Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR MATHEMATIK, INFORMATIK
UND NATURWISSENSCHAFTEN

Comparative studies of protein dynamics in coronaviruses

Dissertation

with the aim of achieving a doctoral degree at the Faculty of Mathematics, Informatics and Natural Sciences

Department of Chemistry

University of Hamburg

submitted by

Kira Schamoni-Kast

born in Meschede

Hamburg, 2025

Thesis reviewers

Prof. Dr. Charlotte Uetrecht

Leibniz Institute of Virology, Hamburg, Germany; Centre for Structural Systems Biology, Hamburg, Germany; University of Lübeck, Lübeck, Germany; Deutsches Elektronen-Synchrotron, Hamburg, Germany.

Prof. Dr. Andrew Torda

Center for Bioinformatics, Head of Research Group Biomolecular Modeling, University of Hamburg, Germany

Examination board

Prof. Dr. Hartmut Schlüter

Institute of Clinical Chemistry and Laboratory Medicine, Section Mass Spectrometry and Proteomics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Prof. Dr. Andrew Torda

Center for Bioinformatics, Head of Research Group Biomolecular Modeling, University of Hamburg, Germany

Prof. Dr. Aymelt Itzen

Center for Experimental Medicine, Institute of Biochemistry and Signal Transduction, University Medical Center Hamburg-Eppendorf, Hamburg, Germany

Thesis submission

12.06.2025

Thesis defense

19.09.2025

This thesis was prepared from 2020 to 2025 under the supervision of Prof. Dr. Charlotte Uetrecht at the Leibniz Institute of Virology, the Centre for Structural Systems Biology, and the University of Lübeck in the working group *Dynamics of Viral Structures*. The second supervisor was Prof. Dr. Hartmut Schlüter, Institute of Clinical Chemistry and Laboratory Medicine, Section Mass Spectrometry and Proteomics, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.

Contents

Publications

Parts of this thesis have been previously published or are currently being prepared for publication. Some of the figures and tables were taken or modified from my own publications or submitted papers.

Chapter 1.1 Coronaviruses

Schamoni-Kast, K., and Uetrecht, C. (2025). From Science to Fiction – Where are We Connecting in Vivo and in Vitro Results in Polyprotein Processing? Preprint at Social Science Research Network, https://doi.org/10.2139/ssrn.5144731 https://doi.org/10.2139/ssrn.5144731.

Simanjuntak, Y., **Schamoni-Kast, K**., Grün, A., Uetrecht, C., and Scaturro, P. (2021). Top-Down and Bottom-Up Proteomics Methods to Study RNA Virus Biology. Viruses *13*, 668. https://doi.org/10.3390/v13040668.

Chapter 3.1 Polyprotein production - 3.1.3 Discussion

Schamoni-Kast, K., Krichel, B., Damjanović, T., Kierspel, T., Toker, S., and Uetrecht, C. (2024). The kinetics of SARS-CoV-2 nsp7-11 polyprotein processing and impact on complexation with nsp16. Preprint at bioRxiv, https://doi.org/10.1101/2024.01.06.574466 https://doi.org/10.1101/2024.01.06.574466.

The revised version has been submitted to Nature Communication.

Schamoni-Kast, K., Krichel, B., Damjanovic, T., Said, F.-A., Kierspel, T., Toker, S., and Uetrecht, C. (2025). The kinetics of nsp7-11 polyprotein processing and impact on complexation with nsp16 among human coronaviruses. (Zenodo). https://doi.org/10.5281/ZENODO.15488266 https://doi.org/10.5281/ZENODO.15488266. (Nat. comm. under revision/accepted)

Chapter 4 Summary and outlook

Schamoni-Kast, K., and Uetrecht, C. (2025). From Science to Fiction – Where are We Connecting in Vivo and in Vitro Results in Polyprotein Processing? Preprint at Social Science Research Network, https://doi.org/10.2139/ssrn.5144731 https://doi.org/10.2139/ssrn.5144731.

Chapter 5.1 - 5.3 and 5.4.2

Schamoni-Kast, K., Krichel, B., Damjanovic, T., Said, F.-A., Kierspel, T., Toker, S., and Uetrecht, C. (2025). The kinetics of nsp7-11 polyprotein processing and impact on complexation with nsp16 among human coronaviruses. (Zenodo). https://doi.org/10.5281/ZENODO.15488266 https://doi.org/10.5281/ZENODO.15488266. (Nat. comm. under revision/accepted)

Chapter 5.6 Data analysis with Python

Schamoni-Kast, K., Krichel, B., Damjanovic, T., Said, F.-A., Kierspel, T., Toker, S., and Uetrecht, C. (2025). The kinetics of nsp7-11 polyprotein processing and impact on complexation with nsp16 among human coronaviruses. (Zenodo). https://doi.org/10.5281/ZENODO.15488266 https://doi.org/10.5281/ZENODO.15488266.

Chapter 6 Supplementary Material

Schamoni-Kast, K., Krichel, B., Damjanovic, T., Said, F.-A., Kierspel, T., Toker, S., and Uetrecht, C. (2025). The kinetics of nsp7-11 polyprotein processing and impact on complexation with nsp16 among human coronaviruses. (Zenodo). https://doi.org/10.5281/ZENODO.15488266 https://doi.org/10.5281/ZENODO.15488266. (Nat. comm. under revision/accepted)

Other Publications

Grün, A.F.R., Said, F.-A., Schamoni-Kast, K., Damjanovic, T., Bosse, J., and Uetrecht, C. (2025). Fast tracking native mass spectrometry: Skipping over buffer exchange. Preprint at bioRxiv, https://doi.org/10.1101/2025.02.22.639503 https://doi.org/10.1101/2025.02.22.639503.

Abbreviations

(+)ssRNA	positive single-strand RNA	MERS-CoV	Middle east respiratory syndrome CoV
AF	AlphaFold	MHV	mouse hepatitis virus
Ar	Argon	mRNA	messenger ribonucleic acid
AUC	area under the curve	MS	mass spectrometry
BU-MS	bottom-up MS	MSA	multiple sequence alignment
CE	capillary electrophoresis	N protein	nucleocapsid proten
CEM	charged ejection model	N_2	(di-)nitrogen
CID	collision induced dissociation	NaCl	sodium chloride
CLEM	correlative light and electron microscopy	NHS	N-hydroxysuccinimide
CoV	coronavirus	NiRAN	Nidovirus RdRp associated nucleotidyl transferase domain
CRM	charged residue model	NMPylation	nucleotidyl-monophosphate transfer
cryo-EM	cryogenic electron microscopy	NMR	nuclear magnetic resonance
cryo-ET	cryogenic electron tomography	NoV	norovirus
CS7/8	cleavage site 7/8	nsp	non-structural proteins
CS8/9	cleavage site 8/9	ORF	open reading frame
CS9/10	cleavage site 9/10	PAE	predicted aligned errors
CS10/11	Cleavage site 10/11	PALM	Photo-Activated Localization Microscopy
Da	Dalton	PDB	protein data bank
DNA	desoxyribonucleic acid	pLDDT	predicted local-distance difference test
DC	direct current	PL^{pro}	papain-like proteases
E protein	envelope protein	pp1a	polyprotein 1 a
E. coli	*Escherichia coli*	pp1ab	polyprotein 1 ab
ESI	electrospray ionization	PTM	posttranslational modification
eV	electron Volt	pTM	predicted template modelling
FACS	fluorescence-activated cell sorter	Q	quadrupole
FPOP	fast photochemical oxidation of proteins	RAP	RNA antisense purification
FRET	Förster resonance energy transfer	RBD	receptor binding domain
FT-ICR	Fourier-transform ion cyclotron resonance	RdRP	RNA-dependent RNA polymerase
GHS	Globally Harmonized System of Classification and Labelling of Chemicals	RF	radio frequency
H^+	proton	RNP	ribonucleoprotein
HCD	Higher-energy C-trap dissociation	RTC	replication/transcription complex
HDX-MS	hydrogen-deuterium exchange MS	S protein	spike protein
He	Helium	SARS-CoV	severe acute respiratory syndrome CoV
HV	High vacuum	SASA	solvent exposed surface area
IEM	ion evaporation model	SDS-PAGE	Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis
IMS	Ion mobility spectrometry	SEC	size-exclusion chromatography
IPTG	isopropyl-β-D-1-thiogalactopyranosid	SILAC	stable isotope labeling by amino acids in cell culture
k	kinetic rate constant	SRM	super-resolution microscopy
K_D	dissociation constant	STED	stimulated emission depletion
LC-MS	liquid chromatography MS	STORM	Stochastic Optical Reconstruction Microscopy

M protein	membrane protein	TD-MS	Top-down MS
m/*z*	mass-to-charge ratio	TGEV	transmissible gastroenteritis virus
TMT	tandem mass tag		
TOF	Time-of-Flight		
Ubl2	ubiquitin-like domain 2		
UHV	ultra high vacuum		
ULP-1	ubiquitin-like-specific protease 1		
WHO	World Health Organization		
Xe	Xenon		
XL-MS	cross-linking MS		
α-CoV	Alphacoronavirus		
β-CoV	Betacoronavirus		
γ-CoV	Gammacoronavirus		
δ-CoV	Deltacoronavirus		

In this work, amino acids are abbreviated according to the single-letter code:

alanine	Ala	A
arginine	Arg	R
asparagine	Asn	N
aspartic acid	Asp	D
cysteine	Cys	C
glutamic acid	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
methionine	Met	M
phenylalanine	Phe	F
proline	Pro	P
serine	Ser	S
threonine	Thr	T
tryptophan	Trp	W
tyrosine	Tyr	Y
valine	Val	V

List of figures

List of tables

List of supplementary figures

List of supplementary tables

Abstract

In coronavirus (CoV) infection, polyproteins (pp1a/pp1ab) are processed into non-structural proteins (nsps), which largely form the replication/transcription complex (RTC). The polyprotein processing and complex formation is critical and offers potential therapeutic targets. However, the interplay of polyprotein processing and RTC-assembly remains poorly understood. In this work, two key aspects were studied: The order of polyprotein processing by viral main protease M^{pro} and its influence on complex formation with the methyltransferase nsp16. Central to this investigation was the establishment of an approach to determine rate constants *k* from cleavage sites in structured CoV polyprotein based on native mass spectrometry (MS). We used this approach for a comprehensive analysis of polyprotein processing in four human CoVs: Severe acute respiratory syndrome CoV 1 and 2 (SARS-CoV-1 and -2), middle east respiratory syndrome CoV (MERS-CoV), and human CoV-229E (HCoV-229E).

Our sensitive and precise native MS approach provided novel insights into polyprotein processing of nsp7-11 region revealing both conserved features and species-specific variations. The experimentally determined rate constants are put into perspective with a comprehensive analysis of primary sequences and structural models.

Kinetic rate constants were determined for the four cleavage sites, CS7/8, CS8/9, CS9/10, and CS10/11, in all four viruses. Based on the presence of intact cleavage sites, processing species were assigned to the cleavage sites, which simplified the multi-reaction to a first-order reaction. This approach allowed us to extract cleavage site kinetics for each site and compare them between CoV species. The kinetics of multi-cleavage reaction revealed that the order and rate of processing are not conserved across species.

Conversion rates at CS7/8 in all four CoVs were substantially slowed down compared to other cleavage sites. The primary structure that influences M^{pro} cleavage efficiency could not explain the different rates alone, suggesting a structural hindrance at CS7/8. AlphaFold prediction models indicated an α-helical fold at this location, which reduces the cleavage efficiency of M^{pro}. In general, the AlphaFold predictions confirmed the experimental data and could provide structural rationale, though local confidence scores at the cleavage sites were low, potentially due to flexible loop regions.

Notably, species-specific differences indicated that cleavage at CS10/11 is not essential. Binding experiments with SARS-CoV-2 nsp16, SARS-CoV-2 nsp10, and the nsp10-11 processing intermediate of MERS-CoV confirmed that cleavage at CS10/11 is not required for nsp16+10 complex formation. However, full cleavage at CS9/10 appears necessary for forming an active methyltransferase complex.

A key advantage of our native MS approach is its ability to capture structural context while directly detecting processing intermediates and protein-protein interactions. This provides significant benefits over traditional peptide-based assays. Moreover, cleavage site kinetics were extracted including protein folding. Our findings offer new mechanistic insights into CoV polyprotein processing and complex assembly, which may inform future antiviral drug development strategies targeting these essential viral processes.

Zusammenfassung

Bei der Infektion mit Coronaviren (CoV) werden Polyproteine (pp1a/pp1ab) zu Nicht-Strukturproteinen (nsps) prozessiert, die größtenteils den Replikations-/Transkriptionskomplex (RTC) bilden. Die Polyproteinprozessierung und Komplexbildung ist entscheidend und bietet potenzielle therapeutische Angriffspunkte. Das Zusammenspiel der Polyproteinprozessierung und des RTC-Zusammenbaus ist jedoch schlecht verstanden. In dieser Arbeit wurden zwei Schlüsselaspekte untersucht: Die Reihenfolge der Polyproteinprozessierung durch die virale Hauptprotease M^{pro} und ihr Einfluss auf die Komplexbildung mit der Methyltransferase nsp16. Außerdem wurde ein Ansatz zur Bestimmung von Geschwindigkeitskonstanten *k* von Schnittstellen in strukturiertem CoV-Polyprotein auf der Grundlage von nativer Massenspektrometrie (MS) entwickelt. In dieser Studie wird eine umfassende Analyse Polyproteinverarbeitung in vier humanen CoVs vorgestellt: Schweres akutes Atemwegssyndrom CoV 1 und 2 (SARS-CoV-1 und 2), mittleres östliches Atemwegssyndrom CoV (MERS-CoV) und human CoV-229E (HCoV-229E).

Unser sensitiver und präziser nativer MS-Ansatz ermöglicht neue Einblicke in die Polyproteinprozessierung der nsp7-11 Region, wobei sowohl konservierte Merkmale als auch artspezifische Variationen sichtbar wurden. Die experimentell ermittelten Geschwindigkeitskonstanten werden durch eine umfassende Analyse der Primärsequenzen und Strukturmodelle relativiert.

Die kinetischen Geschwindigkeitskonstanten wurden für die vier Spaltstellen CS7/8, CS8/9, CS9/10 und CS10/11 in allen vier Viren bestimmt. Aufgrund des Vorhandenseins intakter Spaltstellen wurden den Spaltstellen Prozessierungsspezies zugeordnet, wodurch die Mehrfachreaktion zu einer Reaktion erster Ordnung vereinfacht wurde. Dieser Ansatz ermöglichte es uns, die Kinetik der Spaltstellen für jede Stelle zu extrahieren und sie zwischen den CoV-Spezies zu vergleichen. Die Kinetik der Mehrfachspaltungsreaktion zeigte, dass die Reihenfolge und die Geschwindigkeit der Verarbeitung bei den verschiedenen Arten nicht konserviert sind. Die Umwandlungsraten an CS7/8 waren bei allen vier CoVs im Vergleich zu anderen Schnittstellen erheblich verlangsamt. Die Primärstruktur, die die M^{pro}-Spalteffizienz beeinflusst, konnte die unterschiedlichen Raten nicht allein erklären, was auf ein strukturelles Hindernis an CS7/8 hindeutet. AlphaFold-Vorhersagemodelle wiesen auf eine α-Helixfaltung an dieser Stelle hin, die die Spaltungseffizienz von M^{pro} verringert. Im Allgemeinen bestätigten die AlphaFold-Vorhersagen die experimentellen Daten und konnten eine strukturelle Erklärung liefern, obwohl die lokalen Vertrauenswerte an den Schnittstellen niedrig waren, was möglicherweise auf flexible Schleifenbereiche zurückzuführen ist.

Arten-spezifische Unterschiede wiesen darauf hin, dass die Spaltung an CS10/11 nicht essenziell ist. Bindungsexperimente mit SARS-CoV-2 nsp16, SARS-CoV-2 nsp10 und dem nsp10-11-Prozessierungszwischenprodukt von MERS-CoV bestätigten, dass die Spaltung an CS10/11 für die Bildung des nsp16+10-Komplexes nicht erforderlich ist. Eine vollständige Spaltung an CS9/10 scheint jedoch für die Bildung eines aktiven Methyltransferase-Komplexes erforderlich zu sein.

Ein entscheidender Vorteil unseres nativen MS-Ansatzes ist die Fähigkeit, den strukturellen Kontext zu erfassen und gleichzeitig Prozessierungszwischenprodukte und Protein-Protein-Interaktionen direkt nachzuweisen. Dies bietet erhebliche Vorteile gegenüber herkömmlichen peptidbasierten Assays. Darüber hinaus wurden die Kinetiken der Spaltstellen einschließlich der Proteinfaltung extrahiert. Unsere Ergebnisse bieten neue mechanistische Einblicke in die Prozessierung von CoV-Polyproteinen und den Zusammenbau von Komplexen, die als Grundlage für künftige Strategien zur Entwicklung antiviraler Medikamente dienen können, die auf diese wichtigen viralen Prozesse abzielen.

1. Introduction

1.1. Coronaviruses

Coronaviruses (CoVs) can infect a wide variety of hosts and cause diseases from respiratory tract diseases and gastroenteritis to central nervous system diseases. Over the last two decades, several CoVs transmitted from animals to humans, i.e., zoonotic infections caused by CoVs, have posed an increasing challenge to public health and the economy.

Until the first appearance of severe acute respiratory syndrome CoV (named herein SARS-CoV-1 to have a clearer distinction from SARS-CoV-2) in 2002, CoVs were largely regarded as “virology backwater”. The epidemic caused hundreds of deaths and exhibited fatality rates from 3-8% in China, and up to 27% in Taiwan[1,2]. Since the emergence of SARS-CoV-1, CoVs have received increased attention[3]. This attention has led to the discovery of further human CoVs such as HCoV-NL63 and HCoV-HKU1 in 2004 and 2005, respectively. SARS-CoV-1 was introduced as a zoonotic infection, likely originating from palm civets. Other CoVs had already been circulating among humans for a long time. For example HCoV-OC43 and HCoV-229E have been circulating among humans for at least 60 years[4,5]. In fact, one-third of common cold infections in humans are caused by human CoVs that are globally spread in the population: HCoV-229E, HCoV-NL63, HCoV-OC43 and HCoV-HKU1[4]. In 2012, another coronaviral zoonosis occurred causing initially an epidemic outbreak in Saudi Arabia and then spread over to further countries[6,7]. Middle east respiratory syndrome CoV (MERS-CoV) was transmitted from dromedary camels, which are now known to be a natural and important reservoir for MERS-CoVs. Hence, MERS-CoV has not disappeared completely since its initial outbreak, unlike SARS-CoV-1 which has not been detected since 2005[6,8]. MERS remains a major public health concern in affected countries due to its high fatality rates, which are approximately 35% in primary cases and 20% in secondary cases[7].

Experts have warned of the pandemic potential of CoVs after SARS and MERS epidemics at the latest. In December 2019, a novel pneumonia was reported in Wuhan, Hubei Province and a novel CoV was later identified as the causative virus[9,10]. Due to this newly emerged CoV named SARS-CoV-2 the World Health Organization (WHO) declared a public health emergency of international concern and a pandemic outbreak within four months[11,12]. To this day, the exact origin of the virus remains unknown. A laboratory outbreak of a SARS-CoV-like modified virus is presumed to be unlikely due to several reasons. One of which is that the receptor binding domain (RBD) in the spike (S) protein was not optimal for human infection. Furthermore, SARS-CoV-2 utilizes the polybasic furin cleavage site for cell entry. Given that this feature is unique to SARS-CoV-2 within lineage B and its function was previously unknown, the rationale for a potential deliberate introduction cannot be readily explained[13,14]. Genetic

modifications such as reverse-genetic systems usually leave traces that indicate, for example, a previously used virus backbone. This was not detected[15,16]. The RBD domain of CoVs detected from illegally imported Malayan pangolins showed significant similarity to the RBD in SARS-CoV-2. However, the similarity of the entire genome is not as high as with the genome from RaTG13 bat virus.[13]. Therefore, a zoonotic origin is still the most likely scenario, although the direct progenitor remains to be identified. However, a laboratory origin could not be entirely excluded to date[13,17,18].

Coronaviral infections in humans usually cause mild colds. However, human CoV infections can also be severe in the vulnerable population, i.e., the elderly, children and immunocompromised patients causing pneumonia, bronchiolitis, and laryngitis[19]. Further, HCoV-NL63 infection is associated with croup in children and HCoV-43 seems to play a role in chronic demyelinating disease and acute encephalomyelitis[19–21].

The majority of human coronaviral infections progress mildly, including now infections of SARS-CoV-2. However, natural reservoirs in bats and other mammals persist, posing the risk of recurring cross-species transmission and severe zoonosis[19]. Hence, there is a need for ongoing surveillance as well as basic research to understand the differences in severity of infection[19,22].

1.1.1. Coronavirus biology – polyprotein processing & replication

CoVs are enveloped (+)-sense single-strand RNA viruses ((+)ssRNA) and belong to the order of *Nidovirales*, which comprises four families: *Coronaviridae*, *Toroviridae*, *Roniviridae* and *Arteviridae*. *Coronaviridae* is further divided into two sub-families, one of which is *Orthocoronavirinae*, which in turn can be divided into four genera: *Alpha-, Beta-, Gamma-*, and *Deltacoronaviruses* (α-CoV, β-CoV, γ-CoV, and δ-CoV)[23].

Viruses infecting humans and other mammals are categorized to as α-CoV and β-CoV. CoVs grouped into γ-CoV and δ-CoV genera mostly infect birds. The major part of the replicase gene, the membrane protein M and the nucleocapsid (N) protein are present in all four genera and exhibit only a few point mutations and small insertions/deletions. Next to this conserved part of the genome, *Coronaviridae* have more variable parts that are utilized for classification. In particular, each genus shows a characteristic set of accessory proteins[23].

Coronaviruses consists of the structural proteins spike (S), envelope (E), nucleocapsid (N), membrane (M) and the non-structural proteins (nsp) 1-16. Some virions of β-CoV possess haemagglutinin-esterase as further structural protein. While M and E protein control the incorporation process of the RNA genome in the viral particles, N proteins main function is to bind and facilitates packaging[24]. A second function of N is the regulation of viral gene transcription, which is of higher interest in regards to the scope of this review[25,26]. However,

the biochemical functions of N and its domains are still not fully understood[27]. Interspersed within the ORF for structural proteins are accessory proteins, whose numbers vary among CoV species. These unique viral proteins, which have no known homologs in other viruses, play critical roles in modulating the host immune response, inducing apoptosis, and triggering inflammation[28,29].

Infection takes place via specific binding of coronavirus S protein to the cellular entry receptors and further host factors. Several entry receptors were identified for different CoVs: human aminopeptidase N (HCoV-229E), angiotensin-converting enzyme 2 (ACE2; HCoV-NL63, SARS-CoV-1 and SARS-CoV-2) and dipeptidyl peptidase 4 (MERS-CoV). Depending on the expression of these receptors on the different cell types, tropism and pathogenicity are influenced. In SARS-CoV-2 for example, S protein binds to the cellular receptor ACE2 and attachment factors such as serine protease TMPRSS2, which then promotes either cellular or endosomal membrane fusion. Upon membrane fusion and viral uptake of the cell, the (+)ssRNA genome is released and the two large open reading frames (ORF) are directly translated as polyprotein 1a (pp1a) and pp1ab. These polyproteins are processed by inherent proteases generating mature nsps that finally assemble into a large replication/transcription complex (RTC). Simultaneously with the release of nsps, perinuclear viral replication organelles are formed that create a protective microenvironment in which RNA replication and the transcription characteristic of coronaviruses take place. This microenvironment consists of double membrane vesicles (DMVs) interconnected with folded membranes (CMs) and small open double membrane spheres (DMSs). Finally, budding of virions most likely takes place at ER-to-Golgi intermediate compartements[30].

The event of polyprotein processing occurs in many (+) ssRNA, several double-stranded RNA viruses, and all retroviruses. It is a strategy that is regulated locally and timely, and that is crucial for these viruses[31–33]. In the following sections we are interested in how the combination of *in vivo* and *in vitro* techniques has helped to understand the mechanism of polyprotein processing, particularly in the family of *Coronaviridae.* The family of *Coronaviridae* harbors the members of the largest RNA genomes on average, potentially leading to the most intricately designed replication systems, and therefore putatively less reliant on interactions with host proteins. We aim to highlight, which current techniques could contribute to unresolved molecular mechanisms and, more importantly, how and where these techniques need to evolve to advance polyprotein research in general and in *Coronaviridae* in particular.

After viral entry, when the envelope has fused with the endosomal membrane of the cell, nucleocapsids release the viral genome, which consists of multiple open reading frames (ORFs). One of which, ORF1a/b, is then directly translated into two polyproteins pp1a and pp1ab. The two polyprotein versions pp1a and pp1ab are translated by the host ribosome in

2:1 to 5:1 ratios[34–37], respectively. These ratios are a result of the -1 ribosomal frameshift enabling the ribosome to overcome the termination signal of ORF1a. As a consequence, nsp11 is omitted in pp1ab. The frameshift event is highly conserved[38–40]. Eventually, two huge polyproteins are produced, pp1a comprising nsp1-nsp11 (~ 490 kDa) and pp1ab comprising nsp1-nsp16 (~ 749 kDa). pp1ab is extended by ~ 2700 amino acids compared to pp1a and the conservation of this extended region is much higher[41]. The polyproteins possess two or three viral proteases. Structure and mechanism of both proteases/protease domains are clarified[42–45]. Depending on the CoV strain, they contain one or two papain-like proteases (PL^{pro}) residing in the large nsp3 subunit. PL^{pro} processes nsp1-4. In addition to its processing function, PL^{pro} plays a role in inhibiting host innate immune response and is involved in dysregulated inflammatory responses[46]. PL^{pro} is composed of five domains: fingers, palm, thumb, the ubiquitin-like domain 2 (Ubl2), and a domain preceding Ubl2 and papain-like protease 2[47,48]. The finger domain binds a zinc ion and uniquely connects the two domains of a papain-like fold[49]. Next to PL^{pro}, pp1a and pp1ab include a chymotrypsin-like protease or main protease (M^{pro}, nsp5) that processes nsp4-16[32,42,50]. M^{pro} has a catalytic dyad consisting of Cys145 and His41. It exhibits a non-canonical substrate specificity, requiring a Q residue at the P1 position, as defined by the Schechter and Berger nomenclature[51] (**Figure S1**), and preferring the S, A, G, or N residues at P1′ position. Cleavage efficiency is also influenced by the residues at the flanking P2 and P2′ positions[52,53]. M^{pro} is active as a dimer and autocleaves itself. However, the maturation process of M^{pro} is not well-understood. A recent model suggests that *N*-terminal release occurs within immature dimers (*cis*-activity) and *C*-terminal release occurs between two distinct dimers (*trans* activity)[54]. This model is partially consistent with other models proposed[55–57]. Cleavage in *trans* is more efficient than in *cis*[58]. While cleavage interaction with host proteins for immune evasion purposes are known for both proteases, no host protein interactions have been shown to be required in order to conduct polyprotein processing[59–62].

The following section is intended to serve as a reference point for polyprotein processing research and the formed nsps, highlighting specific techniques used and delineating the known from the unknown.

Although nsp1 is involved in modulating antiviral host immune response, this protein occurs only in α- and β- CoVs[63]. Furthermore, nsp1 and nsp2 are involved in translational regulation and potentially shut-down host protein production[64]. The exact pathway and molecular mechanism are unknown. However, a recent study suggested a mechanistic framework, in which the *C*-terminal domain of nsp1 binds to the mRNA channel of the ribosome and thereby inhibits translation. Viral mRNA can escape this mechanism, which seems to be conserved among β-CoVs[65]. Importantly, nsp1, nsp4 and nsp6, are involved in the formation of double membrane vesicles (DMVs) providing a suitable microenvironment for viral replication.

Moreover, nsp3 possesses membrane domains, associates into a pore structure facilitating shuttling of RNA in and out of the DMVs. Hereby, the extending crown domains are suggested to serve as membrane anchors for the soluble replication transcription complex (RTC)[32,41,66,67]. Nsp7 and nsp8 are processivity enhancing co-factors for the RNA-dependent RNA polymerase (RdRP) nsp12 facilitating RNA replication[68]. The quaternary structure of nsp7 and nsp8 complexes is not conserved and varies within and between α- and β-CoVs, which indicate two different assembly pathways[69]. Next to the interface domain that interacts with nsp7 and nsp8, nsp12 contains two more domains: the RdRP domain consisting of the usual polymerase thumb, finger and palm motifs and the Nidovirus RdRp associated nucleotidyl transferase domain (NiRAN). This domain is a signature domain for the order of *Nidovirales*, to which CoVs belong[70]. While nsp9, an RNA-binding protein, has been identified as a target of the NiRAN domain that undergoes nucleotidyl-monophosphate transfer (NMPylation), further research is needed to fully understand the potential functions of the NiRAN domain. These include the potential functions as ligase, guanosine triphosphate-dependent 5' nucleotidyl transferase, and uracil triphosphate-dependent protein primer initiating RNA replication.[70–73]. Furthermore, nsp9 is an essential co-factor for nsp12 mediated capping of nascent RNA, though the molecular mechanism is not completely understood[74]. Ultimately, nsp9 carries an *N*-terminal posttranslational modification (PTM), and thereby mediates RNA capping, which is essential for host immune evasion[72,75,76]. Another important small nsp from pp1a is nsp10. Nsp10 is a co-factor for nsp14 and nsp16 enhancing their enzymatic functions. Whereas nsp13 executes helicase function, nsp14-16 are involved in capping and the proofreading function of the RTC[77–79]. This function is crucial for avoiding too many mutations during replication and is unique among (+)ssRNA viruses. This proofreading function is not unique to CoVs but is also shared by other nidoviruses, All such as toroviruses and roniviruses[78,80]. Thus, most RNA viruses without proofreading function are naturally limited to < 30 kb, although there are some exceptions for members of the *Flaviviridae*[81].

Compared to CoVs, polyprotein processing in most other (+)ssRNA viruses involves fewer nsps, presumably because they do not carry the information of a complete replication machinery and rely more on interactions with host proteins as is true for flaviviruses and some alphaviruses[31,82]. Independent of the number of nsps, polyprotein processing is a highly regulated process, in which one or more polyproteins have to be processed correctly, in a timely and local fashion. Then the liberated nsps are involved in forming the replication machinery or facilitating viral replication in a broader sense.

In CoVs, polyprotein processing has mainly been studied *in vitro* and has improved the understanding of regulatory processes. Different *in vitro* techniques have shown partly contradictory results. While these studies agree on most cleavage site kinetics, they disagree

on the cleavage sites converted first[83–87]. Once the nsps are released, most of them assemble into the large RTC. The assembly pathway and if or how some nsps are involved in supporting the complex formation is not clear. As CoV polyprotein processing and RTC assembly are potentially less reliant on host protein interactions, these processes are predestined for thorough *in vitro* analyses. While certain regions have been studied, it is still challenging to recombinantly express the whole polyprotein of CoVs (~749 kDa)[88,89]. The two internal proteases autocleave themselves, which makes a controlled initiation of polyprotein processing challenging[64]. In addition, there are some nsps in the polyprotein that contain transmembrane domains, which often places higher demands on the techniques. The role of the structural N protein during polyprotein processing and RTC assembly is not clear either. The underlying molecular details of coronaviral replication fidelity are not well known. While techniques that can provide high-resolution structures are often limited in their ability to capture dynamics, techniques that can capture dynamics often provide much lower structural resolution[90]. Getting the timing right and finding this rare process is often a challenge in *in vivo* studies. Polyprotein processing starts a few hours after infection, continues for several hours and then faints as a study in avian infectious bronchitis virus (IBV) showed[91].

1.1.2. Polyprotein processing in the context of structural biology techniques.

This section highlights available techniques that have been used to study CoV polyprotein processing over the last 30 years. In addition, current techniques that could contribute to unresolved molecular mechanisms and the added value they would bring are presented.

Key events in CoV polyprotein processing and RTC assembly have been studied with *in vivo* and *in vitro* techniques, which contribute differently to the understanding of the underlying mechanism, for example, protease specificity, cleavage site identification and turnover, and the role of proteoforms. Proteoforms include PTMs such as glycosylation, phosphorylation etc., but also sequence variants, splice isoforms and proteolytic processing[92]. Processing alone can give rise to more than 100 potential proteoforms including all possible intermediates in CoVs. *In vitro* methods take advantage of the controlled environment and isolation of mechanisms, which simplifies to pinpoint an effect to a player. *In vivo* methods often have higher biological and physiological relevance but also higher complexity. This complexity of *in vivo* techniques is both a strength and a limitation. Due to the complexity of the system, it can be difficult to isolate complex processes. Thus, validation of *in vitro* findings can be challenging and vice versa. The strengths and limitations of *in vitro* and *in vivo* approaches will be demonstrated based on polyprotein processing events and other steps during replication. *In vivo* approaches comprise studies with infected cells. Not all state-of-the-art techniques have been applied to CoV polyprotein processing or the RTC. If the approach is beneficial, other viruses and even non-viral examples will also be looked at.

For the purpose of the review[93] and this work, the terms *in vivo* and *in vitro* are redefined as follows: *in vivo* studies observe complex biological interactions and processes in living cells (*in cellulo*), be it in cell culture, organoid models or entire organisms mostly in infection context. Furthermore, methods that obtain their samples from such an *in vivo* environment and hence monitor this state are also considered as *in vivo*. For example, protein samples that are extracted from infected cells by co-immunoprecipitation and subjected to mass spectrometry (MS) would also be considered *in vivo*. Consequently, any technique that mimics natural conditions and uses recombinant proteins to study specific biochemical processes or molecular details is considered an *in vitro* study. Furthermore, the term *in situ* is used in the context of microscopy and describes the situation in fixed cells.

1.1.2.1. Discovery and identification of polyproteins and their cleavage sites

Looking at examples in *Coronaviridae*, polyprotein processing intermediates have been identified *in vivo* by immunoprecipitation of lysates from infected cells from mouse hepatitis virus (MHV), HCoV-229E or IBV[86,94–100]. Denison *et al*. discovered large polyproteins in MHV requiring processing at multiple cleavage sites. Furthermore, they detected intermediate products such as p28 and monitored their degradation. Although infected cells were synchronized to accumulate polyprotein and intermediates, it was not possible to monitor the initial cleavage[94]. Ziebuhr and Sidell *et al*.[86] had validated multiple cleavage sites of HCoV-229E by performing cleavage assays that revealed different sized intermediates and demonstrated kinetic differences of cleavage sites at the end of pp1a and the beginning of pp1ab. Further studies revealed more precise intermediate products in MHV and first subcellular localization polyprotein intermediates of HCoV-229E and MHV were shown[97,98,101]. Subcellular distribution of specific proteins may point to possible function. Here, they found nsp7-nsp10 co-localized with the helicase (nsp13) in the cell periphery as well as in perinuclear regions indicating that viral RNA synthesis happens in replication complexes. Thus, localized mature nsp7-nsp10 were attributed roles in replication[101]. Mutagenesis performed in multiple CoVs had often been used to identify cleavage sites and prove their relevance for virus viability. Later, bioinformatics facilitated the prediction of cleavage sites together with, for example, M^{pro} cleavage site specificity. Thus, almost all replicase proteins, nsp1-nsp16, could be detected in SARS-CoV-1 within one *in vivo* study using immunofluorescence and immunoblots[86,96,102,103]. The above examples show how challenging the discovery and identification of polyproteins and their cleavage sites has been experimentally. Bioinformatics tools have accelerated this process and contributed significantly to polyprotein research. Early computational studies suggested that the two polyproteins, pp1a and pp1ab, contain all the proteins necessary for viral replication, including the proteinases in IBV and MHV. In addition, potential cleavage sites were predicted[104,105]. Based on these predictions processing activities were tested and

characterized *in vitro* in IBV and MHV[106,107]. This demonstrated how bioinformatics guided the experimental work and ongoing applications are defining the evolutionary conservation of CoV cleavage sites, proteases, and mature proteins[108]. More importantly, *in silico* work is often shedding light on rather neglected areas. For example, a recent study suggested that genetic transfer between nidovirus species is more common than previously thought[109].

1.1.2.2. Dissecting cleavage efficiency and order of polyprotein processing

Once cleavage sites were discovered and validated, the focus was on studying cleavage efficiency and order. Assays based on synthetic peptides with radioactive or fluorescent labels as substrates were employed in order to demonstrate *in vitro* polyprotein processing for both proteases of MHV, HCoV-229E and IBV[86,96,98,110]. Looking at M^{pro} (nsp5), the self-cleavage sites showed the fasted kinetic rate constants in line with dimer formation being required for activity. Notably, cleavage site 8/9 (CS8/9) showed hardly any processivity in these assays due to peptide folding, which may differ from the secondary structure *in vivo*[86,111]. Reverse genetics in MHV showed that disruption of CS8/9 was lethal, which indicated that this site albeit being suboptimal for M^{pro} is cleaved *in vivo*[112]. The *N*-terminal NNE motif at P1′–P3’ of nsp9 is widely conserved among CoVs and now known to be modified assisting nsp12 in capping of nascent RNA[72]. Fan *et al*. have shown that synthetic peptides spanning CS8/9 are likely to fold into α-helices, which hampers the cleavage[113]. Notably, contrary to expectations from peptide assays, Krichel *et al*. showed that the CS8/9 is processed much faster in the full-length, folded protein[114]. Nevertheless, synthetic peptides and recombinantly expressed enzymes were essential to compare cleavage efficiencies of M^{pro} in different CoV species, namely HCoV-229E, MHV and transmissible gastroenteritis virus (TGEV), and to assess conservation of cleavage sites[86,115]. Later such assays were employed to compare kinetics and to screen for inhibitors[86,111,116,117]. Investigating cleavage efficiencies with such assays is advantageous given their high-throughput nature, but synthetic peptides lack the polyprotein context as explained above.

Straightforwardly designed Förster resonance energy transfer (FRET)-based experiments contributed to the understanding of protease substrate specificity and regulation of polyprotein processing as seen in norovirus and CoV[118–121]. Such FRET-based *in vitro* systems have been developed further. For example, Kenward *et al.*[85] designed a linked protein FRET platform using tethered FRET pair substrates, in which cleavage site organization is closer to the native polyprotein than in previously reported FRET systems. This design overcomes the problem of poor solubility and inner filter effects of former FRET substrate designs, in which kinetic values largely depended on chemical properties and complicated comparisons. This new platform allowed the comparison of all eleven cleavage sites in SARS-CoV-2. Using this advanced FRET assay, they came up with following cleavage order reflecting sheer preference in

cleavage efficiency of M^{pro}: CS9/10 > CS7/8 > CS8/9 > CS10/11[85]. Another way of studying polyprotein processing including the structural cleavage site features is native MS. Native MS mimics a native-like environment, and thereby allows the analysis of a complex biochemical reaction with direct evidence of intermediates including subsequent oligomerization[114]. In SARS-CoV-1, a time-resolved, label-free cleavage assay was performed, showing all intermediate cleavage products at the same time, allowing postulation of a cleavage order (CS9/10 > CS8/9 >> CS7/8). Moreover, the subsequent $nsp7_2+8_2$ heterotetramer formation was monitored alongside.

1.1.2.3. Structural context in time-resolved cleavage assays

A time-resolved study of *in vivo* polyprotein processing does not provide such detailed processing dynamics, but can give information about long lasting intermediates as a study of polyprotein processing of pp1ab has shown in IBV using *in vivo* radiolabeling. For example, they observed a relatively stable intermediate protein of 160 kDa lasting for over 5 hours[91]. In alphaviruses, a similarly conducted *in vivo* radiolabeling assay was put into context with a crystal structural model of nsP2-nsP3, which is already half of the polyprotein in alphaviruses, elucidating structural details and dynamics on polyprotein processing[122].

Due to the large size and intrinsic flexibility of CoV polyproteins, high-resolution structures of the full pp1a/pp1ab (nsp1-nsp11/nsp1-nsp16) are unavailable. Only the cleavage site peptides bound to M^{pro} could be resolved or individual folded domains, which are the nsps connected through flexible linkers comprising the cleavage sites[123].

Structural MS is a toolbox of different MS approaches. Next to native MS, it comprises also hydrogen-deuterium exchange MS (HDX-MS) and crosslinking MS (XL-MS) (**Figure 1**). Both allow the study of structural dynamics and provide local structural information. XL-MS links residues that are in proximity to each other resulting in linked peptides. HDX-MS gains structural information from surface labeling by probing the exchange from backbone amide hydrogen to deuterium. In both approaches, the proteins are labelled or linked and then proteolytically digested to peptides, which are then detected by MS (**Figure 1**). XL-MS is considered a low-resolution technique and provides structural data sufficient to compute 3D models. By combining derived constraints from XL-MS and integrative structural modeling, it can elucidate structural models of e.g., large protein complexes up to atomic resolution[124]. XL-MS always involves covalent reactions of crosslinkers. There are different classes of crosslinkers depicting different strengths and limitations. The N-hydroxysuccinimide (NHS) esters are the most widely used mechanism creating stable bonds with primary and secondary amines such as the free *N*-terminus and the amino groups from lysine side chains. Mentioned limitations would be: unexpected reactions with contaminant ammonium ions or unwanted

reactions with serine and tyrosine residues next to the desired reactions with lysine residues[125]. Furthermore, NHS esters tend to react preferentially with tyrosine residues and free *N*-termini under acidic conditions[126]. This example gives an insight into aspects that need to be considered when choosing the crosslinking reagent for XL-MS experiments. Additionally, chemical labelling can distort the picture of protein structures when the experiments are not carefully conducted. Compared to XL-MS, HDX-MS provides a less perturbing strategy[127] by using HDX, which highly depends on engagement in hydrogen bonding and secondary structure elements. Thus, it can especially supply local structural information, dynamics and conformational changes[128–131]. Ultimately, this method can identify which and how amino acids are involved in ligand binding and provides information on conformational changes upon complex formation[132].

