° rotation, highlighting the active-site cleft, the catalytic zinc, and the surface exposed calcium ions. All structural models were prepared in PyMOL (The PyMOL Molecular Graphics System, n.d.).

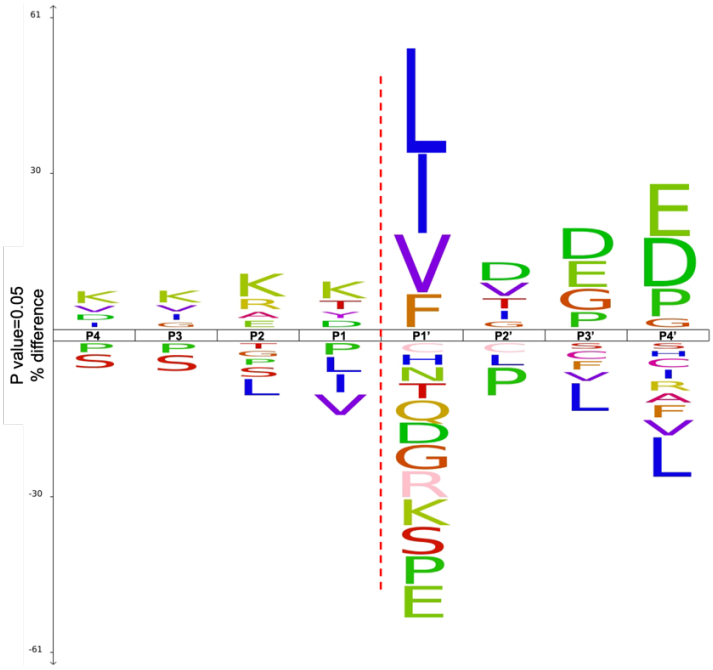


Figure 11: Illustrative IceLogo of thermolysin substrate preference. Image based on data deposited in MEROPS database (Rawlings *et al.*, 2018) and created using IceLogo (Colaert *et al.*, 2009) where preferential substrate residues were coloured according to hydrophobicity at pH 7. Substrate residues favourable for cleavage are shown above the substrate positions indicated (positive y-axis values), while unfavourable residues are displayed below. The red dashed line indicates the scissile bond between P1 and P1'.

Thermolysin is a thermostable enzyme with a bell-shaped pH profile, which shows maximal proteolytic activity at neutral pH values (Feder & Schuck, 1970). When working with this protease, phosphate buffers should be avoided since they inhibit its activity. Moreover, addition of $CaCl_2$ to the buffers is recommended to minimise autolysis (Feder, 1968; van den Burg & Eijsink, 2013). The presence of neutral salts, such as sodium chloride, at high concentrations (1–4 M) increases the thermal stability and catalytic activity of thermolysin (Inouye *et al.*, 1998; Inouye, 1992; Yang *et al.*, 1994), whereas zinc-chelating agents like ortho-phenanthroline reversibly inhibit its activity (Holmquist & Vallee, 1974).

1.4.3. Aureolysin

Among all TLPs, aureolysin caught the attention of researchers due to its involvement in *Staphylococcus aureus* infection. *S. aureus* is a clinically important pathogen involved in a wide range of human infections. This versatile pathogen can occupy numerous niches within the host, varying from a commensal bacterium to an infective opportunist that causes not only superficial lesions such as wound infections and abscesses, but also life-threatening systemic infections such as bacteraemia, endocarditis, meningitis, pneumonia, sepsis, osteomyelitis, and toxic shock syndrome (David & Daum, 2010; Lowy, 1998; Shaw *et al.*, 2004). *S. aureus* infections triggered major concerns since the emergency of methicillin-resistant *S. aureus* (MRSA) strains, which entail important clinical and economic consequences (Ahmad-Mansour *et al.*, 2021; Zhen *et al.*, 2020).

The effectiveness of *S. aureus* infections relies on the concerted production of a vast amount of virulence factors, i.e. protein and non-protein factors that promote infection through colonization and evasion of the immune system, in a spatially, temporally and environmentally controlled manner (Ahmad-Mansour *et al.*, 2021; Cheung *et al.*, 2021). Some of these virulence factors are secreted proteases, such as the cysteine proteases staphopain A (ScpA) and B (SspB), the serine protease V8 (alias SspA) and seven further serine protease-like proteins (SplA–F), as well as the TLP aureolysin (Pietrocola *et al.*, 2017).

The staphylococcal extracellular proteolytic system is driven by a cascade of zymogen activation. Unlike other TLPs, aureolysin zymogen activation seems to occur in

two sequential steps. Pro-aureolysin activation is initiated by autoproteolytic cleavage of the $T_{85}\downarrow L_{86}$ bond within the FTP domain of the pro-peptide and completed by subsequent cleavage at $E_{208}\downarrow A_{209}$, at the junction of the pro- and metallopeptidase domains (Nickerson *et al.*, 2008). Aureolysin activates SspA, which in turn activates SSpB, which further activates ScpA (**Figure 12**). Thus, aureolysin is the key initiator of the activation cascade (Shaw *et al.*, 2004).

Aureolysin activity is also implicated in the evasion of *S. aureus* from the host immune response. It cleaves and thus inactivates the endogenous protease inhibitor α1-antitrypsin, which leads to increased levels of neutrophil elastase, an α1-antitrypsin target, that favours turnover of plasma proteins (Potempa *et al.*, 1986). Moreover, aureolysin directly cleaves plasma proteins, such as prothrombin, which produces active thrombin and stimulates staphylocoagulase activity (Wegrzynowicz *et al.*, 1980). Furthermore, aureolysin circumvents both innate and adaptive mechanisms of the host immune system. For instance, it disturbs T- and B-lymphocyte stimulation by polyclonal activators, inhibits the immunoglobulin production of lymphocytes (Prokešová *et al.*, 1991), cleaves the anti-microbial peptide LL-37 (Sieprawska-Lupa *et al.*, 2004) and the complement components C3 and C3b (Laarman *et al.*, 2011), and protects staphylococci within macrophages and neutrophils upon phagocytosis (Burlak *et al.*, 2007; Kubica *et al.*, 2008).

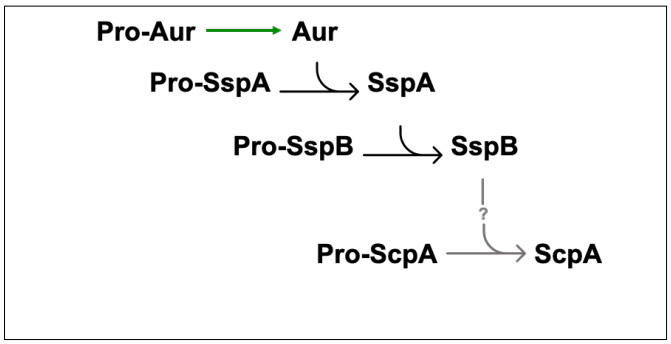


Figure 12: Diagrammatic representation of the activation cascade of *S. aureus* extracellular proteases. Aureolysin (Aur) autolytic activation depicted as a green arrow. The hypothetical activation of Pro-ScpA by SspB is shown by the grey arrows. Adapted from (Shaw *et al.*, 2004).

The aureolysin encoding gene (*aur*) is strongly conserved across *S. aureus* strains and occurs in two allelic forms that give rise to aureolysin type-I and type-II, which differ in eleven residues (Sabat *et al.*, 2000). Aureolysin type-II is two times more active than the type-I variant against azocasein (S. Takeuchi *et al.*, 2002).

The expression of aureolysin is tightly controlled, being activated during the bacterial post-exponential growth phase (Potempa & Shaw, 2013). Mature aureolysin is a single-chain protein that migrates according to a molecular mass of 28 kDa in size-exclusion chromatography and sedimentation equilibrium studies (Arvidson, 1973; Saheb, 1976) or as a 38-kDa protein in SDS-PAGE studies (Drapeau, 1978). Aureolysin type-I shares a sequence identity of only 49% with prototypical thermolysin, but their three-dimensional structures and substrate specificities are very similar (**Figure 13**) (Potempa & Shaw, 2013). However, aureolysin binds only three calcium ions, which are located close to the active site and in the C-terminal domain, and its deep active-site cleft has a more closed conformation than in other TLPs, mostly due to the insertion of five residues above the active site (**Figure 13**). This may also explain the lack of elastinolytic activity, a common feature of thermolysin and TLPs (Potempa & Shaw, 2013).

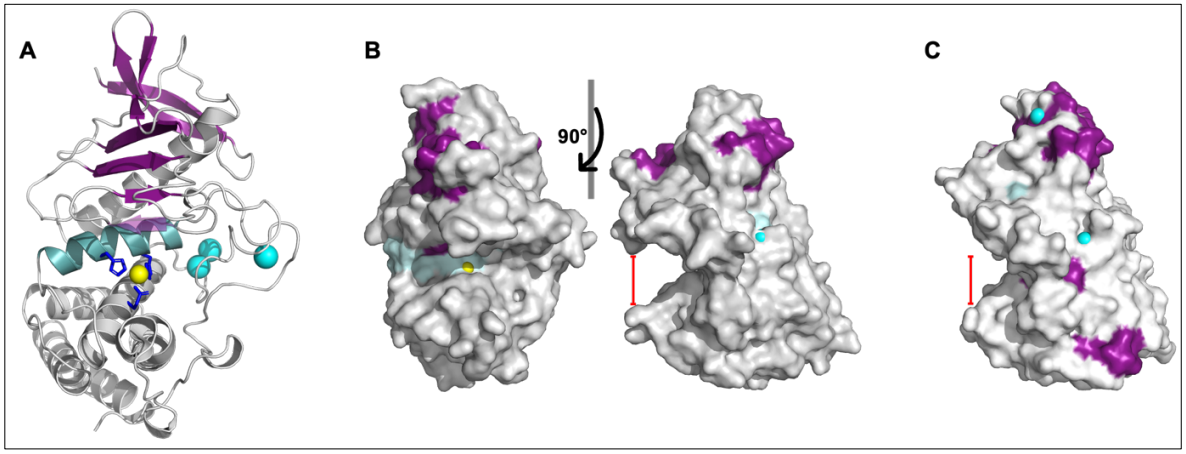


Figure 13: Schematic representation of aureolysin structure. (**A**) Structure of aureolysin (UB 81177; PDB 1bqb; Banbula *et al.*, 1998). β-strands and α-helices are coloured in purple and grey, respectively. Central α-helix, connecting N- and C-terminal domains, is shown in light teal. Zinc ion and zinc binding residues are highlighted in yellow and dark blue, respectively. Aureolysin displays three calcium ions (light blue spheres). (**B**) Aureolysin surface plot in the (left) same orientation as (A) and (B) after clockwise 90° rotation, highlighting the active-site cleft, the catalytic zinc, and the surface exposed calcium ions. (**C**) Thermolysin surface in the exact same orientation as aureolysin to demonstrate the variance of depth of their active site clefts (highlighted with the red bar).

2. Proteolysis regulation

The original perception of proteases as simple indiscriminate digestive enzymes in charge of the degradation of dietary proteins or unfolded and erroneous proteins has changed over recent years as more proteases were discovered and studied in more detail (Puente *et al.*, 2003). Proteases are also key players of limited proteolysis, which is a complex, fine-tuned, irreversible post-translational modification (PTM) described for the first time by Linderstrom-Lang and Ottesen in 1949 according to (Doi *et al.*, 1987; Richards, 1992). Such proteolytic processing is responsible for activation, inactivation or function modification of other proteins, including other proteases (Overall & Blobel, 2007). For example, signal-peptide peptidases are implicated in protein quality control in the endoplasmic reticulum and are responsible for the removal of signal peptides from secreted proteins (Boname *et al.*, 2014; Liaci & Förster, 2021; Schrul *et al.*, 2010). Moreover, proprotein convertases (proteinases functioning in the Golgi apparatus, endosomes, secretory granules or in the cell surface) such as furin are crucial for the activation of proproteins like prohormones or proneuropeptides (Ra & Parks, 2007; Seidah, 2011), in addition to the previously described MMPs. Furthermore, proteolytic enzymes are responsible for the activation at the appropriate time or localization of other proteases synthetized in inactive forms as zymogens (Khan & James, 1998). Some examples of this activation mechanism are the activation of thrombin (Krishnaswamy, 2013), plasma protein C (Stojanovski *et al.*, 2020), caspases (Salvesen & Dixit, 1997) and MMPs (Ra & Parks, 2007). Consequently, proteases are implicated in multitudinous biological processes including cell signalling, cell-cycle regulation (Rhind & Russell, 2012), cell death (namely in apoptosis; Utz & Anderson, 2000), cellular stress response (Fulda *et al.*, 2010), ECM organization (Ricard-Blum & Vallet, 2016), cell migration/invasion (Madri & Graesser, 2000) and regulation of lipid homeostasis (Sam *et al.*, 2019). They are therefore engaged in tissue remodelling (Chen, 1992; Ricard-Blum & Vallet, 2016), organ formation (Sanaei *et al.*, 2021), immune response ranging from antigen presentation (Matthews *et al.*, 2010) to pathogen infection (Marshall *et al.*, 2017) and T-cell maturation (Guerder *et al.*, 2019), homeostasis of the vascular (McCarty & Percival, 2013; Walsh & Ahmad, 2002) and neurological systems (Cenac, 2013; Yagami *et al.*, 2019), as well as in reproductive processes like menstruation, pregnancy (Girardi *et al.*, 2020) and embryogenesis (Rose *et al.*, 2003; Seshagiri *et al.*, 2003).

Limited proteolysis has also been associated with a multitude of roles in procaryotes, among which the control of lipid metabolism and of stress responses (Micevski & Dougan, 2013; Wettstadt & Llamas, 2020), sporulation (Maurizi & Switzer, 1980) and pathogenicity (Culp & Wright, 2017; Frees *et al.*, 2013; Miyoshi, 2013) are most studied.

The myriad of biological processes in which proteases take part demonstrates that control of proteolytic enzymes is of utmost importance. Accordingly, peptidases are regulated at multiple levels, including regulation of gene expression, restriction to specific cellular compartments, and by inactivating pro-domains or endogenous inhibitors (López-Otín & Bond, 2008). It is important to mark that the complexity of biological processes demands a strict control and perfect synchronization of all proteins participating in the respective mechanisms. The fine-tuning of individual proteins is not only attained thought the high specificity and selectivity of a protease and its substrates or by co-regulatory mechanisms at the protease level, but it might also be achieved through the interplay of distinct post-transcriptional modifications at both protease and substrate levels (Boon *et al.*, 2016; Goth *et al.*, 2018). For instance, site-specific O-glycosylation is crucial for the co-regulation of proteolytic events such as ectodomain shedding, proprotein processing and the cleavage of G-protein-coupled receptors (GPCR) and downstream signalling (Goth *et al.*, 2018). The regulation of fibroblast growth factor 23 (FGF23) by proteolysis, site-specific O-glycosylation and phosphorylation is an illustrative example of this concerted regulation (Tagliabracci *et al.*, 2014).

In sum, throughout all forms of life, from single-cell archaea to complex multicellular eukaryotes, peptidases and their inhibitors, present in invading parasites and defending hosts, participate in various (patho)physiological processes and in virulence and invasion, as well as in the countered protective mechanisms (Armstrong, 2006).

2.1. α2-Macroglobulins

The α2Ms are high molecular weight, multi-domain glycoproteins, which are able to inhibit a remarkably broad spectrum of endopeptidases. They belong to the superfamily of the thioester-containing proteins (TEPs) whose members share a common evolutionary origin and exhibit conserved structural and functional features. Other relevant members of this superfamily are complement components C3, C4 and C5; pregnancy zone protein (PZP) and the PZP-like α2M-domain-containing protein 8 (CPAMD8); α1-inhibitor-3 (α1I3) and insect and nematode TEPs. Despite functionally different, the mechanism of action of TEPs entails proteolytic processing and structural rearrangement (Garcia-Ferrer *et al.*, 2017).

The α2Ms are intrinsic components of the innate immune system of eukaryote species, which are found in the haemolymph of invertebrates, plasma of vertebrates, and in the egg white of reptiles and birds (Armstrong, 2006; Buresova *et al.*, 2009; Lim *et al.*, 2011; Nielsen *et al.*, 1994). However, α2Ms are not exclusively found in metazoans. Indeed, they have been identified in several bacterial species, most of them being pathogenic to or colonizers of higher eucaryotes (**Table 5**) (Neves *et al.*, 2012). α2Ms exert similar trapping dynamics for their inhibitory mechanisms, which are triggered by the proteolytic cleavage of a bait region by the prey peptidase, which leads to subsequent conformational modification, independently of their oligomeric state (Arimura & Funabiki, 2022; Enghild *et al.*, 1990; Garcia-Ferrer *et al.*, 2015, 2017; Ikai *et al.*, 1983; Marrero *et al.*, 2012). And even though some α2Ms lack the thioester bond, their inhibitory activity isn't impaired; they just cannot covalently bind the entrapped target peptidase (Nagase & Harris, 1983; Robert-Genthon *et al.*, 2013).

2.1.1. Human α2M

Hα2M, first isolated in 1946, is a 720 kDa homotetrameric glycoprotein (Barrett & Starkey, 1973; Cohn *et al.*, 1946; Travis & Salvesen, 1983). Mainly synthesized in the liver, all mammalian α2Ms are found in blood, which ensures availability throughout the body and, therefore, execution of their innate immunity protection roles (Andus *et al.*, 1983). Hα2M is a unique plasma inhibitor protein that targets endopeptidases regardless of their substrate specificity or catalytic type, and it has been reported to inhibit serine, aspartic,

cysteine and metallopeptidases (Hibbetts *et al.*, 1999; Kantyka *et al.*, 2010). The hα2M inhibition mechanism, known as a "Venus flytrap" mechanism, does not function through blockage of the catalytic site of a target peptidase but rather by peptidase entrapping, following which the imprisoned enzyme remains active but without or limited substrate access (**Figure 14**) (Barrett & Starkey, 1973).

Table 5: Summary table of characterised α2-macroglobulins. The symbols (+), (-) and (~) represent presence, absence or unknown, respectively. Table adapted from (Garcia-Ferrer et al., 2017).

Name	Organism	UniProt Number	Localization	Mass (kDa)	Oligomerization	Thioester bond
α2M	*Homo sapiens*	P01023	Blood serum	720	Tetramer	+
α2ML1		A8K2U0	Epidermis	180	Monomer	+
PZP		P20742	Pregnancy blood serum	720	Tetramer	+
Ovostatin	*Gallus gallus*	P20740	Egg white	780	Tetramer	-
α1M	Rabbit	-	Blood serum	~	Tetramer	+
α2M	*Biomphalaria glabrata* (gastropod mollusc)	-	Haemolymph	800	Tetramer	+
α2M	*Ornithodoros moubata* (tick)	-	Plasma	420	Dimer	+
α1I3	*Rattus norvegicus*	P14046	Blood serum	174	Monomer	+
α1M		Q63041	Blood serum	~	Tetramer	~
α2M	*Limulus polyphemus* (horseshoe crab)	-	Haemolymph and blood cells	354	Dimer	+
α2M	*Erinaceus europaeus* (hedgehog)	-	Plasma	800	Tetramer	~
α2βM		-	Plasma	450-550	Dimer	~
α2M	*Penaeus vannamei* (white shrimp)	-	Hemolymph	360	Dimer	+
IrA2M	*Ixodes Ricinus* (tick)	-	Haemolymph	440	Dimer	+
α2M	*Chelonia mydas japonica* (turtle)	-	Blood serum	~	~	+
Ovostatin		-	Egg white	~	~	-
αM	*Rana catesbiana* (frog)	-	Blood serum	180	Monomer	+
α2M	*Astacus astacus* (crayfish)	-	Haemolymph	390	Dimer	+
α2M	*Octopus vulgaris* (mollusc)	-	Haemolymph	360	Dimer	+
α2M	*Struthio camelus* (ostrich)	-	Blood serum	779	Tetramer	+
α2M	*Helix pomatia* (gastropod mollusc)	-	Haemolymph	697	Tetramer	+
α2M	*Pacifastacus leniusculus* (crayfish)	-	Haemolymph	380	Dimer	+
α2M	*Libinia emarginata* and *Cancer borealis* (crab)	-	Haemolymph	480-460	Dimer	+/-
α2M	*Homarus americanus* (lobster)	-	Haemolymph	342	Dimer	+
α2M	*Cyprinus carpio* (bony fish carp)	-	Blood serum	380	Dimer	+
α2M	*Farfantepenaeus paulensis* (shrimp)	-	Blood serum	389	Dimer	+
ECAM	*Escherichia coli*	P76578	Inner membrane lipoprotein	183	Monomer	+
SA-α2M	*Salmonella enterica* ser. Typhimurium	Q8ZN46	Inner membrane lipoprotein	179	Monomer	+
MagD (YfaS)	*Pseudomonas aeruginosa*	PA4489	Inner membrane lipoprotein	165	Monomer	-

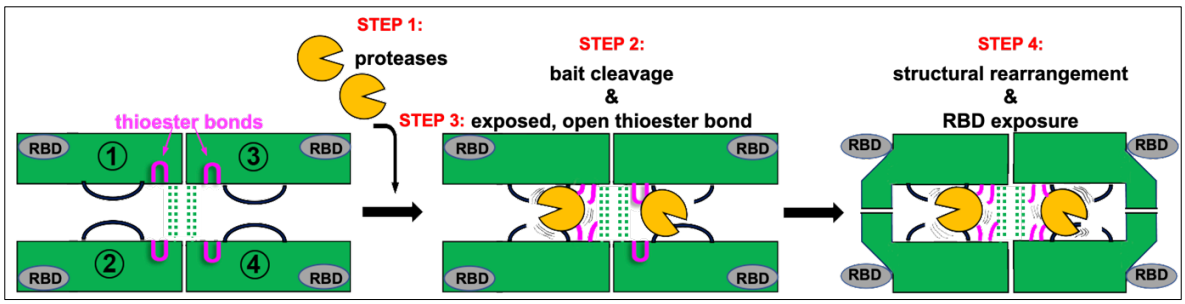


Figure 14: Illustrative representation of the "Venus flytrap" mechanism of inhibition by hα2M. Peptidases enter the inhibitor cavities where they recognize and cleave the bait region segments leading to exposure of the thioester bonds. Lysine residues protruding from the peptidase surface then attack the inhibitor thioester bond, thus covalently binding the peptidase to the inhibitor. Cleavage of the thioester bonds instigate the α2M structural rearrangements marked by exposure of the receptor binding domain (RBD). The top and bottom dimers are built by monomers ① and ② and monomers ③ and ④, respectively. Adapted from (Luque et al., 2022).

The hα2M tetramer is a dimer of dimers, formed by the association of two disulphide-linked dimers in antiparallel orientation (a top and a bottom dimer). Each protomer of 1451 residues is a multidomain molecule comprising 11 domains (**Figure 15**), whose stabilization and solubility are ensured by twelve intramolecular disulphide bonds and eight residues containing N-linked carbohydrate groups (P. E. Jensen & Sottrup-Jensen, 1986; Kolodziej *et al.*, 1996; Marrero *et al.*, 2012; Qazi *et al.*, 2000). The seven N-terminal macroglobulin-like (MG) domains (MG1–MG7) are seven-stranded antiparallel β-sandwiches comprising three- and four-stranded β-sheets. Arrangement of the first six MG domains shapes a central ellipsoid opening (entrance 1) encircled by domains MG3 and MG6 (**Figure 16**). Downstream of the MG7 domain, which constitutes the upper limit of the molecule, and laterally attached to MG2, the CUB domain is formed by two four-stranded antiparallel β-sheets. The thioester domain (TED), which is inserted within the CUB domain, is a helical domain with a α/α-toroid topology and a small β-hairpin. The C-terminus of the protomer features the receptor binding domain (RBD), also called MG8. This domain displays a β-sandwich architecture similar to the MG domains but further includes a β-α-β-motif. The MG1–MG7, CUB, TED and RBD domains frame a cavity (on the back face of the monomer) where the "bait-region domain" (BRD) is embedded.

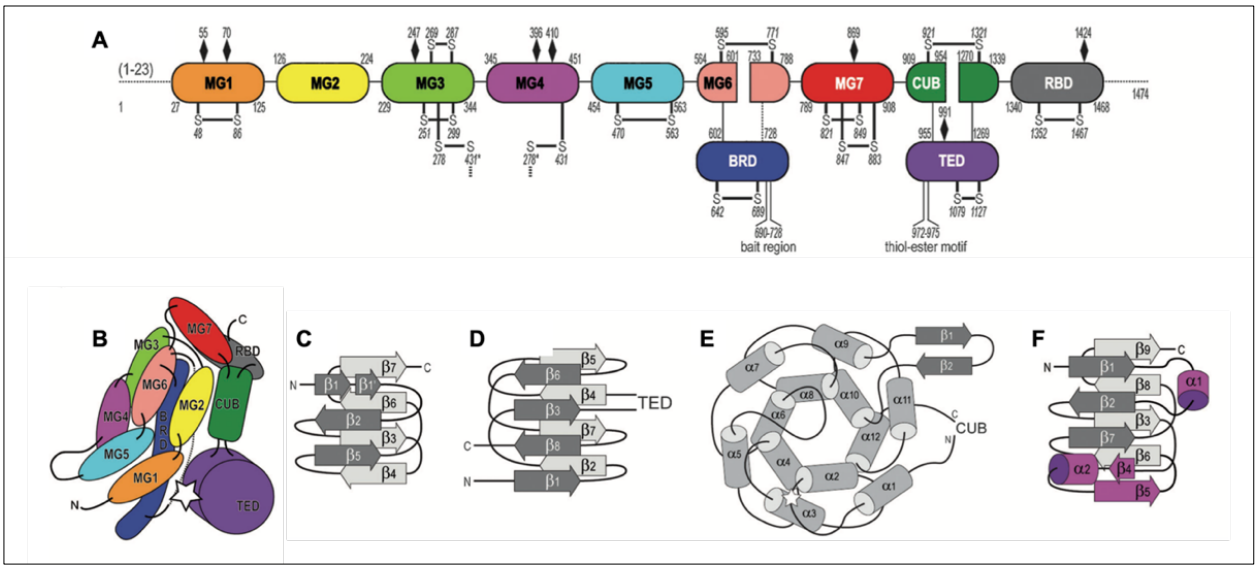


Figure 15: Schematic representation of α2M domains organization and structure. (**A**) Arrangement of human α2M domains with indication of their flanking residues. The disulphide bonds, bait region segment, thioester motif, and N-linked glycosylation sites (◆) protrude from the respective domains. (**B**) Estimated arrangement of α2M protomer domains. The bait region and thioester site are represented with a star symbol and a dashed line, respectively. Topological illustration of the secondary structure of MG, CUB, TED, and RBD domains (**C, D, E and F**, respectively). Adapted from (Marrero *et al.*, 2012).

Each dimer allows access to the central cavity of the tetrameric particle through an "entrance 1" provided by each protomer and through two additional entrances (identified as "entrance 2"), which are framed by the MG2–MG3, MG7, CUB and TED domains, along with the MG4 domain of its disulphide linked monomer (**Figure 16**). A third central cavity ("entrance 3") is created by tetramerization and is framed by the TED of one monomer, part of the BRD of its vicinal protomer and the MG4 domain of its opposite protomer (Marrero *et al.*, 2012). Thus, the homotetramer has a total of twelve major entrances. The size of these entrances prevents the prey from escaping but allows the entrance of small proteins or inhibitors (6–9 kDa) to the central cavity (Barrett & Starkey, 1973). The so called "prey chamber" is located in the centre of the tetramer being delimited at the top and bottom by the MG3 and MG4 segments implicated in the covalent disulphide-dependent dimerization. Plus, this cavity is elongated by the "substrate ante-chambers" formed by the back, concave part of the four monomers. The prey chamber may accommodate two peptidase molecules of 20–30 kDa (one in each disulphide linked dimer) or a single peptidase of higher molecular weight (80–90 kDa), in agreement with the maximal 2:1 stoichiometry of inhibition determined for peptidase binding by tetrameric α2M (Marrero *et al.*, 2012; Sottrup-Jensen, 1989).

Inhibition by α2M is triggered by the entrapped endopeptidase(s) who cleaves the "bait-region segment" (P_{690}-T_{728}) of the BRD, a universal bait for endopeptidases and whose flexibility ensures full accessibility and adaptation to distinct types of active-site clefts (Sottrup-Jensen, 1989). Cleavage of the bait region provokes a major conformational change that leads to the exposure of the up to then buried thioester bond. The exposed thioester bond, which is formed between the cysteine and glutamine side chains of the $C_{972}GEQ_{975}$ motif within the TED, is promptly attacked by surface-located lysine residues from the prey proteinase, leading to covalent entrapment. Moreover, the hydrolysis of the thioester bond might be also triggered by the nucleophilic attack of small amines such as methylamine (MA), ethylamine and ammonia (Barrett *et al.*, 1979; Garcia-Ferrer *et al.*, 2017). In response to hα2M induction and independently of the trigger (proteinase or MA), the RBD of each monomer becomes exposed at the molecule surface, which causes recognition by specific cell-surface receptors and consequential internalization and lysosomal degradation (endocytosis) of the complex. In this way, hα2M ensures clearance of the inhibited proteases from the circulation (Marrero *et al.*, 2012).

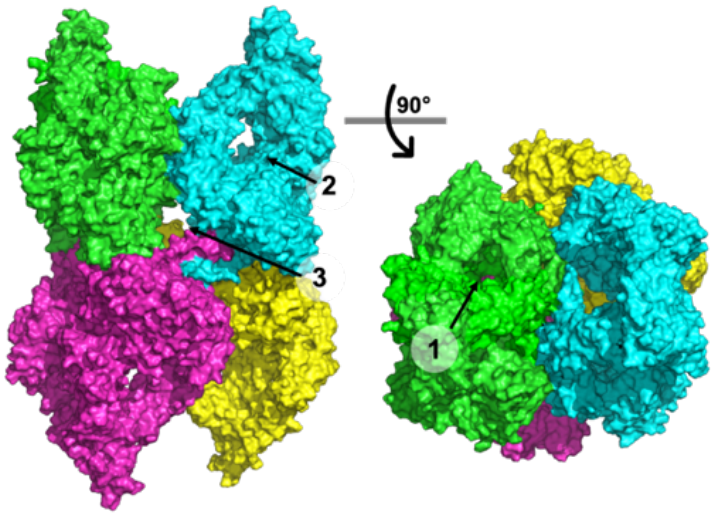


Figure 16: Surface plot of tetrameric α2M. α2M tetramer (PDB 6tav; Luque *et al.*, 2022) is formed by the non-covalent association of the top and bottom dimers, each constituted by two disulphide linked monomers (green plus blue monomers and pink plus yellow monomers, respectively). In the tetramer molecule, each protomer (green for reference) has one vicinal (pink), one opposite (yellow) and one disulphide-linked (blue) protomer. Exemplifying entrances 1, 2 and 3 are indicated by black arrows

Induction of hα2M, either by a proteinase or small nucleophiles, instigates the transition of the native tetramer from an open native conformation to a closed induced one. Therefore, the native and induced hα2M molecules correspond to the electrophoretically "slow" (S) and "fast" (F) forms, respectively (Barrett *et al.*, 1979; Luque *et al.*, 2022).

Finally, it is important to highlight that functions of ha2M are not limited to peptidase binding and inhibition, as it participates in the interaction with several endogenous and exogenous proteins (Garcia-Ferrer *et al.*, 2017; Rehman *et al.*, 2013).

2.2. RECK

The *reck* gene was discovered in 1998 during the screening of human cDNA clones, which induced a flat morphology ("flat reversion") in fibroblasts transformed with the *ras* oncogene. Among the resulting genes, one was found to encode a 110-kDa protein with 971 amino acids, of which 9% were cysteines. Sequence analysis unveiled the presence of three Kazal-like domains, so the protein was dubbed RECK, from reversion-inducing cysteine-rich protein with Kazal motifs (Takahashi *et al.*, 1998). RECK is an extracellular, GPI-anchored protein, ubiquitously expressed in normal tissues throughout all development stages, whose downregulation is associated with tumour and metastasis suppression (Guo & Zou, 2006; Noda & Takahashi, 2007). Notably, in mouse embryos, RECK is expressed in mesenchymal tissues and in the marginal zone of the neuronal tube and large blood vessels (Oh *et al.*, 2001). Moreover, it is also expressed in placenta, from early to term pregnancies, with expression levels increasing over the gestational time (Guo & Zou, 2006). By contrast, RECK expression is abolished or downregulated in tumour-derived cell lines that express oncogenes and in malignant tumours (**Table 6**) (Noda & Takahashi, 2007).

The human *reck* gene was mapped to the short arm of chromosome 9 (9p13→p12), similarly to other tumour suppressor genes (Guo & Zou, 2006; Takahashi *et al.*, 1998). Spanning more than 87 kb, RECK includes 21 exons and 20 introns, and a total of thirteen single nucleotide polymorphisms (SNPs) have been identified in RECK gene: four in gene coding regions (exons 1, 9, 13 and 15) and nine in intragenic regions (Eisenberg *et al.*, 2002). The human RECK gene shows homologies of approximately 98%, 94%, 93% and 86% with the monkey, bovine, mouse, and rat orthologs, respectively (Russell *et al.*, 2021).

The *reck* gene is tightly regulated through epigenetic mechanisms like promotor methylation, histone deacetylation or non-coding RNA-associated gene silencing, mostly exerted by microRNAs (miRNAs) (Russell *et al.*, 2021). Furthermore, oncoproteins like TGAT (trio-related transforming gene in ATL tumour cells), HER-2/neu (human epidermal growth factor receptor 2), and RAS (rat sarcoma), downregulate RECK expression by interaction with the specificity protein 1 (SP1)-binding site within the RECK promotor, triggering the aforementioned epigenetic mechanisms (Y. Chen & Tseng, 2012).

Table 6: List of cancer types who have been associated with RECK downregulation

Cancer type	Reference
Glioma	(Chen & Tseng, 2012; Silveira Corrêa *et al.*, 2010)
Neuroblastoma	(Xu *et al.*, 2015)
Head and neck cancer	(Liu *et al.*, 2012; Xia *et al.*, 2014; Zhang *et al.*, 2015; Zhou *et al.*, 2014)
Gastric cancer	(Liu *et al.*, 2015)
Liver cancer	(Xu *et al.*, 2015)
Biliary tract cancers	(Masui *et al.*, 2003; Yiqing *et al.*, 2005)
Colorectal cancer	(Oshima *et al.*, 2008; Takeuchi *et al.*, 2004)
Lung cancer	(Chang *et al.*, 2007; Qi *et al.*, 2015)
Melanoma	(Jacomasso *et al.*, 2014)
Osteosarcoma	(Kang *et al.*, 2007)
Bladder cancer	(Hirata *et al.*, 2012)
Breast cancer	(Chiang *et al.*, 2013)
Ovarian cancer	(Fejzo *et al.*, 2011)
Cervical cancer	(Discacciati *et al.*, 2015)
Prostate cancer	(Hirata *et al.*, 2013)

At the molecular level, RECK is a glycoprotein comprising five putative cysteine knot motifs (KN) (C_2-X_{7-8}-C-X_3-C-X_{12-22}-C-X_{9-12}-C) followed by two regions encompassing multiple epidermal growth factor-like (EGF-like) repeats and three Kazal-like domains. Such domains are flanked by hydrophobic extremities: the N-terminal extremity corresponds to a secretory signal peptide sequence, while the C-terminus includes a GPI signal responsible for RECK anchoring to the plasma membrane (**Figure 17A**) (Takahashi *et al.*, 1998). The RECK EGF-like repeats present weak homology with EGF, but their function remains unidentified. The Kazal motifs, characteristic for the I1 family of serine protease inhibitors, comprise six cysteine residues forming disulphide bounds in a defined and specific arrangement whose structure consists of a central α-helix surrounded by β-strands (two at the C- and one at the N-terminal side) (Papamokos *et al.*, 1982; Rawlings *et al.*,

2004). In contrast to the first, the second and third Kazal motifs of RECK seem incomplete. The involvement of such motifs in the inhibition of serine proteases through "trapping reactions" or "reversible tight binding interactions" prompted Clark *et al.*, (2007) to suggest the involvement of these domains in a putative inhibitory function of RECK. Additionally, RECK presents five N-glycosylation sites, three of which (N_{86}, N_{297} and N_{352}) have been shown to be crucial for regulation of RECK function (Simizu *et al.*, 2005; Takahashi *et al.*, 1998).

Omura *et al.*, (2009), using transmission electron microscopy, described the RECK structure as a cowbell-like shaped dimer while in our hands the protein behaved as a monomer (Mendes *et al.*, 2020). Despite intensive efforts, crystallisation of RECK has been unsuccessful, which might be explained by glycosylation or intrinsic flexibility of the protein. However, with the advent of the fold-prediction program *AlphaFold* (AF), a homology model of RECK became available (**Figure 17B**).

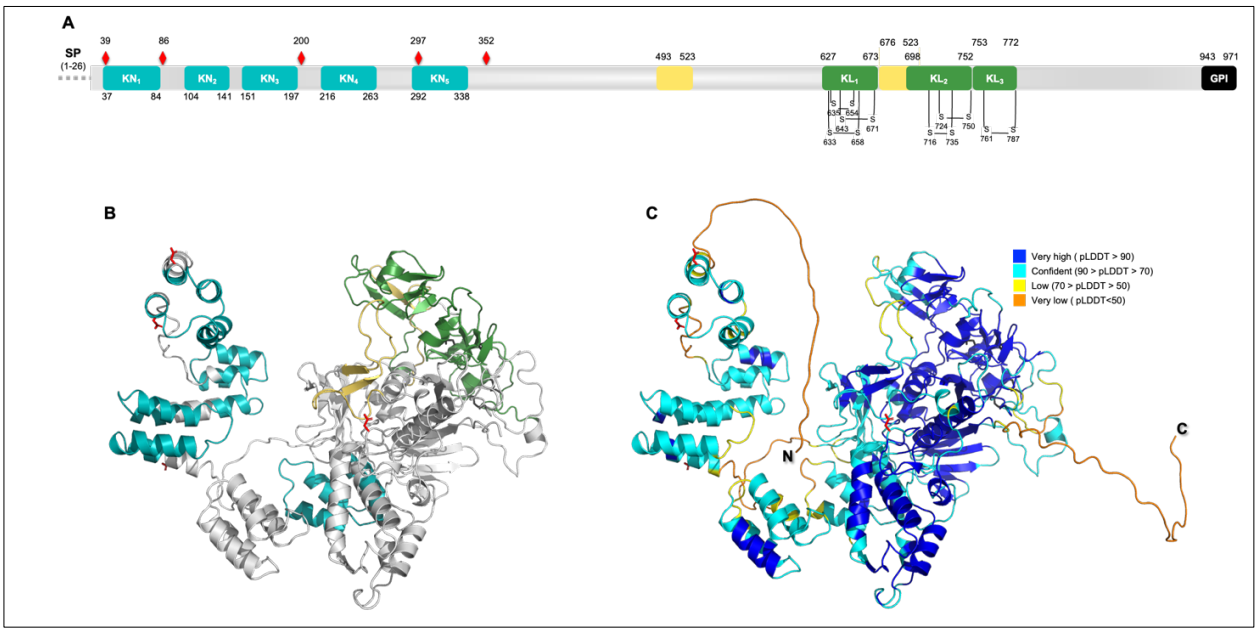


Figure 17: Schematic representation of RECK domains organization and structure. (**A**) Domain arrangement of human RECK (UP O95980) with indication of each domain flanking residues. The disulphide bonds and N-linked glycosylation sites (◆) protrude from the respective domains. The hydrophobic extremities corresponding to the signal peptide (SP) and the GPI-anchoring signal are represented as a grey dashed line and a black square, respectively. RECK Cysteine knot motifs (KN_x), EGF-like repeats and Kazal-like domains (KL_x) are shown in blue, yellow, and green boxes, respectively. Signal Peptide sequence and EGF-like repeats according to (Mendes *et al.*, 2020) and (Takahashi *et al.*, 1998), respectively. Kazal-like domains according to Uniprot. (**B**) RECK structure prediction by AlphaFold (AF; Jumper *et al.*, 2021; Varadi *et al.*, 2022). Cysteine knot motifs, EF-like repeats and Kazal-like domains regions coloured as in (A). Glycosylated asparagine residues and disulphide bonds formed by cysteine residues of the KZ domains are shown in red and black, respectively. SP and GPI-anchoring signal were omitted for clarity. (**C**) RECK structure prediction by AF coloured per-residue confidence score (pLDDT) in a position identical to (B). pLDDT colour legend shown in the top right corner. C- and N-terminal segments (corresponding to the SP, on the left, and to the GPI-anchoring signal, on the right site, respectively are coloured in orange (low pLDDT) which might indicate that they are unstructured segments.

The first clues regarding RECK function were collected by restoring RECK expression in tumour-derived cell lines. *In vitro* and *in vivo* studies demonstrated that recovery of RECK expression levels mitigate the invasiveness of tumour-derived cells and metastasis formation and tumour angiogenesis but did not impact cell viability and growth, chemotactic activity, or motility of these cells (Oh *et al.*, 2001; Rhee, 2002; Takahashi *et al.*, 1998). Furthermore, these studies determined that RECK is involved in the negative regulation of MMP-2, MMP-9, and MMP-14, and, thus in the mediation of tissue remodelling.

Takahashi *et al.*, (1998) reported that expression of both membrane-anchored and secreted forms of recombinant RECK (hRECK and RECKΔC, respectively) in human fibrosarcoma (HT1080) cells repressed their invasiveness *in vitro*. However, RECK location at the plasma membrane is essential for inhibition of pro-MMP-9 secretion, as demonstrated by comparison of the MMP-9 levels in the conditioned medium of cells transfected with RECKΔC and hRECK. Furthermore, they revealed that RECKΔC, partially purified from conditioned medium could bind pro-MMP-9 but not pro-MMP2 and claimed that a purer sample (purity > 95%) could inhibit MMP-9 activity against a synthetic peptide. Posteriorly, Oh *et al.*, (2001) observed that the recombinant expression of RECKΔC also downregulated the levels of active MMP-2 in the conditioned medium of HT1080 cells. Therefore, they tested the inhibitory capacity of purified RECKΔC (sample purity not mentioned) against MMP-2 and MT1-F (an MMP-14 truncated version lacking the transmembrane domain) towards a peptide substrate. On the basis of their results, the authors claimed that RECK can inhibit both MMP-2 and MMP-14. Later, a work developed by Omura *et al.*, (2009) proposed that RECKΔC (with a purity of >90%) inhibits also MMP-7 against the natural protein substrate plasma fibronectin (pFN).

RECK interactions are not limited to members of the MMP family. It has been reported to target two metallopeptidases of the adamalysin family (ADAM-10 and ADAM-17), the serine peptidase urokinase-type plasminogen activator (UPA), the cytokine receptors IL-6 receptor α (IL-6Rα) and gp130, and the epidermal growth factor receptor (EGFR) (Russell *et al.*, 2021).

Due to its function as tissue remodelling mediator, RECK expression is involved in a myriad of physiological functions, namely in skeletal muscle, brain and embryonic development, cartilage differentiation, organogenesis, and angiogenesis and pregnancy (Clark *et al.*, 2007; Guo & Zou, 2006). The high physiological relevance of RECK is demonstrated by the lethality of RECK knock-out mice embryos at E10.5 owing to defects

in collagen fibrils, the basal lamina, and vascular development (Oh *et al.*, 2001). Accordingly, alteration of RECK levels (downregulation or nullification) is associated with pathological conditions like cancer (**Table 6**), inflammation, fibrosis (Dashek *et al.*, 2021) and autoimmune diseases (Hou & Zhang, 2008). There is a vast number of studies reporting RECK altered expression levels with distinct cancer types where, typically, expression levels of RECK and MMPs are proportionally inverse. Furthermore, a great part of these studies suggests RECK as a therapeutic target or as a prognostic or diagnostic marker (Nagini, 2012; C. Zhang *et al.*, 2021), highlighting the potential importance of structural studies of RECK.

2.3. IMPI

Microbial pathogens have developed distinct virulence factors, such as proteolytic enzymes, which stimulate their hosts to create host defence molecules like peptidase inhibitors. While pathogens present a high evolutionary adaptability and are fitted with a broad spectrum of proteolytic enzymes including TLPs, host inhibitors have typically lower genetic plasticity. Despite this, host stimulation has generated an enhanced repertoire of antimicrobial defence molecules in a series of insect species (Vilcinskas, 2010).

Insects, unlike vertebrates, are armed only with an innate immunity system, which is composed of three main pillars. First, anatomical and physiological barriers protect the organism from intruders. When this line of defence is breached, both cellular and humoral response mechanisms are activated. Insect humoral responses are mainly achieved by the synergistic action of several antimicrobial peptides and proteins (AMPs), haemolymph polypeptide components like cecropins, defensins, and proline- and glycine-rich peptides, which possess direct antimicrobial activity (Wojda *et al.*, 2020).

Some AMPs are serine protease inhibitors of the Kunitz, Kazal, serpin and α-macroglobulin families, and they are found in the insect haemolymph, where they not only exert protective functions against microbial proteinases but also regulate endogenous proteases (Kanost, 1999). During immune responses, the larvae of the greater wax moth *Galleria mellonella* produces and secretes distinct peptide and small protein inhibitors of pathogen-associated proteolytic enzymes and other anti-microbial peptides into the haemolymph. Among such proteinase inhibitors are three heat-stable inducible serine

proteinase inhibitors (ISPIs), whose molecular weights range from 6.3 to 9.2 kDa (Fröbius *et al.*, 2000), and the insect metalloproteinase inhibitor (IMPI), which is the first specific microbial metalloproteinase inhibitor discovered in invertebrates (Wedde *et al.*, 1998).

The Vilcinskas group characterised in 1998 native IMPI purified from the haemolymph of *G. mellonella* last-instar larvae after induction of a humoral immune response by bacterial or fungal stimuli. IMPI is a heat-stable and glycosylated peptide with five disulphide bridges and approximately 8.4 kDa molecular mass. The *impi* gene encodes two unique inhibitory peptides (IMPI and IMPI-2), which are obtained by the post-translational cleavage of the 170-amino acid translation product by furin-like proteases (**Figure 18**).

IMPI, which is encoded by the N-terminal part of the gene (upstream of the furin cleavage site), presents a specific activity against TLPs while the C-terminal IMPI-2 is inactive against these metallopeptidases but exhibits weak activity against MMPs (Arolas *et al.*, 2011; Clermont *et al.*, 2004; Wedde *et al.*, 2007). Furthermore, the *impi* gene was found to be activated during *G. melonella* metamorphosis (Altincicek & Vilcinskas, 2006), supporting the notion that the IMPI precursor encodes two unique inhibitory small proteins: IMPI, whose inhibitory activity is inextricable of the insect innate immune system, and IMPI-2, which modulates endogenous MMPs during metamorphosis (Wedde *et al.*, 2007).

The crystal structure of IMPI complexed with thermolysin obtained by Arolas *et al.* (2011) (**Figure 18D**) unveiled that this low-molecular-weight inhibitor displays a spearhead shape with a rhombic base, whose "reactive-centre loop" (RCL) spans from P_{53} to R_{58} and protrudes from the molecular structure, with I_{57} being the tip of the spearhead. IMPI is laterally inserted into active-site cleft of thermolysin, like a slice into a mouth, and the protease-inhibitor interaction is limited to the small surface derived from the RCL. Notably, thermolysin cleaves the IMPI RCL between $N_{56}\downarrow I_{57}$ in a substrate-like manner. However, as IMPI preserves its overall conformation due to its extensive disulphide network, the cleaved version retains full capacity to bind and inhibit thermolysin.

Despite its function as a thermolysin-like metallopeptidase inhibitor, IMPI shares no similarity with tissue inhibitors of metalloproteinases (TIMPs) or any other metallopeptidase inhibitor presently reported for its structure. Unexpectedly, its amino acid sequence showed significant similarity with the trypsin inhibitor-like (TIL) cysteine-rich domain, a domain typical of serine protease inhibitors (Arolas *et al.*, 2011; Clermont *et al.*, 2004). Furthermore, IMPI presents a unique inhibitory mechanism contrasting with other metallopeptidase protein inhibitors who are not typically cleaved by their target protease. This mechanism

resembles those of several serine peptidase inhibitors, which are proteolytically modified by their affiliated proteases (Arolas *et al.*, 2011). These observations indicate that IMPI and some serine peptidase inhibitors might share a common ancestor (Clermont *et al.*, 2004).

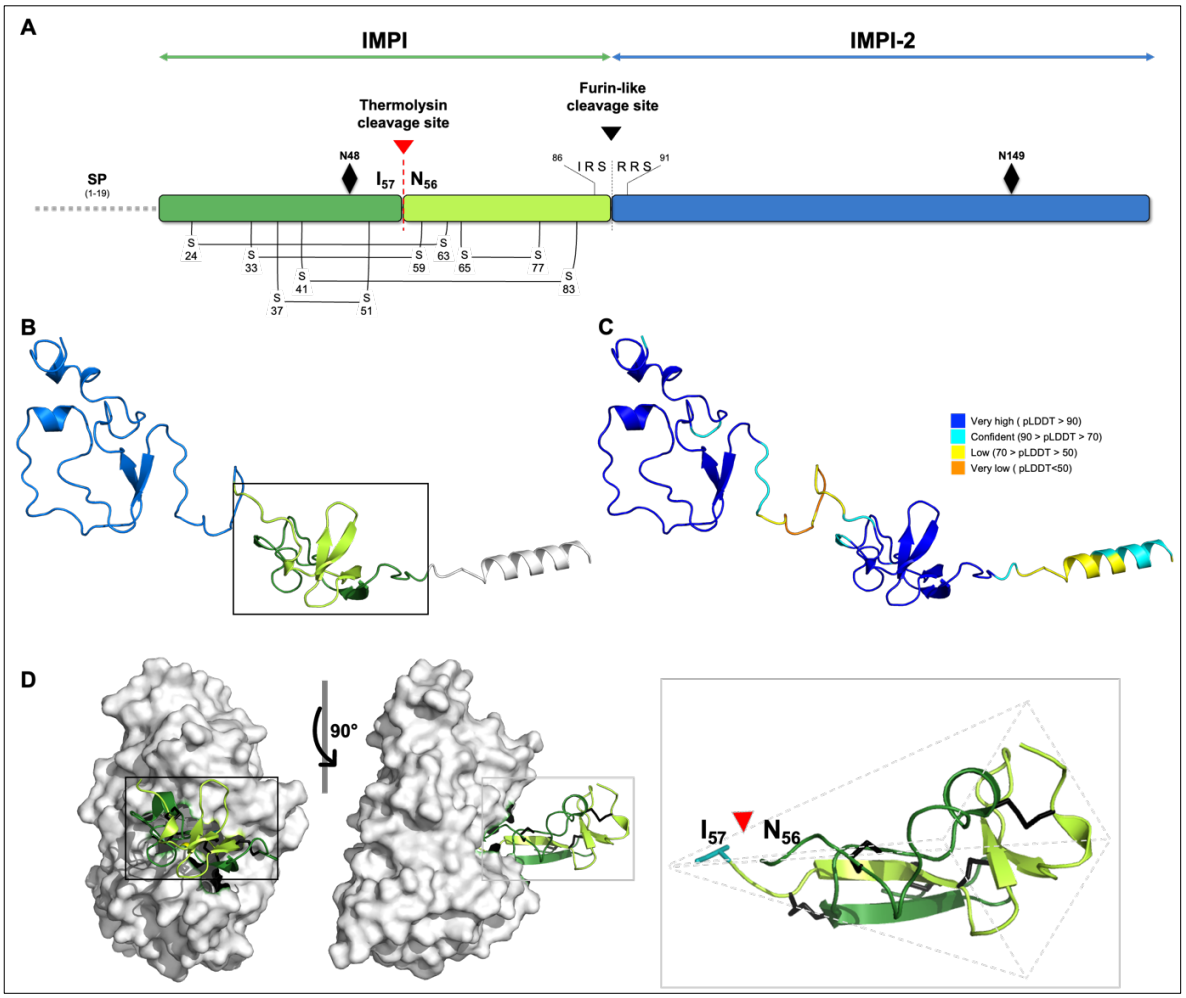


Figure 18: Schematic representation of IMPI domains organization and structure. (**A**) Domain arrangement of IMPI from the greater wax moth *Galleria melonella* (UP P82176). *impi* gene translation product comprises IMPI and IMPI-2 proteins (green and blue boxes, respectively). Furin recognition motif (▼), disulphide bonds and glycosylation sites (◆) protrude from the respective domains with key residues identified. Thermolysin cleavage site (N_{56}↓I_{57}) identified with a red arrow (▼). (**B**) and (**C**) IMPI structure model predicted by AlphaFold (AF; Jumper *et al.*, 2021; Varadi *et al.*, 2022) coloured as in (A) or per-residue confidence score (pLDDT), respectively. Note: The AlphaFold models were oriented to reflect the orientation of IMPI in the thermolysin complex structure further below. **D**) Crystal structure of recombinant IMPI (I_{20}-S_{88}; in green) in complex with prototypical Thermolysin (in grey and in frontal position as in Figure 10) (PDB 3ssb; Arolas *et al.*, 2011). At the left, the complex front-view with IMPI in the same position than in the AF model (black rectangular frame). In the middle, the complex turned 90° along the y axis. On this side-view, the interaction of the peptide inhibitor RSL with the deep thermolysin active-site cleft is shown. At the right, a magnification of IMPI is shown. IMPI is in the exact same position than in the complex-side view (grey rectangular frame). IMPI is spearhead shaped with a rhomboid base (underlined by dashed grey lines) and cleaved when in complex with thermolysin (thermolysin cleavage site indicated by a red arrow (▼). The cysteine residues that form disulphide bonds and the isoleucine residue which shapes IMPI spear tip are shown in black and light blue, respectively.

Proteolytic cleavage of haemolymph constituents by pathogenic (Altincicek *et al.*, 2009) or endogenous (Altincicek & Vilcinskas, 2008) secreted metalloproteases generates low.molecular-weight (MW) degradation products, so-called "protfrags", which strongly induce expression of AMP coding genes, including IMPI (Error! Reference source not found.) (Vilcinskas & Wedde, 2002). In this way, IMPI expression is induced in *G. mellonella* larvae challenged by mammalian or entomo-pathogens regardless of their fungal (Vertyporokh & Wojda, 2017, 2020; Woolley *et al.*, 2020) or bacterial (Asai *et al.*, 2021; Mukherjee *et al.*, 2010; Wojda & Taszłow, 2013) nature.

Similarly, previous exposition to abiotic (temperature or mechanic shock (Mowlds *et al.*, 2008; Mowlds & Kavanagh, 2008; Wojda & Taszłow, 2013) or biotic (non-lethal doses of entomopathogens or antifungal compounds; Bergin *et al.*, 2006; Kelly & Kavanagh, 2011; Vertyporokh & Wojda, 2020) stress instigates expression of IMPI and other AMPs, whose synergistic activity improves the survival outcomes of infected *G. mellonella* larvae.

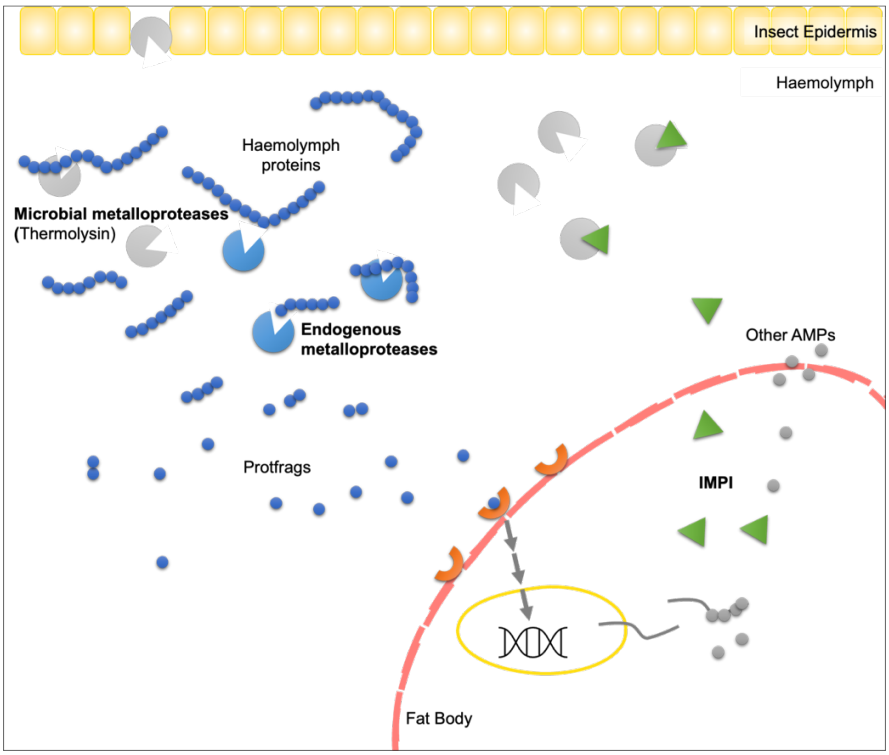


Figure 19: Illustrative representation of induction of IMPI and other AMPs expression in *G. mellonella* larvae. Microbial metallopeptidases secreted into larva haemolymph by entomopathogens that cross the insect physical barriers, together with endogenous metalloproteinases, cleave haemolymph proteins into small fragments of ~3 kDa (Protfrag), which through receptor binding induce expression of antimicrobial peptides (AMPs), such as IMPI, in the larvae fat body. Adapted from (Wojda *et al.*, 2020).

It is important to emphasize that IMPI is the only known specific protein inhibitor of TLPs, which include key virulence factors secreted by human pathogens such as *Staphylococcus aureus* and *Pseudomonas aeruginosa*. Additionally, there are no known members of this family of peptidases in animals, thus making IMPI a potential and promising drug-target candidate (Eisenhardt *et al.*, 2019; Mendes *et al.*, 2022).

Objectives

This thesis compiles the results collected from the different projects in which I was involved during my Ph.D. studies at the IBMB-CSIC in Barcelona (2017-2022): (1) the RECK protein and its interaction with distinct MMPs, (2) the interplay of the protein inhibitor IMPI and the metallopeptidase aureolysin, and (3) the intricate inhibitory mechanism of α2M.

In addition, I significantly contributed also to other projects in the lab, two of which I have enclosed to the thesis as supplementary data due to their recent publication. One was spearheaded by my Ph.D. colleague, Laura del Amo Maestro, who extensively characterised a glutamic endopeptidase, which proficiently degrades the toxic peptides derived from gluten digestion that trigger coeliac disease. After Laura had finished her PhD, I took over the project and optimized purifications and enzymatic characterizations of various mutants (del Amo Maestro L., Mendes S.R., *et al.* 2022, accepted by Nature Communications). In the other project, we teamed up with the Protein Design and Modelling Lab of Enrique Marcos at IBMB. We focused on the *de novo* design of immunoglobulin-like β-sandwich scaffolds as a new class of antibody derivatives, which can be functionalized inserting hypervariable antigen-binding loops. Such designed molecules could tremendously expand our repertoire of protein-based drugs for a variety of diseases (Chidyausiku T.M, Mendes S.R., *et al.* 2022, under second revision at Nature Communications).

Importantly, all of my PhD projects focused on the development of therapeutic proteins or peptides (TPs), with a significant emphasis on the complex interplay of proteases and their inhibitors in health and disease.

RECK function is essential for embryogenesis, organogenesis, and tumour progression, most likely due to its involvement in the regulation of MMPs activity. Of particular interest is its involvement in tumour progression: it has been reported that decreased RECK levels are associated with higher invasiveness and metastasis of tumours and, consequently, with worse prognosis for cancer patients. RECK has been suggested not only as a diagnostic and prognostic marker but also as a potential cancer therapeutic target, which prompted us to develop the first project of this thesis (Project 1) with the following goals:

- Establishment of bacterial and eucaryotic (both insect and mammalian) expression systems for the production of different RECK variants and subsequent establishment of an extensive purification protocol.

- Biochemical assessment of the interplay between MMPs and RECK through the evaluation of its inhibitory capacity against different MMPs (MMP-2, MMP-7, MMP-9, and the catalytic domain of MMP-14) using fluorogenic peptides and natural protein substrates.
- Elucidation of RECK's three-dimensional structure and inhibitory mechanism through X-ray crystallography and/or cryo-EM.

Protease inhibitors are critical not only for the control of endogenous proteases but also for the regulation of proteases produced by invading pathogens as virulence factors. An enthralling example of such interplay is the induced expression of the insect metallopeptidase inhibitor (IMPI) in the greater wax moth after infection or laboratorial infectious stimuli. IMPI is a potent specific protein inhibitor of thermolysin from *Bacillus thermoproteolyticus Rokko* and other thermolysin-like proteases (TLPs). TLPs of particular interest are those secreted by human pathogens, among which aureolysin, a metalloprotease secreted by *Staphylococcus aureus*. The emergence of *Staphylococcus aureus* strains associated with antibiotic resistance makes it imperative to further study the proteases implicated in the virulence mechanisms of these pathogens and their putative inhibitors. With this in mind, we endeavoured to establish the second project presented in this thesis (Project 2) whose main objectives were:

- Establishment of a purification protocol of secreted aureolysin from *S. aureus* strain V8-BC10 cultures.
- Assessment of IMPI as an effective inhibitor of aureolysin and comparison with its inhibitory capacity against thermolysin.
- Elaboration of a set of IMPI variants through single or multiple mutations in the IMPI reactive-centre loop and test of their inhibitory activity in comparison to wild-type IMPI.
- Elucidation of the IMPI inhibitory mechanism against aureolysin through structure determination of the aureolysin-IMPI complex and the inhibitory profiles of the generated IMPI-mutant library.

The main protease inhibitor found in human plasma, α2-macroglobulin (hα2M), has been exhaustively studied along several years in our laboratory. This large homotetrameric glycosylated protein presents a permissive suicidal inhibitory mechanism, known as "Venus

flytrap", which exerts inhibition of up to two peptidase molecules of disparate classes. The native inhibitor, which displays an open conformation, closes upon proteolytic cleavage of a bait region, thereby entrapping the peptidase molecules that, despite remaining catalytically active, cannot access their physiological substrates anymore. The third project of this thesis (Project 3) started as a long collaborative work developed by researchers from different institutions whose main purpose was to structurally elucidate the molecular mechanism of the inhibitory trap unique to hα2M. The results were presented in a prominent publication (Luque *et al.*, 2022).

Influenced by remarks during review of the manuscript, we further compared the activation state of hα2M samples purified from non-frozen versus frozen fresh plasma using a cohort of functional and biophysical experiments. Therefore, in the chapter about Project 3, I present the "ongoing work" in addition to results of the published work, which entailed:

- Demonstration that the quality of hα2M preparations purified from (i) frozen versus (ii) non-frozen fresh plasma (within less than 24h after blood donation) is equivalent and without significant differences at the biophysical and functional level.

Results

MINISTRY OF SCIENCE AND INNOVATION
CSIC HIGHER SCIENTIFIC RESEARCH COUNCIL
MOLECULAR BIOLOGY INSTITUTE OF BARCELONA (IBMB)
Parc Científic de Barcelona
UNIVERSITAT DE BARCELONA

INFORME DEL DIRECTOR DE TESI

30 de Juny de 2022

Per la present vull confirmar que la Soraia Inês dos-Reis Mendes ha contribuït de forma excepcionalment significativa a la ciència durant la seva tesis, fet que ha donat lloc a la publicació dels tres articles següents:

Soraia R. Mendes, Laura del Amo-Maestro, Laura Marino-Puertas, Iñaki de Diego, Theodoros Goulas & F. Xavier Gomis-Rüth (2020). Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases. ***Sci. Rep.*, 10**, 6317.

Soraia R. Mendes, Ulrich Eckhard, Arturo Rodríguez-Banqueri, Tibisay Guevara, Peter Czermak, Enrique Marcos, Andreas Vilcinskas & F. Xavier Gomis-Rüth (2022). An engineered protein-based submicromolar competitive inhibitor of the *Staphylococcus aureus* virulence factor aureolysin. ***Comput. Struct. Biotechnol. J.*, 20,** 534–544.

Daniel Luque, Theodoros Goulas, Carlos P. Mata, **Soraia R. Mendes**, F. Xavier Gomis-Rüth & José R. Castón (2022). Cryo-EM structures show the mechanistic basis of pan-peptidase inhibition by human α_2 -macroglobulin. ***Proc. Natl. Acad. Sci. USA*, 119**, e2200102119.

En tots tres articles, la seva contribució va ser decisiva, en dos d'ells és la primera autora.

Així mateix, és la primera autora de a manuscrit més que es troba en la fase final de redacció i que es presenta a la seva tesi:

Soraia R. Mendes, Theodoros Goulas & F. Xavier Gomis-Rüth (2022). One single plasma freeze-thaw cycle does not affect α_2M homogeneity.

A més a més, també va participar de forma molt rellevant en el següents articles actualment sota revisió a la revista ***Nature Communications***, en que és co-primera autora. Aquests articles no es discuteixen a la seva memòria de tesis.

Laura del Amo-Maestro[#], **Soraia R. Mendes**[#], Arturo Rodríguez-Banqueri, Laura Garzon-Flores, Marina Girbal, María José Rodríguez-Lagunas, Tibisay Guevara, Àngels Franch, Francisco J. Pérez-Cano, Ulrich Eckhard and F. Xavier Gomis-Rüth (2022). Molecular and *in vivo* studies of a glutamate-class prolyl-endopeptidase for coeliac disease therapy.

Tamuka M. Chidyausiku [#], **Soraia R. Mendes**[#], Jason C. Klima, Marta Nadal, Ulrich Eckhard, Jorge Roel-Touris, Scott Houliston, Tibisay Guevara, Hugh K. Haddox, Adam Moyer, Cheryl H. Arrowsmith, F. Xavier Gomis-Rüth, David Baker & Enrique Marcos (2022). *De Novo* design of immunoglobulin-like domains.

Degut a tots aquests mèrits, no hi ha cap dubte que la Sra. Mendes ha realitzat una feina extraordinària, fóra de sèrie, que supera amb escreix el que es d'esperar d'una tesi doctoral.

F. Xavier Gomis-Rüth
Professor d'Investigació CSIC

Proteolysis Lab
IBMB-CSIC
Barcelona Science Park; Helix Building
c/ Baldiri Reixac, 15-21
08028-Barcelona (Catalonia, Spain)
e-mail: xgrcri@ibmb.csic.es
Lab homepage: https://www.ibmb.csic.es/proteolysis

Telf. +34-934 020 186
Fax: +34-934 034 979
VAT: Q2818002D

The **Results** presented here are the compilation of the work developed in three separate projects, all aiming to elucidate the involvement of proteases in pathological conditions and to strengthen the knowledge about protein and peptide inhibitors of proteases with therapeutic potential. The results culminated in three publications, which are here presented as sub-chapters **Project 1**, **Project 2**, and **Project 3**.

An additional manuscript elucidating the effect of a single freeze-thawing cycle of the plasma sample on the structural and functional homogeneity of α2-macroglobulin purified from it is currently in preparation (Mendes S.R. *et al.*). The current status of results is summarized at the project 3 subchapter "Work under development: *α2M samples purified from frozen and unfrozen fresh plasma present no significant structural or functional differences"*.

Furthermore, the work developed during my PhD resulted in two more research manuscripts where I am shared first author, which are included in the Supplementary Materials (del Amo Maestro L., Mendes S.R., *et al.* 2022; and Chidyausiku T.M, Mendes S.R., *et al.* 2022; accepted by and under second revision at *Nature Communications*, respectively).

Project 1

"*Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases*"

"*Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases"*

Soraia R. Mendes, Laura del Amo-Maestro, Laura Marino-Puertas, Iñaki de Diego[1], Theodoros Goulas & F. Xavier Gomis-Rüth *

Proteolysis Laboratory; Department of Structural Biology; Molecular Biology Institute of Barcelona; Higher Scientific Research Council (CSIC); Barcelona Science Park, Helix Building; Baldiri Reixac, 15-21; 08028 Barcelona (Catalonia, Spain).
[1] Present address: ALBA Synchrotron Light Source; Carrer de la Llum, 2-26; 08290 Cerdanyola del Vallés (Catalonia, Spain).
*Corresponding author: e-mail: xgrcri@ibmb.csic.es.

Published in: Scientific Reports (2020)
Impact factor: 4.38
Quartile: Q1

The present publication explored whether this proposed regulator of MMPs can directly inhibit their catalytic activity. Moreover, it discloses three different expression systems established for the expression of RECK variants, a protein with tumour suppressor functions and thus high interest in cancer research, and highlights recommended improvements for protein purifications from Expi293F cells.

In this first project, I cloned several RECK variants into the respective expression vectors and produced them in bacterial, insect, and mammalian cells. Furthermore, after detection of a contaminant with proteolytic activity in partly purified RECK samples, I established an exhaustive purification protocol to ensure removal of the respective contaminant, which was posteriorly used by some of my PhD colleagues that detected the very same background activity in their preparations. Subsequently, I used highly pure RECK protein for evaluation of inhibitory activity against MMP-2, MMP-7, MMP-9 and the catalytic domain of MMP-14, the latter produced and purified by myself in the lab from inclusion bodies). RECK inhibitory capacity against these MMPs was evaluated using both quenched fluorescent peptides (QF-peptides) and natural protein substrates.

Summary

Reversion-inducing cysteine-rich protein with Kazal motifs (RECK) is a glycosylated protein pivotal for embryogenesis and tumour progression. Notably, its involvement in these processes was linked in parts to its reported role as a negative regulator of matrix metalloproteinases (MMPs), a family of metalloproteases whose proteolytic activity against components of the extracellular matrix is essential for a myriad of (patho)physiological processes. Here we established a bacterial, insect, and mammalian expression systems for the production of different RECK proteoforms. We further observed that RECK sample purified from the conditioned medium of Expi293F cells, but not from the other eucaryotic system, comprised a contaminant with proteolytic activity, but which was inhibited by the serine peptidase inhibitor AEBSF. Subsequent adaptation of the purification protocol yielded RECK of highest purity and lacking the AEBSF-sensitive activity, which was then used in the inhibitory assays against MMPs. Importantly, no significant inhibition of the MMP activity was observed, suggesting that RECK isn't a direct inhibitor of MMPs and that the previously reported regulation of these proteases might occur at a different level and/or through other mechanisms.

Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases

Soraia R. Mendes, Laura del Amo-Maestro, Laura Marino-Puertas, Iñaki de Diego[1], Theodoros Goulas & F. Xavier Gomis-Rüth *

Proteolysis Laboratory; Department of Structural Biology; Molecular Biology Institute of Barcelona; Higher Scientific Research Council (CSIC); Barcelona Science Park, Helix Building; Baldiri Reixac, 15-21; 08028 Barcelona (Catalonia, Spain).

[1] Present address: ALBA Synchrotron Light Source; Carrer de la Llum, 2-26; 08290 Cerdanyola del Vallés (Catalonia, Spain).

*Corresponding author: e-mail: xgrcri@ibmb.csic.es.

Matrix metalloproteinases (MMPs) occur in 23 human paralogues with key functions in physiology, and their activity is controlled by protein inhibitors. Reversion-inducing cysteine-rich protein with Kazal motifs (RECK), which is essential for embryogenesis and tumour suppression, has been reported to inhibit MMPs. Here, we developed eukaryotic and bacterial expression systems for different RECK variants and analysed their inhibitory capacity against representative MMPs *in vitro*. We could not detect any significant inhibition. Instead, we found that partially purified RECK from the conditioned medium of transfected Expi293F cells but not that of ExpiCHO-S or *Drosophila* Schneider cells contained a contaminant with proteolytic activity. The contaminant was removed through treatment with a small-molecule serine peptidase inhibitor and additional chromatographic purification. A tantamount contaminant was further detected in an equivalent expression system of the N-terminal fragment of the proteoglycan testican 3, but not in those of two other proteins. These results indicate that previous reports of inhibitory activity of recombinant RECK on MMPs, which were performed with partially purified samples, were probably masked by a coeluting contaminant present in the supernatant of HEK293-derived cells. Thus, RECK is probably not a direct inhibitor of MMP catalytic activity but may still regulate MMPs through other mechanisms

Proteolysis is a post-translational modification of proteins and peptides that is essential for all physiological pathways. It is exerted by peptidases, among which metallopeptidases (MPs) are one of several chemical classes and consist of various clans and families [1]. The matrix metalloproteinases (MMPs; 23 paralogues in humans), as well as the ADAMs/adamalysins (19 in humans) and the more distantly related ADAMTSs (19 in humans) are among the most studied MPs because of their enormous relevance for human health and disease [2-8]. Collectively, they carry out functions as broad degraders during the digestion of intake proteins, turnover of extracellular-matrix components for tissue remodelling and developmental processes, and clearance of obsolete or malfunctioning polypeptides. Moreover, they are fine regulators of shedding, maturation and inactivation of other proteins through limited proteolysis of one or a few peptide bonds [9]. Peptide-bond scission is normally irreparable under physiological

conditions, so MPs must be fastidiously controlled to avoid aberrant cleavage that would cause dysfunction and pathology. This regulation is physiologically carried out for MMPs in humans by four tissue inhibitors of metalloproteinases and the broad-spectrum pan-peptidase inhibitor α_2-macroglobulin [3,10-12]. Another reported inhibitor is the protein RECK [13-17].

RECK, an acronym for reversion-inducing cysteine-rich protein with Kazal motifs, is encoded by a gene that suppresses the transformed phenotype caused by *ras* oncogenes [13,18]. The 971-residue molecule is a membrane-anchored glycoprotein of ~125 kDa, which contains an N-terminal signal peptide for secretion, a region spanning five cystine knots (KNs; KN1-KN5), a region with three repeats similar to Kazal inhibitors of serine endopeptidases (KLs; KL1-KL3) [19,20], and a C-terminal segment (CTS; residues A^{943}–N^{971}; for numbering, see UniProt database entry [UP] O95980) (Fig. 1). The CTS is removed during maturation, which leads to binding of RECK to the plasma membrane through a glycosylphosphatidylinositol anchor attached to S^{942} [13,21]. In addition, *N*-linked glycosylations have been determined at residues N^{200}, N^{86}, N^{297} and N^{352}, and the latter three are essential for function [22].

Physiologically, RECK is critical because knockout mice die during embryonic development with severe tissue, vascular, and neuronal defects [14,23]. It is highly expressed in most normal tissues and non-transformed cells, and is probably involved in myogenesis, chondrogenesis, patterning during embryogenesis, and the establishment of the neuromuscular junction [13,17,24]. It also participates in Notch-dependent neurogenesis, Wnt signalling and brain angiogenesis [25-27]. Moreover, it has been implicated in tumour processes including growth, angiogenesis, invasion, metastasis and relapse [14,28,29]. It is downregulated with poor prognosis for disease outcome in pancreatic, oral, breast, prostate and non-small cell lung cancers, as well as in osteosarcoma [30-35]. Consistently, restored expression of RECK in tumour cells suppresses angiogenesis, invasion and metastasis in animal models, and the level of residual RECK expression in tumour tissues correlates with better prognosis [17]. All these findings suggest that RECK has a potential therapeutic value as a tumour suppressor for the treatment of malignant conditions [21].

At the molecular level, RECK binds ADAMTS-10, which is involved in connective-tissue development[36], and impairs modulation of Notch signalling during neurogenesis mediated by ADAM-10 [23]. RECK without the CTS (RECKΔC) has further been described to prevent cleavage by ADAM-10 of a fragment of protein Delta1 and of a synthetic substrate with an apparent inhibition constant (K_i) of <15 nM [23]. Moreover, RECKΔC has been claimed to directly inhibit cleavage of fluorogenic peptide substrates by MMP-2, MMP-9 and MMP-14 with associated K_i values of 20-80 nM [13-15,17], and of plasma fibronectin by both MMP-2 and MMP-7, the latter with a K_i of ~41 nM [17]. In a report by another group [16], tagged RECKΔC and shorter constructs spanning the two C-terminal Kazal-like motifs (P^{676}–V^{799}), the cysteine knots (residues L^{28}–R^{368}) and all three Kazal-like motifs (residues V^{799}), respectively, were assayed for inhibition of MMP-9 with fluorescein-conjugated gelatine as substrate. The authors have reported that the two former constructs but not the two latter significantly inhibit gelatine cleavage [16]. Based on all these reports, RECK has been included as an MMP inhibitor in the MEROPS database under family I1, which groups the Kazal family of inhibitors of serine endopeptidase families S1 and S8 [19,20,37].

Prompted by these results, we embarked on a long-term project to characterize the structure and function of RECK. To this aim, we here produced several constructs of the protein with the highest purity as a requisite for molecular and

biophysical studies, and we assayed their inhibitory capacity against MMPs *in vitro*.

MATERIALS AND METHODS

Expression vectors – Plasmid pBS-hRECK (for details on constructs, plasmids, vectors and primers, see Table 1) encoding full-length human RECK cDNA in the pBlueScript vector was kindly provided by Makoto Noda, Kyoto (Japan). Constructs encoding fragments RECKΔC (plasmid pS6-RECKΔC; residues G^{27}–S^{942}), KL123 (pCri9a-KL123; S^{621}–S^{797}) and KL23 (pCri9a-KL23; T^{697}–S^{797}) (Fig.1), as well as the coding sequence for residues A^{22}–Q^{313} (UP Q9BQ16) of the N-terminal region of human testican 3 (N-TES; plasmid pS6-NTES) in a synthetic gene (from GenScript) were amplified with primers that introduced sites for directional cloning. For bacterial expression, plasmid pCri9a [38], which adds a C-terminal hexahistidine (His_6)-tag, was used for insertion between the *Nco*I and *Xho*I restriction sites. For expression in mammalian cells, vector pCMV-Sport 6 (Thermo Scientific) with the original signal peptide (for RECKΔC) or with the mouse immunoglobulin κ leader sequence (for N-TES) was used to insert genes between *Sma*I/*Asi*SI and *Bst*EII/*Asi*SI restriction sites, respectively. For expression of RECKΔC in insect cells, the RECK gene was inserted into vector pIEx (Novagen) by restriction-free cloning in frame with the signal peptide of adipokinetic hormone as previously described [39] to yield plasmid pIE-RECKΔC. Primers and DNA-modifying enzymes for polymerase chain reaction (PCR) steps were purchased from Sigma-Aldrich and Thermo Scientific, respectively. PCR was performed with Phusion High Fidelity DNA polymerase (Thermo Scientific) according to the manufacturer's instructions with an extra optimization step by thermal gradient after each reaction. DNA was purified with the OMEGA Biotek Purification Kit (Omega) or GeneJET Plasmid MaxiPrep Kit (Thermo Scientific) according to the manufacturer's instructions, and all constructs were verified by DNA sequencing. Plasmid pET3a-MT1ΔC (see Table 1) encoding the pro- and catalytic domains of MMP-14 (S^{24}–G^{284}; UP P50281) *plus* an extra N-terminal methionine [40] was kindly provided by Yoshifumi Itoh, Oxford (UK).

Eukaryotic cell culture and transient transfection – *Drosophila melanogaster* embryonic Schneider cells (S2; Gibco) adapted to suspension, as well as the HEK293-derived Expi293F cells (Expi; Gibco) and ExpiCHO-S derived from Chinese hamster ovary cells (Expi-CHO; Gibco), were maintained in Sf-900 II SFM and FreeStyle F17 expression medium (Gibco) for insect and human cells, respectively. Both media were supplemented with 0.5 µg/mL amphotericin B (Gibco), 100 units/mL penicillin and 100 µg/mL streptomycin (Sigma).

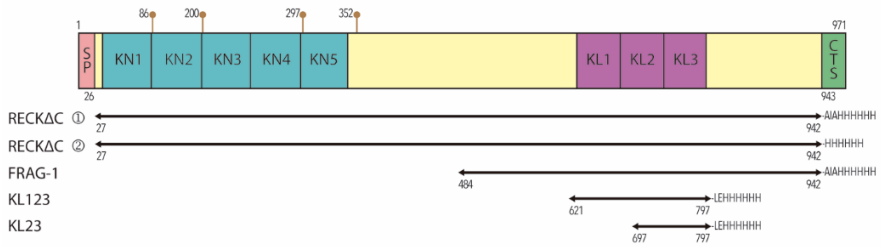


Figure 1. Overview of RECK constructs. Scheme depicting the domain structure of human RECK. SP, signal peptide; KN1–KN5, cysteine knot regions; KL1–KL3, Kazal-like domains; and CTS, C-terminal segment. The distinct constructs that were studied in this work are shown and labelled. 1, RECKΔC produced in Expi or Expi-CHO cells. 2, RECKΔC produced in S2 cells. Brown lollipops pinpoint glycosylation sites according to [22].

Plasmid name	Protein	Parental vector(s)	Forward-primer*	Reverse-primer *	Protein sequence **	Tag***
pS6-RECKΔC	RECKΔC	pBS-hRECK pCMV-Sport6	ATGCCCCGGGATGGCGAC CGTCCGG	GCATGCGATCGCCGATGG CACACTGCTG	G^{27}-S^{942}+**AIA+H_6**	His_6
pIE-RECKΔC	RECKΔC	pBS-hRECK pIEx	TCATCGCTTTCGTCATCAT CGCTGGGCCTGGCTCCGG GCAGTGCGGGTG	AAACTCAATGGTGATGGT GATGATGCGATGGCACAC TGCTGGAGACCTGT	G^{27}- S^{942} + **H_6**	His_6
pS6-NTES	N-TES	Synthetic DNA pCMV-Sport6	ATGCGGTGACCTAGCTGC CGCGGCGGT	GCATGCGATCGCTTGCTG TCTCTGGAAGCA	**L** + A^{22}-Q^{313}+**AIA+H_6**	His_6
pCri9a-KL23	KL23	pS6-RECKΔC pCri9a	ATGCCCATGGTAACGACT TTTGATAA	GCATCTCGAGGCTGTGCT CTGAGAGG	**V**+T^{697}-S^{797}+**LE+H_6**	His_6
pCri9a-KL123	KL123	pS6-RECKΔC pCri9a	ATGCCCATGGTATCAGAA GATGACCG	GCATCTCGAGGCTGTGCT CTGAGAGG	**V**+S^{621}-S^{797}+**LE+H_6**	His_6
pET3a-MT1ΔC	MMP-14 CD	pET3a	–	–	**M**+S^{24}-G^{284}	None

Table 1. Constructs, primers, plasmids and proteins. All constructs are for extracellular expression of the respective proteins. *Restriction-site sequences and overhangs for restriction-free cloning are underlined. **Peptide sequence of the expressed protein. Amino acids derived from the construct are in bold. See also Fig. 1. ***Tag fused to the carboxy-terminus.

Additionally, FreeStyle F17 medium was supplemented with 8 mM L-glutamine and 0.2% Pluronic F-68 (Gibco). Expi and Expi-CHO cells were grown to a density of 3–5×10^6 cells/mL and 4–6×10^6 cells/mL, respectively, and subcultured every 3–4 days by dilution to 0.3–0.5×10^6 cells/mL and 0.2–0.3×10^6 cells/mL, respectively. To this aim, they were incubated at 37°C in a Multitron Cell Shaker Incubator (Infors HT) at 150 rpm in humidified atmosphere with 8% CO_2. Cells were then subcultured to 0.7×10^6 cells/mL and transfected after 24 h at a cell density of 1×10^6 cells/mL with a dropwise added mixture of 1 mg of vector DNA (see Table 1) and 3 mg of linear 25-kDa polyethyleneimine (Polysciences Europe) in 20 mL of Opti-MEM medium (Gibco) per litre of expression medium. The mixture had been previously incubated at room temperature for 15–20 min. After 3 days, the cell-culture supernatant was harvested for protein purification.

S2 cells were grown to a density of 12–16×10^6 cells/mL, subcultured by dilution to 4×10^6 cells/mL every 3–4 days and incubated at 28°C in an Innova 42 Incubator Shaker (New Brunswick Scientific) under agitation at 200 rpm. Cells were subcultured to 6×10^6 cells/mL and transfected after 24 h at a cell density of 12×10^6 cells/mL with a dropwise added mixture of 0.6 µg of DNA (see Table1) and 2 µg of linear 25-kDa polyethyleneimine per 10^6 cells. The mixture had been previously incubated at room temperature for 15–30 min. Transfected cells were diluted to 4×10^6 cells/mL after 1 h incubation at 28ºC under agitation at 200 rpm, and the cell-culture supernatant was harvested after 7 days for protein purification.

Bacterial expression – Plasmids pCri9a-KL123 and pCri9a-KL23 were transformed into competent Lemo21 (DE3) *Escherichia coli* cells (New England Biolabs) and plated on Luria-Bertani (LB) plates. Fifty millilitres of lysogeny broth was inoculated with a single bacterial colony and incubated overnight at 37°C under stirring at 220 rpm. Five millilitres of this preinoculum was used to inoculate 500 mL of lysogeny broth, and cells were left to grow at 37°C until OD_{600}≈0.7. Subsequently, cultures were cooled to 20°C and protein expression was induced with 0.4 mM isopropyl-β-D-1-thiogalactopyranoside (IPTG; Duchefa) for 18–20 h. LB plates and lysogeny broth were supplemented with 50 µg/mL kanamycin (Fisher Bioreagents) and 34 µg/mL chloramphenicol (Fluka).

For the expression of MMP-14 catalytic domain (CD), *E. coli* BL21 (DE3) cells (Sigma) were transformed with plasmid pET3a-MT1ΔC. One hundred millilitres of lysogeny broth was inoculated with a single colony and incubated overnight at 28°C under stirring at 200 rpm. Ten millilitres of this

preinoculum was used to inoculate 500 mL of lysogeny broth, and cells were left to grow at 37°C until OD_{600}≈0.6. Cells were then induced with 0.5 mM IPTG and kept for 5 h at 37°C. LB plates and lysogeny broth were supplemented with 100 µg/mL ampicillin (Apollo Scientific).

Protein purification – For purification of RECKΔC from Expi cells, cell-culture supernatant was cleared at 4°C by centrifugation at 3,500×*g* for 30 min, filter-sterilized and concentrated 20-fold with a VivaFlow 200 Cross Flow Cassette device with a Hydrosart membrane of 30-kDa cutoff (Sartorius). Concentrated supernatant was then dialysed against a 75-fold volume excess of buffer 20 mM Tris·HCl pH 7.5, 150 mM sodium chloride. After addition of 20 mM imidazole to the dialysed supernatant, RECKΔC was captured by nickel-nitrilotriacetic acid (Ni-NTA) affinity chromatography (AC) in a HisTrap HP column (GE Healthcare) previously washed with buffer A (50 mM Tris·HCl pH 7.5, 1M sodium chloride, 500 mM imidazole) and equilibrated with buffer A without imidazole. The protein was washed and eluted with a step gradient of imidazole (2%, 4%, 12% and 60% of buffer A). The presence of a proteolytic impurity in fractions containing RECKΔC was assessed through incubation with 1 mg/mL fibrinogen from human plasma (Sigma) in buffer B (50 mM Tris·HCl pH 7.5, 150 mM sodium chloride, 5 mM calcium chloride, 50 µM zinc chloride) overnight at 37°C. To remove this contaminant and obtain highly pure RECKΔC, the protein was incubated for 1 h at room temperature with 1 mg/mL **4-[2-aminoethyl]benzenesulfonyl fluoride (AEBSF, commercial name** Pefabloc, Sigma) and further purified by size exclusion chromatography (SEC) in a Superdex 200 (GE Healthcare) column equilibrated with buffer A without imidazole.

For the production and purification of RECK construct FRAG-1 (see Fig.1), highly purified RECKΔC was incubated with 20-fold molar excess of MMP-14 CD in buffer B overnight at 37°C. Cleavage fragments were purified by SEC in a Superdex75 (GE Healthcare Life Sciences) column equilibrated with buffer C (50 mM Tris·HCl pH 7.5, 150 mM sodium chloride). Presence of proteolytic activity in the fractions containing FRAG-1 was assessed as above with fibrinogen.

For purification of RECKΔC from S2 or Expi-CHO cells, cleared cell culture supernatant was dialysed against a 17-fold volume excess of buffer 20 mM Tris·HCl pH 7.4, 250 mM sodium chloride, 20 mM imidazole. RECKΔC in the supernatant was captured by AC with Ni-NTA resin (Thermo Scientific) by overnight incubation at 4ºC. It was subsequently loaded onto an open column for batch purification (Bio-Rad), and washed extensively and eluted with 4% and 60% of buffer A, respectively. The presence of proteolytic activity was assessed as above with fibrinogen. Partially purified protein was further purified by SEC in a Superdex 200 10/300 (GE Healthcare) column equilibrated with buffer C (RECKΔC from S2 cells) or buffer A without imidazole (RECKΔC from Expi-CHO cells).

For N-TES purification, cleared cell culture supernatant was supplemented with 20 mM imidazole and incubated for 3-4 h with Ni-NTA resin. It was subsequently loaded onto an open column for batch AC purification (Bio-Rad), and washed extensively and eluted with 4% and 60% buffer A, respectively. Eluted fractions were pooled, desalted and concentrated, and the presence of proteolytic activity was assessed as above with fibrinogen. This activity was suppressed as described above and subsequent purification by SEC followed in a Superdex 75 10/300 (GE Healthcare) column equilibrated with buffer A without imidazole.

For purification of RECK constructs KL23 and KL123, bacterial cells were harvested by centrifugation at 3,500×*g* for 30 min at 4°C and resuspended in cold buffer 50 mM Tris·HCl pH 7.5, 250 mM sodium

chloride, 2 mM ethylenediaminetetraacetate (EDTA). Cells were lysed with a cell disrupter (Constant Systems) at a pressure of 1.35 kBar, and nonclassical inclusion bodies were recovered by centrifugation at 48,000×*g* for 30 min at 4°C and washed first with buffer 100 mM Tris·HCl pH 7.5, cOmplete EDTA-free (inhibitor cocktail; Roche, Sigma), 2 M urea, 2% Triton X-100, and then twice with buffer D (50 mM Tris·HCl pH 7.5, inhibitor cocktail, 2 M urea). The washed inclusion bodies were resuspended in buffer D and kept for 48 h under stirring at room temperature. Non-solubilised protein was removed by centrifugation at 48,000×*g* for 30 min at 4°C, and the supernatant was supplemented with 20 mM imidazole. Protein was captured by AC in a HisTrap HP column (GE Healthcare) previously washed with buffer E (50 mM Tris·HCl pH 7.5, 250 mM sodium chloride, 500 mM imidazole) and equilibrated with buffer 50 mM Tris·HCl pH 7.5, 250 mM sodium chloride, 20 mM imidazole. The RECK fragments were washed and eluted with 20 mM and 300 mM imidazole (0% and 60% of buffer E), respectively. Partially purified proteins were polished by SEC in a Superdex 75 (GE Healthcare) column with buffer C.

Pure MMP-14 CD was obtained from inclusion bodies by adapting a published protocol[41]. Accordingly, bacterial cells were harvested by centrifugation at 3,500×*g* for 30 min at 4°C and washed with 20 mM Tris·HCl pH 8.0, 20% sucrose for 10 min at 37°C under stirring at 220 rpm. Subsequently, cells were resuspended in buffer F (20 mM Tris·HCl pH 8.0) and kept under gentle agitation overnight at room temperature. Afterwards, first deoxycholate (Sigma) at 1.25 mg/mL and then DNase I (Roche) at 1 mg/mL were added to the lysed cells for 3 h. After a further 2 h incubation, inclusion bodies were harvested by centrifugation at 6,500×*g* for 15 min at 4°C and resuspended in buffer F with 0.5% Triton X-100. Inclusion bodies were then dissolved in buffer 20 mM Tris·HCl pH 8.6, 50 µM zinc chloride, 20 mM 1,4-dithio-D,L-threitol (DTT; Thermo Scientific), 8 M urea. They were further purified by ion exchange chromatography (IEC) in a 6-mL Resource Q column (GE Healthcare), previously washed with buffer G (20 mM Tris·HCl pH 8.6, 0.4 M sodium chloride, 50 µM zinc chloride, 1 mM DTT, 8 M urea) and equilibrated with buffer G without sodium chloride. A step gradient of 0%, 25%, 50% and 100% of buffer G was applied and fractions containing protein were pooled. These were then diluted to 0.2 mg/mL with buffer 50 mM Tris·HCl pH 8.6, 150 mM sodium chloride, 5 mM calcium chloride, 100 µM zinc chloride, 1 mM DTT, 6 M urea, supplemented with cystamine (20 mM) and folded in two consecutive dialysis steps at 4°C. The first step was performed overnight against a 10-fold volume excess of buffer 50 mM Tris·HCl pH 8.6, 150 mM sodium chloride, 5 mM calcium chloride, 100 µM zinc chloride, 5 mM β-mercaptoethanol, 1 mM 2-hydroxyethyldisulfide. The second step was performed against a 10-fold volume excess of buffer B, twice for 4 hours and then overnight. This procedure caused activation of MMP-14 under removal of the pro-domain. Precipitated protein was removed by centrifugation at 48,000×*g* for 30 min at 4°C. Subsequently, MMP-14 CD was concentrated and further purified by SEC in a Superdex 75 (GE Healthcare) column with buffer B.

Other procedures applied were similar to those of previous publications of the group, e.g.[42]. In particular, protein identities and purities were assessed by 10-14% Glycine SDS-PAGE gels stained with Coomassie Brilliant Blue, by peptide mass fingerprinting of tryptic protein digests and by N-terminal sequencing through Edman degradation. The latter two analyses were carried out at the Protein Chemistry Service and Proteomics Facilities of the Centro de Investigaciones Biológicas (Madrid, Spain). Ultrafiltration steps were performed with Vivaspin 15, Vivaspin 2 and Vivaspin 500 filter devices of 3-to-30-kDa cutoff (Sartorius Stedim Biotech). Protein concentrations were

generally estimated by measuring the OD_{280} in a spectrophotometer (NanoDrop; GE Healthcare) and applying the respective theoretical extinction coefficients. Particular concentrations were also measured by the BCA Protein Assay Kit (ThermoFisher Scientific) with bovine serum albumin as a standard.