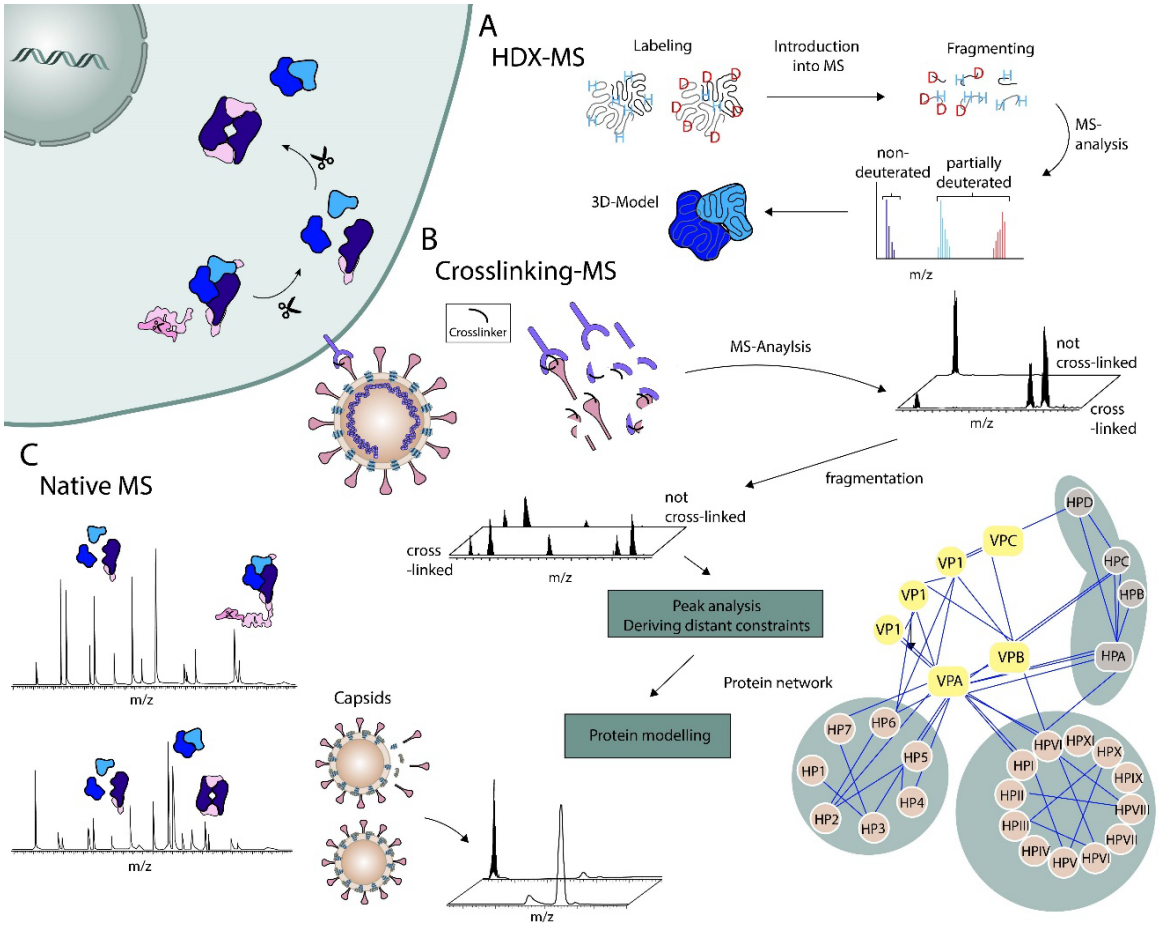


Figure 1 Structural MS for RNA viruses.
All depicted approaches can be conducted as top-down MS whereby HDX-MS **(A)** and XL-MS **(B)** are currently mainly used as bottom-up techniques. For HDX-MS, the workflow starts with labeling of the natively folded protein by exchanging hydrogens to deuterium. The labeled protein is either fragmented or digested to peptide level. Subsequent MS-analyses reveal non-deuterated and partially-deuterated peptides leading to constraints for a 3D model. A similar principle is used in XL-MS experiments **(B)**. First, the protein complex is labeled, which can be done in both ways *in vitro* and *in vivo*, then the sample is fragmented or digested. Distant constraints can be deduced from every successful XL-MS experiment bringing up valuable information for computational modeling and the proposition of a structural model. *In vivo* XL-MS offers the identification of a protein interaction network (bottom right) realizing the ability to unravel important virus-host association. **(C)** Native MS can determine stoichiometries of protein complexes (blue shaded) or measure whole virus capsids.

Processing of the recently emerged SARS-CoV-2 nsp7-11 polyprotein was investigated using integrative computational modelling, incorporating data from pulsed hydrogen-deuterium exchange MS (HDX MS) and crosslinking-MS (XL-MS)[83]. Both techniques provide local structural details, HDX-MS shows local conformational changes based on deuterium incorporation in peptides. Hereby, exposed residues undergo rapid hydrogen-deuterium exchange and shielded areas exchange slower. XL-MS links residues that are in proximity to each other resulting in linked peptides. In both approaches, the proteins are labelled or linked and then proteolytically digested to peptides. After MS detection, peptides can be assigned to the protein sequence, and the labeling provides a basis for drawing structural conclusions. HDX-MS and XL-MS are both bottom-up MS (BU-MS) techniques, in which protein information is reconstructed from identified peptides. How these techniques work and how they are applied in viral research has been explained elsewhere in more detail[133]. Ultimately, this study captured dynamic structural changes upon M^{pro} cleavage and came up with an integrative full nsp7-11 SARS-CoV-2 polyprotein model explaining the observed kinetics. For the integrative modelling, they fed the I-TASSER server[134] with distance constraints from XL-MS data and secondary structure information for nsp7-nsp10 resulting in a few different predictions. They concluded that group B and C models were most consistent with their HDX-MS results showing CS9/10 exposed and most accessible and CS7/8 folded into an α-helix (**Figure 2**)[83]. This study proposed the cleavage order of polyprotein processing using HDX-MS and sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) as follows: CS9/10 > CS8/9 > CS10/11 >> CS7/8[83]. There are indications that there is a temporal regulation in CoV polyprotein processing[84,135]. How and if this is reflected *in vivo* has yet to be demonstrated.

1.1.3. Current techniques that could advance polyprotein research

This section highlights sophisticated techniques that studied CoV nsps or polyprotein processing of other viruses- mostly *in vivo*. Knowledge gaps in CoV polyprotein research and how these could potentially be addressed by current scientific techniques are illustrated.So far, *in vitro* methods have looked at a specific region of the polyprotein, nsp7-10/11, but not the full polyprotein.[83,84] The full polyprotein as a large, multidomain protein is extremely difficult to express. Cell-free systems could solve this problem as e.g. a wheat-germ cell-free expression system was used to produce soluble norovirus (NoV) polyprotein[136]. Native MS can study high masses up to megadalton viral capsids[137], so an *in vitro* assay of full pp1a or pp1ab would in theory be possible, but recombinant expression of full size pp1a/pp1ab has not yet been achieved. Although, membrane proteins can be subjected to native MS[138], these will require adding detergents or lipids limiting resolution. It however appears realistic to look at nsp7-16 excluding transmembrane protein regions and autocleaving proteases. HDX-MS on the other hand will not be limited by a membrane environment[139], instead the sheer amount of different

peptides from the full polyprotein will result in overlaps and likely uninterpretable data. XL-MS is not limited by this and can even be employed *in vivo* and in a time-resolved fashion. Bogdanow *et al*. applied such approaches to strains of virus-like influenza A showing the interactome in virions and using pulsed metabolic labelling and quantitative XL-MS on infected cells[140]. Complex formation of mature polymerase co-factors nsp7 and nsp8 were studied with native MS and compared between α- and β-CoVs. This revealed distinct stoichiometries of nsp7 and nsp8 interaction and showed that, despite high sequence conservation, the quaternary structure is not preserved. Whether this has an impact on RTC stoichiometry or could affect pathogenicity remains to be investigated *in vivo*[69]. Beyond CoV polyprotein processing and nsp complex formation studies, native MS is capable of probing protein-RNA/DNA interactions. A recent native MS study on Lassa virus showed time-resolved ribonucleoprotein (RNP) formation revealing the release of the nucleoprotein from its trimeric storage form via an intermediate bound to a single short RNA. By combining native MS with HDX-MS and structural modelling, Sänger *et al*. were able to reveal the interaction site involving the Z protein, which mediates RNP recruitment to the plasma membrane during the late stages of infection[141]. This and the example in **Figure 2** highlight the strength of combining different techniques. After the first identification *in vivo*, polyprotein processing, cleavage efficiencies, kinetics and order were investigated *in vitro*. Molecular details from *in vivo* experiments would however be desirable. For example, Emmott *et al*. used a cell-based FRET sensor assay to examine the activity of different NoV proteases between different genotypes, providing a comprehensive view of substrate processing and cleavage efficiency[142]. Such a comprehensive cell-based research comparing different genogroups within one study is yet missing for CoVs. The large number of nsps certainly poses a particular challenge here.

The NoV viral protease that processes the polyprotein occurs in several proteoforms, but all NoV protease precursor forms are catalytically active *in vitro*. A study looking at *in vivo* protease activities by using confocal microscopy and FRET labelling showed that these are determined by both substrate and protease localization[143]. Since CoVs have no precursor proteases like NoVs, such a study would not be feasible and would have to be substrate dependent. The individual nsps would have to be labeled with fluorophores, which could influence the cleavage sites or even lead to cleavage failure. Further challenges can be blinking artifacts from single fluorophores, localization uncertainties and detection efficiencies[144].

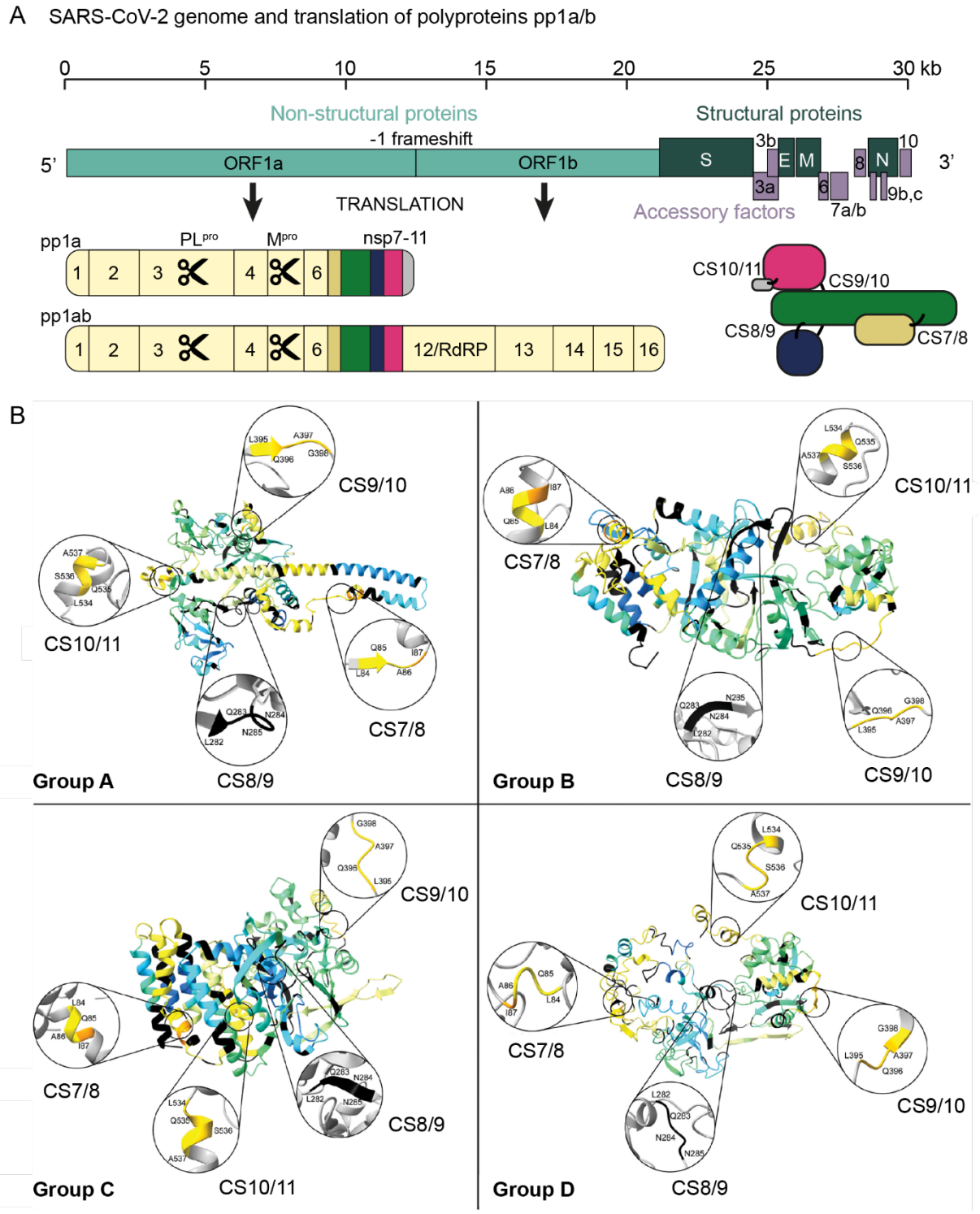


Figure 2 Genome organization and nsp7-11 polyprotein model of SARS-CoV-2.
(A) Schematic illustration of 30 kilobase SARS-CoV-2 genome showing open reading frames 1a and 1ab giving rise to polyprotein 1a and 1ab (pp1a/pp1ab). The two viral proteases, papain-like protease (PL^{pro}) and main protease (M^{pro}) with autocleavage function, are indicated as scissors. Non-structural protein 7-11 (nsp7-11) is highlighted because nsp7-11 has been recombinantly expressed and studied as a coherent polyprotein. A model with a rough arrangement of the nsp domains is shown on the right. The individual nsps are colored as follows: nsp7 (yellow), nsp8 (green), nsp9 (dark purple), nsp10 (pink), nsp11 (grey). **(B)** Yadav *et al*. employed an integrative structural modelling approach using information from structural MS techniques as constraints resulting in four structural groups. The best agreement with the pulsed HDX-MS and SDS-PAGE proteolysis results were found for groups B and C. Secondary structure is colored with 10 s deuterium from HDX-MS analyses. Reprinted from [83], *Copyright © 2022, The American Association for the Advancement of Science*

1.1.3.1. How novel imaging methods could promote polyprotein research

The advent of super-resolution microscopy (SRM), such as STED (stimulated emission depletion), STORM (Stochastic Optical Reconstruction Microscopy) and PALM (Photo-Activated Localization Microscopy), has enabled high spatial resolution through the use of photo-switchable fluorophores[144]. Recent advances in STED allow nanoscale localization, making it possible to monitor viral RNPs during their trafficking along the endosomal pathway in dendritic cells during influenza virus infection[145]. In SARS-CoV-2 infection, the binding mechanism of the structural spike protein was investigated with STORM in more detail[146]. These studies show what SRM is capable of and could contribute to fill the knowledge gaps in CoVs, for example how host immune evasion is mediated by some nsps or how N protein is involved in the RTC template switch during production of subgenomic RNAs[90].

In the series of advanced and innovative microscopy techniques, cryogenic electron microscopy (cryo-EM) has pushed the boundaries of high-resolution imaging enabling structural resolution of heterogenous samples and larger assemblies. Cryo-EM can resolve protein structures and assemblies *in vitro* that could not be captured with X-ray crystallography or nuclear magnetic resonance (NMR). While X-ray crystallography relies on crystallization of proteins, which is particularly difficult for flexible and membrane proteins, NMR has a size limit of ~100 kDa[147]. Therefore, cryo-EM promotes viral research and can fill in knowledge gaps. A challenge is for example solving the surface glycoprotein structures in the fusion state, which are delicate to stabilize, overall flexible and hence in most cases inaccessible by X-ray crystallography. In recent years, several prefusion structures became available for example HIV-1 envelope glycoprotein trimer[148,149], SARS-CoV-2 spike protein[150] or herpes simplex virus glycoprotein B[151]. Although cryo-EM can provide high-resolution structural models of flexible and dynamic proteins, it could not achieve high-resolution reconstruction of the nsp7-10 polyprotein together with M^{pro}[152].

Several structure models promoted the understanding of SARS-CoV-2 RTCs. By using molecular modeling and available structures of all nsps, Perry *et al.* proposed the most complete model comprising a stoichiometry of six copies of nsp7, nsp12, nsp13, nsp14, nsp15, nsp16, twelve copies of nsp8 and nsp10 and two copies of N-protein with leader transcription regulatory sequence oligo (**Figure 3**)[153]. For this RTC superstructure experimentally resolved structures were utilized, which had provided insights into the molecular mechanisms in capping and mRNA synthesis[154–156]. In order to get RTC supercomplex assembly computational docking was applied around a hexameric nsp15 hub (**Figure 3**) [153]. However, the model is not in agreement with the ternary nsp14+16+10 complex[157]. While this model is unlikely, a larger assembly may exist to enable concerted and hence efficient genome replication.

While *in vitro* cryo-EM pushes to higher resolution, there is a rising interest in propelling *in situ* cryogenic electron tomography (cryo-ET). This allowed characterization of SARS-CoV-2 virion assembly, structure and the visualization of budding events[158]. In CoVs, polyprotein subunits are low abundant proteins complicating *in situ* localization with microscopy techniques. However, polyprotein processing research has utilized immunofluorescence confocal microscopy or immunolabeling combined with EM localizing subunits from pp1a in several CoVs. In MHV nsp7, nsp8, nsp9 and nsp10 were localized in the cell periphery by using immunofluorescence microscopy[159,160]. Cryo-ET of infected cells and immunolabeling found viral subunits of pp1a at convoluted membranes (precursors of DMVs) in SARS-CoV-1 and MERS-CoV[161,162]. A further development is the combination of fluorescence microscopy and cryo-EM, which is called correlative light and electron microscopy (CLEM). CLEM allows to research rare events by improving target localization by correlating cryo-sections for cryo-EM using cryo-fluorescence microscopy[149].

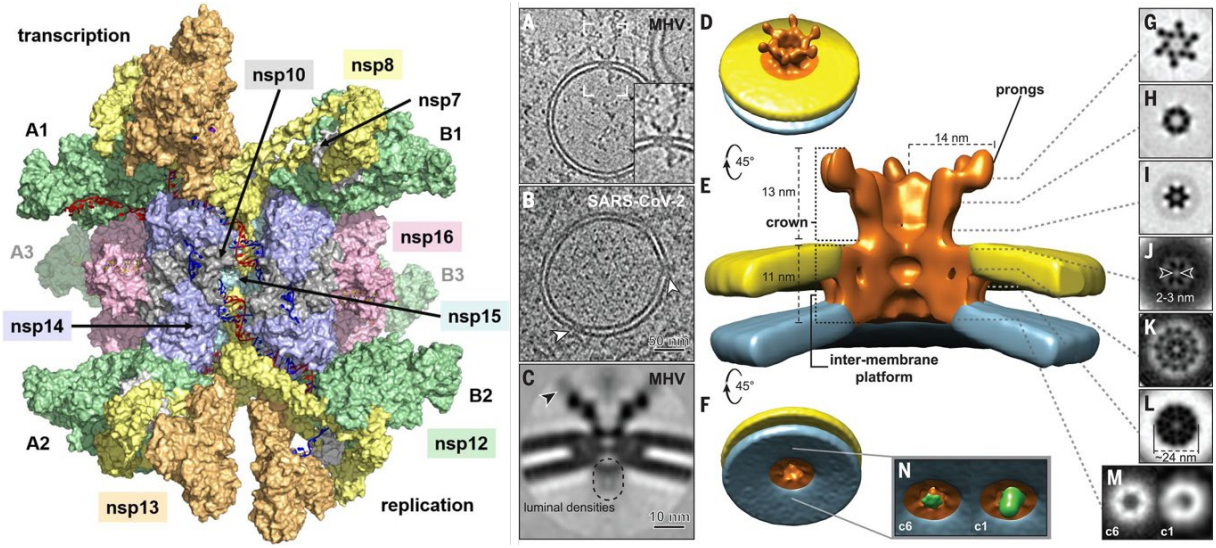


Figure 3: **DMV pore complex and atomistic model of full RTC.**
The left panel shows a model of the complete RTC with an assumed stoichiometry of 6 times nsp7, nsp12, nsp13, nsp14, nsp15, nsp16, 12 times nsp8 and nsp10 and 2 times N-protein. Reprinted from [153]: Copyright © *2021 Published by Elsevier Inc on behalf of American Society for Biochemistry and Molecular Biology.*
The right panel: The architecture of the molecular pore is shown, with tomographic images of DMVs induced by MHV (**A**) and SARS-CoV-2 (**B**). Panel shows the middle slice through the average, suggesting flexible regions near the prongs (black arrow) (**C**). Different views of a 3D surface-rendered model of the DMV pore complex probably consisting of nsp3 is shown. Inner and outer membrane of the DMV is colored in blue and yellow, respectively (**D-F**). Reprinted from [163]: *Copyright © 2020 American Association for the Advancement of Science. No claim to original U.S. Government Works.*

A remarkable cryo-CLEM study revealed six copies of a viral protein, by size most likely nsp3, spanning both membranes of the DMVs as key sites for viral RNA synthesis export in MHV and SARS-CoV-2 (**Figure 3**). Focused ion beam milling of cells and tomographic slicing increased resolution and allowed a 3D surface-rendered model of the pore complex[163]. Despite the advances in microscopy techniques studying low abundant molecules remains challenging

for *in vivo* methods or *in situ* approaches. Therefore, capturing the dynamics of low abundant intermediates in CoV polyprotein processing with cryo-CLEM would still be very challenging.

1.1.3.2. How state-of-the-art mass spectrometry could advance *in vivo* investigations

Polyprotein processing research has benefited equally from *in vitro* and *in vivo* techniques. Now that the techniques are becoming more complex and often more laborious, it remains important to continue down this path. For example, *in vivo* BU-MS can provide a deep host-interaction network illustrating multiple virally perturbed signaling pathways in a cell. Then pathways need to be looked at in more detail. On the one hand, this can mean looking more closely at specific signaling pathways that are mediated and perturbed by a viral protein *in vivo*. On the other hand, it can also mean, for example, studying viral-host protein interactions *in vitro*.

In vivo BU-MS cannot only provide a general idea of global molecular abundances, it also allows the investigation of PTMs and protein turnover[133]. By using cell culture systems and comparing infected to non-infected cells, global perturbations can be quantified with BU-MS. Through this approach, it was shown that SARS-CoV-2 - unlike SARS-CoV-1 - disrupts the transforming growth factor-β pathway, a key player in tissue fibrosis[164]. The study provides a broad overview about host-pathogen interactions of SARS-CoV-2 and highlights biological relevant pathways for more in-depth studies or as potential drug targets. A recent report details the use of high-throughput proteomics and metabolomics in SARS-CoV-2, which illustrates how important this technology is for investigation of pathology[165]. A study in NoV showed how stable isotope labeling by amino acids in cell culture (SILAC)-based quantitative proteomics can be used to investigate infection and translation inhibition of NoV protease NS6. Emmott *et al.* provide a detailed *in vivo* investigation of NS6 interacting and perturbing eukaryotic initiation factors[166]. However, in the context of CoV polyprotein processing, the discovery of individual interactors of nsps or polyprotein intermediates using these approaches is challenging.

Processed intermediate and mature nsp products represent variants arising from the same gene and are considered proteoforms. Proteoforms in processing can be indirectly studied by a new emerging method called *N*-terminomics. Proteins are often modified at their *N*-terminus, for example through acetylation, which can alter their activity or confer distinct properties. Since *N*-terminal acetylation usually takes place co-translationally at the ribosome, polyprotein processing, which is a post-translational event, produces non-acetylated termini. These are coupled to a bait in *N*-terminomics, which can then be enriched by affinity purification. A study in Zika used *N*-terminomics to show how inhibitors impair Zika virus replication. For example, first, they observed no reduction of viral RNA levels upon Regorafenib treatment, and then

they analyzed viral protein synthesis levels. They observed a decrease in structural protein E, but not in NS1 levels, which are both internal parts of a seven-membered polyprotein. In Flaviviruses, NS1 is one of seven proteins that are expressed in a polyprotein and subsequently processed. By using a translation inhibitor and time-dependent measurements, they could observe a block in NS1 secretion[167,168]. Such *N*-terminomics allow *in vivo* investigation and monitoring of nsps and would be ideal for polyprotein processing research in CoVs.

One of the bottlenecks in *in vivo* proteomics is the coverage of peptides and in turn proteins. Thus, emerging microfluidic technologies and improvements in chromatographic separation are striving to enable deeper coverage of complex *in vivo* proteomes and corresponding interactomes. This includes low-input proteomics, low abundant proteins and the detection of PTMs and potential proteoforms. A common approach is to use tandem mass tag (TMT)-labelling allowing quantification of detected species across different conditions or in single cells[169–171]. A study on SARS-CoV-2 infected cells used TMT-labelling and RNA antisense purification coupled with MS (RAP-MS), an *in vivo* technique that identifies proteins interacting with RNA. Not surprisingly, 10 of 16 mature nsps were found to directly interact with RNA, for example nsp12 and nsp8[172]. Here, it would be interesting to know whether nsp8 alone can bind RNA or if it only binds RNA within the RTC. This information gets lost in BU-MS, but native MS could reveal this. At this point, it could be worthwhile to leave the *in vivo* context and validate such hypotheses *in vitro*. Another way to deepen proteome analysis is through *in silico* enrichment approaches, for example GO-enrichment on the SARS-CoV-2-host contactome[169].

Top-down MS (TD-MS) techniques such as native MS have been mentioned several times. TD-MS approaches commonly subject intact denatured proteins to liquid chromatography (LC)-MS setups allowing PTM characterization and quantification with high-throughput. So far TD-MS approaches have rarely been applied in viral contexts in contrast to BU-MS approaches. Given the ‘peptide-to-protein’ inference problem in BU-MS, TD-MS is in general better suited for studying PTMs and hence polyprotein processing.

First, we want to look at some examples of *in vivo* TD-MS. Although these studies are outside a viral context, they highlight the powerful potential of this approach. A study investigated myofilament and Z-disc protein phosphorylation after acute infarction and found a significant reduction in three crucial myocardial proteins. Furthermore, they were able to pinpoint phosphorylation sites by sequencing using multiple fragmentation techniques[173]. In another example using a single-cell TD-MS proteomics approach, PTMs and isoform expression levels between individual muscle cells were compared and showed significant heterogeneity at the proteoform level. Additionally, Melby *et al.* were able to detect multiple isoforms of myosin

heavy chain with a mass of ~ 223 kDa[174]. This highlights the great potential of this technique to investigate phenotypic heterogeneity and functional diversity.

Using TD-MS in a viral context could mean to investigate PTM levels in CoV-infected cells and compare the results to BU-MS studies revealing novel and unambiguously identified proteoforms (i.e. combinations of PTMs). Moreover, detection of the large coronaviral RTC directly from infected cells would advance CoV research and the development of antiviral therapeutics. According to the atomistic model of soluble RTC from Perry *et al.*, this complex would be approximately 4.7 MDa[153]. Compared to the detected ~ 223 kDa myosin heavy chain this has much higher mass and would therefore pose an instrumental challenge. The study from Perry *et al.* neglects the membrane association of the RTC likely through the nsp3 pore. Furthermore, RTCs produce a nested set of subgenomic and genomic RNAs, which may not be performed by one unique complex. Hereby, a native MS approach is particularly suited to reveal the unknown compositions. Currently, there are no examples for detection of such a large complex directly from infected cells by native TD-MS.

In vitro native TD-MS reduces the complexity of the mass spectra due to less cellular background and allows a detailed look at, for example, extensively glycosylated S proteins from SARS-CoV-2. In the S-RBD, eight O-glycoforms were unambiguously identified and their relative molecular abundance quantified. This proteoform analysis serves as a blueprint for the investigation of O-glycoform heterogeneity for SARS-CoV-2 S-RBD variants[175]. Later, this approach was extended to the Omicron variant providing proteoform alterations of O-glycoforms and suggesting how Omicron escapes immunological protection[175]. In this way, it would be possible to monitor intermediates together with their PTMs during the polyprotein processing. This would require mammalian expression systems, which are more likely to result in successful expression of the whole polyprotein pp1ab, and protease inhibition. Alternatively, the whole polyprotein could be extracted from infected cells using immunoprecipitation and protease inhibitors. Immunoprecipitation would help to obtain a large, biologically relevant polyprotein with natural PTMs. A time-dependent study could shed light on unknown modification processes, which would be particularly informative when co-immunoprecipitation is used. Furthermore, co-immunoprecipitation would allow the retrieval of biologically relevant interactors.

1.2. Native mass spectrometry

The principle of a mass spectrometer is to measure the mass-to-charge ratio (*m*/*z*) of an ionized analyte in gaseous form. Therefore, a basic set-up of a mass spectrometer comprises a sample introduction system, an ion source, a mass analyzer, and a detector, all of which are operated *in vacuo* except for the first component.

Over the last two decades native MS has been an emerging field leading to a technique that is nowadays eminently suitable for the analysis of biomolecular issues. It allows far more than just molecular mass determination yielding relevant information for structural biology, protein-protein, protein-ligand interaction, and protein complex stoichiometry. Advances in sample preparation and various components of the mass spectrometer, such as mass analyzers, have contributed greatly to this development[176].

1.2.1. Sample requirements

The most common biological samples are recombinantly expressed proteins. To obtain the best spectra with high quality and well resolved peaks, the sample should ideally have minimal levels of non-volatile salts, contaminants and co-expressed proteins. This is particularly important for complex mixtures. Since protein purification is usually conducted with saline buffers providing better ionic strength and buffering properties than volatile buffers, buffer exchange needs to be performed prior to MS measurements. Depending on the stability of the analyte in MS compatible buffers, buffer exchange can be combined with the last purification step e.g., size-exclusion chromatography (SEC). Otherwise buffer exchange is commonly performed by using centrifugal filters, dialysis and desalting columns. For rapid identification of overexpressed proteins, it is possible to spray from crude cell lysate. However, most application target protein binding studies, in which co-expressed proteins and other contaminants could be disruptive[177].

Native MS compatible buffers should meet requirements such as being non-denaturing, providing ionic strength and having a suitable pH. Ammonium acetate solution is the most widely used allowing an adjustable pH range from four to ten and an ionic strength from zero to several moles per liter providing native-like conditions. However, it should be noted that ammonium acetate is a solution at neutral pH and buffers the pH at 4.75 ± 1 (acetic acid) and 9.25 ± 1 (ammonia).[178] Although non-volatile salts are usually avoided, small quantities of co-factors or bivalent cations can be tolerated and added, which can be e.g. crucial for enzymes. Generally, in the presence of non-volatile salts, the smaller the diameter of the electrospray ionization (ESI) emitter tip, the better the prevention of adduct formation[179]. Membrane proteins require mimicking of the native lipid membrane, which is aimed at by the

addition of detergents that form micelles or designed systems such as nanodiscs or bicelles [180–182].

Furthermore, there are alternative buffers that can be used instead of ammonium acetate. Ammonium bicarbonate provides higher ionic strength, but can lead to unfolding of the protein at neutral pH due to outgassing of CO_2 and subsequent foaming upon ESI process[183]. Thus, ammonium bicarbonate should be used carefully or not at all. Triethylammonium acetate was employed as an additive for charge reduction and Gabelica and Marchand used trimethylammonium acetate to suppress unspecific cations adducts that were used for triggering G-quadruplex formation[184–186]. Triethylammonium bicarbonate can have benefits when analyzing macromolecular species of similar mass[187]. Zhuang and coworkers compared alkylammonium acetate buffers with ammonium acetate and concluded that they can be an alternative with slightly different properties e.g. reducing the charge[188].

Buffer choice and molarity have effects on analyte stability and peak resolution. Another parameter targeting improved quality of mass spectra is to find the best concentration of your analyte. In general, the higher the concentration of your analyte, the higher the signal. However, the higher the concentration, the higher the risk of nonspecific artifacts formation upon ESI process. Therefore, the sample concentration should be evaluated and the varying ionization efficiency depending on the proteins should be considered.

1.2.2. Nano-electrospray ionization

The invention of soft ionization technique such as matrix-assisted laser desorption ionization and ESI have revolutionized the field of mass spectrometry. These techniques opened up the investigation of a larger variety of compounds including larger biomolecules as both of them only cause little or no fragmentation[189–191].

In ESI, the analyte is injected into a mass spectrometer from an electrically conductive emitter by applying a potential of several kV. ESI can be applied in both positive and negative ion mode. In this work positive ion mode is used, which is why this mode is considered in the following. Conventional ESI sample consumption is low and in the range of one and several hundred $\mu L/min^{-1}$, whereas nano-ESI using glass capillaries with diameters below 4 µm resulting in even lower flow rates of 20-50 nl/min. The initial droplet size of nano-ESI is an order of magnitude smaller than conventional ESI. As a consequence, potential high amounts of non-volatile salts are less concentrated[192].

The mechanism how an analyte is transferred from solution into gas phase was unraveled experimentally and computationally via MD smulations[193,194]. The current is applied and the ESI source acts as the anode and the mass spectrometer as the cathode. Upon the high voltage protons are generated at the metal/solution interface inside of the emitter (e.g.,

2 $H_2O \rightarrow 4\ H^+ + 4\ e^- + O_2$). A so-called Taylor cone is formed from which charged droplets are ejected and evaporate rapidly (**Figure 4 A**). Generated droplets carry a positive charge due to the presence of excess ions such as H^+, NH_4^+ and Na^+. These droplets can only carry a certain maximal number of the same charge depending on size and surface tension. Progressive evaporation leads to a scenario where the electrostatic forces of like charges (Coulombic repulsion) are balanced by the force of surface tension. This scenario is called Rayleigh limit (**Equation 1**) and the maximum amount of charge (z_R) can be predicted depending on surface tension (γ), the electrical permittivity of the vacuum (ε_0), the radius (R) of the droplet and the elementary charge (e)[192,193,195].

Equation 1

$$z_R = \frac{8\pi(\gamma\varepsilon_0 R^3)^{\frac{1}{2}}}{e}$$

Prediction of the Rayleigh limit can be simplified (**Equation 2**) by the assumption that the radius of the protein corresponds to its molecular weight and its density equals the density of the water[190].

Equation 2

$$z_R = 0.078\sqrt{M}$$

Droplets at Rayleigh limit yielding a fine mist of droplets via jet fission. Thus, it can be discriminated between initial droplets with radii of a few micrometers and final droplets with radii of a few nanometers. During the ESI process, these cycles of evaporation (initial droplets) and fission (final droplets) are repeated until gaseous ions are obtained that are detectable by MS[192,193].

There are three models that propose different ejection mechanisms depending on the size and shape of the molecule. The ion evaporation model (IEM), charged ejection model (CEM) and the charged residue model (CRM) (**Figure 4 B**). IEM is proposed for low molecular weight species or inorganic ions. Small solvated ions are pushed out to the surface where the repulsion of excess droplet charge is pushing them further away, though still connected via a string of ‘sticky’ molecules. As repulsive forces build, the connecting string breaks, releasing the ion. Natively folded proteins shrink within the nanodroplet during this process. Continued evaporation and fission, consistent with the IEM, reduce the charge until the final shell vanishes and the remaining charge transfers to the analyte (**Figure 4 B**).

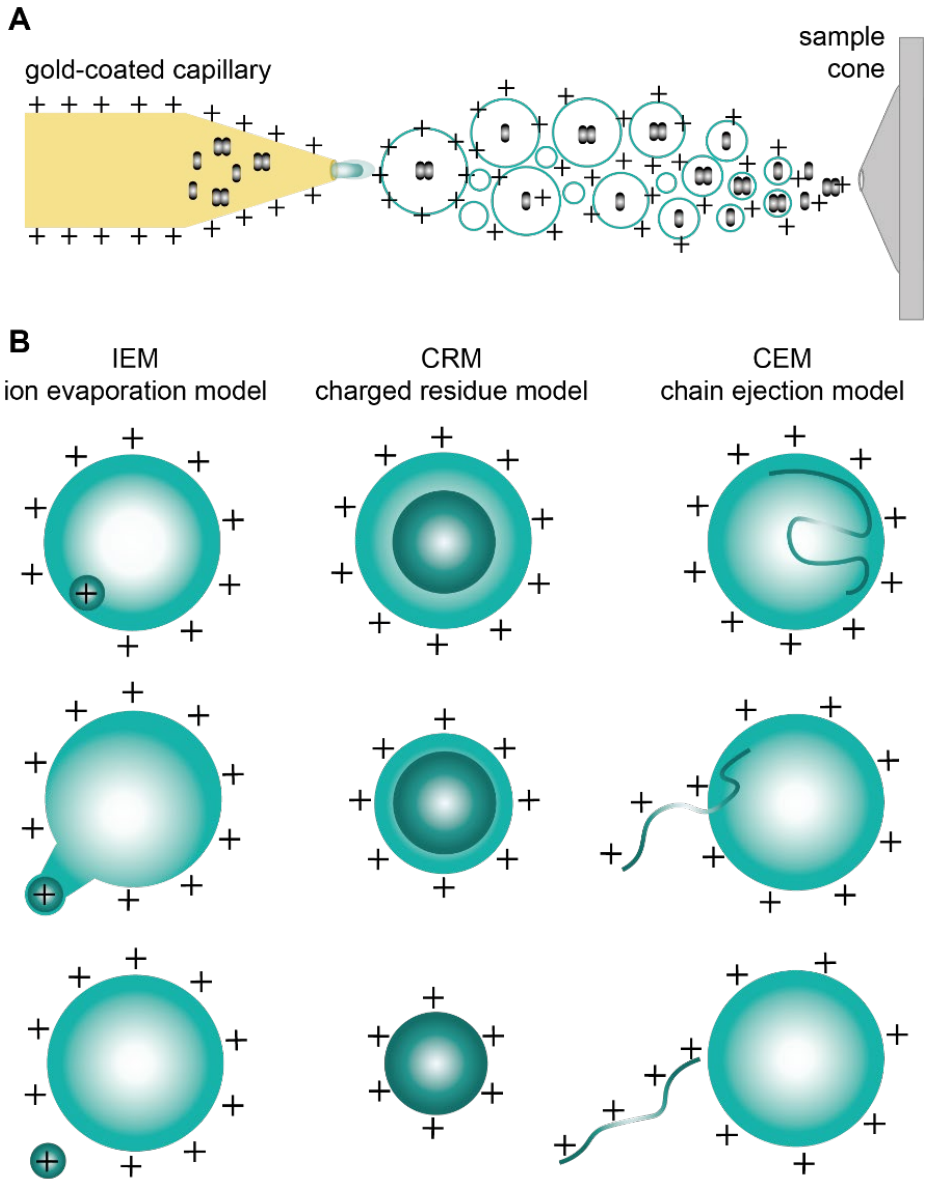


Figure 4: Molecule ejection models of electrospray ionization (ESI).
The application of a high electric potential triggers the formation of the Taylor Cone at the tip of the gold-coated ESI capillary. Positively charged droplets are ejected and undergo evaporation and fission cycles. Finally, the dried out, "naked" analyte reaches the cone and is introduced into the mass spectrometer (**A**). There are three proposed models for the generation of intact ESI analytes: IEM for small molecules, CRM for large, globular proteins and CEM for hydrophobic polymer chains (**B**).

The CEM model applies to polymer chains and proteins under denaturing conditions. Hereby, hydrophobic residues are exposed and the unfolded protein migrates to the droplet surface as its residence within the droplet is unfavorable. Thus, the CEM model comprises protein ejection with a protein-droplet charge equilibrium. First, only a part of the chain is expelled from the droplet and then the protein is stepwise ejected into the gas phase. For natively folded, globular proteins, in which hydrophobic residues are oriented to the interior and thereby making the protein hydrophilic, an alternative model is proposed (**Figure 4 B**). The CRM model is the dominating model for protein ejection into the gas phase in native MS.

Soluble-like proteins exhibit rather lower intensities in native MS compared to denatured proteins. In the case of myoglobin, a native mass spectrum showed nearly two orders of magnitude lower signal intensities[193]. Further, signal degradation increases due to spectral noise and nonspecific adducts. ESI-MS but especially native MS is susceptible to reduction of the signal-to-noise ratio, e.g. by sodiation. Instead of simple protonation ($[M + z_R H]^{z}_{R}$), sodium adducts bind carboxylates on the protein during the end of evaporation and fission cycles raising the following ion species $[M + (z_R - i) H + iNa]^{z}_{R}{}^{+}$, where $i = 0, 1, \ldots, z_R$. The generation of ion species plus one or more nonspecific adducts causing heterogeneity of mass spectra and thereby reducing signal-to-noise ratio[193,196].

1.2.3. Protein folding in the gas phase

The basis of native MS is that protein conformations in aquatic environments are preserved during transfer to the gas phase, allowing biologically relevant protein states to be analyzed. Aqueous environment stabilize protein folding and reduce coulombic interactions. In the gas phase, this hydrophobic effect is absent. It would therefore be logical that proteins would unfold. Surprisingly, early MS analysis of protein-protein interactions, driven by non-covalent interactions, showed that the proteins must adapt a somewhat protein folding allowing protein-protein interactions[197,198]. Several studies suggested the successful transfer of specific solution-phase interactions to gas phase structures and therefore the preservation of non-covalent interactions of protein complexes. For example, avidin (64 kDa) was shown as a tetrameric complex with no present trimer or pentamers, and concanavalin A depicted the known in-solution states: dimer and tetramer[199,200].

The question is therefore, to what extent gas-phase structures resemble in-solution structures. Electron capture dissociation was used to study the unfolding of ubiquitin after transfer to the gas phase. It was shown that unfolding occurs to different extents over time, depending on salt bridges and electrostatic interactions that compensate for the loss of hydrophobic interactions[201]. Ion mobility spectrometry (IMS) separates ions of the same m/z based on their conformational shape. Hereby, IMS quantifies ion species by their collision cross section, which can be directly compared to cross sections from in-solution methods allowing a quantitative assessment of in-solution and gas phase structures. IMS on ubiquitin showed similar collision cross section as in solution. Further, ubiquitin unfolding kinetics within the gas phase require over hundreds of milliseconds to reach a more stable conformation[202]. Thus, evolution of ion species from solution structure to minimum gas phase structure take longer than the time *in vacuo*.

In general, further IMS analyses of standard proteins such as cytochrome c, myoglobin and ubiquitin were in accordance with results from in-solution techniques[176,197]. There are some

exceptions, in which proteins depict non-native-like structures i.e. either unfolded or collapsed upon coulombic repulsion[203–205]. However, the majority of these studies showed a native-like folding in the gas phase.

In summary, metastable conformations, which are theoretically close to in-solution structures, are detected, since ESI experiments last only a few milliseconds[197,201,202]. However, gas phase structures start to partially unfold due to the lack of hydrophobic forces and collapses can occur in rare events[197,203]. Thus, it can be concluded that native MS allows the study of proteins and protein complexes. In-solution conformations are largely preserved and the gas-phase structure reflects many crucial aspects of the native structure.

1.2.4. Peak analysis

Based on the different nano-ESI models, it follows that shape and folding of the protein affects the number of charges that will be transferred. Hereby, more elongated proteins will uptake more charges than globular folded proteins resulting in mass spectra of natively folded proteins and protein complexes that have unique characteristics. This is because of the larger solvent exposed surface area[206,207]. In accordance with the CRM model, globular folded proteins only carry a number of few charges and the charge state distribution follows a gaussian-like distribution, in which the highest peak reflects the most abundant charge state[193,208]. For example, the monomer peak with eleven charges and the dimer peak with 17 charges are the most abundant peaks in the native mass spectrum of SARS-CoV-2 M^{pro} (**Figure 5**).

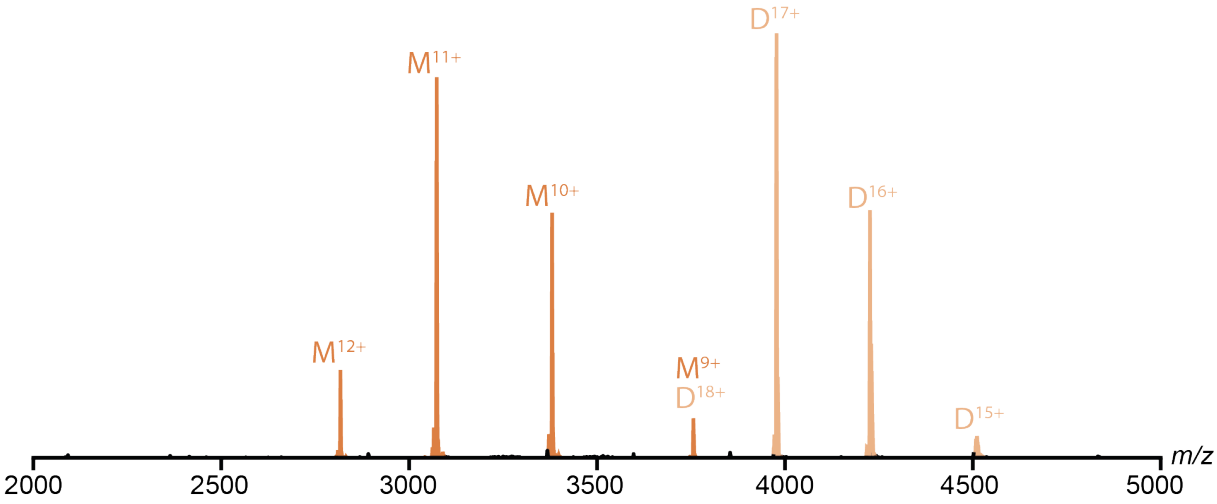


Figure 5: Exemplary native mass spectrum of SARS-CoV-2 M^{pro}.
Monomer species (burnt orange) and dimer species (light copper) are detected in parallel.

There are several ways to determine the molecular mass from the *m*/*z* values given by a mass spectrum. The three key methods are charge state recognition from isotopic patterns, charge state recognition from charge state envelope analysis or from adduct ion analysis. The isotopic approach requires peak resolution, so this approach is often used for peptides[209]. If peaks are *m*/*z* = 1 apart, then the charge is *z* = 1, if the peaks are *m*/*z* = 0.5 apart, then the charge is

z = 2 and so on. Adduct ion analysis requires looking for common ion adducts such as sodium, potassium, or ammonium ions. Once the charge is determined the mass calculation follows the same principle (c.f. **Equation 3**).

In the following the charge state envelope analysis will be explained in more detail as it is the one that mainly utilized later in this work. First, the peaks representing the same molecular species with different charges must be identified. The adjacent ion signals of this peak series have a charge difference of $\Delta z = 1$. The higher m/z value $(m/z)_1$ has one fewer charge than the lower m/z value $(m/z)_2$ (cf. **Figure 5**). Thus, number of charges (n) of adjacent peaks can be expressed as $n_2 = n_1 + 1$. Assuming that all charges are caused by protonation ($[M + z^*H]^{z+}$) leads to the following **Equation 3** :

Equation 3

$$\left(\frac{m}{z}\right)_1 = \frac{M + n_1 m_H}{n_1}$$

Substituting the expression for the charge relation of adjacent peaks results in the following:

Equation 4

$$n_1 = \frac{\left(\frac{m}{z}\right)_2 - m_H}{\left(\frac{m}{z}\right)_1 - \left(\frac{m}{z}\right)_2}$$

The molecular mass can be determined once the charge (n_1) of the higher m/z value is determined. While the approach to determine the charge n_1 is different, the calculation of the molecular mass is the same for all approaches[209].

Equation 5:

$$M = n_1\left(\left(\frac{m}{z}\right)_1 - m_H\right)$$

While the uptake of adducts can be exploited for charge determination, it can be a challenge in data analysis since the additional mass needs to be considered in the formula for mass calculation. Most MS software can handle ion adduction with peak smoothing, peak centering, and baseline correction. These programs allow peak deconvolution resulting in a spectrum that depends on mass[210]. Despite newly developed programs such as UniDec[210], the analysis of very complex native mass spectra remains a challenge and requires manual inspection of the spectra. For example, misleading scores (DScore) can occur due to overfitting, it is therefore advisable to check especially lower Dscores manually[211]. Oftentimes, low abundant ion species have bad confidence scores. In addition, corresponding adduct peaks, e.g. $M + Na^+$, and varying peak widths depending on the analyte should be considered accurately. Peaks of

folded proteins can span several *m*/*z*, hence retrieving the intensity from the area under the curve (AUC) is more precise than considering only peak heights[209,212].

Native MS is a foremost qualitative and semi-quantitative technique. Quantitative analysis must be carried out carefully, as ion efficiencies can vary, especially when proteins differ greatly in size and mass[213]. When ionized proteins are similar, ion intensities can be assumed to be proportional to concentration reflecting the solution phase. Thereby, calculation of relative binding constants can be determined. Determined dissociation constants (K_D) has been shown to be similar to other techniques such as isothermal titration calorimetry and equilibrium thermal denaturation for proteins and nucleotides[214,215].

1.3. Mass analyzer

To determine the mass of molecules, mass analyzers utilize the *m*/*z* of ionized molecules using different approaches. How species especially proteins are ionized in ESI has been explained in detail above. This section is intended to provide a brief overview of mass analyzers before looking at individual ones in more detail below.

In general, a mass analyzer employs a two-step process to determine the mass, namely separate ions based on their *m*/*z*, followed by ion detection. The main principle is that a mass analyzer selectively isolates ions so that ions of a specific *m*/*z* can be detected at a given time[216]. Common mass analyzers are compared in

Table 1 considering their speed, resolution and the mass range. The different properties are more or less suitable for certain applications. Typical applications are listed under 'application'. Looking at the physical principles of how ions are separated, four different approaches can be distinguished: electric fields (e.g quadrupole (Q), orbitrap), magnetic fields (e.g. Fourier-transform ion cyclotron resonance (FT-ICR)), flight time (Time-of-Flight (TOF) analyzer) and ion-trapping (e.g. ion traps, FT-ICR)[217]

Table 1: Common mass analyzers are compared.
Each type has its unique ion separation principle[209,218].

Type	Resolution	Speed	Mass Range	Application
Quadrupole	Moderate	fast	Low/moderate	Routine analysis, quantitative work
TOF	High	Very fast	Wide	Proteomics, large biomolecules
Magnetic Sector	Very high	Slow	Wide	Isotopic studies, precise work
Orbitrap	Ultra-high	Moderate	Moderate	Structural elucidation, complex mixtures
FT-ICR	Ultra-high	Slow	Moderate/wide	Detailed molecular analysis

Qs use oscillating electric fields to stabilize specific ion trajectories and thereby filtering ions of selected *m*/*z*. In the next section, the set-up and working principle of Qs are described in more detail.

Ions with the same charge, but different mass have distinct velocities[218]. Assuming that the ion was at rest, the velocity (v) of a mass of an ion (m_i) after acceleration through an electric field

(U) can be mathematical described as shown in **Equation 6**. TOFs utilize the velocities to separate ions in a drift tube, in which lighter ions travel faster[219].

Equation 6: Velocity of an ion after acceleration in an electric field.

$$v = \sqrt{\frac{2ezU}{m_i}}$$

A similar principle is used from magnetic sectors. Ions are bent in a magnetic field, in which heavy ions curve less than lighter ions. So, the degree of curvature depends on *m*/*z*. A magnetic field is also employed in FT-ICR[209]. Here, ions are excited into different circular orbits perpendicular to the magnetic field and then they circulate based on their cyclotron frequency. Based on the circulating ion, an image current is produced. This signal is translated into a mass spectrum using FT[209,220]. In Orbitraps, ions oscillate in a static electric field and the oscillation frequency correlates to *m*/*z*[217,221]. Orbitraps are described in more detail in 1.3.2.

Followed by the ion separation, the ion signals need to be detected, amplified and processed. Common detectors include electron multipliers, Faraday cups, and microchannel plates. These work in a similar way to electron multipliers[217]. In many systems separation and detection are two distinct steps for instance in Qs and TOFs[209,222]. In contrast to Orbitraps, which integrate both steps seamlessly simplifying the system and improving precision.

Mass analyzers can be combined and sequentially implemented into mass spectrometers. Instruments for native MS often combine Qs with another mass analyzer such as a TOF analyzer or an Orbitrap[218]. Since a mass spectrometer equipped with Qs coupled to an Orbitrap was used in this work, these two will be discussed in the next section as ion separators and the latter as ion detector.

1.3.1. Quadrupole

A Q comprises four parallel cylindrical rods that are symmetrically arranged around a central axis and are about 20 cm long. These rods consist of conductive material like stainless steel and are paired opposite of each other[222]. The opposing pairs are electrically coupled, so either a positive or negative potential is applied (cf. **Figure 6**). The applied potential is a combination of a direct current (DC) and a radio frequency (RF) generating an oscillating electric field. Thus, ions entering to the Q travel through the oscillating electric field. Whereas RF voltages alternate and thereby attract and repel ions, the DC gives a steady force. The combination of both stabilize the ion trajectory[209,222]. As a mass analyzer, voltages in the Q are applied in a way that only ions of a specific *m*/*z* are stabilized, so that other ions with different *m*/*z* are ejected. Depending on how the voltages are applied, Qs can be used as a mass filter. When DC and RF is alternated, but the ratio kept constant, ions within a certain range of *m*/*z* values are

transmitted, the others are non-resonant and will be ejected due to unstable trajectories (cf. **Figure 6**)[222].

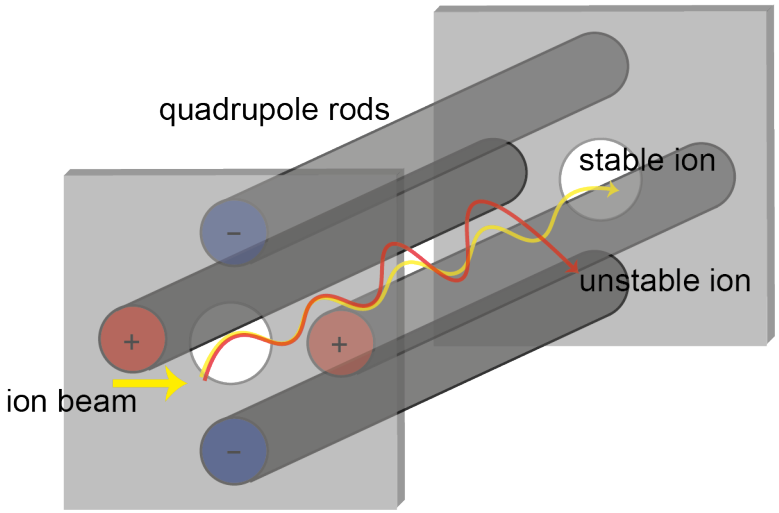


Figure 6: Representation of a quadrupole scheme.
The ions (yellow) enter the quadrupole and the applied voltages of the electronics rods trigger ion oscillation. Only ions with a certain *m*/*z*-value are stabilized (yellow), unstable or non-resonant ions (red) are ejected.

1.3.2. Orbitrap

Orbitraps consist of three electrodes: two outer electrodes and one central electrode. The spindle-like central electrode aligns the two outer dielectric electrodes, together they generate an electric field (**Figure 7**). First ion packets are injected into the orbitrap tangential to the central electrode. Two opposing effects, the radial electric field and the tangential velocity, forces the ions into complex spiral trajectories around the spindle[221,223,224]. Initially, ions do not exhibit well-defined radial motion. However, under appropriate conditions, they tend to stabilize at a specific radial distance, forming a circular trajectory illustrated as ion rings in **Figure 7**. At the same time, they continue to oscillate along the axial direction, which is initiated by the conical shape of the central electrode. Since ions of different masses oscillate at different frequencies both axially and radially, orbitraps can detect different ion species[224,225]. The outer electrode detects these oscillations as an image current at a certain time. Image currents and time information are then amplified and converted into *m*/*z* signals by Fourier transformation. The dynamic range of an orbitrap spans ca. 3-4 orders of magnitude in a single scan, which means that ion abundancies differing by a factor of 1,000- 10,000 are reliably detected[225,226]. Orbitraps mass analyzer allow high mass resolution and accuracy without the need for a superconducting magnet, making high-resolution MS more widely available[221]. FT-ICR mass spectrometers require a lot of space, well-trained personnel and extensive maintenance. The acquisition and energy costs are also high[209]. Instruments optimized for high mass species even allow the analysis of large biological complexes or even virus particles[227–231].

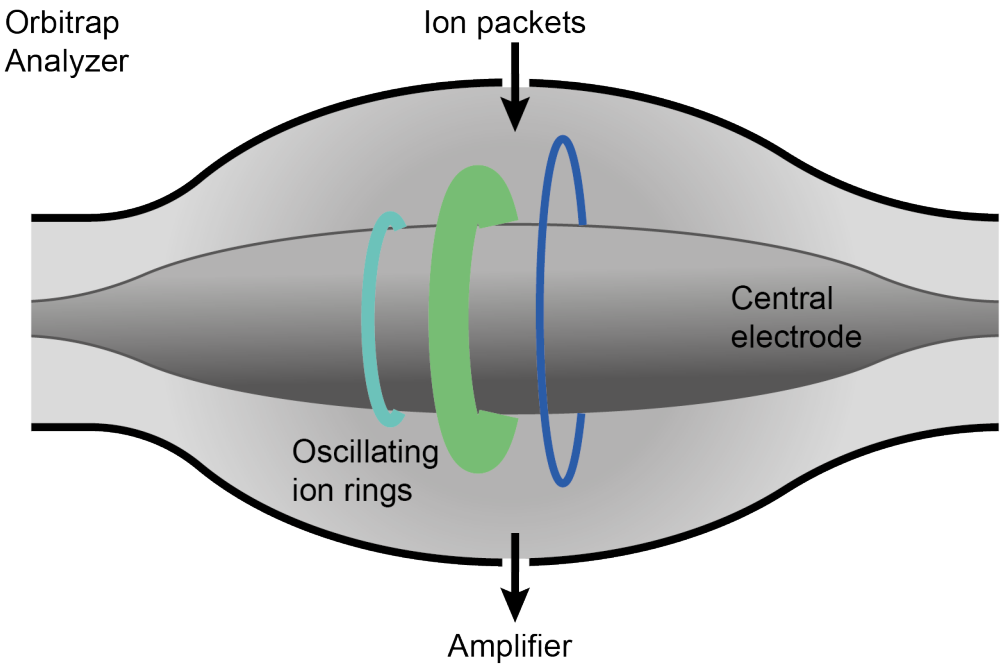


Figure 7: Schematic illustration of Orbitrap mass analyzer.
Nearly circular ion trajectories of three different sized ion species are illustrated in cyan, green and blue. Ions are sent in as ion packets and their oscillation is detected by the outer electrode. Detected signal is then amplified.

1.3.3. Protein complex dissociation

A major advantage of native MS is the ability to detect all species with their stoichiometries in a complex mixture in parallel. To unambiguously assign oligomers or possible overlapping peaks of two species, the ability of dissociation is indispensable. Hereby, a specific peak, the precursor peak, can be isolated and protein complex components can be dissociated by performing collision induced dissociation (CID). In CID, neutral gas atoms such as N_2, He, Ar or Xe collide with the isolated complex transferring kinetic energy, which is partly converted to vibrational energy or internal energy. After a certain number of accumulating collisions, the elevated internal energy within the complex expels the subunit. Collisional activation induces unfolding due to increasingly higher internal energy, and therefore the ejected subunit carries more charges than the remaining, larger part of the complex[199,232]. Hence, the ejected subunit occurs in the lower *m*/*z* region, whereas the residual complex loses charges due to the repulsion and can be detected at higher *m*/*z*.

CID experiments also contain indirect information on protein complex topology. Subunit dissociation is influenced by three key factors. Firstly, smaller subunits are ejected more readily than larger ones. Secondly, subunit located at the periphery are more likely to dissociate early as they generally unfold more easily, although this also depends on their intrinsic structural stability. Thirdly, the overall interaction surface area contributes to complex stability. Thus, early subunit ejection may indicate low binding affinities due to small interaction interfaces. For example, in **Figure 8** the heterotetrameric complex $nsp7_2+8_2$ is isolated and exposed to 75 eV and 100 eV. Subsequently nsp7 and nsp8 subunits dissociate. Nsp7 is more than two times

smaller than nsp8 and is ejected first at 75 eV. Thus, nsp7 sits at the periphery and is not embedded by two nsp8 molecules. When the energy is increased to 100 eV, nsp8 is ejected in addition to nsp7. Therefore, there are two dissociation pathways, which either indicate similar binding contribution of nsp7 and nsp8 or indicate different gas-phase conformations of the heterotetramer[114].

To be able to perform MS^n experiments a combined set-up is required that allows at least two-step mass analysis (MS^2). A common set-up is a collision cell upstream of the second mass analyzer, in which a constant pressure is set and that is filled with neutral gas. The level of dissociation is usually controlled via an adjustable voltage that accelerates the ions entering the collision cell. Depending on the speed ions undergo different numbers of collisions with the inert gas while passing through the collision cell.

In orbitrap-based instruments, collisional activation proceeds in a separate collision cell, a higher-energy C-trap. Higher-energy C-trap dissociation (HCD) follows the same principle like CID, though the energy deposition is considered to be higher compared to CID[233,234].

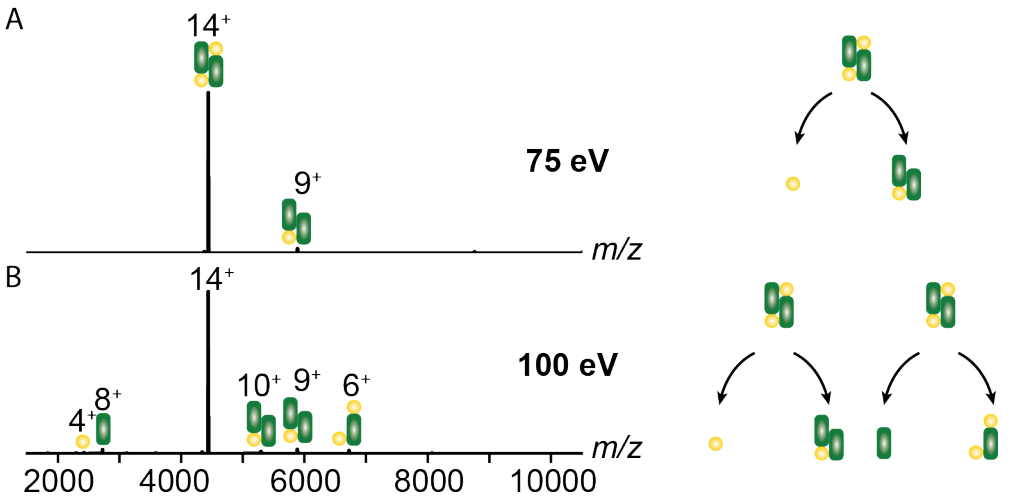


Figure 8: High-energy C-trap dissociation (HCD) of SARS-CoV-2 heterotetramer nsp7$_2$+8$_2$.
The precursor ion 14^+ peak is isolated. (**A**) Dissociation products detected at a collision voltage of 75 eV are nsp7 and nsp7+8$_2$. (**B**) Upon 100 eV, a second dissociation pathway occurs showing dissociation product ions of nsp8 and nsp7$_2$+8 in addition to nsp7 and nsp7$_2$+8.

1.4. AlphaFold

The tertiary structure of proteins is of high interest, given its strong association with function and its utility in designing therapeutic agents[235]. Experimental determination of high-resolution structures is labor-intensive, which is why the number of protein sequences increases more rapidly than the number of structures resulting in the so-called sequence-structure-gap. To address this growing gap, protein structure prediction represents the most efficient alternative method to time-consuming experimental techniques[236]. Considering traditional structure predictions, it could be distinguished between *de novo* prediction, homology modeling, and fold recognition.

Fold recognition, also known as protein threading, identifies structurally resolved proteins that share the same overall fold with the target protein, and uses them as templates for modeling. These template proteins are typically not homologous to the protein of unknown structure. Threading methods are based on the assumption that there exists a finite number of distinct protein folds. In contrast, homology modeling searches for evolutionarily related proteins (homologs) with known structures to serve as templates. This approach relies on the principle that protein structure is more conserved than sequence, meaning that even distantly related proteins are likely to retain similar three-dimensional structures[237]. While homology modeling and fold recognition require available structures as templates, *de novo* e.g., *ab initio* prediction models are mainly based on first principles. Predicting a protein's structure solely from its amino acid sequence, based on physical and chemical principles, demands substantial computational resources[238]. To overcome the computational challenge and to achieve high quality models, protein threading and *de novo* modeling are oftentimes combined, e.g. in Phyre2[239] and I-TASSER[134,240]. For example, I-TASSER combines multiple threading alignments with *ab initio* refinement, and allows additional input from experimental data[134,240,241]. With the ongoing advancement of machine learning, existing protein modeling frameworks have incorporated artificial intelligence (AI) techniques. For example, Rosetta introduced the AI-driven prediction tool trRosetta[242], followed by an improved version, RoseTTAFold[243], which incorporated elements of AlphaFold 2 (AF2)[244]. At the same time, OmegaFold[245] and ESMFold[246] emerged leveraging language learning models to predict protein structures.

The release of AF2 was a major breakthrough in predicting protein structure from sequence with astonishing high accuracy[247]. Consequently, AlphaFold particularly AF2 greatly contributed in closing the sequence-structure gap[248]. For instance, the coverage with protein structures for the foldable regions of the human proteome is considered to be completed[249].

AF uses neural network attention-based architectures that integrate evolutionary, sequential, and spatial information to predict protein structures. By using this approach AF2 outperformed clearly at critical assessment of structure prediction 14 competition. High accuracy of the predicted models were demonstrated by a reliable predicted local-distance difference test (pLDDT) and an accurately determined predicted template modelling score (pTM-score)in AF2 and AF3[250,251].

AF uses both multiple sequence alignment (MSA) representation and pair representation. The MSA representation captures evolutionary relationships among homologous sequences, enabling the model to learn how sequence variation relates to structural constraints. The pair representation builds on this by translating evolutionary signals into spatial relationships, incorporating geometric and physical features such as distance distributions, torsion angles, and residue orientation[250,252].

The key difference between AF2 and AF3 is how the networks, EvoFormer (AF2) and PairFormer (AF3), refine initial structural predictions. Both architectures use attention mechanisms including triangle multiplicative updates. In AF3, the pair representation plays a more central role, enabling improved prediction of structures and interactions while reducing reliance on deep MSAs[252]. Therefore, AF3 usually performs better in predicting proteins with few or no homologs and in predicting biomolecular assemblies.

AF is trained on protein structures retrieved from the protein data bank (PDB), which comprises mostly crystal structures[253–255]. Therefore, AF is not trained on viral protein structures, since fewer than 10% of PDB structures are viral proteins. This makes AF3 superior to AF2 for predicting viral proteins. [255,256]. However, both are clearly not trained for optimal viral protein predictions.

2. Aim and objectives

The aim of this work was to study the kinetics of nsp7-11 polyprotein processing of CoVs and the potential influence on complex formation of nsp16. A quantitative method was to be developed by using native MS, allowing for the precise determination of cleavage site kinetics within their structural context. By examining the coordinated processing in four different CoVs, insights into the assembly and formation of the RTC were expected to be derived.

Polyprotein processing is a common strategy in RNA viruses and the order of cleavage is crucial for viral progeny, and has been studied in CoVs *in vivo* and *in vitro*[31–33,93]. While previous studies have identified polyprotein intermediates *in vivo* and tracked nsp distribution, detailed molecular kinetic studies remain challenging in cellular systems[84,98,159]. *In vitro* approaches in α- and β-CoVs have predominantly relied on peptide-based assays that lack structural context[86,96,98,110]. Although recent MS-based studies of SARS-CoV-1 and -2 have demonstrated that structural features influence cleavage site kinetics, these studies only provided semi-quantitative details on polyprotein processing of nsp7-11[83,114]. Due to the presumable high degree of flexible motion of the polyprotein, it was not possible to resolve a high-resolution structure of the polyprotein to date. There is one study investigating nsp7-11 polyprotein processing in SARS-CoV-2 using HDX-MS, which revealed the cleavage order and structural details of the cleavage sites[83]. However, while this approach provided local structural information, it lacks the ability to directly detect cleavage intermediates and perform quantitative kinetic analysis. Therefore, a quantitative strategy that integrates structural context and enables precise monitoring of polyprotein processing would be highly desirable.

Many nsp functions were characterized and most contributed to the assembly of the large replication machinery[153,257]. However, the stoichiometry and assembly pathway of RTC were not well understood. Complex formation between nsp7 and nsp8 has been widely studied, including during the processing of the nsp7-10 polyprotein in SARS-CoV-1[69,114]. However, the interactions involving processing intermediates and mature nsps remained unexplored. Polyprotein processing is a critical step in nsp maturation, and its intermediates can be considered as proteoforms, potentially exhibiting distinct functions or serving transient regulatory roles compared to their fully processed counterparts.

By using native MS, we aimed to directly detect all cleavage intermediates in parallel and to determine cleavage rate constants, while preserving the structural context of M^{pro} processing. Therefore, the first objective was to establish an approach that optimized and streamlined the workflow of polyprotein processing and enabled the determination of kinetic rates for each cleavage site using native MS. This involved optimizing the design of protein constructs, sample preparation, and measurement strategies that were able to monitor fast and slow

cleavage reactions. Furthermore, multiple processing intermediates generated complex native MS spectra that required a sophisticated data analysis. The second objective was to establish a data analysis approach that extracted precise kinetic rate constants from polyprotein processing, in which multiple reactions occured simultaneously.

Once the novel method was implemented, delivering precise and quantitative analysis of cleavage site kinetics using native MS, we were interested in deepening the comprehension of polyprotein processing and underlying regulatory mechanisms. In order to gain a holistic understanding of polyprotein processing across CoVs, two additional β-CoVs and one α-CoV were included for comparative analysis. To achieve this, nsp7-11 constructs of SARS-CoV-1 and -2, MERS-CoV, and HCoV-229E were designed, and a protein purification protocol was established to obtain nsp7-11 with authentic termini. In the subsequent native MS analysis, we wanted to ensure consistency across species, which is why M^{pro} was used from a single CoV species. Given its relevance in the recent pandemic, SARS-CoV-2 M^{pro} was selected.

To investigate the impact of polyprotein processing on complex formation, we performed *in vitro* interaction studies using nsp16, a methyltransferase that plays a crucial role in the proofreading function of the RTC. The objective was to test the binding potential of the polyprotein by assessing the interaction between processed and unprocessed nsp7-11 polyprotein and nsp16.

Experimental data in conjunction with bioinformatic tools ought to give mechanistic insights into regulatory aspects of M^{pro} cleavage acitivty. To gain deeper molecular insights into polyprotein processing, structural models were predicted using AF. These models were then interpreted alongside the data and sequence alignments.

3. Results and Discussion

3.1. Polyprotein production

3.1.1. Protein production of coronaviral nsp7-11 constructs

In order to investigate the order and kinetics of nsp7-11 polyprotein processing in SARS-CoV-2, we started with two protein constructs containing a polyhistidine-tag (His_6) on either the *N*- or the *C*- terminus (nsp7-11N and nsp7-11C, respectively). Over the course of the experiments, we recognized potential effects on M^{pro} cleavage activity of the non-cleavable His_6-tag. Therefore, we also designed nsp7-11 constructs using an *N*-terminal His_6-SUMO-tag strategy to produce untagged nsp7-11 with authentic termini. Four CoV nsp7-11 constructs were designed: SARS-CoV-2, SARS-CoV-1, HCoV-229E and MERS-CoV.

Protein production for nsp7-11C/nsp7-11N was carried out in a similar way to that for SUMO-His_6-nsp7-11. In order to obtain nsp7-11 with authentic termini, an additional overnight cleavage step had to be included. The SUMO-tag was seamlessly removed by using the SUMO-specific protease ULP-1 (ubiquitin-like-specific protease 1) to produce nsp7-11 with native termini (**Figure 9**).

Proteins were lysed, purified, and directly buffer exchanged into the MS-compatible buffer surrogate ammonium acetate to achieve the best possible polyprotein folding for native MS measurements (**Figure 9 A, B**). SDS-PAGE showed nsp7-11 of the four CoVs with His_6-SUMO-tag around ~ 75 kDa fitting to the size of 73.36 kDa, 73.42 kDa, 73.34 kDa, and 72.75 kDa for SARS-CoV-2, SARS-CoV-1, HCoV-229E and MERS-CoV, respectively. After SEC runs, pure nsp7-11 was obtained with cleaved SUMO-tag. In SEC, ULP-1 was eluted consistently in the range of 17-19 ml, while nsp7-11 was eluted in the range of 10-16 ml peaking slightly different for the four constructs.

During protein purification samples were taken at each step to ensure proper protein purification. Elution fractions from immobilized metal (Nickel resin) affinity chromatography (IMAC) appeared to be pure in SDS-PAGE, so it was checked if an additional IMAC step could pull out His_6-tagged ULP-1 (**Figure S2**). While ULP-1 pull-out was successful, the His_6-SUMO-tag remained in the sample and was visible in native MS (**Figure S3**). Thus, the subsequent SEC-step was necessary to achieve pure, tag-free nsp7-11 with authentic termini.

In the end, the complete processing reactions were verified for all constructs (**Figure 9 B**, **Figure S4**) as done in previous studies[83,258]. It can be concluded that all constructs were produced purely without noteworthy contaminants. Constructs were purified using IMAC followed by SEC, whereby the SUMO-His_6-constructs were digested for 20 h with ULP-1

between IMAC- and SEC-step. All proteins were then directly buffer exchanged into MS buffer without freezing the protein.

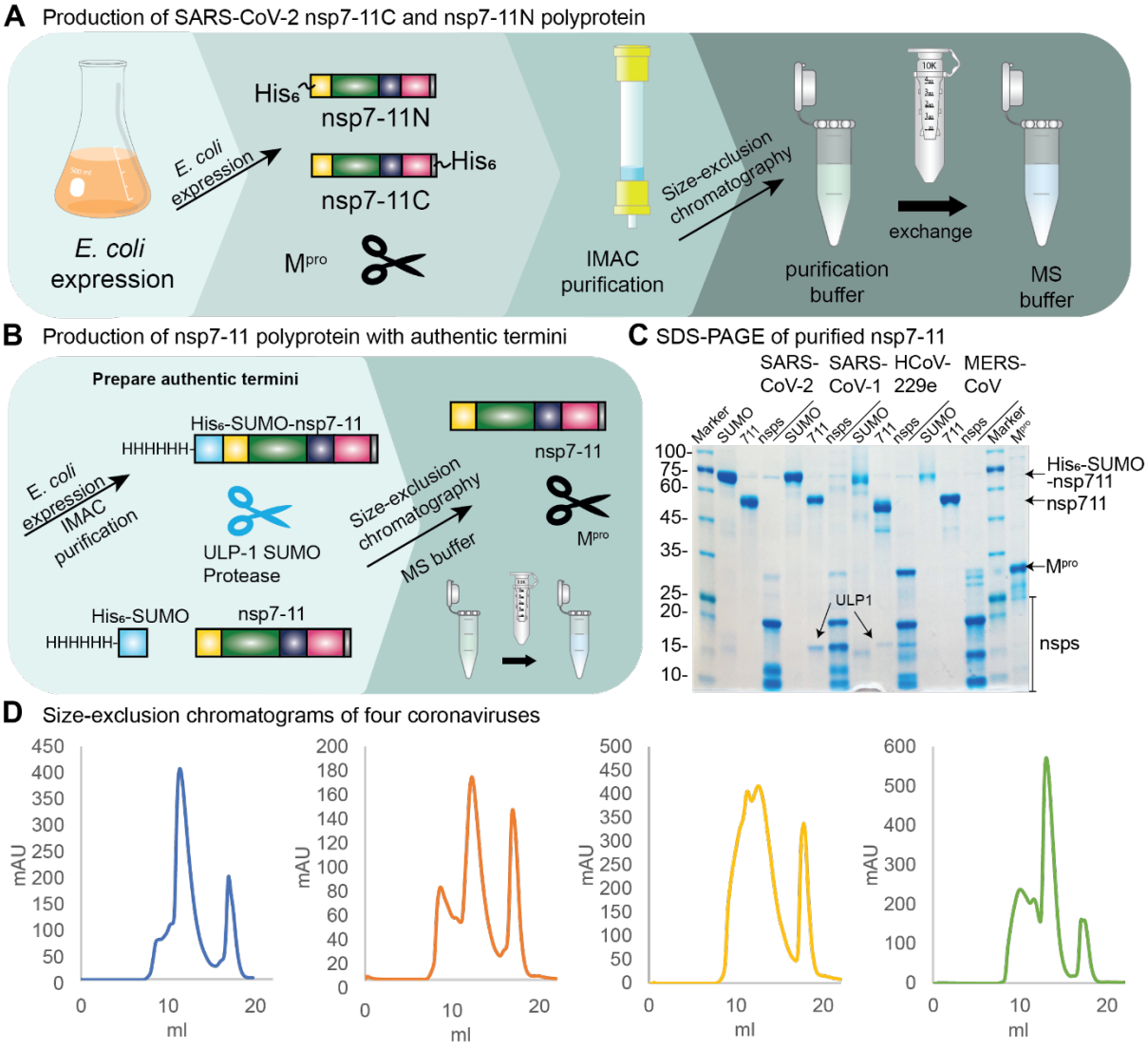


Figure 9: Graphical description of sample preparation of nsp7-11 polyprotein constructs.
(**A**) SARS-CoV-2 nsp7-11 with C- and *N*-terminal His_6-tag were recombinantly expressed in *E. coli*. and purified by IMAC and SEC. Plus, SARS-CoV-2 main protease (M^{pro}) was produced as previously described[259]. Prior to MS analysis the samples were exchanged into the volatile MS buffer surrogate ammonium acetate to prevent mass adduct formation. (**B**) SUMO-His_6-tagged protein constructs were recombinantly expressed in *E. coli* and affinity purified. Samples were also directly exchanged into MS buffer surrogate ammonium acetate. (**C**) SDS-PAGE shows nsp7-11 of the four hCoVs before and after cleavage of the SUMO-tag with ULP-1. For comparison with native MS data, third lane shows nsp7-11 processing products after overnight incubation with M^{pro}. (**D**) SEC was conducted to obtain pure nsp7-11 for SARS-CoV-2 (blue), SARS-CoV-1 (orange), HCoV-229E (yellow) and MERS-CoV (green). SEC runs depict absorbance in mAU from a UV-sensor and show nsp7-11 elution from ~ 10-16 ml and ULP-1 elution from ~17-19 ml.

3.1.2. Complete processing of nsp7-11 polyprotein

Following successful pure protein production, complete digestion of all polyprotein constructs was initiated by adding M^{pro}. This allowed to verify the functionality of the polyprotein processing before moving on to the more complex time-resolved experiments. M^{pro} of SARS-CoV-2 was used as chimera and therefore we wanted to validate complete or near to complete processing. In addition, known and possible new complex formations of nsps were also analyzed and validated.

First, polyprotein processing of all constructs was successfully performed and cleavage products were determined experimentally in native-like conditions confirming the four naturally occurring and conserved cleavage sites: CS7/8, CS8/9, CS9/10 and CS10/11 (**Figure 10**, **Figure 11**). All purified constructs showed pure native mass spectra without noteworthy contaminants alleviating time-resolved polyprotein processing experiments.

Unprocessed polyproteins of nsp7-11C and nsp7-11N showed pure spectra with no contaminants. The charge state distribution of unprocessed nsp7-11C showed a weak bimodal charge state distribution with main charge states from 15+ to 12+ and less abundant charge states from 18+ to 16+. This is in contrast to the monomodal charge state distribution of nsp7-11N, which may indicate a different influence of the His_6-tag on nsp7-11 folding.

The products from nsp7-11C (60.9 kDa) and nsp7-11N (61.1 kDa) differed as expected in the presence of the His_6-tag in either nsp11-His_6 or His_6-nsp7, respectively. Furthermore, we observed the known heterotetrameric protein complex $nsp7_2+8_2$ (62.2 kDa and 65.1 kDa) for either construct (**Figure 10 B**). Hence, the *N*-terminal His_6-tag does not impair complex formation, which indicates proper folding of the proteins. Here upon polyprotein processing, the formation of the heterotetrameric complex of $nsp7_2+8_2$ could be shown despite the heterogenous composition of processing products highlighting the specificity of this complex. The processing products could also be confirmed by SDS-PAGE, albeit with lower resolution and without details on non-covalent complexes (**Figure S4**).

In the nsp7-11N experiments, the His_6-tag at nsp7 sometimes appeared to break off during the native MS measurements, resulting in more mass species, namely nsp7-His_6 and nsp7, and in the heterotetramer depicting either one or two nsp7-His_6 (**Figure 10 B**). Interestingly, an nsp7-His_6/nsp8 heterodimer was not detected, suggesting that heterodimer stability is reduced when formed with nsp7-His_6, thus promoting heterotetramer formation. The higher abundance of heterotetramer formed with nsp7-His_6 supports this hypothesis.

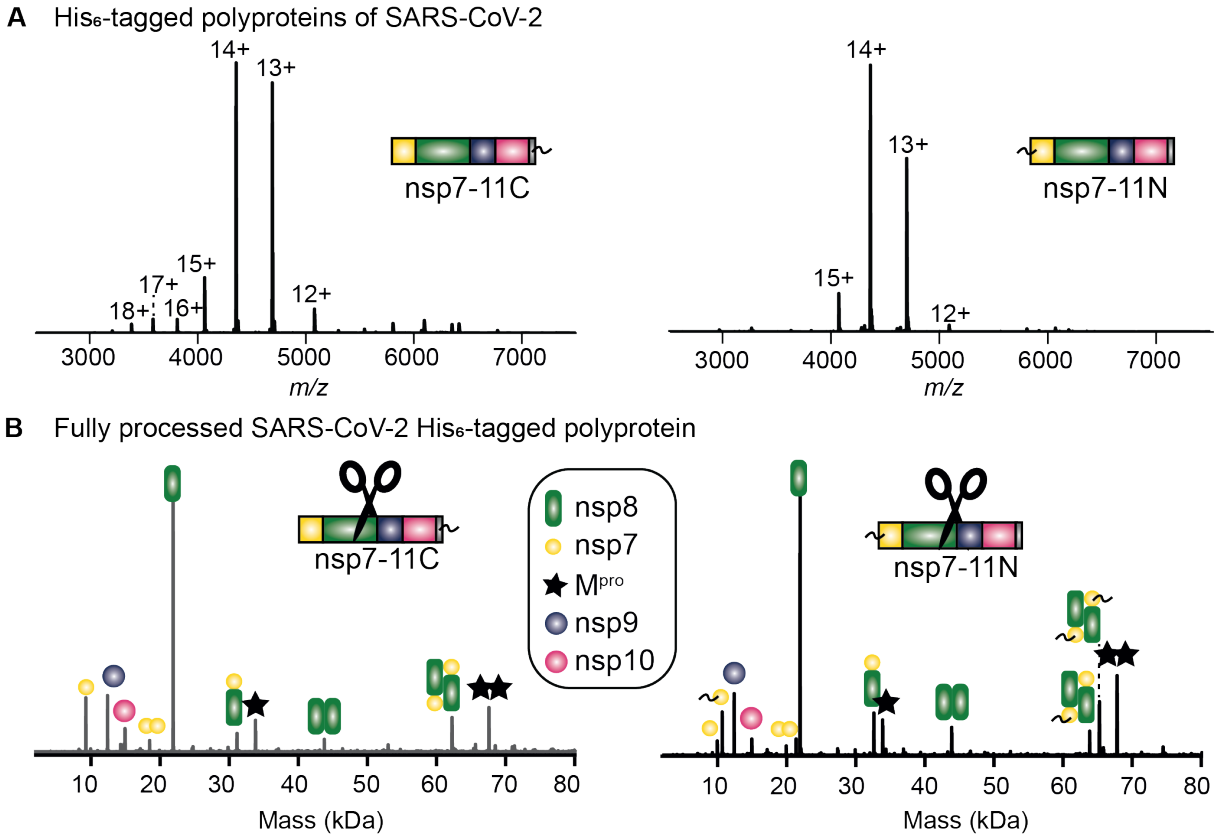


Figure 10: Unprocessed and processed nsp7-11C and nsp7-11N polyprotein in native MS.
(A) Native MS spectra of unprocessed polyprotein nsp7-11C (left, 20 µM) and nsp7-11N (right, 10 µM) are shown. Samples appear to be pure without any contaminations. (**B**) Deconvoluted mass spectra of fully processed nsp7-11C with M^{pro} (left, 20 µM : 10 µM) and nsp7-11N with M^{pro} (right, 20 µM : 10 µM) are shown. Heterotetramer $nsp7_2+8_2$ forms despite the hetergenous composition of processing products, albeit the non-cleavable His_6-tag of nsp7 appears to break off occasionally.

Native MS of unprocessed tag-free nsp7-11 polyprotein showed predominantly pure samples for all CoVs tested, consistent with SDS-PAGE (cf. **Figure 11**, **Figure S4**). It is interesting to note that the charge state distribution is similar in SARS-CoV-1 and SARS-CoV-2, with three to four dominant charge states (16+/15+ to 13+), whereas in MERS-CoV and HCoV-229E the charge state distribution is broader, with charge states ranging from 16+ to 11+. The lower charge states in HCoV-229E and MERS-CoV may indicate more compactly folded nsp7-11 conformations (**Figure 11 A**) The potential presence of different conformational nsp7-11 may imply that HCoV-229E and MERS-CoV polyprotein processing differs from SARS-CoV-1 and SARS-CoV-2.

Mature nsps were detected in all CoV tested, although nsp10 was detected in very low proportions (< 1%) in MERS-CoV and HCoV-229E. MERS-CoV nsp10 occurs mainly still bound in nsp10-11 or nsp9-11 (**Figure 11**, **Figure S5**) Furthermore, heterotetrameric complexes $nsp7_2+8_2$ were detected for all CoVs and $nsp7_2+8$ complexes for MERS-CoV and HCoV-229E were detected as shown in a previous study[69] underlining the specificity of these complexes (**Figure 11**, **Figure S4**).

In summary, we used SARS-CoV-2 M^{pro} to analyze cleavage reactions with two His_6-tagged SARS-CoV-2 polyproteins and four tag-free nsp7-11 of four CoVs. The characteristic charge state distributions of polyproteins imply slight differences in their overall folding. Furthermore, native MS confirmed that all four cleavage sites (CS7/8, CS8/9, CS9/10 and CS10/11) were cleaved in each strain despite the usage of chimeric M^{pro}. After quality assessment and validation of complete processing, time-resolved experiments were started.