Multi-angle laser light scattering – To determine the real molecular mass of RECKΔC, multi-angle laser light scattering (SEC-MALLS) was performed as previously reported [42] in a Dawn Helios II apparatus (Wyatt Technologies) coupled to a SEC Superdex 200 10/300 Increase column equilibrated in buffer 20 mM Tris·HCl pH 7.4, 150 mM sodium chloride at 25ºC at the joint IBMB/IRB Crystallography Platform, Barcelona Science Park (Catalonia, Spain). ASTRA 7 software (Wyatt Technologies) was used for data processing and analysis, for which a typical dn/dc value for proteins (0.185 mL/g) was assumed. All experiments were performed in triplicate.

Proteolytic inhibition assays – Inhibition assays with fluorogenic protein and peptide substrates were performed in a microplate fluorimeter (Infinite M200, TECAN) in 100 µL reaction volumes. Proteolytic activity of MMP-2, MMP-7 and MMP-9 (all from R&D Systems) was measured with the fluorescence-based EnzCheck Assay Kit containing DQ Gelatin (λ_{ex}=490nm and λ_{em}=520nm) as fluorescein conjugate (Invitrogen) at 12.5 µg/mL. Peptidolytic activity of MMP-14 CD was measured with the fluorogenic substrate FS-6 (Mca-K-P-L-G-L-Dnp-Dpa-A-R-NH_2; λ_{ex}=325nm and λ_{em}=400nm; Sigma) at 5 µM. Reactions were carried out at 37°C in buffer H (50 mM Tris·HCl pH 7.5, 150 mM sodium chloride, 10 mM calcium chloride, 50 µM zinc chloride, 0.05% Brij-35) except for MMP-2 and MMP-7, for which buffer H supplemented with 1 mM 4-aminophenylmercuric acetate (APMA, Sigma) was used to activate the respective zymogens by incubation for 1 h at 37ºC. Inhibition was measured after preincubation of a 2-, 5-, 10-, 50- and 100-fold molar excess of tester proteins (KL23, KL123, RECKΔC, FRAG-1 and N-TES) with MMP-2 (0.35 ng), MMP-7 (9.8 ng), MMP-9 (2.5 ng) or MMP-14 CD (50 ng) for 1 h at 37°C. Substrates were added to the reaction mixture and the residual proteolytic activity was measured over a timespan of 3 h. Relative activities of MMP-2, MMP-7 and MMP-9 against fluorogenic protein substrates were determined from the slope of a fluorescence vs. time curve. In contrast, fluorogenic peptides were cleaved too fast for proper slope determination, so the relative activity of MMP-14 in front of FS-6 was determined from the absolute fluorescence values measured between 40 and 50 min after reaction start. Control activity of KL23, KL123, RECKΔC, FRAG-1, N-TES and BSA was measured between 110 and 120 min after reaction start. Bovine serum albumin (BSA; Sigma) at 100-fold molar excess and o-phenanthroline (Fluka) at 5 mM were included as negative and positive controls for inhibition, respectively. In addition, inhibition assays against cleavage of human plasma fibronectin (pFN, MP Biomedicals) were evaluated by Western blot analysis (see below). MMP-2 or MMP-7 (at 40 nM) were incubated in buffer H *plus* 1 mM APMA for activation with pFN (4 nM) at 37°C for 0, 1, 2, 4, 6, 8 or 18 h with or without RECKΔC (400 nM). The broad-spectrum MMP inhibitors marimastat (Sigma), EDTA (Fluka), o-phenanthroline, and batimastat (Calbiochem) were used in controls, as well as the serine peptidase inhibitors AEBSF and phenylmethanesulfonyl fluoride (PMSF; Acros Organics) *plus* the cOmplete EDTA-Free inhibitor cocktail.

Western blot analyses – Protein samples were separated by 10% Glycine SDS-PAGE, transferred to Amersham Protran Premium NC Nitrocellulose Membranes (GE Healthcare Life Sciences) and blocked for one

hour under gentle stirring at room temperature with 50 mL of blocking solution (5% BSA in phosphate buffered saline [PBS] *plus* 0.2% Tween 20 [PBS-T]; Sigma). Fibronectin was detected by overnight incubation at 4ºC with a rabbit polyclonal primary antibody (Abcam) diluted 1:5,000 in PBS-T with 1% BSA and subsequent incubation for 2 h at room temperature with an anti-rabbit HRP-conjugated secondary antibody (Sigma) diluted 1:8,000 with PBS-T. Blots were incubated with mild stripping buffer (1.5% glycine pH 2.2, 0.1% SDS, 1% Tween 20) and further washed with PBS and PBS-T under gentle agitation at room temperature. Blots were re-blocked and re-probed.

His_6-tagged proteins were detected with the monoclonal His-HRP Conjugated Antibody (Santa Cruz Biotechnology) diluted 1:5,000 in PBS-T with 1% BSA incubated overnight at 4°C and subsequently visualized with an enhanced chemiluminescence system (Super Signal West Pico Chemiluminescent; Pierce) according to the manufacturer's instructions. Membranes were exposed to Hyperfilm ECL films (GE Healthcare Life Sciences).

Miscellaneous – Structure prediction calculations through threading were performed with LOMETS [43] and RAPTORX [44] with standard parameters.

RESULTS AND DISCUSSION

Protein preparation – Inhibitory activity of RECK on MMPs *in vitro* has been reported for RECKΔC[13-15,17,23] and a construct spanning the KL2 and KL3 domains [16]. As there were discrepancies in the boundaries of these domains [13,16], we performed structure prediction through threading of segment V^{600}–A^{800}. These calculations suggested that constructs KL123 and KL23 should actually span segments V^{621}–S^{797} and T^{697}–S^{797}, respectively (Fig.1), which do not contain any of the glycosylation sites of the full-length protein [22]. Follistatin (Protein Data Bank [PDB] entries 2P6A, 3HH2 and 2B0U), a regulator of ligands of the transforming growth factor-β superfamily with three Kazal-like repeats [45], and follistatin-like protein 3 (PDB 3B4V) were identified as the closest structural relatives. Both RECK constructs were produced with C-terminal His_6-tags overnight in *E. coli* Lemo21 cells at room temperature and translocated to the periplasm, which provides an oxidizing environment for the formation of disulphide bonds and protein folding. The proteins were obtained in high yields as nonclassical inclusion bodies [46], which were treated under non-reducing conditions with a chaotropic agent and detergent prior to purification by AC and SEC steps (Fig.2A–D). The resulting proteins were soluble and highly pure, did not aggregate when concentrated, and migrated according to 16 kDa (KL23) and 26 kDa (KL123) in calibrated SEC (data not shown), which are consistent with monomeric species. These data suggest that the proteins were well-folded.

We also isolated RECKΔC with a C-terminal His_6-tag (Fig.1) from the conditioned medium of transiently transfected Expi cells by adapting a protocol developed previously for human α_2-macroglobulin [42]. The yield after purification was 0.8 mg per litre of expression medium, and the protein was subsequently purified by AC and SEC (Fig.2E). It had a molecular mass of 111 kDa according to SEC-MALLS (Fig.2F), which is in good agreement with the theoretical protein mass *plus* glycosylation, and indicates that the protein is monomeric. This contrasts with other studies postulating it is a dimer [17]. We further produced RECKΔC from S2 and Expi-CHO systems but the initial yields were significantly lower (0.2 and 0.5 mg/L, respectively) and the proteins required several additional purification steps (data not shown). We next obtained N-TES by the same method from transfected Expi cells (Fig.2H). This protein spans the N-terminal region of the

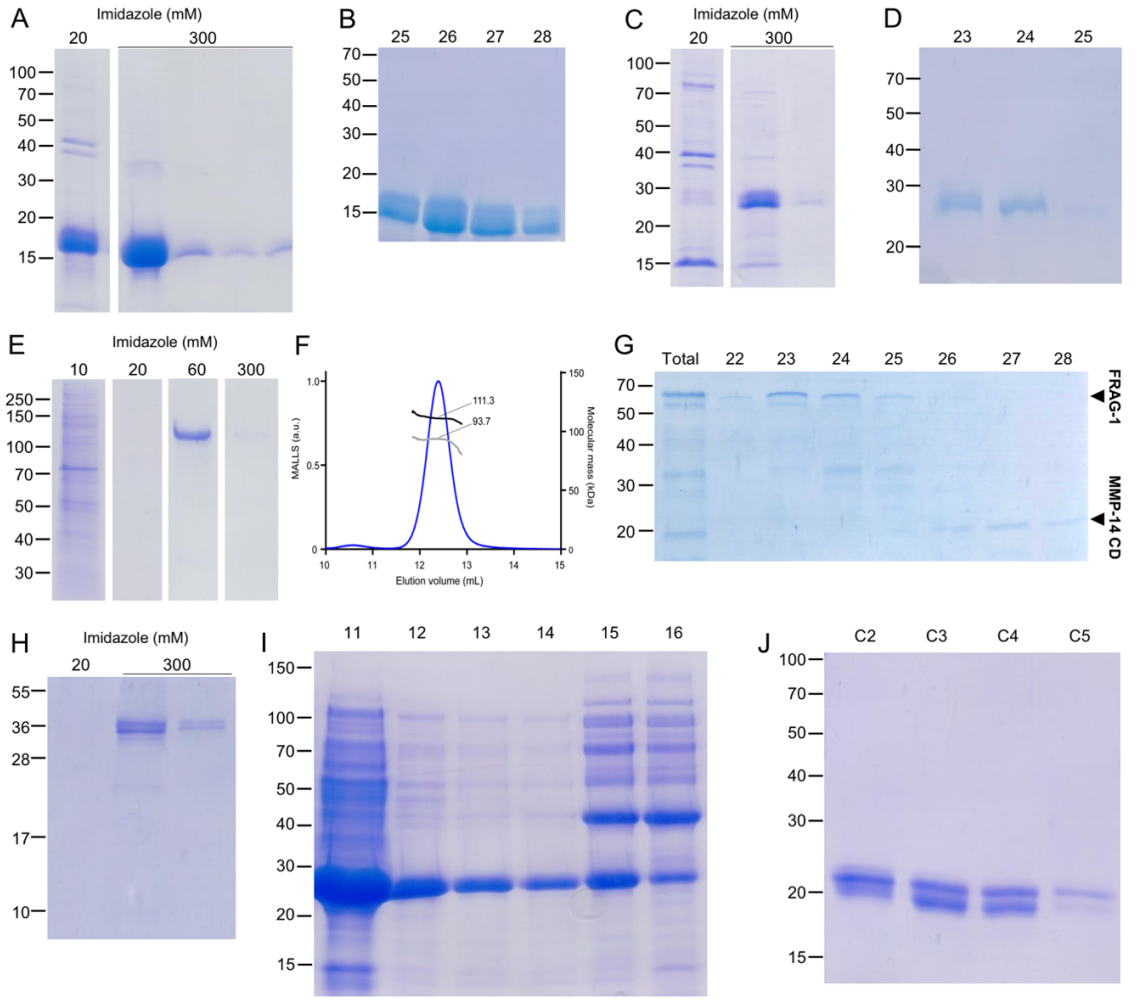


Figure 2. Protein purification. SDS-PAGE of AC and SEC purification steps of KL23 (**A,B**) and KL123 (**C,D**). (**E**) SDS-PAGE of stepwise AC of RECKΔC from Expi cells. (**F**) SEC-MALLS chromatogram of RECKΔC showing it has a total molecular mass of 111.3 kDa of which 93.7 kDa would correspond to protein. (**G**) SDS- PAGE of SEC fractions containing FRAG-1 (22–25) and MMP-14 CD (26–28). (**H**) SDS-PAGE of partially purified N-TES after AC purification. (**I**) SDS-PAGE of AEC purification of MMP-14 CD. (**J**) SDS-PAGE of the SEC fractions containing MMP-14 CD (C2-C5). MMP-14 CD migrates as two bands as previously observed for construct MT1Cat in [41]. Figure panels with lanes/parts from different gels/blots show white separation lines. All original gels can be found in the supplementary materials.

calcium-binding proteolgycan testican 3, which has been reported to bind membrane-type MMPs including MMP-14 and to inhibit pro-MMP-2 activation in HEK293T cells when their respective cDNAs were co-transfected. These results led the authors to suggest that N-TES is an inhibitor of MMP-14 and MMP-16 [47].

Finally, we also produced and purified RECK fragment FRAG-1 resulting from the limited cleavage of RECKΔC by MMP-14 CD, which contained the C-terminal half of the full-length protein including KL1 through KL3 (Figs.1 and 2G). MMP-14 CD was produced by *E. coli* BL21 (DE3) in inclusion bodies, purified by IEC under denaturing conditions, folded by dialysis, and finally purified by SEC by implementing a previous protocol[41] (Figs.2I,J).

Proteolytic contamination and additional purification steps – Recombinant RECKΔC from Expi cells was initially purified by AC (Fig.2E) and SEC. Despite rather high purity (>98%; Figs.2E and 3A), it underwent cleavage over time, which was prevented by an inhibitor cocktail and partially slowed down by the general zinc chelator and MP

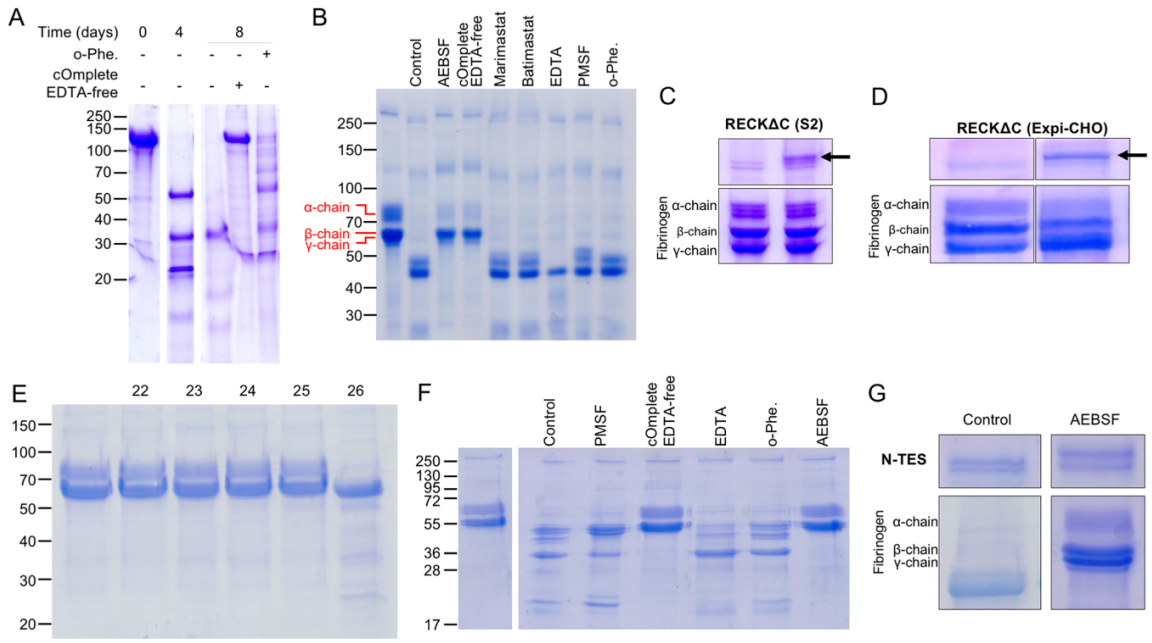


Figure 3. Functional assays. (**A**) Partially purified RECKΔC incubated for up to 8 days at 37 °C with or without o-phenanthroline (o-Phe.) or an inhibitor cocktail (cOmplete EDTA-free). (**B**) Degradation of fibrinogen. *Lane 1*, intact fibrinogen; *lane 2*, *control*, fibrinogen incubated with partially purified RECKΔC for two days shows degradation; *lanes 3–9*, same except for a previous one hour-incubation of RECKΔC with AEBSF, inhibitor cocktail, marimastat, batimastat, EDTA, PMSF or o-phenanthroline. Fibrinogen cleavage does not occur with similarly purified RECKΔC from (**C**) S2 cells or (**D**) Expi-CHO cells (both, *left lane*, fibrinogen alone; *right lane*, fibrinogen *plus* RECKΔC [black arrow]). (**E**) Incubation of fibrinogen (*lane 1*, *control*) with SEC fractions containing only FRAG-1 (22–25) show no degradation. However, the substrate is cleaved by a coeluting MMP-14 CD contamination (fraction 26). (**F**) Incubation of fibrinogen (*lane 1*) with partially purified N-TES without (*lane 2*, *control*) or with (*lanes 3–7*) inhibitors. (**G**) An N-TES preparation purified by SEC cleaved fibrinogen (*control*). This cleavage was abolished with AEBSF. Figure panels with lanes/parts from different gels/blots show white separation lines. All original gels can be found in the supplementary materials.

inhibitor o-phenanthroline (Fig.3A). However, this cleavage did not result in dissociation of RECKΔC in the short term, according to SEC. The cleavage products caused by this impurity (Fig. 3A, *lane 2*) were subjected to N-terminal Edman degradation, which revealed that a ~50 kDa fragment, dubbed FRAG-1, resulted from cleavage before G^{484} and thus comprised RECK domains KL1-KL2-KL3. As partially purified RECKΔC had initially shown slight inhibition of MMP-14 CD (data not shown), we speculated that RECKΔC cleavage might be necessary to yield a species with MMP inhibitory activity. Thus, we included FRAG-1 in subsequent inhibition assays (see below).

To investigate the nature of the proteolytic impurity, we incubated initially purified RECKΔC with the general peptidase substrate human plasma fibrinogen, and found it was cleaved (Fig.3B). We assayed a series of peptidase inhibitors and found that AEBSF abolished the cleavage (Fig.3B). AEBSF is an irreversible small-molecule serine peptidase inhibitor that covalently modifies the catalytic serine of serine endopeptidases, thus blocking them. The same ablation was obtained with the inhibitor cocktail, which contained AEBSF, **but not with PMSF or general MP inhibitors** (**Fig.3B**). This motivated us to include an extra step in the purification protocol of RECKΔC from Expi cells consisting of incubation with AEBSF and final polishing by several cycles of SEC. This protocol yielded RECKΔC of highest purity, incapable of fibrinogen or RECKΔC degradation, for subsequent inhibitory studies. Interestingly, protein produced from S2 or Expi-CHO cells did not show this contaminant and fibrinogen remained intact upon incubation with these RECKΔC species (Fig.3C,D). Finally, FRAG-1 obtained from

highly pure RECKΔC through treatment with MMP-14 CD did not contain the proteolytic contaminant of partially purified RECKΔC but some traces of the MP, which could be separated in SEC (Fig.3E).

To assess whether the AEBSF-sensitive contaminant was a specific feature of the overexpression of RECKΔC, we studied protein N-TES obtained with the same expression system (Fig.2H) and observed similar peptidolytic activity against fibrinogen that was abolished with AEBSF or the inhibitor cocktail (Fig.3F,G). In contrast, two other proteins obtained with the same system did not contain this contaminant (data not shown).

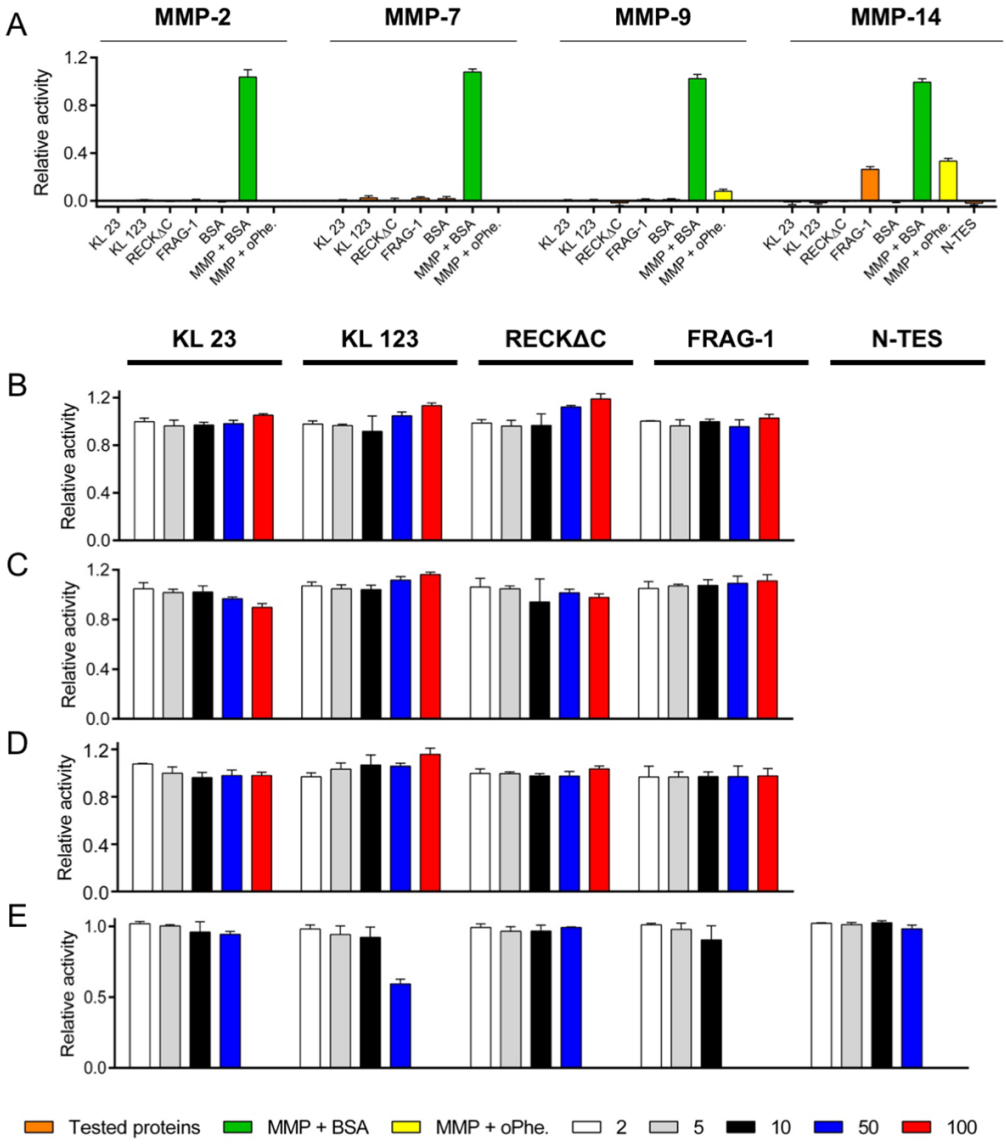


Figure 4. Activity and inhibition assays. (**A**) Hydrolytic activity of pure RECK constructs, N-TES and BSA (tested proteins, orange bars) relative to that of the respective MMPs (MMP-2, -7, -9 and -14) with BSA (100%; positive control, green bar) or o-phenantroline (negative control, yellow bar). Only FRAG-1 evinces slight residual activity against the fluorogenic substrate FS-6, possibly due to a contamination with MMP-14 CD used for its generation (see Fig.2G). (**B**) Residual activity of MMP-2 in the presence of KL23, KL123, RECKΔC or FRAG-1 at various molar rations (1:2, 1:5, 1:10, 1:50 and 1:100; white, gray, black, blue and red bars) after 1 h incubation relative to the activity shown in (A; green bar). (**C**) and (**D**), same as (B) for MMP-7 and MMP-9. (**E**) Same as (B) for MMP-14 but further including N-TES as candidate inhibitor up to 1:50 molar ratio. Only 50-fold excess of KL123 shows a slight inhibitory effect on MMP-14 CD activity. All assays of panels A-E were performed in triplicate.

Inhibition studies of RECK and N-TES constructs – Highly pure RECK variants RECKΔC, KL123, KL23 and FRAG-1, as well as N-TES, were tested for their inhibitory capacity against MMP-2, MMP-7, MMP-9 and MMP-14 activity with peptide and protein substrates up to 100-fold molar excess of the tester proteins (Figs.4 and 5). Fluorescein-conjugated gelatine was used for assays with previously activated MMP-2 (Fig.4B), MMP-7 (Fig.4C) and MMP-9 (Fig.4D), and fluorogenic peptide FS-6 was employed for MMP14 CD (Fig.4E). In addition, inhibition of the activity of MMP-2 (Fig.5A) and MMP-7 (Fig.5B) against plasma fibronectin by RECKΔC at tenfold molar excess was assayed by Western blot analysis. Moreover, inhibition of the activity of MMP-14 CD against a fluorogenic peptide substrate by N-TES was likewise analysed (Fig.4E). As expected, none of the RECK constructs, N-TES or BSA, which was used as a negative control, alone showed relevant peptidolytic activity and o-phenanthroline inhibited the MMPs as predictable (Fig.4A). In addition, BSA at the highest tester concentration (1:100 molar excess) had no influence on MMP

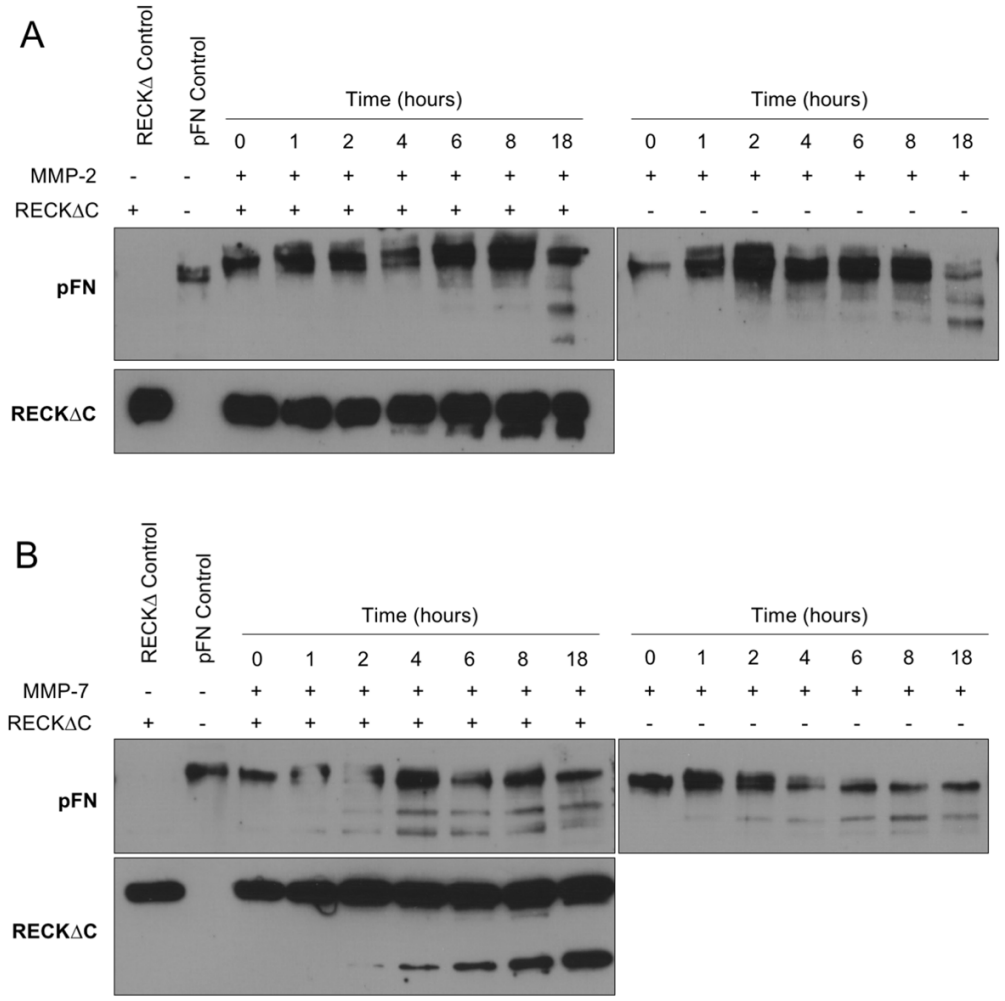


Figure 5. MMP-2 and MMP-7 activity assays against plasma fibronectin. (A) Western-blot analysis of the MMP-2 (40 nM) activity in front of plasma fibronectin (pFN; 4 nM) in the presence or absence of RECKΔC (400 nM) over time (0–18 h). The peptidase eventually cleaves the substrate irrespective of the presence of RECKΔC. **(B)** Same as (A) for MMP-7. In addition, MMP-7 cleaves RECKΔC over time. Figure panels with lanes/parts from different gels/blots show white separation lines. All original gels can be found in the supplementary materials.

activity (Fig.4A). Notably, the experiments revealed that none of the RECK constructs or N-TES showed any significant inhibition of MMPs activity.

CONCLUSIONS

Since its discovery in the 1980s, protein RECK has been found to have pleiotropic roles, for example in embryogenesis and as a tumour suppressor. Among other functions, it has been hailed as an MMP inhibitor. Here, we produced two RECK variants through the mammalian Expi and Expi-CHO systems and through the insect S2 system. We could not detect any significant inhibition with either construct produced with Expi when we assayed MMPs that have been previously reported to be targeted, neither with peptide nor protein substrates. Instead, we found that even quite pure samples of RECKΔC showed proteolytic activity resulting from a contamination by a probable serine peptidase, which could be removed by treatment with AEBSF and additional SEC. This activity was missing when the protein was produced in Expi-CHO or S2 cells. We further established a production and purification protocol for N-TES, which has also been postulated to be an MMP inhibitor. As in the case of RECK, we detected the contaminating peptidase but no inhibitory activity in front of the target MMP-14. In contrast, similar expression systems for two other unrelated proteins did not produce the contaminant in the supernatant. Finally, we made two shorter constructs of RECK spanning KL1-KL2-KL3 and KL2-KL3, respectively, through an *E. coli* system, which also lacked the contaminant. These proteins did not show any inhibitory effect on MMPs either.

These results are consistent with marginal notes in a recent article describing very sensitive gelatine zymography, which revealed gelatinolytic activity associated with recombinant RECKΔC preparations [23]. These corresponded to co-purified peptidases of ~19 and ~28 kDa, which necessarily must have perturbed previously published kinetic assays with partially purified RECKΔC from supernatants of transfected 293F and 293T cells, as well as construct K23 from 293T cells, for which inhibitory activity on MMPs had been reported [13,15-17]. Moreover, the inhibitory activity of RECK on MMPs was requalified as “weak” recently [36].

Taken together, all these findings suggest that RECK, and probably N-TES, are not direct inhibitors of MMP catalytic activity. Instead, RECK may still regulate MMPs *in vivo* at a different level, e.g. through downregulation of MMP transcription, translation or secretion, or by binding and sequestering them, thus preventing them from carrying out their extracellular peptidolytic function.

Finally, we recommend checking for unexpected proteolytic activity associated with recombinant proteins of interest when employing transfected Expi293F or other HEK293-derived cells as protein expression systems, which is absent from systems based on Chinese hamster ovary or fruit-fly S” cells. This activity can be removed through irreversible serine peptidase inhibitors or inhibitor cocktails.

AUTHOR CONTRIBUTIONS

T.G., I.d.D. and F.X.G.R. conceived and supervised the work; S.R.M., L.A.M. and L.M.P. produced and purified proteins, and performed biochemical studies; I.d.D. implemented the S2 overexpression system in the laboratory; and F.X.G.R. wrote the paper with contributions from all authors.

ACKNOWLEDGMENTS

We are grateful to Roman Bonet, Xandra Kreplin and Joan Pous from the joint IBMB/IRB Automated Crystallography Platform and the Protein Purification Service for assistance during purification and SEC-MALLS experiments. Philippe Leone is thanked for advice and provision of positive control target expression cosntructs to set up the *Drosophila melanogaster* Schneider embryonic cell system for protein production in the laboratory. This study was supported in part by grants from Spanish and Catalan public and private bodies (grant/fellowship references BFU2015-64487R, MDM-2014-0435, BES-2015-074583, BES-2016-076877, 2017SGR3 and Fundació "La Marató de TV3" 201815). Yoshifumi Ito from the University of Oxford (UK) and Makoto Noda from Kyoto University (Japan) kindly provided plasmids for the production of human MMP-14 catalytic domain and human RECK, respectively.

CONFLICT OF INTERESTS

The authors declare that they have no financial or non-financial conflicts of interest with the contents of this article.

REFERENCES

[1] Cerdà-Costa, N. & Gomis-Rüth, F. X. Architecture and function of metallopeptidase catalytic domains. *Prot. Sci.* 23, 123-144, doi:10.1002/pro.2400 (2014).

[2] Edwards, D. R., Handsley, M. M. & Pennington, C. J. The ADAM metalloproteinases. *Mol. Aspects Med.* 29, 258-289, doi:10.1016/j.mam.2008.08.00 (2008).

[3] Murphy, G. & Nagase, H. Progress in matrix metalloproteinase research. *Mol. Aspects Med.* 29, 290-308, doi:10.1016/j.mam.2008.05.002 (2008).

[4] Apte, S. S. A disintegrin-like and metalloprotease (reprolysin-type) with thrombospondin type 1 motif (ADAMTS) superfamily: functions and mechanisms. *J. Biol. Chem.* 284, 31493-31497, doi:10.1074/jbc.R109.052340 (2009).

[5] Tallant, C., Marrero, A. & Gomis-Rüth, F. X. Matrix metalloproteinases: fold and function of their catalytic domains. *Biochim. Biophys. Acta - Mol. Cell Res.* 1803, 20-28, doi:10.1016/j.bbamcr.2009.04.003 (2010).

[6] Jobin, P. G., Butler, G. S. & Overall, C. M. New intracellular activities of matrix metalloproteinases shine in the moonlight. *Biochim. Biophys. Acta - Mol. Cell Res.* 1864, 2043-2055, doi:10.1016/j.bbamcr.2017.05.013 (2017).

[7] Klein, T., Eckhard, U., Dufour, A., Solis, N. & Overall, C. M. Proteolytic cleavage-mechanisms, function, and "omic" approaches for a near-ubiquitous posttranslational modification. *Chem. Rev.* 118, 1137-1168, doi:10.1021/acs.chemrev.7b00120 (2018).

[8] Arolas, J. L., Goulas, T., Cuppari, A. & Gomis-Rüth, F. X. Multiple architectures and mechanisms of latency in metallopeptidase zymogens. *Chem. Rev.* 118, 5581-5597, doi:10.1021/acs.chemrev.8b00030 (2018).

[9] Katunuma, N. Regulation of intracellular enzyme levels by limited proteolysis. *Rev. Physiol. Biochem. Pharmacol.* 72, 83-104, doi:10.1007/bfb0031547 (1975).

[10] Tortorella, M. D. *et al.* α_2-Macroglobulin is a novel substrate for ADAMTS-4 and ADAMTS-5 and represents an endogenous inhibitor of these enzymes. *J. Biol. Chem.* 279, 17554-17561, doi:10.1074/jbc.M313041200 (2004).

[11] Nagase, H., Visse, R. & Murphy, G. Structure and function of matrix metalloproteinases and TIMPs. *Cardiovasc. Res.* 69, 562-573, doi:10.1016/j.cardiores.2005.12.002 (2006).

[12] Goulas, T. *et al.* Structural and functional insight into pan-endopeptidase inhibition by α_2-macroglobulins. *Biol. Chem.* 398, 975-994, doi:10.1515/hsz-2016-0329 (2017).

[13] Takahashi, C. *et al.* Regulation of matrix metalloproteinase-9 and inhibition of tumor invasion by the membrane-anchored glycoprotein RECK. *Proc. Natl. Acad. Sci. USA* 95, 13221-13226, doi:10.1073/pnas.95.22.13221 (1998).

[14] Oh, J. *et al.* The membrane-anchored MMP inhibitor RECK is a key regulator of extracellular matrix integrity and angiogenesis. *Cell* 107, 789-800, doi:10.1016/s0092-8674(01)00597-9 (2001).

[15] Miki, T. *et al.* The reversion-inducing cysteine-rich protein with Kazal motifs (RECK) interacts with membrane type 1 matrix metalloproteinase and CD13/aminopeptidase N and modulates their endocytic pathways. *J. Biol. Chem.* 282, 12341-12352, doi:10.1074/jbc.M610948200 (2007).

[16] Chang, C. K., Hung, W. C. & Chang, H. C. The Kazal motifs of RECK protein inhibit MMP-9 secretion and activity and reduce metastasis of lung cancer cells *in vitro* and *in vivo*. *J. Cell. Mol. Med.* 12, 2781-2789, doi:10.1111/j.1582-4934.2008.00215.x (2008).

[17] Omura, A. *et al.* RECK forms cowbell-shaped dimers and inhibits matrix metalloproteinase-catalyzed cleavage of fibronectin. *J. Biol. Chem.*

284, 3461-3469, doi:10.1074/jbc.M806212200 (2009).

[18] Noda, M. *et al.* Detection of genes with a potential for suppressing the transformed phenotype associated with activated *ras* genes. *Proc. Natl. Acad. Sci. USA* 86, 162-166, doi:10.1073/pnas.86.1.162 (1989).

[19] Kazal, L. A., Spicer, D. S. & Brahinsky, R. A. Isolation of a crystalline trypsin inhibitor-anticoagulant protein from pancreas. *J. Am. Chem. Soc.* 70, 3034-3040, doi:10.1021/ja01189a060 (1948).

[20] Laskowski Jr., M. & Kato, I. Protein inhibitors of proteinases. *Annu. Rev. Biochem.* 49, 593-626, doi:10.1146/annurev.bi.49.070180.003113 (1980).

[21] Rhee, J. S. & Coussens, L. M. RECKing MMP function: implications for cancer development. *Trends Cell Biol.* 12, 209-211, doi:10.1016/s0962-8924(02)02280-8 (2002).

[22] Simizu, S., Takagi, S., Tamura, Y. & Osada, H. RECK-mediated suppression of tumor cell invasion is regulated by glycosylation in human tumor cell lines. *Cancer Res.* 65, 7455-7461, doi:10.1158/0008-5472.CAN-04-4446 (2005).

[23] Muraguchi, T. *et al.* RECK modulates Notch signaling during cortical neurogenesis by regulating ADAM10 activity. *Nat. Neurosci.* 10, 838-845, doi:10.1038/nn1922 (2007).

[24] Willson, J. A. & Damjanovski, S. Spatial analysis of RECK, MT1-MMP, and TIMP-2 proteins during early Xenopus laevis development. *Gene Expr. Patterns* 34, 119066, doi:10.1016/j.gep.2019.119066 (2019).

[25] Vanhollebeke, B. *et al.* Tip cell-specific requirement for an atypical Gpr124- and Reck-dependent Wnt/beta-catenin pathway during brain angiogenesis. *Elife* 4, e06489, doi:10.7554/eLife.06489 (2015).

[26] Li, H. *et al.* RECK in neural precursor cells plays a critical role in mouse forebrain angiogenesis. *iScience* 19, 559-571, doi:10.1016/j.isci.2019.08.009 (2019).

[27] Cho, C., Wang, Y., Smallwood, P. M., Williams, J. & Nathans, J. Molecular determinants in Frizzled, Reck, and Wnt7a for ligand-specific signaling in neurovascular development. *Elife* 8, e47300, doi:10.7554/eLife.47300 (2019).

[28] Hill, V. K. *et al.* Genome-wide DNA methylation profiling of CpG islands in breast cancer identifies novel genes associated with tumorigenicity. *Cancer Res.* 71, 2988-2999, doi:10.1158/0008-5472.CAN-10-4026 (2011).

[29] Yoshida, Y., Ninomiya, K., Hamada, H. & Noda, M. Involvement of the SKP2-p27(KIP1) pathway in suppression of cancer cell proliferation by RECK. *Oncogene* 31, 4128-4138, doi:10.1038/onc.2011.570 (2012).

[30] Masui, T. *et al.* RECK expression in pancreatic cancer: its correlation with lower invasiveness and better prognosis. *Clin. Cancer Res.* 9, 1779-1784 (2003).

[31] Span, P. N. *et al.* Matrix metalloproteinase inhibitor reversion-inducing cysteine-rich protein with Kazal motifs: a prognostic marker for good clinical outcome in human breast carcinoma. *Cancer* 97, 2710-2715, doi:10.1002/cncr.11395 (2003).

[32] Takenaka, K. *et al.* Expression of a novel matrix metalloproteinase regulator, RECK, and its clinical significance in resected non-small cell lung cancer. *Eur. J. Cancer* 40, 1617-1623, doi:10.1016/j.ejca.2004.02.028 (2004).

[33] Kang, H. G. *et al.* RECK expression in osteosarcoma: correlation with matrix metalloproteinases activation and tumor invasiveness. *J. Orthop. Res.* 25, 696-702, doi:10.1002/jor.20323 (2007).

[34] Chen, R., Sheng, L., Zhang, H. J., Ji, M. & Qian, W. Q. miR-15b-5p facilitates the tumorigenicity by targeting RECK and predicts tumour recurrence in prostate cancer. *J. Cell Mol. Med.* 22, 1855-1863, doi:10.1111/jcmm.13469 (2018).

[35] Chen, H. C. *et al.* Prognostic role of RECK in pathological outcome-dependent buccal mucosa squamous cell carcinoma. *Oral Dis.* 25, doi:10.1111/odi.13214, doi:10.1111/odi.13214 (2019).

[36] Matsuzaki, T. *et al.* The RECK tumor-suppressor protein binds and stabilizes ADAMTS10. *Biol. Open* 7, bio033985, doi:10.1242/bio.033985 (2018).

[37] Rawlings, N. D. *et al.* The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res.* 46, D624-D632, doi:10.1093/nar/gkx1134 (2018).

[38] Goulas, T. *et al.* The pCri System: a vector collection for recombinant protein expression and purification. *PloS one* 9, e112643, doi:10.1371/journal.pone.0112643 (2014).

[39] van den Ent, F. & Löwe, J. RF cloning: a restriction-free method for inserting target genes into plasmids. *J. Biochem. Biophys. Methods* 67, 67-74, doi:10.1016/j.jbbm.2005.12.008 (2006).

[40] Woskowicz, A. M., Weaver, S. A., Shitomi, Y., Ito, N. & Itoh, Y. MT-LOOP-dependent localization of membrane type I matrix metalloproteinase (MT1-MMP) to the cell adhesion complexes promotes cancer cell invasion. *J. Biol. Chem.* 288, 35126-35137, doi:10.1074/jbc.M113.496067 (2013).

[41] Itoh, Y. *et al.* Homophilic complex formation of MT1-MMP facilitates proMMP-2 activation on the cell surface and promotes tumor cell invasion. *EMBO J.* 20, 4782-4793, doi:10.1093/emboj/20.17.4782 (2001).

[42] Marino-Puertas, L., Del Amo-Maestro, L., Taules, M., Gomis-Rüth, F. X. & Goulas, T. Recombinant production of human α_2-macroglobulin variants and interaction studies with recombinant G-related α_2-macroglobulin binding protein and latent transforming growth factor-β_2. *Sci. Rep.* 9, 9186, doi:10.1038/s41598-019-45712-z (2019).

[43] Wu, S. & Zhang, Y. LOMETS: a local meta-threading-server for protein structure prediction. *Nucleic Acids Res.* 35, 3375-3382, doi:10.1093/nar/gkm251 (2007).

[44] Källberg, M. *et al.* Template-based protein structure modeling using the RaptorX web server. *Nat. Protoc.* 7, 1511-1522, doi:10.1038/nprot.2012.085 (2012).

[45] Lerch, T. F., Shimasaki, S., Woodruff, T. K. & Jardetzky, T. S. Structural and biophysical coupling of heparin and activin binding to follistatin isoform functions. *J. Biol. Chem.* 282, 15930-15939, doi:10.1074/jbc.M700737200 (2007).

[46] Jevsevar, S. *et al.* Production of nonclassical inclusion bodies from which correctly folded protein can be extracted. *Biotechnol. Prog.* 21, 632-639, doi:10.1021/bp0497839 (2005).

[47] Nakada, M. *et al.* Suppression of membrane-type 1 matrix metalloproteinase (MMP)-mediated MMP-2 activation and tumor invasion by testican 3 and its splicing variant gene product, N-Tes. *Cancer Res.* 61, 8896-8902 (2001).

Project 2

"An engineered protein-based submicromolar competitive inhibitor of the Staphylococcus aureus virulence factor aureolysin"

Soraia R. Mendes, Ulrich Eckhard [*], Arturo Rodríguez-Banqueri, Tibisay Guevara, Peter Czermak [1,2], Enrique Marcos [3], Andreas Vilcinskas [1,2,*] and F. Xavier Gomis-Rüth [*]

Proteolysis Laboratory; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC); Barcelona Science Park; Baldiri Reixac 15-21; 08028 Barcelona (Catalonia, Spain).
[1] Department of Bioresources; Fraunhofer Institute for Molecular Biology and Applied Ecology; Ohlebergsweg 12; 35392 Giessen (Germany).
[2] Institute for Insect Biotechnology; Justus-Liebig University Giessen; Heinrich-Buff-Ring 26-32; 35392 Giessen (Germany).
[3] Protein Design and Modelling; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC); Barcelona Science Park; Baldiri Reixac 15-21; 08028 Barcelona (Catalonia, Spain).
*Corresponding authors. E-mail: ueccri@ibmb.csic.es, andreas.vilcinskas@agrar.uni-giessen.de and xgrcri@ibmb.csic.es.

Published in: Computational and Structural Biotechnology Journal (2022)
Impact factor: 7.271
Quartile: Q1

This publication explores the inhibitory capacity of the insect metallopeptidase inhibitor (IMPI) against aureolysin, a major *S. aureus* virulence factor without a specific protein inhibitor known so far.

In this project, I cloned all twelve IMPI mutants, and produced and purified them along with wild-type IMPI from *E. coli* Origami 2 cells. I also purified native aureolysin isoform I from S. aureus (strain V8-BC10) cultures. Subsequently, I performed all activity and inhibition assays, analysed IMPI cleavage by aureolysin through MALDI-TOF and SDS-PAGE and crystallised aureolysin in complex with both wt and $I^{57}F$ IMPI.

Summary

Upon infection, the greater wax *Galleria mellonella* produces a set of defensive proteins with antimicrobial properties, so called antimicrobial peptides and proteins (AMPs). One of those molecules is the insect metallopeptidase inhibitor (IMPI), a low molecular weight protein stabilized by five disulphide bridges which specifically inhibits thermolysin and other members of the thermolysin family (M4). Aureolysin is an attractive target of the M4 family due to its inextricable association with the *S. aureus* infection mechanism, and the growing number of infections caused by antibiotic resistant *S. aureus* strains impel the need for new antibiotic therapeutic drugs. Here we analysed the inhibitory capacity of IMPI against aureolysin and crystallised the protease-inhibitor complex revealing that IMPI inhibits aureolysin through a "standard mechanism of action", which is poorly characterised for metallopeptidases. We then designed a set of fourteen IMPI variants with single or multiple mutations on their reactive-centre loop (RCL); among them, $I^{57}F$ was the best inhibitor of aureolysin with an estimated inhibition constant (Ki) of 346 nM. This work highlighted the inhibitory mechanism of IMPI against the potent virulence factor aureolysin, as well as the structural features of the protease-inhibitor complex, providing valuable information for further development of safe, IMPI-based therapeutic peptides (TP) targeting aureolysin, which might be used in the treatment of antibiotic resistant infections.

An engineered protein-based submicromolar competitive inhibitor of the *Staphylococcus aureus* virulence factor aureolysin

Soraia R. Mendes, Ulrich Eckhard [*], Arturo Rodríguez-Banqueri, Tibisay Guevara, Peter Czermak [1,2], Enrique Marcos [3], Andreas Vilcinskas [1,2,*] and F. Xavier Gomis-Rüth [*]

Proteolysis Laboratory; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC); Barcelona Science Park; Baldiri Reixac 15-21; 08028 Barcelona (Catalonia, Spain).
[1] Department of Bioresources; Fraunhofer Institute for Molecular Biology and Applied Ecology; Ohlebergsweg 12; 35392 Giessen (Germany).
[2] Institute for Insect Biotechnology; Justus-Liebig University Giessen; Heinrich-Buff-Ring 26-32; 35392 Giessen (Germany).
[3] Protein Design and Modelling; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC); Barcelona Science Park; Baldiri Reixac 15-21; 08028 Barcelona (Catalonia, Spain).
* Corresponding authors. E-mail: ueccri@ibmb.csic.es, andreas.vilcinskas@agrar.uni-giessen.de and xgrcri@ibmb.csic.es.

***Keywords*:** bacterial infection, metallopeptidase, protein inhibitor, protein design, therapeutic protein, crystal structure.

Aureolysin, a secreted metallopeptidase (MP) from the thermolysin family, functions as a major virulence factor in *Staphylococcus aureus*. No specific aureolysin inhibitors have yet been described, making this an important target for the development of novel antimicrobial drugs in times of rampant antibiotic resistance. Although small-molecule inhibitors are currently more common in the clinic, therapeutic proteins and peptides (TPs) are favourable due to their high selectivity, which reduces off-target toxicity and allows dosage tuning. The greater wax moth *Galleria mellonella* produces a unique defensive protein known as the insect metallopeptidase inhibitor (IMPI), which selectively inhibits some thermolysins from pathogenic bacteria. We determined the ability of IMPI to inhibit aureolysin *in vitro* and used crystal structures to ascertain its mechanism of action. This revealed that IMPI uses the "standard mechanism", which has been poorly characterised for MPs in general. Accordingly, we designed a cohort of 12 single and multiple point mutants, the best of which ($I^{57}F$) inhibited aureolysin with an estimated inhibition constant (K_i) of 346 nM. Given that animals lack thermolysins, our strategy may facilitate the development of safe TPs against staphylococcal infections, including strains resistant to conventional antibiotics.

Antibiotic resistance is a major global health burden, leading to hundreds of thousands of deaths every year and greatly increasing healthcare costs associated with the treatment of bacterial infections [1-3]. Resistance arises from selection pressure caused by the widespread abuse, overuse and misuse of antibiotics in humans, including premature treatment discontinuation [4], subtherapeutic dosing, and the distribution of counterfeit drugs [5]. Furthermore, ~80% of all antimicrobials used in the USA are administered as prophylactics to farm animals to boost their health and productivity [6]. Once acquired, resistance is spread by horizontal gene transfer, often across species barriers, ultimately giving rise to multidrug-resistant strains [7]. The impact of antibiotic resistance is heightened by the lack of new drugs in the development pipeline, with only two new classes of antibiotics approved in the last 30 years: the oxazolidinones, which target

protein synthesis, and the acidic lipopeptides, which target bacterial membranes [8, 9]. This lack of progress reflects decades of low returns compared with other drug classes, discouraging investment by the pharmaceutical industry [2, 7, 10] and thus posing a serious threat to public health [11]. There are few therapeutic options for the treatment of infections with "superbugs" such as *Acinetobacter baumannii*, *Neisseria gonorrhoeae, Pseudomonas aeruginosa*, *Streptococcus pneumoniae* and *Staphylococcus aureus*, which kill someone every 15 min in the USA [12-14]. Drug-resistant strains of *S. aureus* cause severe endocarditis, pneumonia, sepsis, and toxic shock syndrome [15]. Thus, there is an urgent need for the development of new classes of antibiotics to tackle such infections.

Microbial pathogenesis involves diverse pathways and mechanisms that lead to host colonisation and infection [16]. Virulence factors are secreted by the pathogen to facilitate this process, including peptidases that break down host defence proteins, regulate the availability of other secreted bacterial factors, and provide peptide nutrients for the pathogen. One example is the thermolysin family of bacterial metallopeptidases (MPs), also referred to as the M4 family according to the MEROPS database (www.ebi.ac.uk/merops) [17]. The archetype is *Bacillus thermoproteolyticus* thermolysin, which was the first endo-MP to be structurally resolved [18] and the founding member of the gluzincin clan of MPs [19, 20]. Related MPs produced by human pathogens include *P. aeruginosa* pseudolysin [21], vibriolysin from several *Vibrio* species [22], *Burkholderia cenocepacia* ZmpA/B [23], *Enterococcus faecalis* coccolysin [24], *Legionella pneumophila* Msp [25], *Clostridium perfringens* λ-toxin [26], and aureolysin from *Staphylococcus epidermidis* and *S. aureus* [16, 27-29].

Aureolysin was discovered in *S. aureus* strain V8 [30] and is the product of the *aur* gene, which is located on a monocistronic operon [31] and regulated by the alternative sigma factor σ^B and the staphylococcal accessory regulator SarA [31]. Aureolysin is prevalent in both pathogenic and commensal *S. aureus* strains [32], and peak abundance occurs during post-exponential growth and when the bacterial cells are phagocytosed by human neutrophils [33]. The enzyme accounts for ~50% of the total peptidase activity in culture supernatants [28] and participates in the extracellular peptidase system of *S. aureus* by activating the V8-type serine peptidase SspA, which in turn activates the cysteine peptidase SspB [16]. Together with the cysteine peptidase ScpA, they constitute the four major extracellular peptidases of *S. aureus* [34] known as the "staphylococcal proteolytic cascade" [31]. Moreover, aureolysin recruits nutrients from host proteins [35] and contributes to staphylococcal infections by promoting hypervirulence and the transition from a sessile, biofilm-forming lifestyle to a mobile, invasive phenotype [36, 37]. It degrades the human antimicrobial peptide LL-37 [38] and complement protein C3, while releasing the chemoattractant C5a to prevent complement-mediated killing by neutrophils [39]. It also contributes to the intracellular survival of *S. aureus* in human macrophages [40]. Furthermore, aureolysin highjacks the blood coagulation and fibrinolytic systems by activating prothrombin [41] and inactivating the serpin-type serine peptidase inhibitors α_1-proteinase inhibitor, α_2-antiplasmin, and α_1-antichymotrypsin by cleaving their "reactive-centre loops" (RCLs). This deregulates their targets such as neutrophil urokinase-type plasminogen activator, elastase, and plasmin [42-46]. Finally, aureolysin was shown to trigger osteoblast death and bone destruction in a murine model of osteomyelitis [29], which is a hallmark of *S. aureus* infection in humans [47]. Aureolysin is therefore a promising drug target given its role in the establishment and persistence of infection, which underpins its relevance for bacterial survival *in vivo* [48].

Aureolysin occurs as two isoforms (I and II) across distinct *S. aureus* strains that share 93% sequence identity [49]. It is exported as a 509-residue pre-pro-enzyme (UniProt access code [UP] P81177) comprising a 27-residue signal peptide for secretion, a 181-residue pro-domain (S_{28}–E_{208}; aureolysin residue numbering in subscript), and a 301-residue mature catalytic domain (CD, A_{209}–E_{509}; [27]) with 49% identity to thermolysin [28]. Once secreted, the zymogen is self-processed to yield the mature form [50], which (like other thermolysins [51]) prefers neutral pH and hydrophobic residues in the

substrate $\mathbf{P_1'}$ position [28] (nomenclature of enzyme sub-site and substrate positions on the non-primed and primed sides of the active-site cleft according to [52, 53]). Typically for MPs, the enzyme is inhibited by the general metal chelators EDTA and *o*-phenanthroline, as well as the nonspecific pan-peptidase inhibitor α_2-macroglobulin [28], but no specific small-molecule or protein inhibitors have yet been reported.

Small-molecule drugs are favoured in the clinic because they are often characterised by a long shelf life, oral bioavailability, efficient uptake by cells, and ease of manufacturing [54]. However, they generally have a small surface area for interaction with targets (usually large proteins), and this can limit their specificity and promote off-target effects. In contrast, therapeutic proteins (TPs) have larger surface areas, which result in higher selectivity, fewer toxic side effects, and tuneable dosage [54], often without harmful immune responses [55]. Although most TPs must be injected due to poor gastrointestinal absorption, various systems have been developed to overcome these limitations [56]. Recombinant TPs can also be redesigned to increase their specificity or efficacy. For example, defence proteins produced by one animal host against a class of bacterial virulence factors may be adapted to another host. Overall, this has increased the efficacy and potency of TPs, and they now account for ~10% of the broader pharmaceutical market [54].

The MP inhibitor from *Streptomyces nigrescens* was the first M4 family inhibitor (MEROPS I36) shown to target thermolysin, pseudolysin and griselysin [57, 58], but its mechanism of action remains unknown. In contrast, the mature 68-residue inducible insect metallopeptidase inhibitor (IMPI) from the greater wax moth *Galleria mellonella* (MEROPS I8; UP P82176) is a potent inhibitor of thermolysin, pseudolysin, vibriolysin, bacillolysin, and *Bacillus polymyxa* peptidase, and, importantly, its mechanism is known [59-63]. Moreover, IMPI is currently under development for the therapy of ectopic infections caused by *S. aureus* to cure chronic wounds formulated in poloxamer hydrogels, which caused no side effects in the swine ear model [63, 64]. We therefore sought a protein inhibitor of aureolysin for further development as a TP by designing several IMPI mutants with the ability to block aureolysin, and determined their mechanisms of action by kinetic and structural analysis.

MATERIALS AND METHODS

Expression constructs **–** Plasmid pIMPI-WT contains the sequence of wild-type (wt) IMPI in its mature form (residues I^{20}–S^{88}, superscript numbering based on UP P82176) [62]. It is a modified pET-32a vector, with the IMPI sequence inserted at the BglII and XhoI restriction sites, preceded by an N-terminal His_6-tagged thioredoxin fusion partner and a tobacco etch virus (TEV) peptidase recognition site, placing the peptide sequence G–M–S upstream of I^{20} in the final purified protein. We used pIMPI-WT to generate 13 mutants ($T^{50}N$, $T^{50}Q$, $T^{50}R$, $T^{50}Y$, $I^{54}M$, $I^{55}R$, $I^{55}W$, $I^{55}Y$, $I^{57}F$, $I^{57}Y$, $R^{58}E$, $T^{50}Y+I^{55}R$ and $T^{50}Y+I^{55}R+I^{57}F$). $T^{50}N$ was used only as an intermediate to prepare $T^{50}Y$ and was not tested for activity. The mutants were generated by site-directed mutagenesis with overlapping primers (Table 1) using Phusion high fidelity DNA polymerase (Thermo Fisher Scientific) according to the manufacturer's instructions. Template DNA was digested with *Dpn*I (Thermo Fisher Scientific) and the product was used to transform competent *Escherichia coli* DH5α cells (Thermo Fisher Scientific). Plasmid DNA was purified using the E.Z.N.A. Plasmid DNA Mini Kit I (Omega Bio-Tek) and all constructs were verified by sequencing (Eurofins and Macrogen).

Protein production and purification – The IMPI variants were expressed in *E. coli* BL21 (DE3) Origami2 cells (Novagen) transfected with the corresponding plasmid and grown on lysogeny broth (LB) agar supplemented with 100 µg/mL ampicillin. Single colonies were used to inoculate 25-mL LB starter cultures supplemented with 100 µg/mL ampicillin and 10 µg/mL tetracycline, and were incubated overnight at 37 °C under shaking. The starter cultures (1 mL) were used to inoculate 500 mL of the same medium, followed by cultivation under the same conditions until the OD_{550} reached 0.6. At this point, protein expression was induced with 0.2 mM isopropyl-β-D-1-thiogalactopyranoside (Thermo Fisher

Plasmid	Forward primer	Reverse primer	Template
pIMPI-T50N [a]	CATATACAGAATAAAAATAACTGTCCC	GGGACAGTTATTTTTATTCTGTATATG	***pIMPI-WT***
pIMPI-T50Q	CATATACAGAATAAACAAAACTGTCCC	GGGACAGTTTTGTTTATTCTGTATATG	***pIMPI-WT***
pIMPI-T50R	CAGAATAAACGAAACTGTCCCATC	GATGGGACAGTTTCGTTTATTCTG	***pIMPI-T50Q***
pIMPI-T50Y	CATATACAGAATAAATATAACTGTCCC	GGGACAGTTATATTTATTCTGTATATG	***pIMPI-T50N***
pIMPI-I54M	CTGTCCCATGATTAATATAAGATGTAAT GACAAGTGC	GCACTTGTCATTACATCTTATATTAATCA TGGGACAG	***pIMPI-WT***
pIMPI-I55R	GTCCCATCCGTAATATAAGATGTAATG	CATTACATCTTATATTACGGATGGGAC	***pIMPI-I55W***
pIMPI-I55W	CAAACTGTCCCATCTGGAATATAAGATG TAATGAC	GTCATTACATCTTATATTCCAGATGGGA CAGTTTG	***pIMPI-WT***
pIMPI-I55Y	CAAACTGTCCCATCTATAATATAAGATG TAATG	CATTACATCTTATATTATAGATGGGACA GTTTG	***pIMPI-WT***
pIMPI-I57F	CTGTCCCATCATTAATTTTAGATGTAATG ACAAGTGC	GCACTTGTCATTACATCTAAAATTAATG ATGGGACAG	***pIMPI-WT***
pIMPI-I57Y	CTGTCCCATCATTAATTATAGATGTAAT GACAAGTGC	GCACTTGTCATTACATCTATAATTAATG ATGGGACAG	***pIMPI-WT***
pIMPI-R58E	CATTAATATAGAATGTAATGACAAGTGC	GCACTTGTCATTACATTCTATATTAATG	***pIMPI-WT***
pIMPI-T50Y+I55R	CATATACAGAATAAATATAACTGTCCC	GGGACAGTTATATTTATTCTGTATATG	***pIMPI-I55R***
pIMPI-T50Y+I55R+I57F	CCCATCCGTAATTTTAGATGTAATGACA AGTGC	GCACTTGTCATTACATCTAAAATTACGG ATGGG	***pIMPI-T50Y+I55R***

Table 1. Plasmids and primers for overexpression. [a] This mutant was used as an intermediate to prepare $T^{50}Y$ and was not tested for activity. Only single nucleotides were exchanged in each reaction. For the double and triple mutants, a corresponding ancestral plasmid was used as the template.

Scientific) and overnight incubation at 18°C. Cells were harvested by centrifugation (3500×*g*; 30 min; 4°C), washed twice with cold buffer A (50 mM Tris·HCl, 250 mM sodium chloride, pH 8.0), and resuspended in the same buffer supplemented with 10 mM imidazole, the EDTA-free cOmplete protease inhibitor cocktail (Roche Life Sciences), and DNase I (Roche Life Sciences). Cells were lysed using a cell disrupter (Constant Systems) at a pressure of 135 MPa, and soluble protein was cleared by centrifugation (50,000×*g*; 1 h; 4°C) before passing the supernatant through a 0.22 µm filter (Merck Millipore). For immobilised-metal affinity chromatography (IMAC) [65], protein was captured on a nickel-Sepharose HisTrap HP column (Cytiva), previously washed and equilibrated with buffer A plus 500 or 20 mM imidazole. Each IMPI construct was purified on a separate column to avoid cross-contamination. IMPI was washed and eluted using buffer A supplemented with either 20 or 300 mM imidazole. Protein-containing fractions were dialysed for 4 h at room temperature against a 50-fold excess volume of buffer B (50 mM Tris·HCl, 150 mM sodium chloride, 0.5 mM oxidised glutathione, 3 mM reduced glutathione, pH 8.0) and centrifuged (50,000×*g*; 1 h; 4 °C) to remove precipitated protein. The inhibitors were dialysed overnight with His_6-tagged TEV peptidase (produced in-house) at a peptidase:substrate ratio of 1:20 (w/w) in buffer A at room temperature to remove the fusion partner. After centrifugation (50,000×*g*; 1 h; 4 °C) and 0.22 µm sterile filtration, the soluble fraction was loaded again onto the above HisTrap HP column for reverse IMAC. The flow-through fraction containing untagged inhibitor was collected, whereas TEV, thioredoxin and non-cleaved soluble IMPI aggregates bound to the column were eventually eluted using buffer A supplemented with 300 mM imidazole for column regeneration. The untagged IMPI was recovered after a second round of reverse IMAC, concentrated by exchange into buffer C (20 mM Tris·HCl, 150 mM sodium chloride, pH 8.0) using a HiPrep 26/10 desalting column (Cytiva), and polished by final size-exclusion chromatography (SEC) with buffer C in a Superdex 75 10/300 column (Cytiva) attached to an ÄKTA Purifier 10 apparatus (Cytiva).

Aureolysin isoform I was produced as previously described [66] with slight modifications. *S. aureus* V8-BC10 cells were streaked onto tryptic soy agar plates supplemented with 2.5 g/L glucose and 1% casein. A single colony, surrounded by a halo of digested casein, was then used to inoculate 20 mL of Bacto tryptic soy broth without dextrose (BD Biosciences) supplemented with 2.5 g/L glucose. This pre-inoculum was the same medium, followed by overnight cultivation under the same conditions. The bacterial supernatant was cleared by centrifugation (7000×*g*; 30 min; 4°C) and passed through a 0.22-µm filter. Supernatant proteins were then precipitated in ammonium sulfate (80% saturation) with gentle stirring for 4 h at 4°C, harvested by centrifugation (50,000×*g;* 1 h; 4°C), resuspended in buffer D (20 mM Tris·HCl, 10 mM calcium chloride, pH 7.8), and dialysed at 4°C overnight against

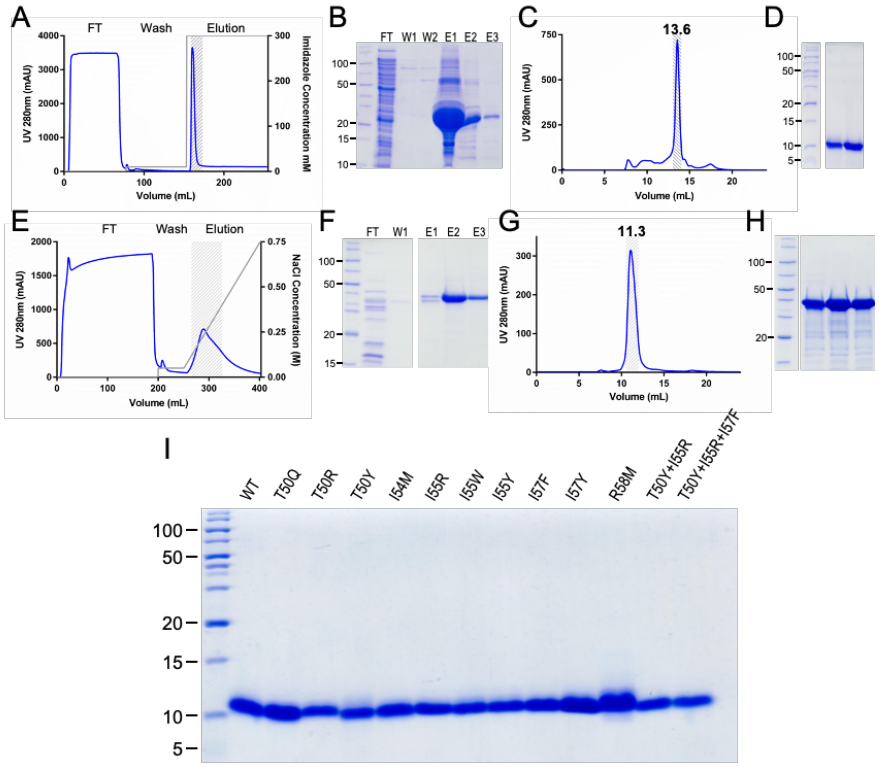


Figure 1. **Protein production and purification. (A)** Representative chromatogram and **(B)** SDS-PAGE analysis of the IMAC purification step of His_6-thioredoxin-tagged wt-IMPI (expected molecular mass ~25 kDa). FT, flow-through; W, wash step; E, elution step. **(C)** Chromatogram and **(D)** SDS-PAGE analysis of the SEC purification step of tag-depleted wt-IMPI (~8 kDa), which migrated as a monomer (13.6 mL). **(E)** Representative chromatogram and **(F)** SDS-PAGE analysis of the anion-exchange chromatography purification step of aureolysin. **(G)** Chromatogram and **(H)** SDS-PAGE analysis of the SEC purification step of aureolysin. Despite the higher-than-expected molecular mass reported by SDS-PAGE (panels F,H), the protein is indeed mature aureolysin (expected mass ~33 kDa), as confirmed by N-terminal sequencing, peptide-mass fingerprinting (Suppl. Fig. 1), and the retention volume in calibrated SEC (panel G; 11.3 mL) corresponding to ~29 kDa. **(I)** 20% Glycine SDS-PAGE showing the purity of wt-IMPI and the 12 mutants (2–5 µg) analysed herein. All constructs behaved similarly to (A-D) during purification and yielded products of comparable purity and molecular mass.

the same buffer. After centrifugation (50,000×*g*; 1 h; 4°C), the supernatant was loaded onto a 5-mL HiTrap Q FF anion exchange column (Cytiva) attached to an ÄKTA Pure 25 apparatus (Cytiva). The column was previously washed and equilibrated with buffer D, with or without 1 M sodium chloride. Protein bound to the column was washed extensively using buffer D supplemented with 50 mM sodium chloride, and eluted in a gradient of 50–750 mM sodium chloride in the same buffer. The purified aureolysin was polished by SEC in a Superdex 75 10/300 column with buffer E (20 mM Tris·HCl, 150 mM sodium chloride, 10 mM calcium chloride, 50 µM zinc chloride, pH 7.8).

Protein purity was assessed by SDS-PAGE on custom-made 14–20% glycine gels followed by staining with Coomassie Brilliant Blue (Sigma-Aldrich). Protein identities were confirmed by peptide mass fingerprinting (Suppl. Fig.1) and N-terminal sequencing (Edman degradation) at the Protein Chemistry Service and the Proteomics Facility of the Centro de Investigaciones Biológicas (CIB-CSIC, Madrid, Spain). Ultrafiltration was carried out using Vivaspin 15 and Vivaspin 2 filter devices with Hydrosart membranes and a 2-kDa cut-off (Sartorius Stedim Biotech). Protein concentrations were determined using the BCA protein assay kit (Thermo Fisher Scientific) by comparison to a dilution series of bovine serum albumin.

Activity and inhibition assays – We tested the proteolytic and peptidolytic activity of aureolysin, thermolysin from *B. thermoproteolyticus* Rokko (Sigma-Aldrich), and ulilysin (produced according to [67, 68]) at 37°C in 100-µL reactions containing buffer F (100 mM Tris·HCl,

Dataset	Aureolysin / wt-IMPI	Aureolysin / wt-IMPI	Aureolysin / I^{57}F-IMPI (2)
Beam line (synchrotron)	XALOC (ALBA)	XALOC (ALBA)	XALOC (ALBA)
Space group / complexes per a.u. [a]	$P4_1$ / 2	$P4_1$ / 2	$P4_1$ / 2
Twinning fraction α (-*k*,-*h*,-*l*)	0.490	0.490	0.536
Cell constants (a and c in Å)	68.14, 166.18	68.14, 166.18	68.08, 166.69
Wavelength (Å)	0.97926	0.97926	0.97926
Measurements / unique reflections	874,126 / 64,323	874,126 / 64,323	398,888 / 99,152
Resolution range (Å) (outermost shell) [c]	52.7 – 1.85 (1.96 – 1.85)	52.7 – 1.85 (1.96 – 1.85)	68.1 – 1.60 (1.70 – 1.60)
Completeness (%) / R_{merge} [d]	100.0 (99.8) / 0.149 (2.772)	100.0 (99.8) / 0.149 (2.772)	99.7 (99.4) / 0.050 (1.069)
R_{pim} [e] **/ $CC(^1/_2)$** [e]	0.042 (0.788) / 0.999 (0.630)	0.042 (0.788) / 0.999 (0.630)	0.029 (0.618) / 0.999 (0.580)
Average intensity [f]	14.7 (1.9)	14.7 (1.9)	14.2 (1.8)
B-Factor (Wilson) (Å²) / Aver. multiplicity	42.2 / 13.6 (13.4)	42.2 / 13.6 (13.4)	34.4 / 4.0 (4.0)
Resolution range used for refinement (Å)	52.7 – 1.85		68.1 – 1.60
Reflections used (test set)	63,598 (724)		98,470 (681)
Crystallographic R_{factor} (free R_{factor}) [d]	0.164 (0.219)		0.158 (0.188)
Non-H protein atoms / ionic ligands /	6467 / 6 Ca^{2+}, 2 Zn^{2+}		6322 / 6 Ca^{2+}, 2 Zn^{2+}
waters / non-ionic ligands per a.u.	709 / 2 PEG, 3 EDO		559 / 5 EDO
***Rmsd* from target values**			
bonds (Å) / angles (°)	0.008 / 1.64		0.008 / 1.76
Average B-factor (Å²)	38.1		32.6
Protein contacts and geometry analysis [b]			
Ramachandran favoured / outliers / all analysed	686 (95.0%) / 0 / 722		691 (95.5%) / 1 / 723
Bond-length / bond-angle / chirality / planarity outliers	0 / 3 / 0 / 2		0 / 2 / 0 / 3
Side-chain outliers	22 (3.6%)		15 (2.5%)
All-atom clashes / clashscore [b]	15 / 1.3		20 / 1.7
RSRZ outliers [b] **/ F_o:F_c correlation**	2 (0.3%) / 0.97 (0.95)		7 (1.0%) / 0.98 (0.97)
PDB access code	7SKM		7SKL

Table 2. Crystallographic data

[a] Abbreviations: EDO, ethylene glycol; PEG, diethylene glycol; RSRZ, real-space R-value Z-score. [b] According to the wwPDB Validation Service (https://wwpdb-validation.wwpdb.org/validservice). [c] Values in parenthesis refer to the outermost resolution shell if not otherwise indicated. [d] For definitions, see Table 1 in [94]. [e] For definitions, see [95, 96]. [f] Average intensity is <I/σ(I)> of unique reflections after merging according to *XSCALE* [70].

150 mM sodium chloride, 10 mM calcium chloride, 50 µM zinc chloride, pH 7.5) in an Infinite M200 microplate fluorimeter (Tecan). As substrates, we used 10 µg/mL of the pig-skin gelatin fluorescein conjugate from the DQ Gelatin EnzCheck assay kit (λ_{ex} = 485nm, λ_{em} = 528nm; Invitrogen, Thermo Fisher Scientific) or 20 µM FRET-4 (Abz-Y-G-K-R-V-F-K[dpn]-OH), an internally-quenched fluorogenic peptide (λ_{ex} = 260nm λ_{em} = 420nm; GenScript).

Inhibition by wt-IMPI was measured using both substrates following the pre-incubation of the inhibitor (up to 200-fold molar excess) with 100 nM aureolysin, 10 nM thermolysin or 10 nM ulilysin for 1 h at room temperature. Inhibition by the IMPI mutants (T^{50}Q, T^{50}R, T^{50}Y, I^{54}M, I^{55}R, I^{55}W, I^{55}Y, I^{57}F, I^{57}Y, R^{58}E, T^{50}Y+I^{55}R, and T^{50}Y+I^{55}R+I^{57}F) was measured using FRET-4 following the pre-incubation of each mutant (up to 100-fold molar excess) with 50 nM aureolysin for 1 h at room temperature. Reactions were carried out at 37°C in buffer G (20 mM Tris·HCl, 150 mM sodium chloride, pH 7.5) in triplicate and the residual proteolytic activity was measured for 3 h. The activity of the inhibitors in the absence of peptidase was monitored for the same duration as a negative control. To determine the relative activity of the IMPI mutants compared to the wild type, initial cleavage velocities of the fluorogenic protein and peptide substrates, without (V_0) and with (V_i) inhibitor, were determined from the slope of the linear range (R^2 > 90%) of the fluorescence *vs* time curve, and (V_0/V_i) was calculated using *GRAPHPAD PRISM* [69].

Complex formation and inhibitor cleavage detection – The complexes of aureolysin (at 100 µM) with wt-IMPI or the I^{57}F-mutant were prepared by incubation in buffer H

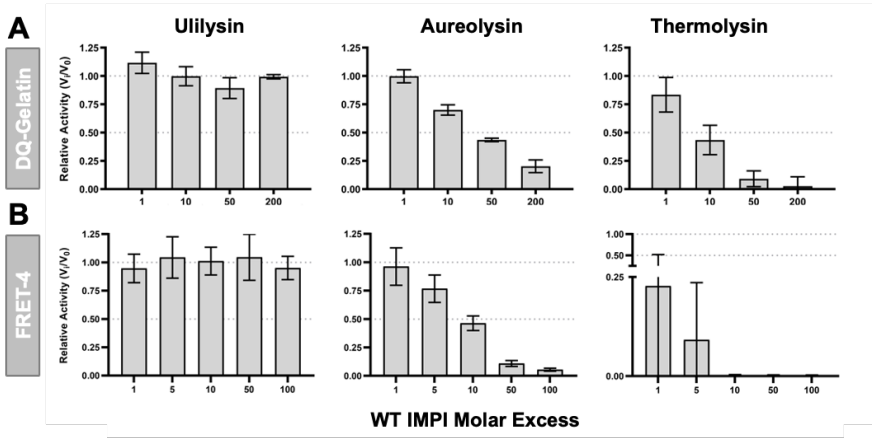


Figure 2. Inhibitory activity of wild-type IMPI. (A) Residual fractional activity as V_i/V_0 relative to the activity in the absence of inhibitor of (left) 10 nM ulilysin, (middle) 100 nM aureolysin, and (right) 10 nM thermolysin after incubation with wt-IMPI at several molar ratios using the DQ gelatin substrate. **(B)** As above, but using the internally quenched fluorescent FRET-4 peptide as the substrate.

(50 mM Tris·HCl, 150 mM sodium chloride, pH 8.0) at a 1:2.5 molar ratio for 30 min at room temperature. The complex was then disrupted by SEC in a Superdex 75 10/300 GL column (GE Healthcare) previously equilibrated in buffer H. The same amounts of aureolysin and inhibitor were processed separately as controls. IMPI cleavage was analysed by SDS-PAGE as above and mass spectrometry in a MALDI-TOF Autoflex III instrument (Bruker). Each sample was desalted using a C18 ZipTip (Millipore), mixed at a 1:1 ratio (v/v) with a matrix solution of 10 mg/mL sinapic acid in 50% acetonitrile, and spotted onto the plate using the dried-droplet method. Mass spectra were acquired in linear-mode geometry. Internal calibration was performed by correction of the average mass of the respective non-treated IMPI control sample (wt-IMPI: 7927.6 Da; I^{57}F-IMPI: 7967.1 Da).

Crystallisation and diffraction data collection – Crystallisation conditions were screened at the joint IRB/IBMB Automated Crystallography Platform using the sitting-drop vapor diffusion method. A Freedom EVO robot (Tecan) prepared screening solutions and dispensed them into the reservoir wells of 96×2-well MRC crystallisation plates (Innovadyne Technologies). A Phoenix/RE robot (Art Robbins) pipetted crystallisation nanodrops containing 100 nL of each protein and reservoir solution into the shallow wells, and plates were incubated in steady-temperature crystal farms (Bruker) at 4°C or 20ºC. Optimal aureolysin crystals complexed with either wt-IMPI or I^{57}F-IMPI formed at 20ºC in solutions containing 5 mg/mL aureolysin and 2.9 mg/mL IMPI (peptidase:inhibitor molar ratio of 1:2.5) in 50 mM Tris·HCl pH 8.0, 150 mM sodium chloride, 1.6 mM calcium chloride, 8.3 µM zinc chloride, which was mixed with reservoir solution consisting of 0.1 M Bis-Tris pH 5.5, 25% (w/v) PEG 3350 or 0.1 M Bis-Tris pH 6.0, 31% (w/v) PEG 2000 MME. Crystals were cryoprotected with reservoir solution plus 10% ethylene glycol, harvested using round LithoLoops of 0.04–0.1 mm (Molecular Dimensions), and flash-vitrified in liquid nitrogen for data collection. X-ray diffraction data were recorded at 100 K on a Pilatus 6M pixel detector (Dectris) at the XALOC beamline of the ALBA synchrotron (Cerdanyola, Catalonia, Spain) and on a Pilatus3 X 2M detector (Dectris) at the ID23-2 beamline of the ESRF synchrotron (Grenoble, France). Diffraction data were processed with programs *XDS* [70] and *XSCALE*, and transformed with *XDSCONV* to MTZ-format for the *PHENIX* [71] and *CCP4* [72] suites of programs. Statistics describing data collection and processing are provided in Table2.

Structure solution and refinement – The structure of the complex of aureolysin and I^{57}F-IMPI was solved by molecular replacement using *PHASER* [73] on a dataset initially processed as space group $P4_12_12$ at 2.05 Å resolution (Table2), with one complex

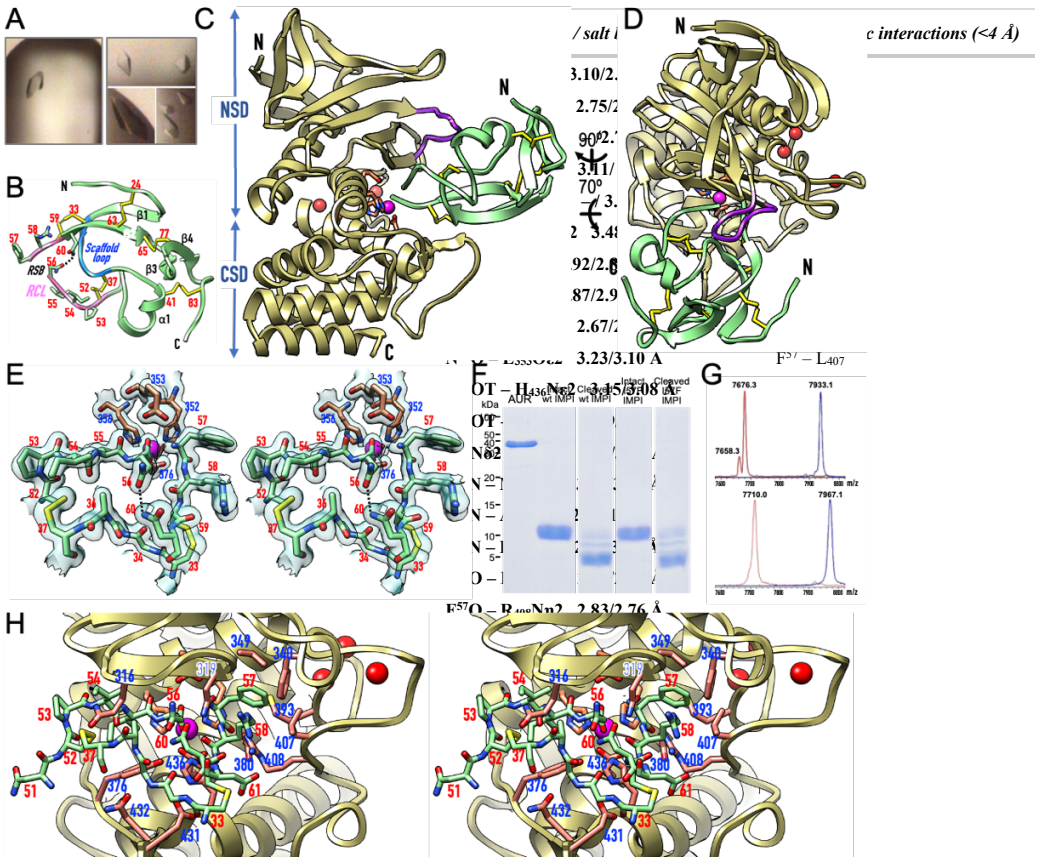


Figure 3. Structure of the IMPI–aureolysin complex. (A) Tetragonal protein crystals of the aureolysin–wt-IMPI (left) and aureolysin–I^{57}F-IMPI complexes (right). **(B)** Ribbon-type plot of I^{57}F-IMPI depicting the four β-strands (β1–β4) and the single helix (α1) of the structure, as well as the five disulfide bridges (with numbered cysteine residues). The scaffold loop is shown in blue, and the reactive-centre loop (RCL) is shown in pink with numbered residues (sticks). The cleaved reactive-site bond (RSB), N-terminus, and C-terminus are labelled. Hydrogen bond N^{56}Oδ1–N^{60}Nδ2 is needed to maintain the position of the **P$_1$** residue in place. **(C)** Ribbon-type plot of the complex between I^{57}F-IMPI (green ribbon, disulfide bonds as yellow sticks) and aureolysin (pale gold ribbon, catalytic zinc and structural calcium cations shown as magenta and red spheres, respectively) viewed along the active-site cleft (vertically rotated 90° counterclockwise away from the traditional "standard orientation" of MPs [53]). The side chains of the zinc-binding MP residues and the general/base acid glutamate are further shown as sticks for reference (carbons in salmon). The N-termini and C-termini are labelled, the characteristic flap is in purple, and the NSD and CSD of the peptidase are indicated. **(D)** Rotated view of (C). **(E)** Close-up in cross-eye stereo showing the RCL and scaffold loop of I^{57}F-IMPI (green carbons) and the zinc site of aureolysin (carbons in salmon) superposed with the final 1.60 Å (2mF_{obs}-DF_{calc})-type Fourier map as a semi-transparent surface contoured at 1 σ in a similar view to (D). The RSB is cleaved, selective inhibitor and MP residues are numbered in red and blue, respectively. Hydrogen bond N^{56}Oδ1–N^{60}Nδ2 is shown as a dashed line. **(F)** *In vitro* proof that binding and inhibition of aureolysin by wt- and I^{57}F-IMPI involves the cleavage of the inhibitor at the RSB (N^{56}–I^{57}) within the RCL as shown by SDS-PAGE analysis of the respective SEC fractions. **(G)** Mass spectra showing analysis of the cleavage of (top) intact wt-IMPI (blue spectrum; 7933.1 Da) to yield the cleaved inhibitor (red spectrum; 7676.3 Da) and (bottom) intact I^{57}F-IMPI (blue spectrum; 7967.1 Da) to yield the cleaved inhibitor (red spectrum; 7710.0 Da). Incubation of both intact species with aureolysin leads to the removal of the N-terminal tag-segment G-M-S (–275 Da) and the addition of a water molecule (+18 Da) due to RSB cleavage. For wt-IMPI, a small fraction of tag-depleted noncleaved inhibitor was detected (7658.3 Da). **(H)** Close-up in stereo of (D), further rotated 25° downwards and 25° leftwards, giving insight into the interactions between I^{57}F-IMPI (sticks with green carbons, residue numbers in red) and aureolysin (sticks with carbons in salmon, residue numbers in blue).

per asymmetric unit (a.u.). The coordinates of the protein part of unbound aureolysin (Protein Data Bank [PDB] access code 1BQB [27]) and wt-IMPI in a complex with *B. thermoproteolyticus* thermolysin (PDB 3SSB [62]) were used as searching models. These calculations yielded unique solutions for the peptidase and inhibitor at Eulerian angles (in °) α = 13.7, β = 29.4, γ = 153.6

Hydrogen bonds / salt bridges (<3.7 Å)	*Hydrophobic interactions (<4 Å)*
$Y^{31}O – K_{430}N\zeta$ 3.10/2.79 Å	$E^{32} – K_{430}$
$E^{32}O\varepsilon2 – D_{431}N$ 2.75/2.82 Å	$I^{54} – I_{326}$
$A^{36}O – Q_{317}N\varepsilon2$ – /2.72 Å	$I^{55} – H_{356}$
$D^{38}N – Q_{317}O\varepsilon1$ 3.11/ – Å	$I^{55} – Y_{367}$
$D^{38}N – Q_{317}N\varepsilon2$ – / 3.48 Å	$F^{57} – F_{340}$
$Q^{47}N\varepsilon2 - Q_{317}N\varepsilon2$ 3.48/ –	$F^{57} – L_{343}$
$I^{55}N – W_{325}O$ 2.92/2.86 Å	$F^{57} – V_{349}$
$I^{55}O – W_{325}N$ 2.87/2.92 Å	$F^{57} – H_{352}$
$N^{56}O – E_{353}O\varepsilon1$ 2.67/2.83 Å	$F^{57} – M_{396}$
$N^{56}O – E_{353}O\varepsilon2$ 3.23/3.10 Å	$F^{57} – L_{407}$
$N^{56}OT – H_{436}N\varepsilon2$ 3.15/3.08 Å	$R^{58} – F_{340}$
$N^{56}OT – Y_{367}O\eta$ 3.69/3.45 Å	$R^{58} – L_{407}$
$N^{56}N\delta2 – A_{323}O$ 2.82/3.00 Å	
$F^{57}N – N_{322}O\delta1$ 3.09/3.25 Å	
$F^{57}N – A_{323}O$ 3.27/3.14 Å	
$F^{57}N – E_{353}O\varepsilon2$ 2.90/3.01 Å	
$F^{57}O – R_{408}N\eta1$ 2.84/2.75 Å	
$F^{57}O – R_{408}N\eta2$ 2.83/2.76 Å	
$R^{58}N – N_{322}O\delta1$ 3.51/3.45 Å	
$R^{58}N\varepsilon – N_{321}O$ 2.85/2.47 Å	
$R^{58}N\eta1 – N_{321}O$ 2.74/ – Å	
$R^{58}O – N_{322}N\delta2$ 2.88/2.83 Å	
$N^{60}O\delta1 – N_{322}N\delta2$ 3.11/2.80 Å	
$K^{62}N\zeta – Q_{317}O\varepsilon1$ – / 2.66 Å	
Ionic interactions	
$N^{56}OT – Zn_{999}$ 2.11/2.08 Å	
$N^{56}O – Zn_{999}$ 2.61/2.38 Å	

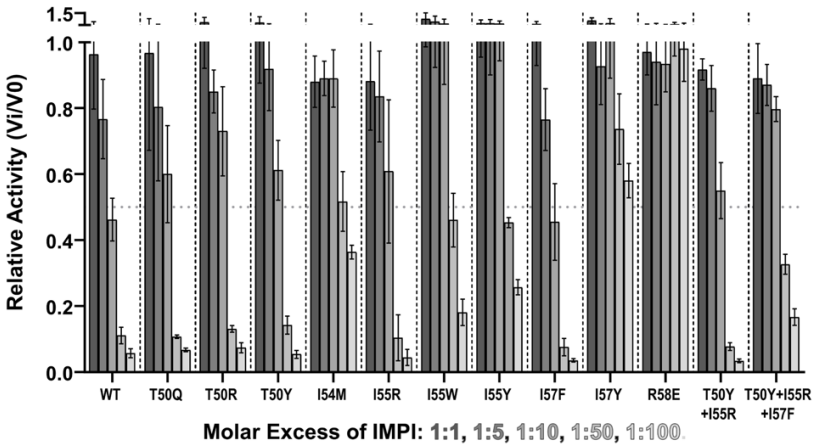
Table 3. Interactions at the I^{57}F-IMPI–aureolysin interface. The first residue/atom belongs to IMPI, the second to aureolysin. The two values for each bond correspond to complexes between protomers A/B and C/D, respectively.

(fractional cell coordinates 0.019, 0.287, 0.972) and α = 166.5, β = 131.6, γ = 104.7 (fractionalcell coordinates 0.751, 0.224, 0.191), respectively. The associated values for the translation functions after refinement were 15.6 and 34.0, and the final log-likelihood gain was 1316. The adequately rotated and translated molecules were refined using the *REFINE* protocol of *PHENIX* [74] and the *BUSTER* [75] program, including translation/libration/screw-motion (TLS) refinement. Unexpectedly, the free R_{factor} stalled at ~30% and the resulting Fourier maps were partially blurred, which together with the analysis of the intensity distribution with *XTRIAGE* [76] in *PHENIX*, and *POINTLESS* [77] in *CCP4*, indicated the presence of merohedral twinning following twin law (-*k*, -*h*, -*l*). At this point, a second dataset for the I^{57}F-IMPI complex with a higher resolution (1.60 Å) became available, which was processed with the actual space group $P4_1$ (Table 2) and solved by Fourier synthesis after rigid-body refinement of the two copies of the partially refined complex structure in the a.u. The structure was manually rebuilt using *COOT* [78] and refined using *REFMAC5* [79] considering twinning, as well as TLS and non-crystallographic symmetry (NCS) restraints. The final model included residues A_{209}–E_{509}, one zinc and three calcium ions of peptidase protomers A and C, as well as I^{20}–I^{86} and I^{20}–P^{84} of inhibitor moieties B and D, respectively, plus five ethylene glycol and 559 solvent molecules. Given that the structure of unbound aureolysin had originally been obtained before the gene sequence was available [27, 49], it contained five erroneous residues at positions 354, 361, 479, 492, and 493, which were corrected in the final model of the complex.

The structure of the wt-IMPI complex with aureolysin was solved at a resolution of 1.85 Å by Fourier synthesis after rigid-body refinement using the coordinates of the refined mutant complex structure. Model completion and refinement were carried out as described above. The final model comprised residues A_{209}–E_{509} and A_{209}–V_{508} of peptidase molecules A and C, plus one zinc and three calcium ions each, as well as I^{20}–I^{86} and I^{20}–K^{85} of inhibitor moieties B and D, respectively. Two diethylene glycol, three ethylene glycol, and 709 solvent molecules completed the model. Table2 provides essential statistics on the final refined models, which were validated using the wwPDB validation service (https://validate-rcsb-1.wwpdb.org/validservice) and deposited at www.pdb.org (access codes 7SKL and 7SKM).

Miscellaneous – Structural superpositions were calculated with *SSM* [80] in *COOT*. Figures were prepared using *CHIMERA* [81]. Protein interfaces and intermolecular interactions were analysed using *PDBEPISA* [82] (www.ebi.ac.uk/pdbe/pisa) and verified by visual inspection. The interacting surface of a complex was taken as half the sum of the buried surface areas of either molecule.