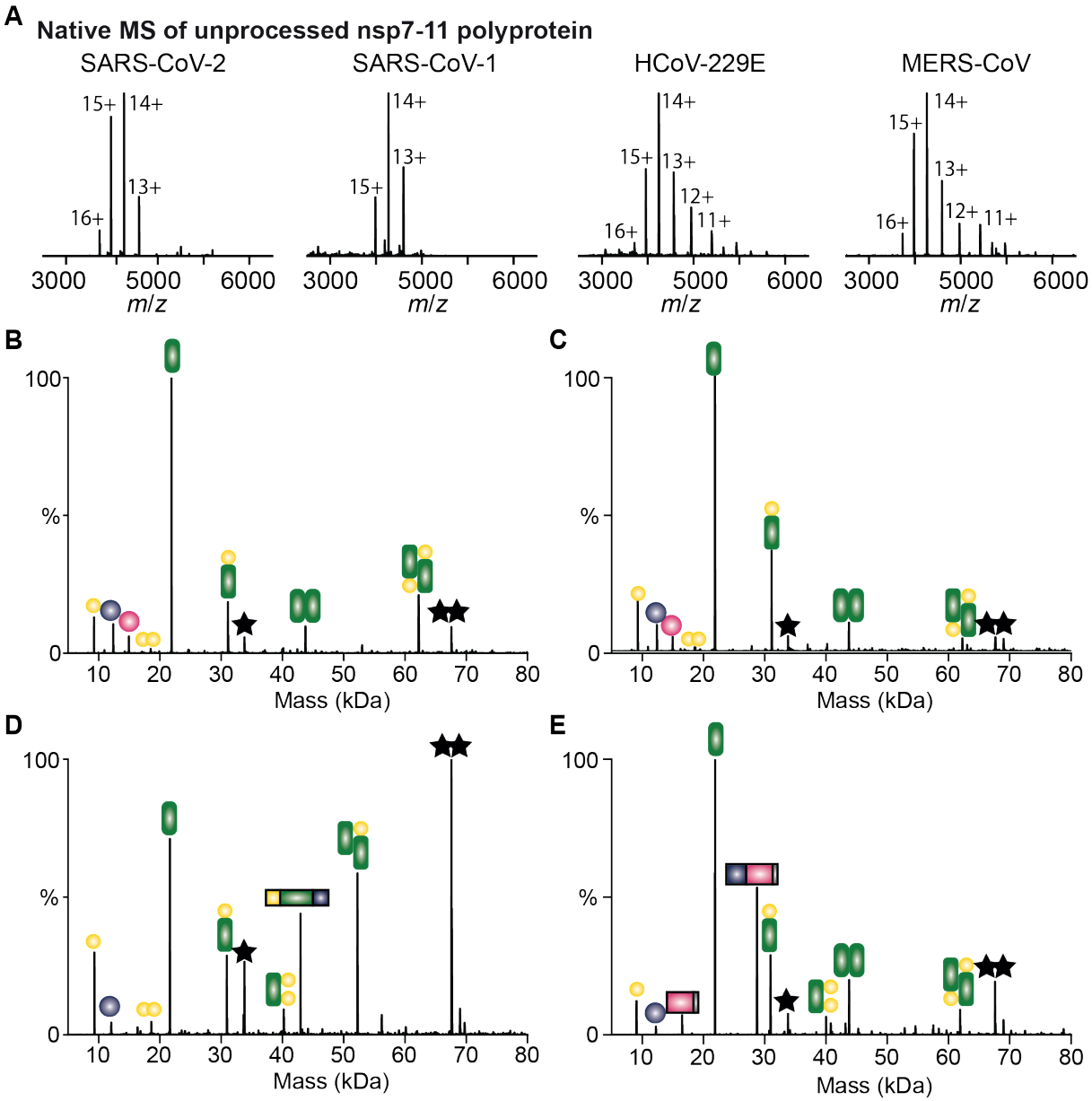


Figure 11: Native MS of unprocessed and processed nsp7-11 polyprotein.
(**A**) Representative native mass spectra of unprocessed nsp7-11 polyprotein of four CoV: SARS-CoV-2, SARS-CoV-1, MERS-CoV and HCoV-229E. Deconvoluted mass spectra of processed nsp7-11 of SARS-CoV-2 (**B**), SARS-CoV-1 (**C**), MERS-CoV (**D**) and HCoV-229E (**E**) after overnight (24 h) incubation with M^{pro} (18 : 3 µM) at 4°C. Mass spectra were deconvoluted with UniDec.

3.1.3. Discussion

In total, six different nsp7-11 proteins were produced. Three different nsp7-11 SARS-CoV-2 proteins were produced, two of which contained an uncleavable His_6-tag (nsp7-11C/nsp7-11N) and one construct was produced with authentic termini. The other three constructs that were produced were tag-free nsp7-11 from SARS-CoV-1, MERS-CoV, and HCoV-229E. After protein production, firstly, all nsp7-11 constructs were tested unprocessed using native MS. In this way, possible distorting contaminant proteins could be detected and were not later confused with processing products during time-resolved measurements. Furthermore, M^{pro} was added and incubated overnight to check fully processed polyproteins.

SEC chromatograms looked similar for MERS-CoV, SARS-CoV-1 and -2 and different for HCoV-229E. SEC elution profiles can provide information on the heterogeneity of a sample e.g. aggregation, conformations, but SEC needs to be equipped accordingly e.g., UV absorbance and light scattering sensors[260]. The SEC chromatograms only depict UV absorbance, which mainly informs about particle concentration[260]. Therefore, it can only be speculated if the shoulder peak prior to the nsp7-11 elution peak depicts protein aggregation, incomplete separation or interaction with other proteins in the sample. While SEC fractions in SDS-PAGE depicted only a band assigned to nsp7-11 (cf. **Figure S6**), native MS of unprocessed nsp7-11 showed low intensity contaminants of tag-free nsp7-11 in SARS-CoV-1, HCoV-229E and MERS-CoV. It is therefore possible that partially interacting proteins such as chaperones are eluted together with nsp7-11, reflected as preceding shoulder peak in the SEC chromatograms. The contaminant proteins detected in native MS match the mass of chaperones e.g., ~ 69 kDa to the mass of DNAK (cf. **Figure S5 C**). This underlines how sensitive native MS is as SEC fractions of the main peak were taken and SDS-PAGE showed clean nsp7-11 bands.

A native MS study has shown that a bimodal charge state distribution reflects unfolded and folded protein states at suboptimal conditions[261]. However, bimodal charge state distributions do not necessarily mean protein unfolding, but often reflect two distinct conformations in solution as most proteins adopt more than one conformation[262]. The weakly bimodal charge distribution of nsp7-11C either indicates partially unfolded protein conformations or just two distinct conformations (**Figure 10**). However, during protein preparation nsp7-11C did not show significantly higher levels of aggregation or clogging on centrifugal filters or within the capillary than the other constructs. Thus, it is likely that the more flexible *C*-terminus of nsp7-11 adopts two distinct conformations in nsp7-11C that are visible in native MS. In contrast to nsp7-11N, in which the His_6-tag is linked to the more rigid nsp7 compared to nsp11. There are a few parameters that define the charge uptake during ESI. The solvent exposed surface area is likely a key parameter dictating the amount of charges during ESI[263,264]. Therefore, the

charge states 11+ and 12+ of untagged nsp7-11 in MERS-CoV and HCoV-229E are noteworthy (cf. **Figure 11**) indicating more compactly folding of nsp7-11.

In our native MS analysis, three different SARS-CoV-2 constructs were examined, all of which yielded a similar profile of detected species. Among these, the nsp8 monomer was the most abundant. Based on prior experiments, nsp8 tended to aggregate when frozen protein samples were used. This suggests that nsp8 could serve as a useful indicator for evaluating the condition of the protein and any changes to its structural state and thereby reducing the ability to form complexes. Therefore, we investigated whether the extended protocol for the SUMO-His_6-nsp7-11 construct would reduce the formation of heterotetramers, but this does not appear to be the case (cf. **Figure S7**). When the relative intensity of these two heterotetramers (His_6-$nsp7_2$+$nsp8_2$ and His_6-nsp7+nsp7+$nsp8_2$) is added together, it is approximately equal to the intensity of the heterotetramer ($nsp7_2$+8_2) from processed nsp7-11 with authentic termini. Heterotetramer formation in nsp7-11C is slightly reduced, though not significantly.

Fully processed nsp7-11N showed nsp7 with and without His_6-tag, approximately 1/3 to 2/3 ratio. However, spectra with unprocessed nsp7-11N showed no mass species without His_6-tag. It is therefore logical to assume that an unintended cleavage of M^{pro} occurred. Looking at the sequence of the His_6-tag (ASRGSHHHHHHGA) there is no Q present, which is a requirement at P1 for M^{pro} canonical cleavage. There are reports of M and H occasionally tolerated at P1, but usually only with very specific P1' partners (A or S)[265,266]. However, the absence of Q and the relatively weak resemblance of the His_6-tag to any known or suspected M^{pro} sites makes an unintended M^{pro} cleavage unlikely. Protease activities during expression and subsequent unspecific carry over is a possibility. However, native MS of unprocessed nsp7-11N could not detect any polyprotein species without His_6-tag. Consequently, only the explanation of the gas phase fragmentation remains. Unintentional gas-phase fragmentation would be conceivable under harsh HCD conditions, but HCD conditions were gentle with little voltage applied (15 eV). Additionally, a clean fragmentation of the entire His_6-tag appears to be a rather unlikely scenario, since even gentle CID/HCD breaks preferentially after P, N or E and rather not after A[267,268]. However, gas phase fragmentation is the more probable explanation.

In the end, six nsp7-11 polyprotein products were successfully produced, showing high quality and purity as confirmed by SDS-PAGE and native MS. Polyproteins of SARS-CoV-1 and -2 were fully processed by M^{pro}, while those from MERS-CoV and HCoV-229E showed near-complete processing. The nsp7-11N construct showed two distinct nsp7 species, one of which lacked the His_6-tag. The mechanism behind this observation remains unclear. However, the complex formation appears to be comparable to that of the other constructs. Overall, all produced polyproteins are suitable for time-resolved polyprotein processing studies. It should

be noted, however, that native MS spectra of nsp7-11N may be more complex due to the presence of two nsp7 species.

3.2. Method establishment

For the time-resolved polyprotein processing experiments M^{pro} was added to nsp7-11. The polyprotein was always in excess, usually in a six-fold excess relative to the protease. The functional M^{pro} started to cleave immediately and native MS data were acquired over time. We conducted two different approaches to measure time-resolved polyprotein processing, which are presented later in this work. The time parameter added another dimension to data analysis, making it more complex. The complete polyprotein processing already showed at least ten distinctive nsp species, and the cleavage reaction was expected to produce more species. Assuming that approximately 15-20 species can occur and that are detected simultaneously, this results in at least about sixty peaks per spectrum.

Time-resolved measurements of polyprotein processing yielded complex MS spectra at different time points, resulting in many species and many spectra that had to be deconvoluted. However, the development of MetaUniDec substantially facilitated high-throughput analysis of native MS data sets, such as the one presented here[269]. Since MetaUniDec allows the deconvolution of mass species and the extraction of peak intensities in a high-throughput manner, it seemed to be an ideal tool for analyzing our data. Hence, we first analyzed our data with MetaUniDec. During data analysis with MetaUniDec we encountered a few problems analyzing our complex data set with this software, which is why we developed a custom Python script in the end. The following chapter describes the reasons why a Python script was developed and it provides a comprehensive overview of the basic principles that were performed by the custom script. The two main aspects were to extract precise intensities for the detected mass species and to further process the data so that kinetic rates are determined for each cleavage sites. For determining kinetic rate constants, an exponential model was fitted to the data that reveal the cleavage activity at each site over time.

3.2.1. Extraction of exact intensities

MetaUniDec offers automatic peak picking in a high-throughput manner and some subsequent graph plotting[269]. Therefore, we initially conducted data analysis using UniDec and MetaUniDec[210]. Here, we describe the challenges we encountered and explain why data anlysis was ultimately conducted using a customized Python script.

Prior to data analysis using MetaUniDec, single spectra were analyzed using UniDec on a sample basis. While UniDec deconvoluted low abundant mass species in single spectrum analysis, MetaUniDec did not deconvolute reliably all of these. For example, UniDec detected nsp10-11 in MS spectra of SARS-CoV-2 nsp7-11 polyprotein processing at 200 min, 270 min and 320 min, but not at the first three time points. Thus, especially low-intensity species have been overlooked in our example, though the peak picking threshold was set to 0.0001. This

setting enables in theory the software to pick peaks with intensity differences of 10^4. In the case of nsp10-11, the difference to the highest peak of nsp7-11 is approximately 10^3 (cf. **Table S1**). Furthermore, removing noise peaks to facilitate automatic peak picking and peak assignment by using an intensity threshold of 0.0003 in the data processing window, did not let the software to detect nsp10-11 peaks.

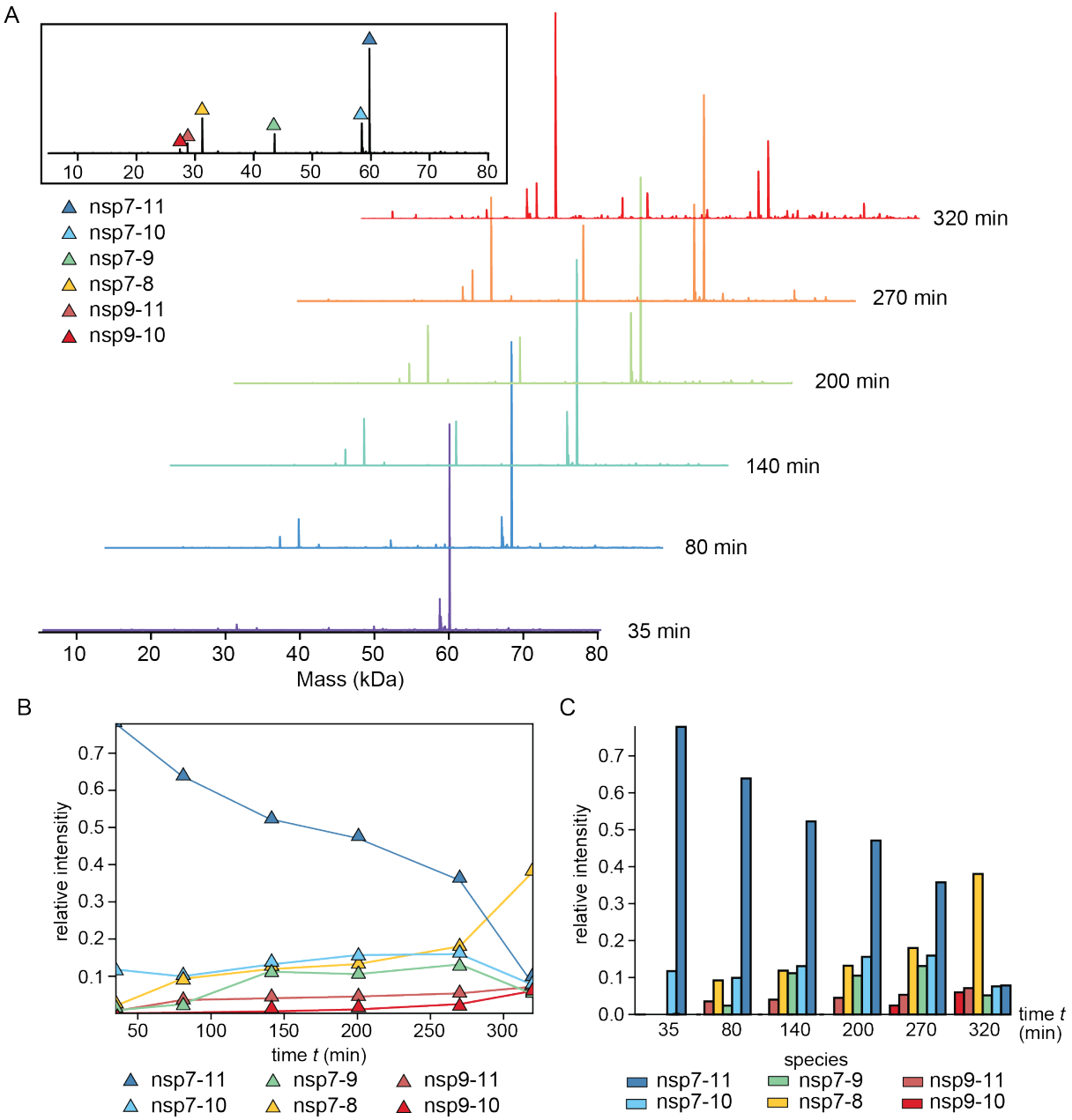


Figure 12: MetaUnidec analysis of SARS-CoV-2 nsp7-11 polyprotein with authentic termini.
This analysis shows one representative run of time-resolved measurements without technical replicates at each time point. (A) Deconvoluted mass spectra (rainbow) show decreasing and increasing relative intensities of mass species at indicated measured time points. The inset black mass spectrum combines all deconvoluted species that were detected over the time. (B) These intensities of all mass species were normalized and plotted over time as line plot. Peak normalization was performed so that the sum of all peaks assigned at each time point is 100%. (**C**) Deconvoluted intensities of all mass species are depicted as bar plot.

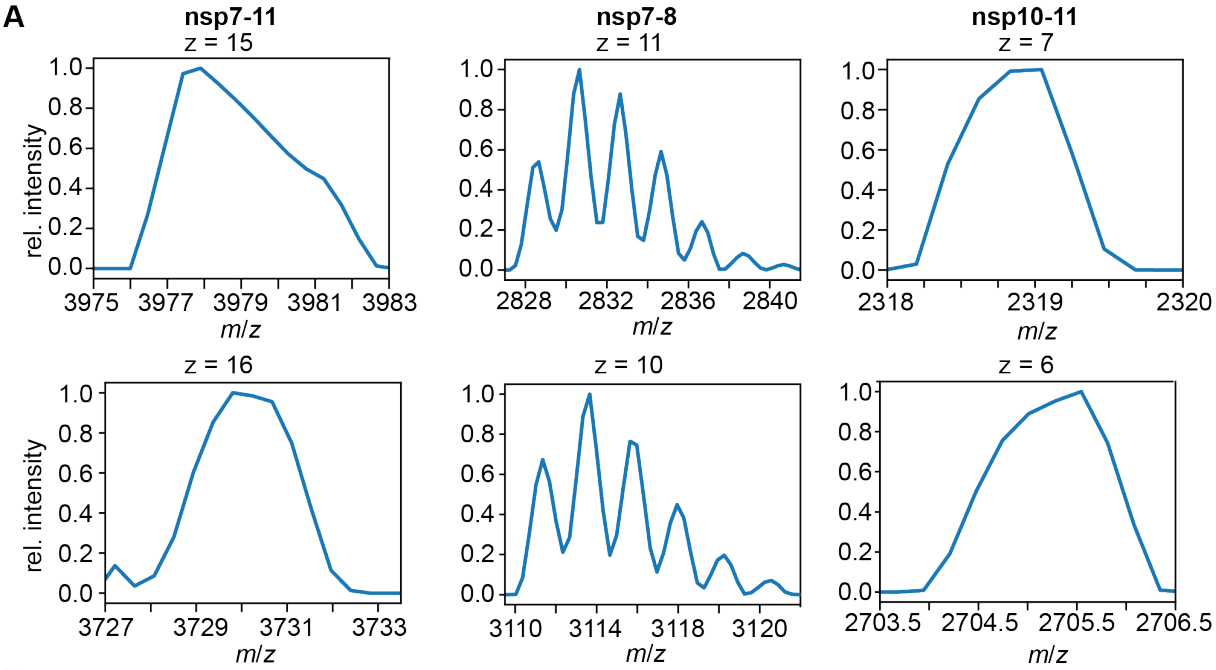
To sum up, while UniDec deconvoluted nsp10-11 in spectra of the three last time points, MetaUniDec did not deconvolute this species at all (cf. **Figure 12**, **Figure S8**). Interestingly, however, Marty and coworkers did not encounter differences in performance when comparing UniDec to MetaUniDec[269]. One possible explanation for this discrepancy could be that because UniDec failed to detect the species at every time point, MetaUniDec likely omitted it from the analysis entirely. Additionally, the edge peaks of a respective peak series were sometimes missed, suggesting that the heterogeneity of the spectra could present a possible limitation for precise peak deconvolution using UniDec.

Furthermore, the analysis of triplicate measurements with MetaUniDec is possible, but did not offer calculation of average and standard error or options to fit data with mathematical models. The latter was especially important for the determination of cleavage rate constants, which is explained in 3.2.3. As the next step was to further process the extracted intensities, it became clear that a custom Python script would be the most appropriate solution to process our complex data sets.Therefore, we established a customized Python script that conducted the following steps: (1) reading acquired data, (2) deconvoluting peaks at each time point using a peak list table, (3) multiplication of species intensities based on protein domains, (4) assignment of substrate species to cleavage sites and normalization, and (5) data fitting to an exponential model. The steps one and two comprise the extraction of the intensity and are described in this section. Steps (3) to (5) are described in 3.2.3.

The first part of the script extracts the intensities for the mass species, for which it required two different inputs: (1) all data files as text files, and (2) a peak list table in comma-separated values file (CSV) format.

The peak list was generated manually with the support of UniDec and by manual peak assignment. It included the columns ‘Name’ for the name of the species, ’Charge states’ for the charge states, and *‘m*/*z* min’ and *‘m*/*z* max’ for defining the peak intervals. The 'Name' and 'Charge States' columns were used to assign unambiguously the extracted AUCs from the peak intervals to the mass species indicated in 'Name'. The peak intervals specify exactly the start (*m*/*z* min) and the end (*m*/*z* max) of the peak as illustrated in **Figure 13 A**. A table of representative species in SARS-CoV-2 nsp7-11, nsp7-8 and nsp10-11 is shown in **Figure 13 B**. By using the peak list for deconvolution, low intensity species such as nsp10-11 were reliably included into the data analysis. As shown in **Figure 13**, the AUCs of two clear peaks of nsp10-11 were extracted and totaled as nsp10-11 species intensity. Thereby, a species that could not be properly deconvoluted with MetaUniDec, was deconvoluted by using the peak list in the custom Python script. Thus, our Python script allowed the inclusion of low species intensities.

A



B

Name	charge states	*m/z* min	*m/z* max
nsp7-11	16	3727	3733.5
nsp7-11	15	3975	3983
nsp7-11	14	4261	4271.5
nsp7-11	13	4589	4603
nsp7-11	12	4970	4982
nsp7-8	11	2827	2841.5
nsp7-8	10	3109.5	3124
nsp7-8	9	3454	3466
nsp10-11	7	2318	2320
nsp10-11	6	2703.5	2706.5

Figure 13: Selected peaks are shown with corresponding peak list and plots of peak intervals.
(**A**) Peak intervals '*m/z* min' to '*m/z* max' are plotted and show exemplary charge states (*z*) of nsp7-11, nsp7-8 and nsp10-11 from which the area under the curve (AUC) is extracted. The extracted AUCs are then assigned to the respective species and totaled as species intensitiies. These species intensities were then further processed. (**B**) A respective peak list as shown here, including 'Name', 'charge states' and peak intervals was generated and fed into a custom python script.

Furthermore, the two plotted charge states of nsp7-8 show that this species can occur with a number of sodium adducts. Adduct peaks contribute to species intensities and should be included as well. However, adduct peaks reduce performance from UniDec[210,211,269]. By cautiously establishing a custom peak list for the polyprotein processing experiments, it is ensured that intensities of adducts peaks and low abundant peaks are included as well.

After feeding the Python script with the data and the peak list, deconvolution and extraction of corresponding intensity was performed at each time point. The script extracted the intensity of each mass species by taking the sum of the AUC for each respective peak of the mass species,

as defined by the peak intervals. Ultimately, the intensity of each mass species was extracted at each time point.

Overall, the first part of the customized script compensates for potential inaccuracies and integrates missing species. In addition, it enables precise control of the extracted intensities via the peak list, thereby allowing control of deconvolution from mass species. Further processing of the data using Python is also facilitated. However, during the creation of the peak list, we identified a few relevant peak overlaps that needed to be corrected.

3.2.2. Peak overlap correction

We detected some substantial peak overlaps that distorted time-resolved polyprotein processing analysis. There are different approaches to estimate proportional intensities for each overlapping species. One of them predicts contribution of the overlapping peak based on expected peak distribution assuming that the respective peak distribution follows a near-Gaussian pattern[270,271]. A more straightforward approach is to estimate proportional peak contribution based on adjacent neighboring peaks, which will be explained more in detail in this section. The implementation into the Python script can be found in 6.3.1.

Representative examples of relevant overlapping peaks in the polyprotein processing of the untagged SARS-CoV-2 and MERS-CoV nsp7-11 polyproteins are shown in **Figure 14 A**. Here, the starting substrate nsp7-11 (z = 15) overlaps with M^{pro} dimer (z = 17). Furthermore, there is a peak overlap in the late stage of polyprotein processing when $nsp8_2$ dimer appears. $Nsp8_2$ dimer (z = 11) also overlaps with M^{pro} dimer (z = 17) (**Table 2**). HCD at m/z 3979 validated both underlying species of the overlapping peak after complete processing of nsp7-11 (**Figure 14 B**).

Assuming that nsp7-11 and M^{pro}_2 dimer have similar ionization patterns, overlapping peaks from both species are proportional to their respective intensities in adjacent non-overlapping peaks (**Figure 14 A**). Thus, the following steps were conducted to estimate the contribution of each species to the overlapping peak.

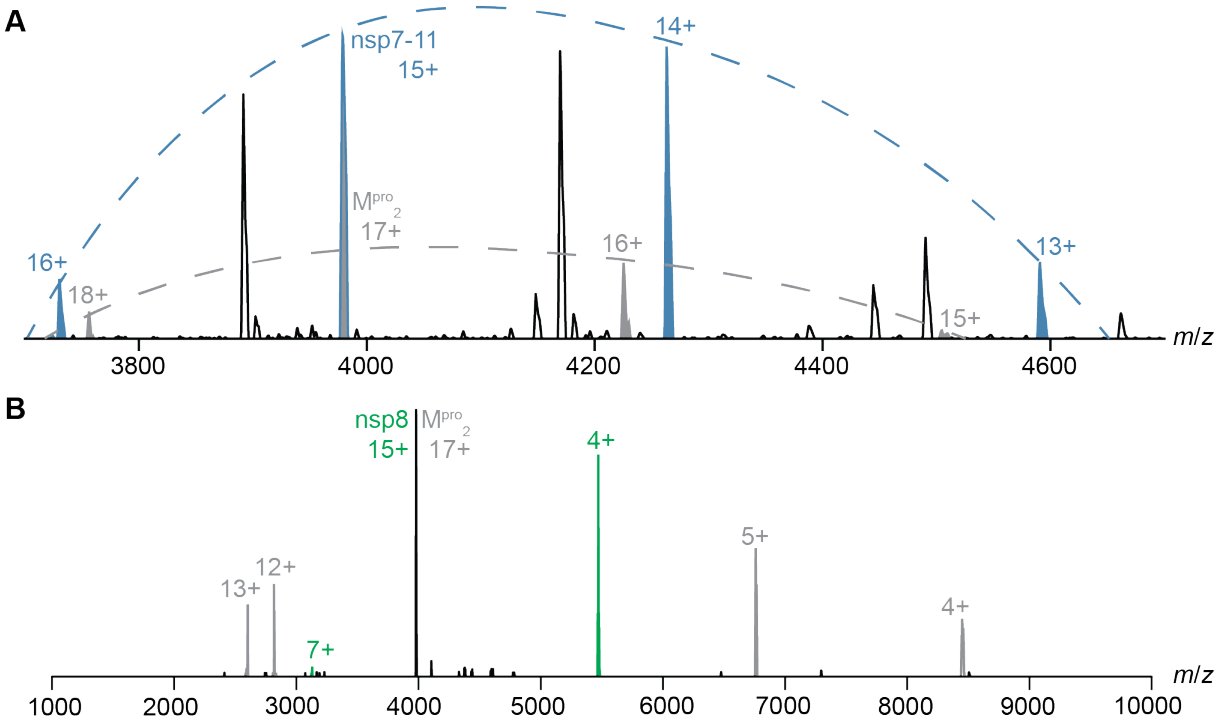


Figure 14: Peak overlap of M^{pro}_2 with SARS-CoV-2 nsp7-11 polyprotein and $nsp8_2$.
(**A**) Peak distribution of SARS-CoV-2 nsp7-11 polyprotein and M^{pro}. Representative mass spectrum at 390 min showing peak overlap between nsp7-11 and M^{pro} during polyprotein processing at time point 390 min. Peaks corresponding to nsp7-11 are colored in light blue, while peaks for M^{pro}_2 dimer are colored in grey. The dotted lines were manually drawn to highlight the near-gaussian distribution of each peak series. Both peak series exhibit four peaks, with two similarily dominant peaks. For M^{pro}, the main peaks correspond to charge states z = 17 and z = 16; for nsp7-11, they correspond to charge states z = 15 and z = 14. The first main peak of each series, z = 17 for M^{pro}_2 and z = 15 for nsp7-11, contributes to the observed overlap. (**B**) Exemplary HCD spectrum of peak overlap from $nsp8_2$ dimer/M^{pro}_2 dimer. Precursor peak *m*/*z* 3979 is either $nsp8_2$ dimer (green), z = 15, or M^{pro}_2 dimer (grey), z = 17. Upon HCD at 187 eV, homodimers of M^{pro}_2 and $nsp8_2$ dissociate into monomers

Firstly, the adjacent equivalent peaks were identified. Therefore, the adjacent peaks used were $z = 16$ for M^{pro}_2 dimer and $z = 14$ for nsp7-11, which means the equivalent adjacent peak was $z = i - 1$ or $z = n - 1$. Secondly, using the peak intervals as described above, AUC of these adjacent peaks were extracted and summed in order to calculate ratios. Ratios and subsequent estimation of peak contribution were calculated for each peak using the following **Equation 7**-**Equation 8**:

Equation 7

$$I_{nsp7-11,\ overlap} = \frac{I_{nsp7-11,\ adjacent}}{I_{nsp7-11,\ adjacent} + I_{Mpro,\ adjacent}} \cdot I_{total,\ overlap}$$

Equation 8

$$I_{Mpro,\ overlap} = \frac{I_{Mpro,\ adjacent}}{I_{nsp7-11,\ adjacent} + I_{Mpro,\ adjacent}} \cdot I_{total,\ overlap}$$

Here, $I_{total,\ overlap}$ is the total AUC of the overlapping peak at *m*/*z* 3979. The calculation of the contribution of each species to the peak overlap could be generalized expressed in **Equation 9**.

Equation 9

$$I_{A,\ z=i} = \frac{I_{A,\ z=i-1}}{I_{A,\ z=i-1} + I_{B,\ z=n-1}} \cdot (I_{A,\ z=i} + I_{B,\ z=n})$$

By using this approach, peak AUCs that contributed to the species intensities were corrected according to their intensity proportionality. This was particularly important for the initial substrate, nsp7-11. As this species is included for the calculation of each cleavage site, precise intensity extraction was essential. With the implementation of this approach in the custom Python script (6.3.1), the initial step, accurately extracting the intensities of each substrate species, has been completed.

Table 2: Species with overlapping peaks are listed with overlapping charge states.

Construct	Species A	Charge state (z)	Species B	Charge state (z)
SUMO-nsp7-11 SARS-CoV-2	nsp7-11	15	M^{pro}_2	17
SUMO-nsp7-11 SARS-CoV-2	nsp7-11	15+	$nsp8_2$	11
SUMO-nsp7-11 SARS-CoV-2	M^{pro}_2	17	$nsp8_2$	11
SUMO-nsp7-11 MERS-CoV	nsp7-11	15	M^{pro}_2	17
SUMO-nsp7-11 MERS-CoV	$nsp8_2$	11	M^{pro}_2	17

3.2.3. Calculation of kinetic rate constants using native MS

In the next step, we wanted to calculate kinetic rates for each cleavage site. By using native MS, all occurring cleavage products can be detected in parallel and hence be monitored over time. This allowed us to simplify the complex processing reaction by assigning the intermediate species to their respective cleavage sites. In this way, conversion rates of the cleavage sites could be investigated and subsequently rate constants could be determined. The second part of the custom Python script performed the required data processing, which comprises the three steps mentioned above: (3) domain correction by factorizing species intensities (6.3.2), (4) assignment of substrate species to cleavage sites and normalization (6.3.3) and (5) data fitting to an exponential model and subsequent determination of kinetic rates (6.3.5). These three steps are explained in more detail in this section.

Domain correction was conducted to improve the quantitative determination of cleavage site kinetics as it more accurately reflects the natural occurrence of the species in solution. Therefore, relative intensities of each intermediate species were multiplied with an individual factor. The factor depends on the number of domains of the species and compensates the different ionization efficiencies. This is required, because upon processing these domains mature to five nsps. Thereby processed nsp7-11 results in five nsps, which are detected by MS and theoretically result together in five times higher intensities than their original species, nsp7-11. However, it would be logical for all end products together to have approximately the same total intensity as the starting product. Therefore, nsp7-11 is multiplied by five as it comprises five domains. **Equation 10** shows a representative multiplication array for untagged nsp7-11 of SARS-CoV-2.

Equation 10

$$m = 5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 3 \cdot I_{nsp9-11} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11}$$

After domain correction, the intensities of all assigned species were normalized to 100% at each time point. This is implemented in the Python script using the following **Equation 11**.

Equation 11

$$a_{t,j,i}^{normalized} = \frac{a_{t,j,i}}{s_{t,j}}$$

Here, $a_{t,j,i}$ is the intensity for each species *i* of repetition *j* at time point *t*. $s_{t,j}$ is the sum of intensities over all species at time *t* and repetition *j*. Thus, each intensity of a species is scaled as a proportion of the total intensitiy per time point. Mean and standard error of the mean were calculated reflecting the spray variation within the technical replicates. Normalized intensities are denoted as relative intensities.

Following normalization, substrate species were assigned to the cleavage sites. This was a necessary step to eventually determine kinetic rates constants for each cleavage site. For this, the relative intensities of all substrates containing the intact specific cleavage site are summed for each time point and plotted over time (**Figure 15 A**). For instance, for CS7/8 the intermediate species nsp7-11, nsp7-10, nsp7-9 and nsp7-8 contain an intact CS7/8, therefore all of them are summed for each time point and plotted over time. For CS10/11, substrate species nsp7-11, nsp7-10 and nsp10-11 contain the intact cleavage site CS10/11 and were thereby considered. The same principle of assignment was employed for CS8/9 and CS9/10 (**Figure 15 B**). Then the relative intensities of the substrate species were summed according to the cleavage sites, and means and errors were calculated from the technical replicates for

each cleavage site. In this way, the multi-reaction process was simplified to a first-order reaction.

In the last step, the data were fitted to an exponential model to determine kinetic rate constants. (**Figure 15**). Mathematically, first-order kinetics can be described using an exponential equation or using a logarithmic equation (**Equation 12**). $[A]_0$ is the initial concentration and $[A]_t$ the concentration to a specific timepoint t.[272]

Equation 12: Algebraic expressions of first-order reactions.
First-order kinetics can be expressed either using exponent (1) or using logarithm (2).

$$(1)\ [A]_{(t)} = [A]_0 e^{-kt} \qquad (2)\ ln[A]_{(t)} = ln[A]_{(0)} - kt$$

This equation can be applied to native MS data, since ion intensity is assumed to be proportional to concentration[273,274]. Therefore, the equation was slightly adjusted (**Figure 15 B**). In order to determine rate constants *k*, the data had to be fitted to the equation in **Figure 15 B**. For this, we employed the least-square method to find the best fit and to extract rate constants *k*, which were calculated based on the slope. To implement this in the custom Python script, we employed 'gmodel' from the model class 'lmfit'[275], which allowed to wrap the pre-defined first-order function as a fitting model. To obtain the fitting parameters from the script, *k* and *k_error* were printed as output. In this way, cleavage site kinetic rate constants were determined.

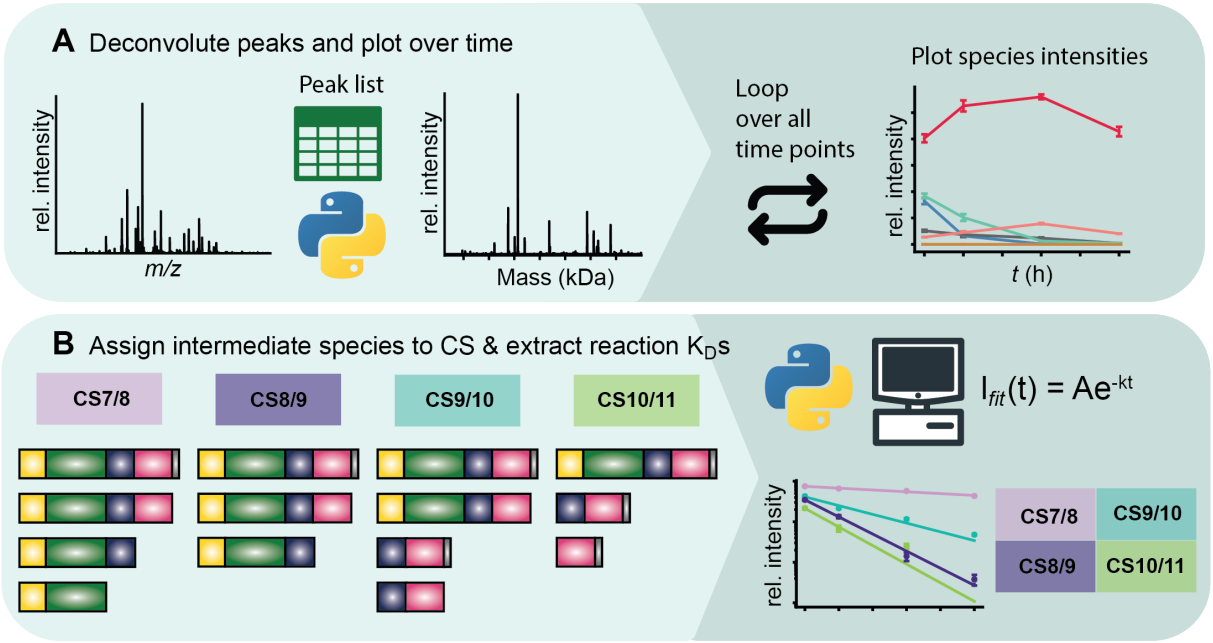


Figure 15: Scheme of two native MS approaches and extraction of cleavage site kinetics.
(**A**) A home-made Python script was fed with acquired spectra (*m/z*-spectra) and a peak list defining respective peak interval for each intermediate species. The AUC was totaled and assigned to mass species at specific time points by looping over the data. Then, the relative intensities of mass species were plotted over time. (**B**) In the next step, the mass species were assigned to their respective cleavage sites, considering only those mass species that had the respective intact cleavage site, as indicated by the mapped species. The relative intensities assigned in this way were summed and plotted as intensities of the cleavage sites over time. The data were fitted with the specified exponential model and kinetic rates were extracted for each cleavage site based on the slope.

3.2.4. Discussion and summary

The analysis of time-resolved polyprotein processing required handling complex native MS spectra containing multiple species. The high dynamic range and resolution of Orbitrap instruments enabled the detection of low-intensity species, which increased the total number of observable species significantly. However, this complexity posed challenges for data analysis.

UniDec and MetaUniDec, developed by Marty and coworkers, provide a sophisticated software for automated deconvolution, including batch processing of large datasets[210,269,276]. Despite their capabilities, the function of automated peak detection in our case failed to fully assign all species and occasionally missed peaks belonging to a given species. Furthermore, it was not possible to compute mean and standard errors within MetaUniDec. Collectively, this led us to opt for a fully integrated approach using a custom Python script.

The deconvolution step, which requires a pre-defined peak list was the primary bottleneck of our script. Accurate definition of peak intervals was critical at this stage, as the precision of intensity extraction directly impacts the reliability of subsequent analysis. Once the peak list was established, data processing proceeded automatically. However, the script could not inherently detect peak overlaps. These were identified during peak list generation and addressed using an intensity proportionality approach as described above. For example, at early time points, the M^{pro}_2 dimer exhibited low intensity due to its 1:6 molar ratio relative to nsp7-11. As processing progressed and nsp7-11 decreased, the M^{pro}_2 dimer contributed increasingly to the shared peak, eventually becoming the major component. Correcting for such overlaps was essential, as each cleavage site's quantification depended on the accurate measurement of nsp7-11 depletion. Similarly, while the nsp8 dimer increased over time, it only became a significant species after complete processing; prior to that, its intensity is negligible.

While the intensity proportionality approach is straightforward, it may be insufficient for resolving overlapping species with substantial size differences. The intensity proportionality approach is based on the assumption that the intensity of an overlapping peak is proportional to that of a respective neighboring, non-overlapping peak. This assumption relies on similar ionization efficiencies of the overlapping species A and B, which becomes increasingly invalid as the size difference between proteins increases. Furthermore, it is important that the adjacent peak is similarly significant within the A and B peak series. For example, it would be problematic when the adjacent peak in peak series A would be the main peak for species A, while the adjacent peak in peak series B is an edge peak. In such a scenario, the assumption of our approach would no longer be valid. Fortunately, this was not the case in the peak overlaps we identified, so it was valid to use this rather crude approach. Thus, a quick

implementation into the custom code was possible. However, employing peak proportion calculations based on Gaussian distribution fitting models would improve accuracy and would be a more robust approach considering such scenarios. The assumption of this approach is that relative intensities of the peaks follow a near-Gaussian distribution and can be therefore fitted with a Guassian fitting model. This would mean to fit the total peak distribution of the overlapping species A and B across their charge states. The fitting can then be used to predict the contribution of the overlapping peak from each species[270,277]. Considering the entire peak series in A and B and subsequent Gaussian peak fitting would solve the problem of non-equivalent adjacent peaks in the proportionality approach.

The second part of the Python script performed factorization based on the domains, normalization, assigned substrate species to the cleavage sites, fitted the data to an exponential model, and extracted kinetic rate constants for each cleavage sites. Factorization based on domains is an approach to approximately compensate for the fact that one nsp7-11 species is processed into five species. Therefore, there is a high probability that the resulting total ion intensity of the five mature species is higher than that of the original one. In addition, ionization efficiency is affected by molecular weight and charge states[278]. This suggests that the intensity of the detected species, which vary in size, is affected differently due to heterogeneity and ionization efficiencies. Therefore, a reference for quantitation purposes would be desirable. One possible approach would be to spike a calibrant solution into the sample, but this would result in more peaks and increase the likelihood of ion suppression[279,280].

After the factorisation step, the substrate species were assigned to the corresponding cleavage sites. This simplifies the multi-reaction to first-order kinetics. This can be easily tested and validated by plotting the fitted data on a logarithmic scale. If this results in a straight line, this will provide proof of first-order kinetics[281].

Ultimately, we developed a custom Python script that performed the entire data analysis workflow, from peak assignment and deconvolution to plotting species intensities over time, fitting kinetic models, and extracting rate constants for each cleavage event. A few peak overlaps were identified in our heterogenous MS spectra. Contribution of the overlapping species were estimated using an intensity proportionality approach. This approach provided an approximate estimate of the proportion of each species contributing to the peak overlap. In this work, using this approach was reasonable as the requirement of using eqivalent adjacent peaks was fulfilled. However, more accurate and robust approaches such as Gaussian fitting are available and implementing them would be desirable. In order to apply a Gaussian fitting model, a more advanced Python code would be required.

3.3. Polyprotein processing of coronaviruses

3.3.1. In-capillary polyprotein processing of SARS-CoV-2

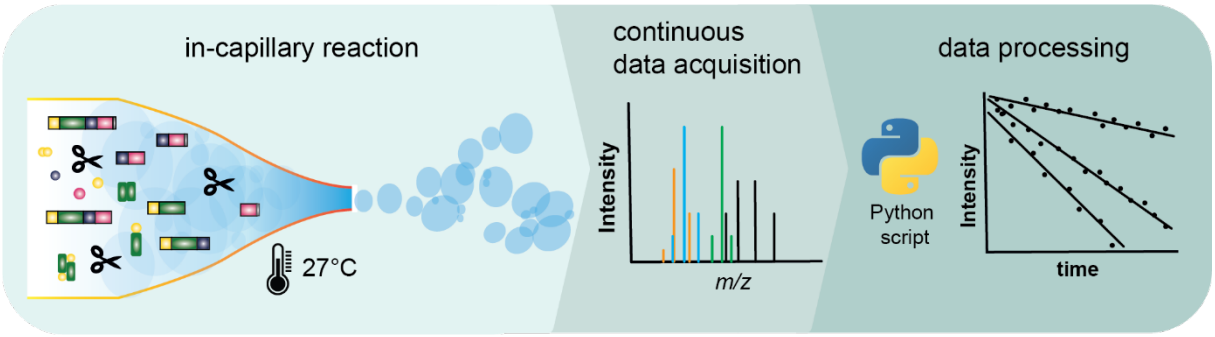


Figure 16: Schematic representation of continuous polyprotein processing measurements.
To monitor the fast multi-cleavage reaction at near-physiological temperature (27 °C), a continuous in-capillary reaction approach was established. In this setup, nsp7–11 (18 µM) was mixed with M^{pro} (3 µM), and data acquisition began one minute after mixing. Mass spectra were then acquired continuously for at least 20 minutes.

When enzymatic reactions take place on different time scales and both fast and slow components are involved, the biggest challenge is to obtain sufficient time points. To address this, we performed both continuous time-resolved measurements for the faster reactions and discontinuous time-resolved measurements for the slower ones. The discontinuous approach will be explained and illustrated in more detail in the next section of this thesis.

The continuous processing approach is particularly well-suited for fast cleavage reactions, as data acquisition was started one minute after initiation of polyprotein processing. In the following, spectra were acquired continuously for 20-30 minutes, during which M^{pro} processed the polyprotein within the capillary at a near-physiological temperatures (27°C). This approach enabled time-resolved measurements with sufficient temporal resolution to monitor fast cleavage events. Tag-free nsp7-11 and nsp7-11C/nsp7-11N were mixed with the same ratio of nsp7-11 (18 µM) to M^{pro} (3 µM).

With the continuous approach, we observed a decrease of intensity of the polyprotein constructs, nsp7-11C (60,950 Da ± 4 Da) or nsp7-11N (61,085 Da ± 1 Da), and an increase of the cleavage intermediates nsp7-9 (43,462 Da ± 1 Da), nsp7-10C (58,360 Da ± 2 Da) or nsp7-10N (59,774 Da ± 3 Da), nsp7-8 (31,103 Da ± 1 Da), nsp7-8N (32,512 Da ± 1 Da), nsp9-10 (27,276.6 Da ± 0.1 Da), nsp9-11 (29,866.2 Da ± 0.2 Da) and nsp10-11C (17,506 Da ± 1 Da) in the first 20 min of the processing reaction (**Figure 17**, **Table S2**). However, cleavage intermediate nsp9-10 did not occur in polyprotein processing of nsp7-11N. Nsp7-11C or nsp7-11N and nsp7-10C or nsp7-10N being the predominant species, suggested that conversion of CS10/11 is slowed down, which in turn pointed to a negative effect of the His_6-tag.

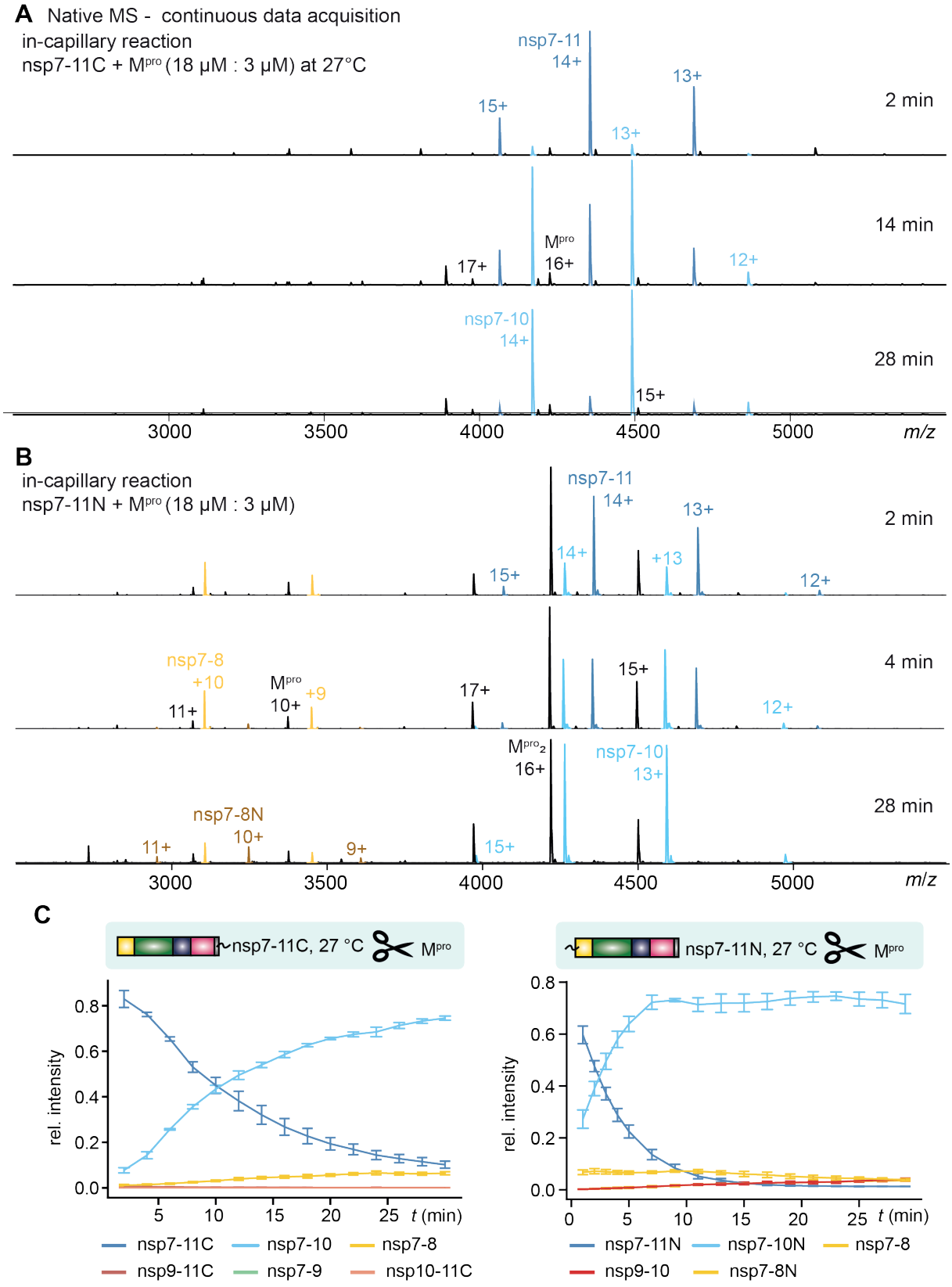


Figure 17: Comparison of continuous processing of SARS-CoV-2 nsp7-11C and nsp7-11N polyprotein. 18 µM nsp7-11 with N- or C-terminal His6-tag were mixed with 3 µM Mpro at 27˚C. (**A**) Representative mass spectra of nsp7-11C showing continuous in-capillary processing at 2 min, 14 min and 28 min. (**B**) Representative mass spectra of nsp7-11N showing continuous in-capillary processing at 2 min, 14 min and 28 min (**C**) Decays of intermediate species of nsp7-11C and nsp7-11N are plotted and dominated by nsp7-11C/nsp7-11N (blue) and nsp7-10/nsp7-10N species (light blue). Data points were simply connected for enhanced visibility.

As the His_6-tag seemed to affect the cleavage efficiency of CS10/11, it was unclear how and to what extent the *N*-terminal His_6-tag affected overall processing. Therefore, continuous polyprotein processing was performed with the tag-free nsp7-11 construct. Fast conversion of CS10/11 was confirmed as nsp7-11 (59,677 Da ± 3 Da) decreased and nsp7-10 (58,366 Da ± 1 Da) increased immediately. A useful point of comparison is the crossing point at which the nsp7-10 species becomes the most abundant. In tag-free nsp7-11 and nsp7-11N, this point occurs after approximately 2.5 minutes, while nsp7-11C takes 10 minutes to reach it. In addition, cleavage intermediates nsp7-9, nsp7-8, nsp9-11, nsp9-10 and nsp10-11 were detected in all three constructs (cf. **Figure 17**, **Figure 18**, **Table S2**, **Table S3**).

Comparing the spectra of the three protein constructs, it is noticeable that peaks of the intermediate nsp7-8 appeared early in nsp7-11N and nsp7-11 processing. The time course of the initial substrate nsp7-11 or nsp7-11N respectively and the early dominant intermediate nsp7-10 were also similar in nsp7-11N and nsp7-11 (cf. **Figure 17**, **Figure 18**).

To improve comparability, cleavage site specific rate constants were determined as explained in 3.2.3. All detected substrates and intermediates comprising a specific cleavage site were summed up and plotted on a logarithmic scale over time.

As nsp9-11, nsp9-10 and nsp10-11 show low intensities below 2%, the rate constants mainly depended on the decrease of nsp7-11 and the increase of nsp7-10. Therefore, in the first 20-30 min, the main conversion was the cleavage of nsp11. The extracted rate constants for CS9/10, CS8/9, and CS7/8 were very similar, as indicated by their overlapping profiles. The kinetic rates were close to zero and show minimal increase over time (cf. **Figure 18 D**). This allowed us to monitor the relatively faster cleavage reactions of two processing sites in detail, CS10/11 and CS9/10 (**Figure 18**, **Table 3**). Continuous polyprotein processing of the three constructs showed very similar kinetic rates for CS10/11 in nsp7-11N and tag-free nsp7-11. This is in contrast to CS10/11 of nsp7-11C, which depicted a three times slower rate (**Figure 18**, **Table 3**).

Table 3: Cleavage sites rate constants of the three SARS-CoV-2 nsp7-11 constructs.
The fitted rate constants in k (min^{-1}) for the three constructs nsp7-11C, nsp7-11N and tag-free nsp7-11 are shown. The standard error of the mean (SEM) is also provided.

Cleavage site	$k_{nsp7\text{-}11C}$ (min^{-1}) at 27°C	$k_{nsp7\text{-}11N}$ (min^{-1}) at 27°C	$k_{nsp7\text{-}11}$ (min^{-1}) at 27°C
CS10/11	0.0816 ± 0.002	0.240 ± 0.005	0.27 ± 0.03
CS9/10	0.003 ± 0.004	0.005 ± 0.001	0.004 ± 0.001

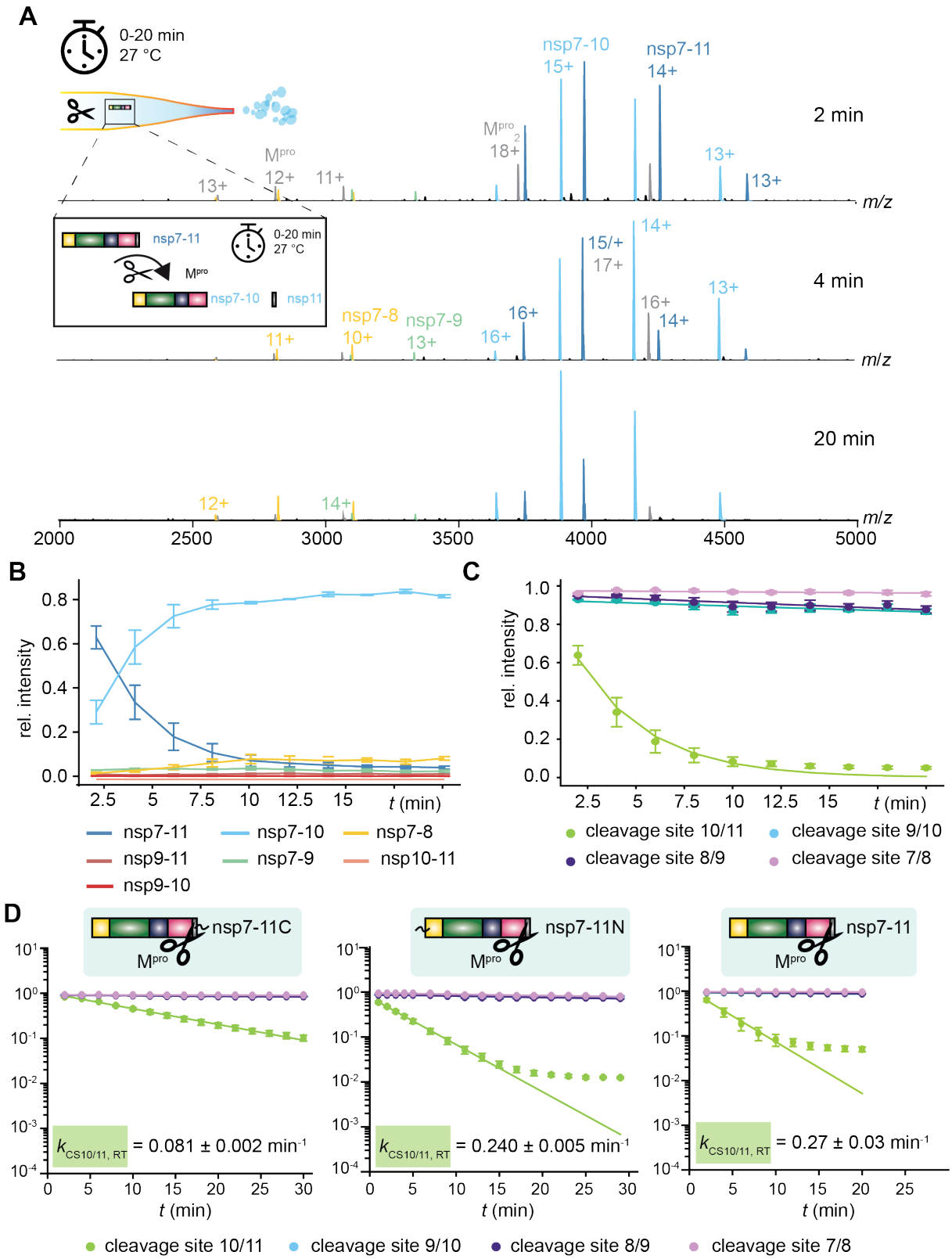


Figure 18: Continuous processing of untagged SARS-CoV-2 nsp7-11 polyprotein.
18 µM nsp7-11 were mixed with 3 µM Mpro at 27˚C. (A) Representative mass spectra of nsp7-11 showing continuous in-capillary processing at 2 min, 14 min and 20 min. (**B**) Decays of intermediate species of nsp7-11 are plotted and dominated by nsp7-11 and nsp7-10 species (light blue). Data points were simply connected for enhanced visibility. (**C**) Decline of all CS for nsp7-11 were plotted and data points were fitted with an exponential model.

In sum, continuous polyprotein processing proved that the His_6-tag affects the speed of the cleavage, but not its order (**Figure 17**, **Figure 18**). Cleavage at CS10/11 is significantly faster than at CS9/10 across all three nsp7-11 constructs: approximately 48-fold faster for nsp7-11N, ~ 68-fold faster for tag-free nsp7-11, and ~ 14-fold faster for nsp7-11C. This indicates that CS10/11 is structurally more accessible in all three constructs (**Table 3**). Overall, the continuous processing approach worked well for early kinetics and had the advantage that a cleavage reaction can be studied near physiologically relevant viral propagation temperatures. However, this approach is time limited due to acidification processes within the capillary[193]. In order to investigate the rate constants *k* at CS8/9 and CS7/8, longer time scales are required

.

3.3.2. Discontinuous processing of SARS-CoV-2 nsp7-11 polyprotein

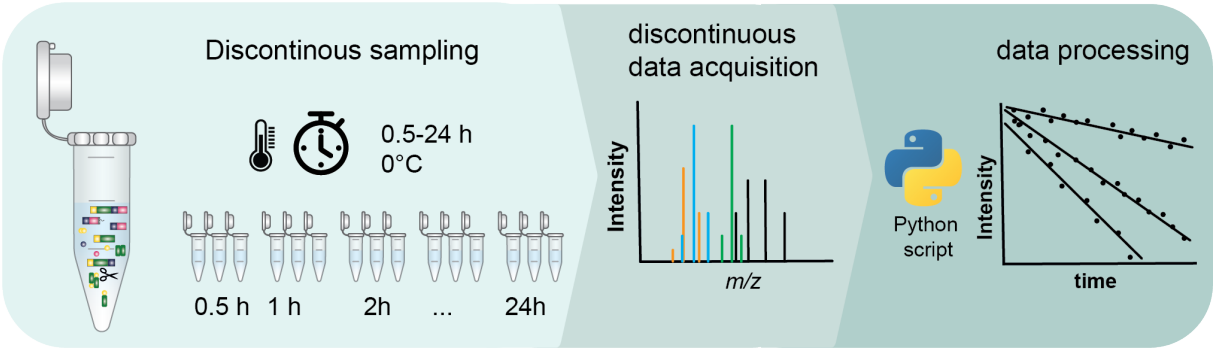


Figure 19: Schematic representation of discontinuous polyprotein processing measurements.
To follow the slow processing reactions, a discontinuous approach was performed. In this set-up, nsp7-11 was mixed with M^{pro} and incubated on ice (0°C). Mass spectrometry data were acquired in triplicates at selected time points.

To monitor slower reactions, we acquired native mass spectra in triplicates at discontinuous time-points from 1 h to 24 h. Over this long period, a more complete picture emerged and allowed determination of kinetic rate constants for all four cleavage sites. Reactions were conducted on ice to ensure protein stability. For SARS-CoV-2, nsp7-11C and nsp7-11 were tested with the discontinuous approach. The His_6-tagged construct was mixed with M^{pro} at a ratio of 2:1, while the tag-free construct was mixed at a ratio of 6:1.

Regarding processing of nsp7-11C, intermediate cleavage products such as nsp7-10, nsp7-9, and nsp9-11 increased at early time-points (> 2 h) and then decreased again. Intermediate products such as nsp9-10, nsp10-11 and nsp7-8 increased for up to 4 h before decreasing again (**Figure 20**). The intermediate nsp7-10 reached its highest intensity after one hour, while other intermediates such as nsp9-10 only reached their highest intensity after two or more hours before they were gradually converted into mature nsps. We determined kinetic rates for the discontinuous nsp7-11C processing as described above.

The fitted rate constants *k* at discrete time points demonstrate the kinetics of nsp7-11C processing, with cleavage at CS8/9 and CS10/11 being the fastest at 0.016 min^{-1} ± 0.001 min^{-1} and 0.018 min^{-1} ± 0.003 min^{-1}, respectively. In contrast, the slowest value at CS7/8 is ten times lower (**Table 4**, **Figure 20 C**). In the continuous approach at 27°C, the data points for nsp7-11N and nsp7-11, as well as nsp7-11C, do not follow exponential decay at t > 15 min and supposedly at t > 30 min (nsp7-11C) due to reduced substrate availability. For nsp7-11C in the discontinuous approach, this would hence be expected after 2 h at 0°C in agreement with our results. This is reflected by the outlier in the data fitting at CS10/11 after 6 h (**Figure 20 C**).

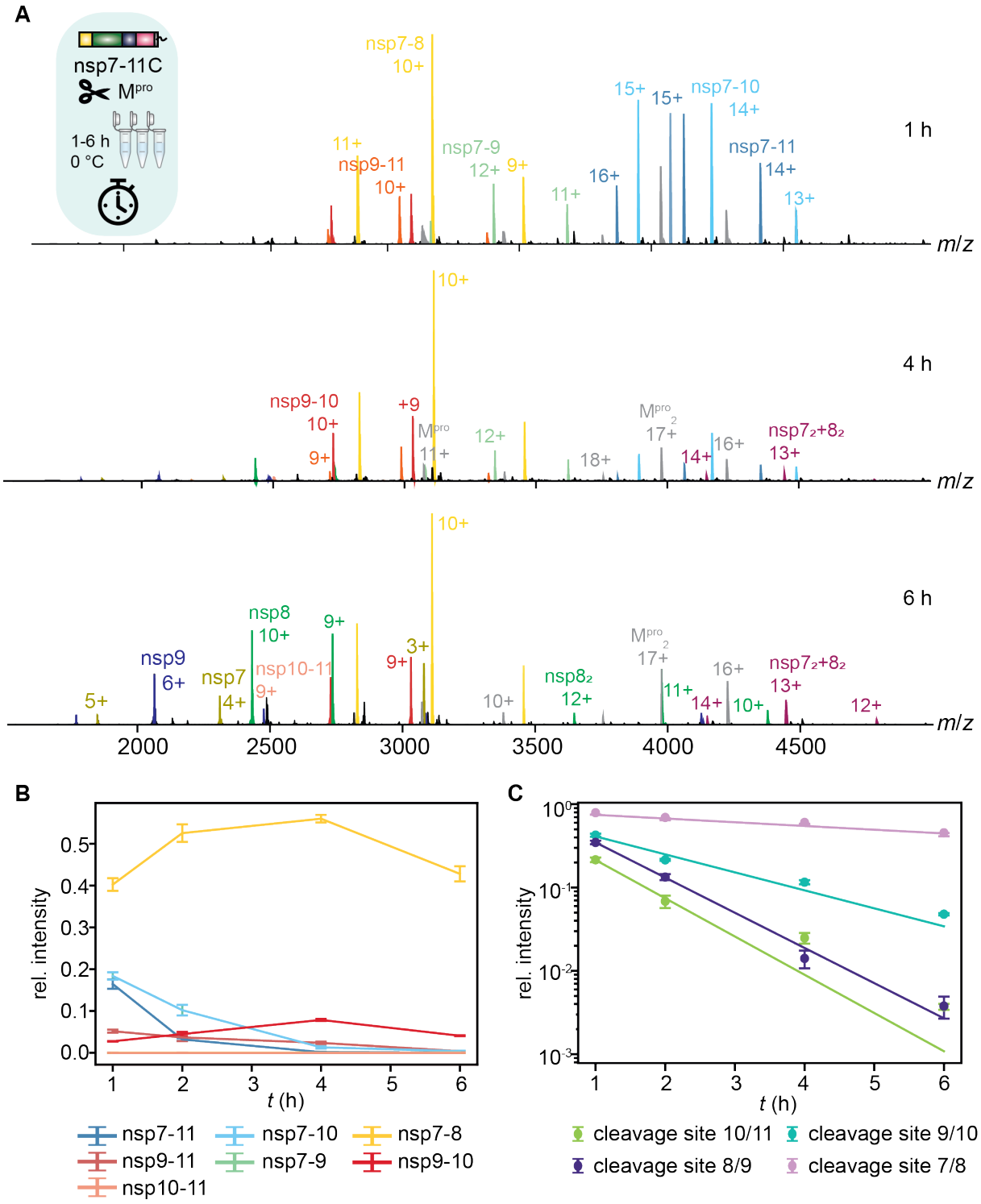


Figure 20: Discontinuous processing approach of nsp7-11C polyprotein.
(**A**) Representative native mass spectra after 1 h, 2 h and 6 h. (**B**) Course of all deconvoluted intermediate species of the initial polyprotein and intermediates at indicated time points. Data points are connected for better visibility. (**C**) Determination of the rate constants k by following the depletion of the substrates corresponding to each cleavage site. The decay is represented as fitted line. After 24 h, all cleavage sites are processed and hence the data points, which are devoid of the plotted species, omitted.

As the continuous polyprotein processing experiments indicated slight effects of the His_6-tag on the cleavage kinetics at CS10/11 and CS9/10, we additionally performed the discontinuous approach using untagged nsp7-11, sampling at discrete time points ranging from 0.5 h to 24 h. Furthermore, the M^{pro} to substrate ratio was reduced from 1:2 to 1:6 to prevent excessive

depletion of nsp7-11, which otherwise could have compromised the accuracy of kinetic rate determination.

Table 4: Cleavage site kinetics of discontinuous polyprotein processing of nsp7-11C and nsp7-11.
Fitted conversion rates k / min^{-1} and the corresponding standard error (SEM) from SARS-CoV-2 polyprotein processing.

Cleavage site	$k_{nsp7\text{-}11C}$ / min^{-1}@ 0°C	$k_{nsp7\text{-}11}$ / min^{-1}@ 0°C
CS10/11	0.018 ± 0.003	0.046 ± 0.004
CS9/10	0.009 ± 0.002	0.026 ± 0.003
CS8/9	0.016 ± 0.001	0.029 ± 0.005
CS7/8	0.0017 ± 0.0002	0.009 ± 0.002

Expectedly, the intensity of cleavage intermediate nsp7-10 peaked first. Then processing quickly led to subsequent cleavage intermediates nsp7-8, nsp7-9, and nsp9-10. At later time points, when the substrate was depleted also lower populated species like nsp9-11 became visible (**Figure 21 A**). Compared to nsp7-11C, similar intermediate cleavage products, i.e., nsp7-10 (58366 Da ± 1 Da), nsp7-9 (43,460 Da ± 2 Da), nsp7-8 (31,102.3 Da ± 0.3 Da), nsp9-11 (28,584.7 Da ± 0.5 Da), nsp9-10 (27,275.9 Da ± 0.2 Da), nsp10-11 (16224 Da ± 1 Da), were observed when processing nsp7-11 polyproteins with native termini (**Table S3**). From these, nsp7-10 was the second most prevalent species after 250 min until nsp7-8 took over at ~300 min. Further species, such as nsp7-9, nsp7-8, nsp9-11, and nsp9-10 were still increasing after the last measurement point (**Figure 21**, **Figure S9**).

During the experiment, we tested whether peaks originated from intermediate product nsp7-8 or nsp7+8 heterodimer. The SARS-CoV-2 precursor ion *m*/*z* 3110 did not dissociate into nsp7 and nsp8, meaning < 1% product ion signal intensity compared to precursor ion intensity originates from the heterodimer (**Figure 22**). Therefore, nsp7 and nsp8 were still too low abundant for complexation and only became significantly populated between 6 h and 24 h.

The fitted rate constants *k* at discrete time points show the kinetics of the processing of the untagged nsp7-11. CS10/11 is converted 75% faster than the next two cleavage sites, CS8/9 and CS9/10, which have a similar conversion rate. This is notable, because nsp7-11C processing exhibited a significantly lower rate in CS9/10 compared to CS8/9. This could be an effect of the His_6-tag. In addition, the examination of nsp7-11 confirmed that CS7/8 has the clearly slowest rate (**Table 4**, **Figure 21 C**). A direct quantitative and temporal comparison of nsp7-11C and untagged nsp7-11 was not possible because the polyprotein/protease ratio is different in the experiments. Nevertheless, it was possible to make a qualitative comparison.

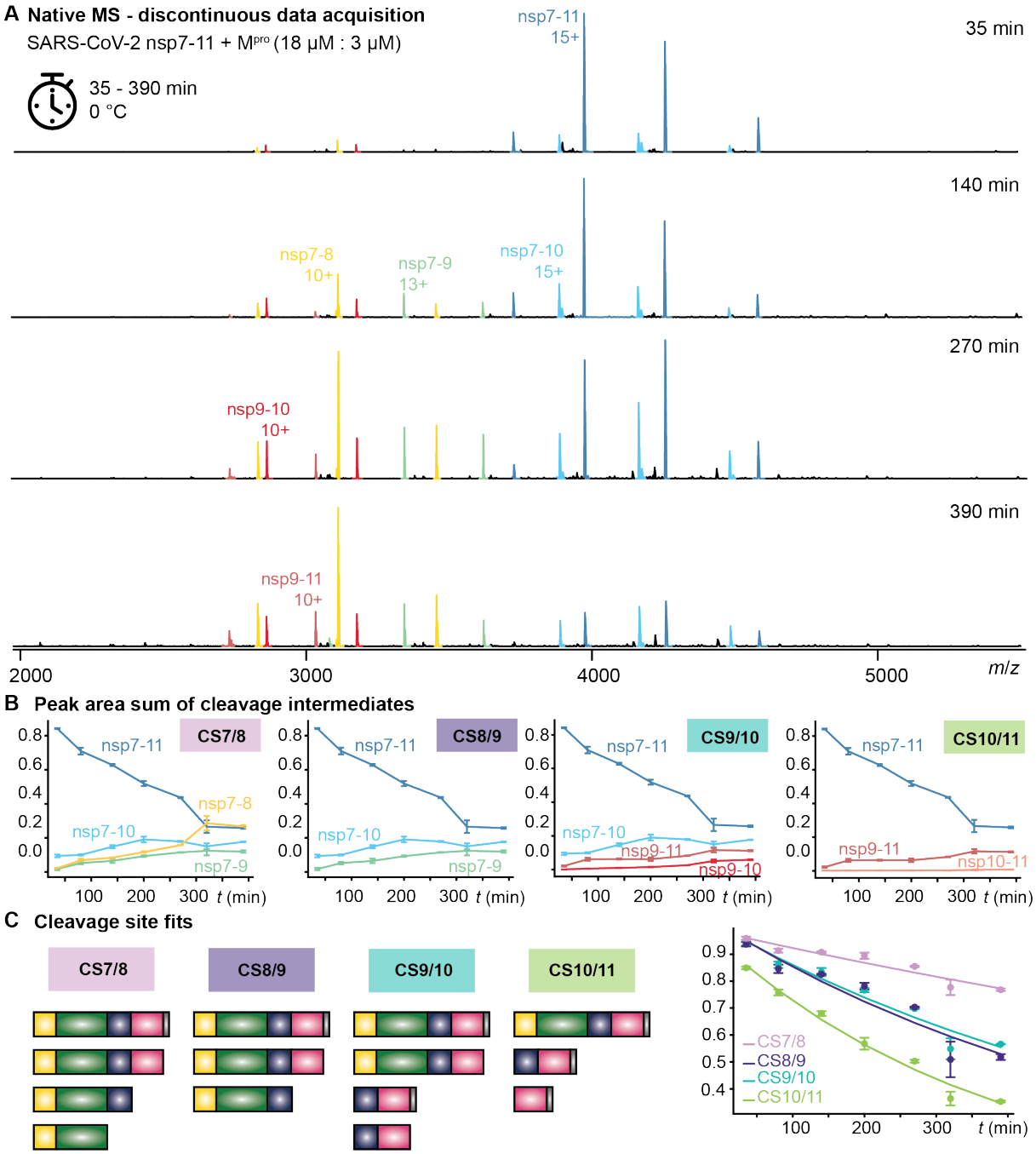


Figure 21: Discontinuous processing approach of SARS-CoV-2 nsp7-11 polyprotein.
19 µM nsp7-11 was mixed with 3.5 µM M^{pro} at 0°C. (**A**) Representative mass spectra at different time points. (**B**) Course of the individual intermediate species of the polyprotein assigned to the four corresponding cleavage sites (indicated on the top right). Data points are connected for better visibility. (**C**) Determination of the rate constants *k* by following the depletion of the substrates corresponding to each cleavage site. The decay is represented as fitted line.

Discontinuous processing of both constructs, nsp7-11C and untagged nsp7-11, showed that *C*-terminal cleavage products can be observed early due to more efficient processing at the *C*-terminal cleavage sites, suggesting a preferred cleavage order from the *C*- to the *N*-terminus. This is in line to what has been shown previously for SARS-CoV-1[258]. However, deviations from this order are observed in both constructs. For instance, intermediate species such as nsp9-11, nsp9-10 and nsp10-11 were detected, whereby nsp7-9, nsp10-11 and nsp9-11

stayed below < 10% (cf. **Figure 20** and **Figure 21**). Notably, there were no intermediate species such as nsp8-9, nsp8-10 or nsp8-11 in either nsp7-11C or nsp7-11, suggesting that CS7/8 cleavage is impaired and occurs at last.

Since CS7/8 and CS9/10 have identical amino acid sequence from P2 to P1'[282], the difference in cleavage kinetics suggests that other mechanisms than the primary sequence are responsible for CS7/8 being a less efficient substrate for M^{pro}. However, the two constructs show a slightly different picture of the cleavage order, which can be derived from the kinetic rates. While discontinuously sampled nsp7-11C processing suggests the cleavage order CS10/11 > CS8/9 > CS9/10 > CS7/8, untagged nsp7-11 suggests the cleavage order CS10/11 > CS8/9 ≈ CS9/10 > CS7/8. It is unclear how large the actual negative influence on the kinetic rates of neighboring cleavage sites, i.e. CS9/10, is. In untagged nsp7-11, CS8/9 appears to be slightly faster than CS9/10, however this difference is not significant (**Table 4**). Ultimately, the His_6-tag appears to hinder processing, implying that the construct with authentic termini could more closely resemble the *in vivo* situation.

In conclusion, we conducted the discontinuous approach and analyzed polyprotein processing of nsp7-11 in SARS-CoV-2. We determined kinetic rates for each cleavage site in two constructs, nsp7-11C and untagged nsp7-11. Based on the kinetic rates of untagged nsp7-11, the cleavage order is CS10/11 > CS8/9 ≈ CS9/10 > CS7/8. Intriguingly, rate constants of CS7/8 were converted much more slowly than all other cleavage sites, despite the similar sequence of CS7/8 and CS9/10, suggesting that structure plays a role in regulating the cleavage order.

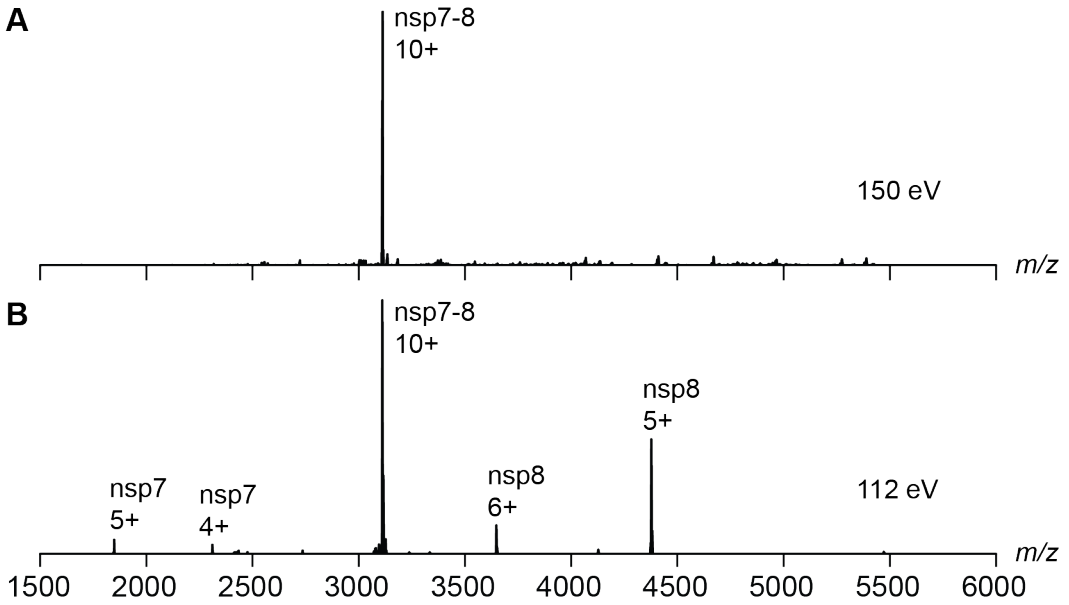


Figure 22: Representative HCD spectra of nsp7-8 or nsp7+8 dimer during SARS-CoV-2 nsp7-11 polyprotein processing.

The precursor ion at m/z 3112 was isolated. (**A**) After 360 min, HCD reveals no subunit dissociation demonstrating that no heterodimer had formed. (**B**) At a later stage of nsp7-11 processing, subunit dissociation observed, confirming the cleavage of nsp7-8 intermediate.

3.3.3. Polyprotein processing of four different hCoVs

After studying nsp7-11 polyprotein processing in SARS-CoV-2 in detail, we extended the research to take a broader view of the underlying mechanism regulating the order of processing. Therefore, we performed the discontinuous approach with three more CoV strains (**Figure 23**-**Figure 25**). To ensure comparability, the ratio of nsp7-11 to M^{pro} was kept at 6:1 for all. We investigated SARS-CoV-1 and MERS-CoV belonging to β-CoVs and HCoV-229E belonging to α-CoVs.

In SARS-CoV-1, early occurring products next to nsp7-10 (58,454 ± 2 Da) were nsp7-9 (43,497 ± 2 Da) and nsp7-8 (31,,115.7 ± 0.2 Da) following nsp9-11 (28,635.3 ± 0.8 Da) later on (**Figure 23**, **Table S4**). However, the observed relative intensities of intermediate cleavage products are largely similar to SARS-CoV-2 nsp7-11, suggesting a similar processing pattern from *C*- to *N*-terminus consistent with previous studies[83,85,258]. This similarity is reflected in the comparable order of rate constants, although SARS-CoV-1 showed slower cleavage rates at CS7/8 and faster cleavage rates at CS9/10 compared to SARS-CoV-2 (**Figure 26** and **Figure 27**). Kinetic rates for SARS-CoV-1 revealed a clear cleavage order from *C*- to *N*-terminus consistent with a previous study[258].

Similar to the observed species in SARS-CoV-1 and 2, the following species were detected in MERS-CoV: nsp7-9 (43,149 ± 1 Da), nsp7-8 (30,931.1 ± 0.2 Da), nsp9-11 (28,741 ± 1 Da) and nsp10-11 (16521.8 ± 0.3 Da). It is notable that nsp9-11 is the second dominant species with 40% in MERS-CoV after 5 h (**Figure 24**, **Table S5**). In SARS-CoV-2, nsp9-11 reached its peak at 5 h with ~ 10%. Furthermore, no species such as nsp8-9, nsp8-10 or nsp8-11 could be found in the three β-CoVs, SARS-COV-1 and -2 and MERS-CoV.

Compared to the three β-CoVs, processing products in HCoV-229E differed. In HCoV-229E the observed dominant early intermediate was nsp7-9 (42,932 ± 2 Da). Intermediate products such as nsp7-8 (30,908 ± 3 Da), nsp7-9 (42,932 ± 2 Da), and nsp7-10 (57440.9 ± 0.4 Da) are also observed, whereas nsp9-10 and nsp9-11 was missing (**Figure 25**, **Table S6**). Strikingly, we observed intermediate products such as nsp8-9 (33,651 ± 1 Da), nsp8-9 dimer (67,300 ± 10 Da) and nsp8-11 (49,992 ± 16 Da), which did not occur in the other three strains. Here, nsp8-11 occurred in low intensities of ~ 2% and the intensities of nsp8-9 and nsp8-9 dimer reached up to ~ 7%.

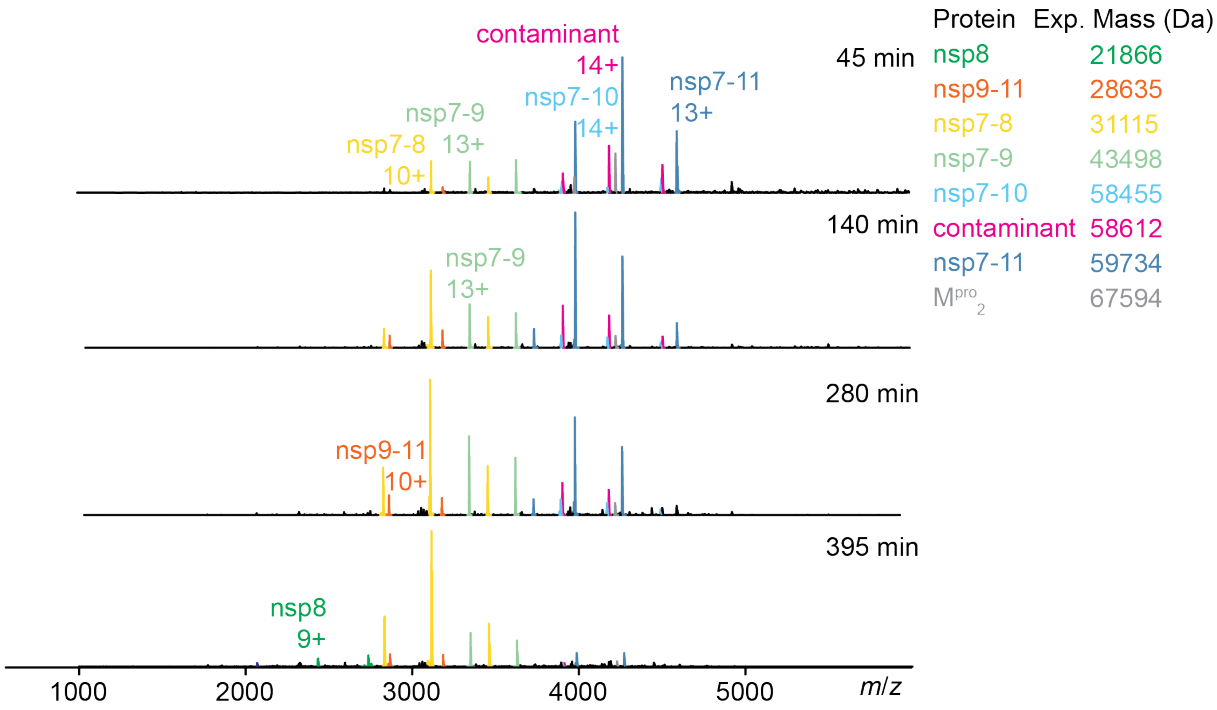


Figure 23: Discontinuous processing of SARS-CoV-1 nsp7-11 polyprotein.
nsp7-11 (19 µM) was mixed with M^{pro} (3.5 µM) and measured at discrete time points. Here representative native mass spectra are shown for every second time point demonstrating the decrease and increase of the initial substrate nsp7-11 and deriving intermediate species. Early detected products were nsp7-10, nsp7-9 and nsp7-8, whereas nsp9-11 occured later in time. While nsp7-10, nsp7-9 and nsp9-11 were already decreasing after 395 min, nsp7-8 wa still accumulating. Measured mass species are indicated on the top right.