A



B

Aureolysin Relative Activity (V_i/V_0) ± SD (in %; n=3)						Inhibitory Constants	
IMPI Molar Excess	1	5	10	50	100	IC50	Estimated Ki
WT	96±17	77±12	46±6	11±3	6±1	468 ±48 nM	373 ±38 nM
T50Q	96±30	81±22	60±15	11±1	7±0	657 ±144 nM	523 ±115 nM
T50R	112±20	85±6	73±14	13±1	8±2	809 ±122 nM	644 ±97 nM
T50Y	111±23	92±13	61±9	14±3	5±1	611 ±91 nM	487 ±73 nM
I54M	88±8	89±5	89±9	52±9	36±2	3.06 ±0.37 µM	2.43 ±0.29 µM
I55R	88±15	84±14	61±22	10±7	4±2	715 ±119 nM	569 ±95 nM
I55W	124±26	114±22	105±18	46±8	18±4	2.19 ±0.36 µM	1.74 ±0.29 µM
I55Y	108±12	106±16	106±11	45±2	26±2	2.34 ±0.24 µM	1.86 ±0.19 µM
I57F	104±11	77±9	45±11	8±2	4±1	434 ±27 nM	346 ±22 nM
I57Y	118±12	93±12	108±18	74±11	58±5	5.67 ±1.97 µM	4.52 ±1.57 µM
R58E	97±7	94±13	93±8	103±8	98±10	n.d.	n.d.
T50Y+I55R	92±3	86±7	55±9	8±1	3±1	581 ±44 nM	463 ±35 nM
T50Y+I55R+I57F	89±10	87±6	80±4	33±3	17±3	1.52 ±0.15 µM	1.22 ±0.12 µM

Figure 4. Inhibitory activity of the IMPI mutants. (A) Relative fractional activity as V_i/V_0 of 50 nM aureolysin after incubation with IMPI mutants, relative to the wild type at molar ratios 1:1, 1:5, 1:10, 1:50 and 1:100 with FRET-4 (at 20 µM) as the substrate. Experiments were performed at least in triplicate, and error bars show standard deviations. **(B)** Tabular representation of the relative aureolysin activity data (in %) shown in (A). Average half-maximal inhibitory concentrations (IC_{50}) were determined using a four-parameter sigmoidal fit in *GRAPHPAD* (see Suppl. Fig. 2), and the inhibitor constant K_i was estimated using the equation $K_i = IC_{50}/([S]/K_M+1)$ [93].

RESULTS AND DISCUSSION

Assessment of wild-type IMPI as an aureolysin inhibitor and initial protein redesign – Wild-type IMPI was expressed in *E. coli* and recovered in a highly pure form (Fig.1A–D; Suppl. Fig.1) to assess its effect on aureolysin, which in turn was purified to homogeneity from cultures of *S. aureus* (Fig.1E–H; Suppl. Fig.1). The inhibitor was tested at molar ratios of 1:1 to 1:200 using a fluorogenic protein (Fig.2A) and a fluorogenic peptide (Fig.2B). We also tested thermolysin (the archetypal M4 family MP) and ulilysin, a metzincin MP from the pappalysin family (MEROPS M43B; [67, 68]) as controls. Thermolysin was efficiently inhibited as expected, whereas ulilysin was not inhibited at all, in agreement with IMPI being a specific inhibitor of M4 family MPs. Aureolysin was also inhibited in a dose-dependent manner, particularly when using the peptide substrate, although not to the same extent as thermolysin.

We superposed the structure of unbound aureolysin [27] onto thermolysin in a complex with wt-IMPI [62] and hypothesised that replacing I^{57} (whose side chain interacts with the MP, see below) with a bulkier residue such as phenylalanine might achieve stronger inhibition. Accordingly, we produced the mutant I^{57}F-IMPI as described above for wt-IMPI (Fig.1I) and used it for further analysis.

Overall structure of the IMPI–aureolysin complex – We crystallised I^{57}F-IMPI and wt-IMPI in complexes with aureolysin (Fig.3A) and used molecular replacement to solve their tetragonal ($P4_1$) crystal structures, which contained two complexes per a.u. Structural

solution and refinement (to 1.60 and 1.85 Å, respectively) was hindered by the presence of merohedral twinning in both crystals, with twinning fractions of 0.536 and 0.490, respectively (Table2). Even so, the structures were refined to final free R_{factor} values of 0.188 and 0.219,respectively, which are considered accurate. This was confirmed by the final Fourier maps (Fig.3E). The two structures were practically indistinguishable upon superposition, so the following discussion focuses on the I^{57}F-IMPI complex (protomers A and B) if not otherwise stated.

The structure of wt-IMPI has been reported in a complex with thermolysin [62]. Briefly, it has a spearhead shape (Fig.3B), whose tip contains a "reactive-site bond" (RBS; N^{56}–I/F^{57}) within a RCL (C^{52}–C^{59}). The latter is anchored to a subjacent "scaffold loop" (C^{33}–C^{37}) via two disulphide bonds, which are part of a set of five that confer structural rigidity. The regular secondary structures of IMPI comprise four β-strands (β1–β4) and one α-helix (α1).

The structure of the aureolysin CD is known for its unbound form [27]. It conforms to that of the thermolysin family and consists of an N-terminal subdomain (NSD; A_{209}–A_{363}, Fig. 3C) featuring an N-terminal β-barrel grafted into a frontal five-stranded mixed β-sheet whose lowermost strand forms the "upper-rim" of the active-site cleft (N_{322}–I_{326}; Fig. 3C,D). This element binds substrates in an extended conformation as an antiparallel β-ribbon. The NSD also contains a "backing helix" and an "active-site helix", encompassing the characteristic motif of the zincin MPs, H_{352}-E-X-X-H_{356} [83, 84]. The two histidine residues are ligands of the catalytic zinc, and the glutamate is the general base/acid of the cleavage reaction [85] (Fig. 3E). The main distinctive structural element of the aureolysin NSD compared to other thermolysins is a "flap" (N_{312}–N_{321}) that precedes the upper-rim strand and protrudes from the surface above the cleft (Fig. 3C,D).

The C-terminal subdomain (CSD; N_{364}–E_{509}; Fig. 3C) starts with the characteristic "glutamate helix" of gluzincins [19, 20], which contains the third zinc-binding protein ligand (E_{376}; Fig. 3E). It is followed by a long "irregular segment" (D_{388}–G_{434}) that shapes the bottom of the active-site cleft on its primed side, including the hydrophobic $\mathbf{S_1'}$ pocket. This pocket confers substrate specificity upon aureolysin and other M4 family MPs, as well as most other MP families [53]. Moreover, the irregular segment embraces three calcium-binding sites, which stabilise the structure [28]. The removal of these ions using chelators therefore causes irreversible inactivation [28, 86]. The CSD also contains a C-terminal four-helix bundle arranged as a Greek-key motif. Remarkably, the aureolysin CSD lacks the conspicuous β-ribbon that protrudes from the last turn of the first of these α-helices in thermolysin.

In the complex, I^{57}F-IMPI inserts like a wedge into the active-site cleft of the peptidase (Fig. 3C,D) and interacts via interfaces of 865 and 849 $Å^2$ (Δ^iG = –5.2 and –4.5 kcal/mol [82]) in complexes A/B and C/D, respectively. This involves 24 hydrogen bonds and salt bridges, plus two metallorganic bonds, as well as hydrophobic interactions between five inhibitor and 10 peptidase residues (Table 3). The main participating elements are the RCL and scaffold loop of the inhibitor, as well as the flap, upper-rim strand, $\mathbf{S_1'}$-pocket shaping residues, and the initial and final stretches of the irregular segment. Diverging from the thermolysin complex, superposition of the aureolysin complexes with wt-IMPI and I^{57}F-IMPI revealed a much smaller spread in the relative orientation between inhibitor and peptidase. The maximum deviation at the cleft-distal site of the inhibitor was ~4°/1.8 Å across the four complexes of the two structures, compared to ~10°/4.8 Å for the two thermolysin complexes in the a.u. [62].

Finally, superposition of IMPI-bound aureolysin with the unbound structure [27] revealed negligible differences between the NTS and CTS. This contrasts with thermolysin, where a 5° relative rotation of the two subdomains distinguishes between the unbound and bound forms [87]. Similar relative motion was proposed for *P. aeruginosa* elastase and *Bacillus cereus* neutral proteinase [27]. Aureolysin therefore does not appear to undergo the closing hinge motion when binding ligands or substrates, in contrast to other M4 family MPs.

IMPI inhibits aureolysin via the standard mechanism – The IMPI RCL runs across the peptidase cleft in the direction of the substrate, blocking $\mathbf{S_4}$–$\mathbf{S_1'}$ with residues P^{53}–I/F^{57} (Fig. 3H). Remarkably, the RSB was cleaved in the

crystals (Fig. 3E), which was verified *in vitro* by incubating both wt-IMPI and $I^{57}F$-IMPI with aureolysin. Indeed, both forms were cleaved at N^{56}–I/F^{57} (Fig. 3F,G). This feature causes the terminal carboxylate oxygen of the $\mathbf{P_1}$ residue, $N^{56}OT$, to bind the catalytic zinc and contribute to a distorted tetrahedral coordination sphere together with protein ligands $H_{352}N\varepsilon2$, $H_{356}N\varepsilon2$, and $E_{376}O\varepsilon2$ (all 2.02–2.11 Å apart in the various structures). $N^{56}OT$ replaces the two solvent molecules found in the unbound structure [27] and further contacts $H_{436}N\varepsilon2$ (3.08–3.15 Å), which is equivalent to H_{231} of thermolysin (thermolysin residues are shown in italics with subscript numbers for clarity). Together with Y_{157}, equivalent to Y_{376} in aureolysin, the residue helps to stabilise the tetrahedral reaction intermediate [85]. Moreover, the other carboxylate oxygen of N^{56} is very close to the general base/acid glutamate ($N^{56}O$–$E_{353}O\varepsilon1$; 2.60–2.67 Å), indicating that one of them must be protonated. On the primed side of the cleft, $\mathbf{P_1'}$ residue I/F^{57} is bound via its α-amino group to $E_{353}O\varepsilon2$ (2.90–3.01 Å) and the upper-rim main-chain carbonyl of A_{323} (3.14–3.27 Å) as well as the side-chain carboxamide of N_{322} (3.09–3.25 Å; Fig. 3E,H).

The inhibition mode described above agrees with the "standard mechanism" or "canonical mechanism" of peptidase inhibition [88, 89]. Remarkably, in standard-mechanism inhibitors (which mostly target serine endopeptidases), the RSB is cleaved very slowly because the cleavage reaction is kinetically unfavourable, so the intact complexes have half-lives of several years [90]. This has been verified by many crystal structures with intact RSBs [91]. In contrast, IMPI represents a unique case of a standard-mechanism MP inhibitor occurring as a cleaved inhibitor, first in its thermolysin complex [62] and now here with aureolysin, whose 69-residue structure is kept rigid through five disulphide bonds that are evenly distributed across the structure.

Finally, in the aureolysin complexes, the cleaved RSB is poised for rejoining, which is another functional requisite of the standard mechanism [91]. This is indicated by the proximity and orientation of the α-amino group of I/F^{57} relative to the carboxylate carbon of N^{56}, which are ideally situated for a nucleophilic attack. Indeed, the angle $I/F^{57}N$–$N^{56}C$–$N^{56}OT$, where $N^{56}OT$ is the oxygen that is not bound to the general base/acid glutamate, is ~110º on average over all four $I^{57}F$-IMPI and wt-IMPI complexes, thus in good agreement with the value postulated for a productive Bürgi-Dunitz geometrical reaction coordinate (105 ± 5º [92]). This is supported by the ability of cleaved wt-IMPI to rejoin *in vitro* following the addition of catalytic amounts of thermolysin [62].

Redesign of IMPI – Based on the IMPI–aureolysin crystal structures described above, we identified positions 50, 54, 55, 57 and 58 of the RCL as ideal for mutagenesis and constructed 11 single, double and triple point mutants in addition to the wt-IMPI and $I^{57}F$-IMPI variants ($T^{50}Q$, $T^{50}R$, $T^{50}Y$, $I^{54}M$, $I^{55}R$, $I^{55}W$, $I^{55}Y$, $I^{57}Y$, $R^{58}E$, $T^{50}Y+I^{55}R$, and $T^{50}Y+I^{55}R+I^{57}F$). All variants were produced and purified as efficiently as described above for wt-IMPI (Fig. 1I), and were compared to wt-IMPI for their ability to inhibit aureolysin at molar ratios of 1:1 to 1:100 using the fluorogenic peptide FRET-4 as the substrate (Fig. 4A,B). $R^{58}E$ did not affect peptidase activity. We tested the mutant with thermolysin, which revealed ~200-fold weaker inhibition than the wild type (Suppl. Fig. 2). We thus conclude that the mutant was properly folded, as suggested by its behaviour during purification, but functionally impaired and thus unable to block thermolysins. The rest of the cohort of mutants achieved the concentration-dependent inhibition of aureolysin. They could be assigned to two groups, one similar to the wild type, with residual activities of 3–8% at the highest molar ratio (Fig. 4B), whereas and the others showing weaker inhibition, with residual activities of 17–58% (Fig. 4B). The derived IC_{50} values enabled us to estimate K_i values of 346–644 nM for the first group and 1220–4520 nM for the second group (Fig. 4B). Notably, mutant $I^{57}F$ (from the initial stage of the project, see above) achieved the highest inhibition among all variants tested (K_i= 346 nM) and would thus provide a suitable lead for further development.

COROLLARY

Aureolysin plays multiple roles during *S. aureus* infections and is a promising target for the development of novel antimicrobials. We

tested the M4-specific inhibitor IMPI, and found that it inhibited the peptidase using the standard mechanism, best described for serine endopeptidases, based on the analysis of crystal structures. We therefore designed a cohort of point mutants, with $I^{57}F$ emerging as the strongest inhibitor. This is, to our knowledge, the first report of a TP candidate that can inhibit one of the major proteolytic virulence factors of *S. aureus*. The only other protein-based inhibitor with this ability is the general pan-peptidase inhibitor α_2-macroglobulin, which has a molecular mass of ~720 kDa and a broad spectrum of targets, making it unsuitable for therapeutic applications. Cell-based and disease challenge studies are now required to confirm the potential of $I^{55}R$-IMPI as a TP for the treatment of *S. aureus* infections.

ACKNOWLEDGMENTS

We are grateful to Laura Company, Xandra Kreplin and Joan Pous from the joint IBMB/IRB Automated Crystallography Platform and the Protein Purification Service at IBMB for assistance during purification and crystallisation experiments, and Carme Quero from the Institut de Química Avançada de Catalunya (IQAC-CSIC) is thanked for assistance with mass spectrometry. We also acknowledge the kind gift of *S. aureus* strain V8-BC10 (for aureolysin production) from Jan Potempa, Jagiellonian University of Kraków, Poland. The authors would also like to thank the ESRF and ALBA synchrotrons for beamtime allocation and the beamline staff for assistance during diffraction data collection. This study was supported in part by Spanish and Catalan public and private bodies that provided funding to the Proteolysis Lab (grants PID2019-107725RG-I00 from MCIN/AEI/10.13039/501100011033, 2017SGR3 from the National Government of Catalonia, and 201815 from Fundació "La Marató de TV3"). S.M.E. acknowledges grant BES2016-076877 from the Spanish State Agency for Research (MCIN/AEI/10.13039/501100011033) and the European Social Fund "ESF invests in your future". U.E. acknowledges a "Beatriu-de-Pinós" COFUND fellowship from the National Government of Catalonia (2018BP00163). A.V. and P.C. acknowledge funding from the German Federal Ministry for Education and Research (BMBF) through project "4-In" (Inhalable Virulence-Inhibitors from Insects for the Therapy of lung infections, ref. 16GW0137K). The authors thank Richard M. Twyman for editing the manuscript.

AUTHOR CONTRIBUTIONS

F.X.G.R. and A.V. conceived, supervised, and funded the project; S.R.M. produced and purified all proteins, prepared the mutants, performed *in vitro* studies with U.E. and P.C., analysed kinetics and mass spectrometry data with U.E., and crystallised proteins with assistance from U.E. and T.G.; S.R.M., U.E. and A.R.-B. collected diffraction data, and U.E. performed initial data analysis; E.M. performed biocomputational calculations; F.X.G.R. solved and refined crystal structures; and F.X.G.R. and A.V. wrote the manuscript with contributions from all authors.

COMPETING INTERESTS

The authors declare no financial or non-financial conflicts of interest with the contents of this article.

REFERENCES

[1] WHO (2014) Antimicrobial resistance: global report on surveillance 2014. . Geneva: WHO. 257 p.

[2] Sabtu N, Enoch DA, Brown NM (2015) Antibiotic resistance: what, why, where, when and how? Br Med Bull 116:105-113.

[3] Bengtsson-Palme J, Kristiansson E, Larsson DGJ (2018) Environmental factors influencing the development and spread of antibiotic resistance. FEMS Microbiol Rev 42:fux053.

[4] Odenholt I, Gustafsson I, Lowdin E, Cars O (2003) Suboptimal antibiotic dosage as a risk factor for selection of penicillin-resistant *Streptococcus pneumoniae*: i*n vitro* kinetic model. Antimicrob Agents Chemother 47:518-523.

[5] Delepierre A, Gayot A, Carpentier A (2012) Update on counterfeit antibiotics worldwide; public health risks. Med Mal Infect 42:247-255.

[6] van Boeckel TP, Brower C, Gilbert M, Grenfell BT, Levin SA *et al.* (2015) Global trends in antimicrobial use in food animals. Proc Natl Acad Sci USA 112:5649-5654.

[7] Livermore D (2004) Can better prescribing turn the tide of resistance? Nat Rev Microbiol 2:73-78.

[8] Diekema DJ, Jones RN (2001) Oxazolidinone antibiotics. Lancet 358:1975-1982.

[9] Strieker M, Marahiel MA (2009) The structural diversity of acidic lipopeptide antibiotics. Chembiochem : a European journal of chemical biology 10:607-616.

[10] Projan SJ (2003) Why is big Pharma getting out of antibacterial drug discovery? Curr Opin Microbiol 6:427-30.

[11] Norrby SR, Nord CE, Finch R, ESCMID f (2005) Lack of development of new antimicrobial drugs: a potential serious threat to public health. Lancet Infect Dis 5:115-119.

[12] Livermore DM (2004) The need for new antibiotics. Clin Microbiol Infect 10 Suppl 4:1-9.

[13] Hawkey PM, Warren RE, Livermore DM, McNulty CAM, Enoch DA *et al.* (2018) Treatment of infections caused by multidrug-resistant Gram-negative bacteria: report of the British Society for Antimicrobial Chemotherapy/Healthcare Infection Society/British Infection Association Joint Working Party. J Antimicrob Chemother 73:iii2-iii78.

[14] Nelson RE, Hatfield KM, Wolford H, Samore MH, Scott RD *et al.* (2021) National estimates of healthcare costs associated with multidrug-resistant bacterial infections among hospitalized patients in the United States. Clin Infect Dis 72:S17-S26.

[15] David MZ, Daum RS (2010) Community-associated methicillin-resistant *Staphylococcus aureus*: epidemiology and clinical consequences of an emerging epidemic. Clin Microbiol Rev 23:616-687.

[16] Martínez-García S, Rodríguez-Martínez S, Cancino-Díaz ME, Cancino-Díaz JC (2018) Extracellular proteases of *Staphylococcus epidermidis*: roles as virulence factors and their participation in biofilm. APMIS 126:177-185.

[17] Rawlings ND, Bateman A (2021) How to use the MEROPS database and website to help understand peptidase specificity. Protein Sci 30:83-92.

[18] Matthews BW, Jansonius JN, Colman PM, Schoenborn BP, Dupourque D (1972) Three-dimensional structure of thermolysin. Nature 238:37-41.

[19] Hooper NM (1994) Families of zinc metalloproteases. FEBS Lett 354:1-6.

[20] Cerdà-Costa N, Gomis-Rüth FX (2014) Architecture and function of metallopeptidase catalytic domains. Prot Sci 23:123-144.

[21] Galdino ACM, de Oliveira MP, Ramalho TC, de Castro AA, Branquinha MH *et al.* (2019) Anti-virulence strategy against the multidrug-resistant bacterial pathogen *Pseudomonas aeruginosa*: pseudolysin (elastase B) as a potential druggable target. Curr Protein Pept Sci 20:471-487.

[22] Miyoshi S-I (2013) Extracellular proteolytic enzymes produced by human pathogenic vibrio species. Front Microbiol 4:339.

[23] Kooi C, Subsin B, Chen R, Pohorelic B, Sokol PA (2006) *Burkholderia cenocepacia* ZmpB is a broad-specificity zinc metalloprotease involved in virulence. Infect Immun 74:4083-4093.

[24] Makinen PL, Makinen KK (1994) The *Enterococcus faecalis* extracellular metalloendopeptidase (EC 3.4.24.30; coccolysin) inactivates human endothelin at bonds involving hydrophobic amino acid residues. Biochem Biophys Res Commun 200:981-985.

[25] Sahney NN, Summersgill JT, Ramírez JA, Miller RD (2001) Inhibition of oxidative burst and chemotaxis in human phagocytes by *Legionella pneumophila* zinc metalloprotease. J Med Microbiol 50:517-525.

[26] Okabe A, Matsushita O (2013) Chapter 113 - Lambda toxin (*Clostridium perfringens*). In: Rawlings ND, Salvesen GS, editors. Handbook of Proteolytic Enzymes. Oxford: Academic Press. pp. 561-563.

[27] Banbula A, Potempa J, Travis J, Fernández-Catalán C, Mann K *et al.* (1998) Amino-acid sequence and three-dimensional structure of the *Staphylococcus aureus* metalloproteinase at 1.72Å. resolution. Structure 6:1185-1193.

[28] Potempa J, Shaw LN (2013) Chapter 114 - Aureolysin. In: Rawlings ND, Salvesen GS, editors. Handbook of Proteolytic Enzymes. Oxford: Academic Press. pp. 563-569.

[29] Cassat JE, Hammer ND, Campbell JP, Benson MA, Perrien DS *et al.* (2013) A secreted bacterial protease tailors the *Staphylococcus aureus* virulence repertoire to modulate bone remodeling during osteomyelitis. Cell host & microbe 13:759-772.

[30] Arvidson S, Holme T, Lindholm B (1972) The formation of extracellular proteolytic enzymes by *Staphylococcus aureus*. Acta Pathol Microbiol Scand B Microbiol Immunol 80:835-844.

[31] Shaw L, Golonka E, Potempa J, Foster SJ (2004) The role and regulation of the extracellular proteases of *Staphylococcus aureus*. Microbiology 150:217-228.

[32] Dubin G (2002) Extracellular proteases of Staphylococcus spp. Biological chemistry 383:1075-86.

[33] Burlak C, Hammer CH, Robinson MA, Whitney AR, McGavin MJ *et al.* (2007) Global analysis of community-associated methicillin-resistant *Staphylococcus aureus* exoproteins reveals molecules produced *in vitro* and during infection. Cell Microbiol 9:1172-1190.

[34] Elmwall J, Kwiecinski J, Na M, Ali AA, Osla V *et al.* (2017) Galectin-3 Is a target for proteases involved in the virulence of *Staphylococcus aureus*. Infect Immun 85:00177-17.

[35] Lehman MK, Nuxoll AS, Yamada KJ, Kielian T, Carson SD *et al.* (2019) Protease-mediated growth of *Staphylococcus aureus* on host proteins Is *opp3* dependent. mBio 10:02553-18.

[36] Martí M, Trotonda MP, Tormo-Más MA, Vergara-Irigaray M, Cheung AL *et al.* (2010) Extracellular proteases inhibit protein-dependent biofilm formation in *Staphylococcus aureus*. Microbes Infect 12:55-64.

[37] Gimza BD, Jackson JK, Frey AM, Budny BG, Chaput D *et al.* (2021) Unraveling the impact of secreted proteases on hypervirulence in *Staphylococcus aureus*. mBio 12:e03288-20.

[38] Sieprawska-Lupa M, Mydel P, Krawczyk K, Wojcik K, Puklo M *et al.* (2004) Degradation of human antimicrobial peptide LL-37 by *Staphylococcus aureus*-derived proteinases. Antimicrob Agents Chemother 48:4673-4679.

[39] Laarman AJ, Ruyken M, Malone CL, van Strijp JA, Horswill AR *et al.* (2011) *Staphylococcus aureus* metalloprotease aureolysin cleaves complement C3 to mediate immune evasion. J Immunol 186:6445-6453.

[40] Kubica M, Guzik K, Koziel J, Zarebski M, Richter W *et al.* (2008) A potential new pathway for *Staphylococcus aureus* dissemination: the silent survival

of *S. aureus* phagocytosed by human monocyte-derived macrophages. PloS one 3:e1409.

[41] Pietrocola G, Nobile G, Rindi S, Speziale P (2017) *Staphylococcus aureus* manipulates innate immunity through own and host-expressed proteases. Front Cell Infect Microbiol 7:166.

[42] Potempa J, Watorek W, Travis J (1986) The inactivation of human plasma α1-proteinase inhibitor by proteinases from *Staphylococcus aureus.* J Biol Chem 261:14330-14334.

[43] Potempa J, Dubin A, Watorek W, Travis J (1988) An elastase inhibitor from equine leukocyte cytosol belongs to the serpin superfamily. Further characterization and amino acid sequence of the reactive center. J Biol Chem 263:7364-9.

[44] Potempa J, Fedak D, Dubin A, Mast A, Travis J (1991) Proteolytic inactivation of α_1-anti-chymotrypsin. Sites of cleavage and generation of chemotactic activity. J Biol Chem 266:21482-21487.

[45] Potempa J, Wunderlich JK, Travis J (1991) Comparative properties of three functionally different but structurally related serpin variants from horse plasma. Biochem J 274 (Pt 2):465-471.

[46] Beaufort N, Wojciechowski P, Sommerhoff CP, Szmyd G, Dubin G *et al.* (2008) The human fibrinolytic system is a target for the staphylococcal metalloprotease aureolysin. Biochem J 410:157-165.

[47] Lew DP, Waldvogel FA (2004) Osteomyelitis. Lancet 364:369-379.

[48] Jusko M, Potempa J, Kantyka T, Bielecka E, Miller HK *et al.* (2014) Staphylococcal proteases aid in evasion of the human complement system. J Innate Immun 6:31-46.

[49] Sabat A, Kosowska K, Poulsen K, Kasprowicz A, Sekowska A *et al.* (2000) Two allelic forms of the aureolysin gene (*aur*) within *Staphylococcus aureus*. Infect Immun 68:973-976.

[50] Nickerson NN, Joag V, McGavin MJ (2008) Rapid autocatalytic activation of the M4 metalloprotease aureolysin is controlled by a conserved N-terminal fungalysin-thermolysin-propeptide domain. Mol Microbiol 69:1530-1543.

[51] Adekoya OA, Sylte I (2009) The thermolysin family (M4) of enzymes: therapeutic and biotechnological potential. Chem Biol Drug Des 73:7-16.

[52] Schechter I, Berger A (1967) On the size of active site in proteases. I. Papain. Biochem Biophys Res Commun 27:157-162.

[53] Gomis-Rüth FX, Botelho TO, Bode W (2012) A standard orientation for metallopeptidases. Biochim Biophys Acta 1824:157-163.

[54] Craik DJ, Fairlie DP, Liras S, Price D (2013) The future of peptide-based drugs. Chem Biol Drug Des 81:136-147.

[55] Dingermann T (2008) Recombinant therapeutic proteins: production platforms and challenges. Biotechnol J 3:90-97.

[56] Bruno BJ, Miller GD, Lim CS (2013) Basics and recent advances in peptide and protein drug delivery. Ther Deliv 4:1443-1467.

[57] Oda K, Koyama T, Murao S (1979) Purification and properties of a proteinaceous metallo-proteinase inhibitor from *Streptomyces nigrescens* TK-23. Biochim Biophys Acta 571:147-156.

[58] Seeram SS, Hiraga K, Oda K (1997) Resynthesis of reactive site peptide bond and temporary inhibition of *Streptomyces* metalloproteinase inhibitor. J Biochem 122:788-794.

[59] Wedde M, Weise C, Kopacek P, Franke P, Vilcinskas A (1998) Purification and characterization of an inducible metalloprotease inhibitor from the hemolymph of greater wax moth larvae, *Galleria mellonella*. Eur J Biochem 255:535-543.

[60] Clermont A, Wedde M, Seitz V, Podsiadlowski L, Lenze D *et al.* (2004) Cloning and expression of an inhibitor of microbial metalloproteinases from insects contributing to innate immunity. Biochem J 382:315-322.

[61] Wedde M, Weise C, Nuck R, Altincicek B, Vilcinskas A (2007) The insect metalloproteinase inhibitor gene of the lepidopteran *Galleria mellonella* encodes two distinct inhibitors. Biol Chem 388:119-127.

[62] Arolas JL, Botelho TO, Vilcinskas A, Gomis-Rüth FX (2011) Structural evidence for standard-mechanism inhibition in metallopeptidases from a complex poised to resynthesize a peptide bond. Angew Chem Int Ed Engl 50:10357-10360.

[63] Eisenhardt M, Schlupp P, Höfer F, Schmidts T, Hoffmann D *et al.* (2019) The therapeutic potential of the insect metalloproteinase inhibitor against infections caused by *Pseudomonas aeruginosa*. J Pharm Pharmacol 71:316-328.

[64] Eisenhardt M, Dobler D, Schlupp P, Schmidts T, Salzig M *et al.* (2015) Development of an insect metalloproteinase inhibitor drug carrier system for application in chronic wound infections. J Pharm Pharmacol 67:1481-1491.

[65] Block H, Maertens B, Spriestersbach A, Brinker N, Kubicek J *et al.* (2009) Immobilized-metal affinity chromatography (IMAC): a review. Methods Enzymol 463:439-473.

[66] Sabat AJ, Wladyka B, Kosowska-Shick K, Grundmann H, van Dijl JM *et al.* (2008) Polymorphism, genetic exchange and intragenic recombination of the aureolysin gene among *Staphylococcus aureus* strains. BMC Microbiol 8:129.

[67] Tallant C, García-Castellanos R, Seco J, Baumann U, Gomis-Rüth FX (2006) Molecular analysis of ulilysin, the structural prototype of a new family of metzincin metalloproteases. J Biol Chem 281:17920-17928.

[68] Huesgen PF, Lange PF, Rogers LD, Solis N, Eckhard U *et al.* (2015) LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. Nat Methods 12:55-58.

[69] Motulsky H, Christopoulos A (2004) Fitting models to biological data using linear and nonlinear regression. A practical guide to curve fitting. New York: Oxford University Press. 352 p.

[70] Kabsch W (2010) XDS. Acta Crystallogr sect D 66:125-132.

[71] Adams PD, Afonine PV, Bunkoczi G, Chen VB, Davis IW *et al.* (2010) PHENIX: a comprehensive Python-based system for macromolecular structure solution. Acta Crystallogr sect D 66:213-221.

[72] Winn MD, Ballard CC, Cowtan KD, Dodson EJ, Emsley P *et al.* (2011) Overview of the CCP4 suite and current developments. Acta Crystallogr sect D 67:235-242.

[73] McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC *et al.* (2007) Phaser crystallographic software. J Appl Crystallogr 40:658-674.

[74] Liebschner D, Afonine PV, Baker ML, Bunkóczi G, Chen VB *et al.* (2019) Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. Acta Crystallogr sect D 75:861-877.

[75] Smart OS, Womack TO, Flensburg C, Keller P, Paciorek W *et al.* (2012) Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. Acta Crystallogr sect D 68:368-380.

[76] Zwart PH, Grosse-Kunstleve RW, Adams PD (2005) Xtriage and Fest: automatic assessment of X-ray data and substructure structure factor estimation. In: Remacle F, editor. CCP4 Newsletter on Protein Crystallography. Daresburt, Warrington (UK): Daresbury Laboratory. pp. 27-35.

[77] Evans PR (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta Crystallogr sect D 67:282-292.

[78] Casañal A, Lohkamp B, Emsley P (2020) Current developments in *Coot* for macromolecular model building of electron cryo-microscopy and crystallographic data. Protein Sci 29:1069-1078.

[79] Kovalevskiy O, Nicholls RA, Long F, Carlon A, Murshudov GN (2018) Overview of refinement procedures within *REFMAC*5: utilizing data from different sources. Acta Crystallogr sect D 74:215-227.

[80] Krissinel E, Henrick K (2004) Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr sect D 60:2256-2268.

[81] Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS *et al.* (2018) UCSF ChimeraX: meeting modern challenges in visualization and analysis. Protein Sci 27:14-25.

[82] Krissinel E, Henrick K (2007) Inference of macromolecular assemblies from crystalline state. J Mol Biol 372:774-797.

[83] McKerrow JH (1987) Human fibroblast collagenase contains an amino acid sequence homologous to the zinc-binding site of *Serratia* protease. J Biol Chem 262:5943-5943.

[84] Bode W, Gomis-Rüth FX, Stöcker W (1993) Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the ´metzincins´. FEBS Lett 331:134-140.

[85] Matthews BW (1988) Structural basis of the action of thermolysin and related zinc peptidases. Acc Chem Res 21:333-340.

[86] Wasylewski Z, Stryjewski W, Wasniowska A, Potempa J, Baran K (1986) Effect of calcium binding on conformational changes of staphylococcal metalloproteinase measured by means of intrinsic protein fluorescence. Biochim Biophys Acta 871:177-181.

[87] Hausrath AC, Matthews BW (2002) Thermolysin in the absence of substrate has an open conformation. Acta Crystallogr sect D 58:1002-1007.

[88] Laskowski Jr. M, Kato I (1980) Protein inhibitors of proteinases. Annu Rev Biochem 49:593-626.

[89] Bode W, Huber R (1992) Natural protein proteinase inhibitors and their interaction with proteinases. Eur J Biochem 204:433-451.

[90] Ascenzi P, Bocedi A, Bolognesi M, Spallarossa A, Coletta M *et al.* (2003) The bovine basic pancreatic trypsin inhibitor (Kunitz inhibitor): a milestone protein. Curr Protein Pept Sci 4:231-251.

[91] Laskowski Jr. M, Qasim MA (2000) What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? Biochim Biophys Acta 1477:324-337.

[92] Bürgi HB, Dunitz JD, Shefter E (1973) Geometrical reaction coordinate. II. Nucleophilic addition to a carbonyl group. J Am Chem Soc 95:5065-5067.

[93] Cer RZ, Mudunuri U, Stephens R, Lebeda FJ (2009) IC_{50}-to-K_i: a web-based tool for converting IC_{50} to K_i values for inhibitors of enzyme activity and ligand binding. Nucleic Acids Res 37:W441-W445.

[94] García-Castellanos R, Marrero A, Mallorquí-Fernández G, Potempa J, Coll M *et al.* (2003) Three-dimensional structure of MecI : Molecular basis for transcriptional regulation of staphylococcal methicillin resistance. J Biol Chem 278:39897-39905.

[95] Weiss MS (2001) Global indicators of X-ray quality. J Appl Cryst 34:130-135.

[96] Karplus PA, Diederichs K (2012) Linking crystallographic model and data quality. Science 336:1030-1033.

SUPPLEMENTARY INFORMATION

"Design of a competitive protein inhibitor for aureolysin, a virulence factor of Staphylococcus aureus"

Soraia R. Mendes, Ulrich Eckhard [*], Arturo Rodríguez-Banqueri, Tibisay Guevara, Peter Czermak [1,2], Enrique Marcos [3], Andreas Vilcinskas [1,2,*] and F. Xavier Gomis-Rüth [*]

A

>sp|P81177|AURE_STAAU Zinc metalloproteinase aureolysin OS=Staphylococcus aureus OX=1280 GN=aur PE=1 SV=2

MRKFSRYAFTSMATVTLLSSLTPAALASDTNHKPATSDINFEITQKSDAVKALKELPKSE
NVKNHYQDYSVTDVKTDKKGFTHYTLQPSVDGVHAPDKEVKVHADKSGKVVLINGDTDAK
KVKPTNKVTLSKDEAADKAFNAVKIDKNKAKNLQDDVIKENKVEIDGDSNKYIYNIELIT
VTPEISHWKVKIDADTGAVVEKTNLVKE**AAATGTGKGVLGDTKDININSIDGGFSLEDLT**
HQGKLSAYNFNDQTGQATLITNEDENFVKDDQRAGVDANYYAKQTYDYYKNTFGRESYDN
HGSPIVSLTHVNHYGGQDNRNNAAWIGDKMIYGDGDGRTFTNLSGANDVVAHELTHGVTQ
ETANLEYKDQSGALNESFSDVFGYFVDDEDFLMGEDVYTPGKEGDALRSMSNPEQFGQPS
HMKDYVYTEKDNGGVHTNSGIPNKAAYNVIQAIGKSKSEQIYYRALTEYLTSNSNFKDCK
DALYQAAKDLYDEQTAEQVYEAWNEVGVE

Observed	Mr (expt)	Mr (calc)	ppm	Start	End	Miss	Ions	Peptide
958.49	957.48	957.46	28.8	457	463	0	41	K.SEQIYYR.A
980.45	979.44	979.43	9.9	283	289	0	---	K.QTYDYYK.N
983.44	982.43	982.42	16.0	329	337	0	---	K.MIYGDGDGR.T
1071.52	1070.52	1070.50	13.1	273	282	0	---	R.AGVDANYYAK.Q
1173.62	1172.62	1172.58	28.9	455	463	1	28	K.SKSEQIYYR.A
1555.75	1554.74	1554.71	18.7	283	294	1	---	K.QTYDYYKNTFGR.E
1704.78	1703.78	1703.74	22.1	408	422	0	---	R.SMSNPEQFGQPSHMK.D
1720.77	1719.76	1719.73	14.8	408	422	0	---	R.SMSNPEQFGQPSHMK.D + Oxidation (M)
2796.28	2795.27	2795.26	1.8	295	319	0	---	R.ESYDNHGSPIVSLTHVNHYGGQDNR.N
3346.54	3345.54	3345.54	-0.6	244	272	1	---	K.LSAYNFNDQTGQATLITNEDENFVKDDQR.

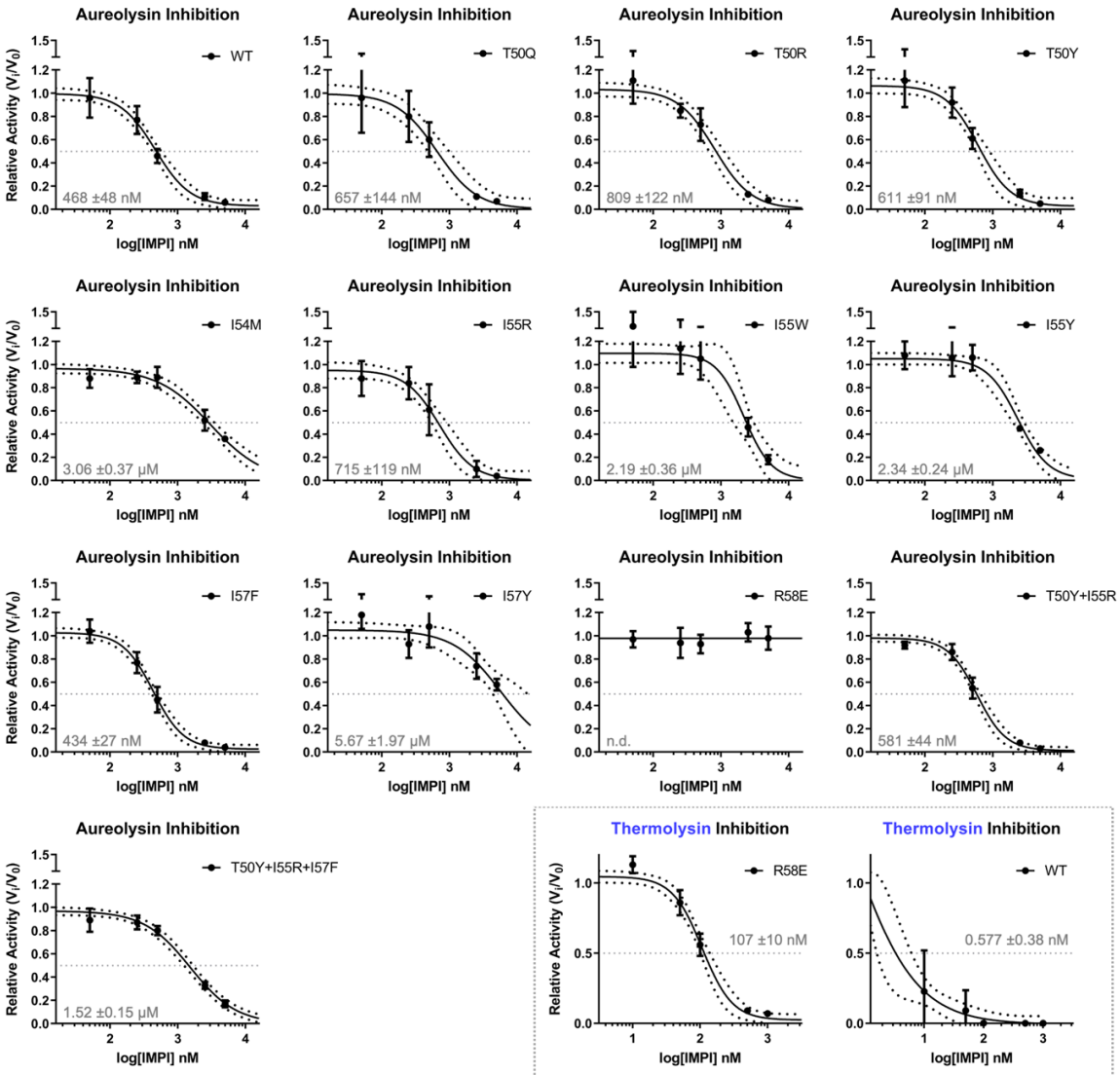
B

>sp|P00000|IMPI_Recomb GMS-IMPI OS=Galleria mellonella OX=7137 GN=IMPI PE=1 SV=2

GMSIVLICNGGHEYYECGGACDNVCADLHIQNK**TNCPIINIR**CNDK**CYCEDGYAR**DVNGK**CI**
PIKDCPKIRS

Observed	Mr (expt)	Mr (calc)	ppm	Start	End	Miss	Ions	Peptide
1100.58	1099.58	1099.58	-3.4	50	58	0	56	K.TNCPIINIR.C
1130.56	1129.55	1129.56	-12.7	77	85	1	45	K.CIPIKDCPK.I
1193.42	1192.41	1192.43	-12.7	63	71	0	54	K.CYCEDGYAR.D

Supplementary Figure 1. **Peptide mass fingerprinting after SDS-PAGE analysis.** The protein identification after in-gel digest with trypsin and carbamidomethylation of cysteines was performed using *MASCOT* (www.matrixscience.com) with a peptide and fragment mass tolerance of 100 ppm and 0.5 Da, respectively. **(A)** Mature aureolysin was identified with a *MASCAT* score of 200 and an E-value of 1.8E-12. Importantly, all 10 MS1 matches mapped to the catalytic domain, and fragmentation of the two most intense ions further increased the identification confidence. The catalytic domain of aureolysin and the identified peptides are highlighted in bold and red, respectively. Residue numbering refers to UniProt entry P81177. **(B)** Recombinant wt-IMPI was identified with a *MASCOT* Score of 200 and an e-value of 5.7E-19, with three MS2 spectra matching the mature IMPI sequence. Identified peptides are highlighted in bold red, and residue numbering refers to UniProt entry P82176.



Supplementary Figure 2. **Inhibitory activity of designed IMPI mutants against aureolysin.** The relative activity of aureolysin at 50 nM is shown as V_i/V_0 in the presence of 50, 250, 500, 2500, or 5000 nM of IMPI mutants and 20 µM FRET-4 as a substrate. The inhibitory activity of mutant $R^{58}E$ and wt-IMPI against thermolysin are shown for reference at the bottom right, framed with dots. All experiments were performed at least in triplicate, averages were plotted with *GRAPHPAD*, and error bars represent standard deviations. Data points were interpolated using a four-parameter sigmoidal curve-fit, with the 95%-confidence band shown as dotted curves flanking the best-fit line. The calculated IC_{50} values are shown with the respective standard deviation in grey, the grey horizontal dotted lines indicate 50% inhibition. Only $I^{57}Y$ (434 nM) showed better inhibition of aureolysin than wt-IMPI (468 nM), while $T^{50}Y+I^{55}R$ (581 nM), $T^{50}Y$ (611 nM), $T^{50}Q$ (657 nM), $I^{55}R$ (715 nM), and $T^{50}R$ (809 nM) were slightly worse. For all other mutants, the derived IC_{50} values were more than 3-fold higher than that of wt-IMPI, with mutant $R^{58}E$ showing no inhibition. To exclude issues with the $R^{58}E$ protein preparation, this mutant was tested also against thermolysin. It actually showed inhibition, though with an IC_{50} of 107 nM, i.e. ~200-fold higher than wt-IMPI.

Project 3

"Cryo-EM structures show the mechanistic basis of pan-peptidase inhibition by human α2-macroglobulin"

"*Cryo-EM structures show the mechanistic basis of pan-peptidase inhibition by human α2-macroglobulin"*

Daniel Luque[#], Theodoros Goulas[1,2,#], Carlos P. Mata[3,#,+], Soraia R. Mendes[1], F. Xavier Gomis-Rüth[1,*] and José R. Castón[4,*]

Spanish National Microbiology Centre, Institute of Health Carlos III, Madrid, Spain.
[1] Proteolysis Lab; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC), Barcelona Science Park, Barcelona, Catalonia, Spain.
[2] Department of Food Science and Nutrition; School of Agricultural Sciences; University of Thessaly, Karditsa, Greece.
[3] AstburyCentre for Structural Molecular Biology, School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK.
[4] Department of Structure of Macromolecules, Centro Nacional de Biotecnología (CNB-CSIC), Campus de Cantoblanco, Madrid, Spain.
[+] Present address: Spanish National Microbiology Centre, Institute of Health Carlos III, Madrid, Spain
[#] These authors contributed equally to this work
*Co-corresponding authors: F. Xavier Gomis-Rüth, fxgr@ibmb.csic.es, and José R. Castón, jrcaston@cnb.csic.es

Published in: The Proceedings of the National Academy of Sciences, PNAS, (2022)
Impact factor:11.205
Quartile: Q1

This publication provides experimental evidence of the, previously theoretical, mechanism of action of the major human blood inhibitor: α2M.

As part of this highly collaborative work developed by researchers of different groups and institutions, I prepared the α2M-plasmin complex and assisted vitrification of those samples in Cryo-EM grids. In more detail, I purified native α2M from frozen fresh plasma, and, as time was of essence, immediately after the last step of native α2M purification, I prepared the α2M-plasmin complex, which was promptly purified again to remove excess of protease, and ensuring sample homogeneity for structural studies.

Summary

Human α2-macrogllobulin (hα2M) is the major inhibitor of human plasma. hα2M is a large homotetrameric glycoprotein who's main but not only function is the inhibition of a myriad of proteases independent of their molecular weight, substrate specificities and catalytic types, and thus α2M is considered a pan-protease inhibitor. Inhibition of such a large repertoire of peptidases is achieved by a unique suicidal inhibitory mechanism described as "Venus' flytrap". Here we examined the hα2M inhibitory mechanism through analysis of eight cryo-electron microscopy (cryo-EM) structures of α2M purified from human plasma – five native forms and three protease-induced conformations. The native forms present an open and expanded conformation resulting from the organization of their monomeric subunits into two flexible modules. Binding of the prey peptidases and consequent proteolytic cleavage of the α2M bait regions triggers the rearrangement of the α2M structure into a closed conformation entrapping the peptidase molecule(s). Upon transition from the open to the closed conformation, the protease(s) gets simultaneously covalently bound to the inhibitor through reaction with a highly-reactive thioester bond, while the receptor-binding domain gets exposed at the inhibitor surface, allowing subsequent recognition of the inhibitory complex (but not the native conformation) by specific cellular receptors, and thus ultimately leading to internalization and clearance from circulation. Altogether, our results provided the long-awaited experimental evidence of the detailed hα2M inhibitory mechanism.

Cryo-EM structures shows the mechanistic basis of pan-peptidase inhibition by human α_2-macroglobulin

Daniel Luque[#], Theodoros Goulas[1,2,#], Carlos P. Mata[3,#,+], Soraia R. Mendes[1], F. Xavier Gomis-Rüth[1,*] and José R. Castón[4,*]

Spanish National Microbiology Centre, Institute of Health Carlos III, Madrid, Spain.
[1] Proteolysis Lab; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC), Barcelona Science Park, Barcelona, Catalonia, Spain.
[2] Department of Food Science and Nutrition; School of Agricultural Sciences; University of Thessaly, Karditsa, Greece.
[3] AstburyCentre for Structural Molecular Biology, School of Molecular and Cellular Biology, Faculty of Biological Sciences, University of Leeds, Leeds, UK.
[4] Department of Structure of Macromolecules, Centro Nacional de Biotecnología (CNB-CSIC), Campus de Cantoblanco, Madrid, Spain.
[+] Present address: Spanish National Microbiology Centre, Institute of Health Carlos III, Madrid, Spain
[#] These authors contributed equally to this work
*Co-corresponding authors: F. Xavier Gomis-Rüth, fxgr@ibmb.csic.es, and José R. Castón, jrcaston@cnb.csic.es

Keywords
α_2-macroglobulin, proteinase, blood proteostasis, cryo-EM, multifunctional complex, conformational states

---

Human plasma α_2-macroglobulin (hα_2M) is a multidomain protein with a plethora of essential functions, including transport of signaling molecules and endopeptidase inhibition in innate immunity. Here we dissected the molecular mechanism of the inhibitory function of the ~720 kDa hα_2M tetramer through eight cryo-EM structures of complexes from human plasma. In the native complex, the hα_2M subunits are organized in two flexible modules with an expanded conformation, which encloses a highly porous cavity in which the proteolytic activity of circulating plasma proteins is tested. Cleavage of bait regions exposed inside the cavity triggers rearrangement to a compact conformation, which closes openings and entraps the prey proteinase. After the expanded-to-compact transition, which occurs independently in the subunits, the reactive thioester bond triggers covalent linking of the proteinase, and the receptor-binding domain is exposed on the tetramer surface for receptor-mediated clearance from circulation. These results depict the molecular mechanism of a unique suicidal inhibitory trap.

SIGNIFICANCE STATEMENT

Human plasma α_2-macroglobulin (hα_2M) is a ~720 kDa homotetrameric particle with pan-peptidase inhibitory functions that transits between an open native conformation and a closed induced state, in which endopeptidases are trapped upon cleavage of an accessible bait region. We determined the molecular mechanism of this function through eight cryo-EM structures, which revealed that the hα_2M subunits are organized in two flexible modules that undergo independent expanded-to-compact transitions. In the induced state, a reactive thioester bond triggers covalent linking of the proteinase, and a receptor-binding domain is exposed on the tetramer surface for binding to its specific cellular receptor for internalization and clearance from circulation. These results elucidate the long-awaited molecular mechanism of a historical suicidal inhibitory trap.

The α_2-macroglobulins are large multi-domain proteins found in animals and selected colonizing bacteria [1-8]. The best characterized is human α_2-macroglobulin (hα_2M], a 1451-residue protein built of 11 domains (for hα_2M domain nomenclature, see Supplementary Fig.1a), which is produced by several cell types including macrophages, astrocytes, and hepatocytes. Four protomers associate to a ~720-kDa polyglycosylated dimer of disulfide-linked homodimers, (hα_2M)$_4$ (Supplementary Fig.1b), which is the largest non-immunoglobulin protein in human plasma and constitutes 2-4% of its total protein content [1, 9]. Its multiple molecular functions include endopeptidase inhibition, as well as sequestration and transport of growth factors, cytokines, and hormones [10, 11]. It is also an acute-phase reactant in rodents and a chaperone that binds misfolded or inactivated proteins, and has many more moonlighting functions such as transglutamination and zinc/copper binding [4, 8, 12].

These disparate functions explain the universal physiological significance of hα_2M: it is part of the innate immune response against pathogens [13, 14] and a major hemostatic regulator of the cardiovascular system through its anticoagulant, procoagulant, and antifibrinolytic activities [4]. It is an early marker of cardiac hypertrophy, as well as a potential diagnostic marker for myocardial infarct and for HIV patients with cardiac pathologies [5]. In amyloidoses, it is prophylactic, as it binds the major component of β-amyloid deposits, the Aβ peptide, and mitigates its neurotoxicity and fibrillogenic capacity [8]. In addition, it has anti-inflammatory, signaling, and apoptotic properties, is engaged in growth and tissue remodeling, and protects joint cartilage [15, 16]. Its deregulation contributes to most major human diseases including Alzheimer's disease, AIDS, inflammatory diseases, diabetes, arthritis, tumor growth and progression, and cardiovascular conditions [8, 15, 17]. No total hα_2M deficiency has been described in humans, which supports the idea that its absence is embryonically lethal [4, 18-20].

The best-characterized molecular function of hα_2M is its indiscriminate inhibitory action on endopeptidases irrespective of catalytic class, whether endogenous or exogenous. This hallmarks it as a unique pan-endopeptidase inhibitor [1, 4, 5, 7, 8, 21]. In the cardiovascular system, it inhibits thrombin, factor Xa, activated protein C, plasma kallikrein and kallikrein-related peptidases, and plasmin [4, 17]. As part of the humoral defense barrier, it inhibits bacterial and viral proteolytic virulence factors such as pseudolysin, HIV-1 proteinase, clostripain, and vibriolysin, as well as snake-venom proteinases [22, 23]. Finally, by inhibiting neutrophil elastase, matrix metalloproteinases, and ADAM/adamalysin metallopeptidases, it participates in inflammatory processes and tissue turnover [15, 16].

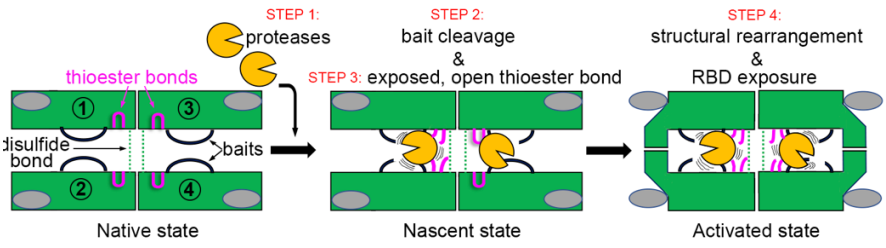


Figure 1. Overview of hα_2M action based on biochemical analysis – In native (hα_2M)$_4$, the four intact bait regions (black) are exposed in the enclosed cavity (prey chamber), while the thioester bonds (pink) and receptor-binding domains (RBD, grey oval) are hidden. Monomer pairs 1 and 2, as well as 3 and 4, are disulfide-linked protomers (dashed green line); pairs 1 and 3, as well as 2 and 4, are vicinal protomers; and pairs 1 and 4, as well as 2 and 3, are diagonally opposite protomers. Each protomer thus has a vicinal, a disulfide-linked, and an opposite protomer. The mechanism of action is illustrated in four steps that result in three states as described in the text.

Whereas the vast majority of peptidase inhibitors act through reversible "lock-and-key" mechanisms that sterically block the active-site cleft of target enzymes [24, 25], hα_2M operates according to a unique suicidal Venus fly-trap [7, 26] or trap-hypothesis mechanism [1], by which prey peptidases diffuse into an unreacted or native (hα_2M)$_4$ tetramer (Fig.1, step 1). Once inside, peptidases access and cleave a flexible, multitarget bait region (Fig. 1, step 2) within an exposed bait-region domain (BRD, see Supplementary Fig.1). Cleavage leads to a nascent state of the inhibitor, in which a buried reactive β-cysteinyl-γ-glutamyl thioester bond becomes exposed on a thioester domain (TED). The bond is attacked by surface lysine amines of the prey peptidase, which thus becomes covalently bound to the inhibitor through an ε-lysyl-γ-glutamyl crosslink (Fig. 1, step 3). Bait-region cleavage also triggers a large, irreversible structural rearrangement of the tetramer, leading to a reacted, induced, or activated state, which engulfs the peptidase without disturbing its active site, similarly to insect capture by the Venus fly-trap plant (Fig.1, step 4). A large reorganization of (hα_2M)$_4$ can also be observed by treatment of the native species with nucleophilic chemicals such as methylamine (MA), which yield an activated species with open thioester bonds but intact bait regions, unable to inhibit peptidases [27-30]. Native and activated states can be distinguished because they have different sedimentation coefficients and mobilities in native gel electrophoresis [13].

Once within the closed trap, the peptidase no longer cleaves large protein substrates, but is still accessible to small inhibitors and substrates through openings in the tetrameric particle. Moreover, the four receptor-binding sequences of the receptor-binding domains (RBD), which are cryptic in the native and nascent species [31], become

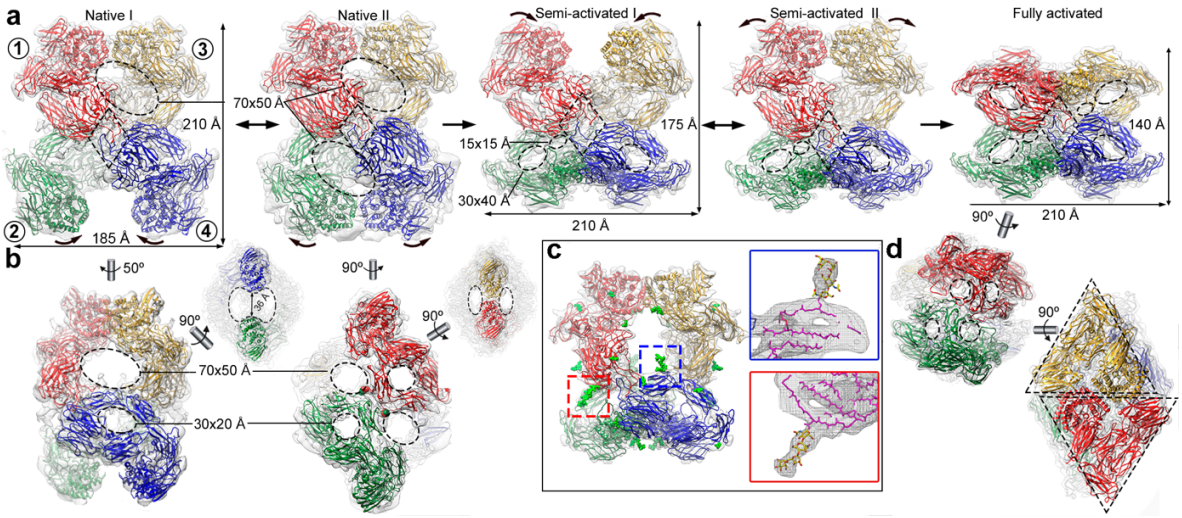


Figure 2. Cryo-EM structures of (hα_2M)$_4$ functional states – (**a**) Five functional states of (hα_2M)$_4$ from human plasma are isolated from cryo-EM analysis, termed native I (left panel) and II (center-left), semi-activated I (center) and II (center-right), and activated (right). Protomer nomenclature [1 (red), 2 (green), 3 (yellow), and 4 (blue)] as described in Fig.1. The red protomer has a disulfide-linked (green), a vicinal (yellow), and an opposite neighbor (blue). Native I and II states have protomers with expanded conformation and are in an equilibrium, at which vicinal dimers are in a distal (native I) or proximal (native II) position (curved arrows). After proteolytic activation, the native state becomes semi-activated states I and II, and one vicinal dimer is built of protomers in compact conformation. Semi-activated states I and II, which correspond to the nascent state described in the literature [46], evolve to the activated state, shown in "H-view", in which all protomers are compact. Openings are indicated as dashed ovals. Tethering loops of opposite protomers (1 and 4) are framed in a dashed rhombus. (**b**) Additional views of the native I (left) and II (right) complexes of panel a. The latter highlights the disulfide-linked residues between protomers 1 and 2 (red and green spheres). (**c**) The semi-activated II complex highlights the N-linked glycosylation sites as green sticks (left). Magnified view of the red and blue boxes that correspond to the same glycan bound to MG4 Asn396 in the compact protomer 4 (blue) and in the expanded protomer 1 (red). (**d**) The activated state is shown in end-view (top) and X-view (bottom), in which the triangular prism profile for each protomer is framed.

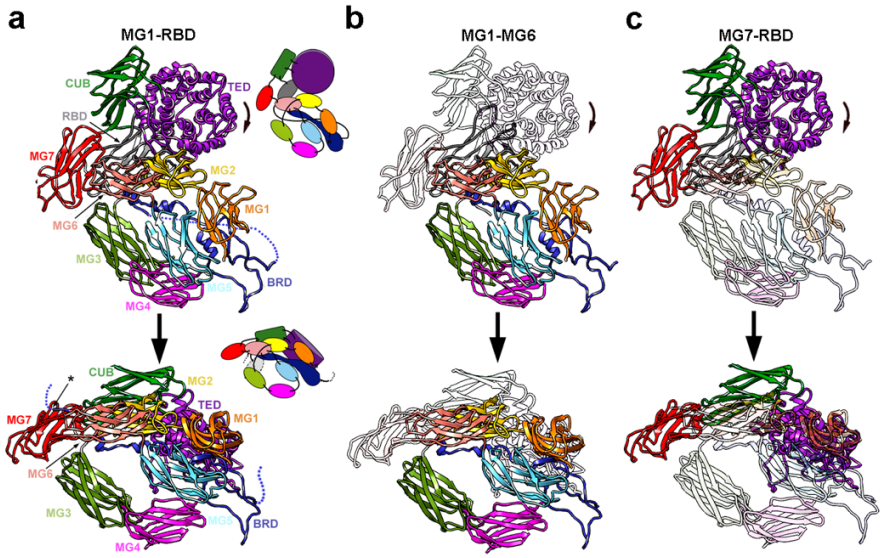


Figure 3. Functional states of $(h\alpha_2M)_4$ are built of expanded and compact protomers – (**a**) Spatial domain organization of the expanded (top) and compact (bottom) protomer conformations in front view (as in Fig.2a). The BRD (dark blue) contains the flexible, intact (top) and broken (bottom) bait region (dashed line). Insets, diagrams of the approximate domain organization of the conformers. (**b**, **c**) Regions equivalent to a, highlighting the N-terminal module (domains MG1-MG6) (**b**) and the C-terminal module (domains MG7-RBD) (**c**). MG, macroglobulin-like domains; BRD, bait region domain; CUB, domain found in C1r/C1s, urchin embryonic growth factor, and bone morphogenetic protein 1; TED, thioester domain; RBD, receptor binding domain.

exposed on the tetrameric particle surface [32] and are recognized by cell-surface receptors such as the low-density lipoprotein receptor-related protein (LRP1) for receptor-mediated endocytosis (Fig. 1, step 4) [33]. Inside the cell, the peptidase:inhibitor complex is cleared in the lysosomes within minutes of complex formation [34].

This sequence of events has been established through painstaking biochemical analyses for decades, but its molecular determinants remain unknown. Attempts have been undertaken to study the structure of tetrameric $h\alpha_2M$ and mammalian orthologs to provide the molecular determinants of this inhibitory mechanism. Nonetheless, owing to the large size and intrinsic flexibility of $h\alpha_2M$, the lack of suitable recombinant expression systems for large-scale production of homogeneous functional protein, and the structural heterogeneity of samples purified from natural sources, only two small domains (the RBD and the macroglobulin-like domain MG2) have so far been described by X-ray crystallography [35-39]. For full-length assemblies, 20–40 Å resolution maps derived from electron microscopy (EM) of negative-stained or vitrified samples, and a crystallographic map of a hypothetically MA-activated $(h\alpha_2M)_4$ to 10 Å resolution, have shown morphological variations of the native and activated states of the $h\alpha_2M$ tetramer, with no details at the domain level [29, 40-43]. Moreover, the 4.3-Å resolution crystal structure of activated $(h\alpha_2M)_4$ showed conserved structural features with the proteolytically activated C3b complement component (derived from the native C3 factor) [26]. The most recent structural analyses of native $h\alpha_2M$ are limited to homology models calculated from the native C3 complement component and docking in low-resolution EM maps [26, 44].

Thus, this study addresses three questions relevant to $h\alpha_2M$ biochemistry: (i) the structure of the native $(h\alpha_2M)_4$ complex when poised to trap plasma endopeptidases and that of the peptidase-activated inhibitor; (ii) the conformational changes and the intermediates on the path between the native and activated states; and (iii) how large proteases are entrapped in a cavity with

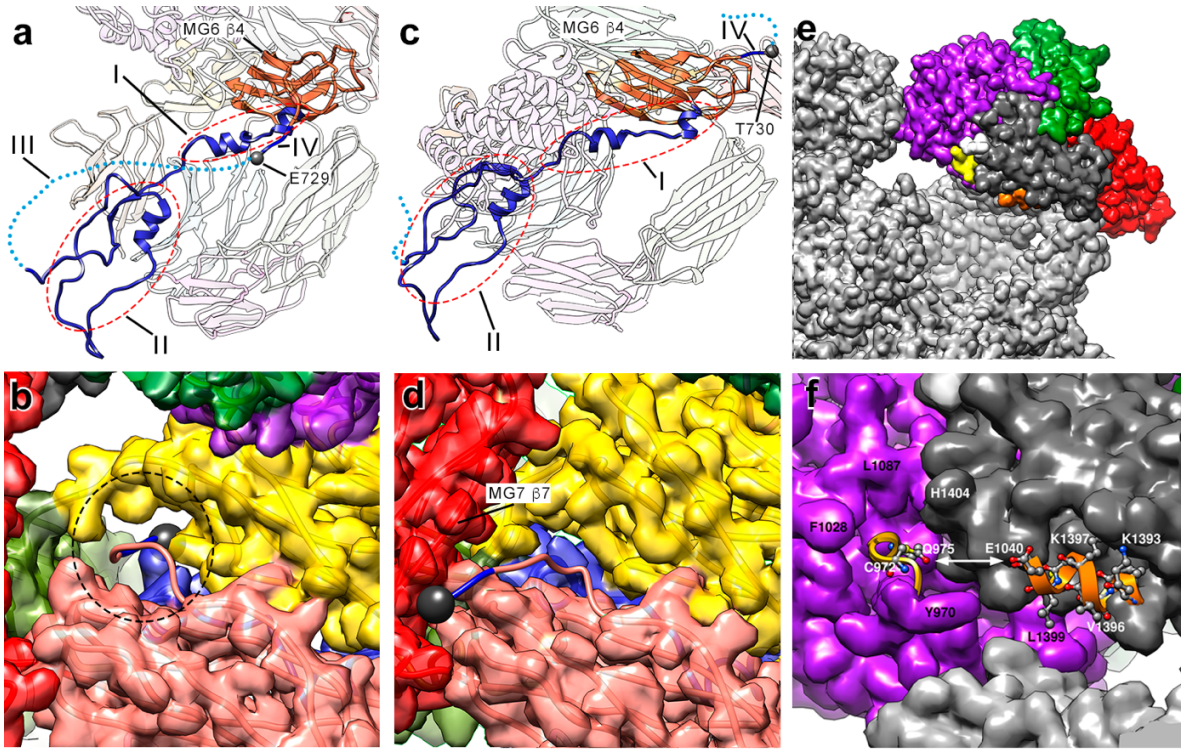


Figure 4. The major players in the conformational shift: BRD, TED and RBD – (**a**) Unprocessed BRD (blue) seen from inside the prey chamber in the expanded conformation of the native I state (upper vicinal protomers 1 or 3). Regions I, II, III (dashed line) and IV are indicated. MG6 β4 (orange) is shown, bound to region IV (Glu729–Arg732). Glu729 is highlighted (grey sphere). (**b**) Close up of the unprocessed segment Glu729–Arg732 passing through the opening (dashed oval) framed by MG6 (pink), MG2 (yellow), MG3 (green) and the MG2-MG3 connecting loop (Glu729, grey sphere). (**c**) The proteolytically processed BRD seen from inside the prey chamber in the compact conformation of the activated state in any of the protomers (1-4), in which region III (dashed line) is discontinuous (similar view as panel a). The last visible residue is Thr730 (grey sphere). (**d**) Close up of the processed segment Thr730–Arg732 (blue), which interacts with the seventh β-strand of MG7 on the outer surface (Thr730, grey sphere). The opening seen in (b) is occluded. (**e**) In the expanded conformation of the native I state, the TED thioester bond (yellow) and α2 helix of the RBD (orange) are 16 Å apart, and are buried in the structure. (**f**) Close up of the region shown in (e). The thioester bond is in a local cavity surrounded by the large hydrophobic side chains of Tyr970, Phe1028, and Leu1087 (from TED), and His1404 (from RBD). Helix α2 from the RBD is partially hidden by segment Ser1428–Thr1432 from the domain's seventh β-strand.

insufficient volume *a priori*. We used authentic human protein to determine eight cryo-EM structures that represent functional states of $(h\alpha_2M)_4$, unbound and in complex with physiologically relevant endopeptidases. The resulting structures show striking conformational rearrangement and provide detailed insight into the molecular mechanism of a unique sequential inhibitory mechanism.

RESULTS AND DISCUSSION

Cryo-EM structure analysis of the $(h\alpha_2M)_4$ complex – Authentic $(h\alpha_2M)_4$ from human plasma was vitrified and imaged by cryo-EM (Supplementary Fig.2a). A total of ~1,625,000 particles collected on 300-keV FEI Titan Krios microscopes were automatically picked. Owing to the intrinsic flexibility and heterogeneity of the structures, they were subjected to several rounds of exhaustive 2D and 3D classification (Supplementary Fig.2b,c). These resulted in five distinct cryo-EM structures at different states of reaction, to resolutions spanning 4.5–7.3 Å as estimated by the values at which the Fourier shell correlation (FSC) coefficient equals 0.143 (Supplementary Fig.3a–e, Supplementary Table1).

To obtain a homogeneous population of fully activated $(h\alpha_2M)_4$ as a control, a purified sample of mostly native protein was incubated with trypsin and its structure

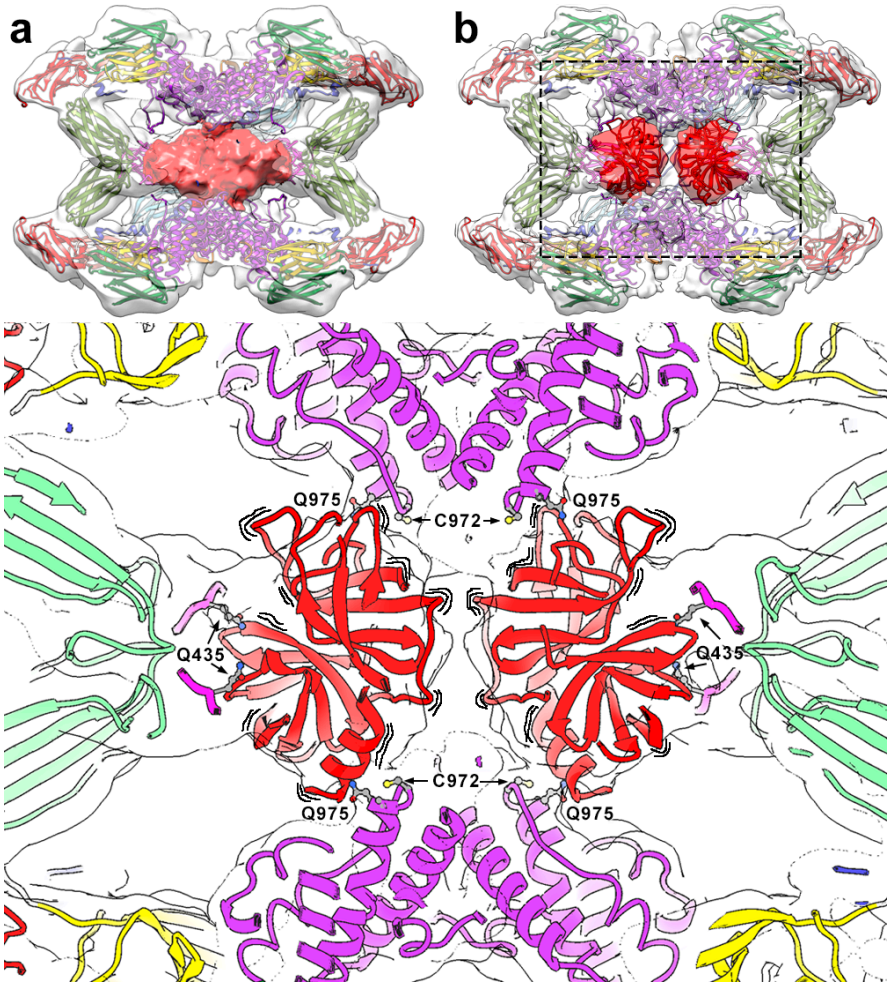


Figure 5. The symmetrically activated state – (**a**) Intrinsically activated (hα_2M)$_4$ from plasma. The front half of the map (transparent surface) was removed to visualize the heterogeneous proteinase density (red). (**b**) Trypsin-activated (hα_2M)$_4$ in a similar view as in (a), with two trypsin molecules tentatively docked into the density (dark red), which is clearly resolved in two separate volumes (clear red). However, the exact orientation of the molecules cannot be determined due to the lowish resolution of the map. (**c**) Close-up of the boxed region in (b). To highlight that the caged trypsin molecules are dynamically oriented, wavy lines are shown around the corresponding red ribbons. Location of major functional residues are indicated in the activated complex. Trypsin is fixed to the TED thioester bonds formed between Cys972 and Gln975 from vicinal protomers. Residues Gln435 of each MG4 moiety are indicated. Color code: MG2, yellow; MG3, green; MG4, pink; TED, purple; trypsin, red.

determined to 3.6 Å resolution (Supplementary Figs.3f and 4). This structure enabled unambiguous assignment of the polypeptide chain of the whole particle, which gives rise to a C2 tetramer featuring a compact cage with several openings of variable size and a large inner cavity, the prey chamber.

Based on protomer conformations, either expanded or compact, on their relative arrangement within the tetramers, and on the volume and occupancy of the prey chambers in the distinct structures (i.e., the presence of entrapped proteinases in its interior chamber), we identified five major states of (hα_2M)$_4$, which we termed native I and II, semi-activated I and II, and fully activated, which corresponded to distinct reaction intermediates (Fig.2a). With maximal dimensions of 210x185x150 Å, the native tetramer is substantially larger than the activated form, which spans 140x210x140 Å (Fig. 2a) and provides an explanation for the higher electrophoretic mobility of the latter [45]. The native states have four large and four small openings of 70x50 Å and 30x20 Å, respectively, and enclose a prey chamber of ~600 nm^3, whereas the activated state has 12

small openings of 30x40 Å and encloses a ~300-nm^3 prey chamber (Fig.2a,b,d).

Each (hα$_2$M)$_4$ protomer has a vicinal, a disulfide-linked, and an opposite neighbor (Fig.2a, left). In the native and semi-activated states, one vicinal dimer was solved to much better resolution than the other (Supplementary Fig.3). For the native I and II states, resolution of the upper pair of vicinal protomers was sufficient to determine unambiguously their subunit organization (Fig.2a, red and yellow). By contrast, the intrinsic flexibility of the lower vicinal protomers (~11–22 Å resolution) required development of a strategy that combined rigid and flexible domain refinement (Fig.2a, green and blue). The semi-activated I state behaved similarly to native states, with the upper pair more flexible than the lower pair (8–25 Å vs. 3.5–12 Å resolution). Finally, the semi-activated II and activated states showed density with more features, which allowed accurate modeling of the four polypeptide chains (Supplementary Fig.5; see Methods for details). Each hα$_2$M subunit has eight surface-located, protruding *N*-linked glycans at positions 55, 70, 247, 396, 410, 869, 991, and 1424 (Fig.2c, indicated for semi-activated state II, green sphere model; see UniProt entry P01023 for sequence numbering). Structural alignment of compact or expanded protomers in all states showed *rmsd* values <1.4 Å, indicating high structural similarity in the respective conformations.

The native (I and II), semi-activated (I and II), and activated states were captured at a ~1:2:1 ratio, although in a specific preparation (preparation P2), more than half the total particles were initially classified as native (Supplementary Fig.2). These data underpin the existence of a substantial population (25–42%) of intrinsically activated (hα$_2$M)$_4$ in preparations assumed to be mainly native, following long-established standard biochemical purification procedures [27].

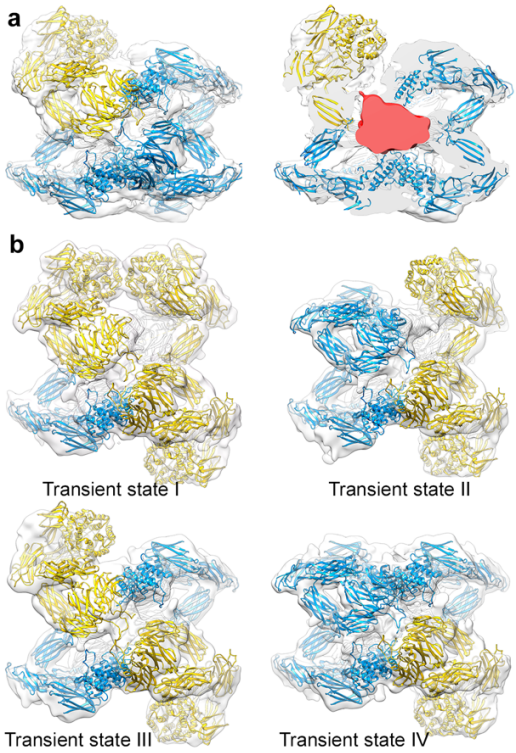


Figure 6. The subunit-mediated activated state – (**a**) State I of plasmin-activated (hα$_2$M)$_4$ with three compact (blue) and an expanded (yellow) subunit (left). After removing the front half of the map, an asymmetric, large density corresponding to plasmin becomes evident (red) (right). (**b**) Analysis by symmetry expansion of native (hα$_2$M)$_4$ reveals arrangements with one (top, left), two (top, right and bottom, left) and three (bottom, right) compact subunits (blue).

The functional tetrameric states are built of expanded and compact protomer conformations – Each hα$_2$M subunit structure consists of 11 or 10 domains for the expanded and compact conformations, respectively, as the C-terminal RBD is flexible and thus not assignable in the latter (Fig.3a; Supplementary Fig.1). The first seven domains (MG1-MG7) are concatenated ~110-residue β-sandwiches with a three- and a four-strand antiparallel β-sheet (Fig.3a). Domains MG1-MG6 form an N-terminal module, which is similarly arranged in all structures as a 1.5-turn right-handed superhelix that encircles an ellipsoidal opening of 30x20 Å (Fig.2b, dashed ellipses). The 127-residue BRD is an extended, flexible domain inserted between the fourth and fifth β-strands of MG6 (Supplementary Fig.1a); the domains downstream of MG6 form a C-terminal module that adopts two different

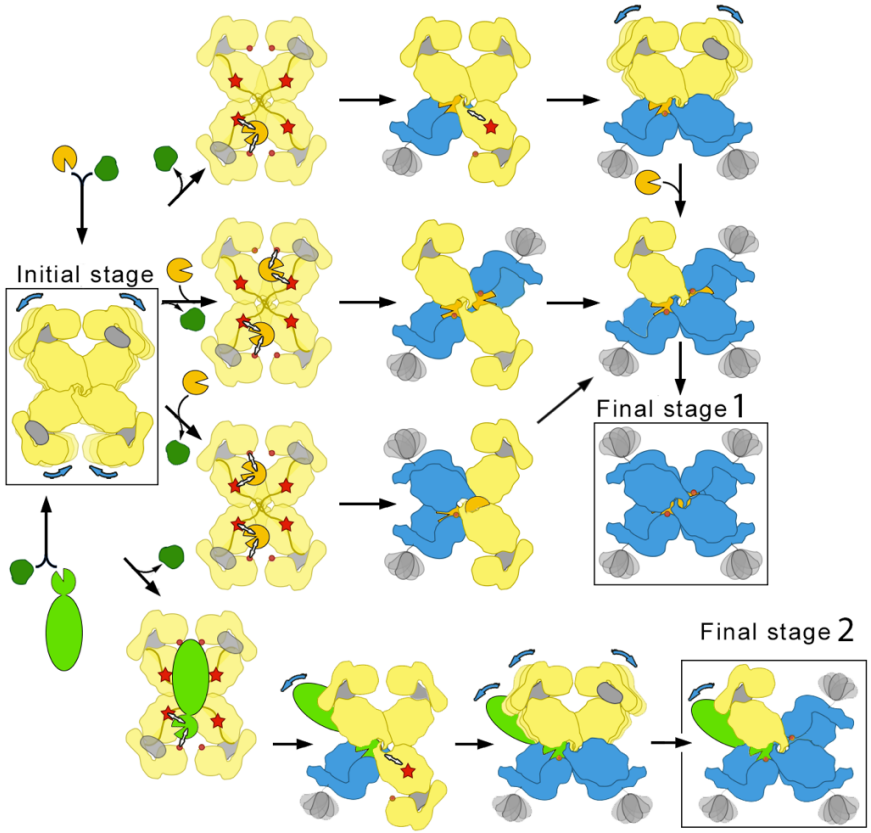


Figure 7. Mechanism of irreversible protease inhibition by (hα$_2$M)$_4$ – In the initial stage, native (hα$_2$M)$_4$ (four yellow expanded subunits) is a molecular sieve of plasma proteins (dark green, light green, and gold). Only peptidases (gold and light green) are retained in the internal cavity after processing of the bait region (red star), which can include two small peptidases (gold) or one large peptidase (light green). The thioester bond (red dot) becomes accessible to covalently fix the prey peptidase, which triggers the expanded-to-compact transition of the protomer (yellow-to-blue) and exposure of the flexible RBD (grey) on the surface, ready for cognate receptor binding (compact blue subunits). As the prey remains active within the trap, it can still cleave a second, third, and fourth protomer to yield the distinct intermediates and ultimately, the final fully compact stage 1 for one or two small peptidases. In the case of large peptidases (bottom row), peptidase regions protruding from the prey chamber prevent complete transition, giving rise to an asymmetric final stage 2 that corresponds to the penultimate step of the pathway for small peptidases.

conformations: expanded or compact. MG7 acts as a hinge domain that connects MG1-MG6 with the 116-residue CUB domain (acronym of complement protein subcomponents C1r/C1s, urchin embryonic growth factor, and bone morphogenetic protein 1), which consists of two four-stranded antiparallel β-sheets. The 315-residue helical TED domain is inserted between the third and fourth β-strands of CUB and includes the scissile thioester bond. TED adopts a thick disk-like structure that consists of six concentric β-hairpins arranged as a six-fold α-propeller around a central shaft (Fig. 3a). The thioester bond is formed between the cysteine side-chain sulfur and the glutamine side-chain carbonyl of the conserved motif Cys972-Gly-Glu-Gln975. Finally, the C-terminal domain of this module is the RBD, a variant of the MG fold with an extra α-β-α unit.

In all five functional states identified in the native preparation, the hα$_2$M monomers adopt either a compact or an expanded conformation for each pair of vicinal dimers. In native states I and II, all subunits are expanded (Fig. 2a). In the native I state, the TED domains of vicinal protomers (within the top dimer) contact each other through their respective fifth helices (Fig.2a, left; subunits 1 and 3). By contrast, the protomers are dynamic in the bottom dimer (Fig.2a, left;

subunits 2 and 4) and swing around the interface between subunits 1 and 2 and subunits 3 and 4. These subunits encompass the inter-protomeric disulfide bond (Fig.2b, indicated in the native II state, red and green spheres) as well as BRD-mediated contacts between opposite monomers of subunits 1 and 4 and subunits 3 and 2 (Fig.2a, dashed rhombuses). As a result, the fifth helices of the bottom TED domains can either be separated by 36 Å (Fig.2a,b, left; green and blue protomers) or in contact, as found in the top domains (Fig.2a, center-left; green and blue protomers). Both expanded protomers, in a distal or proximal position, share an almost identical conformation (*rmsd* < 0.7 Å). This yields the native II state, which is quasi-symmetric for the top and bottom dimers as the bottom dimer is resolved at lower resolution, indicating it is more flexible or dynamic than the top dimer. In semi-activated states I and II, which correspond to the nascent state [46] (Fig.1), the bottom vicinal subunits adopt a compact conformation (Fig.2a, center and center-right; green and blue protomers), while the top vicinal subunits are dynamic and very similar to the native states. In the fully activated state, all subunits show a compact conformation (Fig.2a, right, and Supplemental Movie1).

Whereas the architecture of the N-terminal module is equivalent in the expanded and compact conformations and merely undergoes a 22º-rotation around MG4 (Fig.3b), the C-terminal modules are coordinately shifted, with additional 45º- and 55º-rotations for CUB and TED, respectively, around the hinge-domain MG7 (Fig.3c, Supplementary Movie2). Following this conformational change, all domains except MG3 and MG4 are arranged as a triangular prism in the compact conformation, the hallmark structural motif of the activated state (Fig.2d, dashed triangles). Vicinal TEDs interact through their fifth and third α-helix in the expanded and compact conformation, respectively. TED also contacts MG4 of the respective disulfide-linked subunit, as well as MG1s and BRDs of the respective vicinal subunit.

The conformational rearrangement between the expanded and compact conformations implies notable changes in the contact points within the top and bottom dimers of vicinal protomers (Supplementary Table2). Whereas the internal cavity in the native tetramer is framed by MG1–MG5, BRD, TED and a small part of RBD, in the activated tetramer only MG1–MG5 and TED face the prey chamber as the exposed surfaces of MG1, MG2 and MG5 are substantially reduced. All domains have a very similar fold in both conformations, except for TED. The four MG4 domains in the equatorial plane of the tetramer are nearly invariant in all functional states despite the massive structural changes that underlie the transition. Together with the BRD tethering loops (see next section), these domains form a constant belt responsible for the structural integrity of the tetramer in either state (Supplementary Fig.6).

The major players: BRD, TED and RBD – The BRD spans 75 Å and contains little regular secondary structure. It consists of four regions in the expanded conformation (Fig.4a, I–IV). Region I (Gln602–Leu627) is an N-terminal extended segment that is linked to the third β-strand of MG6 and contains two short α-helices. Region II (Thr628–Gln694) is a compact region that includes the tethering loop (Ile643–Cys689) and interacts with the symmetric region of the opposite protomer. Region III (Pro690–Thr728) is the 39-residue bait segment that is disordered and invisible in our maps. Finally, region IV (Glu729–Arg732) is the C-terminal segment connected with the fourth β-strand of MG6 on the outer surface of the protomer through one of the openings in the expanded conformation. This opening is framed by MG6, MG2, MG3, and the loop connecting MG2 with MG3, and it allows passage of an intact polypeptide chain (Fig.4b, dashed circle; Supplementary Fig.5). In the compact conformation (Fig.4c), the

beginning of the C-terminal segment at Thr730 is on the outer surface and interacts with the seventh strand of MG7. A small rearrangement in this region causes MG7 to fold over MG6, which almost occludes the opening (Fig.4d; Supplementary Fig.6). The BRD-MG6 connecting segment thus regulates the expanded-to-compact transition by acting as a trigger that must be cleaved for displacement of the new N terminus to the exterior surface. This is consistent with the assumption that MA-activated $(h\alpha_2M)_4$, whose bait region is intact, adopts a structure distinct from the peptidase-activated state [32]. These results further indicate that the reported crystal structure of induced $(h\alpha_2M)_4$ (Supplementary Table3), originally thought to be a non-cleaved MA-induced variant [26], cannot contain an intact bait region but is a cleaved, proteinase-induced species. Its overall conformation is equivalent to the structure of the fully peptidase-activated state (see below and Fig.2a, right). At the C terminus, protomers in expanded conformation, both in activated and semi-activated states, have a well-defined RBD, surrounded by MG7, CUB, and TED (Fig.4e, dark grey; Supplementary Fig.5). The second α-helix of RBD, which includes the receptor-binding region [47], points towards the prey chamber and is occluded (Fig.4e,f, orange). Following activation by cleavage of the bait region, the substantial rearrangement of MG7-CUB-TED-RBD disrupts the interaction among these domains. In the activated $(h\alpha_2M)_4$ structures, this causes the RBD to project away from the outer surface and become flexible and thus disordered, except for a small, blurred density for its first three or four residues (Supplementary Fig.7). The receptor-binding region thus becomes accessible for interaction with LRP1. In previous structural studies, the RBD was visible in the crystal structure of activated $(h\alpha_2M)_4$ for only one of the four subunits, probably due to tetramer-tetramer interactions within the crystal, and obviously in a physiologically irrelevant position [26].

The TED thioester bond is 16 Å from the second RBD α-helix in the expanded conformation of the native I state (Fig.4e,f, yellow). Small nucleophiles such as MA can access and open the thioester bond without BRD processing. This would be sensed by the RBD, resulting in exposure of its receptor binding sequence to trigger removal of this inactivated hα2M complex. Alternatively, these two critical regions might be adjacent in the highly porous structure of the native complex to remain simultaneously inaccessible to the external environment. Access to TED is limited in the expanded conformation, as it is found in a cleft surrounded by bulky hydrophobic side chains (Phe1024, Leu1087, Tyr970, and His 1404) (Fig.4f). Moreover, the glycan bound to Asn1424 is nearby (Fig.4e, white), which might prevent thioester-bond opening by surface amines of circulatory proteins and other substances. In contrast, in the compact conformation, thioester bonds are accessible and face the prey chamber, and are thus prepared for reaction with surface lysines of the prey (see below).

We did not identify a native state in which both vicinal dimers displayed simultaneously non-interacting TED domains (Fig.2a, left; green and blue protomers), which suggests that this intermediate is unstable or very short-lived and thus cannot be isolated by cryo-EM. To communicate the separated, interacting TED positions, two small contact areas between vicinal dimers are critical: the opposite BRD tethering loops and the MG3 and MG4 domains. They are symmetrically crosslinked through intermolecular disulfide bonds (Cys278 of MG3 with Cys431 of MG4; see Supplementary Figs.1 and 5). The structural integrity of the native tetramer resides only on the tethering loop-mediated, non-covalent interaction. This TED rigid-body rearrangement might capture non-specific circulatory proteins, which are conveyed to the prey chamber. Most of them will not interact with the $h\alpha_2M$ subunits and will be released directly through the 70x50 Å

openings, which in turn might also be the direct entrance of many soluble proteins. By contrast, those with endopeptidolytic activity will cleave the bait region and trigger conformational rearrangement from the expanded to the compact conformation, which halves the volume of the prey chamber and restricts the openings to maximumally 30x40 Å, thus preventing prey from escaping. This mechanism facilitates scanning of many plasma proteins, and only endopeptidases that must be quickly withdrawn from the circulation would be "swallowed".

The symmetric activated state – Intrinsically activated $(h\alpha_2M)_4$ was resolved as a single complex in which the heterogeneous proteinases trapped were resolved as a featureless density that occupied the prey chamber (Fig.5a, red surface). Since the purified samples contained 2555% native $(h\alpha_2M)_4$, we analyzed a sample activated with bovine trypsin *in vitro* (Supplementary Fig.4). The tetramer in the resulting compact conformation was refined to 3.6 Å resolution (Supplementary Fig.3f), and the density corresponding to the trapped proteinases was resolved as two independent volumes (Fig.5b, red surface). This result indicates that each disulfide-linked dimer participates in the binding of one trypsin molecule, in accordance with an inhibition stoichiometry of two proteinase molecules per inhibitor tetramer [13]. Although the entrapped peptidases are probably not static inside the cavity, they could be tentatively docked (223 residues; PDB 1MTS) into these two densities (Fig.5b,c, red ribbons). The density corresponding to the two peptidase moieties, in both intrinsically and trypsin-activated $(h\alpha_2M)_4$, suggest intimate contact with the two symmetric MG4 loops, in particular with Gln435, and with Gln975 of the vicinal subunit thioester bonds for covalent binding (Fig.5c).

The subunit-mediated activated state – $(h\alpha_2M)_4$ was also activated *in vitro* with plasmin, a 791-residue protease that cannot be accommodated in the internal cavity of the activated tetramer, but is nevertheless inhibited efficiently by $(h\alpha_2M)_4$ following cleavage after Arg719 [48]. Whereas trypsin-incubated $(h\alpha_2M)_4$ adopted an activated state in which the four subunits were in the same compact conformation, plasmin-incubated $(h\alpha_2M)_4$ was resolved in two activated states, with four (plasmin-activated I state) and three (plasmin-activated II state) compact subunits (Supplementary Figs.3g,h and 8), which account for 79 and 21% of the activated complexes, respectively (Fig.6a). The density in the prey chamber attributable to a bound prey was much larger in the complex with three compact subunits (Fig 6a, right), which suggests the absence of the large peptidase in the particles with four compact protomers. We thus hypothesize that acquisition of the compact conformation for the fourth subunit is prevented by steric hindrance through parts of the peptidase protruding from the inhibitor tetramer, even though the bait region is cleaved. Plasmin-activated II state therefore provides a structural explanation for the inhibitory potential of $(h\alpha_2M)_4$ for endopeptidases larger than the prey chamber.

We sought to determine whether both subunits within a vicinal pair must simultaneously undergo the transition from the expanded to the compact conformation. We thus reanalyzed native $(h\alpha_2M)_4$ complexes using a computational approach that included symmetry expansion of the previously calculated map, with imposed C2 symmetry and subsequent 3D classification to detect unique features of each subunit (Supplementary Fig.2d, Supplementary Table4). In addition to the two native states built of expanded subunits and to the fully activated state with four compact subunits, we isolated new tetramer arrangements with one, two, and three activated subunits, which made up 22% of the tetramers included in the 3D classification and were probably partially-activated intermediates (Fig.6b). We hypothesize that once a first bait region is

cleaved by the prey proteinase, the transition from the expanded to the compact conformation exposes the thioester bond to covalently fix the peptidase. This would cause the bait region of the vicinal subunit to be cleaved with higher probability, as the prey cannot escape from the trap, which would make the second protomer to become compact. In some cases, however, the peptidase would not be immediately bound by the thioester bond, and would thus gain access to the bait regions of the opposite and/or disulfide-linked subunit. These cleavages would produce species with three compact protomers. These intermediate structures indicate that conformational transition for each subunit is only dependent on cleavage of its own bait region, independently of the other subunits.

Molecular mechanism of (hα$_2$M)$_4$ function – Based on the structures obtained in this study, we propose a molecular mechanism of plasma endopeptidase inhibition by (hα$_2$M)$_4$ (Fig.7). The native, unreacted tetramer, which is built of subunits in expanded conformation, is extremely porous and flexible, and comprises four large 50x70 Å openings that would fuse to yield a large, irregular cavity. The openings are framed by five of the eight *N*-linked glycans, which contribute to thermal stability and high solubility in blood plasma but not to the inhibitory capacity of hα$_2$M [49]. We hypothesize that this structure could act as a sieving complex in a crowded plasma environment, which is primed with four internal bait regions exposed to the particle lumen for testing of prey endopeptidases. Processing of the bait region leads to a conformational rearrangement of the particular protomer from an expanded to a compact conformation; this exposes the reactive thioester bond from TED for covalent entrapping of the peptidase, and the RBD for recognition by its cognate receptor for subsequent endocytosis. Native subunits become activated as they are cleaved, although for large trapped proteinases, steric constraints or clashes with the prey might prevent the structural transition in some of the processed hα$_2$M subunits. This dynamic complex is stabilized throughout this peptidase inhibition process by an inalterable structural belt formed by the MG4 domains and the BRD tethering loops.

A comparison with the mechanism derived for *Escherichia coli* α$_2$-macroglobulin (ECAM) shows substantial differences, despite a generally similar domain architecture in both inhibitors [50]. While ECAM is monomeric and anchored to the cytosolic membrane of the bacterium facing the periplasm, hα$_2$M is tetrameric and secreted to the plasma in humans. In the former, covalent linkage and steric hindrance of peptidases thus inhibit activity, but only against very large substrates. This modus operandi has been dubbed a "snap-trap mechanism". In contrast, hα$_2$M inhibition is elicited through physical entrapment in a large cage following a "Venus-flytrap mechanism", in which preys are still active against small substrates and inhibitors that can enter the cage through several apertures.

In conclusion, our structural studies report the molecular determinants of the mechanism of action of (hα$_2$M)$_4$, a universal pan-peptidase inhibitor that has been the subject of considerable biochemical and biophysical study since its discovery in 1946 [51]. We document a stepwise Venus-flytrap mechanism unique among peptidase inhibitors in its irreversibility and versatility, which enables to sequester endogenous and exogenous peptidases from plasma and remove them from the circulation. Given the role of these complexes in numerous essential processes, knowledge of their structure provides a starting point for further studies of the many functions of (hα$_2$M)$_4$.

MATERIALS AND METHODS

Cryo-EM data collection – Wild-type authentic native (hα$_2$M)$_4$ was isolated from

thawed frozen plasma from healthy human donors, which was de-identified prior to use in this study. The protein was purified, assessed for peptidase-inhibition competence as described [15, 26, 49], and verified to be equivalent to protein purified from fresh plasma in functional and physiological assays. Aliquots of pure protein (5 µL) were diluted to 0.1 mg/mL, applied to R2/2 300 mesh acetone vapor-treated copper grids, and vitrified using a Leica EM CPC cryofixation unit. Data were collected on FEI Titan Krios electron microscopes operated at 300 kV, and images recorded with Gatan K2-summit cameras in counting mode using the EPU Automated Data Acquisition Software for Single Particle Analysis (ThermoFisher Scientific). The total number of recorded movies, nominal magnification, calibrated pixel size at the specimen level, total exposure, exposure per frame, and defocus range for each specimen are described in Supplementary Table1.

Image processing – Movies were drift-corrected and dose-weighted with Motioncor2 [Zheng *et al.*, 2017], and contrast transfer function (CTF) values were estimated with CTFFIND4.1 [52] using non-dose-weighted micrographs. All subsequent image processing was with RELION 2.1 [53, 54] within Scipion [55], unless otherwise stated. The data processing workflows for $(h\alpha_2M)_4$ complexes purified from plasma, as well as for trypsin- and plasmin-treated complexes, are described in Supplementary Figs. 3, 5 and 9, respectively. Particle statistics are indicated in Supplementary Table1. Class averages from preliminary datasets were used as templates for subsequent automated particle picking with Gautomatch (written by Kai Zhang, https://www2.mrc-lmb.cam.ac.uk/research/locally-developed-software/zhang-software/). Particles were then extracted, normalized, and subjected to several rounds of reference-free two-dimensional (2D) classification to discard particles from 2D classes that did not show secondary structural elements. Selected particles were 3D-classified, imposing C2 symmetry for plasma-purified and trypsin-treated $(h\alpha_2M)_4$ complexes, and C1 symmetry for plasmin-treated complexes. Classes representing equivalent conformational states were pooled and included in a 3D auto-refinement. Particles assigned to a native state were submitted to an additional round of 3D classification without alignment to identify and refine particles corresponding to native I and II states. To identify and classify intermediate conformations between the $(h\alpha_2M)_4$ major states in plasma, the C2 symmetry of these states was expanded [56] and particles were subjected to 3D classification without alignment (particle processing is indicated in Supplementary Table 4). Classes representing equivalent conformational states were pooled and included in a 3D auto-refinement. Local resolution was estimated using MonoRes [57] and unsharpened maps treated by local resolution-based sharpening in LocalDeblur [58].

Model building and refinement – The $h\alpha_2M$ crystal structure [26] (PDB 4ACQ) was first docked manually as a rigid body into the trypsin-activated locally-sharpened density map, and then subjected to real-space fitting with the Fit_in_Map routine of Chimera [59]. A first step of real-space refinement was performed with Phenix [60] applying global minimization, local grid search and atomic displacement parameter refinement (ADP) protocols. The model was then rebuilt manually in Coot [61] to optimize the fit to the density for one set of disulfide-linked subunits (protomers 1 and 2). This asymmetric unit was then C2-symmetrized and further refined in Phenix with similar options as above, and with secondary structure, non-crystallographic symmetry (NCS), side chain rotamer, and Ramachandran restraints, as well as with hydrogens in riding positions. This model was used as starting model to build the compact subunits, and was fitted into the naturally-

activated and plasmin-activated I sample maps, as well as in densities corresponding to protomers 2, 3, and 4 of plasmin-activated II, as a starting point for model building of these states. In Chimera, activated protomer 1 coordinates were flexible fitted in the native I protomer 1 density, considering each domain as a rigid body. These docked domains were used as the starting point for model building of the prototype of an expanded subunit, whose coordinates were fitted in expanded protomers of native I (protomer 2), native II (protomers 1 and 2), semi-activated I (protomer 1), semi-activated II (protomer 1) and plasmin-activated II (protomer 1). Fitted coordinates were checked manually in Coot and C2 symmetrized (to generate protomers 3 and 4). The first step of real-space refinement was performed in Phenix, with morphing and simulated annealing options, followed by the steps described above. Refinement statistics are listed in Supplementary Table1.

Model validation and analysis – The FSC curves between model and map after local sharpening (Model *vs.* Map) are shown in Supplementary Fig.10. For cross-validation against overfitting, the atoms in the final atomic models were displaced by 0.5 Å in random directions using Phenix. The shifted coordinates were then refined against one of the half-maps (work set) in Phenix using the same procedure as for refinement of the final model. The other half-map (test set) was not used in refinement for cross-validation. FSC curves of the refined shifted model against the work set (FSCwork) and against the test set (FSCtest) are shown in Supplementary Fig.10. The FSCwork and FSCtest curves do not diverge markedly, consistent with the absence of overfitting in the final models. The quality of the atomic model was assessed by analysis of the basic protein geometry, Ramachandran plots, and clash analysis, and validated with Coot and MolProbity [62] as implemented in Phenix, and with the Worldwide PDB (wwPDB) OneDep System (https://deposit-pdbe.wwpdb.org/deposition/). Graphics were produced using UCSF Chimera.

Re-refinement of the crystallographic structure of peptidase-activated hα_2M – To obtain a more accurate model of the reported crystal structure of peptidase-activated hα_2M with complete side chains and correct glycan structure, the X-ray diffraction data (PDB 4ACQ; [26]) were reprocessed to 4.2 Å resolution with current versions of the XDS [63] and XSCALE programs [64] (Supplementary Table3). A test set for R_{free} monitoring was chosen in thin shells with SFTOOLS within the CCP4 suite of programs [65]. The coordinates of the hα_2M part of the cryo-EM model of the complex with trypsin were superimposed onto the original crystallographic coordinates and refined in reciprocal space against the newly processed data with Phenix and Buster/TNT [66]; this included hydrogens (set to zero occupancy) at riding positions for the protein residues, non-crystallographic symmetry restraints, and translation/libration/screw-rotation refinement. This refinement alternated with manual model building with Coot (see Supplementary Table3). The eight glycosylation sites of each protomer were rebuilt to match the recently determined glycan structure of hα_2M [67].

Data availability – The atomic coordinates and cryo-EM density maps were deposited in the Protein Data Bank and EM Data Bank with codes 7OTL and EMD-12747 (native I); 7O7M and EMD-12748 (native II); 7O7N and EMD-12750 (semi-activated I); 7O7O and EMD-12751 (semi-activated II); 7O7P and EMD-12752 (activated); 7O7Q and EMD-12753 (trypsin-activated); 7O7R and EMD-12754 (plasmin-activated I); 7O7S and EMD-12755 (plasmin-activated II). The cryo-EM density maps of intermediate structures with one, two [2 maps), and three activated monomers, corresponding to transient I-IV states, were deposited in the EM Data Bank with codes EMD-12941, EMD-12942, EMD-

12943 and EMD-1294. The final X-ray crystallography model was likewise deposited at the Protein Data Bank (PDB 6TAV) and supersedes PDB entry 4ACQ.

ACKNOWLEDGEMENTS

We are grateful to R. Bonet of the IBMB Protein Purification Service and to C. Mark for editorial assistance. We thank members of the Proteomics Facility of the Centro de Investigaciones Biológicas (CIB-CSIC, Madrid) for protein identification; Rocío Arranz and Javier Chichón of the Cryo-EM CNB/CIB-CSIC facility (Madrid) in the context of the CRIOMECORR project (ESFRI-2019-01-CSIC-16); the Diamond Light Source for access to its cryo-EM facility (proposals EM15997 and BI22006); the Astbury BioStructure Laboratory at the University of Leeds for help with cryo-EM data acquisition; and the European Synchrotron Radiation Facility for microscope time (proposal MX-2154). We thank personnel of the High Perfomance Computing Unit of the ISCIII Unidad de Tecnologías de la Información y Comunicación. This work was supported by grants from the Spanish Ministries of Economy and Competitivity (BFU2017-88736-R) and of Science and Innovation (PID2020-113287RB-I00) and the Comunidad Autónoma de Madrid (P2018/NMT-4389) to JRC, and by grants from Catalan and Spanish public and private agencies (BFU2019-107725-RB-I00; 2017SGR00003; Fundació "La Marató de TV3" 201815) to FXGR. TG acknowledges a Juan de la Cierva research contract (JCI-2012-13573) from the MINECO, and SRM an FPI-fellowship (BES2016-076877) from the Ministry of Science and Innovation. The Structural Biology Unit of IBMB was a María de Maeztu Unit of Excellence (2015-2019) and the Centro Nacional de Biotecnología is a Severo Ochoa Center of Excellence (MINECO award SEV 2017-0712), as awarded by the Spanish Ministry of Economy, Industry and Competitiveness. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

AUTHOR CONTRIBUTION

T.G., S.R.M. performed biochemical experiments and purified complexes; D.L., T.G. and C.P.M. acquired the cryo-EM images; D.L. processed images and C.P.M. built the models; T.G., D.L., C.P.M., S.R.M., F.X.G.-R. and J.R.C. analyzed data; F.X.G.-R. and J.R.C. conceived and coordinated the project; F.X.G.-R. and J.R.C. wrote the paper with input from all authors.

CONFLICT OF INTERESTS

The authors declare no financial or non-financial conflicts of interest with the contents of this article.

REFERENCES

[1] A. J. Barrett, P. M. Starkey, The interaction of α_2-macroglobulin with proteinases. Characteristics and specificity of the reaction, and a hypothesis concerning its molecular mechanism. *Biochem. J.* 133, 709-724 (1973).

[2] A. Budd, S. Blandin, E. A. Levashina, T. J. Gibson, Bacterial α_2-macroglobulins: colonization factors acquired by horizontal gene transfer from the metazoan genome? *Genome Biol.* 5, R38 (2004).

[3] N. Doan, P. G. W. Gettins, α-Macroglobulins are present in some Gram-negative bacteria: characterization of the α_2-macroglobulin from *Escherichia coli*. *J. Biol. Chem.* 283, 28747-28756 (2008).

[4] V. Ignjatovic, E. Mertyn, P. Monagle, The coagulation system in children: developmental and pathophysiological considerations. *Semin. Thromb. Hemost.* 37, 723-729 (2011).

[5] A. A. Rehman, H. Ahsan, F. H. Khan, α_2-Macroglobulin: a physiological guardian. *J. Cell. Physiol.* 228, 1665-1675 (2013).

[6] S. G. Wong, A. Dessen, Structure of a bacterial α_2-macroglobulin reveals mimicry of eukaryotic innate immunity. *Nat. Commun.* 5, 4917 (2014).
[7] T. Goulas *et al.*, Structural and functional insight into pan-endopeptidase inhibition by α_2-macroglobulins. *Biol. Chem.* 398, 975-994 (2017).
[8] S. Seddighi, V. Varma, M. Thambisetty, α_2-Macroglobulin in Alzheimer's disease: new roles for an old chaperone. *Biomark. Med.* 12, 311-314 (2018).
[9] P. M. Starkey, A. J. Barrett, Inhibition by α-macroglobulin and other serum proteins. *Biochem. J.* 131, 823-831 (1973).
[10] S. L. Gonias *et al.*, α2-macroglobulin and the α2-macroglobulin receptor/LRP. A growth regulatory axis. *Ann. N. Y. Acad. Sci.* 737, 273-290 (1994).
[11] C. T. Chu, S. V. Pizzo, Interactions between cytokines and α2-macroglobulin. *Immunol. Today* 12, 249 (1991).
[12] N.-M. Liu *et al.*, Transcuprein is a macroglobulin regulated by copper and iron availability. *J. Nutr. Biochem.* 18, 597-608 (2007).
[13] L. Sottrup-Jensen, α-Macroglobulins: structure, shape, and mechanism of proteinase complex formation. *J. Biol. Chem.* 264, 11539-11542 (1989).
[14] P. B. Armstrong, Role of α2-macroglobulin in the immune response of invertebrates. *Invertebrate Surviv. J.* 7, 165-180 (2010).
[15] C. J. Liu, The role of ADAMTS-7 and ADAMTS-12 in the pathogenesis of arthritis. *Nat. Clin. Pract. Rheumatol.* 5, 38-45 (2009).
[16] L. Troeberg, H. Nagase, Proteases involved in cartilage matrix degradation in osteoarthritis. *Biochim. Biophys. Acta* 1824, 133-145 (2012).
[17] J. Schaller, S. S. Gerber, The plasmin-antiplasmin system: structural and functional aspects. *Cell. Mol. Life Sci.* 68, 785-801 (2011).
[18] W. Borth, Alpha 2-macroglobulin, a multifunctional binding protein with targeting characteristics. *FASEB J* 6, 3345-3353 (1992).
[19] R. C. Roberts, Protease inhibitors of human plasma. Alpha-2-macroglobulin. *J Med* 16, 129-224 (1985).
[20] S. Zucker, R. M. Lysik, M. H. Zarrabi, J. J. Fiore, D. K. Strickland, Proteinase-alpha 2 macroglobulin complexes are not increased in plasma of patients with cancer. *Int J Cancer* 48, 399-403 (1991).
[21] I. Garcia-Ferrer, A. Marrero, F. X. Gomis-Rüth, T. Goulas, α_2-Macroglobulins: structure and function. *Subcell. Biochem.* 83, 149-183 (2017).
[22] A. F. Kisselev, K. von der Helm, Human immunodeficiency virus type 1 proteinase is rapidly and efficiently inactivated in human plasma by α_2-macroglobulin. *Biol. Chem. Hoppe Seyler* 375, 711-714 (1994).
[23] E. F. Sánchez, R. J. Flores-Ortiz, V. G. Alvarenga, J. A. Eble, Direct fibrinolytic snake venom metalloproteinases affecting hemostasis: structural, biochemical features and therapeutic potential. *Toxins (Basel)* 9, 392 (2017).
[24] M. Laskowski Jr., I. Kato, Protein inhibitors of proteinases. *Annu. Rev. Biochem.* 49, 593-626 (1980).
[25] N. D. Rawlings, Peptidase inhibitors in the MEROPS database. *Biochimie* 92, 1463-1483 (2010).
[26] A. Marrero *et al.*, The crystal structure of human α_2-macroglobulin reveals a unique molecular cage. *Angew. Chem. Int. Ed.* 51, 3340-3344 (2012).
[27] L. Sottrup-Jensen, T. E. Petersen, S. Magnusson, A thiol-ester in α_2-macroglobulin cleaved during proteinase complex formation. *FEBS Lett.* 121, 275-279 (1980).
[28] J. Travis, G. S. Salvesen, Human plasma proteinase inhibitors. *Annu. Rev. Biochem.* 52, 655-709 (1983).
[29] U. Qazi, P. G. Gettins, D. K. Strickland, J. K. Stoops, Structural details of proteinase entrapment by human a_2-macroglobulin emerge from three-dimensional reconstructions of Fab labeled native, half-transformed, and transformed molecules. *J. Biol. Chem.* 274, 8137-8142 (1999).
[30] P. B. Armstrong, J. P. Quigley, α_2-macroglobulin: an evolutionarily conserved arm of the innate immune system. *Dev. Comp. Immunol.* 23, 375-390 (1999).
[31] M. T. Debanne, R. Bell, J. Dolovich, Characteristics of the macrophage uptake of proteinase-α-macroglobulin complexes. *Biochim. Biophys. Acta* 428, 466-475 (1976).
[32] E. Delain *et al.*, The molecular organization of human a_2-macroglobulin. An immunoelectron microscopic study with monoclonal antibodies. *J. Biol. Chem.* 263, 2981-2989 (1988).
[33] O. M. Andersen *et al.*, Specific binding of α-macroglobulin to complement-type repeat CR4 of the low-density lipoprotein receptor-related protein. *Biochemistry* 39, 10627-10633 (2000).
[34] M. J. Imber, S. V. Pizzo, Clearance and binding of two electrophoretic "fast" forms of human α_2-macroglobulin. *J. Biol. Chem.* 256, 8134-8139 (1981).
[35] L. Jenner, L. Husted, S. Thirup, L. Sottrup-Jensen, J. Nyborg, Crystal structure of the receptor-binding domain of α_2-macroglobulin. *Structure* 6, 595-604 (1998).
[36] W. Huang, K. Dolmer, X. Liao, P. G. W. Gettins, NMR solution structure of the receptor binding domain of human α_2-macroglobulin. *J. Biol. Chem.* 275, 1089-1094 (2000).
[37] T. Xiao, D. L. DeCamp, S. R. Sprang, Structure of a rat α_1-macroglobulin receptor-binding domain dimer. *Protein Sci.* 9, 1889-1897 (2000).
[38] B. J. C. Janssen *et al.*, Structures of complement component C3 provide insights into the function and evolution of immunity. *Nature* 437, 505-511 (2005).
[39] N. Doan, P. G. W. Gettins, Human α_2-macroglobulin is composed of multiple domains, as predicted by homology with complement component C3. *Biochem. J.* 407, 23-30 (2007).
[40] G. R. Andersen, T. J. Koch, K. Dolmer, L. Sottrup-Jensen, J. Nyborg, Low resolution X-ray structure of

human methylamine-treated α_2-macroglobulin. *J. Biol. Chem.* 270, 25133-25141 (1995).
[41] N. Boisset, J. C. Taveau, F. Pochon, J. Lamy, Similar architectures of native and transformed human α_2-macroglobulin suggest the transformation mechanism. *J. Biol. Chem.* 271, 25762-25769 (1996).
[42] S. J. Kolodziej, J. P. Schroeter, D. K. Strickland, J. K. Stoops, The novel three-dimensional structure of native human α_2-macroglobulin and comparisons with the structure of the methylamine derivative. *J. Struct. Biol.* 116, 366-376 (1996).
[43] S. J. Kolodziej, T. Wagenknecht, D. K. Strickland, J. K. Stoops, The three-dimensional structure of the human a_2-macroglobulin dimer reveals its structural organization in the tetrameric native and chymotrypsin a_2-macroglobulin complexes. *J. Biol. Chem.* 277, 28031-28037 (2002).
[44] S. L. Harwood *et al.*, Structural Investigations of Human A2M Identify a Hollow Native Conformation That Underlies Its Distinctive Protease-Trapping Mechanism. *Mol Cell Proteomics* 20, 100090 (2021).
[45] A. J. Barrett, M. A. Brown, C. A. Sayers, The electrophoretically 'slow' and 'fast' forms of the α_2-macroglobulin molecule. *Biochem. J.* 181, 401-418 (1979).
[46] L. Sottrup-Jensen, T. E. Petersen, S. Magnusson, Trypsin-induced activation of the thiol esters in alpha 2-macroglobulin generates a short-lived intermediate ('nascent' a_2-M) that can react rapidly to incorporate not only methylamine or putrescine but also proteins lacking proteinase activity. *FEBS Lett.* 128, 123-126 (1981).
[47] G. A. Jensen *et al.*, Binding site structure of one LRP-RAP complex: implications for a common ligand-receptor binding motif. *J. Mol. Biol.* 362, 700-716 (2006).
[48] L. Sottrup-Jensen *et al.*, Primary structure of the 'bait' region for proteinases in a_2-macroglobulin. Nature of the complex. *FEBS Lett.* 127, 167-173 (1981).
[49]. T. Goulas, I. Garcia-Ferrer, S. Garcia-Pique, L. Sottrup-Jensen, F. X. Gomis-Ruth, Crystallization and preliminary X-ray diffraction analysis of eukaryotic alpha2 -macroglobulin family members modified by methylamine, proteases and glycosidases. *Mol Oral Microbiol* 29, 354-364 (2014).
[50] I. Garcia-Ferrer *et al.*, Structural and functional insights into Escherichia coli alpha2-macroglobulin endopeptidase snap-trap inhibition. *Proc Natl Acad Sci U S A* 112, 8290-8295 (2015).
[51] E. J. Cohn *et al.*, Preparation and properties of serum and plasma proteins; a system for the separation into fractions of the protein and lipoprotein components of biological tissues and fluids. *J. Am. Chem. Soc.* 68, 459-475 (1946).
[52] A. Rohou, N. Grigorieff, CTFFIND4: Fast and accurate defocus estimation from electron micrographs. *J Struct Biol* 192, 216-221 (2015).
[53] S. H. Scheres, RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J Struct Biol* 180, 519-530 (2012).
[54] R. Fernandez-Leiro, S. H. W. Scheres, A pipeline approach to single-particle processing in RELION. *Acta Crystallogr D Struct Biol* 73, 496-502 (2017).
[55]. J. M. de la Rosa-Trevin *et al.*, Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. *J Struct Biol* 195, 93-99 (2016).
[56] S. H. Scheres, Processing of Structurally Heterogeneous Cryo-EM Data in RELION. *Methods Enzymol* 579, 125-157 (2016).
[57] J. L. Vilas *et al.*, MonoRes: automatic and accurate estimation of local resolution for electron microscopy maps. *Structure* 26, 337-344 (2018).
[58] E. Ramirez-Aportela *et al.*, Automatic local resolution-based sharpening of cryo-EM maps. *Bioinformatics* 36, 765-772 (2020).
[59] E. F. Pettersen *et al.*, UCSF Chimera - A visualization system for exploratory research and analysis. *J. Comput. Chem.* 25, 1605-1612 (2004).
[60] P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. sect. D* 66, 213-221 (2010).
[61] P. Emsley, K. Cowtan, COOT: model-building tools for molecular graphics. *Acta Crystallogr. sect. D* 60, 2126-2132 (2004).
[62] V. B. Chen *et al.*, MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. sect. D* 66, 12-21 (2010).
[63] W. Kabsch, XDS. *Acta Crystallogr. sect. D* 66, 125-132 (2010).
[64] W. Kabsch, Integration, scaling, space-group assignment and post-refinement. *Acta Crystallogr. sect. D* 66, 133-144 (2010).
[65] M. D. Winn *et al.*, Overview of the CCP4 suite and current developments. *Acta Crystallogr. sect. D* 67, 235-242 (2011).
[66] O. S. Smart *et al.*, Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr. sect. D* 68, 368-380 (2012).
[67] F. Clerc *et al.*, Human plasma protein N-glycosylation. *Glycoconj. J.* 33, 309-343 (2016).

SUPPLEMENTARY INFORMATION

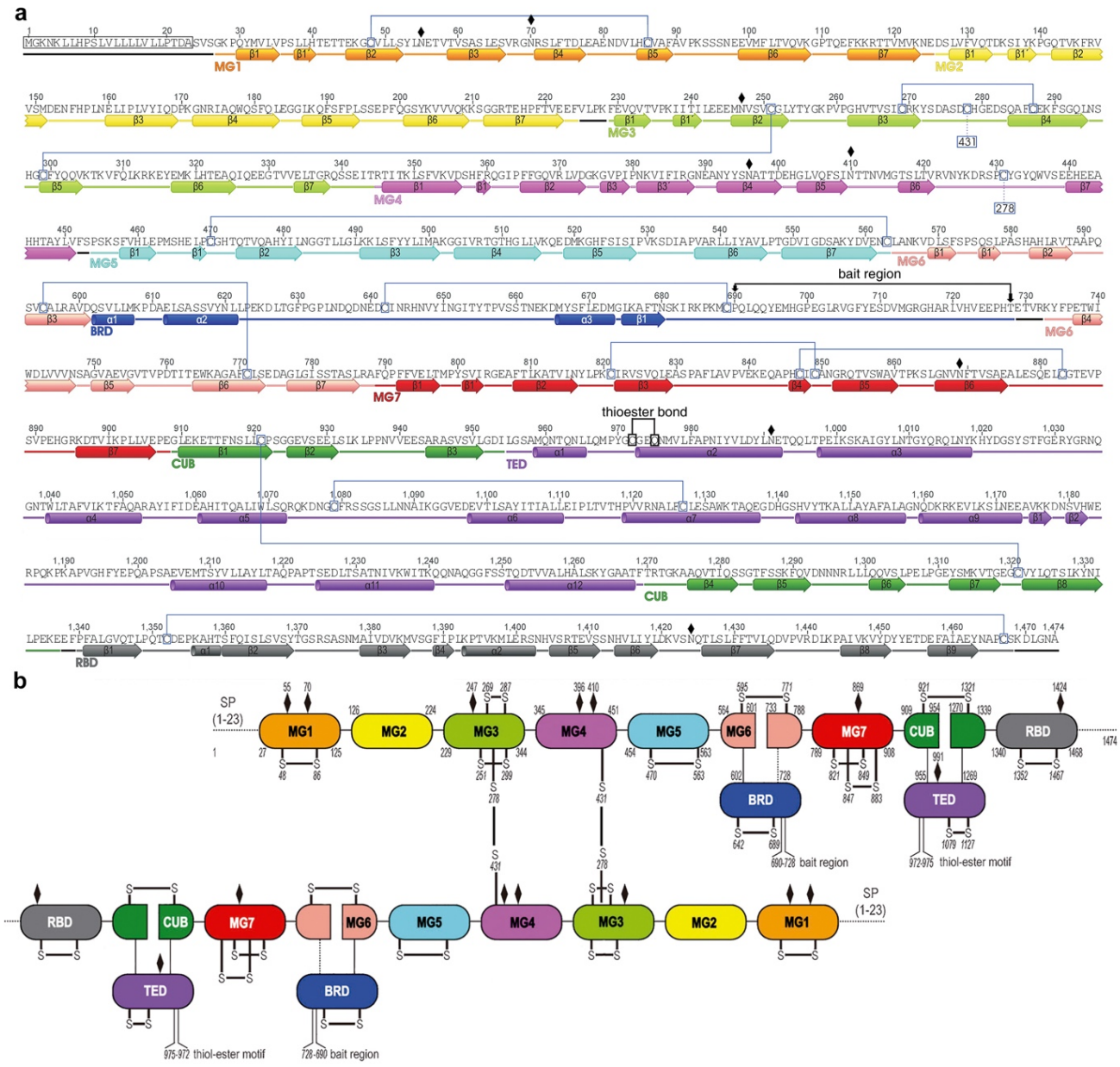
"Cryo-EM shows structural basis of pan-peptidase inhibition by human α_2-macroglobulin"

Daniel Luque[#], Theodoros Goulas[#], Carlos P. Mata[#], Soraia R. Mendes, F. Xavier Gomis-Rüth[*] and José R. Castón[*]

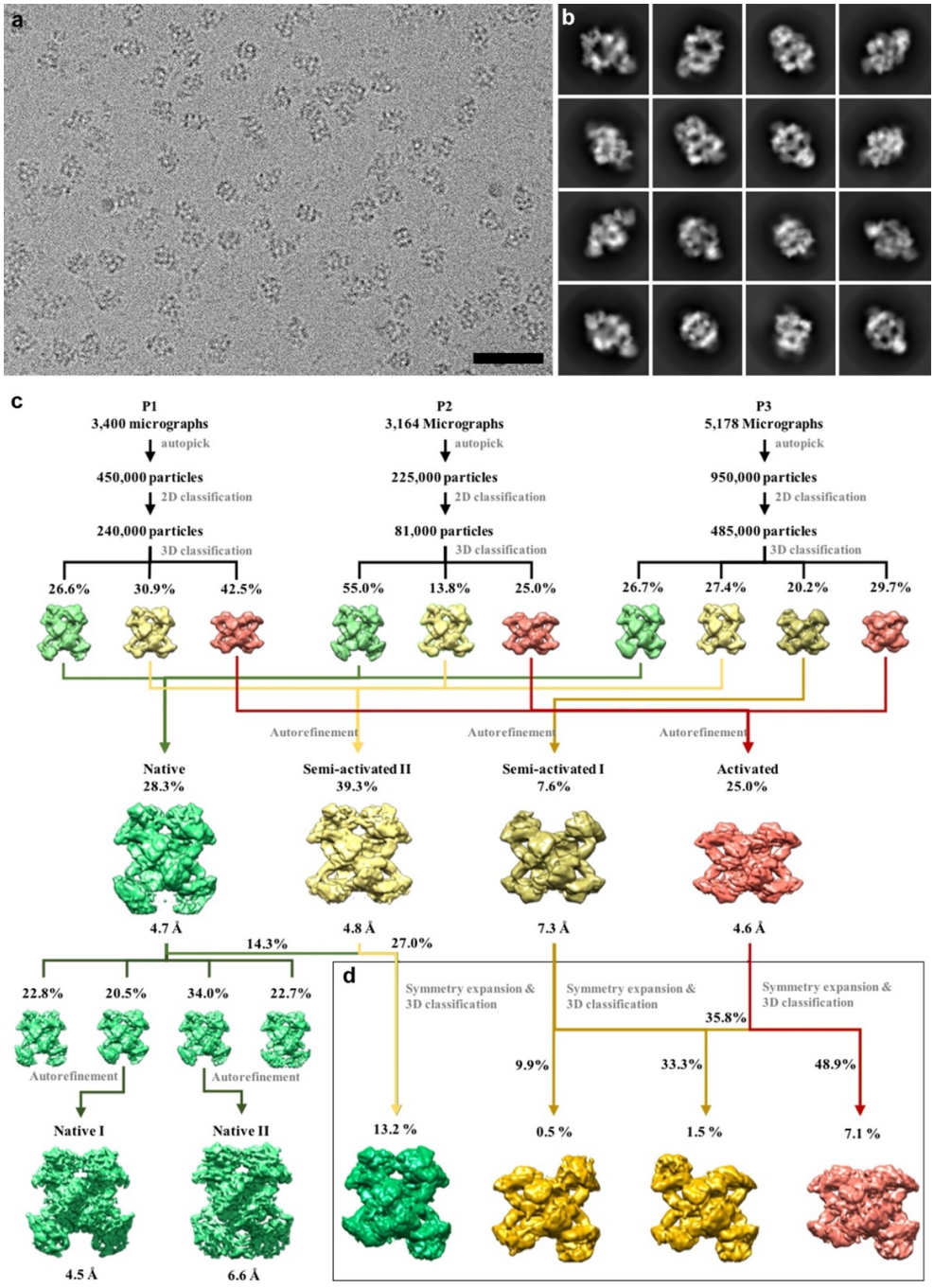
Supplementary Figures: 9

Supplementary Tables: 4

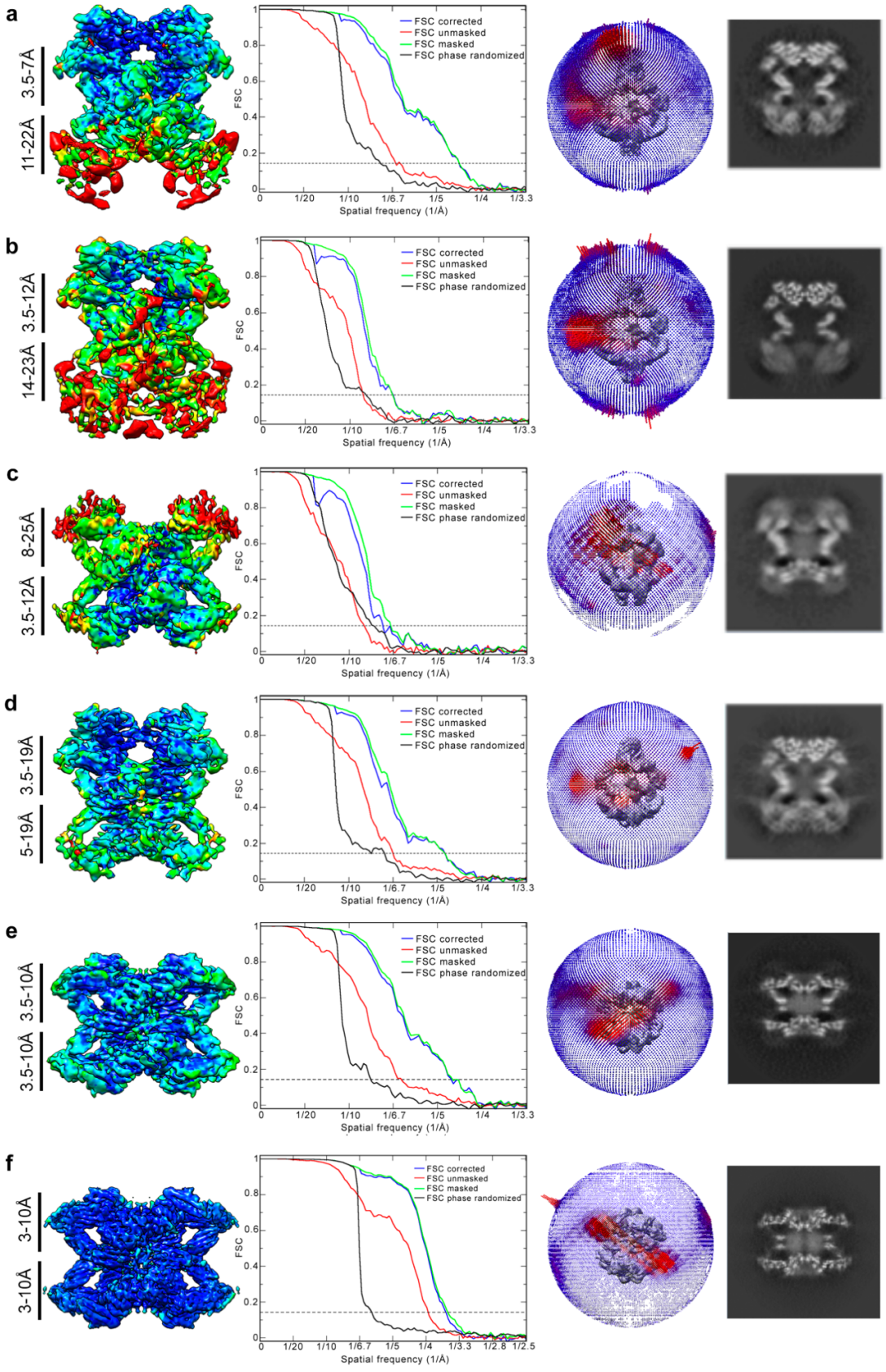
Legends to Supplementary Movies 1 and 2 (uploaded separately)

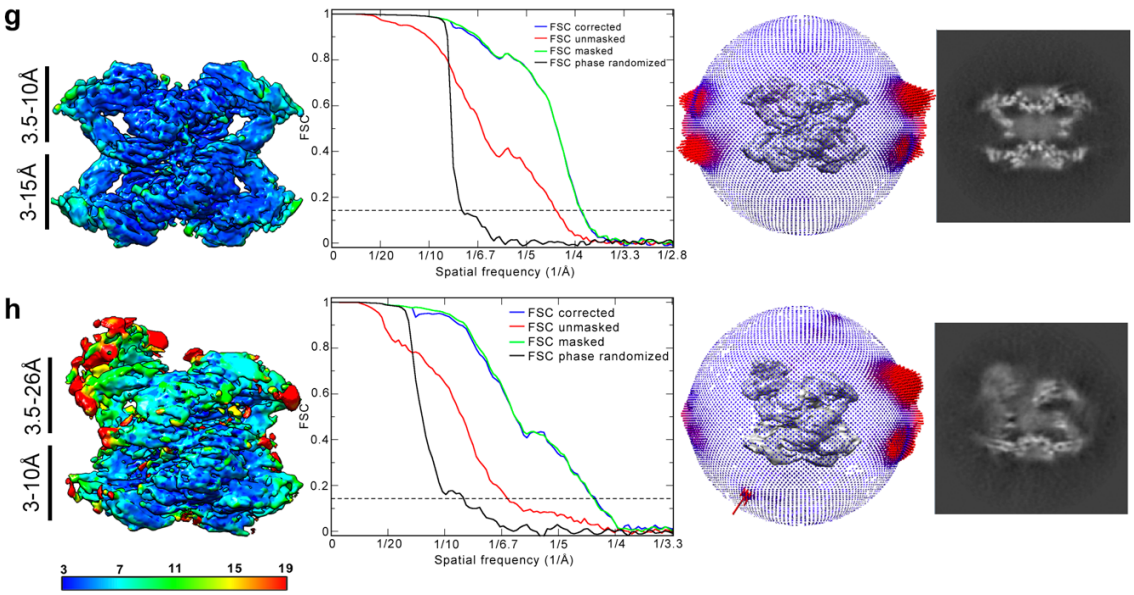


Supplementary Figure S1: Sequence, secondary structure elements, and domain organization of hα_2M. (**a**) Sequence and secondary structure elements of the 1474-residue hα_2M (UniProt code P01023) spanning a 23-residue signal peptide and the 1451-residue secreted protein. In the expanded native conformation, the first and last residues assigned were Ser26 and Ser1468, respectively; the compact activated conformation spans Lys28–Glu1335. The α-helices and β-strands are represented as cylinders and arrows, respectively, colored distinctly for the eleven domains. The bait region and the thioester bond are highlighted. Intra- and intermolecular disulfide bonds are designated with square boxes linked with a solid or dotted line (blue), respectively. N-glycosylation sites are indicated by rhombuses. MG, macroglobulin-like domains; BRD, bait region domain; CUB, domain found in C1r/C1s, urchin embryonic growth factor, and bone morphogenetic protein 1; TED, thioester domain; RBD, receptor binding domain. (**b**) Diagram of the domain organization of the disulfide-linked homodimer. The protomers are bound through two disulfide bonds. Symbols and colors as in (a).

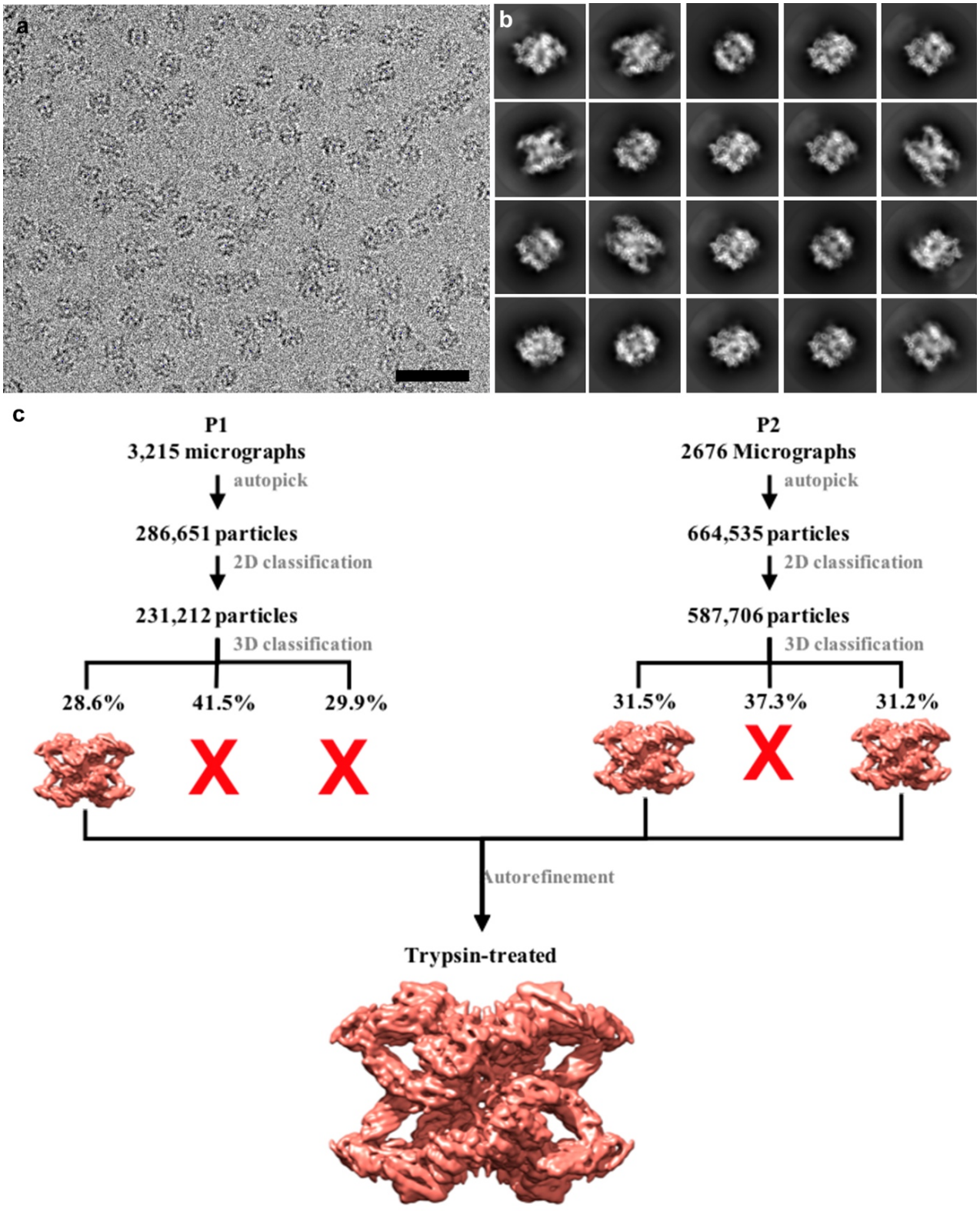


Supplementary Figure S2: Cryo-EM data processing of (hα$_2$M)$_4$ corresponding to the native untreated fraction. (**a**) Representative cryo-EM image of (hα$_2$M)$_4$ complexes (bar, 500 Å). (**b**) 2D class averages of (hα$_2$M)$_4$. (**c**) Data processing workflow and structure determination of the five functional states of (hα$_2$M)$_4$ using three serum preparations (P1-P3). After 3D classification, three (green, yellow, and red for preparations P1 and P2) and four (green, yellow, gold, and red for P3) distinct conformational states were identified (percentages indicated relative to total 3D selected particles of each preparation). Particles of each of the four states were combined into four separate datasets and further refined. The global resolution of each of the resulting 3D reconstructions was 4.7, 4.8, 7.3 and 4.6 Å, respectively, and the percentages relative to the sum of particles in the four classes was 28, 39, 8, and 25%, respectively. The dataset corresponding to truly native (hα$_2$M)$_4$ (green) was further classified and two conformational states were refined, yielding resolutions of 4.5 and 6.6 Å for native I and II states, respectively. (**d**) The C2 symmetry of the native, semi-activated I plus II and activated states was expanded and the particles from each state were subjected to additional 3D classification without alignment. Classes representing equivalent conformational states (percentages relative to the parental state are indicated) were pooled, resulting in 3D reconstructions of four new intermediate transient states.

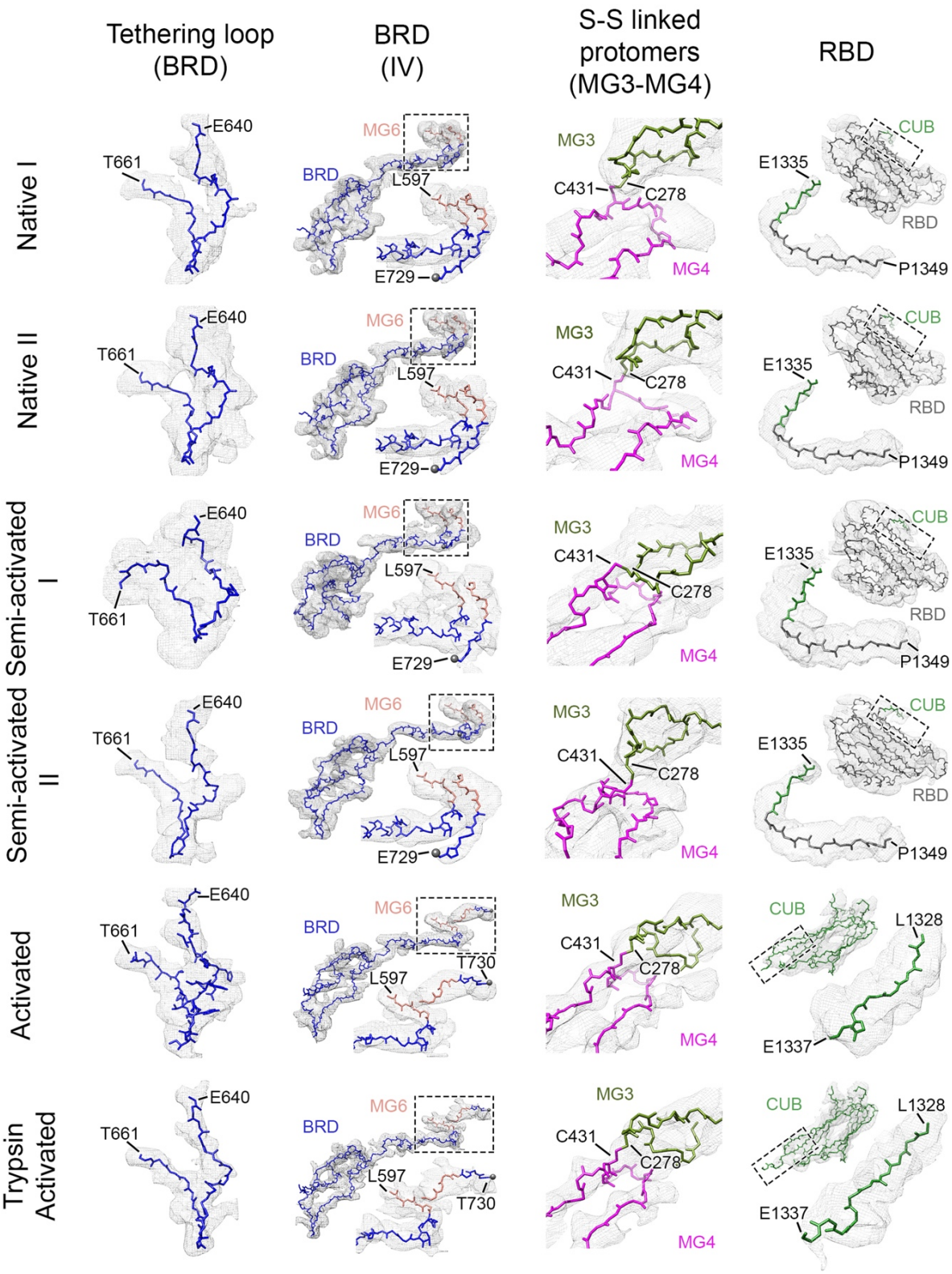

a
3.5-7Å
11-22Å
FSC corrected
FSC unmasked
FSC masked
FSC phase randomized
FSC
Spatial frequency (1/Å)
b
3.5-12Å
14-23Å
c
8-25Å
3.5-12Å
d
3.5-19Å
5-19Å
e
3.5-10Å
3.5-10Å
f
3-10Å
3-10Å



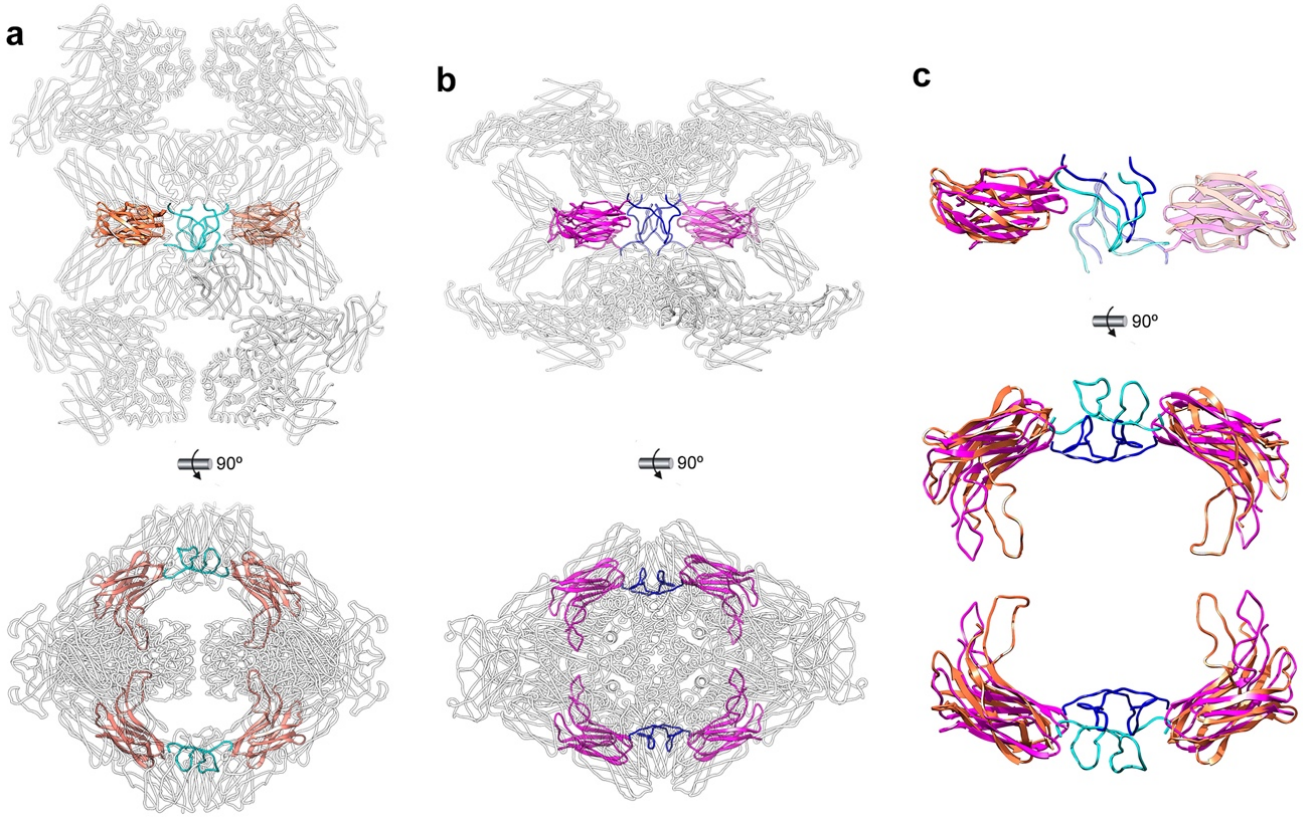
Supplementary Figure S3: Global and local resolution of cryo-EM maps. Local resolution assessment and Fourier shell correlation (FSC) curves for the eight maps calculated in this study: (**a**) native I, (**b**) native II, (**c**) semi-activated I, (**d**) semi-activated II, (**e**) fully activated state, (**f**) trypsin-activated state, (**g**) plasmin-activated I state, and (**h**) plasmin-activated II state. The bar indicates the resolution in Å (bottom). The angular distributions of particles used to compute the final three-dimensional maps are shown (top, right), as well as the longitudinal central section of each map (protein is white, bottom, right).



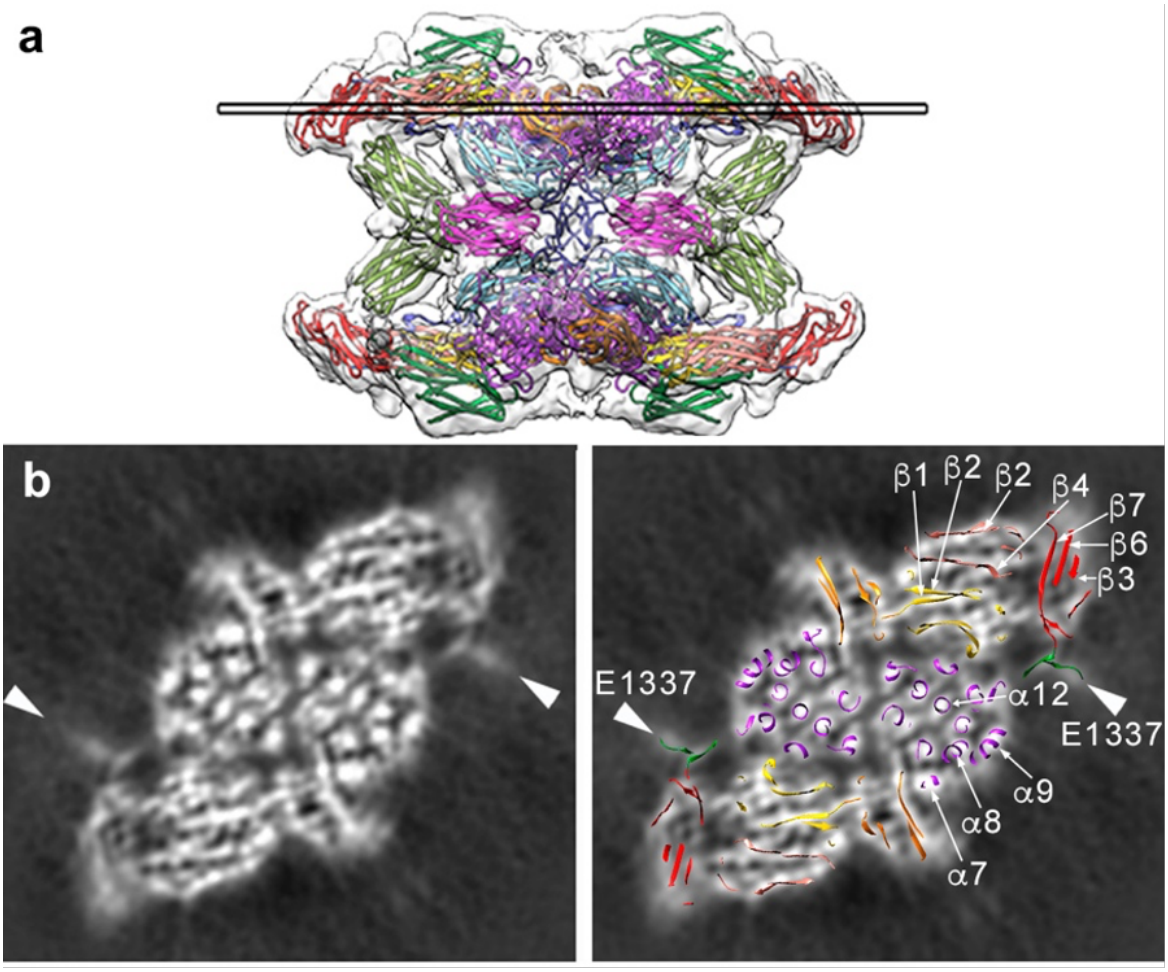
Supplementary Figure S4: Cryo-EM data processing of trypsin-treated (hα$_2$M)$_4$. (**a**) Cryo-EM image of (hα$_2$M)$_4$ complexes after trypsin treatment (bar, 500 Å). (**b**) 2D class averages of trypsin-treated (hα$_2$M)$_4$ particles. (**c**) Data processing workflow and structure determination of the trypsin-treated, activated (hα$_2$M)$_4$ state using two preparations (P1 and P2). After 2D and 3D classifications, a homogenous population of particles was selected for each preparation. Particles for each of these pools were combined and autorefined. Percentages relative to total 2D selected particles of each preparation are indicated. See Methods for further details on cryo-EM data processing.



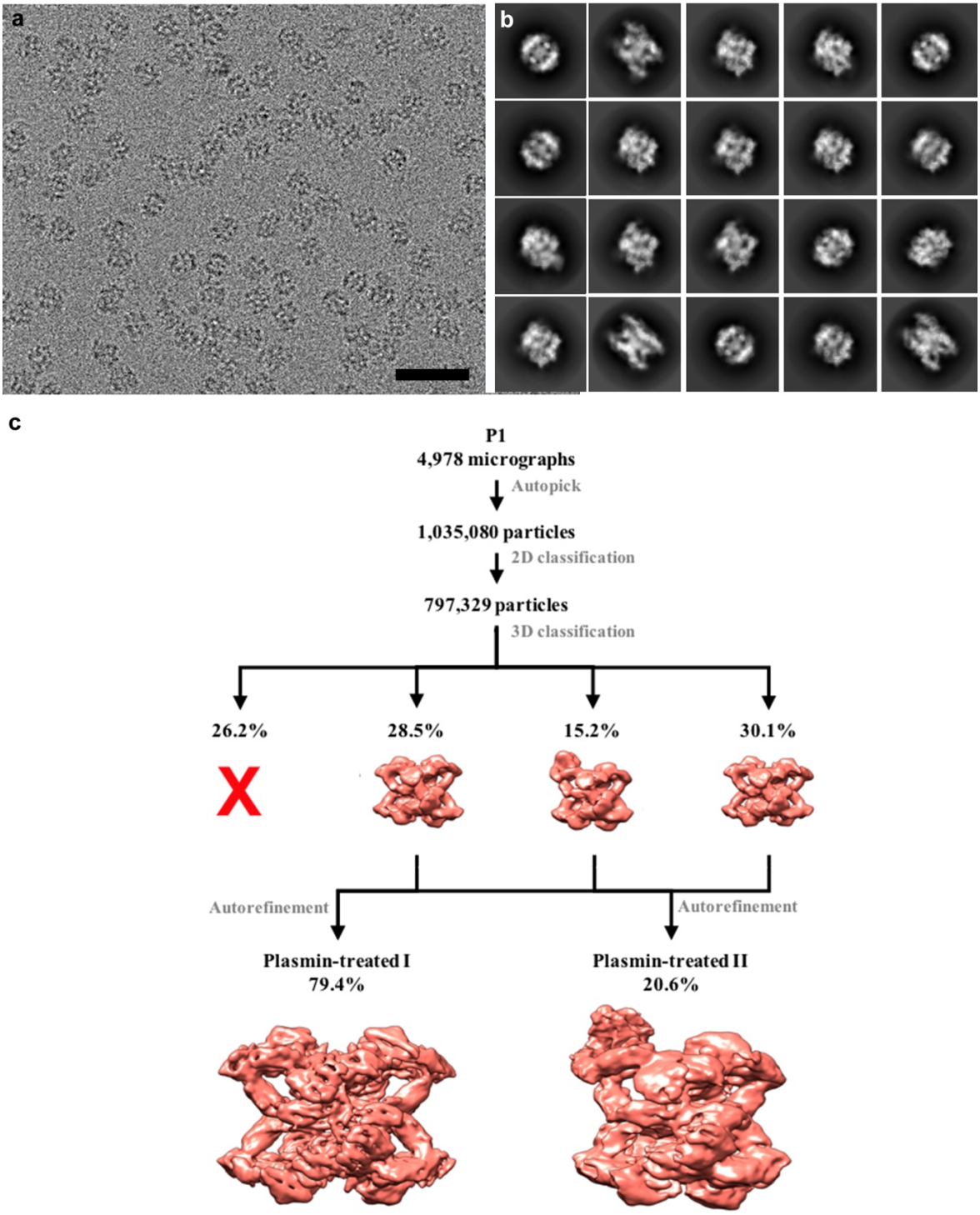
Supplementary Figure S5: Quality of the $(h\alpha_2M)_4$ density maps. Cryo-EM density maps around four different regions (the tethering loop, the BRD region IV, the disulfide-linked MG3-MG4, and the RBD) of $(h\alpha_2M)_4$ structures of native I, native II, semi-activated I, semi-activated II, naturally activated and trypsin-activate states. The cryo-EM density map is shown as a grey mesh with the corresponding atomic model with some selected residues indicated. The box of BRD region IV is shown below in a magnified view for each structure.



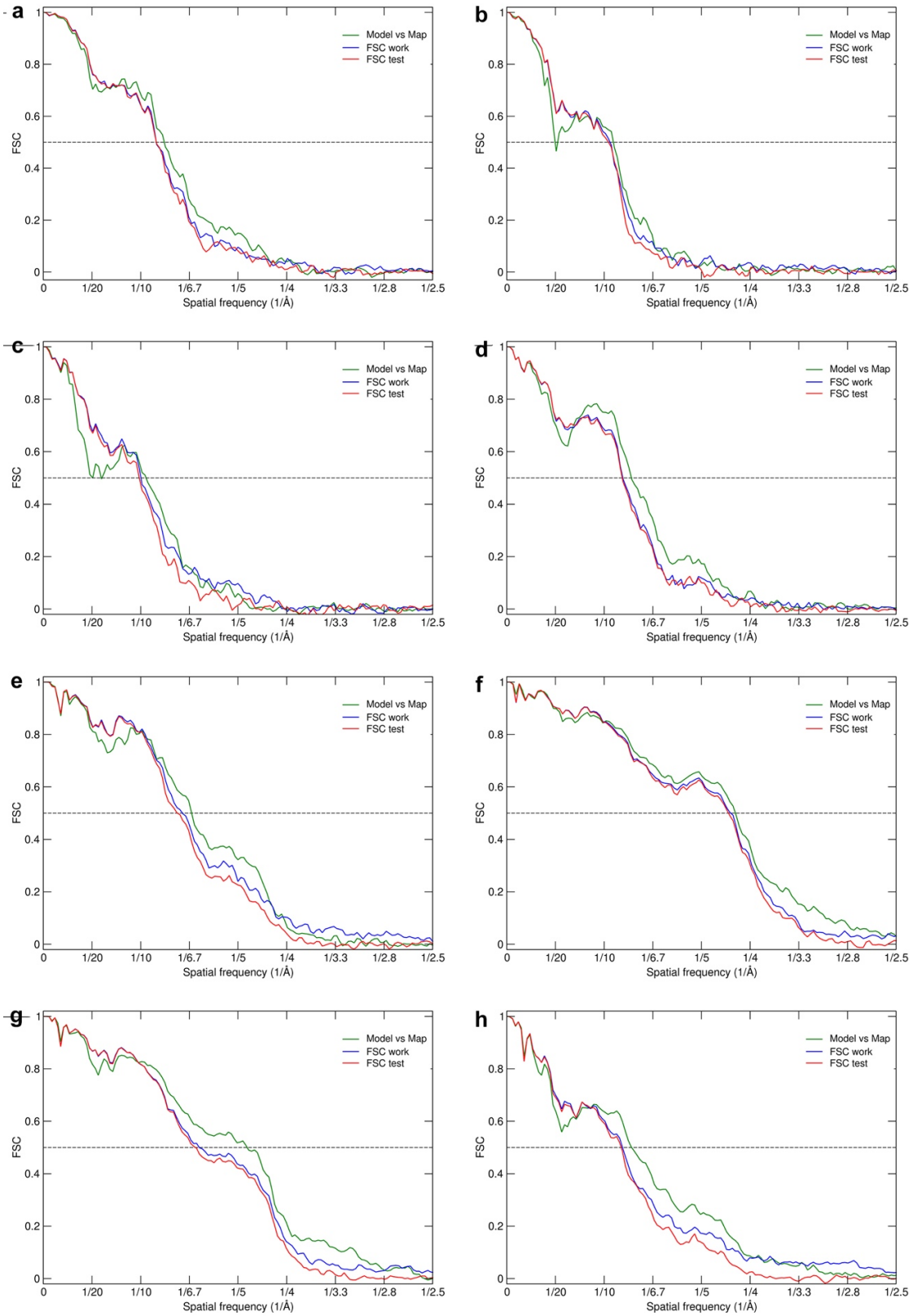
Supplementary Figure S6: Structural belt of MG4 domains and BRD tethering loops. (**a**) Two orthogonal views of the native II state in which the MG4 domains (orange) and the tethering loops (light blue) are highlighted. (**b**) Views as in (a) of the activated state with highlighted MG4 domains (magenta) and tethering loops (dark blue). (**c**) Superimposed MG4 domains and tethering loops of native II and activated states (color code as in a and b).



Supplementary Figure S7: The RBD is flexible in the cryo-EM density map of activated $(h\alpha_2M)_4$. (**a**) Structure of the activated state (domain colors as in Figure S1). The thin rectangle indicates the transversal section shown in (b). (**b**) Section across the cryo-EM map without (left) and with atomic model regions of $(h\alpha_2M)_4$ (right). The last visible $h\alpha_2M$ residue in the model is Glu1337, located in the connecting loop between CUB and RBD, and indicated with two molecules (arrowheads).



Supplementary Figure S8: Cryo-EM data processing of plasmin-treated ($h\alpha_2M)_4$. (**a**) Cryo-EM image of $(h\alpha_2M)_4$ complexes after plasmin treatment (bar, 500 Å). (**b**) 2D class averages of $(h\alpha_2M)_4$ plasmin-treated particles. (**c**) Data processing workflow and structure determination of the plasmin-treated $(h\alpha_2M)_4$ activated state. After 3D classification, two conformational states (I and II) were identified (percentages relative to total 2D selected particles of each preparation are indicated). Particles for each of these states were combined and further refined (percentages indicated relative to total 3D selected particles).



Supplementary Figure S9: Cryo-EM map quality and model validation. FSC of the refined model versus the map (green curve) and FSC work/FSC test validation curves (blue and red curves, respectively) for the eight maps calculated in this study: (**a**) native I, (**b**) native II, (**c**) semi-activated I, (**d**) semi-activated II, (**e**) fully activated, (**f**) trypsin-activated, (**g**) plasmin-activated I, and (**h**) plasmin-activated II states.

Supplementary Table S1. Cryo-EM data collection and refinement statistics.

	Native I **EMD-12747** **PDB: 7O7L**	**Native II** **EMD-12748** **PDB: 7O7M**	**Semiactivated I state** **EMD-12750** **PDB: 7O7N**	**Semiactivated II state** **EMD-12751** **PDB: 7O7O**	**Activated** **EMD-12752** **PDB: 7O7P**	**Trypsin-activated** **EMD-12753** **PDB: 7O7Q**	**Plasmin-activated I state** **EMD-12754** **PDB: 7O7R**	**Plasmin-activated II state** **EMD-12755** **PDB: 7O7S**
Data collection and processing								
Microscope	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios
Detector	K2	K2	K2	K2	K2	K2	K2	K2
Magnification	47.755x	47.755x	47.755x	47.755x	47.755x	47.755x	130,000x	130,000x
Voltage (kV)	300	300	300	300	300	300	300	300
Electron exposure ($e^-/Å^2$)	39.6	39.6	39.6	39.6	39.6	40.0	38,7	38,7
Exposure per frame ($e^-/Å^2$)	0.99-1.27	0.99-1.27	0.99-1.27	0.99-1.27	0.99-1.27	1.12-1.25	0.96	0.96
Defocus range (μm)	-1.00 to -3.25	-1.00 to -3.25	-1.00 to -3.25	-1.00 to -3.25	-1.00 to -3.25	-0.70 to -2.5	-1.30 to -3.70	-1.30 to -3.70
Pixel size (Å)	1.047	1.047	1.047	1.047	1.047	1.047	1.052	1.052
Micrographs collected (no.)	12,143	12,143	12,143	12,143	12,143	6,514	4,978	4,978
Initial particles (no.)	1,625,000	1,625,000	1,625,000	1,625,000	1,625,000	933,186	1,035,080	1,035,080
Final particles (no.)	45,669	30,618	35,993	185,640	118,333	434,851	121,437	466,082
Symmetry imposed	C2	C2	C2	C2	C2	C2	C1	C1
Map resolution (Å)	4.5	6.6	7.3	4.8	4.6	3.6	3.9	4.3
FSC threshold	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.143
Map resolution range (Å)	3.5 – 22.4	3.5 – 23.7	3.5 – 25.8	3.5 – 19.7	3.5 – 9.8	2.2 – 10.1	2.8 – 15.3	2.8 – 25.9
Refinement								
Model resolution (Å)	7.51	8.65	8.17	6.95	6.37	4.10	4.32	6.83
FSC threshold	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
Mask correlation coefficient	0.64	0.57	0.57	0.69	0.69	0.76	0.75	0.58
Map sharpening B factor	LocalDeblur	LocalDeblur	LocalDeblur	LocalDeblur	LocalDeblur	LocalDeblur	LocalDeblur	LocalDeblur
Model composition								
Non-hydrogen atoms	44,614	44,614	42,372	42,642	40,295	40,550	40,530	41,466
Protein residues	5,640	5,640	5,370	5,376	5,100	5,126	5,126	5,236
Ligands								
NAG	38	38	34	42	36	36	36	38
BMA	4	4	0	6	0	4	4	5
MAN	2	2	0	4	0	0	0	1
ADP (B-factors)								
min								
max								
mean								
Protein	143.00	135.21	63.37	164.96	165.62	116.08	131.87	124.45
	819.33	999.99	821.45	806.53	490.80	279.28	415.19	564.41
	375.62	502.63	311.57	328.20	278.28	159.94	201.79	280.43
Ligand	293.91	293.91	238.72	293.91	239.72	164.29	164.29	168.41
	471.01	471.01	446.90	483.22	347.75	211.87	211.87	448.26
	374.75	374.75	329.92	396.34	299.90	187.07	186.77	247.15
R.m.s. deviations								
Bond lengths (Å)	0.007	0.006	0.006	0.010	0.009	0.009	0.013	0.009
Bond angles (°)	1.154	1.019	1.003	1.231	1.143	1.494	1.328	1.276
Validation								
MolProbity score	2.33	2.22	2.09	2.44	2.20	2.16	2.55	2.60
Clashscore	13.26	10.43	8.78	17.31	11.58	9.18	20.68	21.36
Rotamer outliers (%)	0.49	0.36	0.30	0.43	0.13	0.18	1.07	1.09
Ramachandran plot								
Favored (%)	82.49	83.74	87.26	82.15	87.04	84.27	81.39	79.18
Allowed (%)	16.83	15.73	12.33	17.39	12.65	14.44	18.28	20.13
Outliers (%)	0.68	0.53	0.41	0.47	0.31	1.29	0.33	0.69

Supplementary Table S2. Analysis of intra-subunit domains interactions.

	Extended subunit	**Compact subunit**
MG1	MG3, MG5, BRD	MG2, MG5, BRD
MG2	MG1, MG6, BRD, TED	MG1, MG6, BRD, MG7, CUB, TED
MG3	MG4, BRD-MG6 loop, MG7	MG4, MG6, MG7, BRD
MG4	MG3, MG5, BRD	MG3, MG5, BRD
MG5	MG1, MG4, BRD	MG1, MG4, BRD
MG6	MG2, MG3 (BRD-MG6 loop), BRD, MG7	MG2, MG3, BRD, MG7
BRD	MG1, MG2, MG4, MG5, MG6	MG1, MG2, MG3, MG4, MG5, MG6
MG7	MG3, MG6, CUB, RBD	MG2-MG3 loop, MG3, MG6, CUB
CUB	TED, RBD	MG2, TED
TED	MG2, CUB, RBD	MG2, CUB
RBD	MG7, CUB, TED	

Supplementary Table S3. Crystallographic data reprocessing and model re-refinement parameters.

Dataset	:d $(h\alpha_2M)_4$
PDB Access code	
Space group	
Cell constants (a, b, c, in Å)	.8
Wavelength (Å)	
No. of measurements / unique reflections	
Resolution range (Å)	– 4.20) [a]
Completeness (%)	
R_{merge}	
R_{meas}	
$CC^{1/2}$	
Average intensity	
B-Factor (Wilson) (Å^2)	
Aver. multiplicity	
No. of reflections used in refinement [in test set]	
Crystallographic R_{factor} / free R_{factor}	
Correlation coefficient F_{obs}-F_{calc} [test set]	
No. of protein residues / non-hydrogen atoms /	
covalent ligands	, 5 MAN [b]
Rmsd from target values	
bonds (Å) / angles (°)	
Average B-factors (Å^2) (overall // mol. A/ B/ C/ D)	249 / 256
All-atom contacts and geometry analysis [c]	
Protein residues	
in favored regions / outliers / all residues	) (4%) / 5,214
with outlying rotamers / bonds / angles / chirality / torsion	/ 0 / 0
All-atom clashscore	4.3

[a] Data processing values in round brackets are for the outermost resolution shell. Model refinement parameters in square brackets are for the test set of reflection. [b] NAG, *N*-acetyl-D-glucosamine; BMA, β-D-mannose; and MAN, α-D-mannose. [c] According to the wwPDB X-ray Structure Validation Service.

Supplementary Table S4. Cryo-EM data collection.				
	Transient I state EMD-12941	**Transient II state EMD-12942**	**Transient III state EMD-12943**	**Transient IV state EMD-12944**
Data collection and processing				
Microscope	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios	FEI Titan Krios
Detector	K2	K2	K2	K2
Magnification	47.755x	47.755x	47.755x	47.755x
Voltage (kV)	300	300	300	300
Electron exposure ($e^-/Å^2$)	39.6	39.6	39.6	39.6
Exposure per frame ($e^-/Å^2$)	0.99-1.27	0.99-1.27	0.99-1.27	0.99-1.27
Defocus range (μm)	-1.00 to -3.25	-1.00 to -3.25	-1.00 to -3.25	-1.00 to -3.25
Pixel size (Å)	1.047	1.047	1.047	1.047
Micrographs collected (no.)	12,143	12,143	12,143	12,143
Initial particles (no.)	1,625,000	1,625,000	1,625,000	1,625,000
Final particles (no.)	213,866	7,131	23,998	116,074
Symmetry imposed	C1	C1	C1	C1
Map resolution (Å)	5.2	12.0	9.1	9.3
FSC threshold	0.143	0.143	0.143	0.143
Map resolution range (Å)	2.3 – 37.2	4.9 – 47.9	2.1 – 55.9	4.4 – 55.9

Legend to Supplementary Movie S1:
Movie depicting the flexible arrangement of the distinct expanded and compact subunits within the native, intermediate and activated tetramers.

Legend to Supplementary Movie S2:
Movie depicting the transition between expanded and compact conformations of a single entire protomer (left panel), the MG1-MG6 block (central panel), and the MG7-CUB-TED-RBD block (right panel).

Project 3: Work under development

"α2M samples purified from frozen and unfrozen fresh plasma present no significant structural or functional differences"

Sparked by remarks during review of the previous manuscript, we prepared a set of new experiments to compare hα2M purified from frozen and non-frozen fresh plasma. Using both biophysical and functional assays, we provided experimental insights into this highly critical aspect for the very first time. Importantly, the concept of "frozen" versus "non-frozen" plasma must not be mixed with the terms "fresh" versus "non-fresh". The latter refers to the time elapsed since the blood donation, and only plasma that is processed within twenty-four hours after blood collection is considered "fresh".

MATERIALS AND METHODS

Isolation and purification of hα2M

Hα2M was isolated from fresh blood plasma from individual donors and purified essentially as described previously (Goulas *et al.*, 2014). From the initially unfrozen sample, half was used directly for purification and the other half was frozen at -20ºC (approx. within 2 h) until purification (typically for 16 hours), thus mimicking blood bank conditions. Plasma was subjected to sequential precipitation steps with 4–12% PEG 4,000, and the final precipitate containing hα2M was reconstituted in 20 mM sodium phosphate supplemented with 5 mM PMSF (pH 6.8). Partially purified hα2M was captured with a zinc-chelating resin (G-Biosciences), washed with buffer A (50 mM sodium phosphate, 250 mM sodium chloride, 10 mM imidazole, pH 7.2) and eluted in the same buffer but with 100 mM EDTA instead of imidazole. Hα2M-containing samples were buffer exchanged to buffer B (20 mM sodium phosphate, pH 7.4) using a PD-10 column (Cytiva) and further purified by ion exchange chromatography (IEC) in a TSKgel DEAE-2SW column pre-equilibrated with buffer B. A gradient of 2–50% buffer C (20 mM sodium phosphate, 1 M sodium chloride, pH 7.4) was applied over 30 mL, and fractions containing hα2M were pooled. Next, each pool was concentrated and subjected to a final polishing step by size exclusion chromatography (SEC) in a Superose 6 Increase 10/300 column in buffer D (20 mM Tris-HCl, 150 mM sodium chloride, pH 7.4). Importantly, all purification steps were performed at 4ºC. Methylamine (MA)-induced hα2M (MA-hα2M) samples were obtained through overnight incubation at 4°C of native hα2M after the IEC purification step with 200 mM methylamine and 100 mM Tris-HCl pH 8.0. Similar to the native counterpart, the MA-induced sample was subjected to a final SEC polishing step.

Proteolytic inhibition assays

Purified hα2M was used to study protease inhibition after 10 minutes of preincubation at room temperature with two peptidases of interest: the serine peptidase trypsin from bovine pancreas, and the metallopeptidase thermolysin from *Bacillus thermoproteolyticus*. Reactions were carried out in buffer D at molar ratios of 4:1 to 1:4 for tetrameric α2M to peptidase. Samples were used directly to monitor residual proteolytic activity of the tested peptidases against natural and fluorogenic protein substrates at 37°C.

Activity against the fluorogenic BODIPY casein substrate (Invitrogen) was analysed in a microplate fluorimeter (Synergy H1, Biotek; λex = 505nm and λem = 513nm; Invitrogen) The substrate was used at 5 μg/mL in 100μL-reaction volumes and peptidases or peptidase-inhibitor complexes were used at 50 nM final concentration of the peptidase.

The natural protein substrates, namely bovine milk α-casein (35 kDa) and fibrinogen from human plasma (340 kDa), were used at 0.5 mg/mL, and trypsin and thermolysin were used at 100 nM and 5 nM, respectively. The reactions were performed over a period of 10 (for thermolysin) and 90 (for trypsin) minutes and substrate cleavage was assessed by 10–14% Tricine-SDS-PAGE after stopping the reaction by adding a small-molecule inhibitor such as 0.7 mM Pefabloc SC (Roche Life Science) for trypsin and 20 mM EDTA for thermolysin, and subsequent boiling of the samples at 95°C for 5 minutes after addition of SDS sample buffer.

Size Exclusion Chromatography with Multi-angle laser light scattering (SEC-MALLS)

Multi-angle light scattering in a Dawn Helios II apparatus (Wyatt Technologies) coupled to a Superose 6 10/300 Increase column (GE) equilibrated in buffer D was performed at 4°C at the joint IBMB/IRB Crystallography Platform, Barcelona Science Park (Catalonia, Spain) to analyse native and MA-induced hα2M. ASTRA 7 software (Wyatt Technologies) was used for data processing and analysis, for which a dn/dc value typical for proteins (0.185 mL/g) was assumed. All experiments were performed in duplicate.

Thiol quantification

Free thiol groups present in hα2M were determined by reaction of protein samples (at 25.2 mg/mL;156.79 μM of monomer) in buffer D at room temperature for 15 min with Ellman's reagent (5,5'-dithiobis-2-nitrobenzoic acid, DTNB; (Ellman, 1959)), and monitoring the change in absorbance at 412nm (A_{412}) in a microplate spectrophotometer (Power-Wave XS, Biotek). The absorbance signal was measured in 96-well plates containing

220µL reaction volumes (200µL of Ellman's assay reaction plus 20µL of control or test samples) in triplicate. The concentration of free-thiol groups was calculated based on the DTNB molar extinction coefficient (14,150$M^{-1}cm^{-1}$; Riddles *et al.*, 1983), as previously done for α2M samples by others (Steiner *et al.*, 1987). MA-hα2M samples were obtained as described above but a PD-10 column was used for buffer exchange instead of SEC.

Miscellaneous

Protein identity and purity were assessed by 10–15% Tris-Glycine SDS-PAGE stained with Coomassie brilliant blue R250 and using PageRuler Unstained Broad Range Protein Ladder (5–250 kDa) or Unstained Protein Molecular Weight Marker (10–200 kDa), both from Thermo Fisher Scientific, as molecular mass markers. The latter was carried out at the Protein Chemistry Service and the Proteomics Facilities of Biological Research Center (CIB-CSIC) in Madrid, Spain. Native protein samples were also analysed by NuPAGE™ 3 to 8% Tris-Acetate gels (Invitrogen) stained with Coomassie brilliant blue R250 using the NativeMark™ Unstained Protein Standard (20–1200 kDa; Invitrogen) as reference.

Ultrafiltration steps were performed using Vivaspin 15, Vivaspin 2 and Vivaspin 500 filter devices of 50 to 100 kDa cut-off (Sartorius Stedim Biotech). Protein concentrations were estimated by measuring A_{280} in a spectrophotometer (NanoDrop) and applying the respective extinction coefficients ($\varepsilon^{1\%}_{\alpha 2M}$= 9.04; $\varepsilon^{1\%}_{Thermolysin}$=17.65; $\varepsilon^{1\%}_{Trypsin}$= 14.15; $\varepsilon^{1\%}_{\alpha\text{-}casein}$=10.1; $\varepsilon^{1\%}_{Fibrinogen}$= 15.1). Concentrations were also measured by the BCA Protein Assay Kit (Thermo Scientific) with bovine serum albumin as a standard.

Trypsin from bovine pancreas (T1426), thermolysin from *Bacillus thermoproteolyticus* (P1512), bovine milk α-casein (C6780), and fibrinogen from human plasma (F3879) were all purchased from Sigma.

RESULTS

Analysis hα2M structural properties

hα2M circulates exclusively in its native open confirmation as activation of the protein through protease cleavage leads to a closed conformation and the clearance of the complex from the blood. The transition from the open to the closed conformation is a hallmark feature of the hα2M inhibitory mechanism: native hα2M molecules present an open flexible conformation that shifts upon cleavage into a closed arrangement (induced state), thereby trapping the prey peptidase (Luque *et al.*, 2022). Here, we used methylamine (MA) treatment to provoke cleavage of the crucial thioester bond, thereby generating a free thiol group, and inducing the structural rearrangement.

As described by Barrett *et al.* (1979), the native and induced α2M molecules correspond to electrophoretically "slow" and "fast" α2M forms, respectively. This property was used to characterize native and MA-induced α2M (MA-hα2M) purified from both frozen and non-frozen fresh plasma using native PAGE analysis (**Fig.1A**). As expected, native samples (lanes 1 and 3) migrated slower than the MA-induced (lanes 2 and 4) preparations in virtue of their open and extended conformation, the plasma freeze/thaw cycle did not impact their behaviour. Furthermore, native and MA-induced samples were analysed through size-exclusion chromatography combined with multi-angle laser light scattering (SEC-MALLS) (**Fig.1B**). Native samples appeared larger and eluted earlier during gel filtration due to their open conformation when compared with MA-induced protein. More importantly, the SEC-MALLS profiles of native α2M purified from frozen and non-frozen plasma perfectly overlapped, further verifying that the freeze-thaw cycle at the plasma level did not disturb the structural homogeneity of the α2M sample.

Additionally, homogeneity of α2M samples was assessed through quantitative measurement of the free thiol groups, which are only present in α2M-induced molecules after cleavage of the thioester bond, either by surface exposed lysine residues of an entrapped protease, or MA treatment (**Table 1**). Due to the low sensitivity of the assay, hα2M was used at 25.2 mg/mL, corresponding to 156.79 μM of monomeric protein). Notably, the detected average concentration of free thiol groups in the MA-induced sample perfectly corresponded to the protein concentration of hα2M in the assay after correcting for the baseline read-out measured in the native samples, suggesting that all hα2M molecules could be induced by the

MA-treatment. More importantly, no difference was detected between the samples purified from frozen and non-frozen fresh plasma.

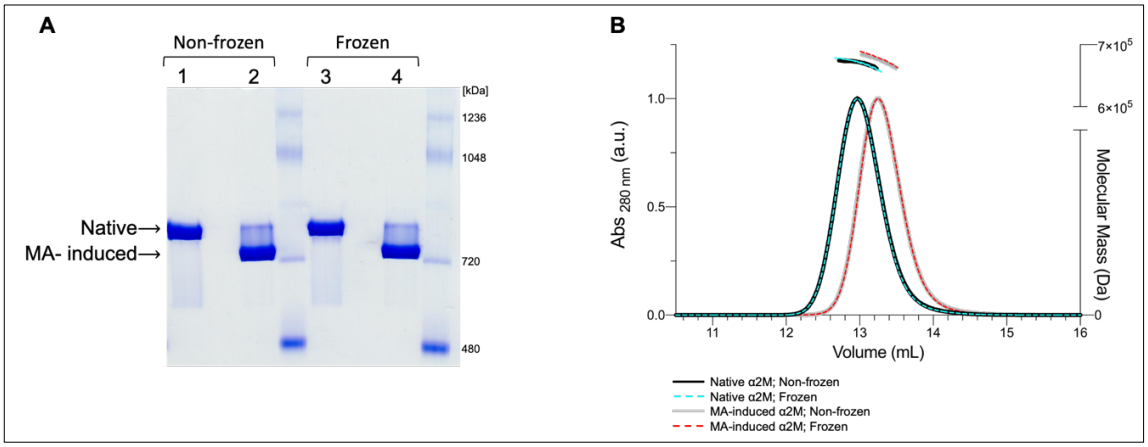


Figure 1: Functional and biophysical studies of α2M preparations. The figures shown are representative for more than three experimental replicates performed. (A) Native PAGE analysis of native (lanes 1 and 3) and MA-induced (lanes 2 and 4) hα2M, and (B) SEC-MALLS of native (black and cyan lines) and MA-induced (grey and red lines) protein purified from frozen (dashed lines) and non-frozen (full lines) plasma. The MALLS determined molecular weights for native (670.8 kDa± 0.36 and 670.7 kDa±0.38) and MA-induced (675.0 kDa± 0.34 and 677.8 kDa±0.34) protein prepared from non-frozen and frozen plasma, respectively, correspond well to the calculated molecular mass of tetrameric hα2M (643.2 kDa) based on ProtParam (Gasteiger et al., 2005) and considering polyglycosylation of the protein. No differences in the two protein preparations could be detected.

Table 1: Concentration of free-thiol groups in hα2M samples. The monomeric hα2M concentration used for the experiment was 156.75 µM. The values presented correspond to the average of three independent experiments ± SD.

	Non-frozen (µM)	**Frozen (µM)**
Native	70.63 ± 23.18	66.19 ± 35.70
MA-induced	212.71 ± 36.25	209.08 ± 12.13

Functional analysis hα2M

Human α2M is a pan-peptidase inhibitor, meaning it can inhibit peptidases irrespective of their catalytic types. Here we evaluated the inhibitory capacity of α2M purified from frozen and non-frozen fresh plasma against the serine peptidase trypsin and the metallopeptidase thermolysin. Activity of treated and non-treated peptidases were analysed towards green-fluorescent BODIPY casein and two natural protein substrates, α-casein and fibrinogen (**Fig.2)**. As induced α2M molecules lost their inhibitory capacity, heterogeneity of native α2M samples (e.g., presence of induced forms) can be studied by assaying the inhibition stoichiometry, considering that one tetrameric hα2M can simultaneously inhibit two medium-sized peptidase molecules such as trypsin and thermolysin (20 to 30 kDa). Thus, we tested the inhibitory capacity at different tetrameric hα2M to peptidase ratios, ranging from 4:1 to 1:4. Prior enzymatic measurement, inhibitor and peptidase were pre-incubated for 10 minutes at room temperature to enable complex formation. The proteolytic activity of the α2M/peptidase sample was then evaluated at 37ºC towards three different substrates.

The inhibitory capacity of α2M against trypsin and thermolysin using BODIPY casein as substrate was calculated after 30 and 10 minutes, respectively. α2M inhibitory capacity was derived from the relative protease activity of α2M-treated sample versus uninhibited peptidase activity. As expected, when α2M/peptidase ratios are 1/2 or higher (e.g., 1/1), there is no significant trypsin activity (**Fig.2A**). However, if only one α2M tetramer is available for four peptidase molecules (1/4 ratio), and considering that the inhibitor can entrap up to two peptidases, two peptidase molecules remain uninhibited and thus will efficiently cleave and deplete the available substrate. Furthermore, α2M covalently entraps the prey peptidases but does not necessarily catalytically inactivate them. Thus, the entrapped peptidases can still cleave smaller substrates capable of accessing the a2M cavity in the closed conformation, and maybe explaining the low residual trypsin activity observed despite excess of inhibitor. Similar results were obtained for the inhibition of thermolysin (**Fig.2B**). As thermolysin showed an overall much higher catalytic efficiency, activities were calculated after a shorter incubation time. Importantly, the inhibitory activity of the α2M samples purified from frozen and non-frozen plasma were almost indistinguishable against both peptidases.

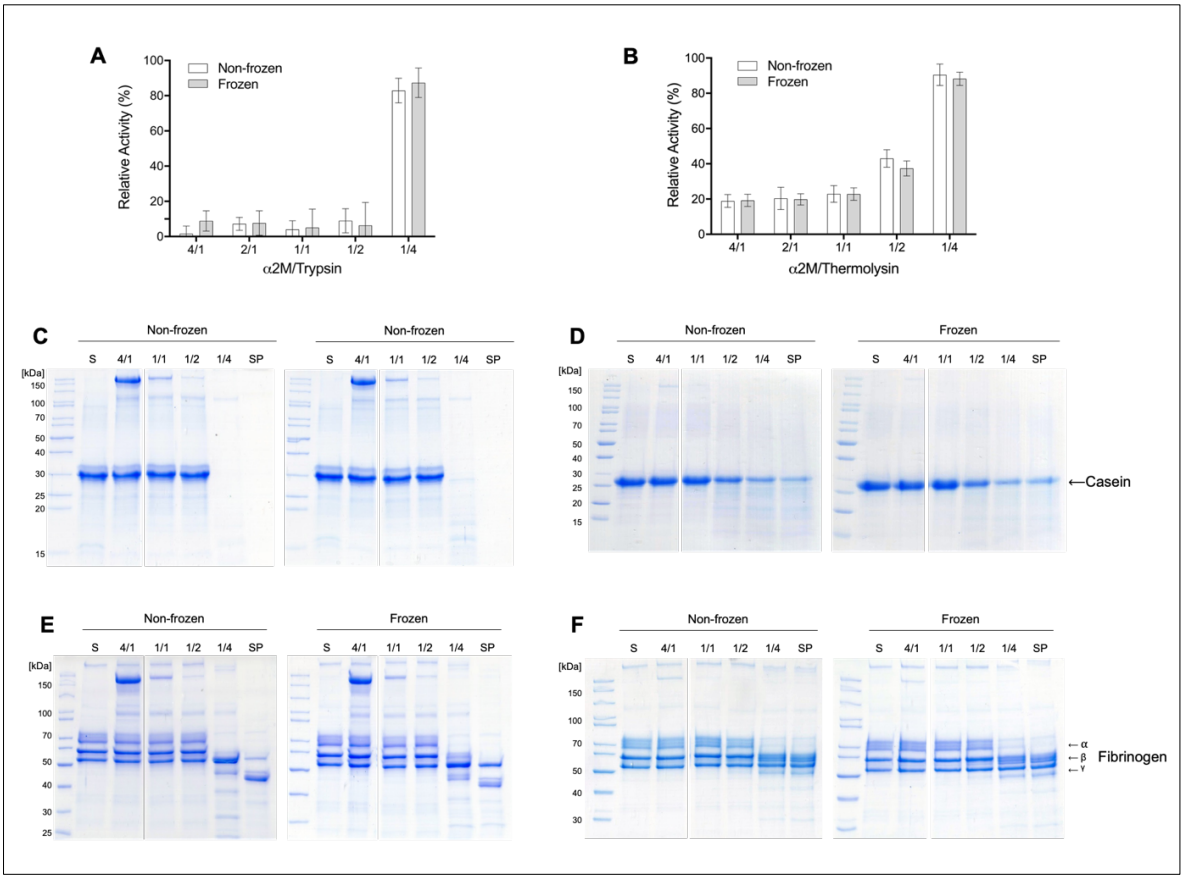


Figure 2: Inhibitory in vitro studies of the α2M-tetramer. Inhibition of (A) trypsin and (B) thermolysin activity against the fluorescence-based BODIPY FL-casein substrate. Percentages of remaining activity are represented as means and standard deviations of three independent experiments. (C to F) SDS-PAGE-based inhibition assay using different ratios of tetrameric hα2M to protease while keeping the protease concentration constant. Inhibition of trypsin activity against (C) α-casein and (E) fibrinogen, and inhibition of thermolysin activity against (D) α-casein and (F) fibrinogen are shown. "S" and "SP" lanes correspond to substrate-alone and substrate incubated with uninhibited protease, respectively.