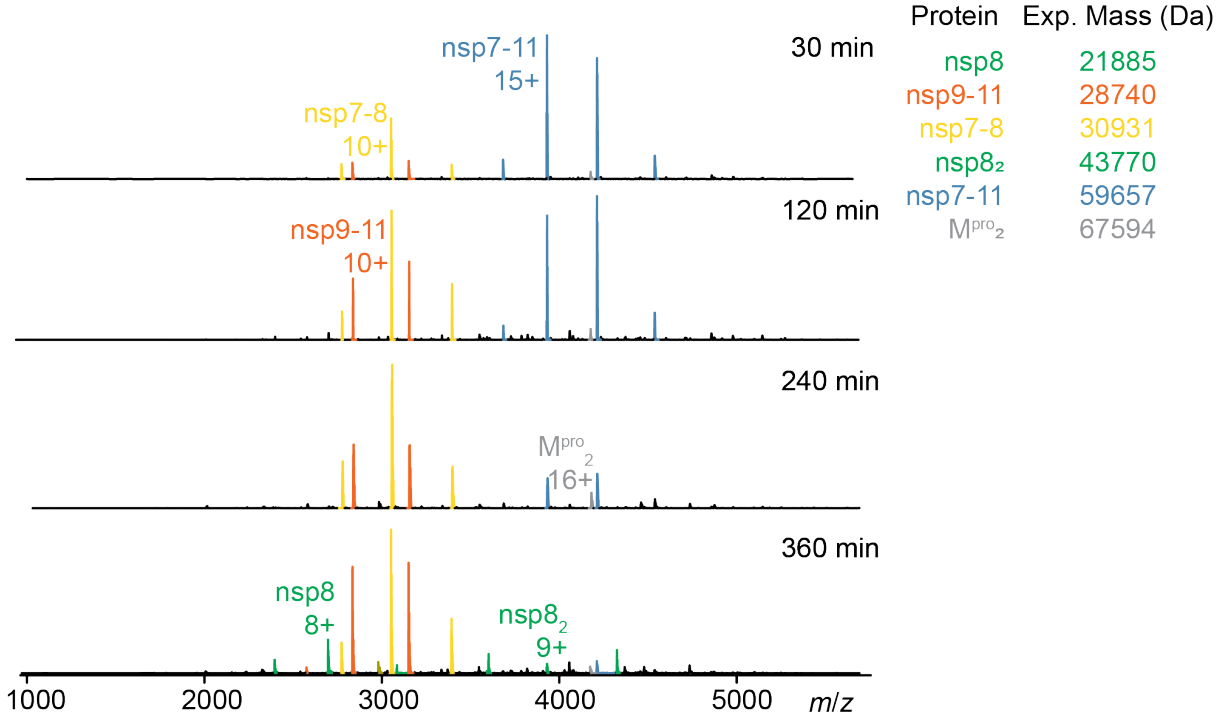


Figure 24: Discontinuous polyprotein processing of MERS-CoV nsp7-11 polyprotein.
nsp7-11 (19 µM) was mixed with M^{pro} (3.1 µM) and measured at discrete time points. Here representative native mass spectra are shown for every second time point demonstrating the decrease and increase of initial substrate nsp7-11 and deriving intermediate species. Intermediate species detected were similar to those in SARS-CoV-1 and -2. Strikingly, nsp9-11 was dominating the spectrum after five hours, suggesting that MERS-CoV nsp7-11 was cut 'in half'. Measured mass species are indicated on the top right.

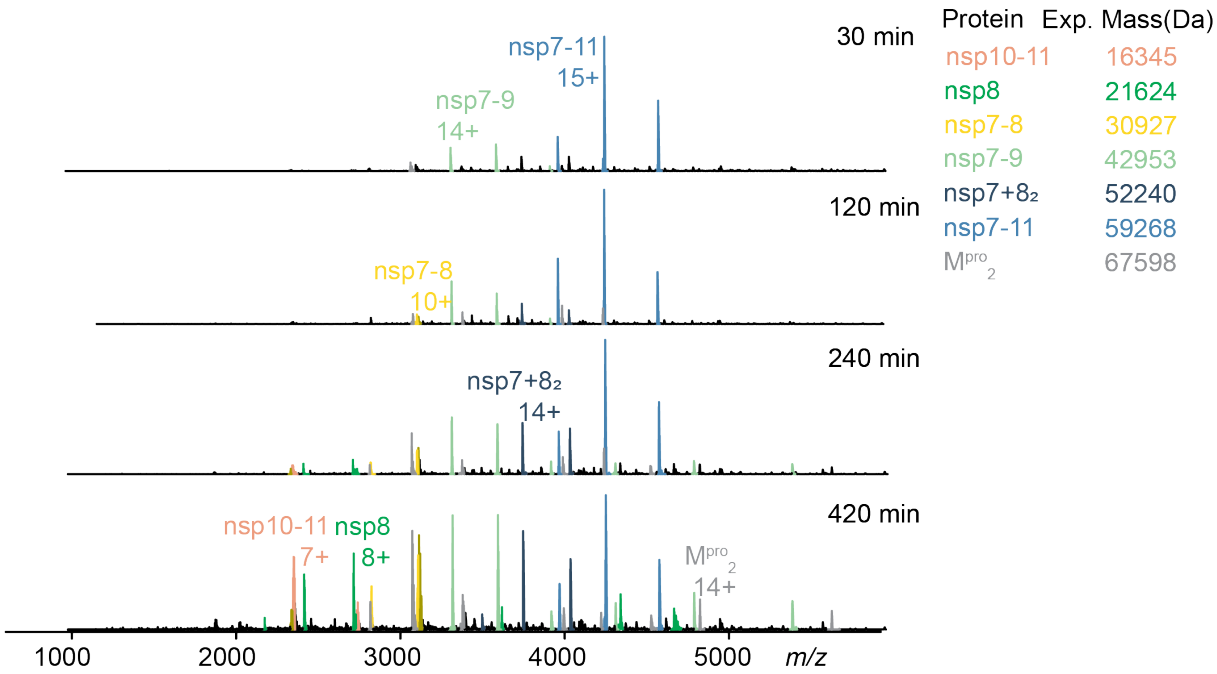


Figure 25: Discontinuous polyprotein processing of HCoV-229E-CoV nsp7-11 polyprotein.
nsp7-11 (17 µM) is mixed with M^{pro} (3.1 µM) and measured at discrete time points. Here representative native mass spectra are shown for every second time point demonstrating the decrease and increase of initial substrate nsp7-11 and deriving intermediate species. Nsp7-9 is the first dominating intermediate species following nsp7-8 and nsp10-11. Measured mass species are indicated on the top right.

Furthermore, in HCoV-229E, nsp7-8 and nsp10-11 increased at later time points, but never accounted for a significant proportion of the intensity. While some early observed products from HCoV-229E nsp7-11 resemble those observed in SARS-CoV-1 and -2, nsp7-10 was essentially absent. It is tempting to state that CS10/11 is hence not addressed first. However, the lack of populated nsp10-11 in the early phases of the reaction rather suggests that CS9/10 and CS10/11 were processed at similar rates, which is corroborated by the linear fits (cf. **Figure 26**, **Figure 27**).

Considering the relative intensities, MERS-CoV exhibited the most distinct intermediate distribution, with nsp7-8 and nsp9-11 emerging as dominant species from the onset throughout the reaction over 500 min (cf. **Figure 24** and **Figure 26**). This unique intermediate pattern effectively results in MERS-CoV nsp7-11 being processed 'in half' at CS8/9. The data suggests an early and efficient cleavage at the CS8/9 site, while CS9/10, CS10/11 and CS7/8 all showed similarly retarded cleavage rates, a pattern distinct from the other three CoVs (**Figure 27**). Indeed, the rate constant for CS8/9 cleavage in MERS-CoV ($k_{cs8/9,\ 0^{\circ}C}$) was approximately twice as fast as any cleavage site rate constant in the other CoVs (**Figure 27Figure 26 B**). The other MERS-CoV cleavage sites were not processed slowest among CoVs tested as the rate constants of cleavage sites CS7/8 ($k_{cs7/8,\ 0^{\circ}C}$) in the other CoVs are two to thirty times slower. This shows that nsp7-11 processing in MERS-CoV appears to be generally faster than in the other three CoVs.

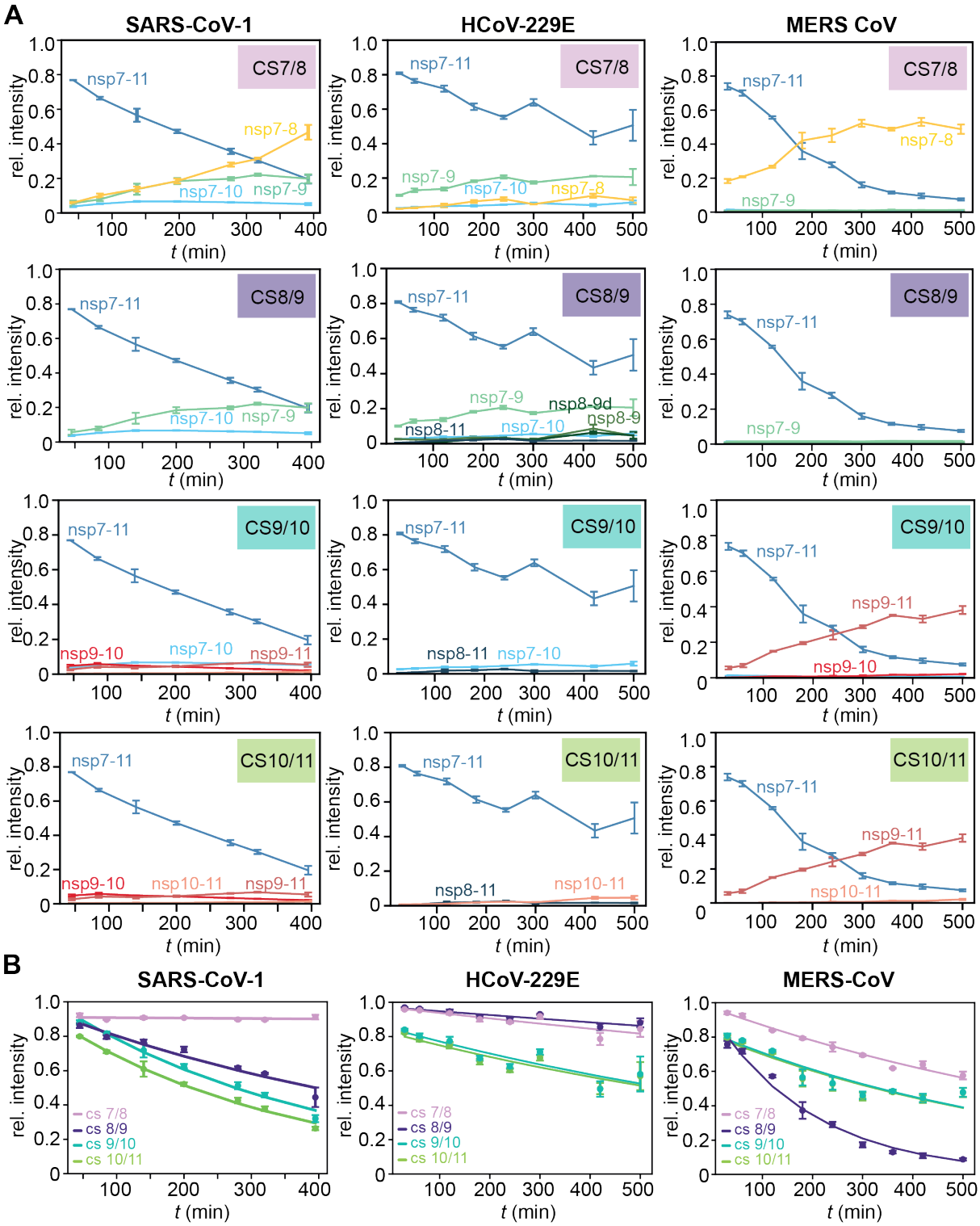
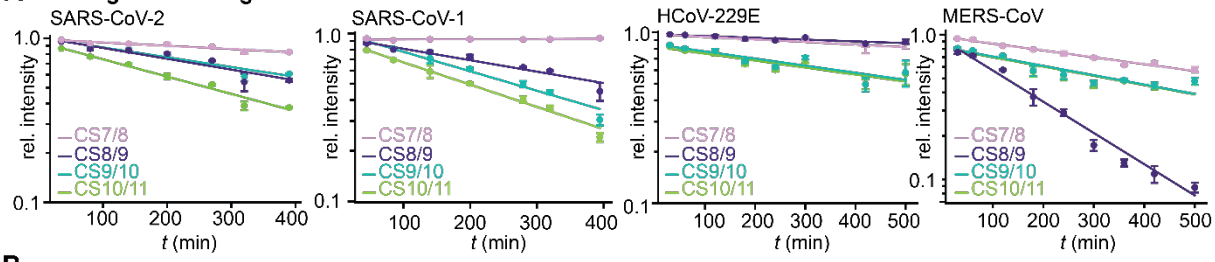


Figure 26: Fits for cleavage site rate constants $k_{0°C}$ and corresponding courses of time-resolved intensities per species.

Polyprotein processing was performed with nsp7-11 and M^{pro} with a ratio of approximately 6:1 at 0°C. All intermediate species containing the specific intact cleavage site were summed in order to monitor the conversion of the specific cleavage site. (**A**) Time-resolved courses of the intermediate species that were considered for each cleavage site are depicted as intensities over time for SARS-CoV-1, HCoV-229E and MERS-CoV. While the intermediate species observed in SARS-CoV-1/2 and MERS-CoV were largely similar, HCoV-229E displayed a distinct pattern of intermediates. For example, nsp9–11 and nsp9–10, which were detected in the β-CoVs, were absent in HCoV-229E. Conversely, nsp8–11 and nsp8–9 were present in HCoV-229E but not observed in SARS-CoV species or MERS-CoV. (**B**) Summed intensities of substrate and intermediates assigned to the corresponding cleavage sites are plotted against time and fitted with first order kinetics

In summary, native MS revealed distinct processing patterns across the four CoVs. SARS-CoV-1 and SARS-CoV-2 CS10/11 was identified as the dominant early cleavage site, while HCoV-229E and MERS-CoV exhibited different patterns. Despite identical sequences at CS7/8 in both SARS species, their rate constant differed by an order of magnitude, suggesting structural rather than sequence effects on cleavage efficiency. The core residues P2 and P1 L and Q are conserved within CS7/8, 8/9 and 9/10 across all species, yet different rates were observed particularly at CS8/9 between HCoV-229E and MERS-CoV, indicating that flanking sequences or structure influence processing. For CS9/10, where the P4 to P3' positions are identical across all species, the varying cleavage rates likely result from structural differences. CS10/11 showed the greatest sequence variability, especially at P2, at which MERS-CoV has P and HCoV-229E has I, possibly explaining their slower kinetics compared to CS10/11 of SARS-CoV-1 and -2. These results reveal that *C*- to *N*-terminal processing of nsp7-11 is not conserved among SARS-CoV-1, SARS-CoV-2, HCoV-229E, and MERS-CoV, though delayed CS7/8 cleavage appears to be a common feature. The non-essential nature of fast CS10/11 cleavage raises the question whether uncleaved intermediates can still function as co-factors in complex formation.

A **Cleavage site fits log**



B

	SARS-CoV-2	SARS-CoV-1	HCoV-229E	MERS-CoV
CS7/8 ***k* [min^{-1}]**	RATLQ|AIASE $0.47\times10^{-3}\pm0.07\times10^{-3}$	RATLQ|AIASE $0.03\times10^{-3}\pm0.04\times10^{-3}$	DSILQ|SVASS $0.34\times10^{-3}\pm0.08\times10^{-3}$	PSVLQ|ATLSE $1.09\times10^{-3}\pm0.01\times10^{-3}$
CS8/9 ***k* [min^{-1}]**	AVKLQ|NNELS $1.52\times10^{-3}\pm0.24\times10^{-3}$	AVKLQ|NNELS $1.60\times10^{-3}\pm0.17\times10^{-3}$	VVKLQ|NNEIM $0.23\times10^{-3}\pm0.05\times10^{-3}$	AVKLQ|NNEIK $4.97\times10^{-3}\pm0.03\times10^{-3}$
CS9/10 ***k* [min^{-1}]**	TVRLQ|AGNAT $1.37\times10^{-2}\pm0.16\times10^{-2}$	TVRLQ|AGNAT $2.56\times10^{-3}\pm0.15\times10^{-3}$	TVRLQ|AGKQT $0.96\times10^{-3}\pm0.02\times10^{-3}$	TVRLQ|AGSNT $1.49\times10^{-3}\pm0.02\times10^{-3}$
CS10/11 ***k* [min^{-1}]**	EPMLQ|SADAQ $2.43\times10^{-3}\pm0.16\times10^{-3}$	EPLMQ|SADAS $2.86\times10^{-3}\pm0.12\times10^{-3}$	RTAIQ|SFDNS $0.94\times10^{-3}\pm0.02\times10^{-3}$	AALPQ|SKDSN $1.48\times10^{-3}\pm0.02\times10^{-3}$

Figure 27: Kinetic rates of polyprotein processing of untagged nsp7-11 from all four CoVs.
(**A**) Relative intensities of respective substrate species were summed for each cleavage site and plotted over time on a logarithmic scale resulting in a linear fit illustrating first order kinetics of the otherwise exponential fitting model. The standard error of the mean (SEM) is also provided. (**B**) Extracted kinetic rates for each cleavage site *k* in min^{-1} at 0°C.

3.3.4. Polyprotein processing of SARS-CoV-2 as prerequisite for nsp16 complex formation

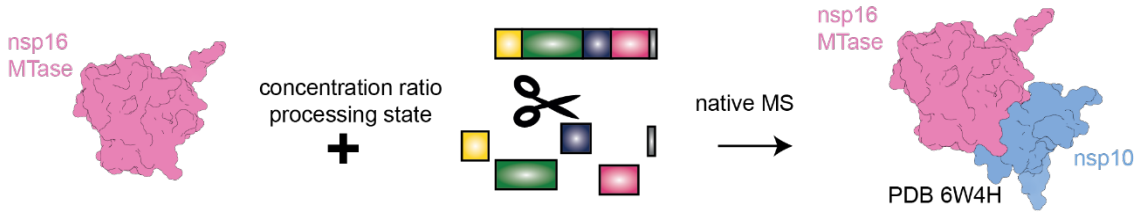


Figure 28: Probing binding dynamics of nsp16 and nsp7-11 polyprotein products.
Recombinantly expressed nsp16-His_6 with an M^{pro}-cleavage site was mixed with nsp7-11 with and without M^{pro}. Thus, the processing states of nsp7-11 were tested for their binding capabilities with methyltransferase (MTase) nsp16. Furthermore, to determine the dissociation constant (K_D), different concentration ratios of nsp7-11 and M^{pro} were tested.

Formation of the RTC requires processing, but whether the RTC incorporates exclusively mature nsps or also immature processing intermediates remains unknown. The functional RTC requires association of several proteins, including nsp10 and nsp16[283–285]. We hypothesize that nsp16+10 complex formation similarly depends on polyprotein processing, specifically the cleavage of CS9/10 and to a lesser extent CS10/11 to release nsp10 from the polyprotein. To test this hypothesis, protein-protein interactions was tested with processed and unprocessed nsp7-11 using native MS (**Figure 28**).

Initially, we tested binding of uncleaved SARS-CoV-2 or MERS-CoV nsp7-11 (59674 ± 3 Da and 59658 ± 4 Da, respectively) to SARS-CoV-2 nsp16 (33323.27 ± 0.14 Da) (**Figure 29**). For SARS-CoV-2 nsp7-11, the predominant signal intensities originated from nsp16 monomer, nsp7-11 monomers and dimers, and low intensities from the nsp7-11+nsp16 complex (~ 2%). Increased levels (< 5%) of SARS-CoV-2 nsp16 complexed with MERS-CoV nsp7-11 were observed despite being a chimeric complex. The complexes were validated using HCD (**Figure 30**, **Figure S10**), which notably revealed the Zn^{2+} binding of nsp10[286]. Here, increased internal energies promoted the migration of zinc ions in the gas phase. This is known to occur occasionally[287].

We then initiated processing of nsp7-11 by adding M^{pro} and incubating with nsp16 overnight. Native mass spectra were distinct for SARS-CoV-2 and MERS-CoV nsp7-11, although both showed high levels of complexation between SARS-CoV-2 nsp16 and nsp10 suggesting specific binding. A heterodimeric complex containing mature nsp10 (nsp16+10, 48,236 ± 1 Da) was apparent in SARS-CoV-2 (**Table S6**). Although $k_{CS10/11,\ 0^{\circ}C}$ in MERS-CoV would suggest complete processing overnight, we observed more than 10% nsp10-11 intermediates and more than 40% heterodimeric nsp16+10-11 as protein complex. Strikingly, no nsp16 with mature nsp10 was observed suggesting that nsp16 exclusively binds to nsp7-11 or nsp10-containing intermediates, which presumingly to protect CS10/11 from further cleavage

in the complex. Given the moderate sequence similarity between MERS-CoV and SARS-CoV-2 (70% for nsp10 and 80% for nsp16), the formation of chimeric nsp16+10 complexes represent an intriguing finding.

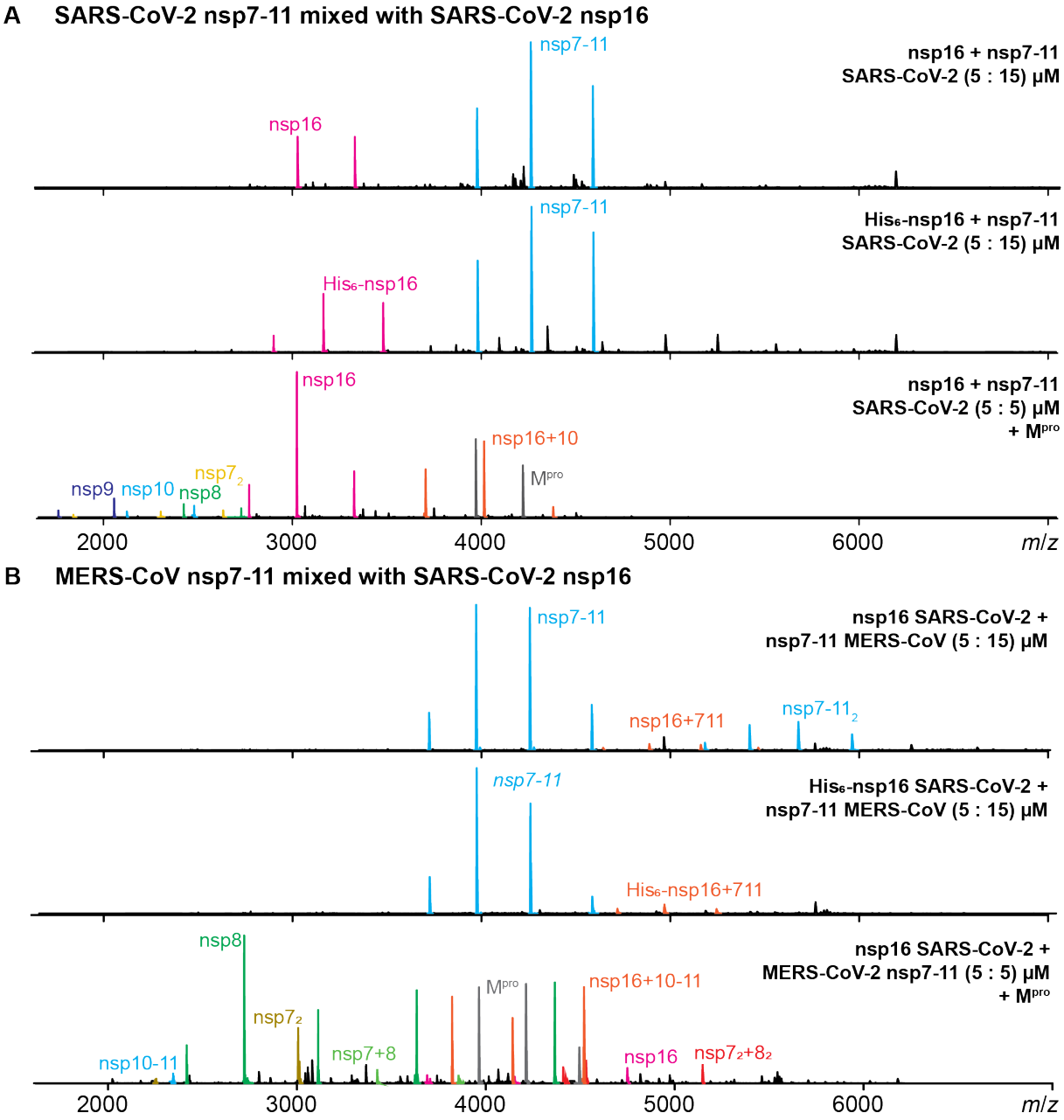


Figure 29: Protein-protein interaction of nsp16 with nsp7-11 polyprotein in native MS.
SARS-CoV-2 nsp16 was probed with processed and unprocessed nsp7-11 of SARS-CoV-2 (**A**) and MERS-CoV (**B**). Concentrations are indicated on the right side. *N*-terminal His_6-tag of nsp16 contains an M^{pro} cleavage site. In order to rule out any artefacts from the His_6-tag, His_6-nsp16 was incubated with low concentration of M^{pro} to cleave the His_6-tag and obtain nsp16 with authentic termini. Then nsp16 was mixed with the polyprotein. Complexes of nsp16 are shown in orange. In (**A**), nsp16 only forms complexes with mature SARS-CoV-2 nsp10 (nsp16+10), whereas in (**B**) SARS-CoV-2 nsp16 forms complexes with MERS-CoV nsp7-11 as well as nsp10-11. Nsp7-11 is colored blue, as well as are its derived products nsp10 or nsp10-11, nsp16 is shown in pink.

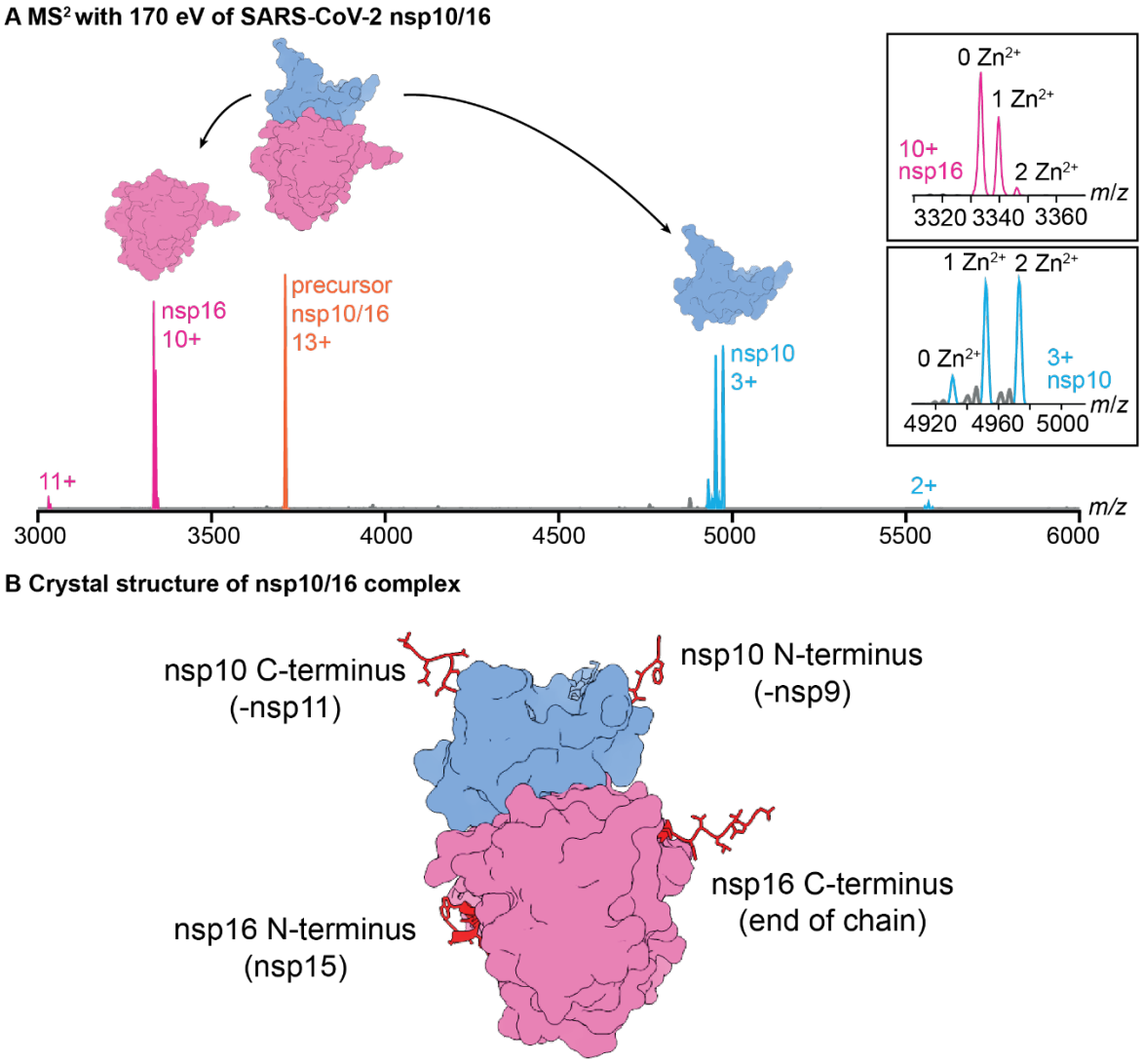


Figure 30: MS² of *m/z* 3712 of nsp16+10 complexes.
(**A**) Precursor complex nsp16+10 (orange) of SARS-CoV-2 dissociating in nsp16 (pink) and nsp10 (blue) at 170 eV is shown. Panels show zoom in on dissociation products nsp16 (z= 10 and 11) and nsp10 (z = 2 and 3) carrying zero to two Zinc ions. (**B**) The crystal structure of the nsp16+10 complex (PDB: 6W4H) shows the surface representation of nsp10 (blue) and nsp16 (magenta), with terminal residues highlighted as ribbon and atom models in red.

To mimic the viral ratio of pp1a to pp1ab, we tested increased proportions of cleaved SARS-CoV-2 nsp7-11 to nsp16, observing similarly increased proportions of nsp16+10 complex formation (**Figure 31**). These experiments yielded a complex dissociation constant K_D of 8 ± 1 µM. In comparison, titration measurements of purified recombinant nsp10 and nsp16 yielded a lower K_D of 1.4 µM[286]. The higher K_D value observed here may result from the complex mixture of polyprotein cleavage products in our experimental system which could lead to signal suppression for the complex[288,289]. Another factor to consider is the method used to calculate K_D. Thibert *et al.* fitted the data using the Hill equation, which is well-suited for capturing potential cooperative binding. However, this approach may be suboptimal in clear 1:1 binding stoichiometry, as it is the case here[286,290].

In summary, nsp16 showed weak binding to immature nsp10 within the polyprotein but strong binding to mostly mature nsp10. Our results indicate complex formation requires *N*-terminal CS9/10 cleavage but not necessarily *C*-terminal CS10/11 cleavage. Available crystal structures of nsp16+10 cannot explain this cleavage site preference, as both nsp10 termini are distant from the nsp16 binding site (**Figure 30 B**). We conclude that while complete processing of nsp7-11 is not essential, it greatly enhances CoV methyltransferase nsp16+10 complex formation.

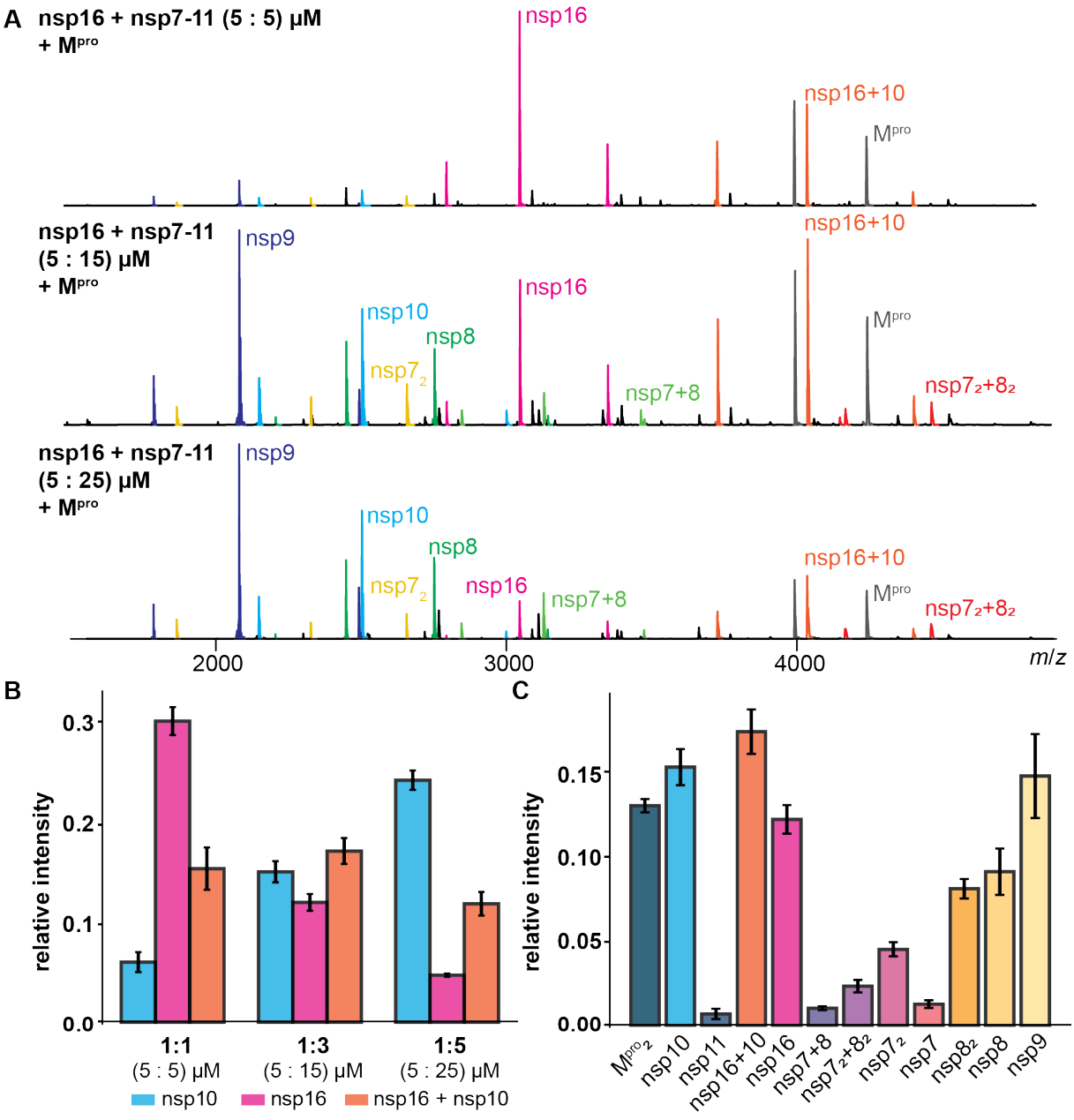


Figure 31: Probing the affinity of SARS-CoV-2 nsp16+10 complex at three different concentrations.
Protein complexation of nsp16+10 was investigated at three different concentrations as indicated in the panels. (**A**) Representative native mass spectra show nsp16+10 complex in orange, nsp16 in magenta and nsp10 in blue. Further nsps are nsp9 (dark blue), nsp7 (yellow), nsp8 (dark green), nsp7+8 heterodimer (light green) and heterotetramer nsp7$_2$+8$_2$ (red). (**B**) The average relative intensity of each species from triplet measurements of the three tested ratios is depicted as barplot. (**C**) The relative intensity of each species is shown in triplcate measurements at a ratio of 5 µM:15 µM.

3.3.5. AlphaFold predictions of nsp7-11 of four hCoVs

To investigate how protein sequence and structure relate to the conversion rates of the four substrates, we analyzed structural models generated by AF2 and AF3$^{250,291–294}$(**Figure 32**-**Figure** 34). For each CoV, the polyprotein sequences of nsp7-11 were fed into AF3 using the standard run. SARS-CoV-2 nsp7-11 was also predicted with AF2.

AF3 and AF2 output files contain pLDDT scores for the top five models, which measure the per-residue confidence in the local structure prediction and estimate how well the prediction agree with an experimental structure. To obtain a general overview of all 20 models predicted by AF3, pLDDT scores per residue were plotted as line plots (**Figure 32**). The top five models for each strain were largely similar, with only minor variations in pLDDT across the protein sequence. For each strain, model 0 was selected, except for MERS-CoV, where model 1 was chosen due to its slightly higher pLDDT scores in the CS7/8 cleavage site region. Model selection was guided by a focused evaluation of prediction quality specifically at the cleavage site regions. The overall folding of the polyprotein was similar in all four CoVs, depicting nsp7-11 as an elongated structure rather than a compact fold (cf. **Figure 33** and **Figure S11**-**Figure S14**). Interestingly, AF2 predicted SARS-CoV-2 nsp7-11 in a more compact form than AF3 (**Figure 33**).

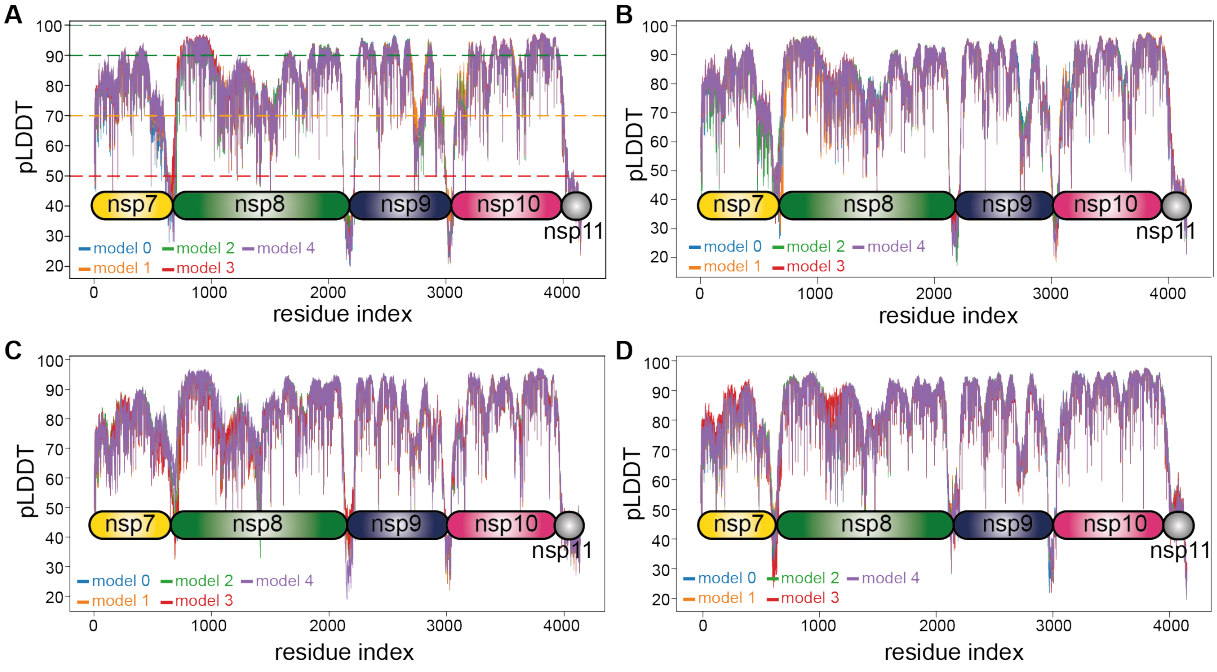


Figure 32: Local confidence scores of all AF3 models are shown for each CoV.
The predicted scores of the local distance difference test (pLDDT) are plotted against the residue index for the five nsp7-11 models of SARS-CoV-2 (**A**), SARS-CoV-1 (**B**), HCoV-229E (**C**) and MERS-CoV (**D**). Polyprotein schemes indicate approximate polyprotein region and emphasize low pLDDT scores across cleavage sites. The green dashed line marks a pLDDT score of 90, indicating a highly confident structural prediction. The orange dashed line corresponds to a pLDDT score of 70, above which the predicted backbone geometry is generally considered reliable. Regions with pLDDT scores below 50 (red) typically represent either intrinsically flexible regions or areas of low model confidence.

The selected AF3 models were subsequently analyzed in more detail. AF3 models with the best performance showed largely similar predictions of all cleavage sites among the four CoVs. There were slight differences at CS8/9 and CS9/10, where a short α-helix is predicted in SARS-CoV-2 and HCoV-229E (cf. **Figure** 34 **B**, **C**). Additionally, short α-helices were predicted at CS8/9 at P1'-P3' (NNE) in SARS-CoV-2 and HCoV-229E, and at P1-P5 (TVRLQ) in SARS-CoV-2 (**Figure** 34). However, AF2 only predicted an α-helix spanning CS7/8 for SARS-CoV-2 (**Figure 33**).

The local confidence scores are stored in the B-factor column of the output files allowing their visualization on the structure. The predicted nsp7-11 AF3 models across all four CoVs showed that the nsp domains nsp7-nsp10 were predicted to be folded proteins with similar local confidence scores. These were usually above 70, indicating medium to high confidence. The pLDDT scores at the cleavage sites and nsp11 regions were below 60 indicating rather low confidence (cf. **Figure 32**, **Figure S11**-**Figure S14**). Consequently, these regoins were predominantly predicted as disordered regions. An exception was the CS7/8 region, which was predicted as an α-helical structure across all species. Here, the pLDDT-values fluctuate between 30-50 representing a low confidence. Although the α-helix shows low confidence scores, it provides a plausible explanation for the slow kinetic rates of CS7/8 together with our experimental data. Notably, the AF2 model showed higher confidence scores (> 70) for the CS7/8 region, while the other three cleavage site regions displayed similarly low confidence scores (30-50), consistent with the AF3 model (cf. **Figure 33**, **Figure S11**, **Figure S15**, **Table S8**).

Next to the pLDDT value, the predicted aligned errors (PAE) are estimating the confidence of how well two residues of two different domains are placed within the predicted structure. PAE scores show low confidence considering the spatial organization of the individual domains pointing to a flexible spatial organization of nsp7-11 domain (**Figure S16**). Altogether, the nsp7-11 models resembled beads on a string, with globular nsp domains linked by flexible cleavage sites, consistent with structures suggested from integrative modelling and previous SARS-CoV-1 cleavage results[258,295].