α2M inhibition of trypsin and thermolysin towards natural substrates was evaluated by SDS-PAGE analysis (**Fig.2C–F**). Inhibition of trypsin activity was evaluated 90 minutes after the addition of the pre-incubated α2M/peptidase mixtures to the α-casein or fibrinogen substrates (**Fig.2C,E**). Uninhibited trypsin degraded all the available α-casein and fibrinogen (**Fig.2C,E**; lanes *SP*). Complete α-casein degradation was also observed for the 1/4 ratio, but not in the samples with sufficiently high inhibitor concentrations. The same pattern was observed for fibrinogen degradation. In the 1/4 ratio, α2M inhibited approximately half of the peptidase activity, thus slowing down substrate degradation ─ evident by the presence of a thicker fibrinogen γ-chain band.

The shorter incubation times and lower amounts of peptidase used for the thermolysin-based assay (**Fig.2D,F**) resulted in incomplete substrate degradation for

uninhibited protease, as seen by fading of the α-casein band and incomplete degradation of the fibrinogen β- and γ-chain. As expected, addition of the α2M tetramer to thermolysin in a 1/4 ratio is not enough to prevent proteolytic cleavage of the substrates (**Fig.2D,F**). Intriguingly, thermolysin partially degraded α-casein but not fibrinogen in the presence of α2M at a 1/2 ratio. This might originate either from (i) inaccuracies during protein concentration measurements, (ii) minor pipetting errors, or most likely, (iii), better access of α-casein to the α2M-cage due to its lower MW compared to fibrinogen. Noteworthy, subtle residual activity could be detected in all scenarios, independent of the source of the α2M preparation, and overall, both α2M preparations showed highly comparable inhibitory profiles.

General Discussion

The present PhD thesis focused on the structural and functional analysis of proteins implicated in the regulation of distinct proteases, namely RECK, IMPI and hα2M. The results obtained significantly advance our knowledge about these key molecules, their roles in distinct (patho)physiological processes, and their intricate mechanisms of action. The study of these protease regulators and their respective protease targets was summarized in three separated chapters (Projects 1, 2 and 3) and their discussions are elaborated below.

D1: "*Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases"*

The first project of the thesis aimed to study the inhibitory capacity of RECK, the reversion-inducing cysteine-rich protein with Kazal motifs, a protein essential for embryogenesis and tumour progression, and described to act as an MMP inhibitor. Reduced RECK levels have been associated with higher tumour invasiveness and metastasis in various tumours. Therefore, RECK has been suggested by several research groups as a useful diagnostic and prognostic marker and even as a potential therapeutic target. These studies motivated us to further investigate the structure of this suggested MMP inhibitor as well as its mechanism of action.

D 1.1: Preparation of protein samples

According to the literature available at the beginning of this project, secreted and membrane-released RECK protein (e.g., after proteolytic removal of the C-terminal GPI anchoring sequence and consequent release of the soluble ectodomain to the extracellular space, and hence termed RECKΔC), along with a RECK variant comprising only the Kazal-like (KL) domains KL2 and KL3, had reported inhibitory activity against MMPs *in vitro* (Chang *et al.*, 2008; Miki *et al.*, 2007; Muraguchi *et al.*, 2007; Oh *et al.*, 2001; Omura *et al.*, 2009; Takahashi *et al.*, 1998). Since there were severe discrepancies in the defined boundaries of the KL domains in those studies (Chang *et al.*, 2008; Takahashi *et al.*, 1998), we performed a structure prediction using RaptorX of the V^{600}–A^{800} RECK segment, concluding that KL123 and KL23 variants must span segments V^{621}–S^{797} and T^{697}–S^{797}, respectively. Both variants, containing a C-terminal His_6-tag, were produced periplasmatically in *E. coli* Lemo21 cells, where the oxidizing environment allows proper

disulphide-bond formation and protein folding. I obtained high yields of non-classical inclusion bodies for both proteins, as previously described by Jevševar *et al.* (2008), which were treated with a chaotropic agent and a detergent under non-reducing conditions and further purified by affinity chromatography (AC) and size exclusion chromatography (SEC). The obtained proteins were highly pure, soluble, did not aggregate when concentrated and behaved as monomers in gel filtration (based on calibrated SEC runs, KL123 and KL23 migrated according to 26 and 16 kDa, respectively; data not shown), indicating that the proteins were well folded.

RECKΔC (G^{27}–S^{942}) with a C-terminal His_6-tag was purified from the conditioned medium of Expi293 cells transiently transfected according to a protocol previously established in the lab (Marino-Puertas *et al.*, 2019). The protein was purified by AC and SEC, with a final yield of 0.8 mg per litre of expression. Pure RECKΔC analysed by SEC-MALLS showed a molecular mass of 111 kDa, which is in agreement with the theoretical protein mass (101.9 kDa, according to Expasy ProtParam (Gasteiger *et al.*, 2005) plus glycosylation (5 N-glycosylation sites are reported) and indicates that also the full-length protein is monomeric. This diverges from other studies reporting that RECK is a dimer (Omura *et al.*, 2009). Additionally, RECKΔC was also obtained from S2 and ExpiCHO expression systems but with lower yields (0.2 and 0.5mg/L expression, respectively), and requiring a more laborious purification protocol (data not shown). Lastly, RECK variant FRAG-1 was produced by limited proteolysis of RECKΔC with the MMP-14 catalytic domain (CD). FRAG-1, corresponding to the C-terminal part of full-length protein (I^{486}–S^{942} and thus containing all KL domains, was further purified by SEC.

A MMP-14 CD purification protocol was adapted from (Itoh, 2001). MMP-14 CD was produced in *E. coli* BL21(DE3) cells as inclusion bodies and purified by IEC under denaturing conditions, refolded under dialysis, and further purified by SEC.

D 1.2: Proteolytic contamination and additional purification steps

After purification by AC and SEC, recombinant RECKΔC produced from Expi293 cells presented high purity (>98%). Nonetheless, it underwent cleavage over time, which could be (i) prevented by the addition of the EDTA-free Roche cOmplete inhibitor cocktail free, (ii) slowed down by o-phenanthroline, a general inhibitor of zinc-metallopeptidases, and (iii) speeded up by storage at 37ºC. The storage of RECKΔC sample at 4ºC prevented proteolysis in the short term (2 to 3 days), as confirmed by SEC. The cleavage products caused by this protease impurity were analysed by N-terminal sequencing using Edman

degradation, which revealed that the C-terminal fragment of approximately 50 kDa results from a cleavage before the G^{484} comprising RECK KL1-KL3 domains. Notably, this cleavage occurs right before the one we identified for MMP-14 CD. Thus, we termed the fragment likewise FRAG-1. Since partially purified RECKΔC had shown in initial assays slight inhibition of MMP-14 CD (data not shown), we hypothesized whether RECKΔC cleavage is essential to yield a species with MMP inhibitory activity. Therefore, FRAG-1 derived from the limited proteolytic cleavage at I^{486} by MMP-14 CD was included in subsequent inhibitory assays.

To unveil the nature of the protease contaminant, I incubated partially purified RECKΔC with human plasma fibrinogen, and indeed, degradation of this general peptidase substrate was observed. Next, I tested various peptidase inhibitors and could show that AEBSF fully abolished fibrinogen cleavage. AEBSF is an irreversible inhibitor of serine peptidases, which covalently reacts with the hydroxyl group of the active-site serine and thus inactivates the enzyme. Similar inhibition was observed using the Roche cOmplete EDTA-free inhibitor cocktail, which likewise includes AEBSF but lacks general metallopeptidase inhibitors. Thus, I adapted the RECKΔC purification protocol by including an incubation step with AEBSF prior the final polishing by SEC. These RECKΔC samples lacking fibrinolytic activity were subsequently used for inhibitory studies. Regarding FRAG-1 derived from RECKΔC processed with MMP-14, these preparations did not contain the proteolytic activity, confirming the complete removal of MMP-14 CD by SEC.

Interestingly, RECKΔC produced from the insect and CHO-based expression systems did not contain this protease contaminant, as evident by the absence of fibrinolytic activity in these RECK preparations. To check if this AEBSF-sensitive protease contaminant is specific to RECKΔC overexpression, I tested also another protein produced in the same expression system at that time, i.e. the N-terminal fragment of the proteoglycan testican-3 (N-TES), a calcium binding protein which has been suggested to be an MMP-14 and MMP-16 inhibitor (Nakada *et al.*, 2001). Notably, we detected a similar proteolytic activity towards fibrinogen, which again could be eradicated by AEBSF and an AEBSF-containing inhibitor cocktail. However, two proteases (ADAMTS-13 and neprosin) studied in our laboratory, obtained from the same expression system and similarly purified did not contain this peptidolytic contaminant (data not shown).

D 1.3: Inhibition studies of RECK

Highly pure RECK preparations (RECKΔC, FRAG-1, KL123 and KL23) were assessed for their inhibitory activity against MMP-2, MMP-7, MMP-9 and MMP-14 CDs using both peptide and protein substrates with up to 100-fold molar excess of the putative inhibitor RECK. Fluoresceine-conjugated gelatin was employed for assaying APMA-activated MMP-2, MMP-7, and MMP-9, while MMP-14 CD activity was tested against the fluorogenic peptide FS-6. Additionally, inhibition of MMP-2 and MMP-7 activity against plasma fibronectin was assayed at tenfold molar excess of RECKΔC by Western blot. Importantly, none of the RECK variants or BSA, which was used as a negative control, displayed proteolytic activity on their own.

However, while o-phenanthroline showed concentration-dependent MMP inhibition, BSA and, strikingly, none of the RECK variants presented any inhibitory capacity against the tested MMPs. Of course, we cannot exclude indirect effects between RECK and MMPs, but based on our data and highly purified protein preparations, despite of RECK cleavage, no interaction between the two proteins could be observed, indicating that RECK is not a direct inhibitor of MMP activity.

D 1.4: Crystallographic study of RECK protein

Highly pure RECKΔC obtained from S2 and Expi293 expression systems was used to set initial crystallisation plates, at both 4ºC and 20ºC, in vapour-diffusion sitting-drop format. We tested both preparations as insect and mammalian glycosylation pathways differ, and as glycosylation can strongly impact crystallisation. We did not succeed in crystallizing RECKΔC. Notably, the currently available predicted model by *AlphaFold* presents various loops with low confidence score, which suggest structural flexibility. Moreover, due to its size and composition of several domains, interdomain flexibility may further hamper crystallisation.

Finally, as I did not manage to crystallize RECKΔC and motivated by the work by Omura *et al.* (2009), we tried to tackle the RECK structure also through cryo-electron microscopy (cryo-EM). We prepared negative staining grids, in which a full field of small proteins could be detected. Despite the negative stanning results were not promising, we attempted to vitrify some grids. Unfortunately, the vitrified grids did not present enough quality to proceed further.

D2: *"An engineered protein-based submicromolar competitive inhibitor of the Staphylococcus aureus virulence factor aureolysin"*

The second project of this thesis focused on the study of the insect metallopeptidase inhibitor (IMPI), a potential inhibitor of the major *S. aureus* virulence factor aureolysin. The advent of antibiotic resistant *S. aureus* strains requires the development of new therapeutic molecules that target those pathogens. Here, we tested the inhibitory capacity of the M4-specific inhibitor IMPI against aureolysin and compared it with the inhibitory efficiency of a cohort of IMPI variants.

D 2.1: Assessment of wild-type IMPI as an aureolysin inhibitor and initial protein redesign

The mature form of wild-type IMPI (I^{20}–S^{88}) was expressed with an N-terminal and TEV-cleavable His6-thioredoxin tag using *E. coli* BL21(DE3) Origami2 cells, while aureolysin was isolated from *S. aureus* (V8 BC-10 strain) cultures. Both proteins were extensively purified to homogeneity.

To determine the effect of IMPI on aureolysin activity, the inhibitor was evaluated at distinct molar ratios ranging from 1:1 to 1:200 using both a protein and peptide substrate. Thermolysin, the prototypical member of the M4 family of metallopeptidases, and ulilysin, a member of the pappalysin family M43 (Huesgen *et al.*, 2015; Tallant *et al.*, 2006), were used as controls. As expected, IMPI efficiently inhibited thermolysin but did not impact ulilysin activity, which is in agreement with its role as a specific inhibitor of M4 metalloproteases. Importantly, aureolysin was likewise inhibited in a dose-dependent manner, especially when using the peptide substrate, but not as efficiently as thermolysin.

The analysis of the superimposed structures of unbound aureolysin (Banbula *et al.*, 1998) with thermolysin in complex with wild-type IMPI (Arolas *et al.*, 2011) made us hypothesize that replacement of the main specificity determinant of IMPI, P1' residue I^{57} with a bulkier residue could potentially lead to stronger inhibition of aureolysin.

Therefore, I produced the IMPI $I^{57}F$ mutant as described previously for the wild-type protein and used it for further analysis.

D 2.2: Overall structure of the IMPI-aureolysin complex

I successfully crystallised aureolysin in complex with wild-type IMPI and the IMPI $I^{57}F$ variant in the tetragonal crystallographic space group $P4_1$. Both crystal structures contained two complexes per asymmetric unit (a.u.) and were solved by molecular replacement. Superimposition of the two structures revealed that they were equivalent.

IMPI presents a spearhead shape whose tip is formed by a "reactive-site bond" (RSB; N^{56}–I^{57}) within the "reactive-centre loop" (RCL; spanning from C^{52} to C^{59}). IMPI is structurally rigid owing to its five disulphide bonds and it exhibits four β-strands and one α-helix. The structure of mature aureolysin in its unbound form was previously described by Banbula *et al.* (1998). It complies with that of M4 family members, so its N-terminal subdomain (NSD; A^{209}-A^{363}) is mostly constituted by β-strands but also comprises a so-called "backing helix" and an "active-site helix". The latter resides at the interface of the NSD and the C-terminal subdomain (CSD), and contains the **H**^{352}**E**xx**H** motif characteristic for metallopeptidases. The CSD on the other hand provides the third proteinaceous zinc ligand, specifically glutamate E^{376}, from its gluzincin helix. Furthermore, glutamate E^{352} acts as a general acid/base for the cleavage reaction (Bode *et al.*, 1993; McKerrow, 1987). The CSD (N^{364} to E^{509}) is mainly constituted by α-helices. It starts with the "glutamate helix" typical for gluzincins ("gluzincin helix") and is followed by a long irregular segment, which encompasses three calcium ions important for protein stabilization. Furthermore, this segment shapes the bottom of the active-site cleft on its primed side, including the hydrophobic S1' pocket, accountable for the substrate specificity of aureolysin and other thermolysin-like metallopeptidases (Gomis-Rüth *et al.*, 2012). Aureolysin and thermolysin differ particularly in two elements: while aureolysin presents a flap in the NSD (N_{312}–N_{321}) protruding from the surface above the active-site cleft, thermolysin exhibits a salient β-ribbon overhang on the CSD surface.

In agreement with the previously described IMPI-thermolysin complex, we observed that the IMPI variants inserted like a wedge into the aureolysin active-site cleft. The main interaction interface is formed between the RCL and the flanking scaffold loop of IMPI with the protease flap and the substrate binding pockets of aureolysin. Notably, while the N- and C-terminal subdomains of unbound thermolysin undergo a relative 5º-rotation upon substrate binding (Hausrath & Matthews, 2002), the superimposed structures of unbound (Banbula *et al.*, 1998) and IMPI-bound aureolysin demonstrated only minor differences in the orientation of the aureolysin subdomains. This indicates that aureolysin, contrary to

thermolysin, *P. aeruginosa* elastase, and *B. cereus* neutral proteinase (Banbula *et al.*, 1998), does not present a closing hinge motion upon substrate or inhibitor binding.

D 2.3: IMPI inhibits aureolysin via the standard mechanism

The IMPI RCL enters the aureolysin active-site cleft by occupying the S4-S1' pockets through residues P^{53}-I/F^{57}. We observed that the IMPI reactive-site bond was cleaved at $N^{56}\downarrow$I/F^{57} in the crystals and confirmed this cleavage *in vitro* through incubation of both wild-type IMPI and I^{57}F-IMPI with aureolysin. We showed that IMPI inhibits aureolysin through the so-called "standard mechanism of inhibition" typical of serine peptidase inhibitors (Bode & Huber, 1992; Laskowski & Kato, 1980), however differing from the latter by exhibiting a more expeditious cleavage speed(Ascenzi *et al.*, 2003). These results confirm the data obtained for the IMPI complex with thermolysin (Arolas *et al.*, 2011) and identify IMPI as a unique inhibitor, which despite being cleaved maintains its rigid overall structure, likely due to its five disulphide. Importantly, as in the thermolysin complex, the cleaved RSB of IMPI is positioned for re-joining also in the complex with aureolysin, as required in the standard mechanism of inhibition (Laskowski & Qasim, 2000).

D 2.4: IMPI redesign

The analysis of the aureolysin-IMPI complex crystal structures determined here revealed five RCL positions (namely T^{50}, I^{54}, I^{55}, I^{57} and R^{58}) ideal for mutagenesis and modulating inhibitor specificity. Consequently, in addition to wild-type and I^{57}F-IMPI, I prepared nine single (T^{50}Q, T^{50}R, T^{50}Y, I^{54}M, I^{55}R, I^{55}W, I^{55}Y, I^{57}Y, R^{58}E) mutants, one double mutant (T^{50}Y+I^{55}R), and one triple mutant (T^{50}Y+I^{55}R+I^{57}F). All IMPI variants were expressed and purified as described for the wild type. The inhibitory efficiency of the IMPI variants against aureolysin at molar ratios of 1:1 to 1:100 was evaluated towards the fluorogenic peptide FRET4 (Abz-Y-G-K-R-V-F-K[dpn]–OH) and compared with the wild type. Notably, the R^{58}E mutant did not inhibit aureolysin at all, although its behaviour during purification and its inhibition of thermolysin (although approximately 200-fold weaker than wild-type IMPI) indicated that it was properly folded although functionally impaired to inhibit M4 peptidases. All other variants presented a concentration-dependent inhibition of aureolysin, some displaying residual activity values comparable to wild-type IMPI (3 to 8% at the highest molar ratio) and others showing weaker inhibition (17% to 58%). The *Ki* values of the variants were derived from calculated IC_{50} values. Notably, the *Ki* values of the IMPI

variants with strongest inhibitory activity were in the low micromolar range (346 to 644 nM) while those of the weaker variants were micromolar (1220 to 4520 nM). IMPI variant $I^{57}F$ presented the highest inhibition among all constructs evaluated (*Ki* = 346nM), thus providing a useful basis for the further development of IMPI-based aureolysin inhibitors.

D3: *"α2M samples purified from frozen and unfrozen fresh plasma present no significant structural or functional differences"*

The third project of this thesis represents a highly collaborative work developed by researchers of complementary backgrounds and aimed to dissect the molecular mechanism of action of the major protein inhibitor of human blood plasma, hα2M. Reviewer comments during the publishing process of our hα2M structures paper in *Proc. Natl. Acad. Sci. USA* (Luque *et al.*, 2022) compelled us to further investigate and compare the quality of our hα2M samples purified from fresh frozen plasma with samples obtained from fresh non-frozen plasma. While it is well accepted in the field that freeze-thaw of purified hα2M protein preparations is detrimental to its inhibitory function, we believe (like many others) that the complex composition of plasma is protective during freeze/thaw and that thus hα2M from frozen fresh plasma is fully functional. To test this, we prepared several experiments that put the structural and functional heterogeneity of differently purified hα2M to the test.

D3.1: Assessment of structural and functional heterogeneity of hα2M samples

The fresh non-frozen plasma samples used in this study were kept at 4º C and used for protein purification within twenty-four hours of blood donation. For the frozen plasma, the freezing method was designed to mimic the freezing protocol at the blood bank, which allows aseptic manipulation and quick sample freezing (-30 ºC in 1h). To do so, the plasma samples were aliquoted in 50 mL tubes at half capacity and placed at -20 ºC with an inclination angle of 60–50º to extend the freezing surface while avoiding contact of the sample with the tube lid. Complete plasma freezing was achieved after two hours.

Native tetrameric hα2M is found in an open and flexible conformation in human plasma, ready to accommodate its peptidase targets. Its inhibitory mechanism relies on a structural rearrangement into a closed conformation, in which the prey peptidases are

entrapped upon inhibitor cleavage and thus no longer have access to their substrates. Nearly simultaneously, surface-located lysine residues of the trapped peptidase attack a prominent and due to the structural rearrangement now accessible thioester bond in hα2M, and thus, a covalent complex between the peptidase and inhibitor is formed. A similar closed conformation can be obtained by incubation of native hα2M with the small nucleophile methylamine (MA), which can diffuse into the hα2M central lumen and directly attack the thioester bond, thereby inducing the aforementioned structural rearrangements, which ultimately lead to the exposure of the recognition binding domain (RBD). Activated hα2M is efficiently removed from the circulation, and thus hα2M purified from human plasma contains almost exclusively molecules in their native form. However, during protein manipulation, hα2M induction might occur and provoke sample heterogeneity, as induced α2M is incapable of entrapping and inhibiting target peptidases, thus impairing the expected inhibition stoichiometry.

Altogether, our experiments demonstrated that α2M samples purified from frozen and non-frozen fresh plasma (i) have identical electrophoretic migration patterns when analysed by Native PAGE, (ii) show indistinguishable UV absorbance profiles and molar mass distributions in SEC-MALLS, and (iii) exhibit equally low concentrations of free-thiol groups quantified using Ellman's reagent. Furthermore, using three different substrates, no differences in inhibitory capacity were detected against the serine peptidase trypsin or the metallopeptidase thermolysin. Thus, we conclude that ha2M preparations from frozen versus non-frozen fresh plasma are indeed identical and that a single plasma freeze-thaw cycle does not detrimentally affect homogeneity and functionality of the α2M preparation, rendering fresh-frozen plasma (FFP) an excellent source of native hα2M.

Conclusions

Project 1: *"Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases"*

- We established a total of four expression systems for the production of highly-pure RECK variants. The main form of interest, RECKΔC, which represents full-length RECK lacking the C-terminal GPI anchoring signal, was produced in one insect and two mammalian expression systems with high yields. Additionally, RECK K123 and K23 were efficiently produced periplasmatically in *E. coli cells*.
- RECK fragments K123 and K23 were produced in non-classical inclusion-bodies, yielding highly pure samples after mild treatment with a chaotropic agent and detergent and further affinity and size exclusion chromatography purification steps.
- Notably, RECKΔC samples prepared from Expi293F cells, but not from S2 or ExpiCHO cells, presented a contaminant with proteolytic activity. This background activity was sensitive to AEBSF, a serine peptidase inhibitor. Consequently, we added an AEBSF incubation step to our RECKΔC purification protocol prior size exclusion chromatography, yielding crystallisation grade protein preparations without detectable proteolytic activity.
- Cleavage of partially purified RECKΔC by the contaminant peptidase generated three main fragments which were analysed by N-terminal Edman degradation. Only the larger fragment of approximately 50 kDa could be identified and pointed to the C-terminal half of RECKΔC containing the KL1-KL3 domains. Thus, we tested whether this proteolytic processing of RECKΔC could be a prerequisite for its full inhibitory activity. To test this hypothesis, we prepared a RECK variant termed FRAG1 (I^{486}–S^{942}) through limited proteolysis with MMP-14 catalytic domain for further use in subsequent inhibitory experiments.
- The inhibitory capacity of RECK variants was tested against various MMPs (MMP-2, MMP-7, MMP-9 and MMP-14 catalytic domain). MMP activity was analysed using (quenched) fluorescent substrates, namely DQ gelatin and FS-6 and Western blotting to analyse cleavage of the natural MMP substrate human plasma fibronectin (pFN).
- Strikingly, none of the analysed RECK variants presented MMP inhibition, indicating that RECK is not a direct inhibitor of MMPs activity. Furthermore, our hypothesis, that RECK is initially translated as an inactive proform that needs proteolytic activation to render its Kazal domains inhibitory active, could not be verified. Thus, despite the

reported molecular interplay between RECK and MMPs published by others, our data clearly show that RECK is not a direct inhibition of MMPs activity.

Project 2: *"An engineered protein-based submicromolar competitive inhibitor of the Staphylococcus aureus virulence factor aureolysin"*

- IMPI is a small protein-based inhibitor specific of thermolysin-like proteases (MEROPS family M4). Here we analysed the inhibitory capacity of wild-type IMPI recombinantly produced in *E. coli* against aureolysin, a M4 TLP and key virulence factor produced by the pathogen *S. aureus*. We observed that wild-type IMPI indeed inhibits aureolysin in a dose-dependent manner. However, its inhibitory efficiency was lower than against thermolysin.
- Analysis of the superimposed structures of aureolysin and thermolysin complexed with IMPI suggested that replacement of residue I^{57} by a bulkier residue might improve aureolysin inhibition. Therefore, we produced and purified IMPI $I^{57}F$ mutant in similar fashion as wild-type IMPI.
- We crystallised both wild-type IMPI and the IMPI $I^{57}F$ variant in complexes with aureolysin, and solved the complex structures, which were almost identical, by molecular replacement. As described previously by our lab for the IMPI-thermolysin complex, both IMPI variants were inserted wedge-like into aureolysin active-site cleft, with I/F^{57} residing at the tip. Interestingly, the comparison of our complex structures with the available structure of unbound aureolysin revealed that the relative orientation of its N- and C-terminal subdomains remains unaltered upon inhibitor binding, while for thermolysin and other members of the M4 family, a distinct hinge bending motion is described.
- Additionally, our crystal structures of IMPI in their aureolysin complexes revealed that IMPI follows the standard mechanism of inhibition, which is typically associated with inhibitors of serine peptidases. This means that the inhibitor is bound in substrate-like manner, but proteolytic turnover is extremely slow, thus rendering the substrate an inhibitor. Furthermore, as the IMPI overall structure is heavily stabilized by five disulphide bonds evenly distributed across the structure, protease cleavage does not impact the structure. This is in perfect agreement with what was reported by Arolas *et*

al. (2011) for the complex of wild-type IMPI and thermolysin. Analysis of the protein-protein interface between IMPI and aureolysin in our complex structures revealed five positions within the IMPI reactive-centre loop that could but mutated to yield a stronger or more specific inhibitor of aureolysin. We prepared a total of eleven IMPI variants through single, double, and triple mutations, and compared them with wild-type IMPI. Notably, only one of them, $R^{58}E$, could not inhibit aureolysin. All others could be divided into two major groups: one with inhibitory activity similar to wt-IMPI (i.e., with *Ki* values ranging from 346 to 644 nM), and one displaying weaker inhibition (i.e., Ki values between 1220 and 4520 nM). Among all IMPI variants, $I^{57}F$-IMPI revealed the highest inhibitory capacity against aureolysin (*Ki*=346nM), and thus represents a good candidate for further development of IMPI-based inhibitors for the key S. aureus virulence factor aureolysin.

Project 3: *"α2M samples purified from frozen and unfrozen fresh plasma present no significant structural or functional differences"*

- We purified α2M from non-frozen and frozen fresh plasma and compared the structural and functional homogeneity of both preparations through a variety of experiments. The biophysical properties were analysed through (i) Native-PAGE, (ii) SEC-MALLS, and (iii) free-thiol content using Ellman's reagent. Furthermore, their functional properties were assessed through inhibitory assays against the serine peptidase trypsin and the metallopeptidase thermolysin using different substrates and both quenched fluorescent and SDS-PAGE-based cleavage assays. None of the experiments performed revealed any significant difference between the analysed α2M preparations, indicating that frozen-fresh plasma (FFP) is an excellent source of native α2M, and that freezing in the context of plasma (but not necessarily as purified protein) does not trigger detrimental activation of α2M

References

Adekoya, O. A., & Sylte, I. (2009). The Thermolysin Family (M4) of Enzymes: Therapeutic and Biotechnological Potential. *Chemical Biology & Drug Design*, *73*(1), 7–16. https://doi.org/10.1111/j.1747-0285.2008.00757.x

Ahmad-Mansour, N., Loubet, P., Pouget, C., Dunyach-Remy, C., Sotto, A., Lavigne, J.-P., & Molle, V. (2021). Staphylococcus aureus Toxins: An Update on Their Pathogenic Properties and Potential Treatments. *Toxins*, *13*(10), 677. https://doi.org/10.3390/toxins13100677

Ahokas, K., Lohi, J., Lohi, H., Elomaa, O., Karjalainen-Lindsberg, M.-L., Kere, J., & Saarialho-Kere, U. (2002). Matrix metalloproteinase-21, the human orthologue for XMMP, is expressed during fetal development and in cancer. *Gene*, *301*(1–2), 31–41. https://doi.org/10.1016/S0378-1119(02)01088-0

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). Protein Function. In *Molecular Biology of the Cell* (4th ed.). Garland Science.

Altincicek, B., Berisha, A., Mukherjee, K., Spengler, B., Römpp, A., & Vilcinskas, A. (2009). Identification of collagen IV derived danger/alarm signals in insect immunity by nanoLC-FTICR MS. *Bchm*, *390*(12), 1303–1311. https://doi.org/10.1515/BC.2009.128

Altincicek, B., & Vilcinskas, A. (2006). Metamorphosis and collagen-IV-fragments stimulate innate immune response in the greater wax moth, Galleria mellonella. *Developmental & Comparative Immunology*, *30*(12), 1108–1118. https://doi.org/10.1016/j.dci.2006.03.002

Altincicek, B., & Vilcinskas, A. (2008). Identification of a lepidopteran matrix metalloproteinase with dual roles in metamorphosis and innate immunity. *Developmental & Comparative Immunology*, *32*(4), 400–409. https://doi.org/10.1016/j.dci.2007.08.001

Andus, T., Gross, V., Tran-Thi, T.-A., Schreiber, G., Nagashima, M., & Heinrich, P. C. (1983). The Biosynthesis of Acute-Phase Proteins in Primary Cultures of Rat Hepatocytes. *European Journal of Biochemistry*, *133*(3), 561–571. https://doi.org/10.1111/j.1432-1033.1983.tb07500.x

Arimura, Y., & Funabiki, H. (2022). Structural Mechanics of the Alpha-2-Macroglobulin Transformation. *Journal of Molecular Biology*, *434*(5), 167413. https://doi.org/10.1016/j.jmb.2021.167413

Armstrong, P. B. (2006). Proteases and protease inhibitors: a balance of activities in host–pathogen interaction. *Immunobiology*, *211*(4), 263–281. https://doi.org/10.1016/j.imbio.2006.01.002

Arolas, J. L., Botelho, T. O., Vilcinskas, A., & Gomis-Rüth, F. X. (2011). Structural Evidence for Standard-Mechanism Inhibition in Metallopeptidases from a Complex Poised to Resynthesize a Peptide Bond. *Angewandte Chemie International Edition*, *50*(44), 10357–10360. https://doi.org/10.1002/anie.201103262

Arvidson, S. (1973). Studies on extracellular proteolytic enzymes from Staphylococcus aureus. *Biochimica et Biophysica Acta (BBA) - Enzymology*, *302*(1), 149–157. https://doi.org/10.1016/0005-2744(73)90017-X

Asai, M., Sheehan, G., Li, Y., Robertson, B. D., Kavanagh, K., Langford, P. R., & Newton, S. M. (2021). Innate Immune Responses of Galleria mellonella to Mycobacterium bovis BCG Challenge Identified Using Proteomic and Molecular Approaches. *Frontiers in Cellular and Infection Microbiology*, *11*. https://doi.org/10.3389/fcimb.2021.619981

Ascenzi, P., Bocedi, A., Bolognesi, M., Spallarossa, A., Coletta, M., Cristofaro, R., & Menegatti, E. (2003). The Bovine Basic Pancreatic Trypsin Inhibitor (Kunitz Inhibitor): A Milestone Protein. *Current Protein & Peptide Science*, *4*(3), 231–251. https://doi.org/10.2174/1389203033487180

Banbula, A., Potempa, J., Travis, J., Fernandez-Catalén, C., Mann, K., Huber, R., Bode, W., & Medrano, F. (1998). Amino-acid sequence and three-dimensional structure of the Staphylococcus aureus metalloproteinase at 1.72 å resolution. *Structure*, *6*(9), 1185–1193. https://doi.org/10.1016/S0969-2126(98)00118-X

Barrett, A. J. (1994). Classification of peptidases. In *Methods in Enzymology* (Vol. 244, pp. 1–15). https://doi.org/10.1016/0076-6879(94)44003-4

Barrett, A. J., Brown, M. A., & Sayers, C. A. (1979). The electrophoretically 'slow' and 'fast' forms of the α2-macroglobulin molecule. *Biochemical Journal*, *181*(2), 401–418. https://doi.org/10.1042/bj1810401

Barrett, A. J., & Rawlings, N. D. (2007). 'Species' of peptidases. *Biological Chemistry*, *388*(11). https://doi.org/10.1515/BC.2007.151

Barrett, A. J., & Starkey, P. M. (1973). The interaction of α2-macroglobulin with proteinases. Characteristics and specificity of the reaction, and a hypothesis concerning its molecular mechanism. *Biochemical Journal*, *133*(4), 709–724. https://doi.org/10.1042/bj1330709

Bergin, D., Murphy, L., Keenan, J., Clynes, M., & Kavanagh, K. (2006). Pre-exposure to yeast protects larvae of Galleria mellonella from a subsequent lethal infection by Candida albicans and is mediated by the increased expression of antimicrobial peptides. *Microbes and Infection*, *8*(8), 2105–2112. https://doi.org/10.1016/j.micinf.2006.03.005

Bock, P. E., Panizzi, P., & Verhamme, I. M. A. (2007). Exosites in the substrate specificity of blood coagulation reactions. *Journal of Thrombosis and Haemostasis*, *5*, 81–94. https://doi.org/10.1111/j.1538-7836.2007.02496.x

Bode, W., Fernandez-Catalan, C., Grams, F., Gomis-Ruth, F.-X., Nagase, H., Tschesche, H., & MASKOS, K. (1999). Insights into MMP-TIMP Interactions. *Annals of the New York Academy of Sciences*, *878*(1 INHIBITION OF), 73–91. https://doi.org/10.1111/j.1749-6632.1999.tb07675.x

Bode, W., Gomis-Rüth, F.-X., & Stöckler, W. (1993). Astacins, serralysins, snake venom and matrix metalloproteinases exhibit identical zinc-binding environments (HEXXHXXGXXH and Met-turn) and topologies and should be grouped into a common family, the 'metzincins.' *FEBS Letters*, *331*(1–2), 134–140. https://doi.org/10.1016/0014-5793(93)80312-I

Bode, W., & Huber, R. (1992). Natural protein proteinase inhibitors and their interaction with proteinases. *European Journal of Biochemistry*, *204*(2), 433–451. https://doi.org/10.1111/j.1432-1033.1992.tb16654.x

Boname, J. M., Bloor, S., Wandel, M. P., Nathan, J. A., Antrobus, R., Dingwell, K. S., Thurston, T. L., Smith, D. L., Smith, J. C., Randow, F., & Lehner, P. J. (2014). Cleavage by signal peptide peptidase is required for the degradation of selected tail-anchored proteins. *Journal of Cell Biology*, *205*(6), 847–862. https://doi.org/10.1083/jcb.201312009

Boon, L., Ugarte-Berzal, E., Vandooren, J., & Opdenakker, G. (2016). Glycosylation of matrix metalloproteases and tissue inhibitors: present state, challenges and opportunities. *The Biochemical Journal*, *473*(11), 1471–1482. https://doi.org/10.1042/BJ20151154

Botelho, T. O., Guevara, T., Marrero, A., Arêde, P., Fluxà, V. S., Reymond, J.-L., Oliveira, D. C., & Gomis-Rüth, F. X. (2011). Structural and Functional Analyses Reveal That Staphylococcus aureus Antibiotic Resistance Factor HmrA Is a Zinc-dependent Endopeptidase. *Journal of Biological Chemistry*, *286*(29), 25697–25709. https://doi.org/10.1074/jbc.M111.247437

Buresova, V., Hajdusek, O., Franta, Z., Sojka, D., & Kopacek, P. (2009). IrAM—An α2-macroglobulin from the hard tick Ixodes ricinus: Characterization and function in phagocytosis of a potential pathogen Chryseobacterium indologenes. *Developmental & Comparative Immunology*, *33*(4), 489–498. https://doi.org/10.1016/j.dci.2008.09.011

Burlak, C., Hammer, C. H., Robinson, M.-A., Whitney, A. R., McGavin, M. J., Kreiswirth, B. N., & DeLeo, F. R. (2007). Global analysis of community-associated methicillin-resistant Staphylococcus aureus exoproteins reveals molecules produced in vitro and during infection. *Cellular Microbiology*, *9*(5), 1172–1190. https://doi.org/10.1111/j.1462-5822.2006.00858.x

Caley, M. P., Martins, V. L. C., & O'Toole, E. A. (2015). Metalloproteinases and Wound Healing. *Advances in Wound Care*, *4*(4), 225–234. https://doi.org/10.1089/wound.2014.0581

Cenac, N. (2013). Protease-Activated Receptors as Therapeutic Targets in Visceral Pain. *Current Neuropharmacology*, *11*(6), 598–605. https://doi.org/10.2174/1570159X113119990039

Cerdà-Costa, N., & Gomis-Rüth, F. X. (2014). Architecture and function of metallopeptidase catalytic domains. *Protein Science*, *23*(2), 123–144. https://doi.org/10.1002/pro.2400

Chang, C.-K., Hung, W.-C., & Chang, H.-C. (2008). The Kazal motifs of RECK protein inhibit MMP-9 secretion and activity and reduce metastasis of lung cancer cells *in vitro* and *in vivo*. *Journal of Cellular and Molecular Medicine*, *12*(6b), 2781–2789. https://doi.org/10.1111/j.1582-4934.2008.00215.x

Chang, H.-C., Cho, C.-Y., & Hung, W.-C. (2007). Downregulation of RECK by promoter methylation correlates with lymph node metastasis in non-small cell lung cancer. *Cancer Science*, *98*(2), 169–173. https://doi.org/10.1111/j.1349-7006.2006.00367.x

Chen, W.-T. (1992). Membrane proteases: roles in tissue remodeling and tumour invasion. *Current Opinion in Cell Biology*, *4*(5), 802–809. https://doi.org/10.1016/0955-0674(92)90103-J

Chen, Y., & Tseng, S.-H. (2012). The potential of RECK inducers as antitumor agents for glioma. *Anticancer Research*, *32*(7), 2991–2998.

Cheung, G. Y. C., Bae, J. S., & Otto, M. (2021). Pathogenicity and virulence of *Staphylococcus aureus*. *Virulence*, *12*(1), 547–569. https://doi.org/10.1080/21505594.2021.1878688

Chiang, C.-H., Hou, M.-F., & Hung, W.-C. (2013). Up-regulation of miR-182 by β-catenin in breast cancer increases tumorigenicity and invasiveness by targeting the matrix metalloproteinase inhibitor RECK. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1830*(4), 3067–3076. https://doi.org/10.1016/j.bbagen.2013.01.009

Clark, J. C. M., Thomas, D. M., Choong, P. F. M., & Dass, C. R. (2007). RECK—a newly discovered inhibitor of metastasis with prognostic significance in multiple forms of cancer. *Cancer and Metastasis Reviews*, *26*(3–4), 675–683. https://doi.org/10.1007/s10555-007-9093-8

Clermont, A., Wedde, M., Seitz, V., Podsiadlowski, L., Lenze, D., Hummel, M., & Vilcinskas, A. (2004). Cloning and expression of an inhibitor of microbial metalloproteinases from insects contributing to innate immunity. *Biochemical Journal*, *382*(1), 315–322. https://doi.org/10.1042/BJ20031923

Cohn, E. J., Strong, L. E., Hughes, W. L., Mulford, D. J., Ashworth, J. N., Melin, M., & Taylor, H. L. (1946). Preparation and Properties of Serum and Plasma Proteins. IV. A System for the Separation into Fractions of the Protein and Lipoprotein Components of Biological Tissues and Fluids [1a,b,c,d]. *Journal of the American Chemical Society*, *68*(3), 459–475. https://doi.org/10.1021/ja01207a034

Colaert, N., Helsens, K., Martens, L., Vandekerckhove, J., & Gevaert, K. (2009). Improved visualization of protein consensus sequences by iceLogo. *Nature Methods*, *6*(11), 786–787. https://doi.org/10.1038/nmeth1109-786

Colloms, S. D. (2004). Leucyl aminopeptidase PepA. In *Handbook of Proteolytic Enzymes* (pp. 905–910). Elsevier. https://doi.org/10.1016/B978-0-12-079611-3.50277-9

Cui, N., Hu, M., & Khalil, R. A. (2017). *Biochemical and Biological Attributes of Matrix Metalloproteinases* (pp. 1–73). https://doi.org/10.1016/bs.pmbts.2017.02.005

Culp, E., & Wright, G. D. (2017). Bacterial proteases, untapped antimicrobial drug targets. *The Journal of Antibiotics*, *70*(4), 366–377. https://doi.org/10.1038/ja.2016.138

Cushing, H. (1935). William Beaumont's Rendezvous with Fame. *The Yale Journal of Biology and Medicine*, *8*(2), 113.b1-126.

Dashek, R. J., Diaz, C., Chandrasekar, B., & Rector, R. S. (2021). The Role of RECK in Hepatobiliary Neoplasia Reveals Its Therapeutic Potential in NASH. *Frontiers in Endocrinology*, *12*. https://doi.org/10.3389/fendo.2021.770740

David, M. Z., & Daum, R. S. (2010). Community-Associated Methicillin-Resistant *Staphylococcus aureus* : Epidemiology and Clinical Consequences of an Emerging Epidemic. *Clinical Microbiology Reviews*, *23*(3), 616–687. https://doi.org/10.1128/CMR.00081-09

de Kreij, A., Venema, G., & van den Burg, B. (2000). Substrate Specificity in the Highly Heterogeneous M4 Peptidase Family Is Determined by a Small Subset of Amino Acids. *Journal of Biological Chemistry*, *275*(40), 31115–31120. https://doi.org/10.1074/jbc.M003889200

Deu, E., Verdoes, M., & Bogyo, M. (2012). New approaches for dissecting protease functions to improve probe development and drug discovery. *Nature Structural & Molecular Biology*, *19*(1), 9–16. https://doi.org/10.1038/nsmb.2203

Discacciati, M. G., Gimenes, F., Pennacchi, P. C., Faião-Flores, F., Zeferino, L. C., Derchain, S. M., Teixeira, J. C., Costa, M. C., Zonta, M., Termini, L., Boccardo, E., Longatto-Filho, A., Consolaro, M. E. L., Villa, L. L., & Maria-Engler, S. S. (2015). MMP-9/RECK Imbalance: A Mechanism Associated with High-Grade Cervical Lesions and Genital Infection by Human Papillomavirus and Chlamydia trachomatis. *Cancer Epidemiology, Biomarkers & Prevention : A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology*, *24*(10), 1539–1547. https://doi.org/10.1158/1055-9965.EPI-15-0420

Doi, E., Koseki, T., & Kitabatake, N. (1987). Effects of limited proteolysis on functional properties of ovalbumin. *Journal of the American Oil Chemists' Society*, *64*(12), 1697–1703. https://doi.org/10.1007/BF02542506

Drapeau, G. R. (1978). Role of metalloprotease in activation of the precursor of staphylococcal protease. *Journal of Bacteriology*, *136*(2), 607–613. https://doi.org/10.1128/jb.136.2.607-613.1978

Eder, J., & Fersht, A. R. (1995). Pro-sequence-assisted protein folding. *Molecular Microbiology*, *16*(4), 609–614. https://doi.org/10.1111/j.1365-2958.1995.tb02423.x

Eijsink, V. G. H., Matthews, B. W., & Vriend, G. (2011). The role of calcium ions in the stability and instability of a thermolysin-like protease. *Protein Science*, *20*(8), 1346–1355. https://doi.org/10.1002/pro.670

Eisenberg, I., Hochner, H., Sadeh, M., Argov, Z., & Mitrani-Rosenbaum, S. (2002). Establishment of the genomic structure and identification of thirteen single-nucleotide polymorphisms in the human RECK gene. *Cytogenetic and Genome Research*, *97*(1–2), 58–61. https://doi.org/10.1159/000064042

Eisenhardt, M., Schlupp, P., Höfer, F., Schmidts, T., Hoffmann, D., Czermak, P., Pöppel, A.-K., Vilcinskas, A., & Runkel, F. (2019). The therapeutic potential of the insect metalloproteinase inhibitor against infections caused by *Pseudomonas aeruginosa*. *Journal of Pharmacy and Pharmacology*, *71*(3), 316–328. https://doi.org/10.1111/jphp.13034

Ellman, G. L. (1959). Tissue sulfhydryl groups. *Archives of Biochemistry and Biophysics*, *82*(1), 70–77. https://doi.org/10.1016/0003-9861(59)90090-6

Enghild, J. J., Thogersen, I. B., Salvesen, G., Fey, G. H., Figler, N. L., Gonias, S. L., & Pizzo, S. v. (1990). .alpha.-Macroglobulin from Limulus polyphemus exhibits proteinase inhibitory activity and participates in a hemolytic system. *Biochemistry*, *29*(43), 10070–10080. https://doi.org/10.1021/bi00495a009

Ezkurdia, I., Juan, D., Rodriguez, J. M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A., & Tress, M. L. (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human Molecular Genetics*, *23*(22), 5866–5878. https://doi.org/10.1093/hmg/ddu309

Fasciglione, G. F., Gioia, M., Tsukada, H., Liang, J., Iundusi, R., Tarantino, U., Coletta, M., Pourmotabbed, T., & Marini, S. (2012). The collagenolytic action of MMP-1 is regulated by the interaction between the catalytic domain and the hinge region. *JBIC Journal of Biological Inorganic Chemistry*, *17*(4), 663–672. https://doi.org/10.1007/s00775-012-0886-z

Feder, J. (1968). A spectrophotometric assay for neutral protease. *Biochemical and Biophysical Research Communications*, *32*(2), 326–332. https://doi.org/10.1016/0006-291X(68)90389-6

Feder, J., & Schuck, J. M. (1970). Studies on the Bacillus subtilis neutral-protease- and Bacillus thermoproteolyticus thermolysin-catalyzed hydrolysis of dipeptide substrates. *Biochemistry*, *9*(14), 2784–2791. https://doi.org/10.1021/bi00816a005

Fejzo, M. S., Ginther, C., Dering, J., Anderson, L., Venkatesan, N., Konecny, G., Karlan, B., & Slamon, D. J. (2011). Knockdown of ovarian cancer amplification target ADRM1 leads to downregulation of GIPC1 and upregulation of RECK. *Genes, Chromosomes and Cancer*, *50*(6), 434–441. https://doi.org/10.1002/gcc.20868

Frees, D., Brøndsted, L., & Ingmer, H. (2013). *Bacterial Proteases and Virulence* (pp. 161–192). https://doi.org/10.1007/978-94-007-5940-4_7

Fritsche, E., Paschos, A., Beisel, H.-G., Böck, A., & Huber, R. (1999). Crystal structure of the hydrogenase maturating endopeptidase HYBD from Escherichia coli. *Journal of Molecular Biology*, *288*(5), 989–998. https://doi.org/10.1006/jmbi.1999.2719

Fröbius, A. C., Kanost, M. R., Götz, P., & Vilcinskas, A. (2000). Isolation and characterization of novel inducible serine protease inhibitors from larval hemolymph of the greater wax moth Galleria mellonella. *European Journal of Biochemistry*, *267*(7), 2046–2053. https://doi.org/10.1046/j.1432-1327.2000.01207.x

Fulda, S., Gorman, A. M., Hori, O., & Samali, A. (2010). Cellular Stress Responses: Cell Survival and Cell Death. *International Journal of Cell Biology*, *2010*, 1–23. https://doi.org/10.1155/2010/214074

Fushimi, N., Ee, C. E., Nakajima, T., & Ichishima, E. (1999). Aspzincin, a Family of Metalloendopeptidases with a New Zinc-binding Motif. *Journal of Biological Chemistry*, *274*(34), 24195–24201. https://doi.org/10.1074/jbc.274.34.24195

Galazka, G., Windsor, L. J., Birkedal-Hansen, H., & Engler, J. A. (1996). APMA (4-Aminophenylmercuric Acetate) Activation of Stromelysin-1 Involves Protein Interactions in Addition to Those with Cysteine-75 in the Propeptide. *Biochemistry*, *35*(34), 11221–11227. https://doi.org/10.1021/bi960618e

Gao, X., Wang, J., Yu, D.-Q., Bian, F., Xie, B.-B., Chen, X.-L., Zhou, B.-C., Lai, L.-H., Wang, Z.-X., Wu, J.-W., & Zhang, Y.-Z. (2010). Structural basis for the autoprocessing of zinc metalloproteases in the thermolysin family. *Proceedings of the National Academy of Sciences*, *107*(41), 17569–17574. https://doi.org/10.1073/pnas.1005681107

Garcia-Ferrer, I., Arêde, P., Gómez-Blanco, J., Luque, D., Duquerroy, S., Castón, J. R., Goulas, T., & Gomis-Rüth, F. X. (2015). Structural and functional insights into *Escherichia coli* α 2 -macroglobulin endopeptidase snap-trap inhibition. *Proceedings of the National Academy of Sciences*, *112*(27), 8290–8295. https://doi.org/10.1073/pnas.1506538112

Garcia-Ferrer, I., Marrero, A., Gomis-Rüth, F. X., & Goulas, T. (2017). *α2-Macroglobulins: Structure and Function* (pp. 149–183). https://doi.org/10.1007/978-3-319-46503-6_6

Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D., & Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. In *The Proteomics Protocols Handbook* (pp. 571–607). Humana Press. https://doi.org/10.1385/1-59259-890-0:571

Girardi, G., Lingo, J. J., Fleming, S. D., & Regal, J. F. (2020). Essential Role of Complement in Pregnancy: From Implantation to Parturition and Beyond. *Frontiers in Immunology*, *11*. https://doi.org/10.3389/fimmu.2020.01681

Gomis-Rüth, F. X., Botelho, T. O., & Bode, W. (2012). A standard orientation for metallopeptidases. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1824*(1), 157–163. https://doi.org/10.1016/j.bbapap.2011.04.014

Goth, C. K., Vakhrushev, S. Y., Joshi, H. J., Clausen, H., & Schjoldager, K. T. (2018). Fine-Tuning Limited Proteolysis: A Major Role for Regulated Site-Specific O -Glycosylation. *Trends in Biochemical Sciences*, *43*(4), 269–284. https://doi.org/10.1016/j.tibs.2018.02.005

Goulas, T., Garcia-Ferrer, I., García-Piqué, S., Sottrup-Jensen, L., & Gomis-Rüth, F. X. (2014). Crystallization and preliminary X-ray diffraction analysis of eukaryotic α 2 -macroglobulin family members modified by methylamine, proteases and glycosidases. *Molecular Oral Microbiology*, *29*(6), 354–364. https://doi.org/10.1111/omi.12069

Gross, J., & Lapiere, C. M. (1962). COLLAGENOLYTIC ACTIVITY IN AMPHIBIAN TISSUES: A TISSUE CULTURE ASSAY. *Proceedings of the National Academy of Sciences*, *48*(6), 1014–1022. https://doi.org/10.1073/pnas.48.6.1014

Guerder, S., Hassel, C., & Carrier, A. (2019). Thymus-specific serine protease, a protease that shapes the CD4 T cell repertoire. *Immunogenetics*, *71*(3), 223–232. https://doi.org/10.1007/s00251-018-1078-y

Guo, J., & Zou, L. (2006). Correlation of RECK with matrix metalloproteinase-2 in regulation of trophoblast invasion of early pregnancy. *Journal of Huazhong University of Science and Technology*, *26*(6), 738–740. https://doi.org/10.1007/s11596-006-0631-3

Gutfreund, H. (1976). Wilhelm Friedrich Kühne; An appreciation. *FEBS Letters*, *62*(S1), E1–E2. https://doi.org/10.1016/0014-5793(76)80846-0

Hartley, B. S. (1960). PROTEOLYTIC ENZYMES. *Annual Review of Biochemistry*, *29*(1), 45–72. https://doi.org/10.1146/annurev.bi.29.070160.000401

Hashimoto, H., Takeuchi, T., Komatsu, K., Miyazaki, K., Sato, M., & Higashi, S. (2011). Structural Basis for Matrix Metalloproteinase-2 (MMP-2)-selective Inhibitory Action of β-Amyloid Precursor Protein-derived Inhibitor. *Journal of Biological Chemistry*, *286*(38), 33236–33243. https://doi.org/10.1074/jbc.M111.264176

Hausrath, A. C., & Matthews, B. W. (2002). Thermolysin in the absence of substrate has an open conformation. *Acta Crystallographica Section D Biological Crystallography*, *58*(6), 1002–1007. https://doi.org/10.1107/S090744490200584X

Heinrikson, R. L. (1977). *[20] Applications of thermolysin in protein structural analysis* (pp. 175–189). https://doi.org/10.1016/0076-6879(77)47022-8

Hibbetts, K., Hines, B., & Williams, D. (1999). An Overview of Proteinase Inhibitors. *Journal of Veterinary Internal Medicine*, *13*(4), 302–308. https://doi.org/10.1111/j.1939-1676.1999.tb02185.x

Hirata, H., Ueno, K., Shahryari, V., Deng, G., Tanaka, Y., Tabatabai, Z. L., Hinoda, Y., & Dahiya, R. (2013). MicroRNA-182-5p Promotes Cell Invasion and Proliferation by Down Regulating FOXF2, RECK and MTSS1 Genes in Human Prostate Cancer. *PLoS ONE*, *8*(1), e55502. https://doi.org/10.1371/journal.pone.0055502

Hirata, H., Ueno, K., Shahryari, V., Tanaka, Y., Tabatabai, Z. L., Hinoda, Y., & Dahiya, R. (2012). Oncogenic miRNA-182-5p Targets Smad4 and RECK in Human Bladder Cancer. *PLoS ONE*, *7*(11), e51056. https://doi.org/10.1371/journal.pone.0051056

Holland, D. R., Tronrud, D. E., Pley, H. W., Flaherty, K. M., Stark, W., Jansonius, J. N., McKay, D. B., & Matthews, B. W. (1992). Structural comparison suggests that thermolysin and related neutral proteases undergo hinge-bending motion during catalysis. *Biochemistry*, *31*(46), 11310–11316. https://doi.org/10.1021/bi00161a008

Holmquist, B., & Vallee, B. L. (1974). Metal Substitutions and Inhibition of Thermolysin: Spectra of the Cobalt Enzyme. *Journal of Biological Chemistry*, *249*(14), 4601–4607. https://doi.org/10.1016/S0021-9258(19)42460-5

Hooper, N. M. (1994). Families of zinc metalloproteases. *FEBS Letters*, *354*(1), 1–6. https://doi.org/10.1016/0014-5793(94)01079-X

Hou, C., & Zhang, Y. (2008). Expression of Reversion-inducing Cysteine-rich Protein with Kazal Motifs in Peripheral Blood Mononuclear Cells from Patients with Systemic Lupus Erythematosus: Links to Disease Activity, Damage Accrual and Matrix Metalloproteinase 9 Secretion. *Journal of International Medical Research*, *36*(4), 704–713. https://doi.org/10.1177/147323000803600412

Huesgen, P. F., Lange, P. F., Rogers, L. D., Solis, N., Eckhard, U., Kleifeld, O., Goulas, T., Gomis-Rüth, F. X., & Overall, C. M. (2015). LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nature Methods*, *12*(1), 55–58. https://doi.org/10.1038/nmeth.3177

Ikai, A., Kitamoto, T., & Nishigai, M. (1983). Alpha-2-Macroglobulin-Like Protease Inhibitor from the Egg White of Cuban Crocodile (Crocodylus rhombifer)1. *The Journal of Biochemistry*, *93*(1), 121–127. https://doi.org/10.1093/oxfordjournals.jbchem.a134145

Inouye, K. I. (1992). Effects of Salts on Thermolysin: Activation of Hydrolysis and Synthesis of N-Carbobenzoxy-L-Aspartyl-L-Phenylalanine Methyl Ester, and a Unique Change in the Absorption Spectrum of Thermolysin. *The Journal of Biochemistry*, *112*(3), 335–340. https://doi.org/10.1093/oxfordjournals.jbchem.a123901

Inouye, K., Kuzuya, K., & Tonomura, B. (1998). Sodium chloride enhances markedly the thermal stability of thermolysin as well as its catalytic activity. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1388*(1), 209–214. https://doi.org/10.1016/S0167-4838(98)00189-7

Itoh, Y. (2001). Homophilic complex formation of MT1-MMP facilitates proMMP-2 activation on the cell surface and promotes tumor cell invasion. *The EMBO Journal*, *20*(17), 4782–4793. https://doi.org/10.1093/emboj/20.17.4782

Itoh, Y. (2015). Membrane-type matrix metalloproteinases: Their functions and regulations. *Matrix Biology*, *44–46*, 207–223. https://doi.org/10.1016/j.matbio.2015.03.004

Iyer, S., Visse, R., Nagase, H., & Acharya, K. R. (2006). Crystal Structure of an Active Form of Human MMP-1. *Journal of Molecular Biology*, *362*(1), 78–88. https://doi.org/10.1016/j.jmb.2006.06.079

Jacomasso, T., Trombetta-Lima, M., Sogayar, M. C., & Winnischofer, S. M. B. (2014). Downregulation of reversion-inducing cysteine-rich protein with Kazal motifs in malignant melanoma: inverse correlation with membrane-type 1-matrix metalloproteinase and tissue inhibitor of metalloproteinase 2. *Melanoma Research*, *24*(1), 32–39. https://doi.org/10.1097/CMR.0000000000000039

Jensen, K., Oestergaard, P. R., Wilting, R., & Lassen, S. F. (2010). Identification and characterization of a bacterial glutamic peptidase. *BMC Biochemistry*, *11*(1), 47. https://doi.org/10.1186/1471-2091-11-47

Jensen, P. E., & Sottrup-Jensen, L. (1986). Primary structure of human alpha 2-macroglobulin. Complete disulfide bridge assignment and localization of two interchain bridges in the dimeric proteinase binding unit. *Journal of Biological Chemistry*, *261*(34), 15863–15869. https://doi.org/10.1016/S0021-9258(18)66643-8

Jeske, L., Placzek, S., Schomburg, I., Chang, A., & Schomburg, D. (2019). BRENDA in 2019: a European ELIXIR core data resource. *Nucleic Acids Research*, *47*(D1), D542–D549. https://doi.org/10.1093/nar/gky1048

Jevševar, S., Gaberc-Porekar, V., Fonda, I., Podobnik, B., Grdadolnik, J., & Menart, V. (2008). Production of Nonclassical Inclusion Bodies from Which Correctly Folded Protein Can Be Extracted. *Biotechnology Progress*, *21*(2), 632–639. https://doi.org/10.1021/bp0497839

Jozic, D., Bourenkov, G., Lim, N.-H., Visse, R., Nagase, H., Bode, W., & Maskos, K. (2005). X-ray Structure of Human proMMP-1. *Journal of Biological Chemistry*, *280*(10), 9578–9585. https://doi.org/10.1074/jbc.M411084200

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kajiwara, K., Fujita, A., Tsuyuki, H., Kumazaki, T., & Ishii, S. (1991). Interactions of Streptomyces Serine-Protease Inhibitors with Streptomyces griseus Metalloendopeptidase II. *The Journal of Biochemistry*, *110*(3), 350–354. https://doi.org/10.1093/oxfordjournals.jbchem.a123584

Kang, H.-G., Kim, H.-S., Kim, K.-J., Oh, J. H., Lee, M.-R., Seol, S. M., & Han, I. (2007). RECK expression in osteosarcoma: correlation with matrix metalloproteinases activation and tumor invasiveness. *Journal of Orthopaedic Research*, *25*(5), 696–702. https://doi.org/10.1002/jor.20323

Kanost, M. R. (1999). Serine proteinase inhibitors in arthropod immunity. *Developmental & Comparative Immunology*, *23*(4–5), 291–301. https://doi.org/10.1016/S0145-305X(99)00012-9

Kantyka, T., Rawlings, N. D., & Potempa, J. (2010). Prokaryote-derived protein inhibitors of peptidases: A sketchy occurrence and mostly unknown function. *Biochimie*, *92*(11), 1644–1656. https://doi.org/10.1016/j.biochi.2010.06.004

Kataoka, Y., Takada, K., Oyama, H., Tsunemi, M., James, M. N. G., & Oda, K. (2005). Catalytic residues and substrate specificity of scytalidoglutamic peptidase, the first member of the eqolisin in family (G1) of peptidases. *FEBS Letters*, *579*(14), 2991–2994. https://doi.org/10.1016/j.febslet.2005.04.050

Kelly, J., & Kavanagh, K. (2011). Caspofungin primes the immune response of the larvae of Galleria mellonella and induces a non-specific antimicrobial response. *Journal of Medical Microbiology*, *60*(2), 189–196. https://doi.org/10.1099/jmm.0.025494-0

Khan, A. R., & James, M. N. G. (1998). Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Protein Science*, *7*(4), 815–836. https://doi.org/10.1002/pro.5560070401

Klein, T., Eckhard, U., Dufour, A., Solis, N., & Overall, C. M. (2018). Proteolytic Cleavage—Mechanisms, Function, and "Omic" Approaches for a Near-Ubiquitous Posttranslational Modification. *Chemical Reviews*, *118*(3), 1137–1168. https://doi.org/10.1021/acs.chemrev.7b00120

Knäuper, V., Docherty, A. J. P., Smith, B., Tschesche, H., & Murphy, G. (1997). Analysis of the contribution of the hinge region of human neutrophil collagenase (HNC, MMP-8) to stability and collagenolytic activity by alanine scanning mutagenesis. *FEBS Letters*, *405*(1), 60–64. https://doi.org/10.1016/S0014-5793(97)00158-0

Kolodziej, S. J., Penczek, P. A., Schroeter, J. P., & Stoops, J. K. (1996). Structure-Function Relationships of the Saccharomyces cerevisiae Fatty Acid Synthase. *Journal of Biological Chemistry*, *271*(45), 28422–28429. https://doi.org/10.1074/jbc.271.45.28422

Krishnaswamy, S. (2013). The transition of prothrombin to thrombin. *Journal of Thrombosis and Haemostasis*, *11*, 265–276. https://doi.org/10.1111/jth.12217

Kubica, M., Guzik, K., Koziel, J., Zarebski, M., Richter, W., Gajkowska, B., Golda, A., Maciag-Gudowska, A., Brix, K., Shaw, L., Foster, T., & Potempa, J. (2008). A Potential New Pathway for Staphylococcus aureus Dissemination: The Silent Survival of S. aureus Phagocytosed by Human Monocyte-Derived Macrophages. *PLoS ONE*, *3*(1), e1409. https://doi.org/10.1371/journal.pone.0001409

Laarman, A. J., Ruyken, M., Malone, C. L., van Strijp, J. A. G., Horswill, A. R., & Rooijakkers, S. H. M. (2011). *Staphylococcus aureus* Metalloprotease Aureolysin Cleaves Complement C3 To Mediate Immune Evasion. *The Journal of Immunology*, *186*(11), 6445–6453. https://doi.org/10.4049/jimmunol.1002948

Laronha, H., & Caldeira, J. (2020). Structure and Function of Human Matrix Metalloproteinases. *Cells*, *9*(5), 1076. https://doi.org/10.3390/cells9051076

Laronha, H., Carpinteiro, I., Portugal, J., Azul, A., Polido, M., Petrova, K. T., Salema-Oom, M., & Caldeira, J. (2020). Challenges in Matrix Metalloproteinases Inhibition. *Biomolecules*, *10*(5), 717. https://doi.org/10.3390/biom10050717

Laskowski, M., & Kato, I. (1980). Protein Inhibitors of Proteinases. *Annual Review of Biochemistry*, *49*(1), 593–626. https://doi.org/10.1146/annurev.bi.49.070180.003113

Laskowski, M., & Qasim, M. A. (2000). What can the structures of enzyme-inhibitor complexes tell us about the structures of enzyme substrate complexes? *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1477*(1–2), 324–337. https://doi.org/10.1016/S0167-4838(99)00284-8

Levene, P. A. (1905). The Cleavage Products of Proteoses. *Journal of Biological Chemistry*, *1*(1), 45–58. https://doi.org/10.1016/S0021-9258(17)46095-9

Liaci, A. M., & Förster, F. (2021). Take Me Home, Protein Roads: Structural Insights into Signal Peptide Interactions during ER Translocation. *International Journal of Molecular Sciences*, *22*(21), 11871. https://doi.org/10.3390/ijms222111871

Ligné, T., Pauthe, E., Monti, J.-P., Gacel, G., & Larreta-Garde, V. (1997). Additional data about thermolysin specificity in buffer- and glycerol-containing media. *Biochimica et Biophysica Acta (BBA) - Protein Structure and Molecular Enzymology*, *1337*(1), 143–148. https://doi.org/10.1016/S0167-4838(96)00142-2

Lim, W., Jeong, W., Kim, J.-H., Lee, J.-Y., Kim, J., Bazer, F. W., Han, J. Y., & Song, G. (2011). Differential expression of alpha 2 macroglobulin in response to dietylstilbestrol and in ovarian carcinomas in chickens. *Reproductive Biology and Endocrinology*, *9*(1), 137. https://doi.org/10.1186/1477-7827-9-137

Liu, W., Song, N., Yao, H., Zhao, L., Liu, H., & Li, G. (2015). miR-221 and miR-222 Simultaneously Target RECK and Regulate Growth and Invasion of Gastric Cancer Cells. *Medical Science Monitor*, *21*, 2718–2725. https://doi.org/10.12659/MSM.894324

Liu, X., Wang, W., Chen, J., Chen, C., Zhou, J., & Cao, L. (2012). Expression of Reversion-Inducing Cysteine-Rich Protein with Kazal Motifs and Matrix Metalloproteinase 9 in Middle Ear Squamous Cell Carcinoma. *ORL*, *74*(1), 16–21. https://doi.org/10.1159/000334243

Loffek, S., Schilling, O., & Franzke, C.-W. (2011). Biological role of matrix metalloproteinases: a critical balance. *European Respiratory Journal*, *38*(1), 191–208. https://doi.org/10.1183/09031936.00146510

López-Otín, C., & Bond, J. S. (2008). Proteases: Multifunctional Enzymes in Life and Disease. *Journal of Biological Chemistry*, *283*(45), 30433–30437. https://doi.org/10.1074/jbc.R800035200

Lowy, F. D. (1998). *Staphylococcus aureus* Infections. *New England Journal of Medicine*, *339*(8), 520–532. https://doi.org/10.1056/NEJM199808203390806

Luque, D., Goulas, T., Mata, C. P., Mendes, S. R., Gomis-Rüth, F. X., & Castón, J. R. (2022). Cryo-EM structures show the mechanistic basis of pan-peptidase inhibition by human α 2 -macroglobulin. *Proceedings of the National Academy of Sciences*, *119*(19). https://doi.org/10.1073/pnas.2200102119

Madri, J. A., & Graesser, D. (2000). Cell Migration in the Immune System: the Evolving Inter-Related Roles of Adhesion Molecules and Proteinases. *Developmental Immunology*, *7*(2–4), 103–116. https://doi.org/10.1155/2000/79045

Mäkinen, P. L., Clewell, D. B., An, F., & Mäkinen, K. K. (1989). Purification and substrate specificity of a strongly hydrophobic extracellular metalloendopeptidase ("gelatinase") from Streptococcus faecalis (strain 0G1-10). *The Journal of Biological Chemistry*, *264*(6), 3325–3334.

Makinen, P. L., & Makinen, K. K. (1994). The Enterococcus faecalis Extracellular Metalloendopeptidase (EC 3.4.24.30; Coccolysin) Inactivates Human Endothelin at Bonds Involving Hydrophobic Amino Acid Residues. *Biochemical and Biophysical Research Communications*, *200*(2), 981–985. https://doi.org/10.1006/bbrc.1994.1546

Marie-Claire, C., Roques, B. P., & Beaumont, A. (1998). Intramolecular Processing of Prothermolysin. *Journal of Biological Chemistry*, *273*(10), 5697–5701. https://doi.org/10.1074/jbc.273.10.5697

Marino-Puertas, L., del Amo-Maestro, L., Taulés, M., Gomis-Rüth, F. X., & Goulas, T. (2019). Recombinant production of human α2-macroglobulin variants and interaction studies with recombinant G-related α2-macroglobulin binding protein and latent transforming growth factor-β2. *Scientific Reports*, *9*(1), 9186. https://doi.org/10.1038/s41598-019-45712-z

Markaryan, A., Lee, J. D., Sirakova, T. D., & Kolattukudy, P. E. (1996). Specific inhibition of mature fungal serine proteinases and metalloproteinases by their propeptides. *Journal of Bacteriology*, *178*(8), 2211–2215. https://doi.org/10.1128/jb.178.8.2211-2215.1996

Marrero, A., Duquerroy, S., Trapani, S., Goulas, T., Guevara, T., Andersen, G. R., Navaza, J., Sottrup-Jensen, L., & Gomis-Rüth, F. X. (2012). The Crystal Structure of Human α2-Macroglobulin Reveals a Unique Molecular Cage. *Angewandte Chemie International Edition*, *51*(14), 3340–3344. https://doi.org/10.1002/anie.201108015

Marshall, N. C., Finlay, B. B., & Overall, C. M. (2017). Sharpening Host Defenses during Infection: Proteases Cut to the Chase. *Molecular & Cellular Proteomics*, *16*(4), S161–S171. https://doi.org/10.1074/mcp.O116.066456

Masui, T., Doi, R., Koshiba, T., Fujimoto, K., Tsuji, S., Nakajima, S., Koizumi, M., Toyoda, E., Tulachan, S., Ito, D., Kami, K., Mori, T., Wada, M., Noda, M., & Imamura, M. (2003). RECK expression in pancreatic cancer: its correlation with lower invasiveness and better prognosis. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *9*(5), 1779–1784.

Matthews, B. W. (1988). Structural basis of the action of thermolysin and related zinc peptidases. *Accounts of Chemical Research*, *21*(9), 333–340. https://doi.org/10.1021/ar00153a003

Matthews, S. P., Werber, I., Deussing, J., Peters, C., Reinheckel, T., & Watts, C. (2010). Distinct Protease Requirements for Antigen Presentation In Vitro and In Vivo. *The Journal of Immunology*, *184*(5), 2423–2431. https://doi.org/10.4049/jimmunol.0901486

Maurizi, M. R., & Switzer, R. L. (1980). *Proteolysis in Bacterial Sporulation* (pp. 163–224). https://doi.org/10.1016/B978-0-12-152816-4.50010-8

McCarty, S. M., & Percival, S. L. (2013). Proteases and Delayed Wound Healing. *Advances in Wound Care*, *2*(8), 438–447. https://doi.org/10.1089/wound.2012.0370

McDonald, A. G., Boyce, S., & Tipton, K. F. (2009). ExplorEnz: the primary source of the IUBMB enzyme list. *Nucleic Acids Research*, *37*(Database), D593–D597. https://doi.org/10.1093/nar/gkn582

McKerrow, J. H. (1987). Human fibroblast collagenase contains an amino acid sequence homologous to the zinc-binding site of Serratia protease. *Journal of Biological Chemistry*, *262*(13), 5943. https://doi.org/10.1016/S0021-9258(18)45517-2

Mendes, S. R., Amo-Maestro, L. del, Marino-Puertas, L., Diego, I. de, Goulas, T., & Gomis-Rüth, F. X. (2020). Analysis of the inhibiting activity of reversion-inducing cysteine-rich protein with Kazal motifs (RECK) on matrix metalloproteinases. *Scientific Reports*, *10*(1), 6317. https://doi.org/10.1038/s41598-020-63338-4

Mendes, S. R., Eckhard, U., Rodríguez-Banqueri, A., Guevara, T., Czermak, P., Marcos, E., Vilcinskas, A., & Xavier Gomis-Rüth, F. (2022). An engineered protein-based submicromolar competitive inhibitor of the Staphylococcus aureus virulence factor aureolysin. *Computational and Structural Biotechnology Journal*, *20*, 534–544. https://doi.org/10.1016/j.csbj.2022.01.001

Micevski, D., & Dougan, D. A. (2013). *Proteolytic Regulation of Stress Response Pathways in Escherichia coli* (pp. 105–128). https://doi.org/10.1007/978-94-007-5940-4_5

Miki, T., Takegami, Y., Okawa, K., Muraguchi, T., Noda, M., & Takahashi, C. (2007). The Reversion-inducing Cysteine-rich Protein with Kazal Motifs (RECK) Interacts with Membrane Type 1 Matrix Metalloproteinase and CD13/Aminopeptidase N and Modulates Their Endocytic Pathways. *Journal of Biological Chemistry*, *282*(16), 12341–12352. https://doi.org/10.1074/jbc.M610948200

Miyoshi, S. (2013). Extracellular proteolytic enzymes produced by human pathogenic vibrio species. *Frontiers in Microbiology*, *4*. https://doi.org/10.3389/fmicb.2013.00339

Moraes, F., & Góes, A. (2016). A decade of human genome project conclusion: Scientific diffusion about our genome knowledge. *Biochemistry and Molecular Biology Education*, *44*(3), 215–223. https://doi.org/10.1002/bmb.20952

Mukherjee, K., Altincicek, B., Hain, T., Domann, E., Vilcinskas, A., & Chakraborty, T. (2010). *Galleria mellonella* as a Model System for Studying *Listeria* Pathogenesis. *Applied and Environmental Microbiology*, *76*(1), 310–317. https://doi.org/10.1128/AEM.01301-09

Muraguchi, T., Takegami, Y., Ohtsuka, T., Kitajima, S., Chandana, E. P. S., Omura, A., Miki, T., Takahashi, R., Matsumoto, N., Ludwig, A., Noda, M., & Takahashi, C. (2007). RECK modulates Notch signaling during cortical neurogenesis by regulating ADAM10 activity. *Nature Neuroscience*, *10*(7), 838–845. https://doi.org/10.1038/nn1922

Nagase, H., & Harris, E. D. (1983). Ovostatin: a novel proteinase inhibitor from chicken egg white. II. Mechanism of inhibition studied with collagenase and thermolysin. *The Journal of Biological Chemistry*, *258*(12), 7490–7498.

Nagase, H., Visse, R., & Murphy, G. (2006). Structure and function of matrix metalloproteinases and TIMPs. *Cardiovascular Research*, *69*(3), 562–573. https://doi.org/10.1016/j.cardiores.2005.12.002

Nagini, S. (2012). RECKing MMP: Relevance of Reversion-inducing Cysteine-rich Protein with Kazal Motifs as a Prognostic Marker and Therapeutic Target for Cancer (A Review). *Anti-Cancer Agents in Medicinal Chemistry*, *12*(7), 718–725. https://doi.org/10.2174/187152012802650237

Nakada, M., Yamada, A., Takino, T., Miyamori, H., Takahashi, T., Yamashita, J., & Sato, H. (2001). Suppression of membrane-type 1 matrix metalloproteinase (MMP)-mediated MMP-2 activation and tumor invasion by testican 3 and its splicing variant gene product, N-Tes. *Cancer Research*, *61*(24), 8896–8902.

Neves, D., Estrozi, L. F., Job, V., Gabel, F., Schoehn, G., & Dessen, A. (2012). Conformational States of a Bacterial α2-Macroglobulin Resemble Those of Human Complement C3. *PLoS ONE*, *7*(4), e35384. https://doi.org/10.1371/journal.pone.0035384

Nguyen, T. T. H., Myrold, D. D., & Mueller, R. S. (2019). Distributions of Extracellular Peptidases Across Prokaryotic Genomes Reflect Phylogeny and Habitat. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.00413

Nickerson, N. N., Joag, V., & McGavin, M. J. (2008). Rapid autocatalytic activation of the M4 metalloprotease aureolysin is controlled by a conserved N-terminal fungalysin-thermolysin-propeptide domain. *Molecular Microbiology*, *69*(6), 1530–1543. https://doi.org/10.1111/j.1365-2958.2008.06384.x

Nielsen, K. L., Sottrup-Jensen, L., Nagase, H., Thøgersen, H. C., & Etzerodt, M. (1994). Amino Acid Sequence of Hen Ovomacroglobulin (Ovostatin) deduced from cloned cDNA. *DNA Sequence*, *5*(2), 111–119. https://doi.org/10.3109/10425179409039712

Noda, M., & Takahashi, C. (2007). Recklessness as a hallmark of aggressive cancer. *Cancer Science*, *98*(11), 1659–1665. https://doi.org/10.1111/j.1349-7006.2007.00588.x

Nomenclature and Symbolism for Amino Acids and Peptides. Recommendations 1983. (1984). *European Journal of Biochemistry*, *138*(1), 9–37. https://doi.org/10.1111/j.1432-1033.1984.tb07877.x

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. v., Mikheenko, A., Vollger, M. R., Altemose, N., Uralsky, L., Gershman, A., Aganezov, S., Hoyt, S. J., Diekhans, M., Logsdon, G. A., Alonge, M., Antonarakis, S. E., Borchers, M., Bouffard, G. G., Brooks, S. Y., … Phillippy, A. M. (2022). The complete sequence of a human genome. *Science*, *376*(6588), 44–53. https://doi.org/10.1126/science.abj6987

Oh, J., Takahashi, R., Kondo, S., Mizoguchi, A., Adachi, E., Sasahara, R. M., Nishimura, S., Imamura, Y., Kitayama, H., Alexander, D. B., Ide, C., Horan, T. P., Arakawa, T., Yoshida, H., Nishikawa, S., Itoh, Y., Seiki, M., Itohara, S., Takahashi, C., & Noda, M. (2001). The Membrane-Anchored MMP Inhibitor RECK Is a Key Regulator of Extracellular Matrix Integrity and Angiogenesis. *Cell*, *107*(6), 789–800. https://doi.org/10.1016/S0092-8674(01)00597-9

Omura, A., Matsuzaki, T., Mio, K., Ogura, T., Yamamoto, M., Fujita, A., Okawa, K., Kitayama, H., Takahashi, C., Sato, C., & Noda, M. (2009). RECK Forms Cowbell-shaped Dimers and Inhibits Matrix Metalloproteinase-catalyzed Cleavage of Fibronectin. *Journal of Biological Chemistry*, *284*(6), 3461–3469. https://doi.org/10.1074/jbc.M806212200

Oshima, T., Kunisaki, C., Yoshihara, K., Yamada, R., Yamamoto, N., Sato, T., Makino, H., Yamagishi, S., Nagano, Y., Fujii, S., Shiozawa, M., Akaike, M., Wada, N., Rino, Y., Masuda, M., Tanaka, K., & Imada, T. (2008). Clinicopathological significance of the gene expression of matrix metalloproteinases and reversion-inducing cysteine-rich protein with Kazal motifs in patients with colorectal cancer: MMP-2 gene expression is a useful predictor of liver metastasis from colorectal cancer. *Oncology Reports*, *19*(5), 1285–1291.

Overall, C. M. (2002). Molecular Determinants of Metalloproteinase Substrate Specificity: Matrix Metalloproteinase Substrate Binding Domains, Modules, and Exosites. *Molecular Biotechnology*, *22*(1), 051–086. https://doi.org/10.1385/MB:22:1:051

Overall, C. M., & Blobel, C. P. (2007). In search of partners: linking extracellular proteases to substrates. *Nature Reviews Molecular Cell Biology*, *8*(3), 245–257. https://doi.org/10.1038/nrm2120

Papamokos, E., Weber, E., Bode, W., Huber, R., Empie, M. W., Kato, I., & Laskowski, M. (1982). Crystallographic refinement of Japanese quail ovomucoid, a Kazal-type inhibitor, and model building studies of complexes with serine proteases. *Journal of Molecular Biology*, *158*(3), 515–537. https://doi.org/10.1016/0022-2836(82)90212-1

Parks, W. C., Wilson, C. L., & López-Boado, Y. S. (2004). Matrix metalloproteinases as modulators of inflammation and innate immunity. *Nature Reviews Immunology*, *4*(8), 617–629. https://doi.org/10.1038/nri1418

Pei, D., Kang, T., & Qi, H. (2000). Cysteine Array Matrix Metalloproteinase (CA-MMP)/MMP-23 Is a Type II Transmembrane Matrix Metalloproteinase Regulated by a Single Cleavage for Both Secretion and Activation. *Journal of Biological Chemistry*, *275*(43), 33988–33997. https://doi.org/10.1074/jbc.M006493200

Pei, D., & Weiss, S. J. (1995). Furin-dependent intracellular activation of the human stromelysin-3 zymogen. *Nature*, *375*(6528), 244–247. https://doi.org/10.1038/375244a0

Pérez-Silva, J. G., Español, Y., Velasco, G., & Quesada, V. (2016). The Degradome database: expanding roles of mammalian proteases in life and disease. *Nucleic Acids Research*, *44*(D1), D351–D355. https://doi.org/10.1093/nar/gkv1201

Pietrocola, G., Nobile, G., Rindi, S., & Speziale, P. (2017). Staphylococcus aureus Manipulates Innate Immunity through Own and Host-Expressed Proteases. *Frontiers in Cellular and Infection Microbiology*, *7*. https://doi.org/10.3389/fcimb.2017.00166

Piskór, B. M., Przylipiak, A., Dąbrowska, E., Niczyporuk, M., & Ławicki, S. (2020). Matrilysins and Stromelysins in Pathogenesis and Diagnostics of Cancers . *Cancer Management and Research*, *Volume 12*, 10949–10964. https://doi.org/10.2147/CMAR.S235776

Ponomarenko, E. A., Poverennaya, E. v., Ilgisonis, E. v., Pyatnitskiy, M. A., Kopylov, A. T., Zgoda, V. G., Lisitsa, A. v., & Archakov, A. I. (2016). The Size of the Human Proteome: The Width and Depth. *International Journal of Analytical Chemistry*, *2016*, 1–6. https://doi.org/10.1155/2016/7436849

Potempa, J., & Shaw, L. N. (2013). Aureolysin. In *Handbook of Proteolytic Enzymes* (pp. 563–569). Elsevier. https://doi.org/10.1016/B978-0-12-382219-2.00114-9

Potempa, J., Watorek, W., & Travis, J. (1986). The inactivation of human plasma alpha 1-proteinase inhibitor by proteinases from Staphylococcus aureus. *Journal of Biological Chemistry*, *261*(30), 14330–14334. https://doi.org/10.1016/S0021-9258(18)67022-X

Prokešová, L., Porwit-Bóbr, Z., Baran, K., Potempa, J., Pospíšil, M., & John, C. (1991). Effect of metalloproteinase from Staphylococcus aureus on in vitro stimulation of human lymphocytes. *Immunology Letters*, *27*(3), 225–230. https://doi.org/10.1016/0165-2478(91)90156-5

Puente, X. S., Sánchez, L. M., Overall, C. M., & López-Otín, C. (2003). Human and mouse proteases: a comparative genomic approach. *Nature Reviews Genetics*, *4*(7), 544–558. https://doi.org/10.1038/nrg1111

Pulkoski-Gross, A. E. (2015). Historical perspective of matrix metalloproteases. *Frontiers in Bioscience*, *7*(1), 429. https://doi.org/10.2741/s429

Qazi, U., Kolodziej, S. J., Gettins, P. G. W., & Stoops, J. K. (2000). The Structure of the C949S Mutant Human α2-Macroglobulin Demonstrates the Critical Role of the Internal Thiol Esters in Its Proteinase-Entrapping Structural Transformation. *Journal of Structural Biology*, *131*(1), 19–26. https://doi.org/10.1006/jsbi.2000.4269

Qi, Q., Lu, N., Li, C., Zhao, J., Liu, W., You, Q., & Guo, Q. (2015). Involvement of RECK in gambogic acid induced anti-invasive effect in A549 human lung carcinoma cells. *Molecular Carcinogenesis*, *54*(S1), E13–E25. https://doi.org/10.1002/mc.22138

Quesada, V., Ordonez, G. R., Sanchez, L. M., Puente, X. S., & Lopez-Otin, C. (2009). The Degradome database: mammalian proteases and diseases of proteolysis. *Nucleic Acids Research*, *37*(Database), D239–D243. https://doi.org/10.1093/nar/gkn570

Ra, H.-J., & Parks, W. C. (2007). Control of matrix metalloproteinase catalytic activity. *Matrix Biology*, *26*(8), 587–596. https://doi.org/10.1016/j.matbio.2007.07.001

Rawlings, N. D. (2013). Identification and prioritization of novel uncharacterized peptidases for biochemical characterization. *Database*, *2013*. https://doi.org/10.1093/database/bat022

Rawlings, N. D. (2016). Peptidase specificity from the substrate cleavage collection in the MEROPS database and a tool to measure cleavage site conservation. *Biochimie*, *122*, 5–30. https://doi.org/10.1016/j.biochi.2015.10.003

Rawlings, N. D. (2020). Twenty-five years of nomenclature and classification of proteolytic enzymes. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1868*(2), 140345. https://doi.org/10.1016/j.bbapap.2019.140345

Rawlings, N. D., & Barrett, A. J. (1993). Evolutionary families of peptidases. *Biochemical Journal*, *290*(1), 205–218. https://doi.org/10.1042/bj2900205

Rawlings, N. D., Barrett, A. J., & Bateman, A. (2011). Asparagine peptide lyases: a seventh catalytic type of proteolytic enzymes. *Journal of Biological Chemistry*, *286*(44), 38321–38328. https://doi.org/10.1074/jbc.M111.260026

Rawlings, N. D., Barrett, A. J., & Finn, R. (2016). Twenty years of the *MEROPS* database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, *44*(D1), D343–D350. https://doi.org/10.1093/nar/gkv1118

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., & Finn, R. D. (2018). The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research*, *46*(D1), D624–D632. https://doi.org/10.1093/nar/gkx1134

Rawlings, N. D., & Bateman, A. (2019). Origins of peptidases. *Biochimie*, *166*, 4–18. https://doi.org/10.1016/j.biochi.2019.07.026

Rawlings, N. D., Tolle, D. P., & Barret, A. J. (2004). Evolutionary families of peptidase inhibitors. *Biochemical Journal*, *378*(3), 705–716. https://doi.org/10.1042/bj20031825

Rehman, A. A., Ahsan, H., & Khan, F. H. (2013). Alpha-2-macroglobulin: A physiological guardian. *Journal of Cellular Physiology*, *228*(8), 1665–1675. https://doi.org/10.1002/jcp.24266

Rhee, J. (2002). RECKing MMP function: implications for cancer development. *Trends in Cell Biology*, *12*(5), 209–211. https://doi.org/10.1016/S0962-8924(02)02280-8

Ricard-Blum, S., & Vallet, S. D. (2016). Proteases decode the extracellular matrix cryptome. *Biochimie*, *122*, 300–313. https://doi.org/10.1016/j.biochi.2015.09.016

Richards, F. M. (1992). Linderstrøm-Lang and the Carlsberg Laboratory: The view of a postdoctoral fellow in 1954. *Protein Science*, *1*(12), 1721–1730. https://doi.org/10.1002/pro.5560011221

Riddles, P. W., Blakeley, R. L., & Zerner, B. (1983). *[8] Reassessment of Ellman's reagent* (pp. 49–60). https://doi.org/10.1016/S0076-6879(83)91010-8

Robert-Genthon, M., Casabona, M. G., Neves, D., Couté, Y., Cicéron, F., Elsen, S., Dessen, A., & Attrée, I. (2013). Unique Features of a Pseudomonas aeruginosa α2-Macroglobulin Homolog. *MBio*, *4*(4). https://doi.org/10.1128/mBio.00309-13

Roderick, S. L., & Matthews, B. W. (1993). Structure of the cobalt-dependent methionine aminopeptidase from Escherichia coli: a new type of proteolytic enzyme. *Biochemistry*, *32*(15), 3907–3912. https://doi.org/10.1021/bi00066a009

Rose, T., LeMosy, E. K., Cantwell, A. M., Banerjee-Roy, D., Skeath, J. B., & di Cera, E. (2003). Three-dimensional Models of Proteases Involved in Patterning of the Drosophila Embryo. *Journal of Biological Chemistry*, *278*(13), 11320–11330. https://doi.org/10.1074/jbc.M211820200

Russell, J. J., Grisanti, L. A., Brown, S. M., Bailey, C. A., Bender, S. B., & Chandrasekar, B. (2021). Reversion inducing cysteine rich protein with Kazal motifs and cardiovascular diseases: The RECKlessness of adverse remodeling. *Cellular Signalling*, *83*, 109993. https://doi.org/10.1016/j.cellsig.2021.109993

Sabat, A., Kosowska, K., Poulsen, K., Kasprowicz, A., Sekowska, A., van den Burg, B., Travis, J., & Potempa, J. (2000). Two Allelic Forms of the Aureolysin Gene (*aur*) within *Staphylococcus aureus*. *Infection and Immunity*, *68*(2), 973–976. https://doi.org/10.1128/IAI.68.2.973-976.2000

Saheb, S. A. (1976). Purification et caractérisation d'une protéase extracellulaire de Staphylococcus aureus inhibée par l'E.D.T.A. *Biochimie*, *58*(7), 793–804. https://doi.org/10.1016/S0300-9084(76)80310-0

Salvesen, G. S., & Dixit, V. M. (1997). Caspases: Intracellular Signaling by Proteolysis. *Cell*, *91*(4), 443–446. https://doi.org/10.1016/S0092-8674(00)80430-4

Sam, P. N., Avery, E., & Claypool, S. M. (2019). Proteolytic Control of Lipid Metabolism. *ACS Chemical Biology*, *14*(11), 2406–2423. https://doi.org/10.1021/acschembio.9b00695

Sanaei, R., Kularathna, P. K., Taghavi, N., Hooper, J. D., Pagel, C. N., & Mackie, E. J. (2021). Protease-activated receptor-2 promotes osteogenesis in skeletal mesenchymal stem cells at the expense of adipogenesis: Involvement of interleukin-6. *Bone Reports*, *15*, 101113. https://doi.org/10.1016/j.bonr.2021.101113

Schechter, I., & Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications*, *27*(2), 157–162. https://doi.org/10.1016/S0006-291X(67)80055-X

Schrul, B., Kapp, K., Sinning, I., & Dobberstein, B. (2010). Signal peptide peptidase (SPP) assembles with substrates and misfolded membrane proteins into distinct oligomeric complexes. *Biochemical Journal*, *427*(3), 523–534. https://doi.org/10.1042/BJ20091005

Seemüller, E., Lupas, A., Stock, D., Löwe, J., Huber, R., & Baumeister, W. (1995). Proteasome from *Thermoplasma acidophilum* : a Threonine Protease. *Science*, *268*(5210), 579–582. https://doi.org/10.1126/science.7725107

Seidah, N. G. (2011). What lies ahead for the proprotein convertases? *Annals of the New York Academy of Sciences*, *1220*(1), 149–161. https://doi.org/10.1111/j.1749-6632.2010.05883.x

Seshagiri, P. B., Lalitha, H. S., Mishra, A., & Sireesha, G. v. (2003). Embryo-endometrial proteases during early mammalian development. *Indian Journal of Experimental Biology*, *41*(7), 756–763.

Shaw, L., Golonka, E., Potempa, J., & Foster, S. J. (2004). The role and regulation of the extracellular proteases of Staphylococcus aureus. *Microbiology*, *150*(1), 217–228. https://doi.org/10.1099/mic.0.26634-0

Silveira Corrêa, T. C., Massaro, R. R., Brohem, C. A., Taboga, S. R., Lamers, M. L., Santos, M. F., & Maria-Engler, S. S. (2010). RECK-mediated inhibition of glioma migration and invasion. *Journal of Cellular Biochemistry*, n/a-n/a. https://doi.org/10.1002/jcb.22472

Simoni, R. D., Hill, R. H., & Vaughan, M. (2002). Urease, the first crystalline enzyme and the proof that enzymes are proteins: the work of James B. Sumner. *The Journal of Biological Chemistry*, *277*(35), 23e.

Sims, A. H., Dunn-Coleman, N. S., Robson, G. D., & Oliver, S. G. (2004). Glutamic protease distribution is limited to filamentous fungi. *FEMS Microbiology Letters*, *239*(1), 95–101. https://doi.org/10.1016/j.femsle.2004.08.023

Sohail, A., Sun, Q., Zhao, H., Bernardo, M. M., Cho, J.-A., & Fridman, R. (2008). MT4-(MMP17) and MT6-MMP (MMP25), A unique set of membrane-anchored matrix metalloproteinases: properties and expression in cancer. *Cancer and Metastasis Reviews*, *27*(2), 289–302. https://doi.org/10.1007/s10555-008-9129-8

Sottrup-Jensen, L. (1989). α-Macroglobulins: structure, shape, and mechanism of proteinase complex formation. *Journal of Biological Chemistry*, *264*(20), 11539–11542. https://doi.org/10.1016/S0021-9258(18)80094-1

Steiner, J. P., Migliorini, M., & Strickland, D. K. (1987). Characterization of the reaction of plasmin with .alpha.2-macroglobulin: effect of antifibrinolytic agents. *Biochemistry*, *26*(25), 8487–8495. https://doi.org/10.1021/bi00399a068

Stojanovski, B. M., Pelc, L. A., & di Cera, E. (2020). Role of the activation peptide in the mechanism of protein C activation. *Scientific Reports*, *10*(1), 11079. https://doi.org/10.1038/s41598-020-68078-z

Strater, N. (1999). X-ray structure of aminopeptidase A from Escherichia coli and a model for the nucleoprotein complex in Xer site-specific recombination. *The EMBO Journal*, *18*(16), 4513–4522. https://doi.org/10.1093/emboj/18.16.4513

Suzuki, K., Enghild, J. J., Morodomi, T., Salvesen, G., & Nagase, H. (1990). Mechanisms of activation of tissue procollagenase by matrix metalloproteinase 3 (stromelysin). *Biochemistry*, *29*(44), 10261–10270. https://doi.org/10.1021/bi00496a016

Tagliabracci, V. S., Engel, J. L., Wiley, S. E., Xiao, J., Gonzalez, D. J., Nidumanda Appaiah, H., Koller, A., Nizet, V., White, K. E., & Dixon, J. E. (2014). Dynamic regulation of FGF23 by Fam20C phosphorylation, GalNAc-T3 glycosylation, and furin proteolysis. *Proceedings of the National Academy of Sciences*, *111*(15), 5520–5525. https://doi.org/10.1073/pnas.1402218111

Takahashi, C., Sheng, Z., Horan, T. P., Kitayama, H., Maki, M., Hitomi, K., Kitaura, Y., Takai, S., Sasahara, R. M., Horimoto, A., Ikawa, Y., Ratzkin, B. J., Arakawa, T., & Noda, M. (1998). Regulation of matrix metalloproteinase-9 and inhibition of tumor invasion by the membrane-anchored glycoprotein RECK. *Proceedings of the National Academy of Sciences*, *95*(22), 13221–13226. https://doi.org/10.1073/pnas.95.22.13221

Takeuchi, S., Saito, M., Imaizumi, K., Kaidoh, T., Higuchi, H., & Inubushi, S. (2002). Genetic and enzymatic analyses of metalloprotease (aureolysin) from Staphylococcus aureus isolated from domestic animals. *Veterinary Microbiology*, *84*(1–2), 135–142. https://doi.org/10.1016/S0378-1135(01)00448-5

Takeuchi, T., Hisanaga, M., Nagao, M., Ikeda, N., Fujii, H., Koyama, F., Mukogawa, T., Matsumoto, H., Kondo, S., Takahashi, C., Noda, M., & Nakajima, Y. (2004). The membrane-anchored matrix metalloproteinase (MMP) regulator RECK in combination with MMP-9 serves as an informative prognostic indicator for colorectal cancer. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, *10*(16), 5572–5579. https://doi.org/10.1158/1078-0432.CCR-03-0656

Tallant, C., García-Castellanos, R., Seco, J., Baumann, U., & Gomis-Rüth, F. X. (2006). Molecular Analysis of Ulilysin, the Structural Prototype of a New Family of Metzincin Metalloproteases. *Journal of Biological Chemistry*, *281*(26), 17920–17928. https://doi.org/10.1074/jbc.M600907200

Tallant, C., Marrero, A., & Gomis-Rüth, F. X. (2010). Matrix metalloproteinases: Fold and function of their catalytic domains. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1803*(1), 20–28. https://doi.org/10.1016/j.bbamcr.2009.04.003

Tang, B., Nirasawa, S., Kitaoka, M., Marie-Claire, C., & Hayashi, K. (2003). General function of N-terminal propeptide on assisting protein folding and inhibiting catalytic activity based on observations with a chimeric thermolysin-like protease. *Biochemical and Biophysical Research Communications*, *301*(4), 1093–1098. https://doi.org/10.1016/S0006-291X(03)00084-6

The PyMOL Molecular Graphics System (Version 1.2r3pre). (n.d.). Schrödinger,LLC. Retrieved July 1, 2022, from The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC.

Theocharis, A. D., Skandalis, S. S., Gialeli, C., & Karamanos, N. K. (2016). Extracellular matrix structure. *Advanced Drug Delivery Reviews*, *97*, 4–27. https://doi.org/10.1016/j.addr.2015.11.001

Titani, K., Hermodson, M. A., Ericsson, L. H., Walsh, K. A., & Neurath, H. (1972). Amino-acid Sequence of Thermolysin. *Nature New Biology*, *238*(80), 35–37. https://doi.org/10.1038/newbio238035a0

Travis, J., & Salvesen, G. S. (1983). HUMAN PLASMA PROTEINASE INHIBITORS. *Annual Review of Biochemistry*, *52*(1), 655–709. https://doi.org/10.1146/annurev.bi.52.070183.003255

Tsuyuki, H., Kajiwara, K., Fujita, A., Kumazaki, T., & Ishii, S. (1991). Purification and Characterization of Streptomyces griseus Metalloendopeptidases I and II. *The Journal of Biochemistry*, *110*(3), 339–344. https://doi.org/10.1093/oxfordjournals.jbchem.a123582

Turk, B. (2006). Targeting proteases: successes, failures and future prospects. *Nature Reviews Drug Discovery*, *5*(9), 785–799. https://doi.org/10.1038/nrd2092

Utz, P. J., & Anderson, P. (2000). Life and death decisions: regulation of apoptosis by proteolysis of signaling molecules. *Cell Death & Differentiation*, *7*(7), 589–602. https://doi.org/10.1038/sj.cdd.4400696

Vallee, B. L., & Auld, D. S. (1990). Active-site zinc ligands and activated H2O of zinc enzymes. *Proceedings of the National Academy of Sciences of the United States of America*, *87*(1), 220–224. https://doi.org/10.1073/pnas.87.1.220

van den Burg, B., & Eijsink, V. (2013). Thermolysin and Related Bacillus Metallopeptidases. In *Handbook of Proteolytic Enzymes* (pp. 540–553). Elsevier. https://doi.org/10.1016/B978-0-12-382219-2.00111-3

van Wart, H. E., & Birkedal-Hansen, H. (1990). The cysteine switch: a principle of regulation of metalloproteinase activity with potential applicability to the entire matrix metalloproteinase gene family. *Proceedings of the National Academy of Sciences*, *87*(14), 5578–5582. https://doi.org/10.1073/pnas.87.14.5578

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., … Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, *50*(D1), D439–D444. https://doi.org/10.1093/nar/gkab1061

Velasco, G., Pendás, A. M., Fueyo, A., Knäuper, V., Murphy, G., & López-Otín, C. (1999). Cloning and Characterization of Human MMP-23, a New Matrix Metalloproteinase Predominantly Expressed in Reproductive Tissues and Lacking Conserved Domains in Other Family Members. *Journal of Biological Chemistry*, *274*(8), 4570–4576. https://doi.org/10.1074/jbc.274.8.4570

Veltman, O. R., Eijsink, V. G. H., Vriend, G., de Kreij, A., Venema, G., & van den Burg, B. (1998). Probing Catalytic Hinge Bending Motions in Thermolysin-like Proteases by Glycine → Alanine Mutations. *Biochemistry*, *37*(15), 5305–5311. https://doi.org/10.1021/bi972374j

Vertyporokh, L., & Wojda, I. (2017). Expression if the insect metalloproteinase inhibitor IMPI in the fat body of Galleria mellonella exposed to infection with Beauveria bassiana. *Acta Biochimica Polonica*, *64*(2). https://doi.org/10.18388/abp.2016_1376

Vertyporokh, L., & Wojda, I. (2020). Immune response of Galleria mellonella after injection with non-lethal and lethal dosages of Candida albicans. *Journal of Invertebrate Pathology*, *170*, 107327. https://doi.org/10.1016/j.jip.2020.107327

Vickery, H. B. (1950). The origin of the word protein. *The Yale Journal of Biology and Medicine*, *22*(5), 387–393.

Vilcinskas, A. (2010). Coevolution between pathogen-derived proteinases and proteinase inhibitors of host insects. *Virulence*, *1*(3), 206–214. https://doi.org/10.4161/viru.1.3.12072

Vilcinskas, A., & Wedde, M. (2002). Insect Inhibitors of Metalloproteinases. *IUBMB Life (International Union of Biochemistry and Molecular Biology: Life)*, *54*(6), 339–343. https://doi.org/10.1080/15216540216040

Walsh, P. N., & Ahmad, S. S. (2002). Proteases in blood clotting. *Essays in Biochemistry*, *38*, 95–111. https://doi.org/10.1042/bse0380095

Wang, L., Main, K., Wang, H., Julien, O., & Dufour, A. (2021). Biochemical Tools for Tracking Proteolysis. *Journal of Proteome Research*, *20*(12), 5264–5279. https://doi.org/10.1021/acs.jproteome.1c00289

Wang, Z., Juttermann, R., & Soloway, P. D. (2000). TIMP-2 Is Required for Efficient Activation of proMMP-2 in Vivo. *Journal of Biological Chemistry*, *275*(34), 26411–26415. https://doi.org/10.1074/jbc.M001270200

Wedde, M., Weise, C., Kopacek, P., Franke, P., & Vilcinskas, A. (1998). Purification and characterization of an inducible metalloprotease inhibitor from the hemolymph of greater wax moth larvae, Galleria mellonella. *European Journal of Biochemistry*, *255*(3), 535–543. https://doi.org/10.1046/j.1432-1327.1998.2550535.x

Wedde, M., Weise, C., Nuck, R., Altincicek, B., & Vilcinskas, A. (2007). The insect metalloproteinase inhibitor gene of the lepidopteran Galleria mellonella encodes two distinct inhibitors. *Biological Chemistry*, *388*(1). https://doi.org/10.1515/BC.2007.013

Wegrzynowicz, Z., Heczko, P. B., Drapeau, G. R., Jeljaszewicz, J., & Pulverer, G. (1980). Prothrombin activation by a metalloprotease from Staphylococcus aureus. *Journal of Clinical Microbiology*, *12*(2), 138–139. https://doi.org/10.1128/jcm.12.2.138-139.1980

Wettstadt, S., & Llamas, M. A. (2020). Role of Regulated Proteolysis in the Communication of Bacteria With the Environment. *Frontiers in Molecular Biosciences*, *7*. https://doi.org/10.3389/fmolb.2020.586497

Wojda, I., Cytryńska, M., Zdybicka-Barabas, A., & Kordaczuk, J. (2020). Insect Defense Proteins and Peptides. In U. Hoeger & J. R. Harris (Eds.), *Vertebrate and Invertebrate Respiratory Proteins, Lipoproteins and other Body Fluid Proteins* (Vol. 94, pp. 81–121). Springer. https://doi.org/10.1007/978-3-030-41769-7_4

Wojda, I., & Taszłow, P. (2013). Heat shock affects host–pathogen interaction in Galleria mellonella infected with Bacillus thuringiensis. *Journal of Insect Physiology*, *59*(9), 894–905. https://doi.org/10.1016/j.jinsphys.2013.06.011

Woolley, V. C., Teakle, G. R., Prince, G., de Moor, C. H., & Chandler, D. (2020). Cordycepin, a metabolite of Cordyceps militaris, reduces immune-related gene expression in insects. *Journal of Invertebrate Pathology*, *177*, 107480. https://doi.org/10.1016/j.jip.2020.107480

Wyatt, A. R., Kumita, J. R., Farrawell, N. E., Dobson, C. M., & Wilson, M. R. (2015). Alpha-2-Macroglobulin Is Acutely Sensitive to Freezing and Lyophilization: Implications for Structural and Functional Studies. *PLOS ONE*, *10*(6), e0130036. https://doi.org/10.1371/journal.pone.0130036

Xia, H., Chen, S., Chen, K., Huang, H., & Ma, H. (2014). MiR-96 promotes proliferation and chemo- or radioresistance by down-regulating RECK in esophageal cancer. *Biomedicine & Pharmacotherapy*, *68*(8), 951–958. https://doi.org/10.1016/j.biopha.2014.10.023

Xu, M., Wang, H.-F., & Zhang, H.-Z. (2015). Expression of RECK and MMPs in Hepatoblastoma and Neuroblastoma and Comparative Analysis on the Tumor Metastasis. *Asian Pacific Journal of Cancer Prevention : APJCP*, *16*(9), 4007–4011. https://doi.org/10.7314/apjcp.2015.16.9.4007

Yagami, T., Yamamoto, Y., & Koma, H. (2019). Pathophysiological Roles of Intracellular Proteases in Neuronal Development and Neurological Diseases. *Molecular Neurobiology*, *56*(5), 3090–3112. https://doi.org/10.1007/s12035-018-1277-4

Yang, J. J., Artis, D. R., & van Wart, H. E. (1994). Differential Effect of Halide Anions on the Hydrolysis of Different Dansyl Substrates by Thermolysin. *Biochemistry*, *33*(21), 6516–6523. https://doi.org/10.1021/bi00187a019

Yiqing, L., Yong, Z., & Qichang, Z. (2005). Expression of RECK gene and MMP-9 in hilar cholangiocarcinoma and its clinical significance. *Journal of Huazhong University of Science and Technology [Medical Sciences]*, *25*(5), 552–554. https://doi.org/10.1007/BF02896015

Zhang, C., Jiang, M., Zhou, N., Hou, H., Li, T., Yu, H., Tan, Y.-D., & Zhang, X. (2021). Use tumor suppressor genes as biomarkers for diagnosis of non-small cell lung cancer. *Scientific Reports*, *11*(1), 3596. https://doi.org/10.1038/s41598-020-80735-x

Zhang, C.-X., Ye, L.-W., Liu, Y., Xu, X.-Y., Li, D.-R., Yang, Y.-Q., Sun, L.-L., & Yuan, J. (2015). Antineoplastic activity of Newcastle disease virus strain D90 in oral squamous cell carcinoma. *Tumor Biology*, *36*(9), 7121–7131. https://doi.org/10.1007/s13277-015-3433-z

Zhen, X., Lundborg, C. S., Zhang, M., Sun, X., Li, Y., Hu, X., Gu, S., Gu, Y., Wei, J., & Dong, H. (2020). Clinical and economic impact of methicillin-resistant Staphylococcus aureus: a multicentre study in China. *Scientific Reports*, *10*(1), 3900. https://doi.org/10.1038/s41598-020-60825-6

Zhou, D.-N., Deng, Y.-F., Li, R.-H., Yin, P., & Ye, C.-S. (2014). Concurrent alterations of RAGE, RECK, and MMP9 protein expression are relevant to Epstein-Barr virus infection, metastasis, and survival in nasopharyngeal carcinoma. *International Journal of Clinical and Experimental Pathology*, *7*(6), 3245–3254.

Supplementary Materials

Supplementary Material 1

"Molecular and in vivo studies of a glutamate-class prolyl-endopeptidase for coeliac disease therapy"

Molecular and *in vivo* studies of a glutamate-class prolyl-endopeptidase for coeliac disease therapy

Laura del Amo-Maestro[#], Soraia R. Mendes[#], Arturo Rodríguez-Banqueri, Laura Garzon-Flores, Marina Girbal[1,2], María José Rodríguez-Lagunas[1,2], Tibisay Guevara, Àngels Franch[1,2], Francisco J. Pérez-Cano[1,2], Ulrich Eckhard and F. Xavier Gomis-Rüth[*]

Proteolysis Laboratory; Department of Structural Biology; Molecular Biology Institute of Barcelona (CSIC); Barcelona Science Park; c/Baldiri Reixac, 15-21; 08028 Barcelona (Catalonia, Spain).
[1] Section of Physiology; Department of Biochemistry and Physiology; Faculty of Pharmacy and Food Science; University of Barcelona; Av. Joan XXIII, 27-31; 08028 Barcelona (Catalonia, Spain).
[2] Research Institute of Nutrition and Food Safety (INSA-UB); University of Barcelona; Av. Prat de la Riba, 171; 08921 Santa Coloma de Gramenet (Catalonia, Spain).
[*] Corresponding author: e-mail: xgrcri@ibmb.csic.es.
[#] Shared first co-authorship.

***Keywords*:** human gastric digestion; coeliac disease; proline-specific endopeptidase; 33-mer peptide; gluten; gliadin; proteolytic mechanism; X-ray crystal structure; zymogen; active-site mutant; glutenase

The digestion of gluten generates toxic peptides, among which a highly immunogenic proline-rich 33-mer from wheat α-gliadin, which trigger coeliac disease. Neprosin from the pitcher plant is a reported prolyl-endopeptidase. We produced recombinant neprosin and 11 mutants, and found that full-length neprosin is a zymogen, which is self-activated at gastric pH by the release of an all-β pro-domain via a pH-switch mechanism featuring a 'lysine plug'. The catalytic domain is an atypical 7+8-stranded β-sandwich with an extended active-site cleft containing an unprecedented pair of catalytic glutamates. Neprosin efficiently degraded both gliadin and the 33-mer *in vitro* under gastric conditions and was reversibly inactivated at pH>5. Moreover, co-administration of gliadin and the neprosin zymogen at 500:1 reduced the abundance of the 33-mer in the small intestine of mice by up to 90%. Neprosin therefore founds a new eukaryotic family of glutamate endopeptidases that fulfils requisites for a therapeutic 'glutenase'

Coeliac disease (CoD) is a chronic autoimmune enteropathy that affects individuals with genetic and environmental sensitization to dietary gluten, a group of cereal prolamin storage proteins rich in proline and glutamine [1, 2]. Prolamins that trigger CoD include gliadin and glutenin in wheat, hordein in barley, and secalin in rye. Intestinal damage can be inflicted by as little as ~10 mg of dietary gluten per day [3], which is <0.1% of the amount found in a typical western diet [2]. CoD is a global health burden across all age ranges, with a worldwide serological prevalence of 1.4% [4] that increases by 7.5% every year [5]. The disease is caused by partially degraded gluten peptides, including a 33-residue fragment of wheat α-gliadin ('33-mer') that is immunogenically the most relevant [2, 6]. These peptides resist further cleavage by gastric, pancreatic and intestinal brush-border membrane peptidases owing to their high proline content (13 in the 33-mer). In coeliacs, they cross the mucosal epithelium of the small intestine, where the glutamine residues are deamidated by tissue transglutaminase. This enhances the affinity of the peptides for the DQ2.5/DQ2.2 and DQ8 alleles of the human

leukocyte antigen (HLA) receptor, which are necessary for the development of CoD [2]. Receptor binding triggers a severe pro-inflammatory autoimmune response mediated by T cells, with intestinal effects including intraepithelial lymphocytosis, crypt hyperplasia, atrophy of small-intestine villi and mucosal inflammation [2]. These lead to the chronic malabsorption of nutrients, diarrhoea, vomiting, bloating, abdominal pain and intestinal lymphomas. Extra-intestinal manifestations include delayed puberty, osteoporosis, axonal neuropathy and cerebellar ataxia [7], which reduce the life expectancy of coeliacs. There is no treatment for CoD, so patients must adhere to a lifelong strict gluten-free diet, which restores the normal architecture of the intestinal villi [2]. However, gluten-free diets do not provide balanced nutrition [7], and many coeliacs suffer intestinal symptoms even with adherence to such dietary restrictions [8, 9]. Moreover, gluten is found in most processed foods and medicines, making dietary compliance challenging in western societies [2]. This has created a demand for effective CoD therapies.

One promising approach is the development of endopeptidases that cleave the toxic peptides and would thus act as *bona fide* 'glutenases' for oral enzyme therapy [10, 11, 12], reminiscent of lactase tablets for lactose intolerance [13]. Such an approach would also benefit patients suffering from non-coeliac gluten sensitivity, which has a worldwide prevalence of up to 13%, and irritable bowel syndrome, with a prevalence of <0.5% [8, 14, 15]. A candidate glutenase must fulfil certain criteria for clinical application. First, it should work in the stomach during digestion, before the gastric bolus passes into the duodenum and initiates the autoimmune response, and thus must remain stable and active in the acidic gastric environment (pH ~2.5) as well as resisting gastric pepsin. Second, a reasonable dose should efficiently digest gliadin and the 33-mer when combined with pepsin under gastric conditions, which requires the processing of large quantities of dietary protein. Third, it should not harm intestinal structures or inhibit nutrient absorption, and thus ideally should be inactive at the slightly acidic postprandial pH of the duodenum [16].

The therapeutic potential of several glutaminyl and prolyl endopeptidases (PEPs) has been assessed, representing various catalytic classes and diverse sources including bacteria, fungi, insects and germinating cereals [7, 10, 11, 12]. These include a serine PEP from *Aspergillus niger* [10, 17]; STAN1, a combination of *A. niger* aspartate aspergillopepsin and *Aspergillus oryzae* serine dipeptidyl-peptidase IV [18]; latiglutenase, a combination of a glutamine-specific cysteine peptidase from barley and a modified serine prolyl-specific oligopeptidase from *Sphingomonas capsulata* [19]; subtilisin-type serine endopeptidases from the natural oral colonizers *Rothia aeria* and *Rothia mucilaginosa* [11]; and the synthetic enzymes KumaMax and Kuma062/TAK-062, developed by the computational redesign of kumamolysin, a serine endopeptidase from the bacterium *Alicyclobacillus sendaiensis* [20]. However, none of these candidates fulfils all of the above requirements. The current frontrunners do not show high activity under gastric pH conditions and/or require extremely high doses or protective modifications such as PEGylation or microencapsulation. Accordingly, clinical trials have not yet achieved significant clinical remission in coeliacs and have not demonstrated the ability of these enzymes to replace a gluten-free diet [12]. Worse, many so-called enzyme preparations currently sold over the counter as CoD dietary supplements do not inactivate toxic gluten peptides and thus represent a hazard for coeliacs [21].

Neprosin is a 380-residue endopeptidase of unknown class and mechanism, currently assigned to family U74 in the MEROPS database (www.ebi.ac.uk/merops; [22]). It is a PEP that was discovered in the digestive fluid of the carnivorous plant *Nepenthes* × *ventrata*, which traps prey animals in its pitcher [23, 24, 25, 26]. The enzyme might have a function in

protein metabolism during prey digestion and/or defence [23]. In combination with other peptidases from the digestive fluid, it has been identified as part of a potential glutenase preparation [24]. Purified neprosin is also considered useful reagent for proteomics [25, 26].

Here, we established a human recombinant production system to produce high yields of neprosin. We determined its mechanism of activation *in vitro* as well as its thermal stability, pH profile, general proteolytic and peptidolytic activities, and susceptibility to a panel of peptidase inhibitors. We also tested cleavage of gliadin and the 33-mer *in vitro* to evaluate the ability of neprosin to act as a solo glutenase. Moreover, we evaluated the effect of recombinant neprosin on the processing of gliadin in mice. Finally, we crystallized and solved the structure of the neprosin zymogen and its mature form in product-mimicking complexes. These data revealed the mechanism of latency, the overall and active-site architectures, catalytic mechanism and peptidase class, which were validated by a cohort of mutants.

2. RESULTS AND DISCUSSION

Heterologous expression, autolytic maturation and stability analysis- Previous studies of neprosin mainly used the enzyme purified from pitcher plant fluid because heterologous expression in *Escherichia coli* produced only a partially impure enzyme with a 'modest yield' [24, 25]. We were unable to reproduce this approach so we developed a system based on human cells, assuming that eukaryotic post-translational processing is required. This yielded ~10 mg/L of pure well-folded full-length protein with a C-terminal hexahistidine (His_6) tag (41 kDa) or ~8 mg/L with a twin-streptavidin (Strep) tag (43 kDa) (Fig.1A,B). The protein was properly folded and remained stable for several weeks at 4°C in a neutral buffer, but lacked proteolytic activity, which we attributed to the full-length protein being the pro-neprosin zymogen. Indeed, it readily underwent autolytic maturation at bond P^{128}–S^{129} (residue numbering of neprosin in superscript; UniProt ID C0HLV2) over time when incubated in a highly acidic buffer, yielding the neprosin catalytic domain (CD) and the excised pro-domain (PD) (Fig.1E). The latter was eventually degraded, and both pro-neprosin and neprosin migrated as monomers when checked by calibrated size-exclusion chromatography (SEC) (Fig.1D).

Differential scanning fluorimetry using the thermofluor approach [27] revealed a midtransition temperature (T_m) of 68°C for the mature enzyme (Fig.2A), which is remarkable for a peptidase that works in an ambient temperature range and is more reminiscent of hyper-thermophilic enzymes [28]. Furthermore, the T_m of the zymogen was 9°C higher (Fig.2A), suggesting the PD promotes stability and, possibly, the correct folding of the full-length protein as reported for other zymogens [29]. This was supported by our inability to express mature neprosin (without the PD) using the same expression system. Finally, thermofluor studies in the presence of a reducing agent revealed an unfolding process with two transitions, the first occurring at 42–44°C (Fig. 2B). This indicated the existence of disulfide bonds that stabilize the protein, as discussed in more detail below.

Proteolytic activity- We investigated the effect of pH on the cleavage of fluorescent bovine serum albumin (BSA) by neprosin, using gastric pepsin, an aspartate peptidase, and pancreatic trypsin, a serine peptidase, for comparison (Fig.2C). The pH optimum of neprosin was 3, close to that of gastric pepsin (pH <2). By contrast, the pH optimum of trypsin was 8, a value at which both neprosin and pepsin were completely inactive. Pepsin was irreversibly inhibited at neutral pH, as previously reported [30], whereas neprosin was reversibly activated and inactivated by switching between pH 2.5 and 9.0. Moreover, neprosin was unaffected by freezing or

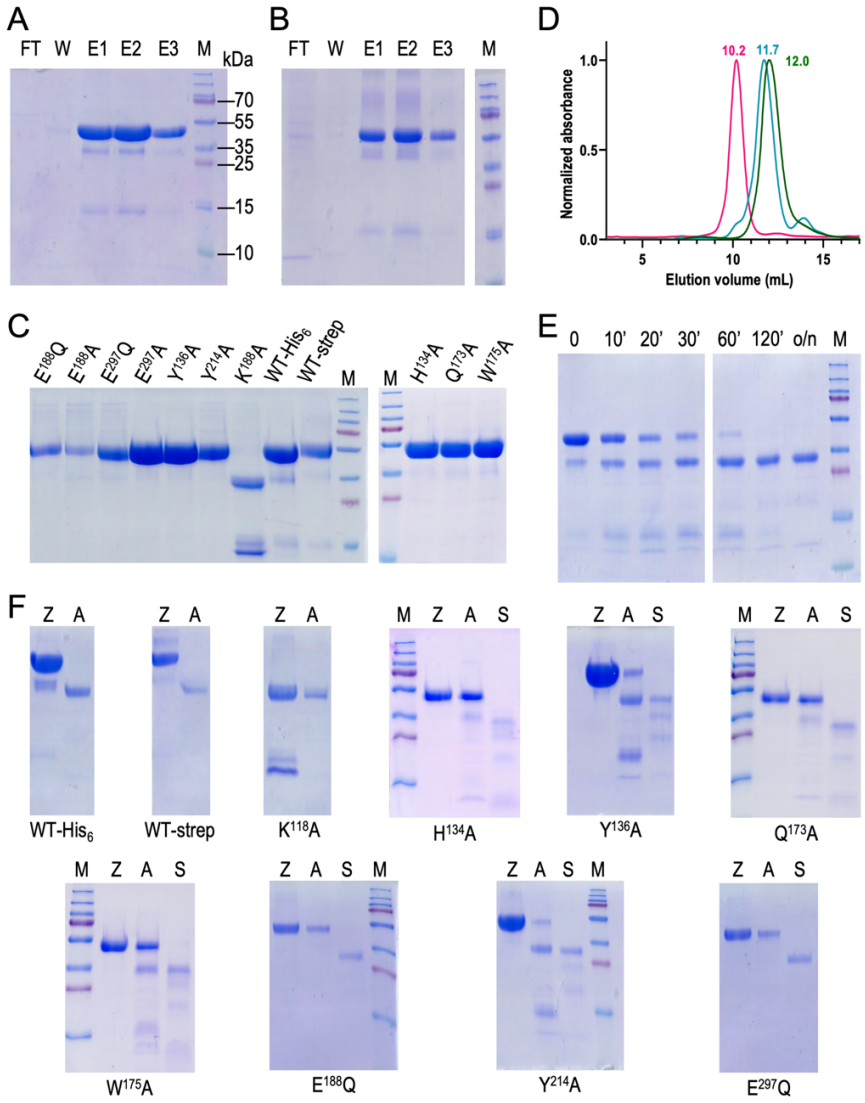


***Figure 1* — Protein purification and activation. (A)** Purification of wild-type (WT) pro-neprosin by His_6- or **(B)** Strep-tag affinity chromatography. The flow-through (FT), wash (W) and elution (E1–E3) fractions were analysed by SDS-PAGE and Coomassie staining, alongside molecular mass markers (lane M). **(C)** Pro-neprosin mutants ($K^{118}A$, $H^{134}A$, $Y^{136}A$, $Q^{173}A$, $W^{175}A$, $E^{188}A$, $E^{188}Q$, $Y^{214}A$, $E^{297}Q$ and $E^{297}A$) after His_6-tag affinity purification compared with the WT forms of (A) and (B). **(D)** Size exclusion chromatography profiles of pro-neprosin with His_6 tag (magenta), neprosin with Strep tag (green) and neprosin with His_6 tag (blue) separated on a Superdex 75 10/300 GL column. Each curve is labelled with the elution volume in mL, representing monomers in all cases. **(E)** Autolytic maturation of pro-neprosin over time at 37 °C in an acidic buffer. **(F)** Activation of pro-neprosin variants (Z lanes) by acidic autolysis (A lanes) or *in trans* by adding Strep-tagged neprosin (S lanes). Mutant $K^{118}A$ (third panel) was obtained as a pre-activated protein after affinity purification, revealing separate PD and mature protein bands (lane Z), which became fully activated by incubation in an acidic buffer (lane A).

lyophilization at pH 7.5 for storage, thus recovering its full activity after thawing or resuspension in an acidic buffer, respectively. Finally, neprosin was insensitive to cleavage by pepsin at acidic pH. This profile of activity, efficiency, stability and robustness was therefore consistent with a digestive enzyme that must work over prolonged timescales under varying conditions, precisely the natural environment in the pitchers of carnivorous plants [31].

To gain further insight into the substrate specificity of neprosin and to guide our cleavage assays, we reanalysed published proteomics data based mainly on purified material, which had identified the enzyme as a *bona fide* PEP [25]. We found 3001 unique cleavage sites spanning P_6–P_6' (substrate and active-site subsite nomenclature based on [32,

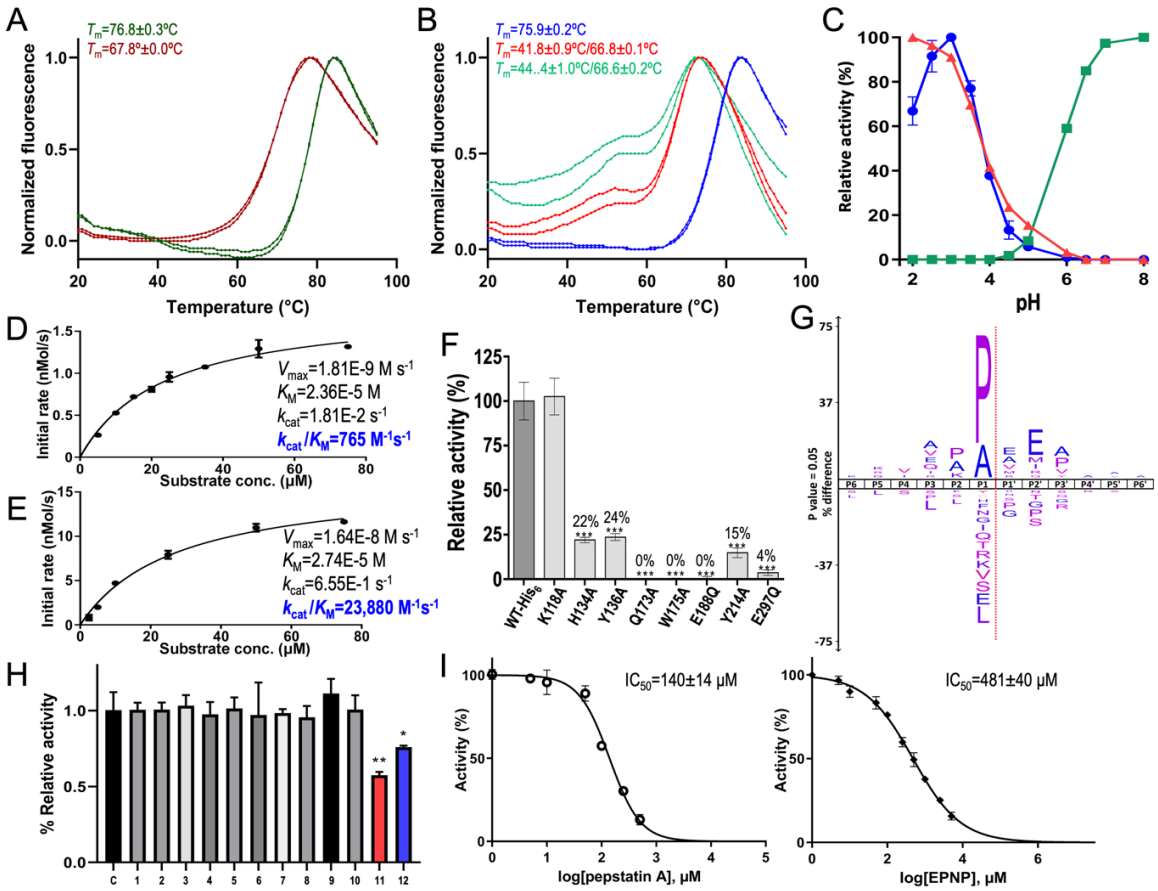


Figure 2 — **Thermal stability, peptidolytic activity and inhibition assays**. **(A)** Differential scanning fluorimetry showing duplicate curves of temperature-dependent fluorescence variation during the thermal denaturation of neprosin (dark red) and pro-neprosin (green). The inset midtransition temperatures (T_m) are the average inflection points of the two respective curves. **(B)** Same as (A), illustrating the effect of TCEP as a reducing agent at 5 mM (red) and 10 mM (green) compared with untreated pro-neprosin (blue). **(C)** The pH-dependent activity of pepsin (red), trypsin (green) and neprosin (blue) on a fluorescent BSA substrate. **(D, E)** Kinetics of the neprosin-mediated cleavage of the fluorogenic peptides (D) FS6 (100 nM neprosin) and (E) FS6-QPQL (25 nM neprosin). The insets show the corresponding V_{max}, k_{cat}, K_M and k_{cat}/K_M values. **(F)** Peptidolytic activity of wild-type (WT) neprosin and mutants on the fluorogenic FS6-QPQL peptide. Data in (D–F) are means ± SEM (n = 3). Statistical significance determined by Student's *t*-test (*p = 0.05, **p = 0.01, ***p = 0.001). **(G)** Logo depicting the substrate preference of neprosin based on reanalysis of deposited data [25]. **(H)** Effect of the test molecules or mixtures (*1*) 1,10-phenathroline, (*2*) AEBSF, (*3*) phosphoramidon, (*4*) marimastat, (*5*) cOmplete, (*6*) **BGP, (*7*) captopril, (*8*)** DAN, (*9*) **BEOPC, (*10*) AMP, (*11*) pepstatin A and (*12*)** EPNP compared to the WT control (C). Only the last two compounds achieve significant inhibition. Data are means ± SEM (n = 3). Statistical significance determined by Student's *t*-test (*p = 0.05, **p = 0.01, ***p = 0.001). **(I)** Plot of the inhibitory activity of pepstatin A (left) and EPNP (right) showing tester concentrations with the derived IC$_{50}$ values. Data are means ± SEM (n = 3).

33]), 1863 (62%) of which featured a proline residue in P_1 (Fig.2G). Proline was also enriched twofold over its natural abundance at P_2 and P_3', but was strongly disfavoured at P_1' and P_2'. Glutamate and methionine were enriched threefold at P_2', alanine was readily accepted throughout P_6–P_6', but glycine was significantly disfavoured at P_1–P_3'. These data revealed a strong preference for substrates with proline at P_1 and that specific positions within P_6–P_6' were unsuited for certain amino acids (Fig.2G).

Cleavage of gliadin and the 33-mer- We investigated the ability of neprosin to digest gliadin in the presence and absence of pepsin by SDS-PAGE and turbidimetry, compared to pepsin alone (Fig.3A,B). Both enzymes efficiently degraded gliadin separately at concentrations below ~5 µM, the physiological threshold of pepsin [34], but optimal results were achieved when both enzymes were combined. Remarkably, the

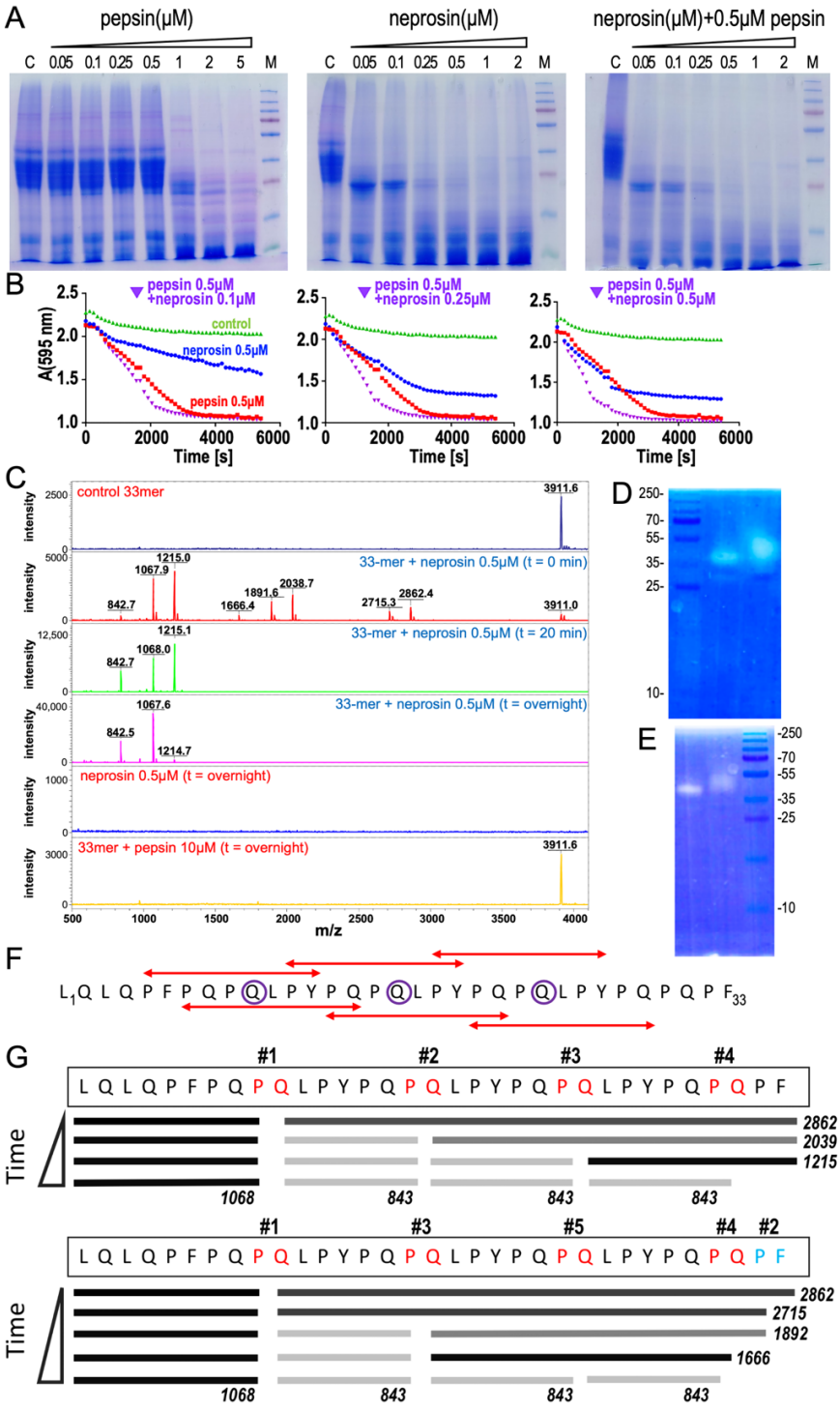


***Figure 3* — Neprosin activity against molecules relevant for coeliac disease. (A)** SDS-PAGE analysis of gliadin exposed to increasing concentrations of pepsin (left), neprosin (centre) or neprosin plus pepsin (right). **(B)** Curves depicting gliadin cleavage as in (A) over time measured by turbidimetry. **(C)** Mass spectra of, top to bottom, the 33-mer peptide (3912 Da); the 33-mer peptide after incubation with 0.5 μM neprosin for 0 min, 20 min and overnight; neprosin alone; and the 33-mer peptide after overnight incubation with 10 μM pepsin, which leaves the peptide intact. **(D)** Gliadin zymogram depicting the activity of neprosin (left lane) and the mature enzyme resulting from pro-neprosin self-activation (right lane). **(E)** Same as (D) but showing gelatin zymography. **(F)** Sequence of the 33-mer and extent of the six overlapping HLA-DQ2.5-binding epitopes as highlighted by red double arrows [7]. Glutamines susceptible to deamidation by transglutaminase are shown in purple circles [11]. The peptide corresponds to segment L^{76}–F^{108} of α-gliadin (UniProt ID P18573). **(G)** Cleavage of the 33-mer peptide by neprosin over time proceeds according to two pathways (top and bottom).

optimal concentration of neprosin was similar to that of gastric pepsin, and orders of magnitude lower than that required for current glutenase candidates. Zymography showed that neprosin degraded gliadin and gelatine, also a dietary protein, with similar efficiency (Fig.3D,E).

Next, we investigated cleavage of the 33-mer, which includes three glutamine residues that are deamidated by transglutaminase and six overlapping immunogenic HLA-DQ2.5 T-cell epitopes [6, 11, 35], by mass spectrometry (Fig.3F). We found that 250 µM of the peptide was efficiently degraded by 0.5 µM neprosin, a 500:1 ratio, after 20 min at pH 3 (Fig.3C). No autolytic cleavage products were detected even after overnight incubation, which confirmed the stability of the mature enzyme under acidic conditions. By contrast, pepsin failed to cleave the peptide even after overnight incubation at a 20-fold higher concentration than neprosin, which confirmed the resistance of the 33-mer against digestive peptidases. Analysis of the peptide cleavage fragments generated by neprosin revealed two final products: Q–L–P–Y–P–Q–P (843 Da) and L–Q–L–Q–P–F–P–Q–P (1068 Da). By monitoring the reaction over time (Fig.3G), we found that cleavage only occurred immediately downstream of five specific proline residues among the 13 present in the 33-mer, preferably at P–Q–P*Q–L–P and always with P–Q–P at the P_1–P_3 subsites, which qualifies the simple specificity for proline at P_1 deduced from indiscriminate proteomics and is in line with disfavouring large hydrophobic residues (leucine, phenylalanine and tyrosine) in P_2 as discussed above [25]. Overall, our results demonstrate that the 33-mer is degraded at multiple sites featuring the Q–P*Q–L motif. Remarkably, two P–Q dipeptides are also found in BSA, together with five equally favoured P–E sites (see above), which explains why albumin is a suitable substrate for neprosin at low pH.

Finally, we tested the cleavage of a cohort of fluorogenic peptides. We found that peptide FS6 containing a P–L bond (Mca–K–P–L–G–L–Dpa–A–R–NH_2), which is a substrate of matrix metalloproteinases and adamalysins [36], was cleaved with modest efficiency according to kinetic analysis (k_{cat}/K_M = 765 $M^{-1}s^{-1}$; Fig.2D). In contrast, peptide variant FS6-QPQL, redesigned to include the neprosin cleavage site of the 33-mer (Mca–Q–P–Q–L–Dpa–A–R–NH_2), was cleaved 30-fold more efficiently, mainly due to k_{cat} increase (k_{cat}/K_M = 23,880 $M^{-1}s^{-1}$; Fig.2E). Accordingly, neprosin can be defined as a PEP with a more constrained specificity than P–X that efficiently degrades the 33-mer under gastric-like conditions.

Inhibitory profile- Given the unknown catalytic class of the enzyme, we next tested a panel of peptidase inhibitors for their ability to block FS6-QPQL cleavage by neprosin **(Fig.2H)**. We also followed an approach recently applied to find inhibitors of pyrroline-5-carboxylate reductase [37], whose product is proline, and tested a series of **proline-containing/mimicking compounds. We found that only pepstatin A and** 2-[(4-nitrophenoxy)methyl]oxirane (**EPNP) weakly but significantly inhibited neprosin (Fig.2H), with half-maximal inhibitory concentration (IC_{50}) values of 140 and 480 µM, respectively (Fig.2I). Given that pepstatin and** EPNP-like epoxides are inhibitors of pepsin-type aspartate endopeptidases [38, 39], which share no sequence similarity with neprosin, this **pointed to an unexpected peptidase type and mechanism of catalysis for neprosin.**

Evaluation of neprosin activity *in vivo*- To investigate the activity of neprosin *in vivo*, mice were fed a bolus of gliadin 5 min after receiving either the zymogen at a very low mass ratio (1:500 w/w) or vehicle. After 2.5 h, we harvested the contents of three upper gastrointestinal tract segments and measured the concentration of the 33-mer by enzyme-linked immunosorbent assay (Fig.4). The peptide was substantially less abundant in all

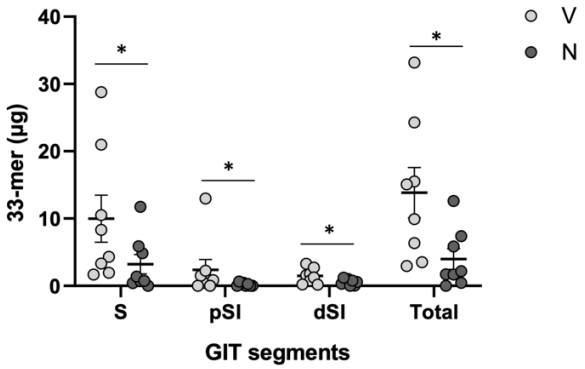


***Figure 4* —Analysis of neprosin activity against gliadin *in vivo*. (A)** Amount of 33-mer (µg) in the total contents of the stomach (S), proximal small intestine (pSI) and distal small intestine (dSI) of mice receiving neprosin zymogen (N) or vehicle (V) prior to a bolus of gliadin. Results are means ± SEM (n = 8 animals per group). The statistical significance was determined by the F-test (*, $p < 0.05$, N *vs* B).

segments of the treated animals (61-91%) and by 71% overall. The inactive zymogen is therefore activated upon reaching the stomach and efficiently helps to break down gliadin (and particularly the 33-mer) *in vivo* while remaining resistant to physiological digestive enzymes. This occurs at much lower concentrations than those of candidate glutenases, and without protective strategies such as PEGylation or microencapsulation. These results are consistent with a previous study reporting that sensitized NOD/DQ8 mice showed a significant decrease in inflammatory markers when fed gliadin that was pre-digested with pepsin and *Nepenthes* pitcher fluid, which included among other components neprosin and nepenthesin [24].

Structural analysis of latent and mature neprosin- We crystallized pro-neprosin in an orthorhombic space group (Fig.5A and Suppl. Table1) and found that the polypeptide was cleaved at the physiological maturation site (P^{128}–S^{129}). The crystals therefore contained the zymogenic complex of the cleaved PD and the CD (Fig.5A). We solved the structure by single-wavelength anomalous diffraction, collecting data at the lutetium L_{III} absorption edge wavelength from a crystal soaked in Lu-Xo4 [40] (Fig.5B). This soaking led to significant variation in one of the crystal cell axes when compared to native crystals while keeping good diffraction of X-rays (Suppl. Table1). The final refined model of the derivative complex was used to solve the native pro-neprosin structure by molecular replacement. Moreover, mature neprosin produced two different monoclinic crystal forms, I and II (Fig.5A and Suppl. Table1), whose structures were likewise solved by molecular replacement.

Pro-neprosin is a compact oblong molecule of ~55 × ~45 × ~40 Å (Fig.5C). The N-terminal PD (R^{25}–P^{128}) is defined in the final Fourier map from A^{29} onwards, and features a globular part (A^{29}–G^{112}) followed by a linker (L^{113}–P^{128}) to the downstream CD (S^{129}–Q^{380}). Segment (N^{122}–N^{131}), which includes the cleaved maturation site, is flexible. The PD features an antiparallel three-stranded β-sheet in which the central strand is bisected by the insertion of the leftmost strand (Fig.5C,D). The two right strands are connected by a long segment on the top, which includes two short α-helices, a disulfide bond (C^{52}–C^{98}), and a disordered 10-residue segment (Y^{77}–N^{86}) at the back of the molecule. The latter probably results from the protruding glycan chain attached to N^{152} within a back strand of the CD (Fig.5C). A second glycan is attached to N^{145} from a cross-over loop on top of the CD. Beyond the last strand of the PD, the chain undergoes a 90° turn and enters the PD/CD linker, which runs in extended conformation along the front surface of the CD.

Atypically for peptidases, which are generally α/β-proteins [41], the CD is an antiparallel β-sandwich, with a seven-stranded strongly-curled front sheet and an eight-stranded back sheet, which provides a scaffold for the former (Fig.5C,D). Both sheets are inter-connected by nine cross-over loops, including long hairpin β12β13, and two further disulfide bonds (C^{219}–C^{224} and C^{358}–C^{379}) on either side of the sandwich (Fig.5D). All these elements contribute to a compact and sturdy structure, which explains the remarkable pH stability of neprosin and its ability to resist pepsin

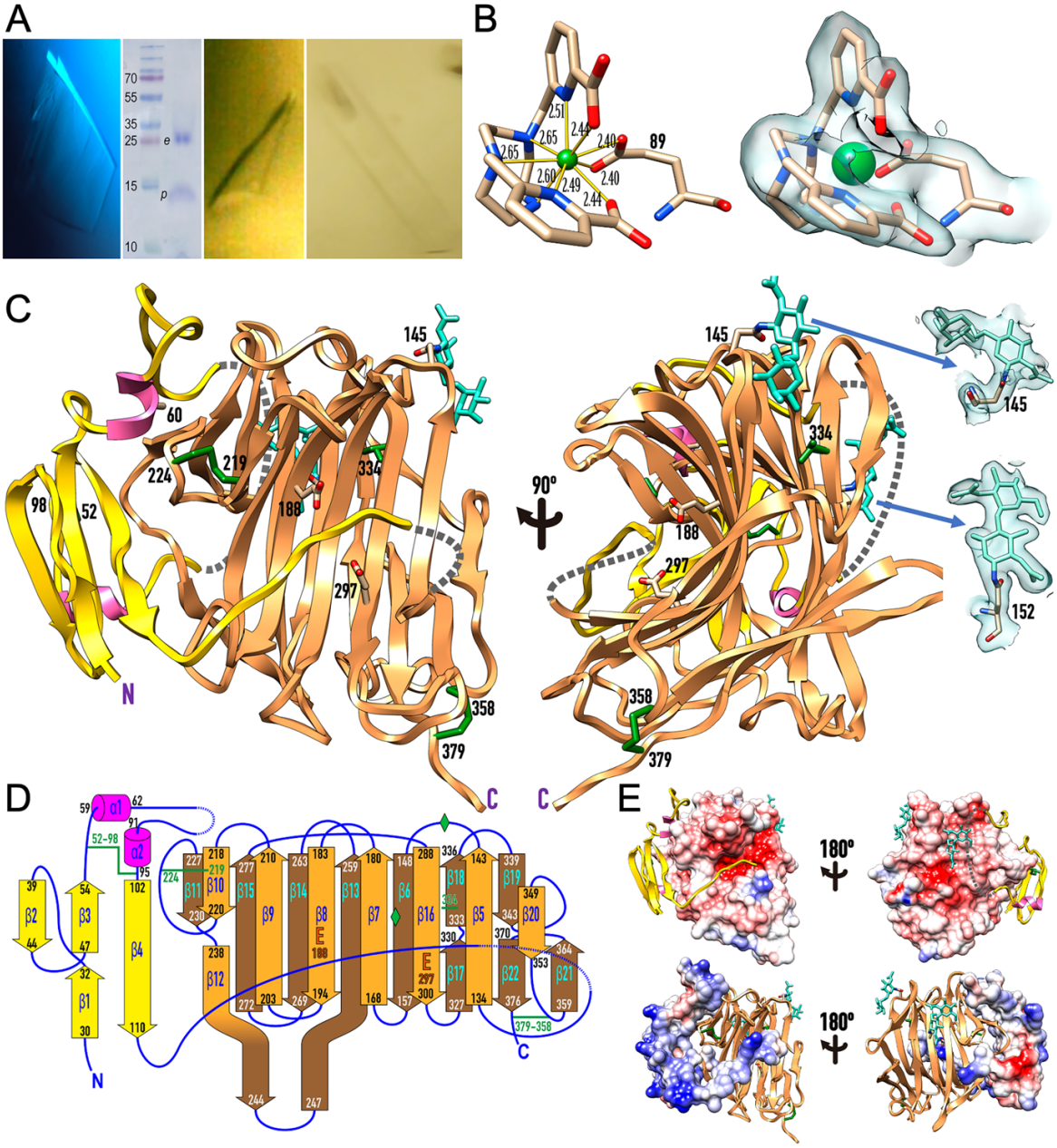


***Figure 5* — Structures of pro-neprosin and neprosin. (A)** Orthorhombic crystals of pro-neprosin (left panel) contained the cleaved PD (*p*) and the CD (*e*) (centre-left panel). Mature enzyme crystals were monoclinic (centre-right panel, crystal form I; right panel, crystal form II). **(B)** The structure of pro-neprosin was solved using a lutetium derivative. At one site (left panel), the Lu^{3+} cation (green sphere) was nona-coordinated by two carboxylate oxygens plus five nitrogen atoms from the organic scaffold and the carboxylate oxygens of protein residue E^{89} at distances spanning 2.40–2.65 Å. The cation-binding site was unambiguously defined in the final ($2mF_{obs}$-DF_{calc})-type Fourier map of the derivative contoured at 1.3 σ (right panel). **(C)** Ribbon-type plot of pro-neprosin viewed from the frontal (left panel) and lateral (right panel) perspectives. The PD is gold with magenta helices. The mature enzyme is shown in salmon. Disordered/cleaved segments are indicated by grey dashed lines. The two glycosylation sites at N^{145} and N^{152}, the seven cysteines, A^{60}, and the two catalytic glutamates (E^{188} and E^{297}) are shown with their side chains and labelled. The final Fourier map around the two glycan chains is depicted at 0.6 σ. **(D)** Topology of pro-neprosin with strands as arrows (labelled β1–β22) and the two short helices (α1 and α2) as magenta rods. The terminal residues of each secondary structure element are indicated. The PD has yellow strands and magenta helices, the front sheet of the mature enzyme moiety is in orange, and the back sheet is in brown. The seven cysteines are further indicated in green, the glycans are shown as green rhombi. The catalytic glutamates are marked for reference. **(E)** The top row shows the front view of pro-neprosin as in (C) (left) and the back view (right), both depicting the PD as yellow ribbon and the Coulombic surface of the CD (*red*, –10 kcal/mol·*e*; *blue*, +10 kcal/mol·*e*) computed with *Chimera* [85]. The calculated pI of the mature enzyme component is 4.3. The bottom row shows the same except that here the PD is shown for its Coulombic surface (pI = 9.5) and the CD as salmon ribbon.

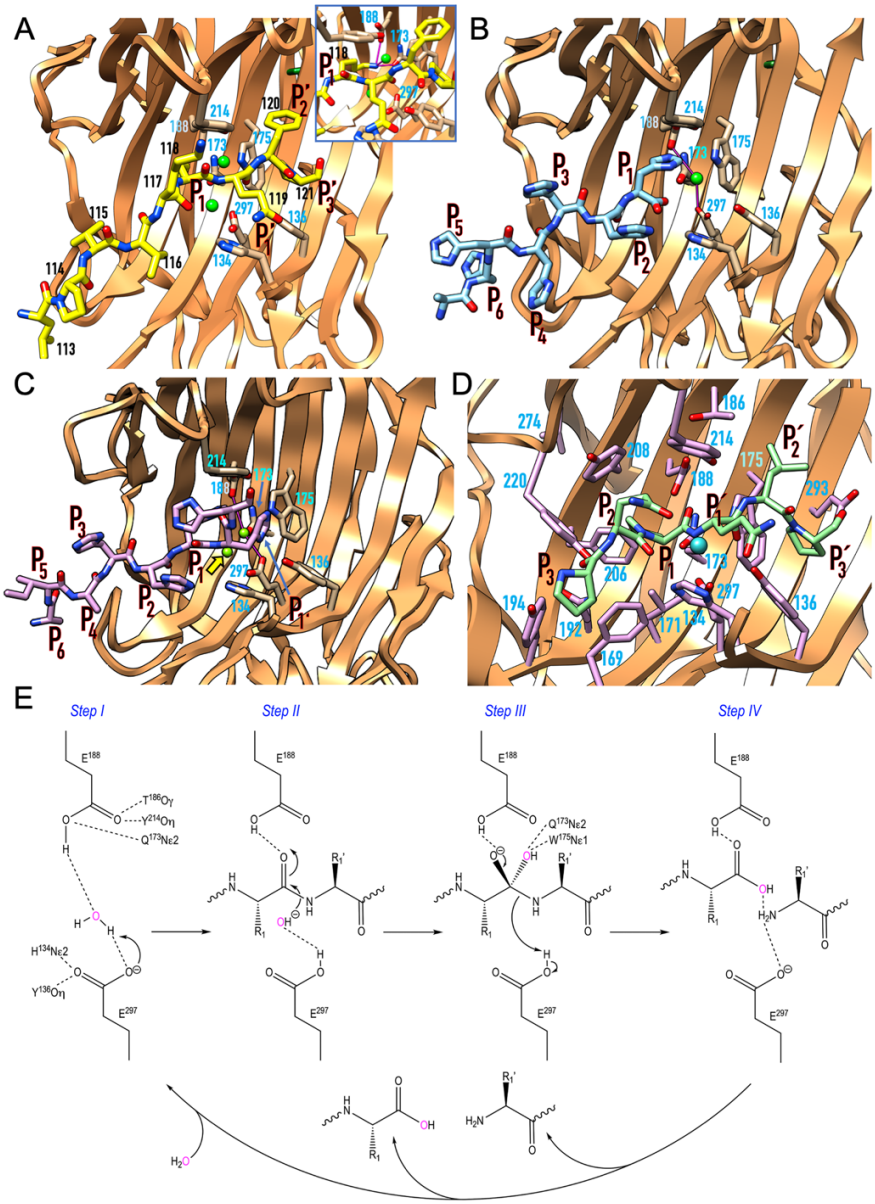


***Figure 6* — The active site and proposed mechanism. (A)** Close-up of Fig. 5C depicting the final segment (L^{113}–P^{121}) of the PD defined in the final Fourier map as a stick model with yellow carbons and black residue numbers running across the active-site cleft. The likely $\mathbf{P_1}$ and $\mathbf{P_1'}$–$\mathbf{P_3'}$ residues are labelled. In addition, selected residues of the active site are depicted for their side chains with carbons in tan and numbered in light blue. Two solvent residues potentially relevant for catalysis are shown as green spheres. The inset provides a slightly rotated close-up view to highlight the interaction (magenta lines) of K^{118} with E^{188}, Q^{173} and a solvent molecule. **(B)** Same as (A) depicting the product complex of mature neprosin (crystal form I), with the C-terminal tail from a symmetry mate spanning A^{403} and the His_6-tag residues (H^{404}–H^{409}) as a stick model with carbons in cyan featuring substrate subsites $\mathbf{P_6}$–$\mathbf{P_1}$. A solvent molecule potentially relevant for catalysis (green sphere) bridges E^{297} and E^{188} (magenta sticks). **(C)** Same as (B) for crystal form II. The C-terminal tail from a symmetry mate spanning A^{401} and part of the His_6-tag (H^{404}–H^{408}) is shown as a stick model with carbons in plum, probably covering subsites $\mathbf{P_6}$–$\mathbf{P_1'}$. A solvent molecule potentially relevant for catalysis (green sphere) bridges E^{297} and E^{188} (magenta sticks). A second solvent molecule (yellow arrow) probably occupies the position of the scissile carbonyl oxygen in the Michaelis complex. The polypeptide chains of both crystal forms overlap for tag residues H^{404}–H^{409} (crystal form I) and A^{401}–H^{406} (crystal form II) upon superposition of the respective CDs. **(D)** Model of the likely Michaelis complex between a substrate spanning residues P–Q–P*Q–L–P (green carbons) at positions $\mathbf{P_3}$–$\mathbf{P_3'}$ and the active site of neprosin. Selected residues are displayed for their side chain (plum carbons) and labelled. The catalytic solvent is depicted as a cyan sphere. **(E)** Proposed chemical mechanism of substrate cleavage by neprosin.

digestion. By contrast, the disulfide bonds are not deeply buried in the structure, which explains its sensitivity to reducing agents. The structure of mature neprosin crystal form I (Suppl. Table1) proved practically identical to the equivalent part of the zymogen, with a core root mean square deviation (RMSD) of 0.62 Å. The only significant difference was encountered at N^{232}–Y^{233}, which is folded outward in the zymogen to accommodate I^{103} at the beginning of the rightmost strand of the PD. Crystal form II, in turn, was practically indistinguishable from crystal form I (core RMSD = 0.66 Å) except for the tip of loop Lβ21β22, which was spaced apart by 3.8 Å, and the C-terminal tag, which was reoriented owing to crystal packing. Thus, the mature enzyme component is essentially preformed in the zymogen as seen in most peptidases, with the notable exception of chymotrypsin-type serine peptidases [42, 43, 44].

The active site- We hypothesized that the active-site cleft would be delineated by the PD linker (Fig.6A) as found in other zymogens [42, 44]. Moreover, in the structures of mature neprosin crystal forms I and II, the C-terminal segment, which spanned an alanine–isoleucine–alanine tripeptide followed by the His_6-tag, ran along the surface of a symmetry mate, thus mimicking product complexes. Both crystal forms were monoclinic but with different cell constants (Suppl. Table1), which resulted in variable crystal packing. Even so, the C-terminal tag penetrates the cleft in a similar manner in both crystallographic arrangements but is shifted by three positions, so that H^{404}–H^{409} from crystal form I overlaps A^{401}–H^{406} from crystal form II (Fig.6B,C). Accordingly, neprosin would possess an extended active-site cleft traversing the concave face of the sheet, which is oblique to the direction of the front-sheet β-strands by ~55° (Figs.6A-C).

On the search for possible catalytic residues, we were inspired by the functionally analogous pepsin-type aspartic peptidases, which despite their disparate architecture are likewise mainly β-proteins and operate at extremely acidic pH [45]. Moreover, the only (weak) neprosin inhibitors we could find are also known to inhibit aspartate peptidases (see above). These enzymes use a pair of aspartic residues bridged by a solvent molecule for catalysis [46]. Indeed, we found a striking pair of glutamate residues (E^{188} and E^{297}) bridged by a solvent molecule pinching the bound peptides in the product complexes (Fig.6B,C). In crystal form II, a clearly resolved second solvent molecule would replace the scissile carbonyl oxygen of a substrate (Fig.6C). The glutamate pair was similarly arranged in the zymogen structure, albeit slightly farther apart (Fig.6A). We therefore produced $E^{188}Q$, $E^{188}A$, $E^{297}Q$ and $E^{297}A$ point mutants of His_6-tagged pro-neprosin for testing (Fig.1C). These variants did not autoactivate when incubated at acidic pH (Fig.1F), so activation was triggered *in trans* using catalytic amounts of mature Strep-tagged wild-type neprosin. Finally, we obtained well-folded and intact mature variants of the $E^{188}Q$ and $E^{297}Q$ mutants, but not $E^{297}A$ or $E^{188}A$ (Fig.1F), and these indeed were catalytically inactive (Fig.2F). E^{188} and E^{297} may therefore act as a catalytic dyad, revealing that neprosin is a glutamate peptidase, a catalytic class that (in contrast to the aspartate peptidases) has been studied very poorly [47]. This is in agreement with very recent predictions based on bioinformatics studies but not validated experimentally [48]. Our results suggest that the mature neprosin structures mimic upstream product complexes collectively occupying subsites S_6 to S_1' (Fig.6B,C), with the two catalytic glutamates plus the bridging solvent molecule poised for reaction. As the PD linker extends for further residues on the right side of the cleft in the zymogen (Fig.6A), these would correspond to positions up to P_3'. Thus, together with the extra space in the cleft beyond S_3', neprosin would feature an extended cleft, probably spanning up to 11 subsites (S_6–S_5'), which explains the need for extended peptides beyond the scissile bond (see above).

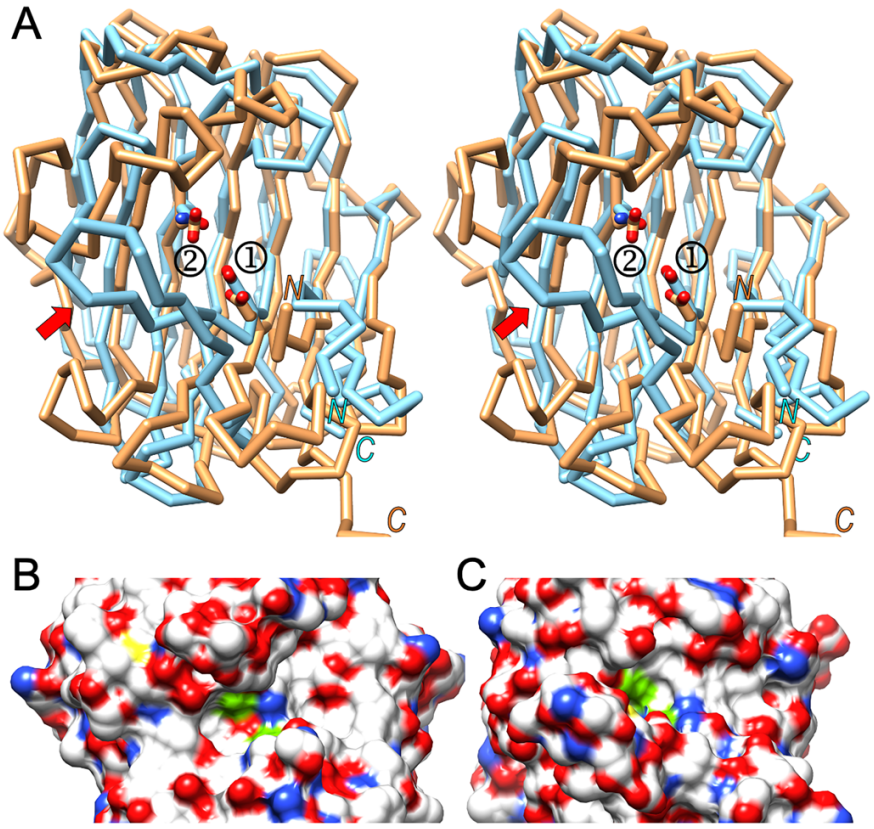


***Figure 7* — Structural similarity of neprosin and eqolysins. (A)** Superposition of the Cα-traces of neprosin (salmon) and SCP-B (pale blue) in stereo, with the respective catalytic residues shown as sticks and labelled (□, E^{297}/E_{190} of neprosin/SCP-B; □, E^{188}/Q_{107} of neprosin/SCP-B). Note the unique flap of SCP-B covering the active-site cleft (red arrow). The N-terminus and C-terminus are indicated. **(B)** Close-up of the active-site cleft of neprosin shown for its Connolly surface in the orientation of (A). The two catalytic residues are shown (green patches). **(C)** Same as (B) for SCP-B.

In the absence of a substrate complex, we constructed a model for the Michaelis complex of neprosin with the P–Q–P*Q–L–P peptide based on the zymogen and product–complex structures (Fig.6D). This model revealed further residues in the proximity of the catalytic glutamates with potential binding or catalytic functions. We therefore mutated residues H^{134}, Y^{136}, Q^{173}, W^{175} and Y^{214} (Fig.6A-D) by replacing them with alanine, and purified the corresponding proteins (Fig.1C). All the mutants required activation *in trans* as discussed above (Fig.1F). The activity of $H^{134}A$, $Y^{136}A$ and $Y^{214}A$ was ~80% lower than the wild-type enzyme, whereas mutants $Q^{173}A$ and $W^{175}A$ were totally inactive (Fig.2F). We conclude that H^{134}, Y^{136} and Y^{214} are relevant but not critical for catalysis, possibly playing an ancillary role in the catalytic mechanism, whereas Q^{173} and W^{175} are essential (see below).

Mechanisms of latency and activation- The PD attaches laterally to the left side of the mature enzyme so that its central β-sheet is rotated ~90° away from the plane of the front sheet. The inter-domain surface has a solvation free energy gain upon interface formation (Δ^iG) of –25.8 kcal/mol [49], indicating a very strong interaction. Furthermore, the complex buries 2176 $Å^2$, which exceeds the reported average value of 1910 $Å^2$ for protein–protein complexes [50]. The PD (theoretical pI = 9.5; Fig.5E, *bottom*) is crescent-shaped and snuggly embraces the CD (pI = 4.3; Fig.5E, *top*) under electrostatic complementation, which contributes to activity repression and zymogen stability (pI

= 5.9) at neutral or slightly acidic pH values. Moreover, the intimate zymogenic interaction further explains the remarkable stability of pro-neprosin in thermal shift assays (see above). Finally, the importance of the PD was further assessed by testing point mutant $A^{60}R$, designed to destabilize the interface (Fig.5C, *left panel*). This mutation prevented the isolation of a folded protein.

Once secreted to the acidic digestive fluid, the protonation of negatively charged residues leads to repulsion of net positive charges so that the zymogen falls apart under liberation of the preformed mature moiety and the active-site cleft. The $\mathbf{S_1}$ position of the cleft is occupied by K^{118} from the PD linker in the zymogen structure, which was obtained at pH 7.5. This residue forms a strong salt bridge with catalytic E^{188} and a hydrogen bond with Q^{173}Nε2, which is essential (see above and Fig.6A, *inset*). We therefore produced and tested mutant $K^{118}A$, which was efficiently overexpressed but underwent partial autolytic maturation in a neutral buffer, conditions under which the wild-type enzyme and other mutants remained intact (Fig.1C). Subsequent incubation at pH 2.5 completed the activation process (Fig.1F). As expected, the activity of the mature mutant was similar to that of the wild-type enzyme (Fig.2F).

Based on the above, we propose that the K^{118}–E^{188} pair features a 'latency plug' that may be weakened once the zymogen reaches acidic environment by following a pH-switch mechanism, so the PD linker is pulled out for maturation cleavage. This is reminiscent of the digestive aspartate peptidases pepsin and gastricsin, which feature a lysine residue functionally equivalent to K^{118} [43, 51], and of the lysosomal peptidase legumain [52]. The pH-switch mechanism, and the fact that the scissile P_1'–P_1 peptide bond is sandwiched by the Y^{214} side chain so that it is not accessible for cleavage (Fig.6A), explains why the zymogen linker can bind in the direction of a substrate to the cleft at neutral pH without being cleaved. This contrasts with most zymogens, including digestive aspartate peptidases, in which pro-segments interact in a non-substrate-like manner with the mature enzyme residues as a mechanism to prevent untimely activation [42, 43, 44]. Finally, given that the scissile-bond position in the cleft is occupied by K^{118}–Q^{119} but maturation occurs at P^{128}–S^{129}, activation probably occurs *in trans* by a second enzyme molecule once the PD linker is released from the cleft.

Proposed catalytic mechanism- Based on the preceding results, the catalytic cleavage mechanism of neprosin would proceed as follows. The solvent bridging the E^{188} and E^{297} carboxylates in the product complexes would represent the catalytic water in the ground state (Fig.6E, *step I*). The water is closer to E^{297}, which suggests that E^{188} may be protonated, as reported for one of the two catalytic aspartates in pepsin-type acidic peptidases [46]. E^{297} is kept in place by hydrogen bonding with H^{134}Nε2 and Y^{136}Oη, whereas E^{188} is kept in place by hydrogen bonding with Y^{214}Oη, T^{186}Oγ and Q^{173}Nε1. During the reaction, the substrate would bind to the active-site cleft in its extended conformation (Fig.6E, *step II*), with the $\mathbf{S_3}$, $\mathbf{S_1}$ and $\mathbf{S_3'}$ subsites of the cleft being shaped by Y^{194}, Q^{192} and F^{169}; Y^{208}, Y^{206}, L^{171}, E^{188} and Q^{173}; and Y^{136}, W^{175} and E^{293}, respectively, which are ideal for the accommodation of prolines (Fig.6D). The substrate main chain would be fixed by hydrogen bonds between its carbonyls and Y^{220}Oη in P_3, H^{134}Nε2 in P_2 (enabled by a 180º rotation around χ_2 upon by substrate binding), and Y^{136}Oη in P_1' (Fig.6D). Substrate insertion would shift the catalytic solvent further towards E^{297}, which would act as a general base and abstract a proton from the solvent to enhance its nucleophilicity. The protonated E^{188} carboxylate, in turn, would bind the scissile carbonyl oxygen (Fig.6D,E). Thereafter, the polarized solvent would perform a nucleophilic attack on the *si*-face of the scissile carbonyl carbon, which would result in a tetrahedral *gem*-diolate reaction intermediate (Fig.6E, *step III*). The latter would be stabilized by indispensable W^{175}Nε1

and Q^{173}Nε1, in the critical role of an oxyanion hole [53]. The intermediate would then resolve by breaking the scissile C–N bond. At this stage, E^{297} would act as a general acid and protonate the new α-amino nitrogen (Fig.6E, *step IV*). Finally, the two cleavage products would leave the cleft and the enzyme would be poised for a new round of catalysis.

Structural similarity with eqolysins- Peptidases were originally assigned to five mechanistic classes: the serine, cysteine, threonine, aspartate, and metal-dependent peptidases [54]. In 2004, the first glutamate peptidase was structurally characterized, namely scytalidocarboxyl peptidase B (SCP-B) from the dematiaceous fungus *Scytalidium lignicolum* [55, 56, 57]. Since that pioneering report, only the closely related aspergilloglutamic peptidase (~50% identical to SCP-B) has been structurally analysed [58, 59], and seven others have been functionally assessed, mostly from fungi [60, 61, 62, 63, 64] but one from a bacterium [65]. They are assigned to family G1 in the MEROPS database and are informally known as the 'pepstatin-insensitive fungal carboxylpeptidase group' [66] or eqolysins [55]. They are thermophilic and pepstatin-insensitive enzymes that function under acidic conditions [65] and feature a catalytic glutamate acting as a solvent-polarising general base, which is E_{190} in SCP-B (see UniProt ID P15369 for residue numbering in subscript according to the full-length protein, and subtract 54 for the commonly used mature enzyme numbering [55, 56]). The glutamate is assisted by a glutamine (Q_{107} in SCP-B), hence the family name eqolysins [55]. These residues are invariant within the family and are flanked by very similar residues [66, 67].

Archetypal SCP-B is a 7+7 antiparallel β-sandwich that shows overall similarity with the neprosin CD (Fig.7A). Superposition of neprosin and the bound mature form of SCP-B (Protein Data Bank [PDB] ID 2IFR [56]), whose zymogenic structure is unknown, revealed 140 aligned residues with a rather large core RMSD of 3.0 Å and a sequence identity of only 11%. There are remarkable differences in the connecting loops and the active site, e.g. a large disulfide-linked protruding β-hairpin inserted in the fungal enzyme following the β-strand equivalent to β16 in neprosin (Fig.7A). Within the active site, the only conserved residue is the catalytic glutamate (E^{297} in neprosin and E_{190} in SCP-B), as well as the position of the catalytic assistant (E^{188} in neprosin and Q_{107} in SCP-B), which lead to variable active-site clefts with disparate substrate trajectories and surface profiles (Fig.7B,C). Moreover, cross-mutants Q_{107}E of SCP-B and E^{188}Q of neprosin, which mimic each other's catalytic dyad, are completely inactive, as discussed above and reported in [68]. This explains the different substrate specificities, which in SCP-B leads to the cleavage of F–F, L–Y and F–Y bonds in insulin but not proline-flanking bonds [68].

Corollary- Current glutenases have limitations in meeting the stringent criteria for efficient oral enzyme therapy against CoD. Here, our *in vitro* and *in vivo* studies showed that recombinant neprosin is a robust pepsin-resistant enzyme that very efficiently degrades gliadin and its 33-mer under laboratory-simulated gastric conditions and in the mouse stomach. Low doses of the enzyme therefore complement gastric pepsin during digestion. Our results demonstrate that the Q–P*Q–L motif of the 33-mer is readily cleaved, which removes all six overlapping immunogenic epitopes by generating peptides too small to stimulate the division of gliadin-specific T cells [69]. The cleavage efficiency of neprosin *in vitro* under simulated gastric conditions is orders of magnitude higher than that of other glutenases [18, 20, 24, 70, 71, 72]. The zymogen is produced at neutral pH, at which it remains stable and is lyophilizable for transport and storage. It only becomes activated after ingestion in the stomach and cleaves toxic components of gluten. Once the gastric bolus exits to the slightly acidic postprandial pH duodenum, it becomes

inactive again. Neprosin is therefore a highly promising candidate for further therapeutic development against gluten-sensitive conditions.

Structural and functional studies backed by mutants and activity assays identified neprosin as a pepstatin-sensitive PEP and the first glutamate endopeptidase found in higher eukaryotes. It features a hitherto undescribed pair of catalytic glutamates that are analogous to the aspartates of the otherwise unrelated pepsin-type acidic endopeptidases. Neprosin is produced and secreted as a zymogen, which is activated only in its strongly acidic natural environment, the pitcher plant digestive fluid. Maturation follows a pH-switch mechanism that releases a lysine-mediated latency plug.

Finally, neprosin is pepstatin-sensitive but shares its overall fold with the pepstatin-resistant glutamate peptidases of the eqolysin family, which possess a glutamate–glutamine dyad and are represented by the archetype SCP-B. However, there are differences in the size of the PD and the CD, the active-site environment, the substrate-binding modus and specificity, as well as the chemical mechanism of catalysis. Furthermore, whereas eqolysins are restricted to fungi and bacteria [67, 73], potential neprosin orthologues with ~35–40% sequence identity are widely found in (and restricted to) plants, including gluten-containing crops. This suggests the neprosin family may have originated from a SCP-B ancestor by horizontal gene transfer from a bacterium or fungus to a plant, as previously described for other proteins [74]. Transfer would have been followed by divergent evolution within the plant kingdom to modify one of the catalytic residues and the loops decorating the central β-sandwich to adapt to new substrates. By analogy to the eqolysins, neprosin family members could be named ‘eelysins’.

MATERIALS AND METHODS

Protein production and purification- A synthetic gene encoding wild-type neprosin from *Nepenthes* × *ventrata*, which is 91% identical to the orthologue from *Nepenthes alata* (UniProt ID A0A1L7NZU4), was inserted into vector pET-28a(+) by GenScript to produce vector pET-28a(+)-proNEP. The coding sequence was transferred to vector pCMV (kindly provided by Jan J. Enghild, Aarhus University, Denmark) to produce vector pS6-proNEP. This conferred ampicillin resistance and added a C-terminal hexahistidine (His_6) tag. The encoded protein is described herein as pro-neprosin. The same plasmid was modified by annealed oligonucleotide cloning to (a) replace the His_6-tag with a twin Strep tag (pS6-proNEP-Strep) for the expression of pro-neprosin-strep, and (b) to remove the PD (pS6-NEP) for the expression of the neprosin CD (S^{129}–Q^{380}) plus the C-terminal His_6-tag. The QuikChange Site-Directed Mutagenesis Kit (Stratagene) or inverse PCR-based site-directed mutagenesis were used to generate variants of pS6-proNEP with point mutations $A^{60}R$, $K^{118}A$, $H^{134}A$, $Y^{136}A$, $Q^{173}A$, $W^{175}A$, $E^{188}A$, $E^{188}Q$, $Y^{214}A$, $E^{297}Q$ and $E^{297}A$. Plasmids were purified with the GeneJET Plasmid MaxiPrep Kit (Thermo Fisher Scientific), and constructs were verified by DNA sequencing.

Proteins encoded by the pS6-proNEP, pS6-proNEP-Strep and pS6-NEP plasmids, as well as the 11 point mutants, were assessed for overexpression in human 293 cells grown in a Multitron cell shaker incubator (Infors HT) at 37°C. The cells were transfected with plasmid DNA and harvested after several days for protein purification. Cell-conditioned medium was cleared by centrifugation and supplemented with imidazole, incubated with nickel-nitrilotriacetic acid (Ni-NTA) resin (Invitrogen), subjected to batch affinity chromatography purification (AC), and washed extensively with buffer containing 20 mM imidazole. Proteins were eluted with the same buffer containing 300 mM imidazole. For pro-neprosin-strep, the Ni-NTA resin was replaced with Strep-Tactin XT

Superflow suspension resin (IBA Life Sciences), and proteins were eluted in buffer containing 50 mM D-biotin (VWR Life Science). Fractions containing the protein were pooled and concentrated before size-exclusion chromatography (SEC) in a Superdex 75 10/300 GL column (GE Healthcare), which was attached to an ÄKTA Purifier liquid chromatography system (GE Healthcare).

Proteins were concentrated by ultracentrifugation in Vivaspin filter devices (Sartorius Stedim Biotech). Approximate protein concentrations were determined by measuring the absorbance at 280 nm (A_{280}) using a BioDrop-DUO Micro Volume (Biochrom), and applying the appropriate theoretical extinction coefficients. Moreover, protein purity was assessed by sodium dodecylsulfate polyacrylamide gel electrophoresis (SDS-PAGE) followed by staining with Coomassie (Thermo Fisher Scientific). Protein identity was determined by peptide mass fingerprinting and N-terminal Edman sequencing at the Protein Chemistry Service and the Proteomics Facility of the Centro de Investigaciones Biológicas (Madrid, Spain), respectively. Finally, mature wild-type neprosin was lyophilized, stored at –20°C, and reconstituted by dissolving in Milli-Q water.

For activity assays, the filtered conditioned medium of wild-type neprosin and the point mutants was supplemented with 3 mM reduced glutathione and 0.3 mM oxidized glutathione, the pH was adjusted with 20 mM Tris·HCl pH 8.0, and the mixture was incubated with cOmplete His-Tag Purification Resin (Roche). The resin was collected in an open column and the bound protein was washed with 10 mM Tris·HCl pH 7.0, 300 mM sodium chloride, and was then eluted with 100 mM glycine pH 2.5, 300 mM sodium chloride.

Autolytic activation of pro-neprosin- Wild-type and mutant mature forms of neprosin or neprosin-strep were obtained by autolysis. Protein samples eluted from Ni-NTA or Strep-Tactin columns were dialysed against buffer, diluted twofold with 100 mM glycine pH 2.5, and incubated at 37°C for up to 16 h. Reactions were stopped at specific time points (0 min, 10 min, 20 min, 30 min, 1 h, 2 h and overnight) by boiling aliquots in reducing/denaturing SDS sample buffer, followed by SDS-PAGE. Mature neprosin was buffer-exchanged to 20 mM Tris·HCl pH 7.5, 250 mM sodium chloride in a PD10 column followed by SEC in a Superdex 75 10/300 GL column with the same buffer. Protein purity and identity were assessed as stated above.

Trans-activation of pro-neprosin mutants- To obtain mature neprosin point mutants from zymogens that do not autoactivate, the purified pro-proteins ($H^{134}A$, $Y^{136}A$, $Q^{173}A$, $W^{175}A$, $E^{188}A$, $E^{188}Q$, $Y^{214}A$, $E^{297}Q$ and $E^{297}A$) were incubated with activated neprosin-strep at a 20:1 weight ratio overnight at 37°C. Pro-neprosin-strep was previously buffer-exchanged to 100 mM glycine pH 3.0, 150 mM sodium chloride for activation. Cleaved samples were buffer-exchanged and purified by reverse affinity chromatography, concentrated and purified by SEC.

Protein stability assays- Pro-neprosin and mature neprosin were analysed by differential scanning fluorimetry using an iCycler iQ real-time PCR detection system (Bio-Rad). Samples were prepared at 0.5 mg/mL, in the presence or absence of 5 or 10 mM tris(2-carboxyethyl)phosphine (TCEP) as a reducing agent, and supplemented with 5× SYPRO Orange Protein Stain (Thermo Fisher Scientific). The temperature of midtransition I was determined as the average of duplicate measurements of the midpoint value of the stability curve.

Proteolytic activity and pH profile- We incubated 10 µM of the fluorescent protein substrate DQ Red BSA (Thermo Fisher

Scientific) with 0.15 µM neprosin in 100 µL buffer at pH 2–8. Fluorescence was monitored using an Infinite M2000 microplate fluorimeter (Tecan) at 37°C. We tested 0.5 µM bovine trypsin (Sigma-Aldrich) and porcine pepsin (Fluka) for comparison. Each assay was carried out in triplicate.

Cleavage studies with fluorogenic peptides and determination of kinetic parameters- The kinetic parameters of FS6-QPQL peptide (Mca–Q–P–Q–L–Dpa–A–R–NH_2; GenScript) cleavage by wild-type neprosin (25 nM final enzyme concentration), as well as those of the FS6 peptide (Mca–K–P–L–G–L–Dpa–A–R–NH_2; Sigma-Aldrich) by neprosin at 100 nM final enzyme concentration, were determined in reactions containing 100 mM glycine pH 3.0 and substrate concentrations of 1–75 µM (FS6-QPQL) or 2.5–75 µM (FS6) at 37°C. The fluorescence signal, representing cleavage product formation, was recorded over time for each substrate concentration and the initial rate (v_0) was derived from the slope of the linear part of the curve. Using a range of substrate concentrations and a surplus of peptidase, we measured the fluorescence signal generated after full substrate turnover and calculated the corresponding fluorescence units per picomole of cleaved substrate. These values were plotted against substrate concentration and fitted to the hyperbolic Michaelis-Menten equation ($v = V_{max}\cdot[S]/\{K_M+[S]\}$) by nonlinear regression using *GraphPad* [75] and *SigmaPlot* [76] to determine the maximum velocity (V_{max}), the Michaelis substrate affinity constant (K_M), the turnover rate ($k_{cat} = V_{max}/[E_{total}]$), and the catalytic efficiency (k_{cat}/K_M**)** of17leavageavage reaction. All experiments were carried out in triplicate.

The peptidolytic activity of wild-type neprosin was compared to the mutants $K^{118}A$, $H^{134}A$, $Y^{136}A$, $Q^{173}A$, $W^{175}A$, $E^{188}Q$, $Y^{214}A$ and $E^{297}Q$ (140 ng) using 10 µM of the fluorogenic FS6-QPQL peptide in 100 mM glycine pH 3.0, 150 mM sodium chloride at 37°C, shaking in a Synergy H1 microplate reader (BioTek). To ensure identical sample treatment, all protein variants were activated with neprosin-strep, which was then removed by reverse affinity chromatography as stated above. The protein concentration was estimated from the surface of the A_{280} SEC curves and corrected based on the ε_{280} values. Fluorescence values after 30 min were used as activity endpoints. Experiments were carried out in triplicate and differences were analysed for statistical significance using *GraphPad*.

Cleavage of gliadin in vitro- Wheat gliadin (Sigma-Aldrich) was prepared in 100 mM glycine pH 2.5 and variable concentrations of pepsin (0.05–10 µM) from porcine gastric mucosa (Fluka), neprosin (0.05–2 µM), or mixtures of 0.5 µM pepsin and 0.05–2 µM neprosin were used to digest 10 mg/mL gliadin slurries. Reactions were monitored by turbidimetry in 96-well plates (Corning) at 37°C in a microplate spectrophotometer (BioTek). Reactions were quenched by boiling in SDS sample buffer before analysis by SDS-PAGE. Gliadin degradation by neprosin was also analysed by zymography using SDS-PAGE gels containing either wheat gliadin or teleostean gelatin (Sigma-Aldrich), which was used as a control, at 0.1 mg/mL. Pro-neprosin was also tested, and became activated to the mature form during the assay. Proteins were renatured by washing the zymograms with 2.5% Triton X-100 in 100 mM glycine pH 2.5, 200 mM sodium chloride. After further washes with the same buffer plus 0.02% Brij-35, the zymograms were incubated overnight in the same buffer, rinsed briefly with water, and stained with Coomassie.

Cleavage of the 33-mer peptide in vitro- Cleavage of the 33-mer peptide of wheat α-gliadin (LQLQPFPQPQLPYPQPQLPYPQPQLPYPQPQPF, 3911 Da) was monitored using an AutoFLEX III MALDI-TOF mass spectrometer. The peptide (from GenScript)

was dissolved in water to a concentration of ~20 mg/mL and stored at –20 °C. The cleavage reaction was carried out with ~1 mg/mL (~250 µM) substrate in 100 mM glycine pH 3.0 at 37°C by adding 0.5 µM neprosin or 10 µM pepsin. Reactions were stopped at different time points (0 min, 10 min, 20 min, 45 min, 1 h and overnight) and samples were then diluted 1:10 with water, mixed with an equal volume of the 2,5-dihydroxybenzoic acid matrix at 10 mg/mL in a solution containing 30% acetonitrile and 70% 0.1% trifluoroacetic acid, and spotted on a ground steel plate (Bruker). Mass spectra were acquired in positive reflectron mode at 21 kV total acceleration voltage.

Liquid chromatography-mass spectrometry (LC-MS/MS) data analysis- We reanalysed the cleavage specificity data of endogenous neprosin or recombinant material obtained from *Escherichia coli* deposited at Chorus (Project ID 1262 [25]). LC-MS/MS raw files were converted to MGF format, and data were processed using *TANDEM*, *Comet* and *MS-GF+*, as implemented in *SearchGUI* [77]. Results were evaluated using *PeptideShaker* [78] with a false discovery rate of 1%. Data were non-specifically searched for hits against the human proteome in UniProt (March 2020) using a mass tolerance of 20 ppm for both MS1 and MS2, fixed cysteine carbamidomethylation, and variable methionine oxidation. Up to 50 missed cleavages or a maximum of 5500 Da were tolerated for the parental peptide mass.

Inhibition assays- On the search for neprosin inhibitors, we assayed the broad-spectrum cOmplete Inhibitor Cocktail (Roche); the metallopeptidase inhibitors 1,10-phenathroline, phosphoramidon, marimastat, and captopril (all from Sigma-Aldrich); the serine peptidase inhibitor **4-(2-aminoethyl)-benzenesulfonyl fluoride (AEBSF;** Sigma-Aldrich**); the aspartate peptidase inhibitors pepstatin A (**Sigma-Aldrich**),** methyl-2-[(2-diazoacetyl)amino]hexanoate (DAN; Chemical Abstracts Service (CAS) 7013-09-4; Bachem 4010441)**, and** ENPN (CAS 5255-75-4; Apollo Scientific OR26560); **as well as the proline-containing/mimicking compounds 2-acetyl-1-methylpirrole (AMP; CAS 932-16-1; Sigma-Aldrich 160865); (*S*)-*tert*-butyl-2-(3-ethoxy-3-oxopropanoyl)pyrrolidine-1-carboxylate (BEOPC; CAS 109180-95-2; Fluorochem 387901); and *N*-boc-glycylproline (BGP; CAS** 14296-92-5; **Bachem 4003703). Inhibition of the cleavage of the FS6-QPQL peptide was investigated by pre-incubating 100 nM neprosin in** 100 mM glycine pH 3.0 **with 100 µM of each tester compound for >1 h at 37°C. We then added 10 µM of the substrate and the residual activity was monitored for 4h as an increase in fluorescence.** Differences were analysed for statistical significance using *GraphPad***. The positive control in the absence of inhibitors (100% activity) contained the same final concentration of dimethyl sulfoxide that was used to solubilize the inhibitors. In addition, half-maximal inhibitory concentration (IC_{50}) values were determined for pepstatin A and ENPN by measuring the activity of 50 nM neprosin in the presence of 10 µM of substrate and inhibitor concentrations of 5–500 µM and 5–5000 µM, respectively, to obtain the inhibition curves. These curves were analysed by nonlinear regression using *GraphPad*.**

Evaluation of gliadin processing by neprosin in vivo- Experimental procedures involving mice followed the institutional guidelines for the care and use of laboratory animals and the ARRIVE guidelines. Protocols were approved by the Ethical Committee for Animal Experimentation of the University of Barcelona (CEEA-UB/Ref. 186/20-P2) and the Government of Catalonia (PAMN/Ref. 11485), which followed Directive 2010/63/EU for the protection of animals used for scientific purposes. The sample size was estimated by the Appraising

Project Office's program from the Universitat Miguel Hernández of Elx (Alacant, Spain). We used 5-week-old male and female C57BL/6 mice (n=16) purchased from Janvier and housed at the animal facility of the Faculty of Pharmacy and Food Science of the University of Barcelona in a controlled environment (20–24°C, 40–60% relative humidity) and a 12-h photoperiod, with lights on at 8 a.m. and lights off at 8 p.m. Animals were housed in cages with large Souralit 1035 fibrous particles as bedding (Bobadeb), and tissue paper (Gomà-Camps) and cardboard climbing structures for cage enrichment. Animals had free access to water and RM3 (P) SQC diet (Special Diet Services).

After 1 week for acclimation, two groups of mice were randomly selected, each comprising four males and four females (n=8 per group) and were marked neprosin (N) or vehicle (V). Animals were not fasted to account for the physiological transit time, and food and water were removed only 1 h before oral gavage. Group N mice were fed 50 µL pro-neprosin in vehicle (0.2 mg/mL in 20 µM Tris-buffered saline pH 7.5, 150 µM sodium chloride), whereas group V mice were fed 50 µL vehicle alone. After 5 min, all mice were fed 50 µL gliadin slurry containing 5 mg wheat gliadin (Sigma-Aldrich) at 100 mg/mL in 10% ethanol solution using small-volume Hamilton syringes and adapted oral probes. The enzyme:gliadin ratio (1:500) was calculated based on our *in vitro* results, which had shown that neprosin digests gliadin at a 1:500–1000 ratio at 37°C over a period of 90 min. Given that gastrointestinal transit in mice causes a bolus to reach the small intestine after 1–3 h, with some content already entering the large intestine [79], we selected 2.5 h as the optimal endpoint to assess the degradation of gliadin in the upper gastrointestinal tract. Animals were then euthanized by cervical dislocation and the contents of the stomach, proximal small intestine and distal small intestine were removed, weighed, and frozen at –20°C.