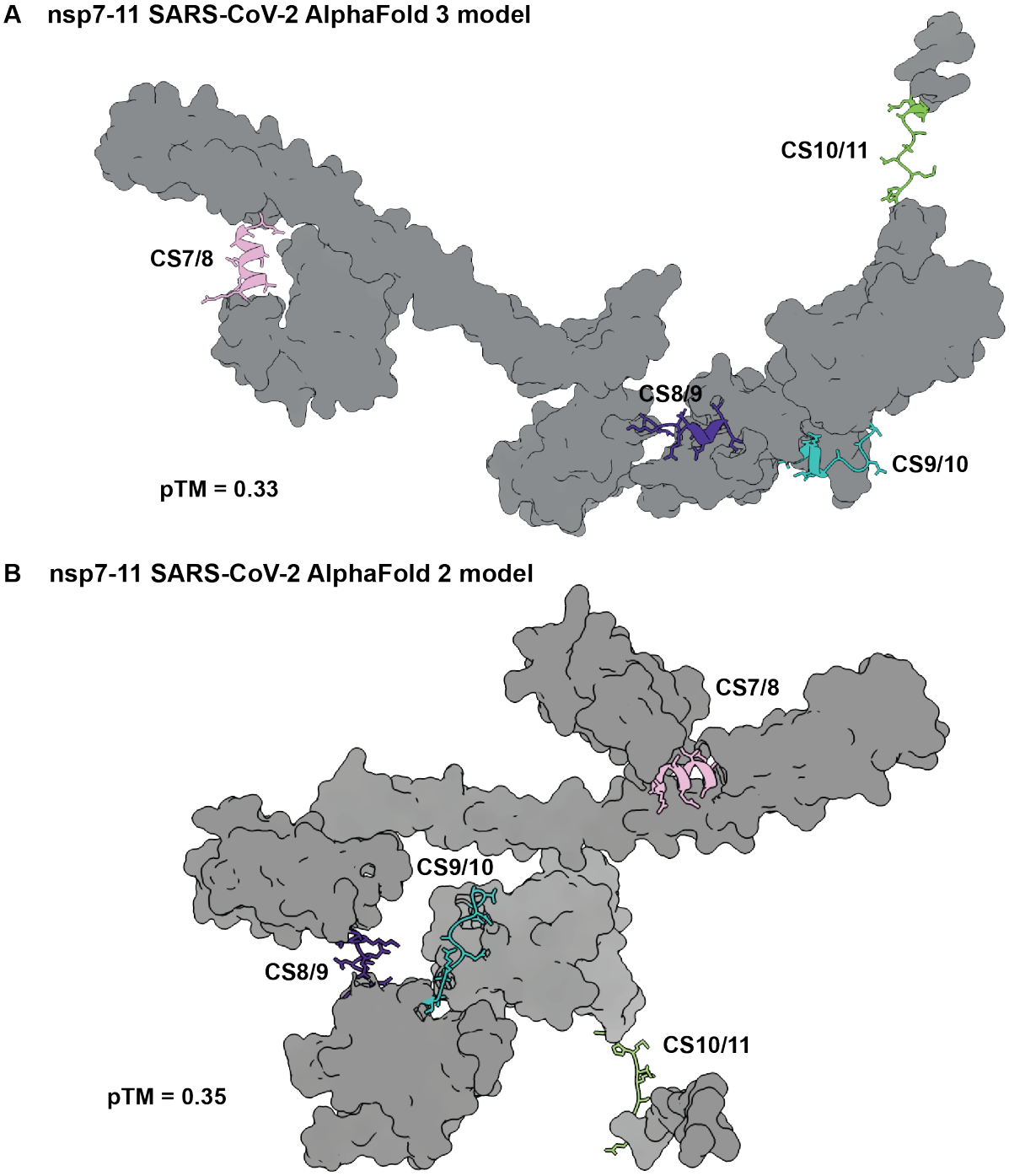


Figure 33: Structure prediction by AF of SARS-CoV-2 nsp7-11 polyprotein.
Predicted overall folding of nsp7-11 of SARS-CoV-2 with four cleavage sites: CS7/8 (pink), CS8/9 (purple), CS9/10 (cyan) and CS10/11 (green). Two prediction models are depicted, showing the best model from AF3 prediction (**A**) and AF2 prediction (**B**).

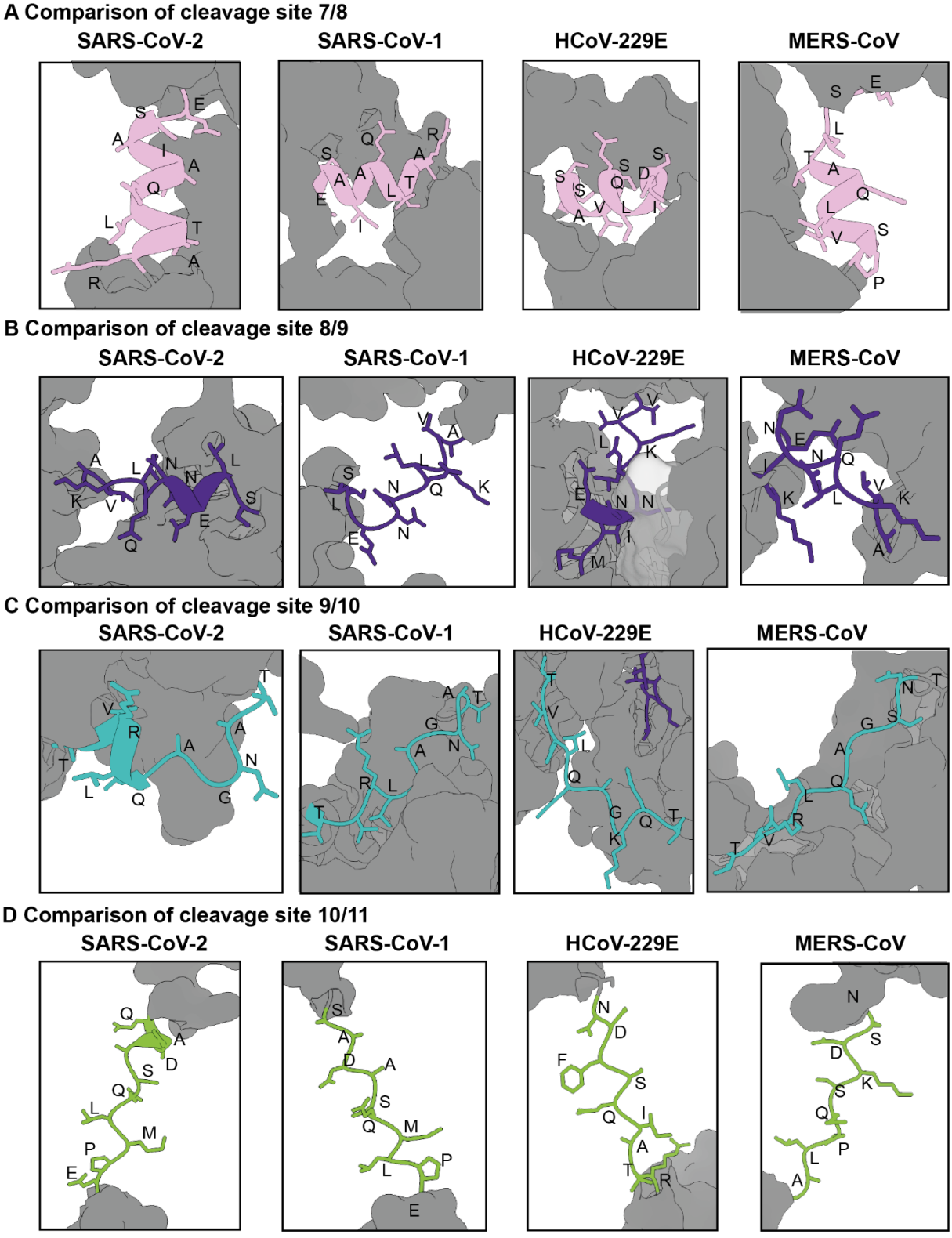


Figure 34: Comparison of the four cleavage sites between the four different hCoVs.
Zoom in on each predicted cleavage site region P6 to P6'. (**A**) Predicted α-helical structure spanning CS7/8 (pale pink) is shown in comparison. (**B**) CS8/9 is shown in purple and predicted to be partly α-helical depending on the CoV species. (**C**) CS9/10 is depicted in cyan and is predicted to be partly folded as α-helix in SARS-CoV-2. (**D**) CS10/11 in green is predicted to be disordered and elongated in all four CoVs.

3.4. Discussion

In this work, we characterized the nsp7-11 polyprotein processing kinetics of four CoV species SARS-CoV-1, SARS-CoV-2, HCoV-229E and MERS-CoV. Using native MS, we quantified multi-reaction kinetics and determined rate constants *k* for all four contained M^{pro} cleavage sites simultaneously. Our analysis revealed both conserved and unique features in nsp7-11 processing reactions.

For fast rates, we used a continuous monitoring approach, where the ongoing enzyme reaction was performed 'in-capillary' at 27°C. For slow rates at 0˚C (on ice), we used a discontinuous approach, with processing reactions performed in a test tube, and sampled discontinuously over time. The two measurement approaches were the basis for a detailed analysis of M^{pro}-mediated processing (**Figure 15 A**). While time-resolved measurements have been done previously for other samples[296], having identical starting points was an essential requirement to extract kinetic data of high quality. After establishing a custom Python script, we used it to extract rate constants *k* of CS7/8, CS8/9 CS9/10 and CS10/11.

In the following discussion, the experimental results, the predicted structural models and the primary structure are assessed and compared to get a comprehensive picture of polyprotein processing in human CoVs. Furthermore, technical aspects, advances and limitations are also discussed.

3.4.1. AlphaFold models reveal structural features of polyprotein processing

We predicted in total twenty AF3 models for the four CoVs and selected the best four models for each strain with a specific focus on the local confidence score at cleavage site regions. We observed overall low confidences for the cleavage sites considering pLDDT and PAE. AF3 is a powerful tool, but it has limitations when predicting intrinsically flexible regions or domains[250,297]. Nsp7-11, as a polyprotein, appears to be inherently built to exhibit a certain degree of flexibility[298]. Therefore, the low confidence scores in the predictions are not entirely unexpected. In the following, prediction models are compared to each other and discussed together with our experimental data. Overall, the models were similar across the four CoV species, though the spatial arrangement of the protein subdomains were slightly different between the top models of each CoV species. In the light of our experimentally determined kinetic rate constants, particular attention was given to the cleavage sites and their surrounding structural environments.

There were slight differences at CS8/9 and CS9/10, where SARS-CoV-2 was predicted to have a short α-helix at both sites, while HCoV-229E was predicted to have a short α-helix at CS8/9. Additionally, short α-helices were predicted at CS8/9 at P1'-P3' (NNE) in SARS-CoV-2 and HCoV-229E, and at P1-P5 (TVRLQ) in SARS-CoV-2 (Figure 34). In spite of the very low local

confidence values at the cleavage sites, these predictions were largely in agreement with the experimental results. For instance, CS9/10 of MERS-CoV appeared to be the most accessible and the most elongated cleavage site, which matched the observation that the nsp9-11 intermediate species was clearly the most abundant in MERS-CoV compared to the other strains. Furthermore, CS9/10 was predicted to comprise an α-helix (P1-P5) in SARS-CoV-2, but not in SARS-CoV-1, which corroborates the significantly faster processing of CS9/10 in SARS-CoV-1 compared to SARS-CoV-2. However, while CS8/9 in SARS-CoV-2 and SARS-CoV-1 exhibited similar conversion rates, the models predicted a short α-helix for SARS-CoV-1 but not for SARS-CoV-2. The existence of such a short α-helix would be expected to reduce conversion at CS8/9 in SARS-CoV-1, yet the kinetic rates were nearly identical for both SARS-CoV-1 and SARS-CoV-2.

The local confidence scores indicate that the structural predictions of the short α-helices should be interpreted with caution. Particularly, considering that AF2 model of SARS-CoV-2 nsp7-11 only predicted an α-helix at CS7/8. Nevertheless, apart from the predicted α-helix at CS8/9 in SARS-CoV-1 and -2, the predictions are consistent with experimental observations and help to visualize and interpret the observed kinetic rates. These predicted models thus serve as useful complementary tools for understanding structure-function relationships, particularly when integrated with the experimental kinetic data. Interestingly, according to the pTM, the AF2 model performed slightly better than the AF3 model. However, both SARS-CoV-2 nsp7-11 models showed a pTM below 0.35. AF2 structure exhibits a more compact folding similar to the I-TASSER models from Yadav *et al.* that included constraints from experimental data[83]. Therefore, it is tempting to say that the more compact folding reflects the 'real' polyprotein structure. However, protein structures are not rigid, but rather dynamic, particularly within the cellular environment[299]. A study on AF2 models concluded that low pLDDT scores not necessarily represent low confident predictions, but rather reflect a high degree of flexibility at these positions[300]. Given the similar pTM scores and the similar pLDDT scores of the AF3 and AF2 models, none of them can be favored over the other. It is plausible that both structures represent alternative conformers of the nsp7-11 polyprotein. Given the flexible and dynamic nature of this region, both AF2 and AF3 approached the limits of their predictive capabilities[297,301,302]. Additionally, AF performance depends on its training data, which comprise fewer than 10% viral protein structures[255,256,301]. AlphaFold Structural Database (AFDB) contains more than 200 million AF2 predicted structures from UniProtKB[248,303]. However, viral proteins were excluded, probably due to the reason that viral protein strucutres are underrepresented in PDB[255,256]. Consequently, prediction accuracy is oftentimes low as in our study.

A low PAE indicates low confidence in the relative positioning of two residues, which may reflect intrinsic structural flexibility or insufficient data for accurate prediction. Assuming PAE scores indicate a flexible spatial organization of the nsp7-11 polyprotein, this would align with the apparent difficulty in crystallizing nsp7-11, as no crystal structure has been resolved to date. Additionally, it has not been possible to obtain a high-resolution structure of the full nsp7-11 using cryo-EM. Only the segment captured at a stalled M^{pro} cleavage site has been structurally characterized[152].

Analysis of the determined conversion rates revealed notable correlations with structural predictions. The largest variation in conversion rates occurred at CS8/9, CS9/10 and CS10/11, where AF3 predicted largely disorder in the corresponding linker regions. Across all four strains, CS7/8 had relatively slow cleavage rates and was predicted α-helical (**Figure** 34). As α-helices generally serve as poor protease substrates, a significant structural transition would be required explaining slower kinetics[304–306]. This is further supported by CS7/8 adopting a disordered conformation when crystallized with M^{pro}, indicating that the structure indeed has to adapt to the protease binding grove[307,308].

These findings help to explain the observed differences in conversion rates: The disordered regions at CS8/9, CS9/10, and CS10/11 can rapidly undergo structural changes to accommodate M^{pro}, while the stable α -helical structure at CS7/8 requires more reorganization and more time to fit the protease active site. However, our $k_{CS7/8,\,0°C}$ suggest that this α-helical structure exhibits distinct unfolding dynamics with lower stability in MERS-CoV and higher stability in SARS-CoV-1. Moreover, nsp8 undergoes partial conformational changes during processing, which may contribute to the observed delayed cleavage kinetics at its associated cleavage sites[83].

Eventually, despite sequence similarities between CS7/8 and CS9/10, our experimental data showed reduced constant rates $k_{cs7/8,\,0°C}$. AF3 models of all four CoVs predicted an α-helix at CS7/8 and thus providing a logical explanation for reduced $k_{cs7/8,\,0°C}$. This is consistent with other model predictions and data in literature[307]. It can therefore be concluded that the structural feature of an α-helix most likely leads to a reduced conversion rate for CS7/8. This suggests that nsp7 and nsp8 are liberated and mature later than expected, acting as co-factors of nsp12 (RdRP). This is likely to be a regulatory aspect of polymerase activity. For instance, this could ensure that proof-reading and capping functions, for which nsp9 and nsp10 are required, are available before polymerase processivity ramps up.

The predicted structure of SARS-CoV-2 nsp7-11 offered insights into its interaction with nsp16. In our experiments, a low-abundant nsp7-11+16 complex was observed, while strong binding occurred only after nsp10 release. Existing nsp16+10 crystal structures, do not explain this requirement. Structural alignment of nsp10 from the crystal structure and nsp10 from AF3

model revealed that the α-helix downstream of CS9/10 in nsp7-11 clashes with the nsp16 interface (cf. **Figure 35**). This *N*-terminal region is absent in the crystal structure (6W4H[309]) as it is based on truncated nsp10 (blue). Given the dynamic nature of proteins in solution, a transient state likely permits weak binding of nsp7-11 to nsp16 prior to nsp10 release, and is therefore detected with native MS.

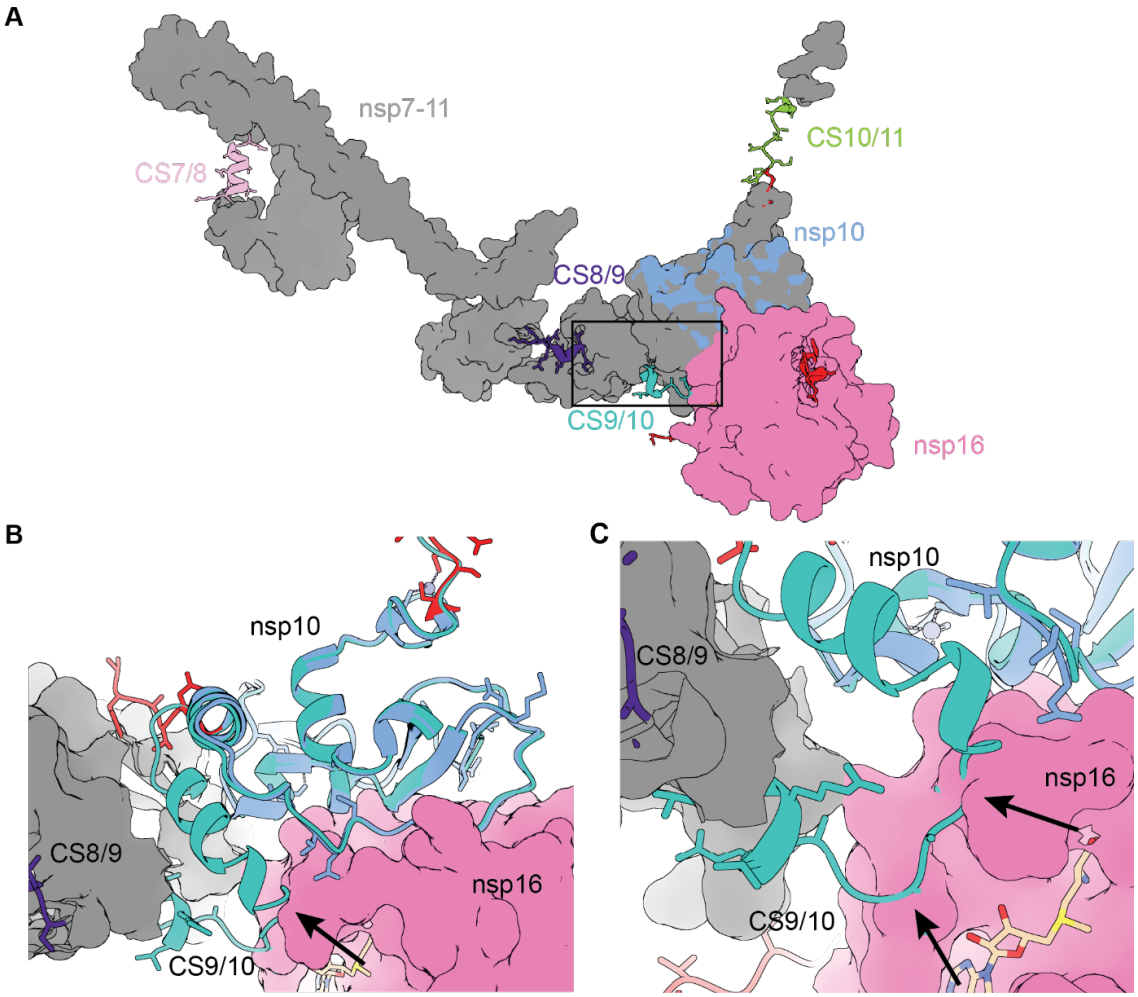


Figure 35: SARS-CoV-2 nsp7-11 AF3 model aligned with crystallized nsp16+10 complex structure. (**A**) Predicted model of SARS-CoV-2 nsp7-11 (grey) is aligned with crystallized nsp16+10 complex (blue/magenta, 6W4H)[309]. *C*- and *N*-terminal residues of nsp16+10 crystal structure are colored in red. Cleavage site regions of nsp7-11 (grey) are depicted. Black frame indicates the approximate region zoomed in on in (B) and (C). (**B**) Zoom in on aligned nsp10 (turquoise) of nsp7-11 prediction model and nsp16+10 complex (blue/magenta)). Black arrow indicates the α-helix (turquoise) that clashes with nsp16 interface. (**C**) Zoom in on *N*-terminal nsp10 region. Black arrows highlight the clash of nsp10 (from nsp7-11) and nsp16 (magenta).

3.4.2. Primary structure, polyprotein folding and implications on RTC assembly

While the prediction models provided a reasonable explanation for the reduced kinetic rates at CS7/8, they could not explain the significant differences at CS10/11. This is particularly relevant given the predicted elongation and accessibility of CS10/11. Here, a closer examination of sequences, predicted structures, and our determined conversion rates provided insights into the cleavage mechanism at each site. In the following, a more detailed look at CS8/9, CS9/10 and CS10/11 is provided. Additionally, we demonstrated that while complete

processing of nsp7-11 is not essential, it greatly enhances methyltransferase complex assembly. Here, we discuss structural implications of our findings and evaluate our approach against conventional techniques.

Across all tested species, cleavage sites CS7/8, CS8/9, and CS9/10 contained the typical M^{pro} consensus sequence elements: L at P2 and Q at P1[111,310–313]. Despite these primary structure similarities, their conversion rates varied significantly. For CS7/8, as discussed above, its secondary structure appeared to be the key factor inhibiting cleavage. At CS9/10, all tested strains shared sequence conservation from P5 to P2', and cleavage occurred with relatively high efficiency (cf. **Figure 27**, **Figure 36**). Previous studies demonstrated that residues up to P6 in CS9/10 interact tightly with M^{pro}, explaining the consistent processing efficiency across strains at this site. The conversion rates of CS8/9 showed unexpected variability between strains, exhibiting the highest measured rate in MERS-CoV but slow rates in HCoV-229E. This variability was particularly surprising given its highly conserved non-canonical sequence, notably the NNE at P1'-P3' being essential for nsp9 PTMs in CoV transcription[314]. The only differences among species at CS8/9 specific to HCoV-229E that could explain an altered interaction with M^{pro} was a P5 A-to-V substitution, suggesting these substantially inhibit CS8/9 cleavage. MERS-CoV however shares the P5-P1 with SARS-CoV-1 and -2 suggesting that other flanking residues or structural aspects determine this highest conversion rate. Hence, our findings indicate that variations in conversion rates likely arise from specific structural features within or surrounding the cleavage sites, rather than primary structure alone.

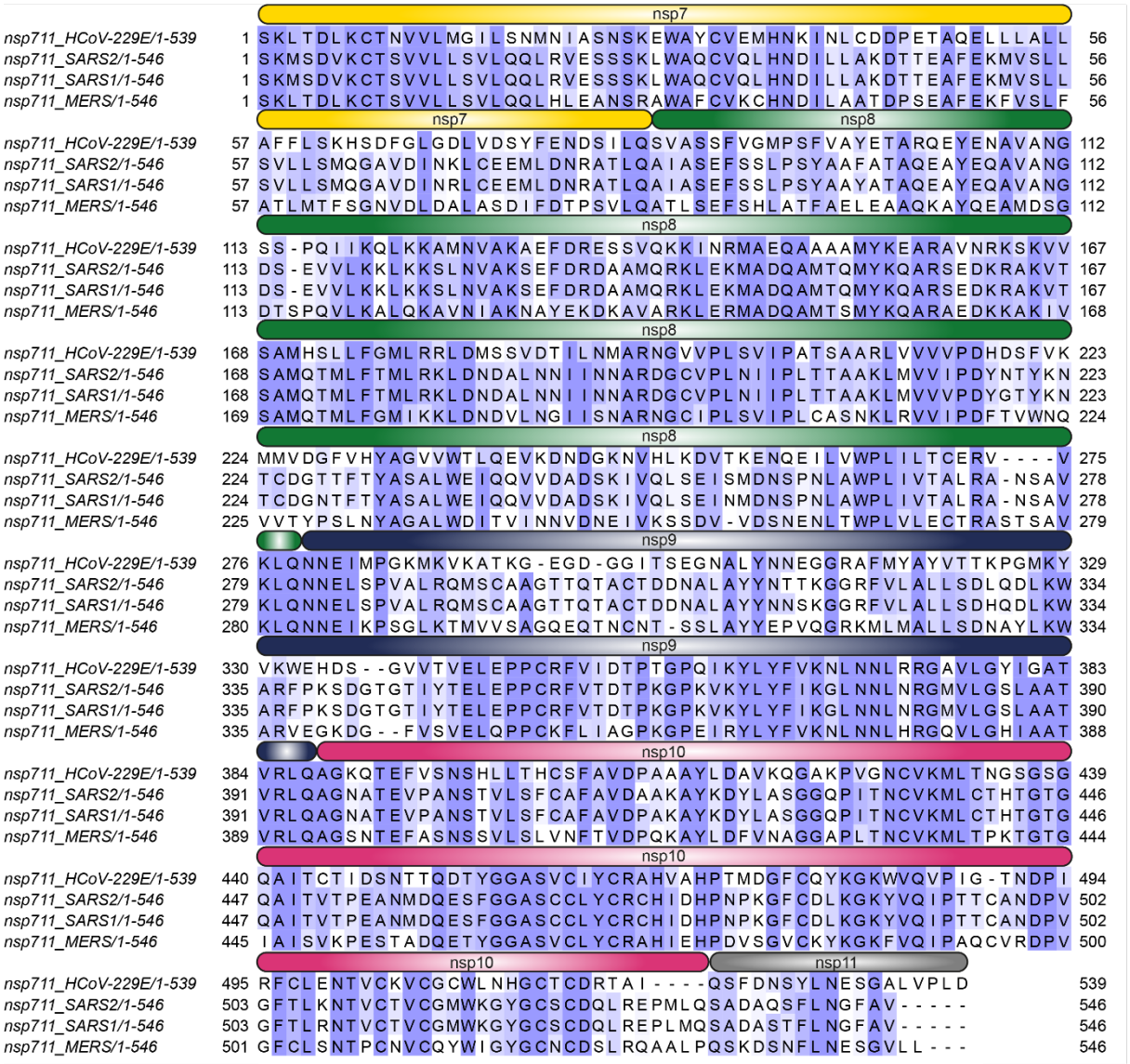


Figure 36: Multiple sequence alignment (MSA) of the four hCoVs tested.
MSA is colored with Blosum62 Score scheme. Gaps are colored white and residues matching the consensus sequence are colored dark blue. If the residue does not match the consensus sequence, but the Blosum62 matrix gives a positive score, it is colored light blue.

The CS10/11 provided particularly valuable insights. CS10/11 showed high-efficiency cleavage in SARS-CoV-1 and -2, but low efficiency in HCoV-229E and in MERS-CoV. CS10/11 showed relatively low sequence conservation among tested species, with notable substitutions at the P2 position of the M^{pro} consensus sequence (**Figure 36**, **Figure S17**). In this position, the typically conserved L is replaced with M in SARS-CoV-1, I in HCoV-229E, and P in MERS-CoV. A detailed study of M^{pro} substrate specificity supports our findings, showing that M at position P2 still permits moderate cleavage efficiency, while other substitutions result in lower efficiency[315]. In our structural models, CS10/11 occupies a peripheral, exposed position, which likely enables M^{pro} substrate recognition despite unfavorable sequence motifs.

It is important to note that all nsp7-11 substrates from the four CoVs (SARS-CoV-2, SARS-CoV-1, HCoV-229E and MERS-CoV) were cleaved using SARS-CoV-2 M^{pro}. While M^{pro}

consensus sequences are generally conserved across CoVs and structural superimposition of M^{pro} of the four CoVs tested show very similar structures (**Figure 37**), species-specific differences in protease-substrate interactions cannot be entirely excluded[311,316,317]. Future studies using matched species-specific M^{pro}-substrate pairs could further refine our understanding of these processing kinetics. However, this would further complicate the assignment of observed differences to specific features in the polyproteins.

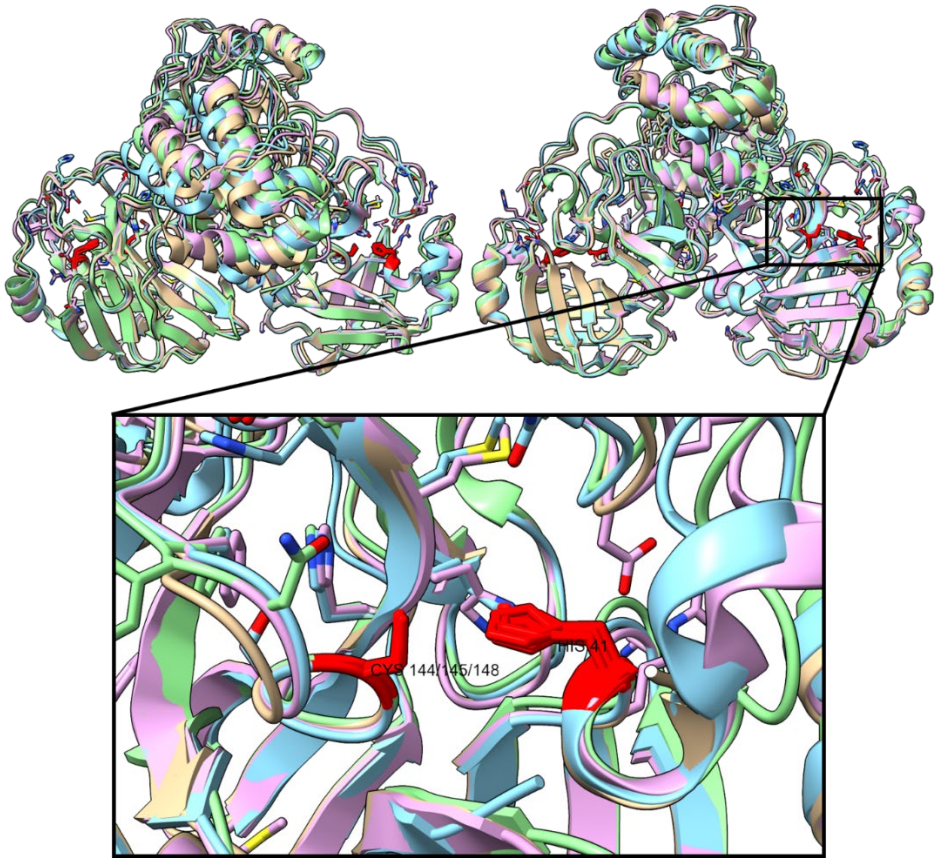


Figure 37: Superimposed crystal structures of M^{pro} from the four hCoVs tested.
Top panel shows M^{pro} dimers from two sites. Lower panel shows zoom in on catalytic site with the two catalytic residues marked in red. PDB structures: 6LZE in pink (SARS-CoV-2)[318], 1Q2W in gold (SARS-CoV-1)[319], 5WKK in light blue (MERS-CoV)[320], 2ZU2 in light green (HCoV-229E)[321].

Nevertheless, the literature provides useful insights into previous approaches using chimeric M^{pro} constructs that are worthy of attention. Denison and coworkers investigated M^{pro} activity across genogroups using chimeric MHV encoding M^{pro} from α- and β-CoVs *in vivo*[322]. Chimeric MHV using M^{pro} within the same genogroup lineage showed efficient replication, although viral fitness was reduced compared to the originating virus. In contrast, engineered chimeric MHV from α-CoVs showed no viral recovery. The results indicate that a possible difference in timing and cleavage efficiency of M^{pro} in a heterologous background could be the reason for reduced viral recovery[322]. However, it remains unknown to what extent these findings affect *in vitro* processing with chimeric M^{pro} as *in vitro* experiments usually simplify heterogeneity. Therefore, additional species-specific processing kinetics would give useful information on the chimeric

use of M^{pro} in *in vitro* cleavage assays. In the end, using SARS-CoV-2 M^{pro} was important to achieve better comparability between nsp7-11 substrate of the four CoVs.

There was a rapid conversion of CS10/11 in both SARS-CoV-2 and SARS-CoV-1. However, this was not observed for MERS-CoV and HCoV-229E, where low kinetic rates were found instead. The prediction models indicate an elongated structure in the vicinity of CS10/11, with optimal accessibility for M^{pro} binding. Importantly, the low efficiency CS10/11 proved non-essential for formation of the RTC sub-complex nsp16+10-11 (**Figure 29**). This is consistent with a low sequence conservation among the four CoVs (**Figure 36**, **Figure S17**). In a study by Yadav *et al.*[83], HDX-MS experiments revealed that nsp11 shields its adjacent nsp10 region from deuterium uptake, suggesting it effectively covers this interface in the native structure in SARS-CoV-2. Subsequently, integrative modeling by using constraints from HDX-MS and XL-MS data showed an α-helix at CS10/11, which suggests a reduced conversion rate of CS10/11 in SARS-CoV-2 (cf. **Figure 2 B**)[83].

In comparision, the findings in this work suggest that CS10/11 cleavage efficiency is neither conserved across hCoVs nor required for complexation with nsp16. This possibly reflects CS10/11's origin as a secondary product from the CoV ORF1a/ab RNA frameshift region that is required for the translation of pp1ab (nsp1-nsp16). Notably, nsp10 is followed by nsp12 in approximately one-third of the CoV polyprotein that is expressed as pp1ab. Since nsp12 shares the same *N*-terminus as nsp11, the cleavage sites CS10/11 and CS10/12 are structurally similar by design. This raises the intriguing possibility that a long-lived nsp10-12 intermediate could exist and form a complex with nsp16, resulting in an nsp10-12+16 super-complex.

The cleavage order proposed also illustrates the ongoing debate regarding the initial cleavage site. Based on pulsed HDX-MS and SDS-PAGE, Yadav *et al.* suggested the following cleavage order: CS9/10 > CS8/9 > CS10/11 >> CS7/8[83]. However, our findings contradict these data showing a preferred cleavage order with CS10/11 as SARS-CoV-2 fastest cleavage site: CS10/11 > CS8/9 ≈ CS9/10 >> CS7/8. Next to the cleavage order proposed by structural MS-techniques, an advanced FRET assay using a linked protein platform suggested a third variant cleavage order: CS9/10 > CS7/8 > CS8/9 > CS10/11. Collectively, there are three different biophysical techniques proposing three different cleavage orders of nsp7-11 polyprotein processing. The study using pulsed HDX-MS and integrative modelling and our study using native MS employed full-length polyprotein and thereby included polyprotein folding. Both techniques agree on a slowly converted CS7/8, but not on whether CS9/10 or CS10/11 is targeted first[83–85]. In fact, CS10/11 is claimed to be the fastest cleavage site in this study, but clearly not in the studies using the novel FRET-platform or using HDX-MS. However, while the novel FRET-platform provides cleavage site expressed in a linked protein, the cleavage order

still reflects the preference on cleavage efficiency and primary structure. Furthermore, compared to HDX-MS that give indirect information of intermediate products, native MS gives direct evidence by detection of all cleavage products in parallel.

As cleavage of CS10/11 occurs rapidly in SARS-CoV-2 in our work, it could be possible that the depletion of the polyprotein had progressed already so far that the experimental design of Yadav *et al.* missed out the fast cleavage of CS10/11. Especially as they have used a higher ratio of polyprotein nsp7-11 to M^{pro} (1:1) and incubated at RT. We have observed a similar scenario during the processing of nsp7-11C using the discontinuous approach (cf. **Figure 20**). Here, the rate of CS10/11 is no longer the fastest due to the depletion of the initial substrate, nsp7-11. Another interesting aspect is that nsp7-11 was frozen and thawed for the HDX-MS measurements. A discontinuous polyprotein processing experiment using frozen nsp7-11 from SARS-CoV-1, under identical conditions, revealed a different cleavage order based on the kinetic rates, with CS10/11 exhibiting a slower rate likely due to advanced substrate depletion (**Figure S18**). These results may be not transferred one to one as polyprotein processing was conducted in different buffer conditions. However, the negative impact of freezing on the folding of polyproteins is evident.

Our research illuminates the sophisticated relationship between CoV polyprotein processing and nsp complex formation. Notably, functional chimeric complexes can even form between components from different CoV species, such as SARS-CoV-2 and MERS-CoV. Coordinated processing generates multiple proteoforms with distinct functions, a common viral strategy[143,323]. The CoV methyltransferase complex demonstrates remarkable flexibility. While it can bind unprocessed polyprotein nsp7-11, it clearly prefers mature and half-mature products nsp10 or nsp10-11. The integrative model of Yadav *et al.* for SARS-CoV-2 nsp7-11 suggests that nsp9 and nsp10 are in close proximity, which could explain why the polyprotein is a poor binder of nsp16[83]. Our observation of consistently delayed CS7/8 cleavage across all species hints at a possible regulatory mechanism for the assembly of CoV polymerase complexes.

This spatiotemporal coordination likely orchestrates the sequential formation of various functional assemblies. The processivity-enhancing nsp7 and nsp8 subunits would join nsp12 later to form the polymerase complex while nsp9 and nsp10 would become available earlier in the viral lifecycle. This is particularly interesting considering the critical nsp10-dependent complexes, including the proofreading nsp14+10, methyltransferase nsp16+10 and the recently reported ternary complex nsp10+14+16[324,325]. Such regulated processing would ensure that RNA capping and proof-reading are in place first, potentially allowing the virus to fine-tune RNA synthesis and modification. The ordered and regulated polyprotein processing parallels cleavage processes in alphaviruses, another group of enveloped positive-sense RNA

viruses[326]. In both virus families, timely and precise polyprotein processing is crucial for viral replication, yet depends on factors beyond sequence accessibility[327,328].

3.4.3. Technical novelty, approach and limitations

This section outlines the key advances of the established approach compared to conventional approaches that are usually based on peptide cleavage assays. Furthermore, the experimental approach is discussed highlighting the rationale behind key technical choices and addressing methodological considerations such as ion suppression, dynamic range, and the optimization of protein concentrations.

A key advantage of our approach was the use of folded polyprotein substrate, presenting cleavage sites in their native structural context. By using tag-free nsp7-11 with natural amino acid sequence, we aimed to replicate authentic M^{pro}-polyprotein interaction dynamics. In contrast, conventional techniques typically use either natural libraries or artificial substrates such as labelled or unlabeled peptides or short cleavage site sequences expressed between reporter proteins. These methods offer advantages in throughput and automated readout, enabling broad sequence space and condition sampling[315,329,330]. However, in a previous study it was shown that peptide-based cleavage sites not reflect native structural dynamics, potentially producing misleading results[258]. For example, FRET-based assays of SARS-CoV-1 and MERS-CoV suggested high conversion rates for CS7/8, indicating a different processing order than in our findings[331,332]. Thus, our approach offers a detailed and quantifiable perspective to dynamic reactions and serves as blueprint for multi-cleavage reactions in general. However, a development regarding higher throughput would be desirable.

Even though the temperature used in our experiments is lower than the typical temperature at which the catalytic constant k_{cat} is measured, the *k*-values determined using native MS are lower than those obtained using peptide-based cleavage assays[258,331,332]. The main reason for this is that the peptide substrates are usually added in excess, and therefore the enzyme is always fully occupied. M^{pro} is part of pp1ab, released by autocleavage and then becomes active as a dimer resulting in a twofold excess of the substrate *in vivo*[333]. Thus, the competition of different cleavage sites for the protease active site and similar concentrations of substrate and enzyme reflect the *in vivo* situation much better. We therefore conclude that our approach using native protein sequences in folded nsp7-11 polyprotein better represents authentic processing reactions.

Native MS enabled characterization of the dynamic landscape of protein species, including non-covalent complexes[334]. However, this approach required using ammonium acetate as a buffer surrogate and low temperatures to prevent nsp8-mediated complex aggregation[335]. Our in-capillary experiments at a temperature of 27°C showed decreasing reaction rates for

CS10/11 over time, likely from natural substrate depletion during the ongoing process. However, influences from the elevated temperature or in-capillary acidification, which happens during prolonged nano-ESI[336], cannot be entirely excluded. Nevertheless, linearization of data clearly reveals when the assumption of first-order kinetics is no longer valid and *k* can simply be extracted from these data points. Undersampling or too little data points is fortunately not an issue in the continuous approach as spectra were recorded at rates far higher than 10 Hz at the employed resolution settings of 6250. Additionally, temperature-controlled experiments could yield Arrhenius plots, providing enthalpic and entropic energies of the reactions[337]. To rule out capillary-based biases, similar multi-cleavage experiments could benefit from automated LC online-buffer exchange, which would standardize the sampling timepoints and allow the reaction to be performed in standard buffer systems until the moment of sampling[338]. However, this approach would require more complex instrumentation and sample handling. Direct mass detection of intact protein intermediates enabled both extraction of conversion rates and comprehensive insights into multi-cleavage reaction kinetics. A complementary study using XL-MS and HDX-MS examined SARS-CoV-2 polyprotein processing, revealing novel spatial and dynamical information about M^{pro}-nsp7-11 interactions[83].

Considering the detection of both low- and high-abundance species, the question arises whether the dynamic range limit of the Orbitrap instrument has been reached. Sample concentration directly impacts the dynamic range of a mass spectrometer for several reasons. The most abundant ion species can saturate the detector, ion suppression effects increase especially in complex mixtures and space charge effects become more pronounced[288,289]. Space charge effects occur when the density of ions becomes high enough that repulsive forces (Coulombic repulsion) between ions significantly affect their behavior. To assess this, ion intensities for nsp7, nsp10-11, and nsp7-11 were examined (**Table S1**). Given the Orbitrap's wide dynamic range and our observation that ion suppression was evident at ~ 30 µM but not at ~ 20 µM total protein concentration (**Figure 31**), the concentrations used (< 25 µM) appear appropriate. Nonetheless, to further minimize potential artifacts from space charge and ion suppression, particularly near adjacent peaks, slightly lower concentrations may be advisable.

Orbitrap systems provide a wide range of adjustable settings, allowing optimal tuning for specific analytes. In this study, polyprotein processing yielded over 15 distinct mass species, ranging from 1.5 kDa (nsp11) to 67 kDa (M^{pro}_2). For these experiments, we were particularly interested in intermediate species, so Tune settings were optimized for masses in the ~ 20-70 kDa range. Parameters such as an interior heated capillary temperature of 100-150 °C, 15 eV in the source region, and 25 eV in the HCD cell are not ideal for preserving protein-protein interactions in smaller dimers, such as the nsp7 dimer (~ 18 kDa). As a result,

HCD products were observed in MS1 spectra, including dissociated nsp7 (cf. **Figure S5 C**). However, the use of the HCD cell was not required during the processing measurements, but only for the identification of selected peaks such as nsp7-8 polyprotein and nsp7+8 heterodimer. Both the Orbitrap and C-trap have high space charge capacities, but to eliminate any potential risk of space charge repulsion, it would have been advisable to omit the HCD cell[221,339]. More importantly, this did not affect the extracted cleavage site kinetics, as mature nsps were excluded from the analysis, and gentler conditions were applied where necessary, such as during K_D measurements.

In summary, we directly monitored M^{pro}-mediated polyprotein processing, capturing the structural context of the cleavage sites. This offers an advantage over conventional peptide-based assays. Cleavage kinetics were extracted using two complementary approaches: one enabling measurements at near-physiological temperatures, and the other providing detailed kinetic rate analysis. While the usage of HCD cell could have been omitted from each measurement, we established a balanced experimental setup optimizing protein concentrations to support nsp interactions while maintaining suitable sample conditions with respect to dynamic range and ion suppression.

4. Summary and outlook

In this work, we analyzed CoV polyprotein processing across four human CoV species, revealing dynamic intermediate products, cleavage site conversion rates, and the interconnection between processing and complex formation. A sensitive and precise native MS approach was established, providing novel insights into processing kinetics of CoVs, and enabling determination of kinetic rate constants for the four cleavge sites, including the structural context of the nsp7-11 polyprotein. Processing kinetics of the four human CoVs demonstrated both conserved features and species-specific variations in nsp7-11 processing (**Figure 38**, **Figure 39**). We established that while complete processing enhances nsp16+10 complex formation, it is not essential, and showed that functional complexes can form even between divergent human CoVs.