Samples were suspended in phosphate-buffered saline (pH 7.2) at a concentration of 200 mg/mL, homogenized with a Kimble Pellet Pester Cordless Motor (DWK Life Sciences), and extracted first with buffer at 50°C for 40 min and then with 80% ethanol at 20–25°C for 1 h. The mixtures were centrifuged (2000×*g*, 10 min, 4°C), and the aqueous layer between the particulate and fat layers was removed. The 33-mer content in each diluted extract was analysed using the AgraQuant Gluten G12 ELISA test kit (Romer Labs), which has a detection limit of 2 ppm, according to the manufacturer's instructions. The G12 antibody detects the 33-mer but no other gliadin degradation fragments [80]. Final amounts were normalized taking into account the sample weight and results were expressed as mean ± SEM. The *Statistical Package for Social Sciences* (SPSS v22.0; IBM) was used for statistical analysis. The data showed homogeneity of variance (Levene's test) and followed a normal distribution (Shapiro-Wilk test), so we applied conventional one-way analysis of variance (ANOVA).

Crystallization and diffraction data collection- We screened for crystallization conditions at the joint IBMB/IRB Automated Crystallography Platform using the sitting-drop vapour diffusion method. Optimal pro-neprosin crystals (~20 mg/mL in 20 mM Tris·HCl pH 7.5, 150 mM sodium chloride) were obtained at 20°C with 0.1 M sodium acetate pH 4.0, 22% polyethylene glycol (PEG) 6000, 10% isopropanol as the reservoir solution. Crystals were harvested using cryo-loops (Molecular Dimensions), rapidly passed through a cryo-buffer consisting of reservoir solution plus 15% (v/v) glycerol, and flash-vitrified in liquid nitrogen for data collection. A lutetium derivative of pro-neprosin was obtained by soaking native crystals for 5 min in cryo-buffer supplemented with 100 mM of the Lu-Xo4 'crystallophore' (Polyvalan) [40] and flash-vitrifying them without back soaking. X-ray diffraction data were collected

from native crystals at 100 K on a Pilatus 6M-F pixel detector at beamline I04-1 of the Diamond Light Source (Harwell, UK). Lutetium derivative data were recorded on a Pilatus 6M detector at beamline XALOC of the ALBA synchrotron (Cerdanyola, Catalonia, Spain).

The mature neprosin–product complex (crystal form I) was obtained at a protein concentration of ~16 mg/mL in 20 mM Tris·HCl pH 7.5, 250 mM sodium chloride at 4°C using 10% PEG 1000, 10% PEG 8000 as the reservoir solution. Crystals were cryo-protected with the same reservoir solution plus 15% (v/v) glycerol prior to flash-vitrification in liquid nitrogen. X-ray diffraction data at 100 K were collected at beamline ID30B of the ESRF synchrotron (Grenoble, France) using a Pilatus 6M detector. The mature neprosin–product complex in crystal form II was obtained at the same protein concentration but in 0.1 M glycine pH 3.0, 150 mM sodium chloride at 20°C using 0.1 M sodium citrate tribasic pH 5.6, 0.5 M ammonium sulfate, 1 M lithium sulfate as the reservoir solution. Crystals were cryo-protected with a solution containing 20% (v/v) glycerol. Diffraction data were collected at beamline XALOC on a Pilatus 6M detector.

Diffraction data were processed using *Xds* [81] and *Xscale*, and were transformed to MTZ format using *Xdsconv* for the *Phenix* [82] and *Ccp4* [83] program suites. All crystals contained a monomer in the crystal asymmetric unit and Suppl. Table1 provides essential statistics on data collection and processing.

Structural solution and refinement- The structure of pro-neprosin was solved by single-wavelength anomalous diffraction using data collected from a lutetium derivative crystal at the L_{III}-absorption peak wavelength (1.34 Å) by applying the *Autosol* protocol of the *Phenix* package. The resulting Fourier map was then subjected to further density modification with *wARP/ARP* [84]. A starting model for Lu-Xo4 was obtained by energy minimization applied to the coordinates of the metal-chelating moiety of the compound as found in its complex with Tb^{3+} (Protein Data Bank [PDB] ID 6FRO, residue name 7MT) using *Chimera* [85]. The resulting coordinates in PDB format were combined with a Lu^{3+} ion for model building. Thereafter, several rounds of manual model building in *Coot* [86] alternated with crystallographic refinement using the *Refine* protocol of *Phenix* and the *BUSTER* [87] program. The final model comprised pro-neprosin residues A^{29}–Q^{380} except S^{76}–Y^{85} and N^{122}–N^{131} plus three extra C-terminal residues from the purification tag (A^{401}–I^{402}–A^{403}); two Lu-Xo4 moieties at roughly half occupancy; two *N*-linked glycan chains totalling five sugar residues attached to N^{145} and N^{152}, respectively; two acetate molecules; and 180 solvent molecules.

The structure of native pro-neprosin was solved by molecular replacement using the *Phaser* crystallographic software within *Ccp4* and the protein coordinates of the lutetium derivative crystal structure. Subsequent model building and refinement proceeded as described above. The final model included residues A^{29}–Q^{380} except Y^{77}–Y^{85} and N^{122}–N^{131} plus two extra C-terminal residues from the purification tag (A^{401}–I^{402}), two *N*-linked glycan chains totalling four sugar residues, as well as eight acetate, one isopropanol, four glycerol and 257 solvent molecules.

The structure of a product complex of native mature neprosin in crystal form I was also solved by molecular replacement, using the coordinates of fragment T^{132}–I^{402} from native pro-neprosin. Subsequent model building and refinement proceeded as described above. The final model spanned residues T^{132}–Q^{380} plus the entire C-terminal tag (A^{401}–I^{402}–A^{403}+H^{404}–H^{409}), two *N*-linked glycan chains totalling seven sugar residues, plus one triethylene glycol and 171 solvent molecules. The structure of a product complex of native mature neprosin in crystal form II was also solved by molecular replacement, using fragment T^{132}–Q^{380} of the crystal form I complex. The final model

included residues T^{132}–Q^{380} plus the C-terminal tag except H^{409} (A^{401}–I^{402}–A^{403}+H^{404}–H^{408}), as well as two *N*-linked glycan chains totalling four sugar residues plus one nickel cation, three sulfate anions, one tetraglycine and one glycine, as well as 250 solvent molecules. The nickel ion, presumably from the Ni-NTA resin used for purification, was tentatively assigned based on short liganding distances to two histidine residues (~1.8 Å) ,which were closer to those reported for tetrahedrally-coordinated nickel ions (1.88 Å on average) than to those of the more abundant lithium (2.03 Å) from the reservoir solution [88]. A tetraglycine was tentatively placed in an adequate density region based on the capacity of this amino acid to oligomerize under certain conditions [89]. Suppl. Table1 provides essential statistics on the final refined models, which were validated using the *wwPDB Validation Service* at https://validate-rcsb-1.wwpdb.org/validservice and deposited with the PDB at www.pdb.org (access codes 7ZU8, 7ZVA, 7ZVB and 7ZVC).

Miscellaneous- Structural superpositions and structure-based sequence alignments were calculated using the *SSM* program within *Coot*. Figures were prepared using *Chimera*. Structure-based similarity searches were performed with *Dali* [90]. Protein interfaces were calculated with *PDBePISA* at www.ebi.ac.uk/pdbe/pisa. The interacting surface of a complex was defined as half the sum of the buried surface areas of either molecule.

ACKNOWLEDGMENTS

We are grateful to Laura Company, Roman Bonet, Xandra Kreplin and Joan Pous from the joint IBMB/IRB Automated Crystallography Platform and the Protein Purification Service for assistance during purification and crystallization. Plasmid pCMV was kindly provided by Jan J. Enghild, Århus University, Denmark. The authors also would like to thank the ESRF, DIAMOND, and ALBA synchrotrons for beamtime and the respective beamline staff for assistance during diffraction data collection. This study was supported in part by grants from Spanish and Catalan public and private bodies (grant/fellowship references PID2019-107725RG-I00, BES-2016-076877, BES-2015-074583, "Beatriu de Pinós" 2018BP00163, 2017SGR3 and Fundació "La Marató de TV3" 201815). The authors thank Richard M. Twyman for editing the manuscript.

AUTHOR CONTRIBUTIONS

F.X.G.R. conceived and supervised the project; L.d.A.M., T.G., L.G.-F. and S.R.M. produced and purified proteins, generated mutants, and performed *in vitro* studies; L.d.A.M., A.R.B. and T.G. crystallized proteins; A.R.B. and U.E. collected diffraction data; U.E. performed experiments, analysed data and supervised workers; F.X.G.R. solved and refined crystal structures; F.J.P.C., M.G., M.J.R.L., and À.F. performed animal experiments, and F.X.G.R. wrote the manuscript with contributions from all authors.

COMPETING INTERESTS

The authors declare no financial or non-financial conflicts of interest with the contents of this article.

DATA AVAILABILITY STATEMENT

All data and reagents are freely available from the authors upon reasonable request and signature of non-disclosure and material transfer agreements for non-profit usage by academic groups. The PDB coordinates of the structures solved in this work have been

deposited with the Protein Data Bank, the respective codes are provided in Suppl. Table1.

REFERENCES

[1] Gee S. On the coeliac affection. *St Barth Hosp Rep* 24, 17-20 (1888).

[2] Lindfors K*, et al.* Coeliac disease. *Nat Rev Dis Primers* 5, 3 (2019).

[3] Monachesi C*, et al.* Quantification of accidental gluten contamination in the diet of children with treated celiac disease. *Nutrients* 13, 190 (2021).

[4] Singh P*, et al.* Global prevalence of celiac disease: systematic review and meta-analysis. *Clin Gastroenterol Hepatol* 16, 823-836 (e822) (2018).

[5] King JA*, et al.* Incidence of celiac disease Is increasing over time: a systematic review and meta-analysis. *Am J Gastroenterol* 115, 507-525 (2020).

[6] Hausch F, Shan L, Santiago NA, Gray GM, Khosla C. Intestinal digestive resistance of immunodominant gliadin peptides. *Am J Physiol Gastrointest Liver Physiol* 283, G996-G1003 (2002).

[7] Balakireva AV, Zamyatnin AA. Properties of gluten intolerance: gluten structure, evolution, pathogenicity and detoxification capabilities. *Nutrients* 8, (2016).

[8] El-Salhy M, Hatlebakk JG, Gilja OH, Hausken T. The relation between celiac disease, nonceliac gluten sensitivity and irritable bowel syndrome. *Nutr J* 14, 92 (2015).

[9] Daveson AJM*, et al.* Baseline quantitative histology in therapeutics trials reveals villus atrophy in most patients with coeliac disease who appear well controlled on gluten-free diet. *GastroHep* 2, 22-30 (2020).

[10] Shan L*, et al.* Structural basis for gluten intolerance in celiac sprue. *Science* 297, 2275-2279 (2002).

[11] Wei G, Helmerhorst EJ, Darwish G, Blumenkranz G, Schuppan D. Gluten degrading enzymes for treatment of celiac disease. *Nutrients* 12, 2095 (2020).

[12] Kivelä L, Caminero A, Leffler DA, Pinto-Sanchez MI, Tye-Din JA, Lindfors K. Current and emerging therapies for coeliac disease. *Nat Rev Gastroenterol Hepatol* 18, in press (2021).

[13] Suchy FJ*, et al.* National Institutes of Health Consensus Development Conference: lactose intolerance and health. *Ann Intern Med* 152, 792-796 (2010).

[14] Molina-Infante J, Santolaria S, Sanders DS, Fernández-Banares F. Systematic review: noncoeliac gluten sensitivity. *Aliment Pharmacol Ther* 41, 807-820 (2015).

[15] Creed F. Review article: the incidence and risk factors for irritable bowel syndrome in population-based studies. *Aliment Pharmacol Ther* 50, 507-516 (2019).

[16] Clarysse S, Tack J, Lammert F, Duchateau G, Reppas C, Augustijns P. Postprandial evolution in composition and characteristics of human duodenal fluids in different nutritional states. *J Pharm Sci* 98, 1177-1192 (2009).

[17] König J, Holster S, Bruins MJ, Brummer RJ. Randomized clinical trial: effective gluten degradation by *Aspergillus niger*-derived enzyme in a complex meal setting. *Sci Rep* 7, 13100 (2017).

[18] Ehren J, Moron B, Martin E, Bethune MT, Gray GM, Khosla C. A food-grade enzyme preparation with modest gluten detoxification properties. *PloS one* 4, e6313 (2009).

[19] Kulkarni A, Patel S, Khanna D, Parmar MS. Current pharmacological approaches and potential future therapies for Celiac disease. *Eur J Pharmacol* 909, 174434 (2021).

[20] Pultz IS, Leffler DA, Liu T, Winkle P, Vitanza JM, Hill M. AGA Abstracts 1125. Kuma062 effectively digests gluten in the human stomach: results of a phase 1 study. *Gastroenterology* 158, S-218 (2020).

[21] Krishnareddy S, Stier K, Recanati M, Lebwohl B, Green PH. Commercially available glutenases: a potential hazard in coeliac disease. *Therap Adv Gastroenterol* 10, 473-481 (2017).

[22] Rawlings ND, Bateman A. How to use the MEROPS database and website to help understand peptidase specificity. *Protein Sci* 30, 83-92 (2021).

[23] Lee L, Zhang Y, Ozar B, Sensen CW, Schriemer DC. Carnivorous nutrition in pitcher plants (*Nepenthes* spp.) via an unusual complement of endogenous enzymes. *J Proteome Res* 15, 3108-3117 (2016).

[24] Rey M*, et al.* Addressing proteolytic efficiency in enzymatic degradation therapy for celiac disease. *Sci Rep* 6, 30980 (2016).

[25] Schräder CU*, et al.* Neprosin, a selective prolyl endoprotease for bottom-up proteomics and histone mapping. *Mol Cell Proteomics* 16, 1162-1171 (2017).

[26] Schräder CU, Ziemianowicz DS, Merx K, Schriemer DC. Simultaneous proteoform analysis of histones H3 and H4 with a simplified middle-down proteomics method. *Anal Chem* 90, 3083-3090 (2018).

[27] Ericsson UB, Hallberg BM, DeTitta GT, Dekker N, Nordlund P. Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem* 357, 289-298 (2006).

[28] Vieille C, Zeikus GJ. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. *Microbiol Mol Biol Rev* 65, 1-43 (2001).

[29] Eder J, Fersht AR. Pro-sequence-assisted protein folding. *Mol Microbiol* 16, 609-614 (1995).

[30] Raufman JP. Pepsin. In: *Encyclopedia of Gastroenterology.* (ed Johnson LR). 1st edn. Academic Press - Elsevier (2004).

[31] Butts CT, Bierma JC, Martin RW. Novel proteases from the genome of the carnivorous plant *Drosera capensis*: structural prediction and comparative analysis. *Proteins* 84, 1517-1533 (2016).

[32] Schechter I, Berger A. On the size of active site in proteases. I. Papain. *Biochem Biophys Res Commun* 27, 157-162 (1967).

[33] Gomis-Rüth FX, Botelho TO, Bode W. A standard orientation for metallopeptidases. *Biochim Biophys Acta* 1824, 157-163 (2012).

[34] Roberts NB, Sheers R, Taylor WH. Secretion of total pepsin and pepsin 1 in healthy volunteers in response to pentagastrin and to insulin-induced hypoglycaemia. *Scand J Gastroenterol* 42, 555-561 (2007).

[35] Qiao SW*, et al.* Antigen presentation to celiac lesion-derived T cells of a 33-mer gliadin peptide naturally formed by gastrointestinal digestion. *J Immunol* 173, 1757-1762 (2004).

[36] Neumann U, Kubota H, Frei K, Ganu V, Leppert D. Characterization of Mca-Lys-Pro-Leu-Gly-Leu-Dpa-Ala-Arg-NH_2, a fluorogenic substrate with increased specificity constants for collagenases and tumor necrosis factor converting enzyme. *Anal Biochem* 328, 166-173 (2004).

[37] Christensen EM*, et al. In crystallo* screening for proline analog inhibitors of the proline cycle enzyme PYCR1. *J Biol Chem* 295, 18316-18327 (2020).

[38] Umezawa H, Aoyagi T, Morishima H, Matsuzaki M, Hamada M. Pepstatin, a new pepsin inhibitor produced by *Actinomycetes*. *J Antibiot (Tokyo)* 23, 259-262 (1970).

[39] Abell AD, Hoult DA, Bergman DA, Fairlie DP. Simple *cis*-epoxide-based inhibitors of HIV-1 protease. *Bioorg Med Chem Lett* 7, 2853-2856 (1997).

[40] Engilberge S*, et al.* Crystallophore: a versatile lanthanide complex for protein crystallography combining nucleating effects, phasing properties, and luminescence. *Chem Sci* 8, 5909-5917 (2017).

[41] Stawiski EW, Baucom AE, Lohr SC, Gregoret LM. Predicting protein function from structure: Unique structural features of proteases. *Proc Natl Acad Sci USA* 97, 3954-3958 (2000).

[42] Khan AR, James MN. Molecular mechanisms for the conversion of zymogens to active proteolytic enzymes. *Prot Sci* 7, 815-836 (1998).

[43] Khan AR, Khazanovich-Bernstein N, Bergmann EM, James MNG. Structural aspects of activation pathways of aspartic protease zymogens and viral 3C protease precursors. *Proc Natl Acad Sci USA* 96, 10968-10975 (1999).

[44] Arolas JL, Goulas T, Cuppari A, Gomis-Rüth FX. Multiple architectures and mechanisms of latency in metallopeptidase zymogens. *Chem Rev* 118, 5581-5597 (2018).

[45] Fujinaga M, Chernaia MM, Tarasova NI, Mosimann SC, James MNG. Crystal structure of human pepsin and its complex with pepstatin. *Protein Sci* 4, 960-972 (1995).

[46] Wlodawer A, Gutschina A, James MNG. Chapter 2 – Catalytic pathways of aspartic peptidases. In: *Handbook of Proteolytic Enzymes* (eds Rawlings ND, Salvesen GS). 3rd edn. Academic Press (2013).

[47] Rawlings ND, Barrett AJ. Chapter 1 - Introduction: aspartic and glutamic peptidases and their clans. In: *Handbook of Proteolytic Enzymes* (eds Rawlings ND, Salvesen GS). 3rd edn. Academic Press (2013).

[48] Ting T-Y, Baharin A, Ramzi AB, Ng C-L, Goh H-H. Neprosin belongs to a new family of glutamic peptidase based on in silico evidence. *Plant Physiol Biochem* 183, 23-35 (2022).

[49] Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol* 372, 774-797 (2007).

[50] Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. *Q Rev Biophys* 41, 133-180 (2008).

[51] Herriott RM, Bartz QR, Northrop JH. Transformation of swine pepsinogen into swine pepsin by chicken pepsin. *J Gen Physiol* 21, 575-582 (1938).

[52]Dall E, Brandstetter H. Mechanistic and structural studies on legumain explain its zymogenicity, distinct activation pathways, and regulation. *Proc Natl Acad Sci USA* 110, 10940-10945 (2013).

[53] Robertus JD, Kraut J, Alden RA, Birktoft JJ. Subtilisin; a stereochemical mechanism involving transition-state stabilization. *Biochemistry* 11, 4293-4303 (1972).

[54] Boon L, Ugarte-Berzal E, Vandooren J, Opdenakker G. Protease propeptide structures, mechanisms of activation, and functions. *Crit Rev Biochem Mol Biol* 55, 111-165 (2020).

[55] Fujinaga M, Cherney MM, Oyama H, Oda K, James MNG. The molecular structure and catalytic mechanism of a novel carboxyl peptidase from *Scytalidium lignicolum*. *Proc Natl Acad Sci USA* 101, 3364-3369 (2004).

[56] Pillai B, Cherney MM, Hiraga K, Takada K, Oda K, James MN. Crystal structure of scytalidoglutamic peptidase with its first potent inhibitor provides insights into substrate specificity and catalysis. *J Mol Biol* 365, 343-361 (2007).

[57] Kondo MY*, et al.* Studies on the catalytic mechanism of a glutamic peptidase. *J Biol Chem* 285, 21437-21445 (2010).

[58] Sasaki H*, et al.* The three-dimensional structure of aspergilloglutamic peptidase from *Aspergillus*

niger. *Proc Jpn Acad Ser B - Phys Biol Sci* 80, 435-438 (2004).

[59] Sasaki H*, et al.* The crystal structure of an intermediate dimer of aspergilloglutamic peptidase that mimics the enzyme-activation product complex produced upon autoproteolysis. *J Biochem* 152, 45-52 (2012).

[60] Jara P*, et al.* Cloning and characterization of the *eapB* and *eapC* genes of *Cryphonectria parasitica* encoding two new acid proteinases, and disruption of *eapC*. *Mol Gen Genet* 250, 97-105 (1996).

[61] Poussereau N, Creton S, Billon-Grand G, Rascle C, Fevre M. Regulation of *acp1*, encoding a non-aspartyl acid protease expressed during pathogenesis of *Sclerotinia sclerotiorum*. *Microbiology* 147, 717-726 (2001).

[62] Moon JL, Shaw LN, Mayo JA, Potempa J, Travis J. Isolation and properties of extracellular proteinases of *Penicillium marneffei*. *Biol Chem* 387, 985-993 (2006).

[63] O'Donoghue AJ*, et al.* Inhibition of a secreted glutamic peptidase prevents growth of the fungus *Talaromyces emersonii*. *J Biol Chem* 283, 29186-29195 (2008).

[64] Rolland S, Bruel C, Rascle C, Girard V, Billon-Grand G, Poussereau N. pH controls both transcription and post-translational processing of the protease BcACP1 in the phytopathogenic fungus *Botrytis cinerea*. *Microbiology* 155, 2097-2105 (2009).

[65] Jensen K, Ostergaard PR, Wilting R, Lassen SF. Identification and characterization of a bacterial glutamic peptidase. *BMC Biochem* 11, 47 (2010).

[66] Oda N, Gotoh Y, Oyama H, Murao S, Oda K, Tsuru D. Nucleotide sequence of the gene encoding the precursor protein of pepstatin insensitive acid protease B, scytalidopepsin B, from *Scytalidium lignicolum*. *Biosci Biotechnol Biochem* 62, 1637-1639 (1998).

[67] Stocchi N, Revuelta MV, Castronuovo PAL, Vera DMA, ten Have A. Molecular dynamics and structure function analysis show that substrate binding and specificity are major forces in the functional diversification of Eqolisins. *BMC bioinformatics* 19, 338 (2018).

[68] Kataoka Y, Takada K, Oyama H, Tsunemi M, James MNG, Oda K. Catalytic residues and substrate specificity of scytalidoglutamic peptidase, the first member of the eqolisin in family (G1) of peptidases. *FEBS Lett* 579, 2991-2994 (2005).

[69] van de Wal Y, Kooy YM, Drijfhout JW, Amons R, Koning F. Peptide binding characteristics of the coeliac disease-associated DQ(α1*0501, β1*0201) molecule. *Immunogenetics* 44, 246-253 (1996).

[70] Siegel M*, et al.* Rational design of combination enzyme therapy for celiac sprue. *Chem Biol* 13, 649-658 (2006).

[71] Gass J, Bethune MT, Siegel M, Spencer A, Khosla C. Combination enzyme therapy for gastric digestion of dietary gluten in patients with celiac sprue. *Gastroenterology* 133, 472-480 (2007).

[72] Mitea C, Havenaar R, Drijfhout JW, Edens L, Dekking L, Koning F. Efficient degradation of gluten by a prolyl endoprotease in a gastrointestinal model: implications for coeliac disease. *Gut* 57, 25-32 (2008).

[73] Sims AH, Dunn-Coleman NS, Robson GD, Oliver SG. Glutamic protease distribution is limited to filamentous fungi. *FEMS Microbiol Lett* 239, 95-101 (2004).

[74] Keeling PJ. Functional and ecological impacts of horizontal gene transfer in eukaryotes. *Curr Opin Genet Dev* 19, 613-619 (2009).

[75] Swift ML. GraphPad Prism, data analysis, and scientific graphing. *J Chem Inf Comput Sci* 37, 411-412 (1997).

[76] Kornbrot D. Statistical software for microcomputers: SigmaPlot 2000 and SigmaStat2. *Br J Math Stat Psychol* 53 (Pt 2), 335-337 (2000).

[77] Barsnes H, Vaudel M. SearchGUI: a highly adaptable common interface for proteomics search and *de novo* engines. *J Proteome Res* 17, 2552-2555 (2018).

[78] Vaudel M*, et al.* PeptideShaker enables reanalysis of MS-derived proteomics data sets. *Nat Biotechnol* 33, 22-24 (2015).

[79] Padmanabhan P, Grosse J, Asad AB, Radda GK, Golay X. Gastrointestinal transit measurements in mice with 99mTc-DTPA-labeled activated charcoal using NanoSPECT-CT. *EJNMMI Res* 3, 60 (2013).

[80] Morón B*, et al.* Toward the assessment of food toxicity for celiac patients: characterization of monoclonal antibodies to a main immunogenic gluten peptide. *PloS one* 3, e2294 (2008).

[81] Kabsch W. XDS. *Acta Crystallogr sect D* 66, 125-132 (2010).

[82] Adams PD*, et al.* PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr sect D* 66, 213-221 (2010).

[83] Winn MD*, et al.* Overview of the CCP4 suite and current developments. *Acta Crystallogr sect D* 67, 235-242 (2011).

[84] Langer G, Cohen SX, Lamzin VS, Perrakis A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat Protoc* 3, 1171-1179 (2008).

[85] Huang CC, Meng EC, Morris JH, Pettersen EF, Ferrin TE. Enhancing UCSF Chimera through web services. *Nucl Acids Res* 42, W478-W484 (2014).

[86] Casañal A, Lohkamp B, Emsley P. Current developments in *Coot* for macromolecular model building of electron cryo-microscopy and crystallographic data. *Protein Sci* 29, 1069-1078 (2020).

[87] Smart OS*, et al.* Exploiting structure similarity in refinement: automated NCS and target-structure restraints in BUSTER. *Acta Crystallogr sect D* 68, 368-380 (2012).

[88] Kuppuraj G, Dudev M, Lim C. Factors governing metal-ligand distances and coordination gemoetries of metal complexes. *J Phys Chem B* 113, 2952-2960 (2009).

[89] Campbell TD, Febrian R, Kleinschmidt HE, Smith KA, Bracher PJ. Quantitative analysis of glycine oligomerization by ion-pair chromatography. *ACS Omega* 4, 12745-12752 (2019).

[90] Holm L, Laakso LM. Dali server update. *Nucleic Acids Res* 44, W351-W355 (2016).

Supplementary Material 2

"De Novo Design of Immunoglobulin-like Domains"

De Novo Design of Immunoglobulin-like Domains

Tamuka M. Chidyausiku[1,2,7,†,‡], Soraia R. Mendes[3†], Jason C. Klima[1,2,§], Marta Nadal[4], Ulrich Eckhard[3], Jorge Roel-Touris[4], Scott Houliston[5,6], Tibisay Guevara[3], Hugh K. Haddox[2], Adam Moyer[2], Cheryl H. Arrowsmith[5,6], F. Xavier Gomis-Rüth[3*], David Baker[1,2,7,*], Enrique Marcos[4*]

[1] Department of Biochemistry, University of Washington, Seattle, WA 98195, USA
[2] Institute for Protein Design, University of Washington, Seattle, WA 98195, USA
[3] Proteolysis Laboratory, Department of Structural Biology, Molecular Biology Institute of Barcelona (IBMB-CSIC), Baldiri Reixac 15, 08028 Barcelona, Spain
[4] Protein Design and Modeling Lab, Department of Structural Biology, Molecular Biology Institute of Barcelona (IBMB-CSIC), Baldiri Reixac 15, 08028 Barcelona, Spain
[5] Structural Genomics Consortium, University of Toronto, Toronto, ON, M5G 1L7, Canada
[6] Princess Margaret Cancer Centre and Department of Medical Biophysics, University of Toronto, Toronto, ON, M5G 2M9, Canada
[7] Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA
[†] These authors contributed equally to this work
[‡] Present address: Novartis Institutes for BioMedical Research Inc., San Diego, CA 92121, USA § Present address: Encodia, Inc., San Diego, CA 92121, USA
[*] Corresponding authors: embcri@ibmb.csic.es, dabaker@uw.edu, xgrcri@ibmb.csic.es

---

Antibodies and antibody derivatives such as nanobodies contain immunoglobulin-like (Ig) β-sandwich scaffolds which anchor the hypervariable antigen-binding loops and constitute the largest growing class of drugs. Current engineering strategies for this class of compounds rely on naturally existing Ig frameworks, which can be hard to modify and have limitations in manufacturability, designability and range of action. Here we develop design rules for the central feature of the Ig fold architecture – the non-local cross-β structure connecting the two β-sheets – and use these to *de novo* design highly stable Ig domains, confirm their structures through X-ray crystallography, and show they can correctly scaffold functional loops. Our approach opens the door to the design of a new class of antibody-like scaffolds with tailored structures and superior biophysical properties.

Immunoglobulin-like (Ig) domain scaffolds have two sandwiched β-sheets that are well-suited for anchoring antigen-binding hypervariable loops, as in antibodies and nanobodies. To date, approaches to engineering antibodies rely on naturally occurring Ig backbone frameworks, and mainly focus on optimizing the antigen-binding loops and/or multimeric formats for improving targeting efficiency or biophysical properties. Despite their exponential advance as protein therapeutics, engineered antibodies have significant limitations in terms of stability, manufacturing, size and structure, among others. Several alternative antibody fragments, such as Fab (antigen binding fragment) and scFv (single-chain variable fragment), and antibody-like scaffolds such as nanobodies have been engineered to address some of these limitations (*1–3*). The β-sheet geometry in these antibody alternatives are kept very close to naturally existing Ig structures because it is much harder to modify the β-sheet structure than the variable loops. *De novo* designing Ig domains with a wider

range of core structures could expand the scope of antibody-engineering applications, but the design of β-sheet proteins remains a formidable challenge due to their structural irregularity and aggregation propensity (*4*). Recent understanding of design rules controlling the curvature (*5*, *6*) and loop geometry in β-sheets (*7*, *8*) have enabled the design of β-barrels (*6*, *9*) and double-stranded β-helices (*8*), but the design principles for Ig domains and β-sandwiches in general are still poorly understood.

We set out to *de novo* design new Ig fold structures, and began by considering the key aspects of the fold. The basic Ig domain (*10*, *11*) is a β-sandwich formed by 7-to-9 β-strands arranged in two antiparallel β-sheets facing each other, and connected through β-hairpins (within the same β-sheet) and β-arches (*12*) (crossovers between two opposing β-sheets). Natural Ig domains are structurally very diverse, often containing extra secondary structure elements and complex loop regions, but they all share a protein core with a super-secondary structure "cross-β" motif that is common to most β-sandwiches: two antiparallel and interlocked β-arches (*13*) in which the first β-strands of each β-arch form one β-sheet, and the following β-strands cross and pair in the opposing β-sheet (Fig. 1). The four constituent cross-β strands (S2, S3, S5, S6) correspond to the B, C, E and F β-strands that build the common structural core of Ig domains found in nature (*10*, *11*), and for which some sequence signatures related to stability or function have been reported – e.g. a disulfide bridge between the B and F β-strands, a buried tryptophan in β-strand B (*11*, *14*) or the tyrosine corner (*15*) between β-strand C and the loop connecting β-strands E and F. The non-local cross-β structure (Fig. 1a) comprises two Greek key super-secondary structures (*16*, *17*) involving four consecutive β-strands in which the first is paired to the last (Fig. 1b). Once the cross-β structures – which associate portions of the peptide chain distant along the linear sequence – are formed or designed, assembling the remainder of the peripheral β-strands is straightforward as it is only necessary to extend sequence-local β-hairpins out from the cross-β strands (Fig. 1b). Peripheral β-strands form later in the folding of Ig-like proteins (*14*, *18*), and are variable in number and structure across the different subtypes of Ig domains found in nature (*10*, *11*). The cross-β motif also controls the overall β-sandwich geometry, which can be conveniently described by the rigid-body transformation parameters relating the two constituent β-sheets – i.e., the distance and rotation along a vector connecting the two centers of the two opposing β-sheets, and the rotations around the two orthogonal vectors (Fig. 2a).

RESULTS

Principles for designing cross-β motifs- We began by investigating how the structural requirements associated with cross-β motifs constrain the geometry of the two β-arches connecting the β-strands. Since β-arch connections have four possible sidechain orientation patterns (*8*) ("Out-Out", "Out-In", "In-Out" and "In-In") depending on whether the Cα-Cβ vector of the β-strand residues preceding and following the β- arch connection point inwards ("In") or outwards ("Out") from the β-arch (Fig.2b; Supplementary Fig.1), there are sixteen possible cross-β motif connection orientations in total. For example, the "Out-Out/In-In" cross-β connection orientation means that the first and second β-arch connections have the "Out-Out" and "In-In" orientations, respectively. Due to the alternating pleating of β- strands, the cross-β connection orientation and the length of the β-strands in the two β-sheets are strongly coupled: if paired β-strands have no register shift, they must be odd-numbered in four of the possible cross-β orientations, even-numbered in four of the other possible cross-β orientations, and odd-numbered in one of the two β-sheets and even-numbered in the other β-sheet in the remaining

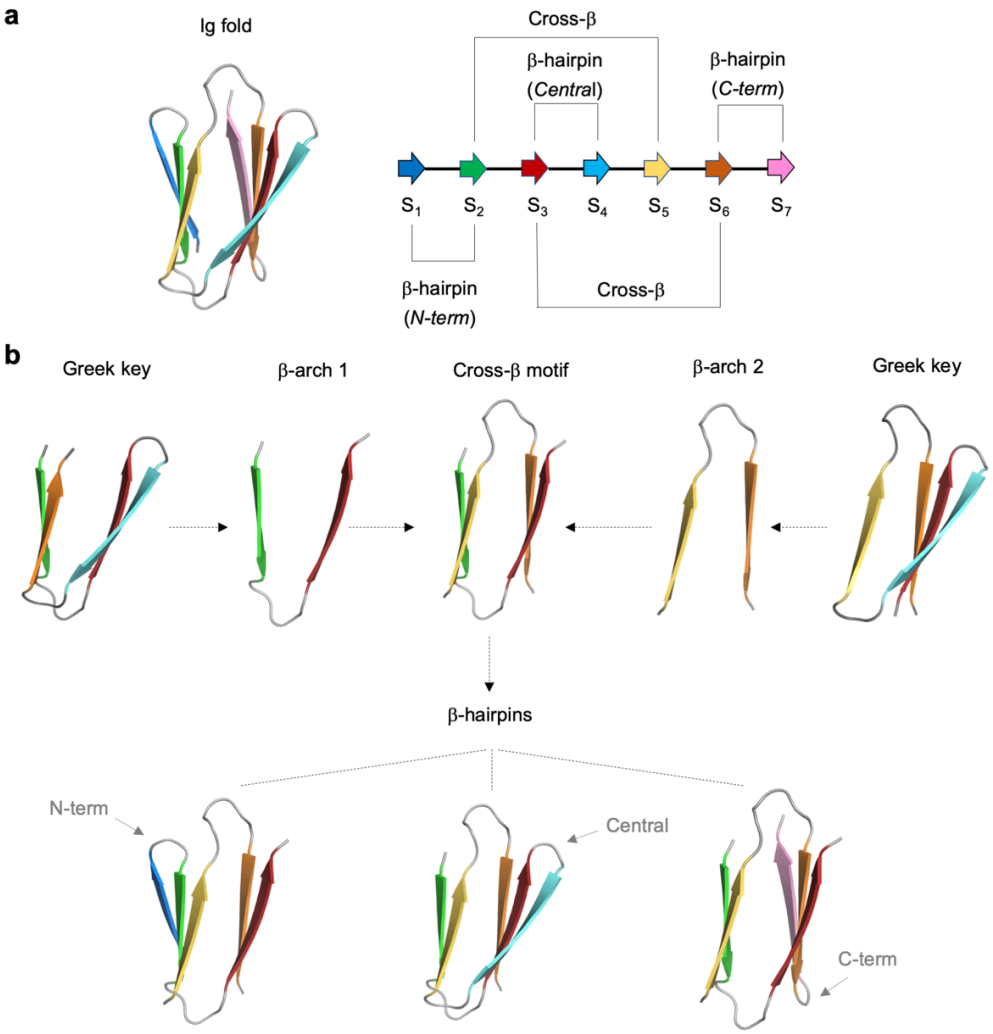


Fig. 1. Topology of immunoglobulin-like domains. a, Three-dimensional cartoon representation of an Ig structure formed by seven β-strands *(left)*; and backbone hydrogen bond patterns *(annotated thin lines)* between paired β-strands along the sequence *(right)*. Cross-β interactions have higher sequence separation (and higher contact order) than β-hairpins, which slows down folding. **b**, β-arches of the cross-β motif belong to two contiguous and distinct Greek key motifs: with 2 β-strands in each β-sheet *(left);* and with 3 β-strands in one β-sheet and 1 β-strand in the other *(right)*. From the folding and design perspective, the main limiting factor for correctly assembling the Ig structure is formation of the cross-β motif, since the three β-hairpins can form independently of one another.

eight cases. Guided by this principle, we studied the efficiency in forming cross-β motifs of highly structured β-arch connections; too flexible β-arches can hinder folding as they increase the protein contact order (*19*) – the average sequence separation between contacting residues – which slows down folding. The cross-β motif is the highest contact order part of the Ig fold architecture, and thus the rate of formation of this structure likely determines the overall rate of folding and thus contributes to the balance between folding and aggregation; once the cross-β motif is formed, folding is likely completed rapidly as the remaining β-hairpins are sequence-local (Fig. 1b).

We generated cross-β motifs exploring combinations of short β-arch loops frequently observed in naturally occurring proteins and spanning the sixteen possible

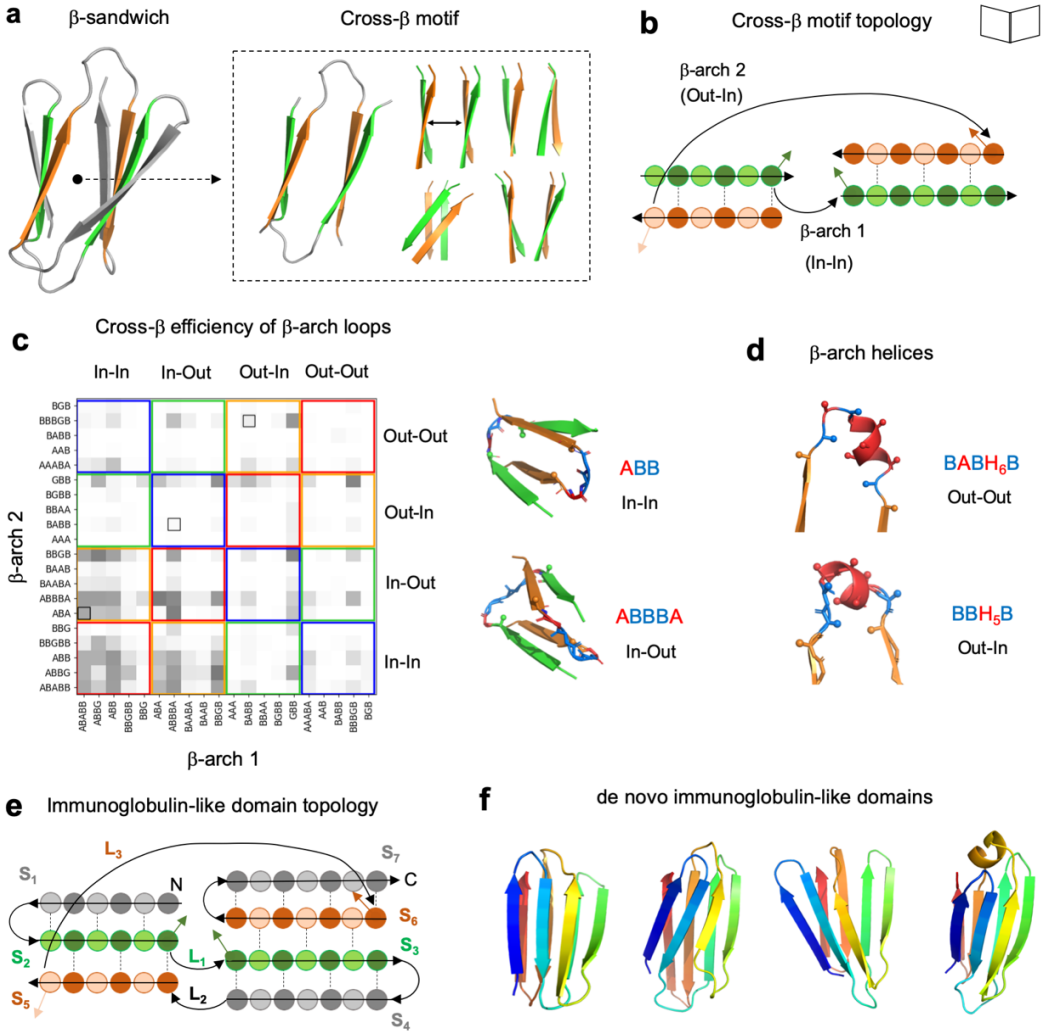


Fig. 2. Design rules for cross-β motifs in β-sandwiches. a, Cartoon representation of a 7-stranded immunoglobulin-like domain model formed by two β-sheets packing face-to-face, and the corresponding cross-β motif, which generates translations and rotations between the two opposing β-sheets. **b**, Topology diagram of a cross-β motif with circles and arrows representing β-strand residue positions and connections, respectively. Dark- and light-colored circles correspond to residues with sidechains pointing inwards or outwards from the β-sandwich, respectively. **c**, Efficiency of pairs of common β-arch loop geometries (described with ABEGO backbone torsions) in forming cross-β motifs obtained from Rosetta folding simulations (gray shaded squares). geometries were classified in four groups according to the sidechain directions of the adjacent residues. Colored squares group pairs of loops that, due to their sidechain orientations, have different requirements in β-strand length: in red or blue, if all β-strands need an odd or even number of residues, respectively; in green, if the β-strands of the first and second sheet need an odd and even number of residues, respectively; and in yellow for the opposite case (even and odd number of residues for the first and second sheet, respectively). Black-outlined boxes highlight loop combinations observed in natural Ig domains. On the right, examples of changes in cross-β motif geometry linked to β- arch loop geometry**. d**, β-arch helices are formed by a short α-helix connected to the adjacent β-strands with short loops, and are complementary to β-arch loops for connecting cross-β motifs. **e,** Topology diagram of a 7-stranded Ig domain. β-arch loops are indicated as L*i*, where *i* is the β-arch number. **f**, Examples of *de novo* designed Ig backbones generated with different geometries and β-arch connections following the described rules, colored from N-terminus *(blue)* to C-terminus *(red)*.

sidechain orientations (Supplementary Fig. 1), along with β-strand length, using Rosetta folding simulations with a sequence-independent model (*7*, *20*) biased by the ABEGO torsion bins specifying desired loop geometries (*21*) (Fig. 2c). It is convenient to describe the backbone geometry of loop residue positions with ABEGO torsion bins representing different areas of the Ramachandran plot ("A", right-handed α-helix region; "B", extended region; "E", extended region with positive *Φ*; "G", left-handed α-helix region; and "O", if the peptide bond deviates from planarity) (see Supplementary Fig. 1a for a definition). For cross-β motifs to form, the geometry of the two β-arch loops must allow the concerted spanning of the proper distance along the β-sheet pairing direction and along an axis connecting the two opposing β-sheets so that the two following β-strands cross and switch the order of β-strand pairing in the opposite β-sheet (Supplementary Fig. 2). Multiple pairs of β- arch loops with the same or different ABEGO torsion bins were found to fulfill these geometrical requirements (Fig. 2c), with sampled ranges of cross-β geometrical parameter values similar to or broader than those found in naturally occurring Ig domains (Supplementary Fig. 3). For example, β-arch loops "ABB" and "ABBBA" strongly favor cross-β motifs but with twist rotations (Supplementary Fig. 4) in opposite directions (Fig. 2c, right). Of the short β -arch loops we considered for design, only a few are present in the cross-β motifs of naturally occurring Ig domains (Fig. 2c), which are mostly built by longer or hypervariable loops (as is the case of the first β-arch). We next explored the efficiency of short α-helices (spanning 4-6 residues) connecting the two β-strands through short loops (of 1-3 residues) which we refer to as "β-arch helices". For cross-β motifs formed with β-arch helices, we identified efficient loop-helix-loop patterns (i.e. helix length together with adjacent loop ABEGO-types) for the four possible β-arch sidechain orientations (Supplementary Fig. 5). Overall, the formation and structure of cross-β motifs can in this way be encoded by combining β-arch loops and/or β-arch helices of specific geometry with β- strands compatible in terms of length and sidechain orientations.

Computational design of Ig domains - Based on these rules relating β-arch connections with cross-β motifs, we *de novo* designed 7-stranded Ig topologies (Fig. 2e, f). We generated protein backbones by Rosetta Monte Carlo fragment assembly using blueprints (*7*, *20*) specifying secondary structures and ABEGO torsion bins, together with hydrogen bond constraints specifying β-strand pairing. We explored combinations of β-strand lengths (between 5 and 8 residues) and register shifts between paired β- strands 3 and 6 (between 0 and 2 residues). β-arches 1 and 3 are those involved in the cross-β motif, and their connections were built with loop ABEGO-types having high cross-β propensity, as described above. We reasoned that β-arch helices may fit better in β-arch 3 than in β-arch 1 (Fig. 2e), which by construction is more embedded in the core, and explored topology combinations combining β-arch 1 loops with β-arch 3 helices. The three β-hairpin loops were designed with two residues for proper control of the orientation between the two paired β-strands according to the ββ-rule (*7*). Those topology combinations with β-strand lengths incompatible with the expected sidechain orientations of each β-arch and β-hairpin connection were automatically discarded. We then carried out Rosetta sequence design calculations (*22*, *23*) for the generated backbones. Loops were designed using consensus sequence profiles derived from fragments with the same ABEGO backbone torsions. Cysteines were not allowed during design to avoid dependence of correct folding on disulfide bond formation (in contrast to most natural Ig domains). As an implicit negative design strategy against edge-to-edge interactions promoting aggregation, we incorporated at least one inward-facing polar or charged amino acid (TQKRE) (*24*) into each solvent-exposed edge β-strand. Sequences were ranked based on energy and sidechain packing metrics, as well as local sequence-structure compatibility assessed by 9-mer fragment quality analysis (*4*). Folding

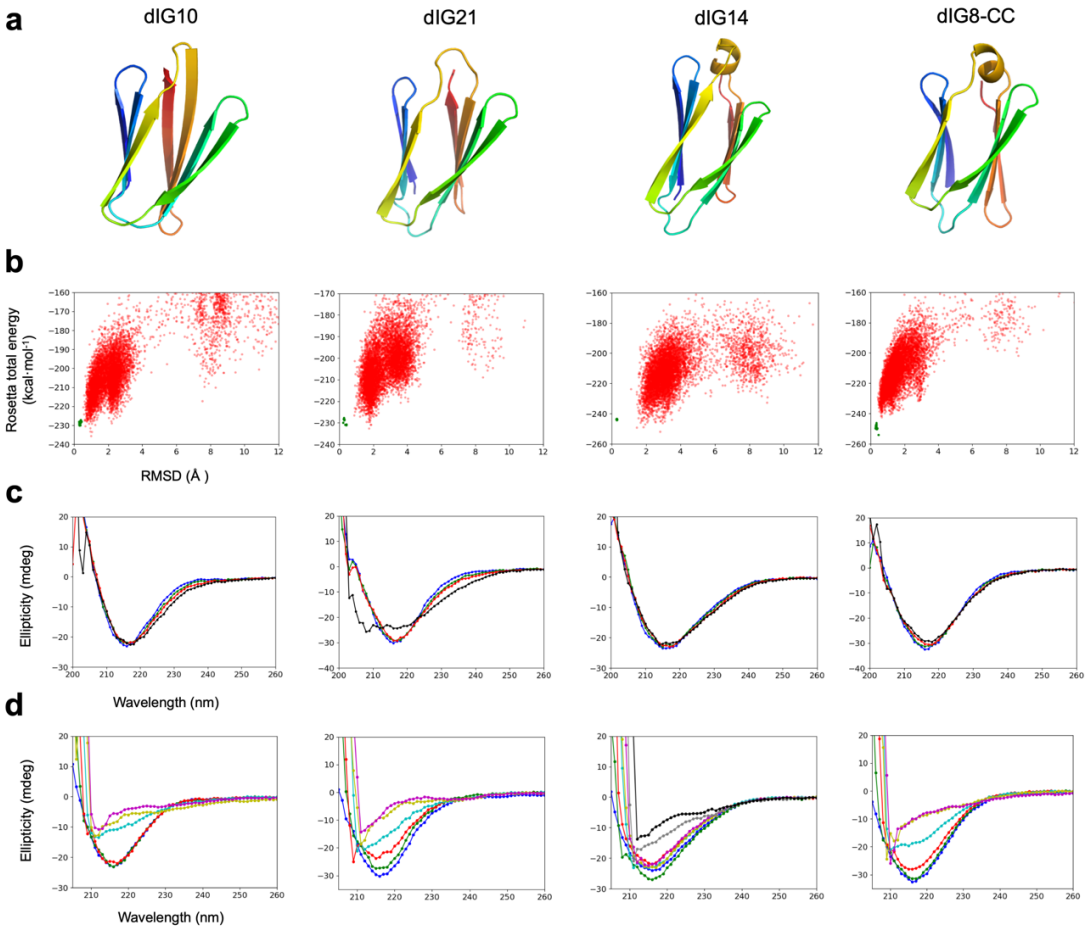


Fig. 3. Folding and stability of designed proteins. a, Examples of design models. **b**, Simulated folding energy landscapes, with each dot representing the lowest energy structure obtained from *ab initio* folding trajectories starting from an extended chain *(red dots)* or local relaxation of the designed structure *(green dots)*. The x-axis depicts the Cα-RMSD from the designed model and the y-axis, the Rosetta all-atom energy. **c,** Far-ultraviolet circular dichroism spectra (blue: 25 °C; green: 55°C; red: 75°C; black: 95°C). **d**, Far-ultraviolet circular dichroism spectra at different guanidine hydrochloride concentrations and at 25°C (blue: 0 M; green: 1 M; red: 2 M; cyan: 3 M; yellow: 4 M; magenta: 5 M; gray: 6 M; black: 7 M).

of the top-ranked designs was quickly screened by biased forward folding simulations (*5*), and those with near-native sampling were subjected to Rosetta *ab initio* folding simulations from the extended chain (*25*). The extent to which the designed sequences encode the designed structures was also assessed through AlphaFold (*26*) or RoseTTAFold (*27*) structure prediction calculations (see below).

Biochemical characterization of the designs- For experimental characterization, we selected 31 designs predicted to fold correctly by *ab initio* structure prediction (Fig. 3a, b); 29 of which had AlphaFold or RoseTTAFold predicted models with pLDDT > 80 and Cα atom root mean square deviations (Cα-RMSDs) < 2 Å to the design models (Supplementary Table 1). The designed sequences contain between 66 and 79 amino acids and are unrelated to naturally occurring sequences, with Blast (*28*) (E-values > 0.1) and more sensitive sequence-profile searches (*29*, *30*) finding very weak or no remote homology (E-values > 0.003) (Supplementary Table 2). The designs also differ substantially from natural Ig domains in global structure (with an average ± s.d.TM-score (31) of 0.54 ± 0.06; Supplementary Fig. 6), and cross-β twist rotation (close to zero, which are infrequent in natural Ig domains;

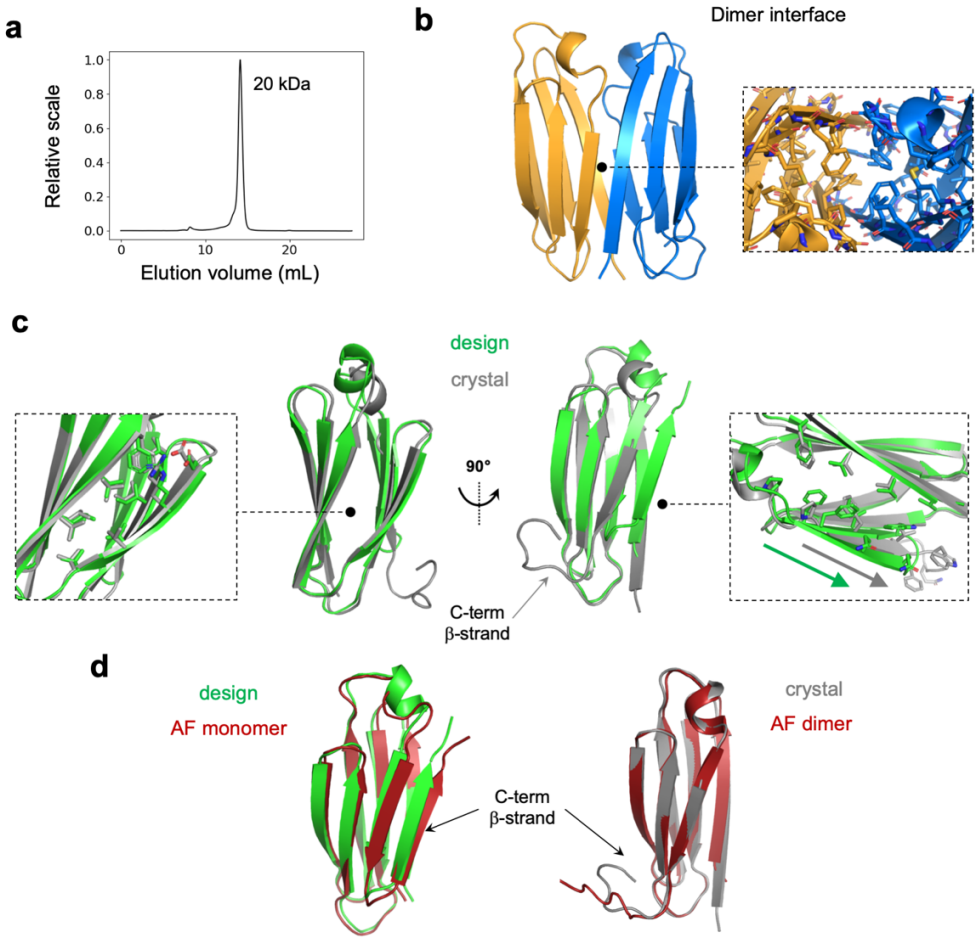


Fig. 4. Crystal structure of the dIG14 dimer. a, SEC-MALS analysis of dIG14 estimates a molecular weight corresponding to a dimer (M_W monomer = 9.7 kDa). **b**, Crystal structure of the homodimer interface formed by antiparallel pairing between β-strands 1 and 6 enabled by flipping out of the C-terminal β-strand; the monomer core becomes more accessible and the interface is primarily formed by hydrophobic contacts *(right inset)*. PDB accession code of the dIG14 crystal structure: 7SKP . **c**, dIG14 design model (*green*) in comparison with the crystal structure (*gray*, chain B). Sidechain packing interactions in the non-terminal edge β-strands were well recapitulated in the crystal structure *(left inset)*. A shift in β-strand pairing register observed in the crystal structure is highlighted by the two colored arrows *(right inset)*. **d**, The AlphaFold monomer prediction *(left)* superimposes well with the design model (Cα-RMSD 1.0 Å); while AlphaFold- Multimer *(right)* correctly predicts the monomer subunits in the crystal structure (Cα-RMSD 0.6 Å, except for the C-terminal disordered β-strand).

Supplementary Table 3). We obtained synthetic genes encoding for the designed amino acid sequences (design names are dIG*n*, where "dIG" stands for "designed ImmunoGlobulin" and "*n*" is the design number). We expressed them in *Escherichia coli*, and purified them by affinity and size-exclusion chromatography. Overall, 24 designs were present in the soluble fraction and 8 were monodisperse, had far-UV circular dichroism spectra compatible with an all-β protein structure, and were thermostable (T_m > 95 °C, except for dIG21 with T_m > 75 °C) (Fig. 3c, Supplementary Table 4, and Supplementary Fig. 7 and 8). In size-exclusion chromatography combined with multi- angle light scattering (tic), five designs were dimeric, one was monomeric (dIG21) and another one (dIG8) was found in equilibrium between monomer and dimer (Fig. 4a, Supplementary Fig. 7, 8 and 9). The

monomeric design had a well-dispersed ^{1}H-^{15}N HSQC nuclear magnetic resonance (NMR) spectrum consistent with a well-folded β-sheet structure (Supplementary Fig. 10).

Structural characterization of a dimeric *de novo* Ig design- The most stable design, dIG14, remained folded at 5 M guanidine hydrochloride (GdnCl) (Fig. 3d), had a well-dispersed ^{1}H-^{15}N HSQC spectra (Supplementary Fig. 10) and was found to be dimeric by SEC-MALS (Fig. 4a). To gain structural insight on its dimerization mechanism, we solved a crystal structure at 2.4 Å resolution (Fig. 4b-c, Supplementary Table 5) and found it was in excellent agreement with the computational model over the first five β-strands and their connections (Cα-RMSD of 0.8 Å; Fig. 4c). By contrast, the C-terminal region had three main differences: β-arch 3 helix was found in a different orientation, the register between paired β-strands 6 and 3 shifted by two β-strand positions (Fig. 4c, right inset), and the C-terminal β-strand flipped out of the structure, being disordered. This conformational difference altered the cross-β structure, exposed the protein core and formed an edge-to-edge dimer interface mediated by two antiparallel β-strand pairs (between β-strands 1 and 6 of each protomer), overall forming a 12- stranded β-sandwich (Fig. 4b). AlphaFold and RoseTTAFold predictions recapitulated the design model and did not predict these conformational differences, but the pLDDT values in the β-arch helix were quite low compared with the rest of the structure (Fig. 4d; Supplementary Fig. 11). Rosetta *ab initio* folding simulations sampled conformations closer to the crystal structure with energies similar to the design (Supplementary Fig. 11). Structure prediction of dIG14 as a homodimer with AlphaFold-Multimer (*32*) generated models closer to the crystal structure (Fig. 4d) despite formation of an incorrect dimer interface (Supplementary Fig. 11); the conformational differences between the design and crystal structure may be driven at least in part by the energetics of dimer interface formation.

Structural characterization and functionalization of a monomeric *de novo* designed Ig scaffold- For the dIG8 design, crystallization trials yielded no hits, but we reasoned that a disulfide bond could further rigidify the structure and promote crystallization. As disulfide bonds with high sequence separation are more stabilizing due to greater unfolded state entropy reduction, we computationally designed disulfide bonds between β-strands not forming a β-hairpin using a hash-based disulfide placement protocol (*33*) which searches for transformations between pairs of residue positions compatible with naturally occurring disulfide bond geometries (see Methods). We designed the double mutant dIG8-CC (V21C, V60C) (Fig. 5a), which, like the parental protein (Supplementary Fig. 7), was well-expressed, thermostable and was found in an equilibrium between monomers and dimers by SEC-MALS (Fig. 5b). We were able to obtain two crystal structures of dIG8-CC in two different space groups, with data to 2.05 and 2.30 Å resolution by molecular replacement using the design and RoseTTAFold predicted models (Supplementary Table 5). The asymmetric unit of both crystal structures contained four protomers, and all of them closely matched the computational model with Cα-RMSDs ranging between 1.0 and 1.3 Å (Fig. 5c). The designed cross-β motif combines a β-arch loop (ABABB) with a β-arch helix (BB-H5-B), and both were well recapitulated (Cα-RMSDs ranging between 0.7 and 1.0 Å for the two connections) across the eight monomer copies, suggesting high structural preorganization of the designed connections (Fig. 5d). The sidechain of residue C21 was found in two different conformations, disulfide-bonded with C60 as in the design and unbound (Supplementary Table 6), which suggests low stability of the disulfide bond (Supplementary Fig. 12) and that it is not essential for proper folding of dIG8-CC. This is consistent with the high stability determined for parental dIG8 without the disulfide bridge (Supplementary Fig. 7).

The crystal structures also revealed an edge-to-edge dimer interface between the N-

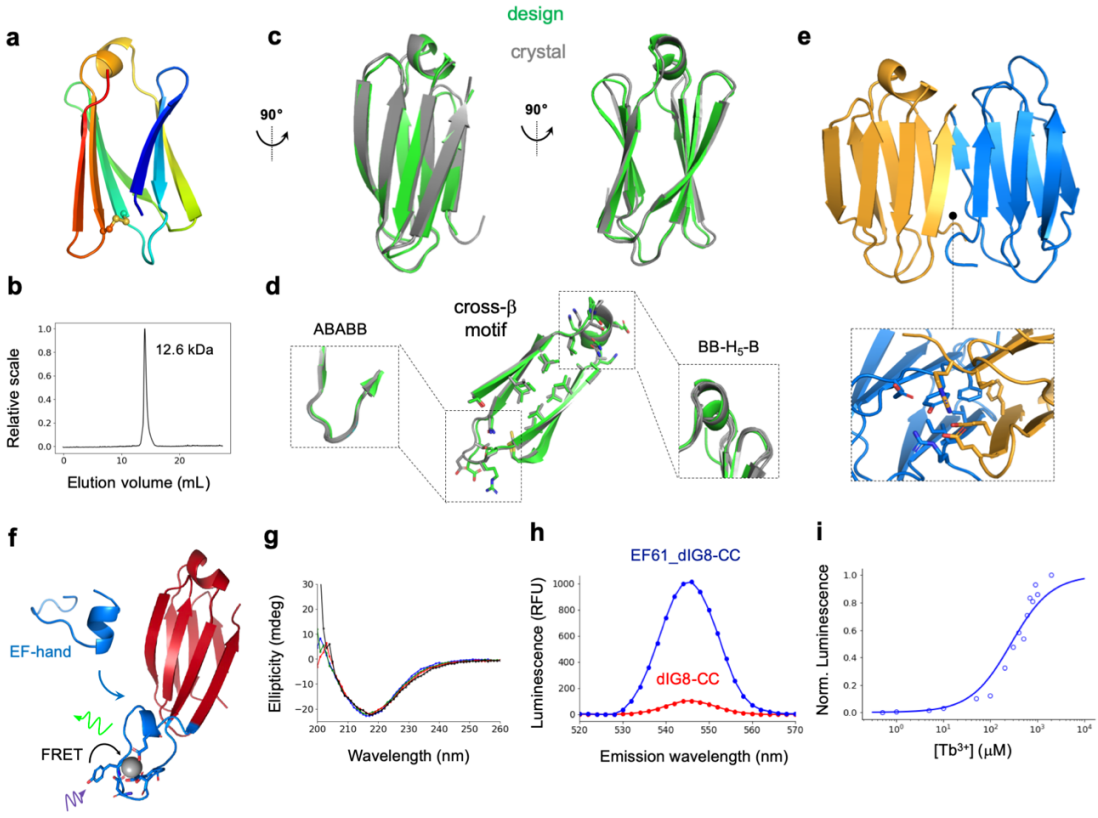


Fig. 5. Crystal structure of dIG8-CC and functional loop scaffolding. **a**, Design model of dIG8-CC with a disulfide bridge (*spheres*) between β-strands 3 and 6. **b**, SEC-MALS analysis of dIG8-CC estimates a molecular weight between monomer (8.3 kDa) and dimer (16.6 kDa). **c**, Design model (*green*) in comparison with the crystal structure with PDB accession code 7SKP (*gray,* chain C). **d**, Cross-β motif connections and core sidechain interactions in the design and the crystal structure. The β-arch helix and loop conformations are well preserved across monomer copies in the crystal asymmetric units *(insets)*. **e**, Crystal homodimer interface by parallel pairing between the two terminal β-strands, which are stabilized through hydrophobic and salt-bridge interactions *(inset)*. **f**, Computational model of dIG8-CC with a grafted EF-hand motif (design EF61_dIG8-CC, *cartoon)*, showing Tb^{3+} *(sphere)* bound to EF-hand motif residues *(sticks)*. Tb^{3+} luminescence is sensitized by absorption of light (*purple*) by a proximal tyrosine residue on the EF-hand motif with subsequent fluorescence resonance energy transfer (FRET) to Tb^{3+}, resulting in Tb^{3+} luminescence *(green)*. **g**, Far-ultraviolet circular dichroism spectra of EF61_dIG8-CC without Tb^{3+} (blue: 25 °C; green: 55 °C; red: 75 °C; black: 95 °C). **h**, Time-resolved luminescence emission spectra in 100 μM Tb^{3+} final concentrations for EF61_dIG8-CC and dIG8-CC at 20 μM. **i**, Tb^{3+} concentration- dependent time-resolved luminescence intensity of 20 μM EF61_dIG8-CC using excitation wavelength λ_{ex} = 280 nm and emission wavelength λ_{em} = 544 nm. Normalized intensities are fit to a one-site binding model by non-linear least squares regression (K_d = 267 μM).

and C- terminal β-strands, overall forming a 14-stranded β-sandwich (Fig. 5e). Docking calculations on dIG8-CC suggested that the β-sandwich edge formed by the two terminal β-strands is more dimerization-prone than the opposite edge (Supplementary Fig. 13), mainly due to a more symmetrical backbone arrangement and complementary hydrophobic and salt-bridge interactions in the former, and the presence of more inward-pointing charged residues in the latter. In contrast to dIG14, AlphaFold correctly predicted the dIG8-CC monomer crystal structure with very high confidence across all residues and did not

change that prediction in the context of the homodimer. The closest Ig structural analogues found across the PDB and the AlphaFold Protein Structure Database (*34*) had a TM-score ≤ 0.65 (Supplementary Fig. 14); and contained more irregular β- strands, longer loops, and differences in the β-strand pairing organization.

We next sought to investigate whether *de novo* designed immunoglobulins could be functionalized by scaffolding ligand-binding loops. We set out to computationally graft an EF- hand calcium-binding motif (PDB accession code 1NKF) into the β-hairpins of dIG8-CC. To facilitate motif grafting, we designed N-terminal linkers containing between 0 and 3 residues with an extended backbone conformation, and C-terminal linkers containing between 0 and 10 residues keeping the α-helical secondary structure of the C-terminal side of the EF-hand motif. We selected 12 designs for experimental testing with minimal linker lengths and spanning the three insertion sites. Design EF61_dIG8-CC (Fig. 5f), with the EF-hand motif grafted at the C-terminal β-hairpin of dIG8-CC after residue 61, was the best expressed and monodisperse by size-exclusion chromatography, and was found to be thermostable by far-UV circular dichroism (Fig. 5g), as was the parent design dIG8-CC. Since EF-hand motifs generally bind terbium, we assessed ligand- binding by terbium luminescence, which can be sensitized by energy transfer (*35*) from a proximal tyrosine residue on the grafted EF-hand motif upon excitation at 280 nm wavelength. For increasing luminescence signal-to-noise ratio, we carried out time-resolved luminescence measurements taking advantage of the long luminescence lifetime of terbium (*36*, *37*). EF61_dIG8-CC mixed with 100 μM $TbCl_3$ displayed a 10-fold higher luminescence emission intensity at 544 nm than dIG8-CC without the EF-hand motif (Fig. 5h). Tb^{3+} titrations in the presence of EF61_dIG8-CC displayed a hyperbolic increase in luminescence with increasing Tb^{3+} concentrations (Fig. 5i; Supplementary Fig. 15a). In competitive binding titrations, Tb^{3+} luminescence intensity decreased with increasing Ca^{2+} concentrations, showing that Ca^{2+} competes with Tb^{3+} for the grafted EF-hand motif (Supplementary Fig. 15b).

DISCUSSION

Since initial attempts in the early 90's (*38–40*), the *de novo* design of globular β-sheet proteins with high-resolution structural validation had remained elusive until very recently, when they were enabled by considerable advances in our understanding of how to program the curvature of β-sheets and the orientation of their connecting loops into an amino-acid sequence. Here, we describe the first successful *de novo* design of an immunoglobulin-like domain with high stability and accuracy, which was confirmed by crystal structures. This success became possible by elucidating the requirements for effective formation of cross-β motifs, which establish the non- local central core of Ig folds by structuring β-arch connections through short loops and helices, while favoring sidechain orientations compatible with the length and pleating of the sandwiched β-sheets.

The cross-β motifs of our designs differ from natural ones in several ways. Our cross-β motifs are formed by combining short β-arch loops not seen in natural Ig domains (Fig. 2c), which generally have more complex loops (including a complementarity-determining region (CDR) in the first β-arch of the cross-β motif found in antigen-binding regions of antibodies), and are stabilized by hydrophobic interactions without incorporating sequence motifs typically found in the core strands B, C, E and F of natural Ig domains. For example, the disulfide bond of dIG8-CC is between two β-strands paired in the same β-sheet in contrast to the sheet-to-sheet disulfide bridge found between strands B and F in many Ig domains. The tyrosine corner which stabilizes Greek keys in many natural β-barrels and β-sandwiches (*15*, *18*) was also not needed in our designs. These differences in sequence requirements reflect the substantial structural differences between our designs and natural Ig domains. The designs contain cross-β motifs less twisted than those from natural Ig domains, and their overall structural (average TM-score of 0.54) and sequence (Supplementary Table 2) similarity is very low (HHPred did identify matches to short segments of β -sandwiches, including one Ig domain (PDB accession code 2R39), with locally similar alternating

patterns of hydrophobic and polar amino acids typical of β-strands).

Several of the designs tended to dimerize in solution, highlighting design challenges in preventing self-interactions between β-sheets. Solvent-exposed β-strand edges favor intermolecular β-strand pairing through backbone hydrogen bonds (between the unpaired NH- and CO- groups) and hydrophobic interactions at the interface between monomers. As in previous *de novo* β-sheet design studies (*5*, *7*, *8*), we used an implicit negative design strategy to disfavor association by favoring polar or charged amino acids at inward-facing positions of the edge β- strands to weaken interface sidechain interactions. Explicit negative design against possible edge- to-edge dimer interfaces is an alternative, but remains challenging as it requires enumerating many possible negative states: the crystal structures of two designs show two possible interfaces (one including structural rearrangement of the monomer), and we cannot rule out the possibility that other dimer interfaces formed in designs that were not crystallized (via parallel or antiparallel edge-strand pairing with varied register shifts). Alternatively, negative design against edge-to-edge interfaces can be encoded in protein backbone irregularities – e.g. β-bulges, prolines or short protective β-strands – disfavoring the ideal geometry for hydrogen-bonded β-strand pairing (*41*).

The edge-to-edge dimer interfaces in the crystal structures of our designs differ from those found between the heavy- and light-chains of antibodies, which are arranged face-to-face. For engineering antibody-like formats presenting several loops targeting one or multiple epitopes, designing dimeric Ig interfaces through the β-sandwich edge formed by the terminal β-strands has the advantage over face-to-face dimers of decreasing the number of exposed β-strand edges, thereby reducing aggregation-propensity. It will likely be useful to custom-design both edge-to- edge and face-to-face dimers from our *de novo* Ig domains; these would present loops from the two monomers in different relative orientations, and depending on the target structure and the loops involved, one of these two arrangements will likely be better suited than the other for designing shape-complementary binding interfaces. Another advantage of controlling the N- and C-termini of the two monomeric subunits can be positioned in close proximity to allow fusion through short or compact connections into rigid and hyperstable single-chain constructs –similar in spirit to single-chain variable fragments (scFvs) but with greater structural control and higher stability. The high stability of our designs opens up exciting possibilities for grafting functional loops, as shown for the EF-hand terbium-binding motif inserted into the C-terminal β-hairpin of dIG8-CC. The β-hairpins in our scaffolds can be readily extended to incorporate ligand- and protein-binding motifs, functional peptide motifs, or complementarity-determining regions (CDRs) of antibodies or nanobodies (it is likely more straightforward to insert functional loops into β-hairpins than into β-arches, since the latter tend to form more slowly and need to be highly structured, but this remains to be studied and may vary depending on the loop to be inserted). In antibodies, the CDRs are located on one side of the β-sandwich (at the bottom given the orientation displayed in Figs. 1-5), and we inserted the terbium binding motif on this side, but the robustness of our scaffolds could allow insertions on the other side as well. Ultimately, achieving the structural control over the Ig backbone together with the high expression levels and stability of *de novo* designed proteins in general should lead to a versatile generation of antibody-like scaffolds with improved properties.

Acknowledgements

We are grateful to Laura Company and Joan Pous from the joint IBMB/IRB Automated Crystallography Platform and the Protein Purification Service for assistance during SEC-MALS, purification procedures, and crystallization experiments. We thank Lauren Carter and Cameron Chow for assistance with SEC-MALS experiments at the Institute for Protein Design and shipment of protein samples for NMR. We also thank Minkyung Baek for assistance with structure predictions with RoseTTAFold. The authors would further like to thank the ESRF and ALBA synchrotrons for beamtime allocation and the

respective beamline staff for assistance during diffraction data collection. We acknowledge computing resources provided by Rosetta@Home volunteers, the Galicia Supercomputing Center (CESGA), and the Red Española de Supercomputación (grants BCV-2021-1-0014 and BCV-2021-3-0010). This research was supported by grants from the Spanish Ministry of Science and Innovation (RYC2018-025295-I, EUR2020-112164 and PID2020-120098GA-I00). This study was also supported in part by grants from Spanish and Catalan public and private bodies (grant/fellowship references MCIN/AEI/10.13039/501100011033/PID2019-107725RG-I00, 2017SGR3 and Fundació "La Marató de TV3" 201815). S.R.M. acknowledges grant BES2016-076877 from the Spanish State Agency for Research (MCIN/AEI/10.13039/501100011033) and the European Social Fund "ESF invests in your future". U.E. was funded by a Beatriu de Pinós post-doctoral fellowship (AGAUR-MSCA COFUND 2018BP00163. J.R.T. was supported by an EMBO postdoctoral fellowship (under grant agreement ALTF 145-2021). J.C.K. was supported by a National Science Foundation Graduate Research Fellowship (grant DGE-1256082). D.B. and T.M.C. acknowledge the Howard Hughes Medical Institute. We thank the Princess Margaret Cancer Centre for funding of the NMR facility. The Structural Genomics Consortium is a registered charity (no: 1097737) that receives funds from Bayer AG, Boehringer Ingelheim, Bristol Myers Squibb, Genentech, Genome Canada through Ontario Genomics Institute [OGI-196], EU/EFPIA/OICR/McGill/KTH/Diamond Innovative Medicines Initiative 2 Joint Undertaking [EUbOPEN grant 875510], Janssen, Merck KGaA (aka EMD in Canada and US), Pfizer and Takeda. The content herein is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author contributions

E.M., T.M.C., F.X.G.R. and D.B. designed the research. T.M.C. carried out design calculations, protein expression, purification and CD experiments. S.R.M. cloned, expressed, purified and characterized proteins. S.R.M., T.G., and U.E. crystallized proteins, and U.E. collected and analyzed diffraction data. T.M.C. and J.C.K. designed and experimentally tested EF- hand terbium-binding loops. J.R.T. carried out docking calculations. M.N. expressed, purified and performed CD and terbium-binding experiments. F.X.G.R. solved crystal structures. H.K.H. analyzed design structural diversity. A.M. provided crosslinking scripts for disulfide bridging. S.H. and C.H.A. carried out NMR spectroscopy. E.M. set up the design methods, carried out design calculations and performed the structural analyses. E.M., T.M.C., F.X.G.R. and D.B. prepared the manuscript with input from all authors.

References

1. C. Jost, A. Plückthun, Engineered proteins with desired specificity: DARPins, other alternative scaffolds and bispecific IgGs. Curr Opin Struct Biol. 27, 102–112 (2014).
2. J. R. Kintzing, M. V. Filsinger Interrante, J. R. Cochran, Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment. Trends Pharmacol Sci. 37, 993–1008 (2016).
3. F. Sha, G. Salzman, A. Gupta, S. Koide, Monobodies and other synthetic binding proteins for expanding protein science: Monobodies and Other Synthetic Binding Proteins. Protein Sci. 26, 910–924 (2017).
4. E. Marcos, D. Silva, Essentials of de novo protein design: Methods and applications. WIREs Comput Mol Sci. 8 (2018), doi:10.1002/wcms.1374.
5. E. Marcos *et al.*, Principles for designing proteins with cavities formed by curved β sheets. Science. 355, 201–206 (2017).
6. J. Dou *et al.*, De novo design of a fluorescence-activating β-barrel. Nature. 561, 485–491 (2018).
7. N. Koga *et al.*, Principles for designing ideal protein structures. Nature. 491, 222–227 (2012).
8. E. Marcos *et al.*, De novo design of a non-local β-sheet protein with high stability and accuracy. Nat Struct Mol Biol. 25, 1028–1034 (2018).

9. A. A. Vorobieva *et al.*, De novo design of transmembrane β barrels. Science. 371 (2021), doi:10.1126/science.abc8182.
10. P. Bork, L. Holm, C. Sander, The Immunoglobulin Fold. J Mol Biol. 242, 309–320 (1994).
11. D. M. Halaby, A. Poupon, J.-P. Mornon, The immunoglobulin fold family: sequence analysis and 3D structure comparisons. Protein Engineering, Design and Selection. 12, 563–571 (1999).
12. J. Hennetin, B. Jullian, A. C. Steven, A. V. Kajava, Standard Conformations of β-Arches in β-Solenoid Proteins. J Mol Biol. 358, 1094–1105 (2006).
13. A. E. Kister, A. V. Finkelstein, I. M. Gelfand, Common features in structures and sequences of sandwich-like proteins. Proc Natl Acad Sci USA. 99, 14137–14141 (2002).
14. J. Clarke, E. Cota, S. B. Fowler, S. J. Hamill, Folding studies of immunoglobulin-like β- sandwich proteins suggest that they share a common folding pathway. Structure. 7, 1145– 1153 (1999).
15. J. M. Hemmingsen, K. M. Gernert, J. S. Richardson, D. C. Richardson, The tyrosine corner: A feature of most greek key β-barrel proteins. Protein Sci. 3, 1927–1937 (1994).
16. J. S. Richardson, in Advances in Protein Chemistry (Elsevier, 1981; https://linkinghub.elsevier.com/retrieve/pii/S0065323308605203), vol. 34, pp. 167–339.
17. E. G. Hutchinson, J. M. Thornton, The Greek key motif: extraction, classification and analysis. Protein Eng Des Sel. 6, 233–245 (1993).
18. S. J. Hamill, A. Steward, J. Clarke, The folding of an immunoglobulin-like greek key protein is defined by a common-core nucleus and regions constrained by topology. Journal of Molecular Biology. 297, 165–178 (2000).
19. K. W. Plaxco, K. T. Simons, D. Baker, Contact order, transition state placement and the refolding rates of single domain proteins. J Mol Biol. 277, 985–994 (1998).
20. J. K. Leman *et al.*, Macromolecular modeling and design in Rosetta: recent methods and frameworks. Nat Methods. 17, 665–680 (2020).
21. Y.-R. Lin *et al.*, Control over overall shape and size in de novo designed proteins. Proc Natl Acad Sci USA. 112, E5478–E5485 (2015).
22. B. Kuhlman, D. Baker, Native protein sequences are close to optimal for their structures. Proc Natl Acad Sci USA. 97, 10383–10388 (2000).
23. B. Kuhlman *et al.*, Design of a Novel Globular Protein Fold with Atomic-Level Accuracy. Science. 302, 1364–1368 (2003).
24. J. S. Richardson, D. C. Richardson, Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA. 99, 2754–2759 (2002).
25. P. Bradley, Toward High-Resolution de Novo Structure Prediction for Small Proteins. Science. 309, 1868–1871 (2005).
26. J. Jumper *et al.*, Highly accurate protein structure prediction with AlphaFold. Nature. 596, 583–589 (2021).
27. M. Baek *et al.*, Accurate prediction of protein structures and interactions using a three- track neural network. Science. 373, 871–876 (2021).
28. C. Camacho *et al.*, BLAST+: architecture and applications. BMC Bioinformatics. 10, 421 (2009).
29. M. Remmert, A. Biegert, A. Hauser, J. Söding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods. 9, 173–175 (2012).
30. L. Zimmermann *et al.*, A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J Mol Biol. 430, 2237–2243 (2018).
31. Y. Zhang, TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res. 33, 2302–2309 (2005).
32. R. Evans *et al.*, "Protein complex prediction with AlphaFold-Multimer" (preprint, Bioinformatics, 2021), , doi:10.1101/2021.10.04.463034.

33. S. Yao *et al.*, De novo design and directed folding of disulfide-bridged peptide heterodimers. Nat Commun. 13, 1539 (2022).
34. K. Tunyasuvunakool *et al.*, Highly accurate protein structure prediction for the human proteome. Nature. 596, 590–596 (2021).
35. S. C. Zondlo, F. Gao, N. J. Zondlo, Design of an Encodable Tyrosine Kinase-Inducible Domain: Detection of Tyrosine Kinase Activity by Terbium Luminescence. J Am Chem Soc. 132, 5619–5621 (2010).
36. S. Pandya, J. Yu, D. Parker, Engineering emissive europium and terbium complexes for molecular imaging and sensing. Dalton Trans., 2757 (2006).
37. A. M. Lipchik, L. L. Parker, Time-Resolved Luminescence Detection of Spleen Tyrosine Kinase Activity through Terbium Sensitization. Anal. Chem. 85, 2582–2588 (2013).
38. T. P. Quinn *et al.*, Betadoublet: de novo design, synthesis, and characterization of a beta- sandwich protein. Proc Natl Acad Sci USA. 91, 8747–8751 (1994).
39. Y. Yan, B. W. Erickson, Engineering of betabellin 14D: Disulfide-induced folding of a β- sheet protein. Protein Sci. 3, 1069–1073 (1994).
40. M. H. Hecht, De novo design of beta-sheet proteins. Proc Natl Acad Sci USA. 91, 8729– 8730 (1994).
41. X. Hu, H. Wang, H. Ke, B. Kuhlman, Computer-Based Redesign of a β Sandwich Protein Suggests that Extensive Negative Design Is Not Required for De Novo β Sheet Design. Structure. 16, 1799–1805 (2008).

METHODS

Structural analysis of β-arch loops- β-arch loops of less than 9 residues were collected from a non-redundant set of 5,857 PDB structures with sequence identity < 30% and resolution ≤ 2.0 Å. They were identified by first assigning the secondary structure with DSSP (*42*), and ensuring they were connecting β-strands with no hydrogen-bond pairing between them (the first and last residue of each assigned β-strand were considered the end residues connecting to the loops). The ABEGO torsion bins of each loop position was assigned based on their φ/ψ backbone dihedrals as defined in Supplementary Fig. 1a. The sidechain orientations of the two residues (*i* and *j*) preceding and following the β-arch loop are a function of the relative orientation between their Cα-Cβ vector and the translation vector (v1) connecting their Cα atoms, as shown in Supplementary Fig. 1b. The β-arch sliding distance was calculated as the dot product between v1 and the CO vector of the preceding residue (v1 • CO_i), which points along the β-sheet hydrogen bond direction. If the dot product between v1 and the Cα- Cβ (i) vector of the preceding residue is negative, then the sliding distance is calculated as v1 • -CO_i. The β-arch twist was calculated as the dihedral between positions Cα (*i*-2), Cα (*i*), Cα (*j*), and Cα (*j*+2).

Cross-β motif analysis- To extract the cross-β geometrical parameters we calculated the rigid body transformation between two reference frames defined at the two β-sheets comprising the cross-β motif. For the first β-sheet (formed by the two N-terminal strands, 1 and 3, of the motif), the reference frame was built with the vectors $S1$, which defines the direction of β-strand 1 (from N to C-termini), $S31$, which connects the centers of the two strands (Supplementary Fig. 2), and P_N as the vector orthogonal to the β-sheet calculated as the cross product between the $S1$ and $S31$ vectors ($P_N = S1 \times S31$). For the second β-sheet (formed by the two C-terminal strands, 2 and 4, of the motif), the reference frame was calculated in the same way with the equivalent vectors $S4$, $S24$ and P_C. To minimize the dependence of cross-β parameters on differences in the internal geometry of β-strands from the two different β-sheets, we pre-generated a template antiparallel strand dimer that, before calculating the transform, is superimposed on each of the two strand dimers of the cross-β motif. The transform rotational angles were calculated as the Euler angles of the transform (twist, roll and tilt). The cross-β motif distance was calculated between the centers of the two strand dimers. The β-arch sliding distance in a cross-β motif was calculated as the dot product between the translation vectors and the vector $S31$.

Structural analysis of naturally occurring immunoglobulin-like domains- We searched for Ig-like domains classified in SCOP (*43*) as "Ig-like beta-sandwich" folds (SCOP ID 2000051) and selected those with X-ray resolution ≤ 2.5 Å, yielding a total of 467 annotated domains.

Protein backbone generation and sequence design- We specified blueprint files for each target protein topology and constructed poly-valine backbones with the RosettaScripts (*44*) implementation of the Blueprint Builder (*7*) mover, which carries out Monte Carlo fragment assembly using 9- and 3-residue fragments picked based on the secondary structure and ABEGO torsion bins specified at each residue position. We used the *fldsgn_cen* centroid scoring function with reweighted terms accounting for backbone hydrogen bonding (lr_hb_bb) and planarity of the peptide bond (omega).

For constructing cross-β motifs, we followed a two-step procedure. First, the two N- terminal strands of the motif (strands 1 and 3) were generated as antiparallel β-strand dimers of desired length from φ/ψ values typical of β-strands (extended region of the Ramachandran plot) and relaxed using hydrogen-bond pairing restraints. Second, the cross-β loops and C-terminal strands (strands 2 and 4) were then appended by fragment assembly using the Blueprint Builder, as described above, combined with a strand pairing energy bonus between strands 2 and 4. We assign the two N-terminal strands to different chains (A and B), and the resulting jump between the two chains allows to fold the two C-terminal strands independent of each other. Then, the secondary structures of the resulting backbones were calculated by DSSP (*42*) and those with a secondary structure identity to that defined in the

blueprints below 90% were discarded to guarantee correct strand pairing formation. The filtered backbones needed to fulfil two additional properties to be considered a cross-β motif: (1) the two C-terminal strands must form antiparallel strand pairing with each other, but not with any of the N-terminal strands (to guarantee β-sandwich formation); (2) the two β-arches must cross. For the latter, we checked crossing based on the relative orientation between the two vectors orthogonal to each of the two β-sheet planes packing face-to-face. The P_N vector orthogonal to the β-sheet formed by the two N-terminal strands is calculated as the cross product between the S_1 and S_{31} vectors ($P_N = S_1 \times S_{31}$) as described above. The P_C vector orthogonal to the β-sheet formed by C-terminal strands is calculated similarly as $P_C = S_4 \times S_{24}$ as described above. If the two orthogonal vectors are parallel (if $P_N \bullet P_C > 0$) the two β- arches were considered to cross.×

For designing 7-stranded Ig backbones, we carried out hundreds of independent blueprint- based trajectories folding each target topology in one step followed with a backbone relaxation using strand pairing constraints. We encouraged correct formation of strand pairs using custom python scripts writing distance and angle constraints specifying backbone hydrogen bond pairing at each pair of residue positions. The generated backbones were subsequently filtered based on their match with the secondary structure and ABEGO torsion bins specified in the corresponding blueprint files, and their long-range backbone hydrogen bond energy (lr_hb_bb score term). We carried out *FastDesign* (*45*) calculations using the Rosetta all-atom energy function *ref2015* (*46*) to optimize sidechain identities and conformations with low-energy, efficiently packing the protein core, and compatible with their solvent accessibility. Designed sequences were filtered based on the average total energy, Holes score (*47*), buried hydrophobic surface, and sidechain-backbone hydrogen bond energy (for better stabilizing β-arch geometry). For loop residue positions, we restricted amino acid identities based on sequence profiles derived from naturally occurring loops with the same ABEGO torsion bins (*5*).

Sequence-structure compatibility evaluation- The local compatibility between the designed sequences and structures was evaluated based on fragment quality. Sequence-structure pairs were considered locally compatible if for all residue positions at least one of the picked 9-mer fragments (based on sequence and secondary structure similarity with the design) had a RMSD below 1.0 Å. For designs fulfilling this requirement, we assessed their folding by Rosetta *ab initio* structure prediction in two steps. We started screening hundreds of designs quickly with biased forward folding simulations (*5*) (BFF) using the three 9- and 3-mers closer in RMSD to the design. Those designs with a substantial fraction (>10%) of BFF trajectories sampling structures with RMSDs to the design below 1.5 Å were then selected for standard Rosetta *ab initio* structure prediction (*25*). We ran AlphaFold (*26*) and the PyRosetta version of RoseTTAFold (*27*) with a local installation and using default parameters.

Docking calculations- HADDOCK (*48*) was used for the evaluation of the crystallographic interface of the design. We picked the first chain from the dIG8-CC crystal structure and used two copies of this monomer for all two-body docking simulations. Taking advantage of the ability of HADDOCK to build missing atoms, we constructed the mutants by renaming and removing all atoms but those forming the backbone (N, Cα, C, O) and the Cβ (to maintain sidechain directionality). For the simulations targeting the crystallographic interface, we selected all residues pertaining to the first and seventh strands (segments 1-7 and 65-70) as active residues to drive the docking. For the ones aiming to the opposite interface, all residues from the third and fourth strands (segments 30-35 and 39-45) were instead used as active residues. For all docking simulations, we defined two different sets of symmetry restraints as follows: (1) we applied C2 symmetry restraints to assure a 180o symmetry axis between both molecules, and (2) enabled non-crystallographic restraints (NCS) to enforce identical intermolecular contacts. All remaining docking and analysis parameters were kept as default. In terms of analysis, the generated models were evaluated by the default HADDOCK scoring function. This mathematical approximation is a weighted linear combination of different

energy terms including: van der Waals and electrostatic intermolecular energies, a desolvation potential and a distance restraint energy term. The scoring step is followed by a clustering procedure based on the fraction of common contacts, and the resulting clusters are re-ranked according to the average HADDOCK score of the best 4 cluster members. For comparison purposes, we used the exact same set of parameters for all docking simulations and selected the top model from the best ranked cluster.

Design of disulfide bonds- The identification of the position of disulfide bonds was carried out with a novel motif hashing protocol (*33*). 30,000 examples of native disulfide geometries were extracted from high resolution protein crystal structures in the PDB. The relative orientation of the backbone atoms was calculated by determining the translation and rotation matrix between the two sets of backbone atoms. These translation and rotation matrices were hashed and stored in a hash table with the associated conformation of the sidechains. Once the hash table has been completed by including all of the examples of disulfides from the PDB, the hash table can be utilized to place disulfides into de novo proteins by evaluating the relative orientation within a designed protein to find which residue pairs match an example from the hash table. All of the code necessary to generate the hash tables and run the disulfide placement protocol can be found in https://github.com/atom- moyer/stapler.

Design of EF-hand calcium binding motifs- A minimal EF-hand motif from Protein Data Bank (PDB) accession code 1NKF (*49*) was generated by truncating the PDB file 3-dimensional coordinates to the minimal Ca^{2+}-binding sequence DKDGDGYISAAE. RosettaRemodel (*50*) blueprint files were generated from the 3- dimensional coordinates of the dIG8 computational model and minimal EF-hand motif, and an in- house script used to write RosettaRemodel blueprint files for domain insertion of the minimal EF- hand motif into dIG8. 132 blueprint files were generated to insert the EF-hand motif after residues 8, 28, and 61 of dIG8 while systematically sampling N-terminal linker lengths of 0-3 residues with β-sheet secondary structure and C-terminal linker lengths of 0-10 residues with α-helical secondary structure. RosettaRemodel was run three times for each blueprint file using the pyrosetta.distributed and dask python modules (*51–53*). Linker compositions were *de novo* designed in RosettaRemodel using specific sets of amino acids defined in the blueprint files at each position of the N-terminal and C-terminal linkers while preventing repacking of EF-hand motif sidechain rotamers required for chelating Ca^{2+}. Out of 396 domain insertion simulations, 86 successfully closed the N-terminal and C-terminal linkers producing single-chain decoys. On each decoy, a custom PyRosetta script was run to append a Ca^{2+} ion into the EF-hand motif. Decoys were then relaxed via Monte Carlo sampling of protein sidechain repacking and protein sidechain and backbone minimization steps with a full-atom Cartesian coordinate energy function(*46*) with coordinate constraints applied to the aspartate and glutamate residues chelating the Ca^{2+} ion. The 86 resulting designs were scored in RosettaScripts (*44*) with an in-house XML script. Concomitantly, each of the 86 designs were forward folded (*25*) after temporarily stripping out the Ca^{2+} ion from each decoy, and the ff_metric algorithm used to evaluate funnels (*54*). To select designs for experimental validation, the following computational protein design metric filters were applied: buns_all_heavy_ball $\leq$ 1.0; buns_all_heavy_ball_interface $\leq$ 1.0; total_score_res $\leq$ -3.7; geometry = 1.0. Filtered designs were ranked ascending primarily on buns_all_heavy_ball, ascending secondarily on ff_metric, and ascending tertiarily on total_score_res. To experimentally test designs at the three domain insertion sites, the top three ranked designs at each of the three domain insertion sites were selected. To experimentally test designs with the shortest N-terminal and C-terminal linkers, the top three ranked designs with up to a 3-residue N-terminal linker and up to a 2-residue C-terminal linker were selected. 12 designs in total were selected for experimental characterization after mutating positions compatible with disulfide bonds to cysteines.

Recombinant expression and purification of the designed proteins for biophysical studies- Synthetic genes encoding for the

selected amino acid sequences were ordered from Genscript and cloned into the pET-28b+ expression vector, with the genes of interest inserted within NdeI and XhoI restriction sites and the pET28b backbone encoding an N-terminal, thrombin-cleavable His6-tag. Escherichia coli BL21 (DE3) competent cells were transformed with these plasmids, and starter cultures from single colonies were grown overnight at 37°C in Luria- Bertani (LB) medium supplemented with kanamycin.Overnight cultures were used to inoculate 50 ml of Studier autoinduction media (*55*) with antibiotic as done in a previous study (*56*). Cells were harvested by centrifugation and resuspended in a 25 mL lysis buffer (20 mM imidazole in PBS containing protease inhibitors), and lysed by microfluidizer. PBS buffer contained 20 mM NaPO4, 150 mM NaCl, pH 7.4. After removal of insoluble pellets, the lysates were loaded onto nickel affinity gravity columns to purify the designed proteins by immobilized metal-affinity chromatography (IMAC). The expression of purified proteins was assessed by SDS- polyacrylamide gel; and protein concentrations were estimated from the absorbance at 280 nm measured on a NanoDrop spectrophotometer (ThermoScientific) with extinction coefficients predicted from the amino acid sequences using the ProtParam tool (https://web.expasy.org/protparam/). Proteins were further purified by size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column.

Circular dichroism- Far-UV circular dichroism measurements were carried out with a JASCO spectrometer. Wavelength scans were measured from 260 to 195 nm at temperatures between 25 and 95 °C with a 1 mm path-length cuvette. Protein samples were prepared in PBS buffer (pH 7.4) at a concentration of 0.3-0.4 mg/mL. GdnCl solutions were prepared by dissolving GdnCl salt into PBS buffer and checking the refractive index.

Size-exclusion chromatography coupled to multiple-angle light scattering (SEC-MALS)- To ascertain the oligomerisation state of dIG proteins, SEC-MALS was performed in a Dawn Helios II apparatus (Wyatt Technologies) coupled to a SEC Superdex 75 Increase 10/300 column. The column was equilibrated with PBS or buffer B at 25 oC and operated at a flow rate of 0.5 mL/min. A total volume of 100-165 µL of protein solution at 1.0-3.0 mg/mL was employed for each sample. Data processing and analysis proceeded with *Astra 7* software (Wyatt Technologies), for which a typical dn/dc value for proteins (0.185 mL/g) was assumed.

Protein production for crystallization studies- The original thrombin site of plasmids pET28-dIG8-CC and pET28-dIG14 was replaced with a Tobacco-Etch-Virus peptidase (TEV) recognition site via *NcoI* and *Nde* employing forward and reverse primers (Eurofins). The generated plasmids, pET28*-dIG8-CC and pET28*-dIG14, were mixed at 100 mg each in Takara buffer (50 mM Tris-HCl, 10 mM magnesium chloride, 1 mM dithiothreitol, 100 mM sodium chloride, pH 7.5), annealed by slowly cooling down the sample to room temperature following 4 minutes at 94 °C, and ligated into the doubly digested plasmid. For pET28*-dIG14, the original thrombin-cleavable N-terminal His6-tag was removed and four histidine residues were added to the protein C-terminus by PCR using *NcoI* and *XhoI* sites. Of note, due to the cloning strategy, dIG18-CC and dIG-14 proteins were preceded by a G–H–M and a M–G motif, respectively. All PCR reactions and ligations were performed using Phusion High Fidelity DNA polymerase and T4 Ligase, and ligation products were transformed into chemically competent *E. coli* DH5-α cells for multiplication (all Thermo Fisher Scientific). Plasmids were purified with the E.Z.N.A. Plasmid Mini Kit I (Omega Bio-Tek) and verified by sequencing (Eurofins and Macrogen).

For protein expression, competent *E. coli* BL21 (DE3) cells (Sigma) were transformed with the pET28*-dIG8-CC and pET28*-dIG14 plasmids and grown on LB plates supplemented with 100 µg/mL kanamycin. Single colonies were selected to inoculate 5-mL starter cultures of this medium and incubated overnight at 37 °C under shaking. Respective 1-mL aliquots were used to inoculate 500 mL of the same medium. Once cultures reached $OD_{600}\approx0.6$, protein expression was induced with 0.5 mM IPTG (Fisher Bioreagents), and cultures were

incubated overnight at 18°C. Cells were harvested by centrifugation (3,500×*g,* 30 min, 4 °C) and resuspended in cold buffer A (50 mM Tris·HCl, 250 mM sodium chloride, pH 7.5), supplemented with 10 mM imidazole, EDTA-free cOmplete Protease Inhibitor Cocktail (Roche Life Sciences), and DNase I (Roche Life Sciences). Cells were lysed using a cell disrupter (Constant Systems) operated at 135 MPa, and soluble protein was clarified by centrifugation (50,000×*g*, 1 h, 4°C) and subsequently passed through a 0.22-µm filter (Merck Millipore).

For immobilised-metal affinity chromatography (IMAC (*57*)), proteins were captured on nickel-sepharose HisTrap HP columns (Cytiva), which had previously been washed and pre- equilibrated with buffer A plus either 500 mM or 20 mM imidazole, respectively. Column-bound dIG14 was extensively washed with a gradient of 20-to-150 mM imidazole in buffer A and eluted with a gradient of 200-to-300 mM imidazole in buffer A. Column-bound dIG8-CC was washed and eluted with buffer A containing 20 mM and 300 mM imidazole, respectively.

Fractions containing the dIG8-CC protein were then buffer-exchanged to buffer B (20mM Tris·HCl, 150 mM sodium chloride, pH 7.5) in a HiPrep 26/10 desalting column (GE Healthcare), and incubated overnight at 4°C with inhouse-produced His6-tagged TEV peptidase at a peptidase:substrate ratio of 1:20 (w/w) and 1mM dithiothreitol for fusion-tag removal. After centrifugation (50,000×*g,* 1 h, 4°C) and filtration (0.22-µm), the clarified dIG8-CC protein was loaded again onto the HisTrap HP column for reverse IMAC with buffer A plus 20 mM imidazole, which retained tagged protein and TEV, and had untagged dIG8-CC in the flow-through. The bound proteins were eventually eluted with buffer A plus 300 mM imidazole for column regeneration.

Untagged dIG8-CC and dIG14 were polished by size-exclusion chromatography (SEC) with buffer B in a Superdex 75 Increase 10/300 GL column (Cytiva) attached to an ÄKTA Purifier 10 apparatus. Protein purity was assessed by 20% SDS-PAGE stained with Coomassie Brilliant Blue (Sigma). PageRule Unstained Broad Range Protein Ladder and PageRuler Plus Prestained Protein Ladder (both Thermo Fisher Scientific) were used as molecular-mass markers. To concentrate protein samples, ultrafiltration was performed using Vivaspin 15 and Vivaspin 2 Hydrosart devices (Sartorius Stedim Biotech) of 2-kDa molecular-mass cutoff. Protein concentrations were determined either by the BCA Protein Assay Kit (Thermo Fisher Scientific) with bovine serum albumin as a standard or by A280 using a BioDrop Duo+ apparatus (Biochrom). Supplementary Fig. 16 provides proof of the effective protein purification procedures.

Protein crystallization- Crystallization screenings using the sitting-drop vapor diffusion method were performed at the joint IRB/IBMB Automated Crystallography Platform (www.ibmb.csic.es/en/facilities/automated-crystallographic-platform) at Barcelona Science Park (Catalonia, Spain). Screening solutions were prepared and dispensed into the reservoir wells of 96×2-well MRC crystallization plates (Innovadyne Technologies) by a Freedom EVO robot (Tecan). These reservoir solutions were employed to pipet crystallization nanodrops of 100 nL each of reservoir and protein solution into the shallow crystallization wells of the plates, which were subsequently incubated in steady-temperature crystal farms (Bruker) at 4ºC or 20ºC.

After refinement of initial hit conditions, suitable dIG14 crystals appeared at 20oC in drops consisting of 0.5 µL protein solution (at 1.9 mg/mL in buffer B) and 0.5 µL reservoir solution (0.1 M sodium acetate, 0.2 M calcium chloride, 20% w/v polyethylene glycol [PEG] 1500, pH 5.5). Crystals were cryoprotected with reservoir solution supplemented with 20% glycerol, harvested using 0.1–0.2 mm nylon loops (Hampton), and flash-vitrified in liquid nitrogen. The best tetragonal dIG8-CC crystals were obtained at 20 oC in drops containing 0.5 µL protein solution (at 30 mg/mL in buffer B) and 0.5 µL reservoir solution (0.1 M Bis-Tris, 0.2 M calcium chloride, 20% w/v PEG 3350, 10% v/v ethylene glycol, pH 6.5). Crystals were directly harvested using 0.1–0.2 mm loops, and flash-vitrified in liquid nitrogen. Proper orthorhombic dIG8-CC crystals resulted from the same condition as the tetragonal ones except that magnesium chloride and glycerol replaced calcium chloride and ethylene glycol, respectively. Furthermore, 0.25 mL of 5% n-

dodecyl-N,N-dimethylamine-N-oxide (w/v) was included as an additive. These crystals were cryoprotected with reservoir solution supplemented with 20% glycerol, harvested with elliptical 0.02–0.2 mm LithoLoops (Molecular Dimensions), and flash-vitrified in liquid nitrogen.

Diffraction data collection and structure solution- X-ray diffraction data were recorded at 100 K on a Pilatus 6M pixel detector (Dectris) at the XALOC beamline (*58*) of the ALBA synchrotron (Cerdanyola, Catalonia, Spain) and on an EIGER X 4M detector (Dectris) at the ID30A-3 beamline (*59*) of the ESRF synchrotron (Grenoble, France). Diffraction data were processed with programs *Xds* (*60*) and *Xscale*, and transformed with *Xdsconv* to MTZ-format for the *Phenix* (*61*) and *CCP4* (*62*) suites of programs. Analysis of the data with *Xtriage* (*63*) within *Phenix* and *Pointless* (*64*) within *CCP4* confirmed the respective space groups and indicated absence of twinning and translational non-crystallographic symmetry. Supplementary Table S5 provides essential statistics on data collection and processing.
The structure of dIG8-CC, both in its tetragonal ($P4_12_12$; 2.30 Å) and orthorhombic ($C222_1$; 2.05 Å) space groups, was solved by molecular replacement with the *Phaser* (*65*) program employing the coordinates of the designed structure. The tetragonal crystals contained four protomers (chains A–D) in the asymmetric unit (a.u.) arranged as two dimers, and the calculations gave final refined values of the translation function Z-score (TFZ) and log-likelihood gain (LLG) of 14.5 and 307, respectively. Subsequently, the adequately rotated and translated molecules were subjected to successive rounds of manual model building with the *Coot* program (*66*) alternating with crystallographic refinement with the *Refine* protocol of *Phenix* (*67*), which included translation/libration/screw-motion (TLS) refinement and non-crystallographic symmetry (NCS) restraints. The final model included residues R^{1}–G^{70} of each protomer preceded by M^{0}, H^{-1}, and, in chain D only, G^{-2} from the upstream linker, as well as 22 solvent molecules. The orthorhombic crystals were solved as the tetragonal ones with final refined TFZ and LLG values of 11.9 and 263, respectively. Model building and refinement proceeded as above. The final model encompassed residues R^{1}–G^{70} of each protomer preceded by M^{0} and H^{-1}, plus one magnesium cation and 34 solvent molecules. Cysteines C^{21} and C^{60} were present in both disulfide-linked and unbound conformations in all protomers of both crystal forms. The occupancy of the disulfide bond in the two crystal structures ranges between 0.00 and 0.67 across the eight protomers, with a mean occupancy of 0.47 and 0.41 in each of the structures (Supplementary Table 6).

The structure of dIG14 in a yet different space group ($P4_32_12$; 2.50 Å) with two molecules per a.u. was likewise solved by molecular replacement, with final refined TFZ and LLG values amounting to 17.4 and 269, respectively. The phases derived from the adequately rotated and translated molecules were subjected to a density modification and automatic model building step under twofold averaging with the *Autobuild* routine (*68*) of *Phenix*, which produced a Fourier map that assisted model building as aforementioned. Crystallographic refinement was also performed as above except that both *Phenix* and the *BUSTER* package (*69*) were employed. The final model comprised R^{1}–G^{68} of protomer A and R^{1}–F^{74} of protomer B, either preceded by G^{0} and M^{-1} from the upstream linker, as well as 15 solvent molecules.
Supplementary Table 4 provides essential statistics on the final refined models, which were validated through the *wwPDB V alidation Service* at https://validate-rcsb-1.wwpdb.org/validservice and deposited with the PDB at www.pdb.org with accession codes: 7SKN (design: dIG8-CC; space group: $P4_12_12$), 7SKO (design: dIG8-CC; space group: $C222_1$), and 7SKP (design: dIG14; space group: $P4_32_12$).

Tb^{3+} binding luminescence measurements- To measure the Tb^{3+} luminescence of samples dIG8-CC and EF61_dIG8-CC (in buffer 20 mM Tris, 50 mM NaCl, pH 7.4), time-resolved luminescence emission spectra and intensities were measured on a Synergy H1 hybrid multi-mode reader (BioTek) in flat bottom, black polystyrene, 96-well half-area microplates (Corning 3694). A stock solution

of terbium(III) chloride ($TbCl_3$) (Sigma-Aldrich, 451304-1G) was prepared in the same protein buffer. Time-resolved luminescence intensities were measured using excitation wavelength λ_{ex} = 280 nm and emission wavelength λ_{em} = 544 nm with a delay of 300 µs, 1 ms collection time and 100 readings per data point. Time-resolved luminescence emission spectra between 520 nm and 570 nm was collected in 2 nm increments and smoothed with a Savitzky-Golay filter of order 3 (Fig. 5h). For Tb^{3+} titrations, samples were incubated for 3 hours and the collected time-resolved luminescence emission intensities at λ_{em} = 544 nm were normalized to obtain protein bound fractions, and the normalized data was fit to the equilibrium binding equation with a Hill coefficient of 1 using non-linear least squares regression (Fig. 5i; Supplementary Fig. 15a). Ca^{2+} binding was measured by titrating $CaCl_2$ prepared in the same protein sample buffer into 20 µM EF61_dIG8-CC and 100 µM Tb^{3+}, and measuring the decrease of time-resolved luminescence emission intensity at λ_{em} = 544 nm (Supplementary Fig. 15b).

Protein expression of isotopically labeled proteins for NMR- Plasmids were transformed into BL21 (DE3) expression strain of *E. coli* (Invitrogen) and grown in 50 mL of Luria Broth containing 50 µg/mL of kanamycin and grown at 37°C with shaking overnight. After approximately 18 hours, the 50 mL starter culture was used to inoculate 500mL of minimal labeling media (M9), containing N15 labeled Ammonium Chloride at 50 mM and C13 glucose to 0.25% (w/v), as well as trace metals, 25 mM Na_2HPO_4, 25 mM KH_2PO_4, and 5 mM Na_2SO_4. The culture was returned to 37°C, at 250 rpm and allowed to reach OD_{600} ~0.7- 1.0. To induce expression 1mM of IPTG was added and the temperature was reduced to 25°C to allow the culture to express overnight. Cells were harvested by centrifugation at 4000 rpm for 20 minutes then resuspended with 40 mL of Lysis Buffer (20 mM Tris 250 mM NaCl 0.25% Chaps pH 8) and lysed with a Microfluidics M110P Microfluidizer at 18,000 psi. The lysed cells were clarified using centrifugation at 24,000 ×*g* for 30 minutes. The labeled protein in the soluble fraction was purified using Immobilized Metal Affinity Chromatography (IMAC) using standard methods (QIagen Ni-NTA resin). The purified protein was then concentrated to 2 mL and purified by FPLC size-exclusion chromatography using a Superdex 75 10/300 GL (GE Healthcare) column into 20 mM $NaPO_4$ 150 mM NaCl pH 7.5. The efficiency of labeling was confirmed using mass spectrometry.

Nuclear magnetic resonance spectroscopy- NMR data were acquired at 30 °C on Bruker spectrometers operating at 600 or 800 MHz, equipped with cryogenic probes. His-tagged double-labeled (^{15}N, ^{13}C) dIG21 and ^{15}N-labeled dIG14 constructs were dissolved in PBS buffer (pH 7.5, 150 mM NaCl) at concentrations of ~ 150-200 µM. For dIG21, triple-resonance backbone spectra, and a 3D NH-NOESY spectrum, were acquired with non-uniform sampling schemes in the indirect dimensions and were reconstructed by the multi-dimensional decomposition software qMDD (*70*), interfaced with NMRPipe (*71*), as described previously (*72*). The spectra were analyzed using SPARKY (*73*), and the automated in-house program FMCGUI/ABACUS (*74*) was used to aid the assignment of backbone resonances.

Data availability

Coordinates and structure factors have been deposited in the Research Collaboratory for Structural Bioinformatics Protein Data Bank with the accession codes 7SKN (dIG8-CC, tetragonal space group), 7SKO (dIG8-CC, orthorhombic space group) and 7SKP (dIG14). All the designed protein structures experimentally tested are available as Supplementary Dataset 1, and their corresponding sequences are provided in Supplementary Table 2. Further structural analyses (for loops, cross-b motifs and Ig designs), biochemical and biophysical characterization of the designs, structure prediction calculations, sequence analysis and X-ray crystallography statistics are provided as Supplementary Figures and Tables. Other data are available from the corresponding authors upon request.

Code availability

The Rosetta macromolecular modelling suite (http://www.rosettacommons.org) is freely available to academic and non-commercial users. Computational protocols used for analyzing and designing protein structures are available at https://github.com/emarcos/immunoglobulin_design.

Methods references

42. W. Kabsch, C. Sander, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 22, 2577–2637 (1983).

43. A. Andreeva, E. Kulesha, J. Gough, A. G. Murzin, The SCOP database in 2020: expanded classification of representative family and superfamily domains of known protein structures. Nucleic Acids Research. 48, D376–D382 (2020).

44. S. J. Fleishman *et al.*, RosettaScripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite. PLoS ONE. 6, e20161 (2011).

45. G. Bhardwaj *et al.*, Accurate de novo design of hyperstable constrained peptides. Nature. 538, 329–335 (2016).

46. R. F. Alford *et al.*, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. J. Chem. Theory Comput. 13, 3031–3048 (2017).

47. W. Sheffler, D. Baker, RosettaHoles2: A volumetric packing measure for protein structure refinement and validation: RosettaHoles2 for Protein Structure. Protein Science. 19, 1991–1995 (2010).

48. G. C. P. van Zundert *et al.*, The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. Journal of Molecular Biology. 428, 720–725 (2016).

Other data are available from the corresponding

49. M. Siedlecka *et al.*, Alpha-helix nucleation by a calcium-binding peptide loop. Proceedings of the National Academy of Sciences. 96, 903–908 (1999).

50. P.-S. Huang *et al.*, RosettaRemodel: A Generalized Framework for Flexible Backbone Protein Design. PLoS ONE. 6, e24109 (2011).

51. A. S. Ford, B. D. Weitzner, C. D. Bahl, Integration of the Rosetta suite with the python software stack via reproducible packaging and core programming interfaces for distributed simulation. Protein Science. 29, 43–51 (2020).

52. K. H. Le *et al.*, PyRosetta Jupyter Notebooks Teach Biomolecular Structure Prediction and Design. The Biophysicist. 2, 108–122 (2021).

53. M. Rocklin, (Austin, Texas, 2015; https://conference.scipy.org/proceedings/scipy2015/matthew_rocklin.html), pp. 126–132.

54. T. Brunette *et al.*, Modular repeat protein sculpting using rigid helical junctions. Proc Natl Acad Sci USA. 117, 8870–8875 (2020).

55. F. W. Studier, Protein production by auto-induction in high-density shaking cultures. Protein Expr Purif. 41, 207–234 (2005).

56. I. Anishchenko *et al.*, De novo protein design by deep network hallucination. Nature (2021), doi:10.1038/s41586-021-04184-w.

57. H. Block *et al.*, in Methods in Enzymology (Elsevier, 2009; https://linkinghub.elsevier.com/retrieve/pii/S0076687909630275), vol. 463, pp. 439–473.

58. J. Juanhuix *et al.*, Developments in optics and performance at BL13-XALOC, the macromolecular crystallography beamline at the Alba Synchrotron. J Synchrotron Rad. 21, 679– 689 (2014).

59. D. von Stetten *et al.*, ID30A-3 (MASSIF-3) – a beamline for macromolecular crystallography at the ESRF with a small intense beam. J Synchrotron Rad. 27, 844–851 (2020).

60. W. Kabsch, XDS. Acta Crystallogr D Biol Crystallogr. 66, 125–132 (2010).

61. P. D. Adams *et al.*, PHENIX: a comprehensive Python-based system for macromolecular

structure solution. Acta Crystallogr D Biol Crystallogr. 66, 213–221 (2010).

62. M. D. Winn *et al.*, Overview of the CCP4 suite and current developments. Acta Crystallogr D Biol Crystallogr. 67, 235–242 (2011).

63. P. H. Zwart, R. W. Grosse-Kunstleve, P. D. Adams, CCP4 Newsletter on Protein Crystallography Vol. 43 (Winter 2005) (ed F. Remacle) 27-35 (Daresbury Laboratory) (2005).

64. P. R. Evans, An introduction to data reduction: space-group determination, scaling and intensity statistics. Acta Crystallogr D Biol Crystallogr. 67, 282–292 (2011).

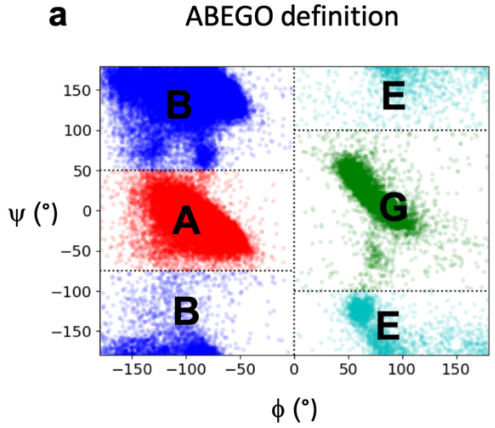
65. A. J. McCoy *et al.*, Phaser crystallographic software. J Appl Crystallogr. 40, 658–674 (2007).

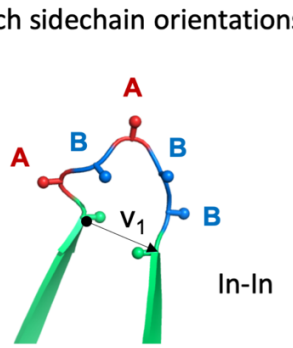
66. A. Casañal, B. Lohkamp, P. Emsley, Current developments in Coot for macromolecular model building of Electron Cryo-microscopy and Crystallographic Data. Protein Science. 29, 1055–1064 (2020).
67. D. Liebschner *et al.*, Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. Acta Crystallogr D Struct Biol. 75, 861–877 (2019).
68. T. C. Terwilliger *et al.*, Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. Acta Crystallogr D Biol Crystallogr. 64, 61– 69 (2008).
69. BUSTER version 2.10 (Global Phasing Ltd., Cambridge (UK) (2017).
70. K. Kazimierczuk, V. Yu. Orekhov, Accelerated NMR Spectroscopy by Using Compressed Sensing. Angew. Chem. Int. Ed. 50, 5556–5559 (2011).
71. F. Delaglio *et al.*, NMRPipe: A multidimensional spectral processing system based on UNIX pipes. J Biomol NMR. 6 (1995), doi:10.1007/BF00197809.
72. A. Lemak *et al.*, A novel strategy for NMR resonance assignment and protein structure determination. J Biomol NMR. 49, 27–38 (2011).
73. T. D. Goddard, D. G. Kneller, Sparky 3. University of California, San Francisco.
74. A. Lemak, C. A. Steren, C. H. Arrowsmith, M. Llinás, Sequence specific resonance assignment via Multicanonical Monte Carlo search using an ABACUS approach. J Biomol NMR. 41, 29–41 (2008).

Supplementary Information

De Novo Design of Immunoglobulin-like Domains

Tamuka M. Chidyausiku[1,2,7,†‡], Soraia R. Mendes[3†], Jason C. Klima1[.2'§], Marta Nadal[4], Ulrich Eckhard[3], Jorge Roel-Touris[4], Scott Houliston[5,6], Tibisay Guevara[3], Hugh K. Haddox[2], Adam Moyer[2], Cheryl H. Arrowsmith[5,6], F. Xavier Gomis-Rüth[3*], David Baker[1,2,7*], Enrique Marcos4*





c

In-In

ABEGO	Frequency
ABB	118
BBG	81
ABABB	79
BBGBB	54
ABBG	24

In-Out

ABEGO	Frequency
BBGB	216
BAAB	29
ABA	25
ABBBA	24
BAABA	18

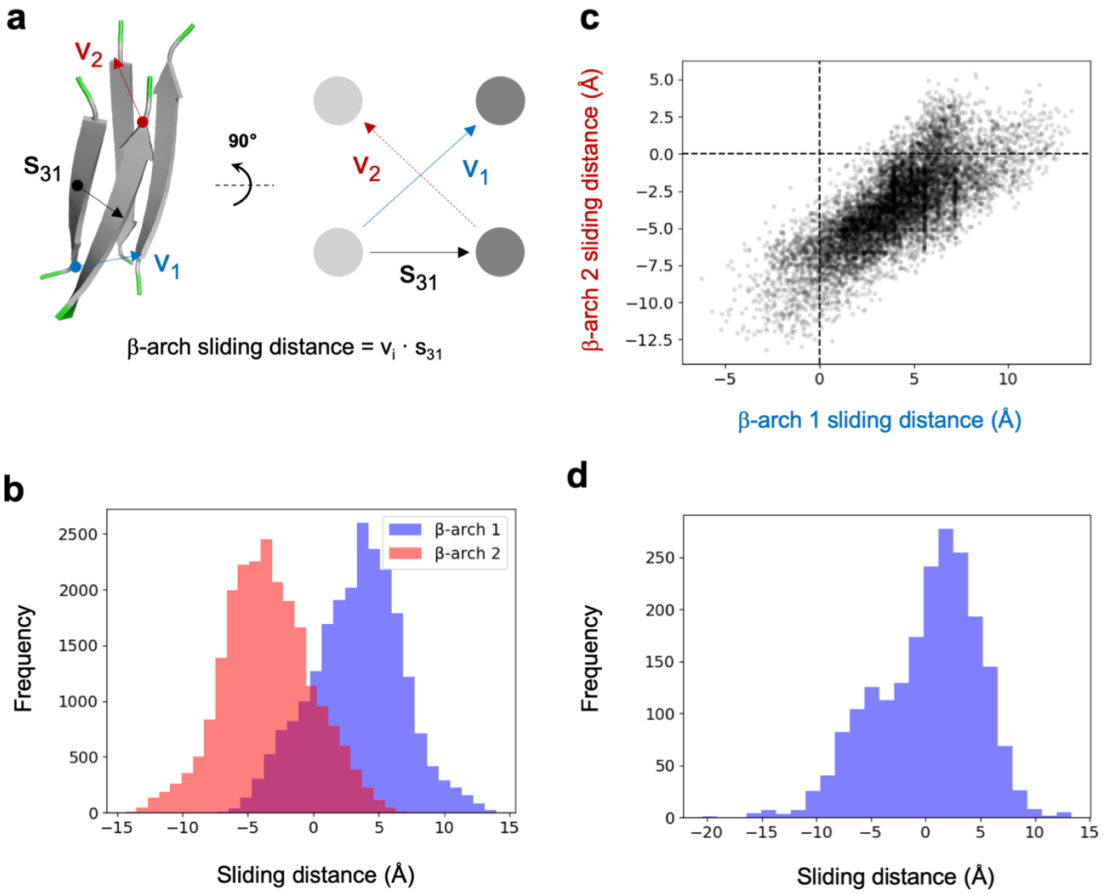
Out-Out

ABEGO	Frequency
BGB	68
AAB	34
BBBGB	33
AAABA	25
BABB	22

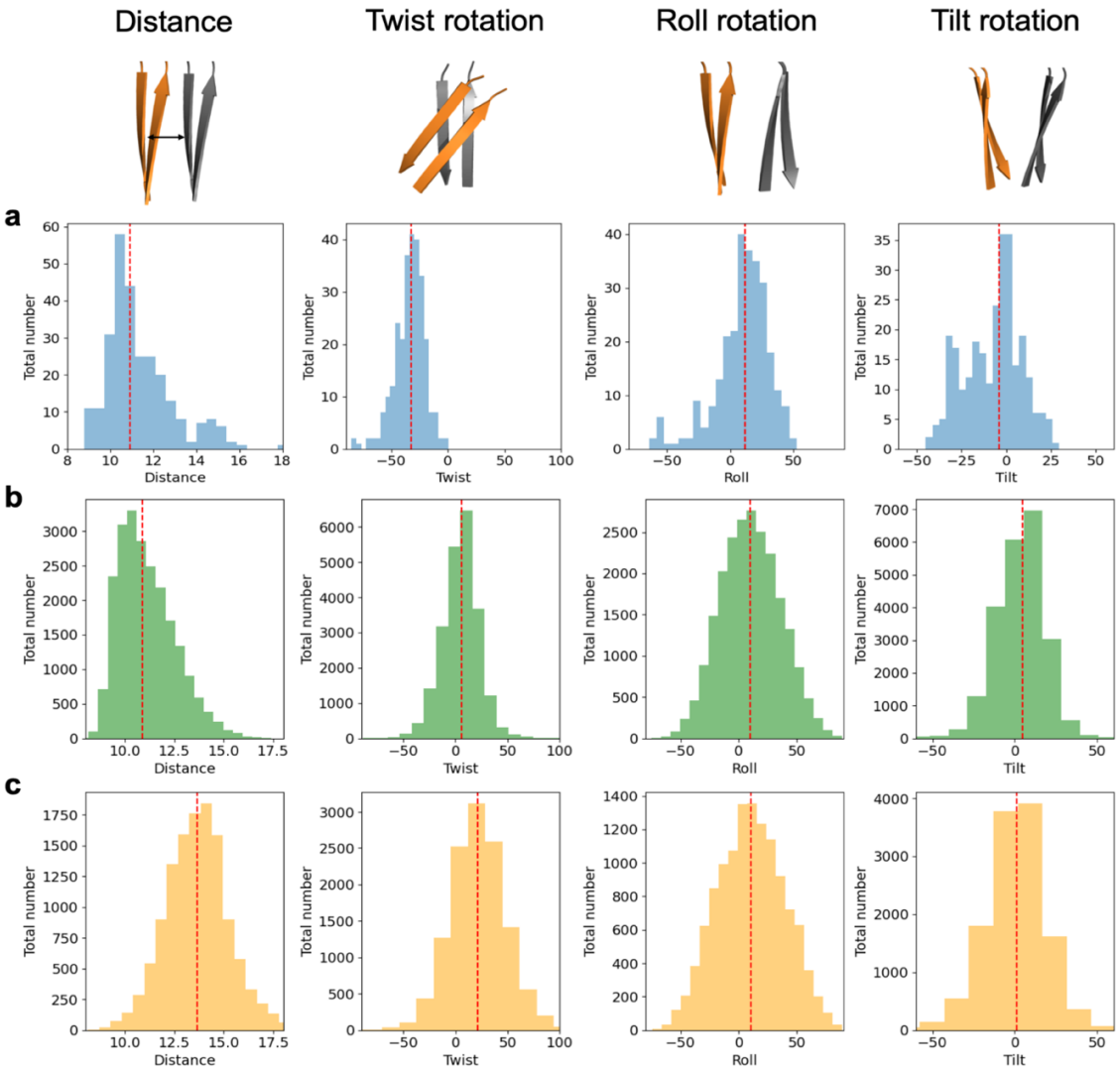
Out-In

ABEGO	Frequency
BABB	44
BGBB	34
AAA	22
GBB	21
BBAA	14

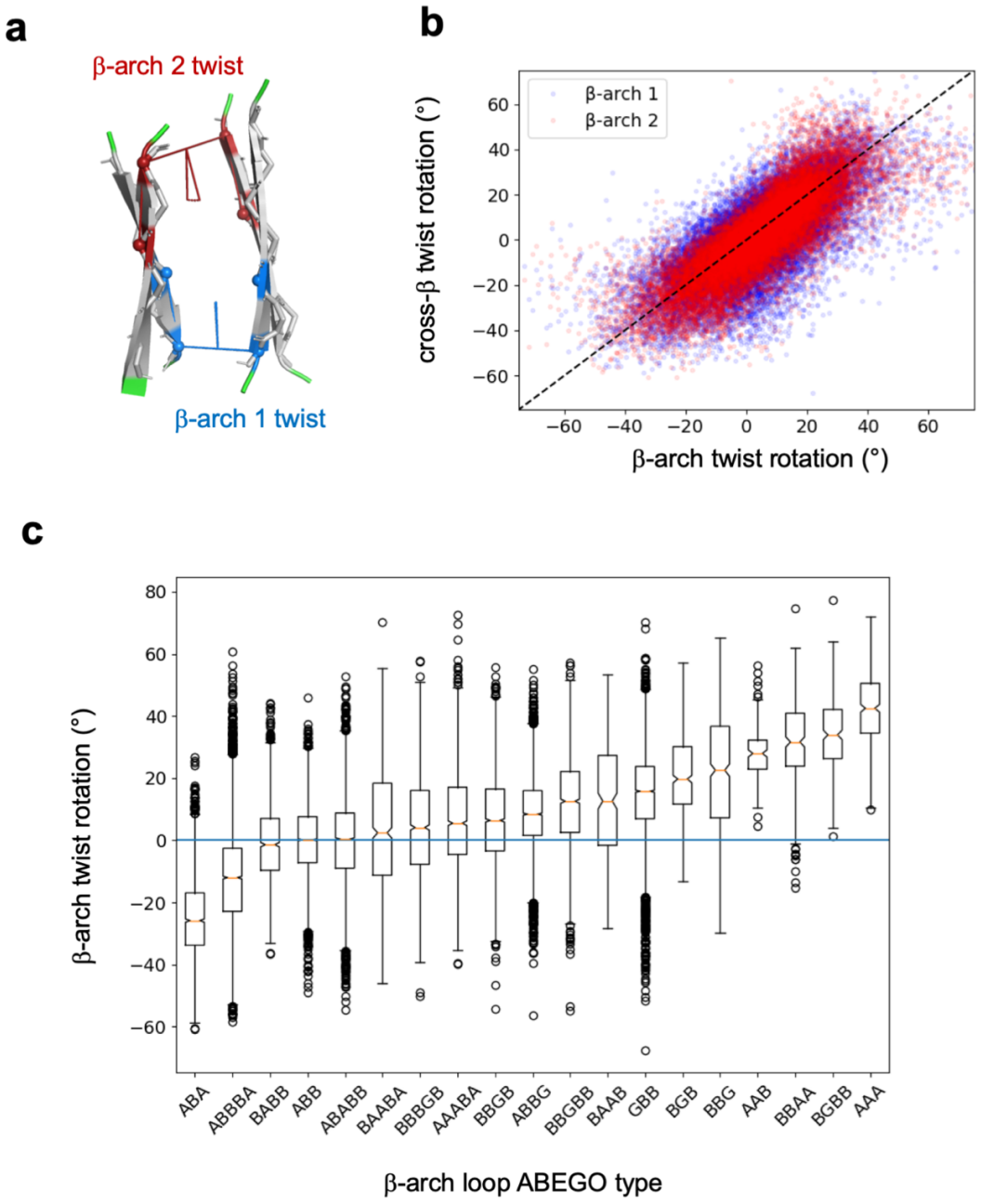
Supplementary Fig. 1. Frequently observed β-arch loops in naturally occurring protein structures. a, The Ramachandran plot is conveniently discretized in ABEGO torsion bins describing local backbone geometry at the residue level ("A", right-handed α-helix region (*red*); "B", extended region (*blue*); "E", extended region with positive f (*cyan*); "G", left-handed α-helix region (*green*); and "O", if the peptide bond dihedral angle (w) deviates from planarity). **b**, Definition of the β-arch sidechain orientation based on the relative orientation between the translation vector (v_1) and the C_α-C_β vector of the two adjacent β-strand residues. If the C_α-C_β vector of the preceding residue is oriented in the same direction as v_1, then the sidechain orientation is considered to point inwards ("In"), otherwise i tis considered to point outwards ("Out"). The same applies to the residue following the loop but considering -v_1 as the translation vector. Loop positions are colored according to their ABEGO bin, as shown in (**a**). **c**, β-arch loops (ranging between 3 and 5 residues) spanning the four possible sidechain orientations that are most frequently observed in a non-redundant set of naturally occurring protein structures.



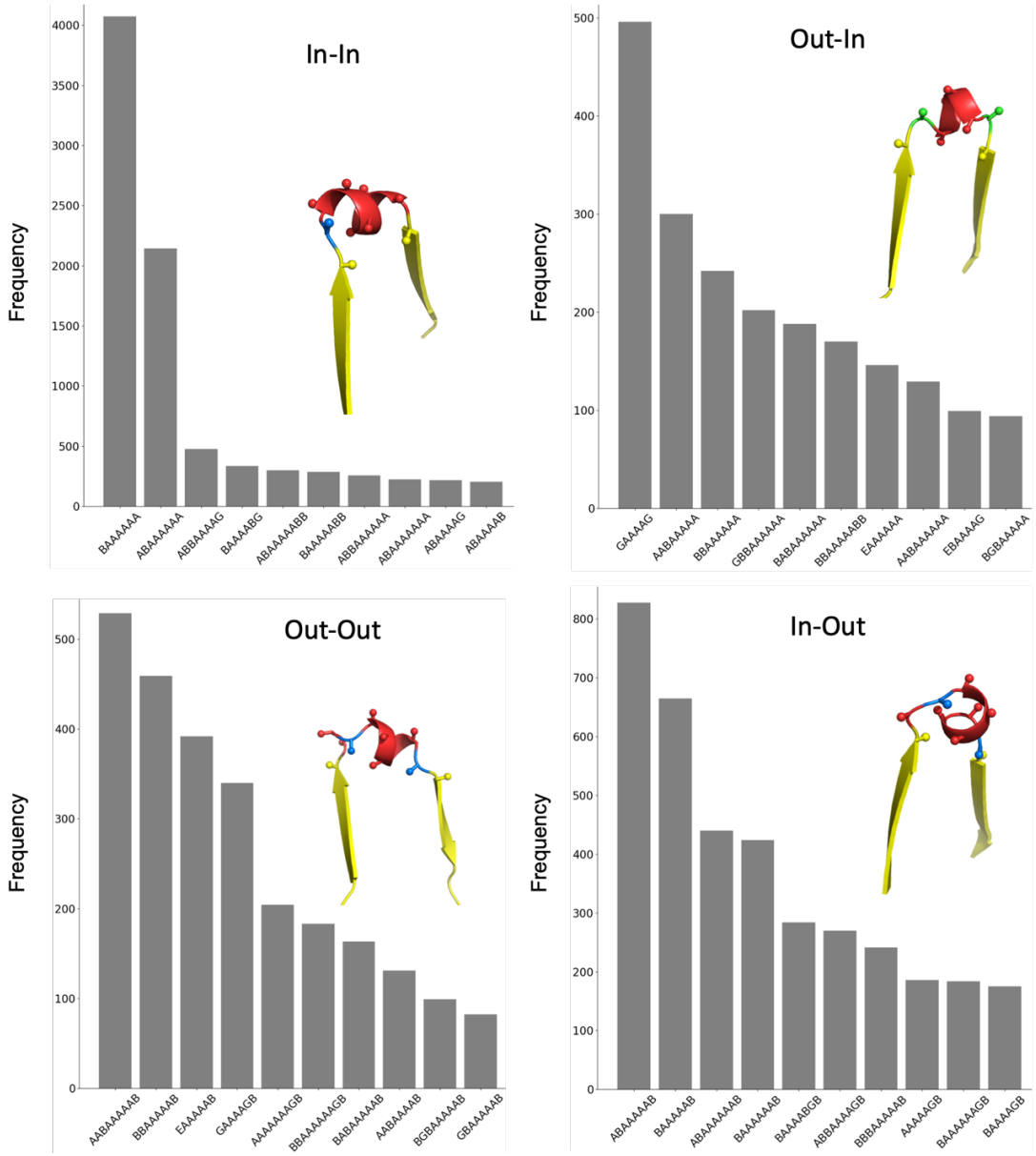
Supplementary Fig. 2. Coupling between the two β-arches forming cross-β motifs. a, β-arch sliding distance definition. Cartoon representation *(left)* and diagram *(right)* of a cross-β motif. We define v_1 and v_2 as the translation vectors connecting the C_α atoms of the residues preceding and following β-arch loops 1 and 2, respectively; and the S_{31} vector between the centers of the two N-terminal β-strands (1 and 3). The sliding distance is the projection of the β-arch translation vectors onto the S_{31} vector. **b,** Distribution of β- arch sliding distances in cross-β motifs generated by Rosetta folding simulations. In general, cross-β motifs tend to have positive and negative sliding distances for β-arches 1 and 2. **c**, Correlation between the two β- arch sliding distances in simulated cross-β motifs with low twist rotations (between -10° and 10°). **d**, Distribution of β-arch sliding distances in β-arch loops from naturally occurring protein structures.



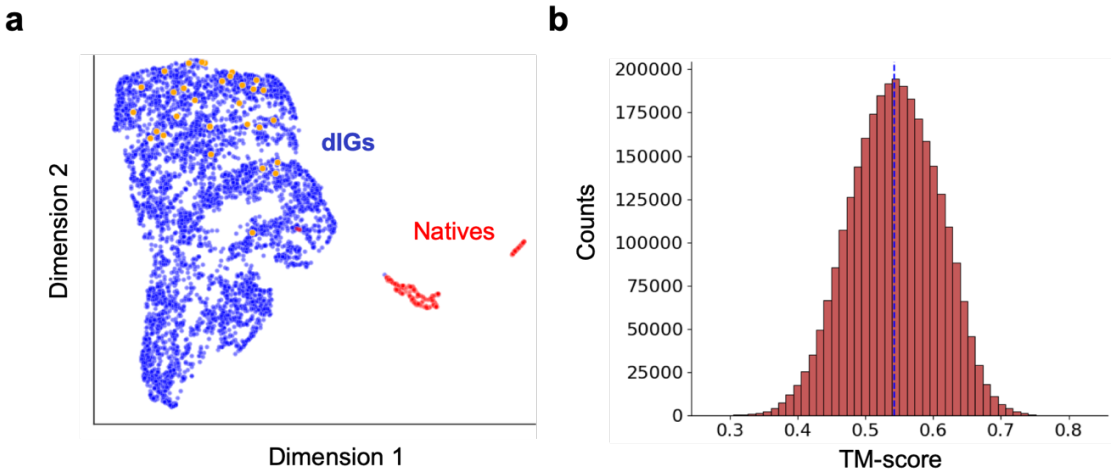
Supplementary Fig. 3. Distributions of cross-β geometrical parameters obtained from naturally occurring Ig domains and Rosetta folding simulations. a, Median (*red dotted line*) and median absolute deviations for each parameter: distance (10.9 ± 0.8 Å), twist (-32.1 ± 7.7°), roll (12.0 ± 12.2°) and tilt (-4.0 ± 11.1°). Distributions correspond to a set of 275 natural Ig domains with sequence identity below 40%. **b**, Median (*red dotted line*) and median absolute deviations for each parameter: distance (10.9 ± 1.0 Å), twist (5.7 ± 11.0°), roll (9.7 ± 18.0°) and tilt (4.5 ± 9.6°). Distributions correspond to 22,507 cross-β motif models generated by Rosetta folding simulations exploring different combinations of strand lengths (5-7 residues) and frequently observed β-arch loops (3-5 residues). **c**, Median (*red dotted line*) and median absolute deviations for each parameter: distance (13.7 ± 1.0 Å), twist (21.1 ± 17.2°), roll (10.3 ± 20.4°) and tilt (1.2 ± 11.2°). Distributions correspond to 12,335 cross-β motif models generated by Rosetta fragment assembly simulations exploring different combinations of strand lengths (5-7 residues) and β-arch helices (3-5 residues).



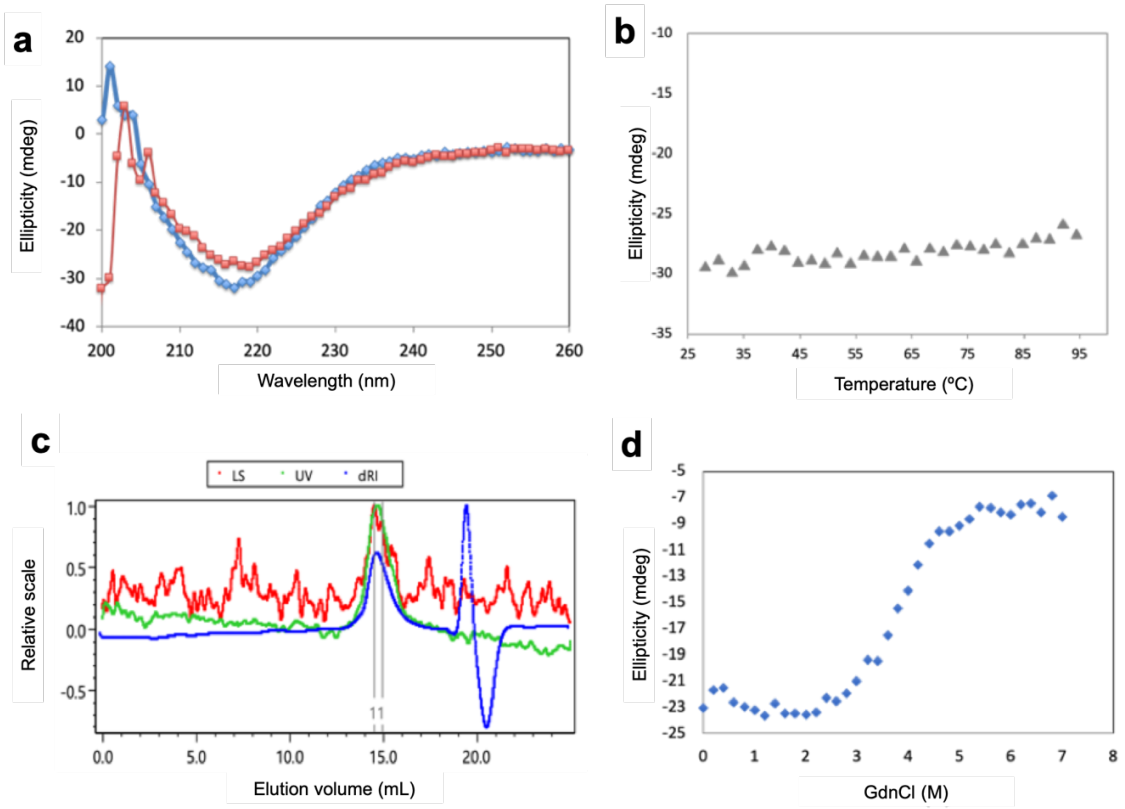
Supplementary Fig. 4. β-arch loops can twist cross-β motifs in different directions depending on their geometry. **a**, Definition of β-arch twist based on the dihedral angle formed between the a-carbons $C_\alpha(i\text{-}2)$, $C_\alpha(i)$, $C_\alpha(j)$ and $C_\alpha(j{+}2)$; where *i* and *j* correspond to the residues preceding and following the β-arch loop. **b**, Correlation between β-arch loop twisting and the cross-β twist rotation obtained from Rosetta folding simulations. **c**, Distributions of β-arch twist values for loops with frequently observed ABEGO torsion bins forming cross-β motifs in Rosetta folding simulations, sampling both positive and negative rotations.



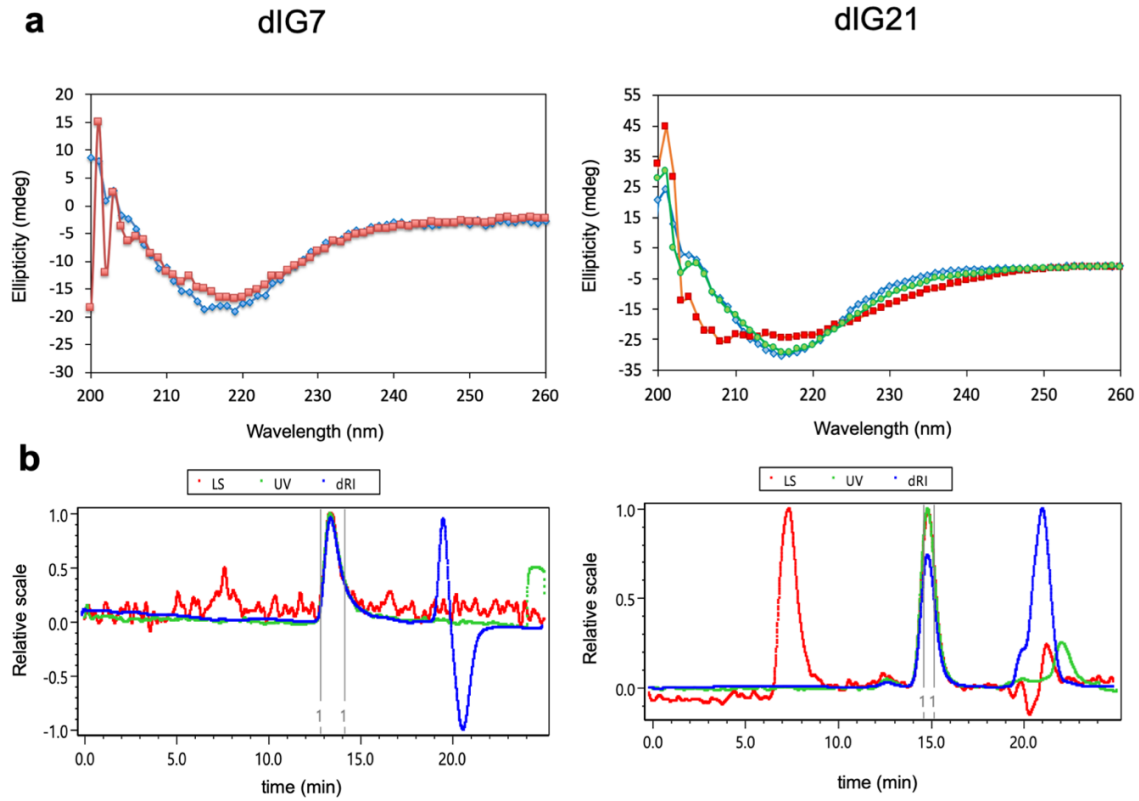
Supplementary Fig. 5. β-arch helices favoring cross-β motifs obtained from Rosetta folding simulations. The 10 most frequently observed loop-helix-loop ABEGO patterns of each possible sidechain orientation are shown on the horizontal axes. Frequencies are calculated as the total number of counts across β-arches from all generated cross-β motifs by Rosetta folding simulations with a sequence-independent model. Cross-β motif examples for the most frequently observed ABEGO pattern of each sidechain orientation is shown and color-coded as in Supplementary Fig. 1a, with the preceding and following β- strands in yellow. Most ABEGO patterns have a "B" torsion in the residue preceding the helix, which is typically observed at the start of α-helices as it provides N-terminal hydrogen bond capping.



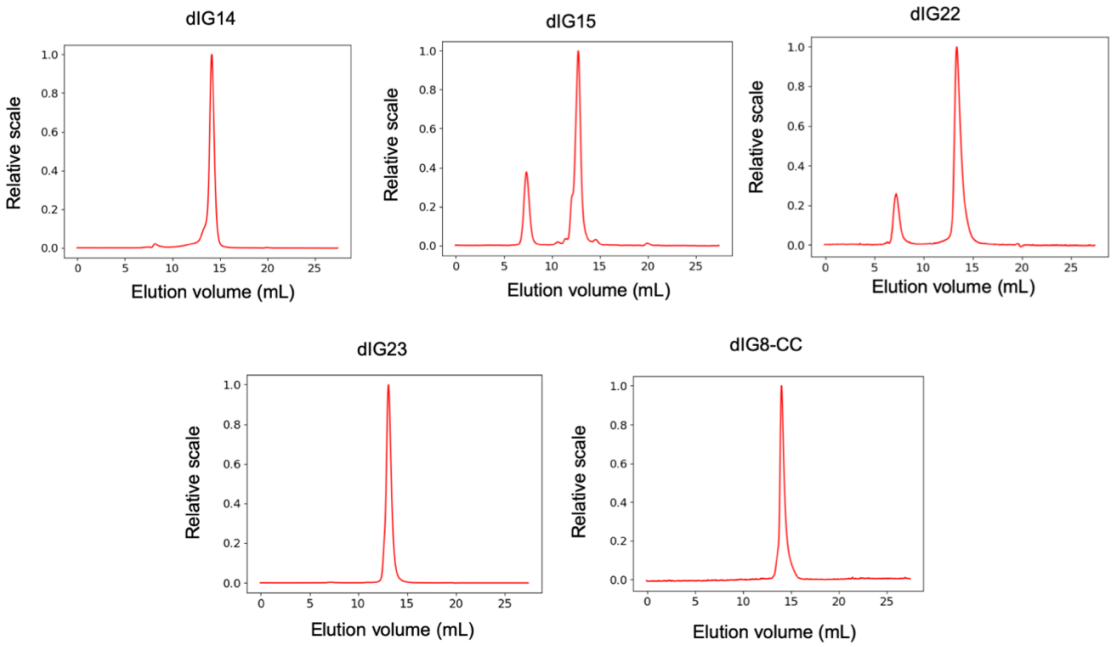
Supplementary Fig. 6. Structural diversity of the designed proteins and their comparison to natural Ig-like domains. a, Uniform manifold approximation and projection (UMAP) analysis of computationally designed *(blue)* and naturally occurring *(red)* Ig domains based on pairwise distances calculated as the TM- score. Experimentally tested designs are shown in orange. The designs broadly sample a structural space distinct from natural Ig proteins. **b**, Distribution of TM-scores between designs and natural Ig domains (mean 0.54 ± 0.06 s.d. *(dashed blue line)*).



Supplementary Fig. 7. Biochemical characterization of the dIG8 design. a, Far-ultraviolet circular dichroism spectra (blue: 25 °C; red: 95 °C). **b**, Thermal denaturation monitored at 220 nm wavelength by circular dichroism. The design denatures at temperatures above 95 °C. **c**, SEC-MALS analysis showing light scattering (LS) *(red)*, ultraviolet (UV) *(green)*, and differential refractive index (dRI) *(blue)* signals. The protein is monodispersed and has an estimated molecular weight of 16.6 kDa, which lies between that corresponding to the theoretical monomer (10.3 kDa) and dimer (20.6 kDa). The protein includes the thrombin cleavage site and the hexa-histidine purification tag, which adds 2.3 kDa to the design. **d**, Chemical denaturation with guanidine hydrochloride (GdnCl) monitored at 220 nm wavelength by circular dichroism. The cooperative unfolding transition indicates that the protein is well-folded. All experiments were carried out in PBS buffer.

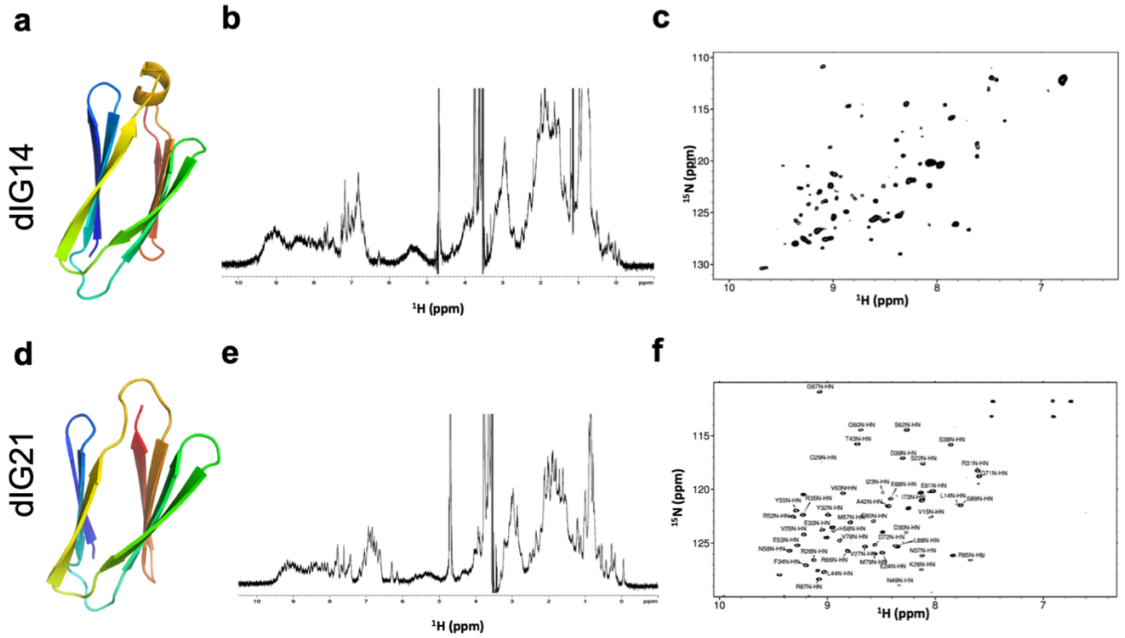


Supplementary Fig. 8. Biochemical characterization of the dIG7 and dIG21 designs. a, Far-ultraviolet circular dichroism spectra (blue: 25 °C; green, 75 °C; red: 95 °C). **b**, SEC-MALS analysis showing light scattering (LS) *(red)*, ultraviolet (UV) *(green)*, and differential refractive index (dRI) *(blue)* signals. dIG7 and dIG21 are monodispersed and have estimated molecular weights of 23.4 and 10.1 kDa, which correspond to dimer and monomer respectively. All experiments were carried out in PBS buffer.

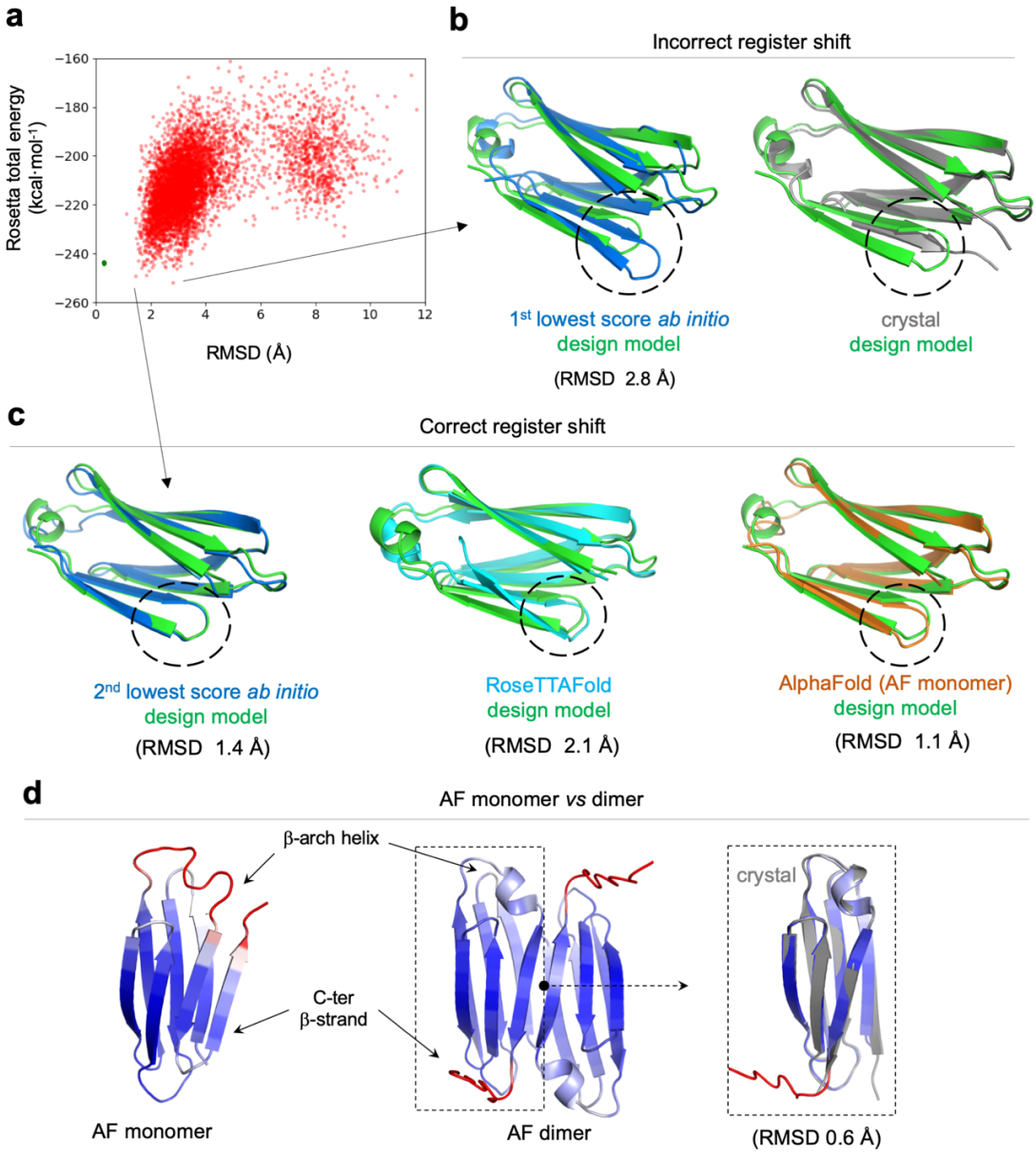


Design name	Theoretical M_w (kDa)	Estimated Mw (kDa)
dIG14	9.7	20.0±0.1
dIG15	8.8	17.5±0.3
dIG22	10.5	20.5±0.1
dIG23	10.6	20.1±0.1
dIG8-CC	8.3	12.6±0.3

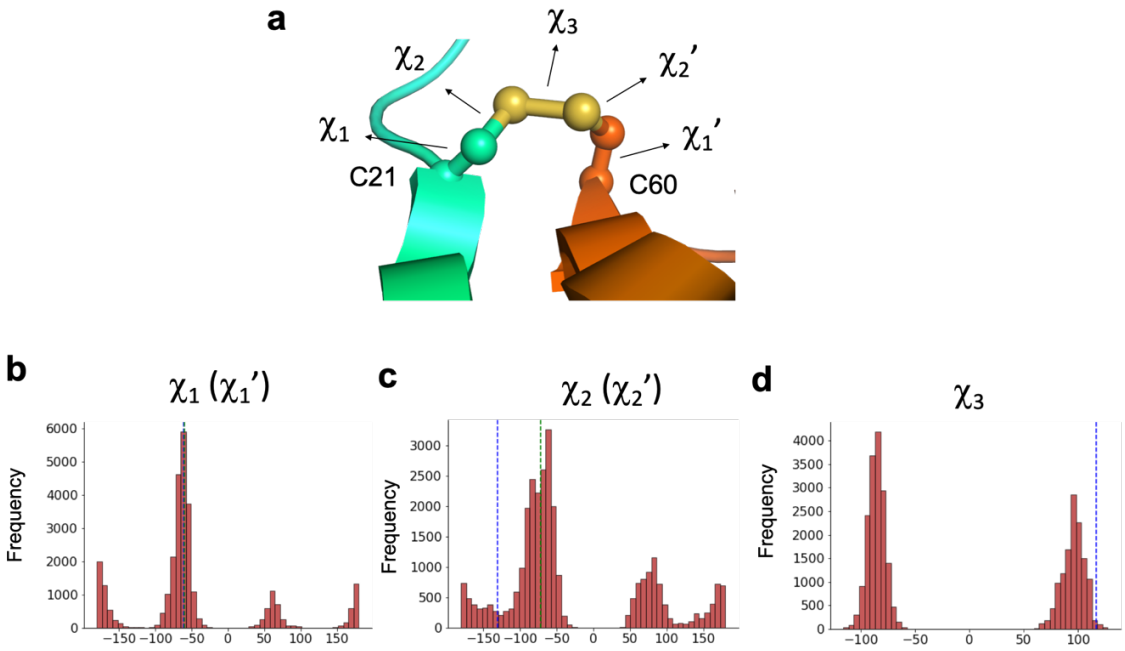
Supplementary Fig. 9. Size-exclusion chromatography combined with multi-angle scattering data. dIG8-CC has a predicted molecular weight (M_w) that corresponds to between the monomer and dimer molecular weights, suggesting an equilibrium between both states. dIG14 and other representative designs are predicted to be dimers in solution. Samples were prepared in 20 mM Tris·HCl, 150 mM sodium chloride, pH 7.5.



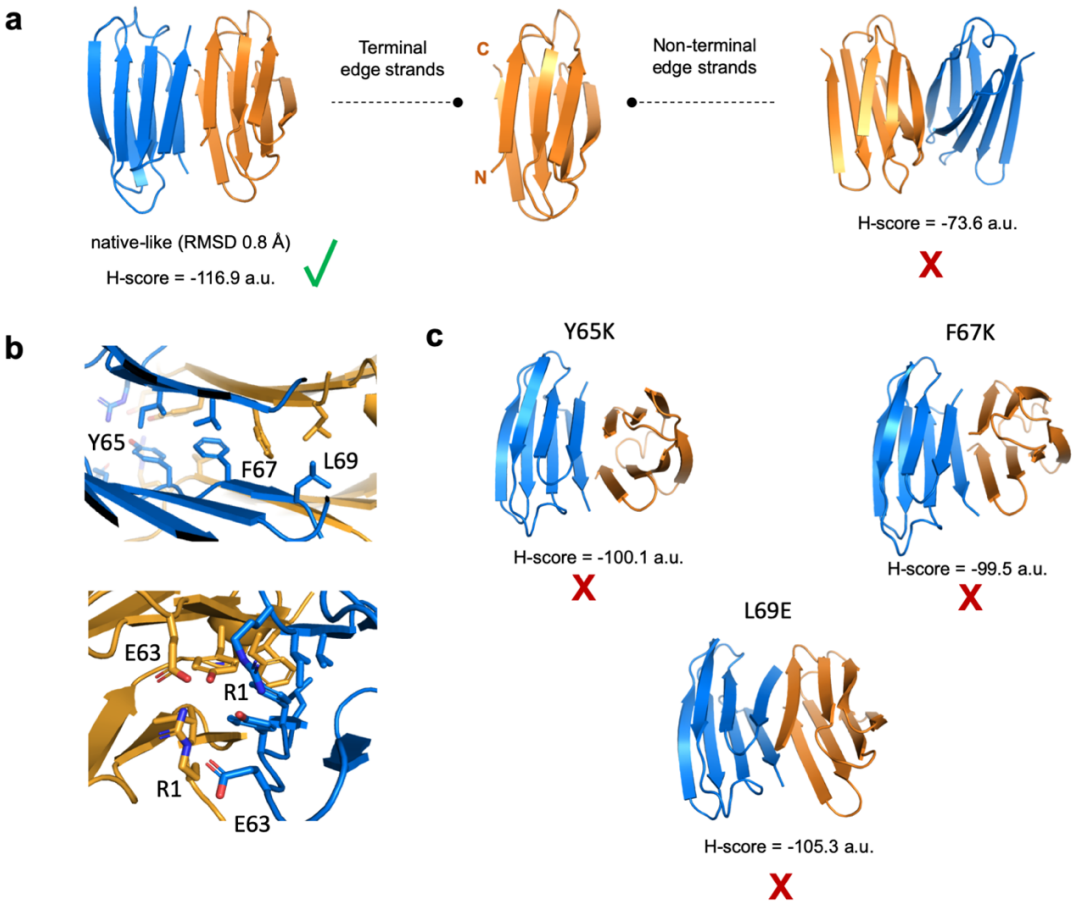
Supplementary Fig. 10. Nuclear magnetic resonance characterization of designs dIG14 and dIG21. a, Cartoon representation of the dIG14 design model. **b**, ^{1}H-1D spectrum showing dispersed amide (~7.5 to 9.5 ppm) and aliphatic (shifted upfield to ~ -0.2 ppm) resonances, indicative of a well-folded construct. **c**, ^{1}H-^{15}N HSQC spectrum showing considerable dispersion, consistent with the presence of significant β- extended secondary structure. **d**, Cartoon representation of the dIG21 design model. **e**, ^{1}H-1D spectrum showing dispersed amide (~ 7.5 to 9.5 ppm) and aliphatic (shifted upfield to ~ -0.1 ppm) resonances, indicative of a well-folded construct. **f**, ^{1}H-^{15}N HSQC spectrum (which are labeled where assigned) show considerable dispersion, consistent with the presence of significant β-extended secondary structure. We assigned ~ 70% of the backbone NH, C_{α}, C_{β}, CO, H_{α} and H_{β} resonances, and these assignments are in agreement with the designed secondary structure (based on predicted φ/ψ dihedral angles using TALOS+ (*1*) chemical shift analysis).



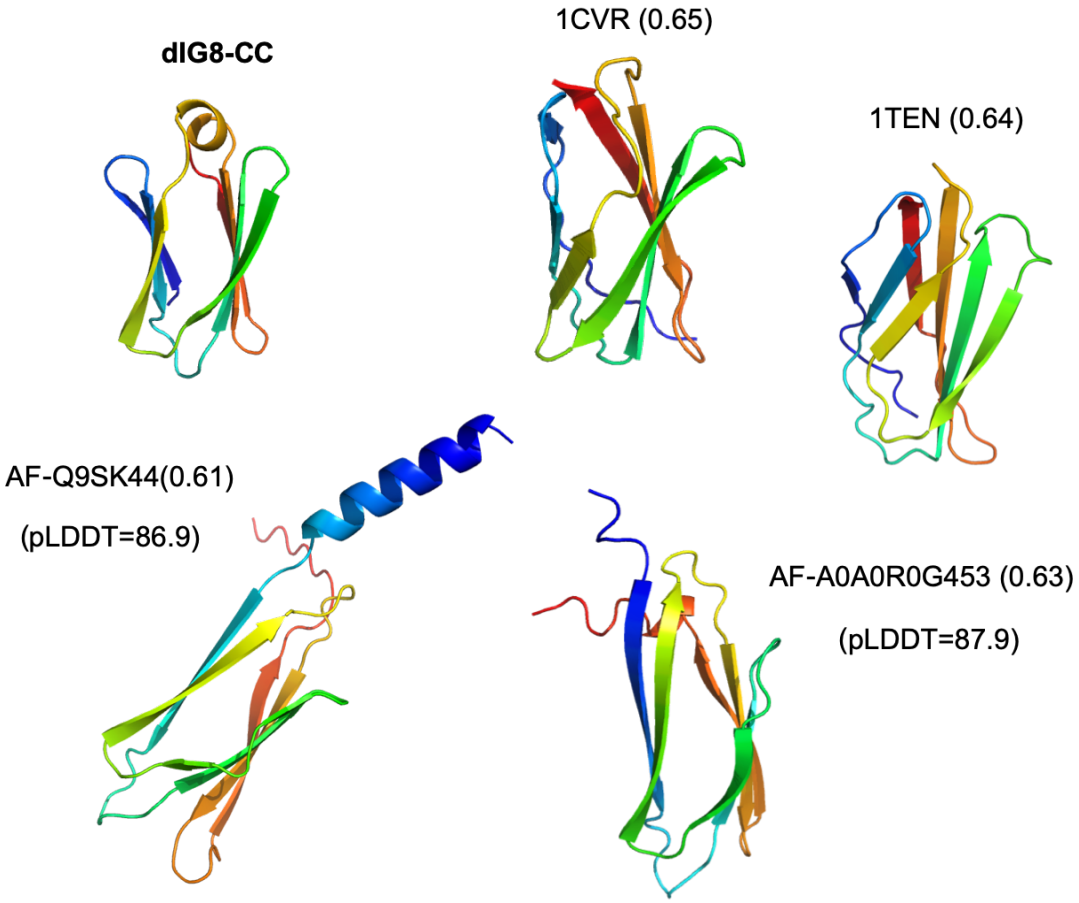
Supplementary Fig. 11. Structures predicted for design dIG14 by Rosetta *ab initio* folding simulations, RoseTTAFold and AlphaFold. a, b, Rosetta *ab initio* folding simulations revealed that the pairing between β-strands 3 and 6 has two conformational states very close in energy, one as designed and the other as observed in the crystal structure. **c**, RoseTTAFold and AlphaFold predict the register shift and the C-terminal strand as designed, which disagrees with the experimental structure. None of the methods predict the C-terminal strand flip out as observed in the crystal, but all predict a conformational rearrangement of the designed β-arch helix. **d**, (*left*) Top AlphaFold monomer prediction colored by pLDDT (from red to blue increasing in pLDDT) highlights a sequence-structure mismatch in the β-arch helix area. (*center*) Top AlphaFold dimer prediction, with monomer subunits having high pLDDT across all residues (except for the C-terminal strand residues) and matching closely the crystal structure monomer subunits (*right*). The predicted interface differs from the crystal structure (Fig. 4b) by a rotation of 180o between the two monomer subunits.



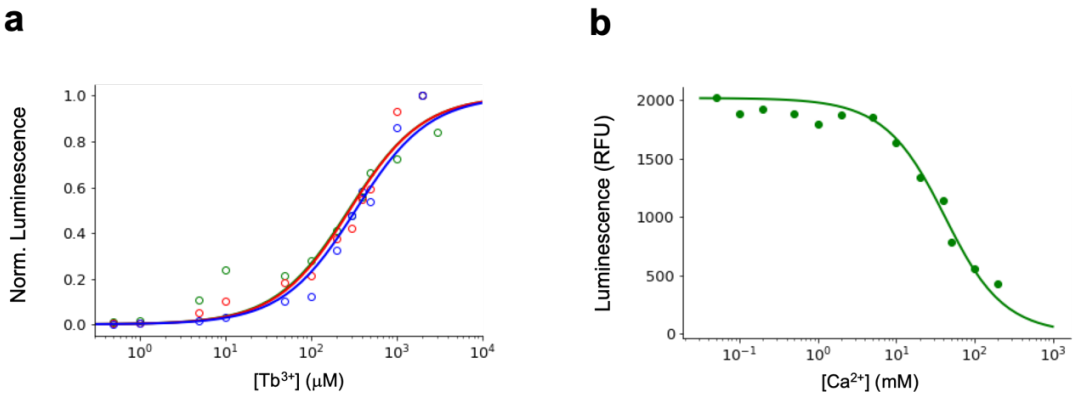
Supplementary Fig. 12. Dihedrals of the designed dIG8-CC disulfide bond in comparison with natural distributions. a, Five dihedrals describing the geometry of the designed disulfide bond (*spheres* and *sticks*) between C21 and C60. **b**, Distribution of c_1 (and c_1') dihedral angles obtained from a database of ~30,000 native disulfide bond geometries that was used for design (see Methods). The corresponding dihedral angles of the dIG8-CC design are represented as dashed vertical lines (c_1 = -60.3° in *blue* and c_1' = -59.9° in *green*). **c**, Distribution of c_2 (and c_2') dihedral angles obtained from the database of native disulfide bond geometries. The corresponding dihedral angles of the dIG8-CC design are represented as dashed vertical lines (c_2 = -130.6° in *blue* and c_2' = -72.0° in *green*). **d**, Distribution of the c_3 dihedral angle obtained from the database of native disulfide bond geometries. The corresponding dihedral angle of the dIG8-CC design is represented as dashed vertical lines (c_3 = 117.1° in *blue*). Two of the five disulfide dihedral angles (c_2 and c_3) are not frequently observed in distributions from naturally occurring disulfides, which is likely associated with the low disulfide bond stability suggested by the crystal structures.



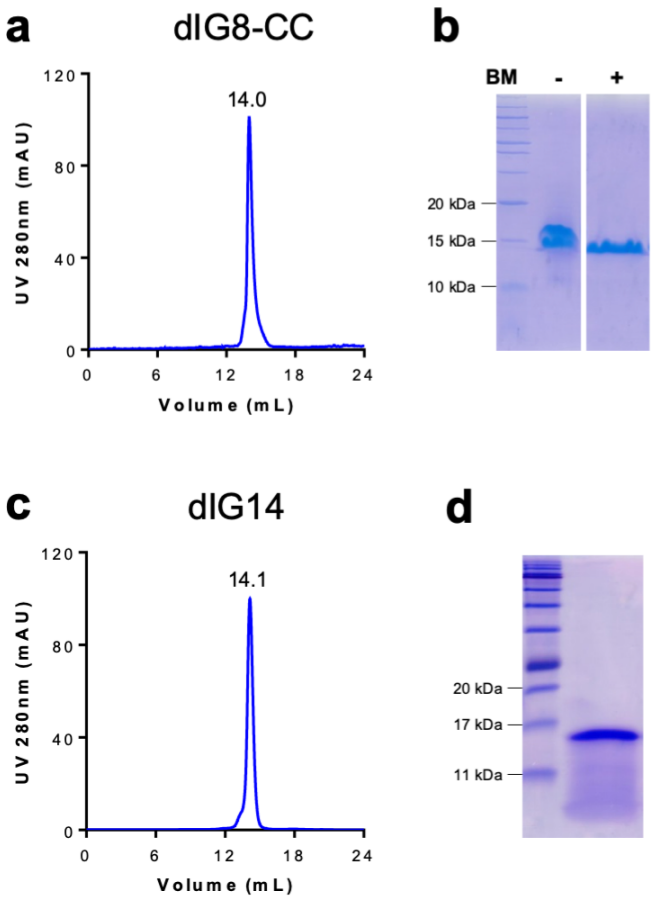
Supplementary Fig. 13. Docking calculations on the dIG8-CC homodimer interface. a, Docking calculations of two dIG8-CC monomers using ambiguous restraints between terminal edge strands recapitulate the parallel interface observed in the crystal structure *(left)*. Docking restrained toward the opposite edge predicts dimer orientations with disrupted edge-to-edge strand pairing and worse docking scores *(right)*; overall supporting that the terminal edge strands are more dimerization-prone. **b**, The crystal dimer interface is formed primarily by hydrophobic *(top)* and salt bridge *(bottom)* interactions. **c**, Docking calculations for single-point mutants replacing interface hydrophobics by lysine or glutamate, both of which are known to efficiently disrupt edge-to-edge interfaces as inward-pointing charged residues. All mutants are effective in disrupting the native interface. Some mutants flip the dimer orientation to form antiparallel interfaces with diminished backbone hydrogen-bonded strand pairing and overall higher docking scores. The lowest-score decoy for the most populated cluster of each simulations is shown. Docking scores (H- score), calculated with the HADDOCK docking software, are provided in arbitrary units (a.u.).



Supplementary Fig. 14. Naturally occurring protein structures most similar to design dIG8-CC found across the PDB and the AlphaFold Protein Structure Database. The closest structural analog with experimental structure available (PDB ID: 1CVR) was identified by a TM-align search over a curated dataset of immunoglobulin-like domains as identified by SCOP (those under the Ig β-sandwich fold classification and with X-ray structure resolution < 2.5 Å). Closest structural analogues in the AlphaFold Protein Structure Database with confident predictions (pLDDT > 85) are also shown. Normalized TM- scores are indicated in parentheses.



Supplementary Fig. 15. Terbium and calcium concentration-dependent luminescence of EF61_dIG8- CC. a, Normalized time-resolved luminescence intensity for Tb^{3+} titrations with EF61_dIG8-CC at three different concentrations (5 μM, *green*; 10 μM, *red*; 20 μM, *blue*). The three protein concentrations, each below the Tb^{3+} K_d (Fig. 5i), result in nearly identical normalized binding curves. **b**, Time-resolved luminescence intensity in relative fluorescence units (RFU) for Ca^{2+} titrations with 20 μM EF61_dIG8-CC and 100 μM Tb^{3+}, showing Ca^{2+} competition with Tb^{3+} for the EF61_dIG8-CC Tb^{3+} binding site. **a**, **b**, For each protein concentration, luminescence intensities are fit to a one-site binding model by non-linear least squares regression *(lines)*. The excitation wavelength used was λ_{ex} = 280 nm and the emission wavelength used was λ_{em} = 544 nm.



Supplementary Fig. 16. Protein purification of dIG8-CC and dIG14 for crystallization studies. Representative final size-exclusion chromatograms of dIG8-CC (**a**) and dIG14 (**c**). Retention volumes in mL are indicated above the respective peak. Subsequent SDS-PAGE analysis of dIG8-CC (**b**) and dIG14 (**d**) after concentration, with (dIG8-CC) or without (dIG8-CC and dIG14) β-mercaptoethanol (BM).

Supplementary Table 1. Deep-learning-based structure prediction of the designed proteins. Three metrics from the highest-confidence AlphaFold and RoseTTAFold predicted models are reported: RMSD to the design model, mean pLDDT over all residues and the minimum pLDDT value.

	AlphaFold			RoseTTAFold		
Design	**RMSD (Å)**	**pLDDT**	**min(pLDDT)**	**RMSD (Å)**	**pLDD**	**min(pLDDT)**
dIG1	0.9	92.6	73.6	0.9	90.3	81.4
dIG2	8.9	76.3	58.7	2.2	84.1	76.3
dIG3	0.6	89.2	74.4	1.1	83.9	68.9
dIG4	0.7	92.6	69.7	0.9	86.4	61.0
dIG5	0.8	87.8	74.1	1.7	86.0	74.7
dIG6	1.1	91.0	70.5	1.3	87.8	76.0
dIG7	1.3	88.7	71.6	1.4	85.6	70.5
dIG8	1.0	90.3	77.8	0.9	86.6	66.0
dIG9	2.1	88.6	74.0	1.2	87.9	77.0
dIG10	1.2	88.6	72.3	2.3	82.0	69.5
dIG11	0.9	85.3	60.2	1.2	86.4	65.7
dIG12	3.1	84.0	52.5	1.3	88.1	77.0
dIG13	1.4	85.1	66.3	2.4	83.8	75.4
dIG14	1.2	83.6	64.8	2.0	84.0	63.1
dIG15	0.9	84.6	63.7	1.3	87.7	77.6
dIG16	1.2	90.8	73.6	2.4	78.4	55.4
dIG17	1.7	89.3	63.4	1.3	87.1	74.3
dIG18	1.6	86.0	67.7	1.8	86.2	71.0
dIG19	0.9	94.7	81.4	1.5	86.8	71.9
dIG20	2.1	83.2	61.3	1.8	79.2	49.0
dIG21	1.0	91.5	84.7	2.1	72.6	51.6
dIG22	0.8	92.6	80.8	1.2	83.5	62.8
dIG23	1.3	92.5	82.2	1.9	79.4	58.9
dIG24	0.7	88.2	76.4	2.0	86.3	69.4
dIG25	1.3	85.2	67.6	2.3	85.0	69.8
dIG26	0.9	85.7	65.7	2.1	83.8	58.1
dIG27	1.2	85.2	52.7	1.5	86.5	76.4
dIG28	1.1	89.3	72.4	2.0	87.4	69.4
dIG29	0.7	92.7	73.7	2.3	78.0	68.3
dIG30	1.2	84.1	67.8	1.9	85.4	74.0
dIG31	1.0	84.7	69.5	1.4	86.2	71.2

Suplementary Table 2. Designed protein sequences in comparison with naturally occurring ones.

The lowest E-values obtained from BLAST (against the NCBI nr database of non-redundant protein sequences), and more sensitive sequence-profile searches with HHBlits (against the UniRef30 database) and HHPred are reported. The PDB ID of the lowest E-value hit identified with HHPred is also shown in parentheses.

Design	Amino acid sequence	Blast	HHBlits	HHpred
dIG1	TVEVRIRKNGNEYEVEVENRSDRPAEVRFHYDGTTETYTVP PGTRLRYRTKLTKPMRIEVRAGNTTYEYTVS	0.1	0.073	0.057 (2r39)
dIG2	EIHVELRKEGDRVEVRVENRSSQPGTVEIEVDGQRYEFTANP GERIQFEARGKTPVRVEVVYGNTTYRYEVR	3.8	0.19	0.056 (2r39)
dIG3	RVRVEVKNNKIEVENNSDQPAEIHLEFGGRRFTYTGNKGERI EVQISPEEAKNARIEIKVGDKKLEYQYH	3.5	2.9	4.1 (4ktp)
dIG4	RVEVRISGNTIRVENRSDRPARVEFEYGGRREEYTAPPGSEL RVTISPEELKNARVEIEYGGQRYRFEVT	0.73	1.6	6.7 (1r0u)
dIG5	KIRIEVRSSGNTIHVEVENNSDRPVRIRVTAPGTTLETTANPG ERVRFEFRGVPPGGEVEVEVKAGDEKVRTRYRS	0.7	0.067	0.54 (2x3c)
dIG6	TVEVRITEKNGQWEVRIRNRSSQPARVEVEEGGRREEYTLN PGDELELHFTSPKPVRITVEVGGQRYTYTLR	1.2	0.5	0.92 (4xin)
dIG7	RMEVRVSNGRVEIENKSSQPGRVEVRFNGKRYEYTANPGER VEVEVSPEELKNLRVRLEYDGKTEETQYS	2.1	0.056	0.47 (6w0p)
dIG8	RIEVRVDNGRVRVRNGTDRPVRVRVTAGGETREYTVNPGT ELEVELSPEQQNNAEVEVEVGNEKYRFQLG	3.8	0.47	3.0 (6w0p)
dIG9	SIRVEIEKRGDSYRVEVENRSDQPAEIEVRWNGRRERYEAN KGETVEVEVRAPSPVEVRVRAGNTEVRVEQR	1.0	0.47	0.46 (2r39)
dIG10	RVEVRISGNTIEIRSEGPGRLELEYNGQREEYTLNPGTRIEFE GRPGEEVRVEVEMNGQRYTFEVRF	1.6	0.32	0.44 (6w0p)
dIG11	RLEVRMEGKKVEVRNNSDRPMRVEFTWNGQRERYHVNPG ETLEVEVQPGARVEVRVQSGDTQYRYEFEL	2.1	0.061	0.089 (6e5c)
dIG12	SLEVRVRKSGNTFEVEIRNKSDRPAEVRLEIGGRRETYTVPP GSTLRLRGPGKRPGRVEIKAGDAKYEVELR	0.62	0.11	0.012 (2r39)
dIG13	YVEIRYKGEKVHIRTNGPVTLEVEFEGKRERYTLNPGEELEI RIRARRIRVEVQEGDRKIETELTF	0.31	0.2	0.41 (6e5c)
dIG14	RVEVRVEFEGDKMRVRLRNDSSTPVEVHIKVGDEKRTVTV NPGEEVEVTFSANDPHKFNRPQFTIEWGGQRQHFQHH	0.72	0.0028	0.26 (4ay0)
dIG15	RPKVQLELHGNKMRVRLRNDSSTPVEVHIKVGDEKRTVTV NPGEEVEVTFSTTDPRELKNATIQLHQGDQTVEYRVD	0.2	0.0031	0.63 (2r39)
dIG16	EVEIEVRTKNGKIEVRVTNRSDRPVEVRMEKGGQRETYTAP PGSTVRVEFSPSDDRQKRPTVEVTVNGRRYEVRVH	2.5	0.22	0.044 (5ngl)
dIG17	RVEFRLREEGDRYRLEIRTDRPGTIEIEVNGRRERYTANPGT TITVEGTRGEEVEVTVEYDGKRERWRFRM	1.0	0.72	7.7 (6ex6)
dIG18	RVRWTWRISGNTIEFRFENNSDRPARVEIEVDGQRREYTVN PGERLELHFQAGAREIRVEVEVGKEKYEVRIRF	0.51	0.056	0.37 (2r39)
dIG19	RVEVRIREEGDKYELRIRNRSDRPAEVRIEKGGKRETYTVNP GEELRIEFPPGAPPGRVEVQVGDKKYEYTVK	1.8	0.065	0.39 (6i60)
dIG20	VVEVRLEGERIRVRNNSDRPATVHVEKDGQRETYTVNPGEE LEITSPDSSQNKGLRLRIHVEVNGQRFTFEFTM	0.51	0.024	3.6 (6w8u)
dIG21	SIEVRVKGDRYEFRNNSDKPATLEVEKNGKREEYHMNPGES VEVRGEPGQDIRFEMVMEGTTYRYRLS	0.61	0.044	1.7 (7agw)
dIG22	SIEVRVKGDRYEFRNNSDKPATLEVEKNGKREEYHMNPGES VEVRGPPGQDIRFEMTMDGTTYRYRLS	3.6	0.31	1.7 (7agw)
dIG23	DLEVRRKDGKFEFRNNSDKPATLEVEKDGQREEYRMNPGE TIEVQAPPGQDVRFTVEMPGREYRYKLD	1.0	0.021	0.16 (3q48)
dIG24	TFEVRVQWSGNTIRVTVENQSDRPATVRIEYGNTTYQRTINP GDRLTVEFTGGPGEVHVEVEINGKREERTFTK	4.5	0.032	0.35 (3sd2)

dIG25	EVQMRVEISGDTIRVEVRNNSDRPGRVEFEVGGVRTSYTMN PGERIEVEVTVSTAEKQGIKVEVHVEAGDEKRTYEFQM	2.3	0.99	0.83 (6fjy)
dIG26	RVEVRVQEKNGKVEIRVRSDGPVRVEVEVGGQRREY TGNPGEEVEIEVTADQPVRVEVKAGDKKFTYTVSE	2.1	0.36	15 (6w0p)
dIG27	MFRVEVREKNGRVEVRVENRSDRPGTVEVEVGGVRL RFTVNPGEELEIRMDVPNGRRVEIEIVGKGVKYSYEYT V	1.3	0.13	1.6 (2wnw)
dIG28	SWEVRVRWKNGRLEVEIRNNSSQPGKVRIEFDGKRHE VHLNPGESTKWRFENPGGEFH VEAGKEKYTYTV	2.5	0.017	1.3 (2wnw)
dIG29	RVEVRQSGNTIEIRSEGPGRLELEYNGQREEYTLNPGT RYEYEGRPGEEVRVEVEMNGQ RYTYEVRS	2.8	1.0	1.5 (5bvq)
dIG30	RSEVHVRFEGERIEIQIHNGTDKPARVEMEVNGQRYEY HMPPNSKMEYRVPLRQEIRFEVEVGGQRFTYRYTS	2.8	0.65	2.3 (1v7w)
dIG31	RVEVRVTYKGNRVEVRVRNNSDRPVRFRVVGPGAKY ELKGNPGTEMRVEIRVPNAREIEVEVNGQRQRYQM	4.1	0.94	1.8 (6ywf)
dIG8-CC	RIEVRVDNGRVRVRNGTDRPCRVRVTAGGETREYTV NPGTELEVELSPEQQNNAEVEVECGNEKYRFQLG	3.3	0.033	2.1 (6w0p)
EF61_dIG8-CC	RIEVRVDNGRVRVRNGTDRPCRVRVTAGGETREYTV NPGTELEVELSPEQQNNAEVEVECTVDDKDGDGYISA AEAA VEKYRFQLG	8.7	0.006	0.0018 (6ohh)

Supplementary Table 3. Cross-β geometrical parameters calculated for the designed proteins. For comparison, median and median absolute deviation values for cross-β parameters calculated from naturally occurring Ig domain structures are also provided, as shown in Supplementary Fig.3a.

Design	Distance (Å)	Twist (°)	Roll (°)	Tilt (°)
dIG1	11.5	-3.4	12.6	-6.4
dIG2	10.7	-9.2	14.6	6.8
dIG3	11.3	-12.4	20.5	27. 1
dIG4	11.0	-10.8	10.6	7.2
dIG5	10.2	-24.7	-5.7	-8.3
dIG6	11.2	5.5	4.3	-4.8
dIG7	10.4	-18.7	-7.4	-18.8
dIG8	10.0	-16.6	6.8	3.1
dIG9	10.3	-13.9	11.6	8.3
dIG10	11.0	-8.6	6.6	-2.7
dIG11	11.1	-17.1	5.2	8.1
dIG12	10.7	2.0	11.0	4.0
dIG13	11.5	-19.2	-4.7	-10.4
dIG14	11.2	-13.8	9.9	4.4
dIG15	11.0	-16.8	11.3	-5.1
dIG16	10.5	-14.5	-4.1	-12.5
dIG17	9.8	-2.6	-4.9	-0.3
dIG18	12.2	-1.8	-0.1	-2.0
dIG19	11.2	-5.7	19.2	1.8
dIG20	10.8	11.9	-9.8	-2.9
dIG21	10.9	-18.4	-17.6	-19.0
dIG22	10.7	4.5	9.9	-5.8
dIG23	10.8	0.3	3.2	-1.1
dIG24	10.3	-20.7	-8.8	-12.5
dIG25	10.0	-21.5	-0.7	-17.4
dIG26	12.2	0.6	20.0	16.1
dIG27	10.8	-15.7	6.1	10.7
dIG28	10.4	1.2	5.5	1.3
dIG29	11.0	-8.6	6.7	2.7
dIG30	10.8	-18.2	6.5	-7.1
dIG31	11.9	-7.9	13.5	26.5
Natural Ig domains	10.9±0.8	-32.1±7.7	12.0±12.2	4.0±11.1

Supplementary Table 4. Summary of the experimental characterization of designs.

dIG	Soluble expression	Monodisperse	CD* spetra (25ºC)	Tm‡ (ºC)	Oligomeric state†
1	No	-	-	-	-
2-6,9,11-13, 16-19,24-31	Yes	No	-	-	-
10,20	Yes	Yes	β	>95ºC	High
7,14,15,22,23	Yes	Yes	β	>95ºC	D
21	Yes	Yes	β	>75ºC	M
8,8-CC	Yes	Yes	β	>95ºC	M/D

* 'CD', circular dichroism. ‡ 'T_m', melting temperature. † Oligomeric state of the dominant species determined with size-exclusion chromatography with multi-angle light-scattering (SEC-MALS) ('M', monomer; 'D', dimer).