The structural analysis of cleavage sites was performed using AF models, sequence alignments and the kinetic rates constants, and revealed, how their structural environment contributed to processing efficiency. Cleavage rates were low at CS7/8 in all four CoVs species. Together with the AF models, this finding suggested that an α-helix was a structural hindrance and was likely a regulatory secondary structure that controlled the delayed release of nsp7 and nsp8. In conjunction with our kinetic data, the AF models provided further insights into the structural context of the cleavage sites, thereby enhancing our understanding of polyprotein processing.

Our methodology demonstrated that native MS is a versatile tool for investigating enzyme kinetics. Assets over conventional techniques are consideration of structural context and label-free substrates that are cheap to produce. A further advantage is the direct feedback on complex formation and stoichiometric ratios in one experiment. This improved mechanistic understanding of CoV polyprotein processing and complex formation may inform future antiviral drug development strategies targeting these essential viral processes.

There are a few experiments that could be realized quickly from here onwards and would significantly contribute to unravel the role of processing intermediates. Binding experiments with processed and unprocessed nsp7-11 and nsp16 suggested a potential regulatory role for the intermediates. A logical next step would be to investigate nsp16 complex formation in a time-resolved manner using the method established here. It is possible that processing intermediates would also interact with nsp16, given the observed interaction between nsp7-11 and nsp16, albeit at a low level. Testing the activity of a potential intermediate+nsp16 complex might reveal whether processing intermediates have regulatory roles beyond serving as precursors to mature nsps. Thus, further nsps such as the RdRP nsp12 or the helicase nsp13 could also be tested in time-resolved binding experiments to reveal the possible functions of

the processing intermediates. However, protein production protocols would need to be established for these nsps.

Our novel approach showed promise in advancing drug development strategies through its unique ability to provide a quantitative analysis of the kinetics of viral polyprotein processing. While the current approach had proven effective, several methodological enhancements could have further increased its impact and applicability. Two main areas for development have been identified: streamlining the native MS measurement process and optimizing protein purification workflows.

For this work, native MS measurements were conducted manually, limiting the number of samples that could be screened. These labor-intensive experiments would have benefitted from an automated nano-ESI platform, enabling high-throughput screening of various ratios using both continuous and discontinuous approaches. Additionally, such automation would have improved reproducibility[340,341]. Higher-throughput measurements would result in significantly larger datasets that would required efficient analysis. Our custom Python script, tailored to our specific scientific question, required manual peak list creation for each new construct and at least a control check for each measurement to validate the peak list. Additionally, the script did not inherently detect overlapping peaks. Therefore, a software would be needed that offers a peak-picking algorithm capable of automatically deconvoluting mass species, identifying overlapping peaks, and accurately resolving them using a peak distribution fitting model. MetaUniDec offered some of these features. However, for applications similar to this study, the software would have needed to be improved, for example, an enhanced control over peak picking and additional features for more in-depth data processing and export.

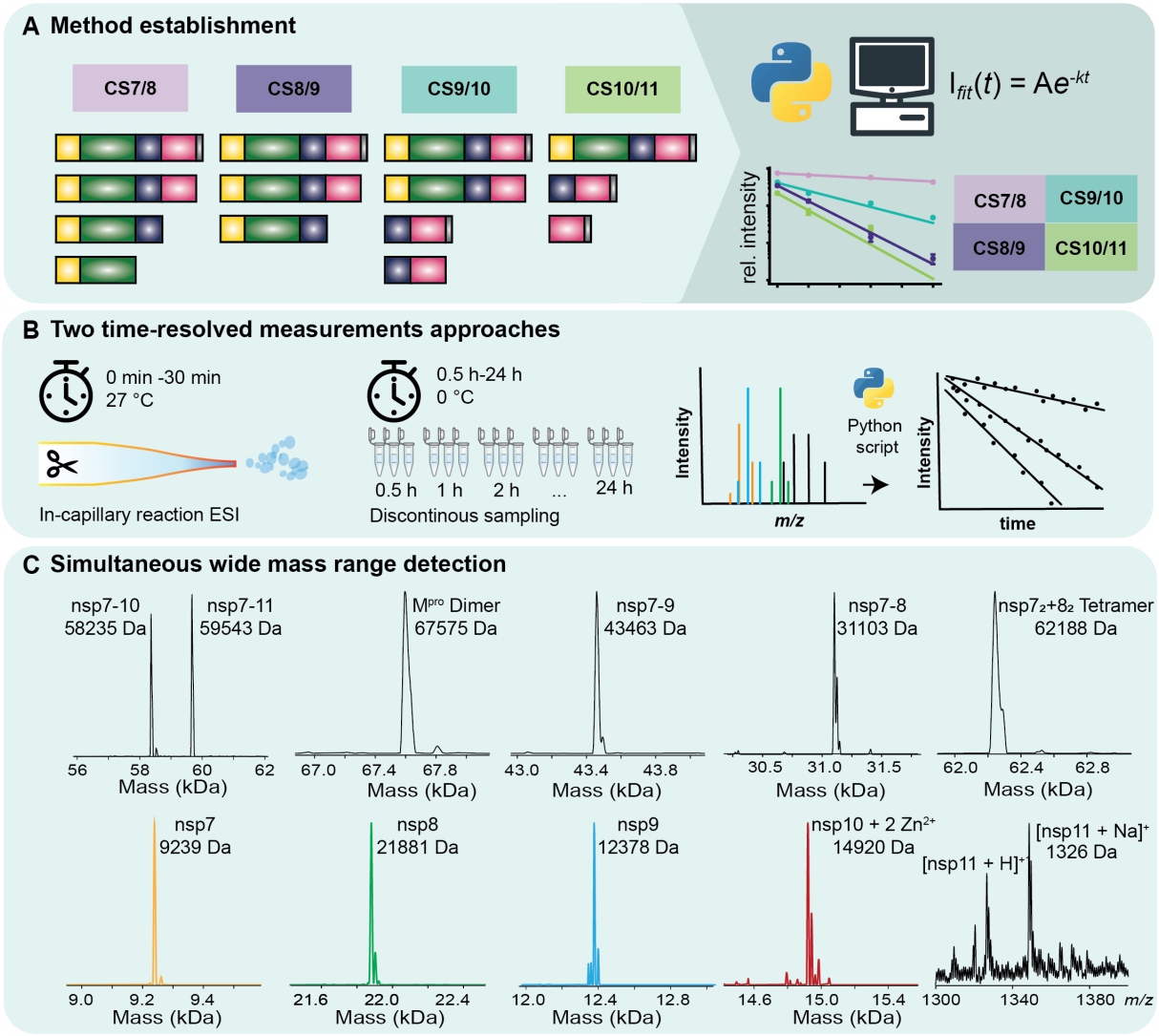


Figure 38: Graphical summary of the method establishment.
(**A**) A custom Python script was developed to import MS data, deconvolute peaks, and extract kinetic cleavage rate constants for four cleavage sites: CS7/8, CS8/9, CS9/10, and CS10/11. The resulting data were fitted using the specified exponential model. (**B**) To capture both rapid and slow cleavage events, two complementary measurement strategies were employed. Data from both approaches were processed uniformly using the developed Python script. (**C**) A major advantage of monitoring polyprotein processing with native MS is the simultaneous detection of all cleavage products. This is enabled by the broad mass detection range. For example, low-molecular weight species such as nsp11 (1.3 kDa) can be detected alongside larger complexes like the nsp7$_2$+8$_2$ tetramer (62.2 kDa).

The second development would aim for an optimized protein purification protocol. Although the polyproteins could be expressed in high yields, the samples showed a tendency to aggregate and behaved unpredictably when frozen[335]. As a result, samples had to be freshly prepared, significantly reducing efficiency. To address this, our group recently developed an elegant, streamlined protocol for handling low-yield or difficult proteins[342]. The so-called fast-track method saves time and enables the analysis of proteins that do not tolerate buffer exchange. Such rapid protocols are furthermore important if many nsps or partial polyproteins are to be provided at the same time to follow RTC assembly. If expression of the transmembrane devoid nsp7-16 is successful in mammalian cells or cell-free expression, yields will likely be low, so fewer purification steps will minimize losses. Therefore, it appears

realistic to look at processing of larger polyprotein constructs *in vitro* in the near future. This would enable the study of polyprotein processing and subsequent RTC assembly in real time. Combination with minimal RNA substrates could even be feasible, hence enabling precision biochemistry of a concerted set of enzymatic reactions driven viral replication. To achieve this, constructs should be designed with an M^{pro} cleavage site so that the purification tag is cleaved off during the processing reaction, enabling the investigation of nsp samples with authentic termini. It is still unclear whether the individual intermediates have an altered function during processing or whether the processing to mature nsps is merely a temporal regulation to 'switch on' the function of the liberated nsps. Native MS experiments designed in this way could provide valuable insights into the key stages of the RTC assembly pathway.

This work suggested that the order of cleavage is not conserved among CoVs. Studying the cleavage kinetics of nsp7-11 in more α- and β-CoVs, as well as in γ- and δ-CoVs might have revealed greater diversity or shared features. Significant divergence exists among CoV genera in the replicase gene, but structural insights remain limited in these genera[344,345]. However, interest appears to be increasing, as a recent study underscores showing genus-specific characteristics in the IBV nsp12 structure within the RTC[346]. Advancing our understanding of CoV replication across genera is essential for future outbreak preparedness, as γ-CoVs have demonstrated the capacity for cross-species transmission[345]. Therefore, applying our method to γ- and δ-CoVs would enhance our understanding of polyprotein processing in CoVs.

Further *in vitro* work should focus on the RTC assembly pathway and stoichiometry, as this remains poorly understood despite its critical importance for viral replication. Current debates surrounding RTC assembly center on the roles of PTMs, heterogeneous stoichiometry, and how different stoichiometric configurations relate to distinct functional states. Understanding this heterogeneity is crucial because different proteoforms and structural assemblies likely serve specialized functions within the replication machinery. Mapping the relationship between specific structural configurations and their corresponding functions would enhance our ability to identify targetable conformations. This knowledge would enable the rational design of small molecules that selectively disrupt essential RTC functions by targeting specific structural states, rather than attempting broad inhibition of the entire complex.

While the nsp7-11 polyprotein region of SARS-CoV-2 had been studied extensively *in vitro*, the next logical steps should either return to *in vivo* validation of these findings or investigate larger polyprotein constructs that better recapitulate the full complexity of viral polyprotein processing. The next section discusses further polyprotein processing experiments that can be performed with existing technologies and that go beyond native MS.

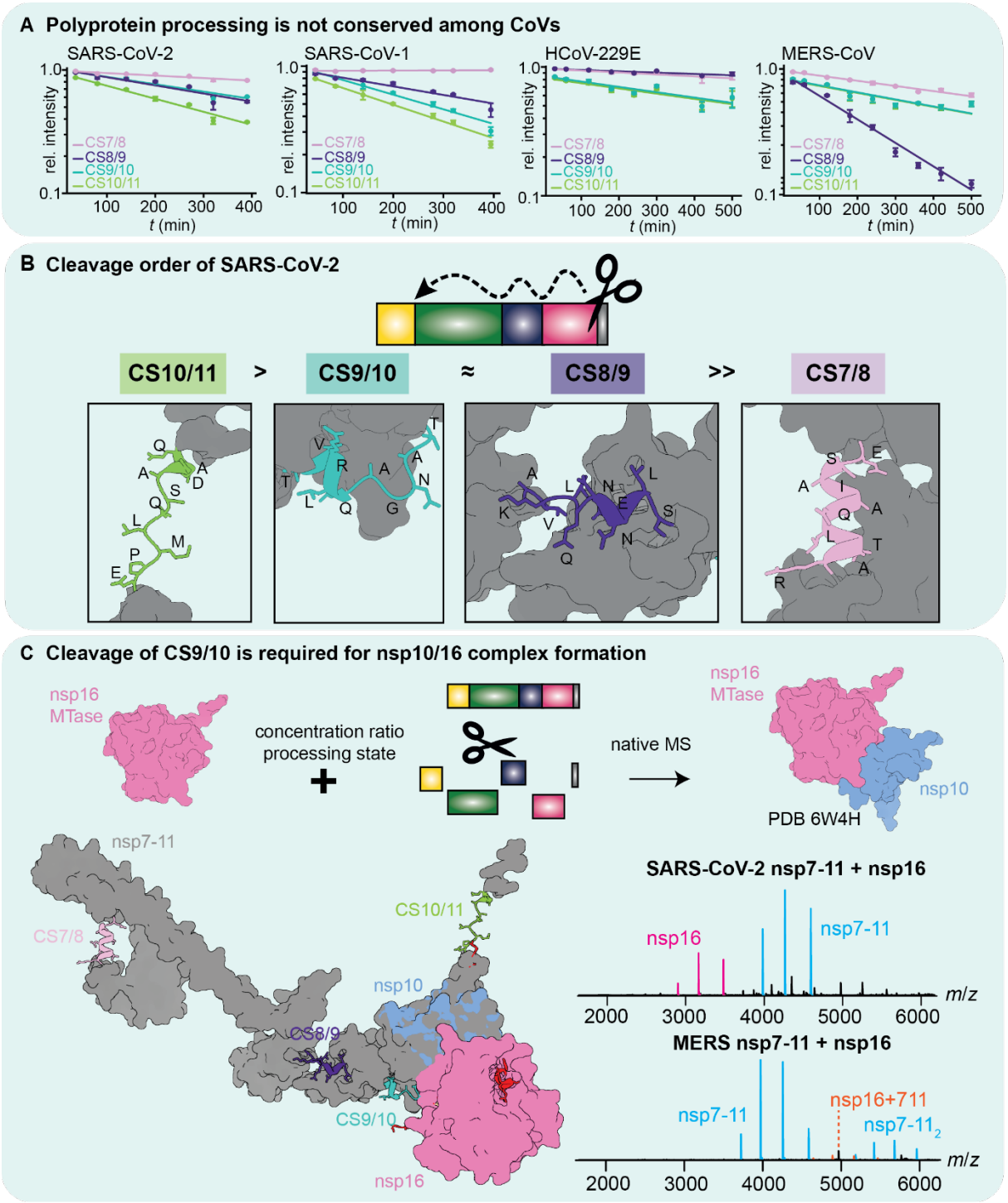


Figure 39: Graphical summary of polyprotein processing and nsp16 binding studies.
(**A**) Comparative analysis of the four human CoVs: SARS-CoV-2, SARS-CoV-1, HCoV-229E, and MERS-CoV revealed distinctive processing patterns, suggesting that processing is not strictly evolutionarily conserved. However, certain features, such as the delayed cleavage of CS7/8, appear to be conserved. These observations support the hypothesis that polyprotein processing plays a critical regulatory role, particularly in the temporal control of nsp maturation. (**B**) As all processing reactions were carried out using SARS-CoV-2 M^{pro}, species-specific effects cannot be excluded. Therefore, we propose a cleavage order only for SARS-CoV-2. In this context, M^{pro} appears to cleave preferentially from the *C*-terminus to the *N*-terminus, with CS9/10 and CS8/9 being processed at comparable rates. (**C**) Interaction binding studies with methyltransferase nsp16 demonstrated strong interaction upon release of nsp10. Nonetheless, binding was also observed with the nsp7-11 polyprotein, including a chimeric nsp7-11+16 complex between MERS-CoV and SARS-CoV-2.

A complementary *in vivo* technique for this work could be *N*-terminomics. Intermediates of the polyprotein processing could be monitored based on the *N*-terminal modifications and analyzed to resolve the cleavage order of nsp7-11, providing insight into all the other cleavage sites in pp1a and pp1ab. However, rapid kinetics and low abundance of intermediate species may be a challenge requiring precise timing and optimal enrichment methods. Here, further improvements in enrichment and analytical depth may be needed.

In the following paragraphs, a dream destination is described for investigating polyprotein processing using *in vivo* native TD-MS. It is a hybrid approach that combines complementary novel methods from MS. Here, important aspects include optimized sample preparation, advancements in mass analyzer technology to enhance dynamic range, and the use of enrichment and separation methods. Additionally, labeling techniques such as fast photochemical oxidation of proteins (FPOP) play a crucial role in enabling detailed structural analysis.

For *in vivo* TD-MS, the dynamic range problem is currently limiting, which means that low abundant species such as polyprotein intermediates are difficult to detect. Considering the investigation of SARS-CoV-2, a human cell line would be employed, which means a high dynamic range proteome of 10^8 to 10^{12}.[349] To address this, a suitable protocol needs to be established to pull nsps or polyprotein containing species from infected cells in a fast fashion by using antibodies. This would allow looking at species present *in vivo* with native TD-MS, and would likely bring along natural RNA substrates and co-factors (**Figure 40 A**).

Lysates would need to be inactivated as sophisticated MS systems are usually not situated in Biosafety level 3 (**Figure 40 A**)[350]. Furthermore, the electrospray ionization employed generates aerosols, which have to be avoided with infectious material. Most standard inactivation procedures would interfere with the native structure, however UV-C will less affect proteins and is suited for SARS-CoV-2 lysates[351].

In the future, considering the analysis of mammalian or human lysates this would require an experimental set-up, in which the complex mixture is separated to overcome dynamic range limitations. A presorting of infected and non-infected cells by using fluorescence-activated cell sorter (FACS) would enrich the polyprotein as it occurs only in the infected cells[352]. Separation could also be achieved by a combination of advanced separation techniques such as capillary-electrophoresis (CE) and ion mobility coupled to MS (**Figure 40 B**)[353]. CE with its low flow-rates lends itself as online chromatography system that preserves folding and interactions such as recently shown for an online nanoflow ion-exchange chromatography[353,354]. Furthermore, advances in dynamic range capacities of mass detectors would be promoting.

A desirable native TD-MS approach for the analysis of polyprotein processing or the RTC would be a set-up, in which a surface labeling could be employed on demand, so that local structural information and conformational changes can be monitored. An online FPOP system would be a good option, as it covalently labels amino acid side chains using hydroxyl radicals while maintaining the native conformation. Furthermore, surface labelling could be conveniently performed on demand by simply switching the laser on and off (**Figure 40 D**)[355]. In order to be able to fragment reliably down to peptide level an Omnitrap platform would be suitable to conduct multiple-stage fragmentation in MS^n [356,357](**Figure 40 C**). The Omnitrap platform provides a range of fragmentation techniques, allowing users to tailor their choice to the experimental design and scientific objectives, given the distinct characteristics of each method[358].

An alternative approach to deduce structure from such high complexity samples, is using native MS setups for online purification. When protein complex ions enter the gas phase, they become isolated precluding re-equilibration between states as observed in solution separation techniques. Mass or shape selected ions can then be soft-landed and subjected to cryo-EM[359] or subjected to X-ray diffraction on the fly[360,361].

With these future technologies, it will be possible to resolve transient states of the RTC and relevant PTMs in great molecular detail. Furthermore, heterogenous complexes with distinct functions can be detected and structurally investigated. Improved resolution in cryo-ET and computational modelling will help to put these findings into the right cellular context. In the future, highly complementary structural biology techniques, such as structural mass spectrometry, will continue to be essential for understanding dynamic molecular mechanisms.

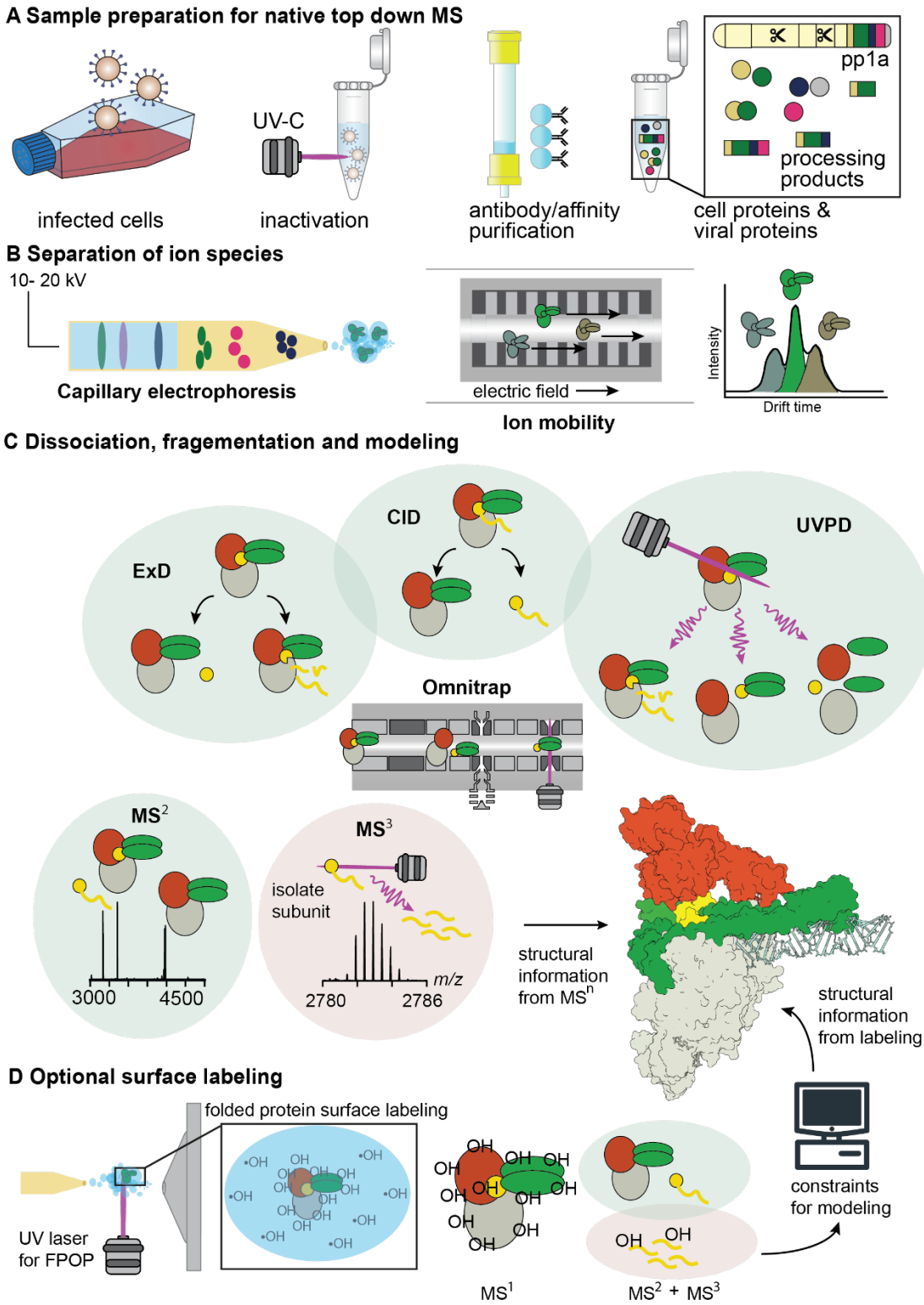


Figure 40: Graphical overview of an advanced native top-down MS (nTD-MS) platform for polyprotein processing analysis.

(**A**) Infected cell lysates are UV-C treated to inactivate viruses. Polyprotein intermediates and products are enriched via antibody or affinity purification. (**B**) Samples are introduced into an nTD-MS-optimized system with capillary electrophoresis for gentle separation prior to ionization. Ion mobility and quadrupole filtering (not shown) enable further gas-phase separation. (**C**) Fragmentation occurs in the Omnitrap system using MS^2 methods with different characteristics: ExD cleaves peptide backbones while preserving inter-subunit interfaces; CID disrupts weaker bonds, often causing subunit dissociation; UVPD generates extensive backbone fragmentation. An exemplary MS^3 could be CID followed by UVPD on an isolated fragment. (**D**) An optional UV laser enables hydroxyl (OH) -radical labeling during electrospray, highlighting solvent-accessible regions. Combined, the platform could deliver deep sequence coverage and structural insights for computational modeling (structure shown from PDB: 7RE2[362]).

5. Methods

5.1. Protein production

5.1.1. Protein constructs

Gene sequences for nsp7-11C and nsp7-11N were taken from “Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1” as published in January 2020 (replaced by NCBI LOCUS NC_045512) and commercially synthesized (GenScript). The synthetic gene sequence for nsp7-11C/nsp7-11N with suitable overhangs were cloned with Type IIS restriction enzymes into either pASK35+ and pASK33+ (IBA life sciences), generating a plasmid with *C*- and *N*-terminal His_6-tag, respectively.

Gene sequences for SARS-CoV-1, MERS-CoV and HCoV-229E were taken from the following NCBI LOCI: R1A_SARS, NC_038294.1 and R1A_MERS1. Genes were commercially synthesized and sub-cloned via restriction enzymes NcoI/XhoI into the vector pET-28a (+). All constructs contain a His_6-Strep2-SUMO-tag and are called SUMO- His_6-tagged for simplicity. Sequences of all constructs used are provided (**Table 5**).

The plasmid for nsp16 was synthesized as full-length nsp16 with *N*-terminal His_6-tag in pET22b (+) vector (**Table 5**). The His_6-tag is followed by a short linker SAVLQ enabling cleavage of viral protease M^{pro}.

The plasmid for SARS-CoV-2 M^{pro} in PGEX-6p-1 was generously provided by Prof. Rolf Hilgenfeld.

Table 5: Amino acid sequences of protein constructs and their theoretical mass (in Da).
The constructs that have a cleavable tag are shown with their authentic sequences. All of them are expressed with an *N*-terminal Sumo-Strep2-His6-tag that is listed once for all.

Protein	Sequence	Theoretical Mass (in Da)
nsp7-11C	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTE AFEKMVSLLSVLLSMQGAVDINKLCEEMLDNRATLQAIASEFSSLPSY AAFATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEFDRDAAMQRKL EKMADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALN NIINNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNTCDGTTFTYASALW EIQQVVDADSKIVQLSEISMDNSPNLAWPLIVTALRANSAVKLQNNELS PVALRQMSCAAGTTQTACTDDNALAYYNTTKGGRFVLALLSDLQDLK WARFPKSDGTGTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRG MVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYKDYLASG GQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHI DHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWKGYG CSCDQLREPMLQSADAQSFLNGFAVSARGSHHHHHH	60,824
nsp7-11N	ASRGSHHHHHHGASKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCV QLHNDILLAKDTTEAFEKMVSLLSVLLSMQGAVDINKLCEEMLDNRAT LQAIASEFSSLPSYAAFATAQEAYEQAVANGDSEVVLKKLKKSLNVAK SEFDRDAAMQRKLEKMADQAMTQMYKQARSEDKRAKVTSAMQTML FTMLRKLDNDALNNIINNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNT CDGTTFTYASALWEIQQVVDADSKIVQLSEISMDNSPNLAWPLIVTAL RANSAVKLQNNELSPVALRQMSCAAGTTQTACTDDNALAYYNTTKG GRFVLALLSDLQDLKWARFPKSDGTGTIYTELEPPCRFVTDTPKGPKV KYLYFIKGLNNLNRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFA	60,953

	VDAAKAYKDYLASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQES FGGASCCLYCRCHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLKN TVCTVCGMWKGYGCSCDQLREPMLQSADAQSFLNGFAV	
Sumo-Strep2-His6-tag	MGSSHHHHHHSSGWSHPQFEKGGMSDSEVNQEAKPEVKPEVKPET HINLKVSDGSSEIFFKIKKTTPLRRLMEAFAKRQGKEMDSLRFLYDGIRI QADQTPEDLDMEDNDIIEAHREQIGG	13,832
nsp7-11 SARS-CoV-2	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTE AFE KMVSLLSVLLSMQGAVDINKLCEEMLDNRATLQAIASEFSSLPSYAAF ATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEFDRDAAMQRKLEK MADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALNNII NNARDGCVPLNIIPLTTAAKLMVVIPDYNTYKNTCDGTTFTYASALWEI QQVVDADSKIVQLSEISMDNSPNLAWPLIVTALRANSAVKLQNNELSP VALRQMSCAAGTTQTACTDDNALAYYNTTKGGRFVLALLSDLQDLKW ARFPKSDGTGTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNLNRGM VLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDAAKAYKDYLASGG QPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCRCHID HPNPKGFCDLKGKYVQIPTTCANDPVGFTLKNTVCTVCGMWKGYGC SCDQLREPMLQSADAQSFLNGFAV	59,543
nsp7-11 SARS-CoV-1	SKMSDVKCTSVVLLSVLQQLRVESSSKLWAQCVQLHNDILLAKDTTE AFEKMVSLLSVLLSMQGAVDINRLCEEMLDNRATLQAIASEFSSLPSY AAYATAQEAYEQAVANGDSEVVLKKLKKSLNVAKSEFDRDAAMQRKL EKMADQAMTQMYKQARSEDKRAKVTSAMQTMLFTMLRKLDNDALN NIINNARDGCVPLNIIPLTTAAKLMVVVPDYGTYKNTCDGNTFTYASAL WEIQQVVDADSKIVQLSEINMDNSPNLAWPLIVTALRANSAVKLQNNE LSPVALRQMSCAAGTTQTACTDDNALAYYNNSKGGRFVLALLSDHQ DLKWARFPKSDGTGTIYTELEPPCRFVTDTPKGPKVKYLYFIKGLNNL NRGMVLGSLAATVRLQAGNATEVPANSTVLSFCAFAVDPAKAYKDYL ASGGQPITNCVKMLCTHTGTGQAITVTPEANMDQESFGGASCCLYCR CHIDHPNPKGFCDLKGKYVQIPTTCANDPVGFTLRNTVCTVCGMWKG YGCSCDQLREPLMQSADASTFLNGFAV	59,606
nsp7-11 MERS-CoV	SKLTDLKCTSVVLLSVLQQLHLEANSRAWAFCVKCHNDILAATDPSEA FEKFVSLFATLMTFSGNVDLDALASDIFDTPSVLQATLSEFSHLATFAE LEAAQKAYQEAMDSGDTSPQVLKALQKAVNIAKNAYEKDKAVARKLE RMADQAMTSMYKQARAEDKKAKIVSAMQTMLFGMIKKLDNDVLNGII SNARNGCIPLSVIPLCASNKLRVVIPDFTVWNQVVTYPSLNYAGALWD ITVINNVDNEIVKSSDVVDSNENLTWPLVLECTRASTSAVKLQNNEIKP SGLKTMVVSAGQEQTNCNTSSLAYYEPVQGRKMLMALLSDNAYLKW ARVEGKDGFVSVELQPPCKFLIAGPKGPEIRYLYFVKNLNNLHRGQVL GHIAATVRLQAGSNTEFASNSSVLSLVNFTVDPQKAYLDFVNAGGAPL TNCVKMLTPKTGTGIAISVKPESTADQETYGGASVCLYCRAHIEHPDV SGVCKYKGKFVQIPAQCVRDPVGFCLSNTPCNVCQYWIGYGCNCDS LRQAALPQSKDSNFLNESGVLL	59,527
nsp7-11 HCoV-229E	SKLTDLKCTNVVLMGILSNMNIASNSKEWAYCVEMHNKINLCDDPETA QELLLALLAFFLSKHSDFGLGDLVDSYFENDSILQSVASSFVGMPSFV AYETARQEYENAVANGSSPQIIKQLKKAMNVAKAEFDRESSVQKKINR MAEQAAAAMYKEARAVNRKSKVVSAMHSLLFGMLRRLDMSSVDTILN MARNGVVPLSVIPATSAARLVVVVPDHDSFVKMMVDGFVHYAGVVW TLQEVKDNDGKNVHLKDVTKENQEILVWPLILTCERVVKLQNNEIMPG KMKVKATKGEGDGGITSEGNALYNNEGGRAFMYAYVTTKPGMKYVK WEHDSGVVTVELEPPCRFVIDTPTGPQIKYLYFVKNLNNLRRGAVLGY IGATVRLQAGKQTEFVSNSHLLTHCSFAVDPAAAYLDAVKQGAKPVG NCVKMLTNGSGSGQAITCTIDSNTTQDTYGGASVCIYCRAHVAHPTM	59,134

	DGFCQYKGKWVQVPIGTNDPIRFCLENTVCKVCGCWLNHGCTCDRT AIQSFDNSYLNESGALVPLD	
nsp16	MHHHHHHSAVLQSSQAWQPGVAMPNLYKMQRMLLEKCDLQNYGDS ATLPKGIMMNVAKYTQLCQYLNTLTLAVPYNMRVIHFGAGSDKGVAP GTAVLRQWLPTGTLLVDSDLNDFVSDADSTLIGDCATVHTANKWDLII SDMYDPKTKNVTKENDSKEGFFTYICGFIQQKLALGGSVAIKITEHSW NADLYKLMGHFAWWTAFVTNVNASSSEAFLIGCNYLGKPREQIDGYV MHANYIFWRNTNPIQLSSYSLFDMSKFPLKLRGTAVMSLKEGQINDMI LSLLSKGRLIIRENNRVVISSDVLVNN	34,776
M^{pro} SARS-CoV-2	SGFRKMAFPSGKVEGCMVQVTCGTTTLNGLWLDDVVYCPRHVICTS EDMLNPNYEDLLIRKSNHNFLVQAGNVQLRVIGHSMQNCVLKLKVDT ANPKTPKYKFVRIQPGQTFSVLACYNGSPSGVYQCAMRPNFTIKGSF LNGSCGSVGFNIDYDCVSFCYMHHMELPTGVHAGTDLEGNFYGPFV DRQTAQAAGTDTTITVNVLAWLYAAVINGDRWFLNRFTTTLNDFNLVA MKYNYEPLTQDHVDILGPLSAQTGIAVLDMCASLKELLQNGMNGRTIL GSALLEDEFTPFDVVRQCSGVTFQ	33,669

5.1.2. Expression and purification

For expression and purification of nsp7-11 (either nsp7-11C, nsp7-11N or SUMO-His_6-tagged constructs), transformed BL21 Rosetta2 (Merck Millipore) were grown to OD_{600} 0.4-0.6 in 2xYT medium and then induced either with 50 µM anhydrotetracycline or with 500 µM isopropyl ß-D-1-thiogalactopyranoside (IPTG) for 16 h at 20°C. Cells were harvested using at 4,000 × g for 10 min at 4°C. Subsequently, cell pellets were washed in 40 mM phosphate buffer, 300 mM NaCl (pH 8,0) and stored at -20°C.

Lysis was performed in the same way for all protein constructs. Lysis buffer (40 mM phosphate, 300 mM NaCl, 15 mM imidazole (pH 8,0), 0.2 mg/ml lysozyme and cOmplete™, EDTA-free protease inhibitor) was added to the thawing pellet and sonication was performed for 2-3 min using the Micro tip, 70% power, on 1 s, off 5 s (*Branson digital sonifier SFX 15)*. Subsequently, the lysate was centrifuged at 20,000 × g for 40 min at 4°C. After centrifugation the supernatant (crude extract (CE) sample for SDS-PAGE) was collected and the pellet was discarded. Nsp7-11N and nsp7-11C and SUMO-His_6-tagged nsp7-11 proteins were purified via Ni-NTA affinity chromatography (PureCube Ni-NTA Agarose, *Cube Biotech*) and Superdex10/300 (*Cytiva*) size exclusion chromatography (SEC)[258]. Whereby, the SUMO-His_6-tagged constructs were incubated with an in-house made SUMO-protease (0.1 mg protease per 1 mg target protein) and dialyzed overnight followed by SEC.

For nsp16 expression, the plasmid was transformed in BL21 Rosetta2. Cells were grown until an OD_{600} of 0.4-0.6, cooled on ice and induced with 0.5 mM IPTG and then incubated overnight at 20°C. Nsp16 was purified via Ni-NTA affinity chromatography (PureCube Ni-NTA Agarose, *Cube Biotech*) and Superdex10/300 (*Cytiva*) size exclusion chromatography.

M^{pro} was transformed in BL21 and expression culture was induced with 0.5 mM IPTG between an OD_{600} of 0.4-0.6. Expression culture was incubated overnight and harvested as described above. M^{pro} was purified using Ni-NTA (HisPur™ Ni-NTA Resin) affinity chromatography and

Superdex10/300 (*Cytiva*) size exclusion chromatography. To cleave the His_6-tag from M^{pro}, it was transferred into Slide-A-Lyzer dialysis cassette (*Thermo Fisher*) with MWCO 10,000 and digested overnight in PreScission protease cleavage buffer (50 mM Tris HCl, 100 mM NaCl, 1 mM EDTA, 1 mM DTT, pH 8.0). PreScission protease was pulled out using GST sepharose beads to obtain pure M^{pro}. Nsp16 and M^{pro} were flash-frozen and stored at -80°C.

5.2. Preparations for native mass spectrometry

5.2.1. Buffer exchange

Freshly purified samples were exchanged into a structure preserving MS compatible buffer surrogate ammonium acetate at 300 mM AmAc (99.99% purity, Sigma-Aldrich), 1 mM DTT, pH 8.0. To adjust the pH without introducing salts, Ammonium hydroxide (extra pure, 25% solution in water; *Thermo Fisher Scientific*) and acetic acid (100%, extra pure, *Roth*) were used.

M^{pro} was exchanged into the buffer surrogate by applying two cycles of centrifugal gel filtration (Biospin mini columns, 6000 MWCO, Bio-Rad). nsp16 and nsp7-11 were exchanged by six rounds of dilution and concentration in centrifugal filter units (Amicon Ultra, 10 K MWCO, Merck Millipore). Below is a detailed description of the buffer exchange device protocols.

Micro Bio-Spin 6 Columns (*Bio-rad)*

The column was prepared by resuspending the gel matrix and removing the tip. The packaging buffer was then removed by gravity drainage followed by centrifugation at 1,000 × g for 2 min at 4 °C. Subsequently, 500 µl of MS buffer was added, briefly drained by gravity, and centrifuged for 1 min at 1,000 × g at 4 °C. This step was repeated three additional times. Depending on the sample concentration, 20-75 µl of the sample was loaded onto the column. Elution was performed by centrifugation at 1,000 × g for 4 min at 4 °C.

The eluted sample was applied to a new column that was prepared as described above to ensure quantitative desalting.

Amicon® Ultra Centrifugal Filter (*Merck Millipore Solutions*)

The filter membrane was equilibrated with 400 µl of MS buffer by centrifugation at 13,000 × g for approximately 7 min. A 50 µl sample was then applied, followed by the addition of 350 µl of MS buffer. The sample was concentrated to 50 µl by centrifugation at 13,000 × g for 7-12 min at 4 °C. An additional 350 µl of MS buffer was added and the process was repeated. This step was performed four to five times to ensure effective desalting. Samples were recovered by inverting the filter tube and centrifuging for 1 min at 1,000 × g into a new collection tube.

5.2.2. Production of home-made nano-ESI capillaries

Gold-coated capillaries were employed to introduce the prepared sample into the mass spectrometer during native MS experiments. A micropipette puller (P-1000, Sutter Instruments) was used to produce nano-ESI capillaries in a two-step program from borosilicate capillaries (1.2 mm and 0.68 mm outer and inner diameter, respectively, World Precision Instruments) using a squared-box filament (2.5 mm × 2.5 mm). Capillaries were gold-coated by using a

sputter coater (CCU-010, Safematic, 5.0×10^{-2} mbar, 30.0 mA, 120 s, three runs to vacuum limit 3.0×10^{-2} mbar argon). Capillaries were opened using tweezers under a microscope.

5.3. Native mass spectrometry

For the processing experiments, native MS was performed on a Q Exactive UHMR Orbitrap from *Thermo Scientific*. Positive ion mode was used by applying capillary voltages of 1.2-1.7 kV, 100-150°C capillary temperature, 15 eV in-source disociation and 25 eV in HCD cell. Trapping gas pressure optimization was set to 5 or 7.

For the time-resolved polyprotein processing measurements the ion transfer *m/z* optimization were adapted as follows: Inj. Fl. RF Ampl. to 300, Bent. Fl. RF Ampl. to 940, Trans. MP and HCD-cell RF Ampl. to 900 and C-Trap Ampl. to 2750. Tandem MS was always conducted by stepwise increase of the HCD voltage of 10-20 V. Detector optimization was set to "*low m/z*".

For the nsp16 interaction studies standard settings of ion transfer *m*/*z* optimization were used either in "*low m*/*z"* or "*high m*/*z"* mode.

Setting the trapping gas pressure to a specific value does not define an exact pressure. Pressure reading can vary from instrument to instrument. Therefore, pressure readings of high vacuum (HV) and ultra-high vacuum (UHV) are indicated below.

5.3.1. Continuous polyprotein processing

For the continuous approach, M^{pro} was added to a final concentration of approximately 3 µM to nsp7-11/nsp7-11C/nsp7-11N (final concentration 18 µM), then the sample was briefly mixed by pipetting before transferring 1-2 µL to the capillary. Data acquisition was started 1 min after mixing. At least three replicates were conducted. The temperature of the capillary housing was 27°C and the temperature of the interior heated capillary was 150°C. Representative pressure readings from the conducted experiments are shown in the **Table 6**.

Table 6: Pressure readings during continuous polyprotein processing experiments.
Pressures of HV and UHV are listed.

Construct	HV (mbar)	UHV (mbar)
nsp7-11N	3.01×10^{-9}	2.06×10^{-10}
nsp7-11C	3.01×10^{-9}	1.99×10^{-10}
nsp7-11	3.31×10^{-9}	2.25×10^{-10}

5.3.2. Discontinuous polyprotein processing

For the discontinuous approach, the final concentrations of nsp7-11/nsp7-11C and M^{pro} are given in **Table 7**. The mixture was incubated on ice and triplicate measurements were taken at selected time points.

Table 7: Final concentrations of discontinuous processing experiments.

construct	strain	nsp7-11 concentration	M^{pro} concentration
nsp7-11	SARS-CoV-1	19 µM	3.5 µM
nsp7-11	SARS-CoV-2	19 µM	3.5 µM
nsp7-11C	SARS-CoV-2	20 µM	10 µM
nsp7-11	MERS-CoV	19 µM	3.1 µM
nsp7-11	HCoV-229E	17 µM	3.1 µM

The temperature of the capillary housing was 27 °C and the temperature of the interior heated capillary was set to 100 °C for polyprotein processing of untagged nsp7-11 and set to 150 °C for polyprotein processing of nsp7-11C. Instrument pressure readings are indicated in the following **Table 8**.

Table 8: Pressure readings during discontinuous polyprotein processing experiments.
Pressures of high vacuum HV and UHV are listed.

Construct	strain	HV (mbar)	UHV (mbar)
nsp7-11	SARS-CoV-1	3.30×10^{-9}	2.48×10^{-10}
nsp7-11C	SARS-CoV-2	3.01×10^{-9}	2.04×10^{-10}
nsp7-11	SARS-CoV-2	3.32×10^{-9}	2.52×10^{-10}
nsp7-11	MERS-CoV	3.27×10^{-9}	2.35×10^{-10}
nsp7-11	HCoV-229E	3.27×10^{-9}	2.33×10^{-10}

5.3.3. Binding studies of nsp7-11 and nsp16

Interaction studies of nsp7-11 of SARS-CoV-2 and MERS-CoV with authentic termini and nsp16 with cleaved and uncleaved His_6-tag were conducted on the Q Exactive UHMR Orbitrap. Here, fresh and frozen protein were used.

For SARS-CoV-2, interaction studies were conducted with the following final concentrations: nsp7-11 with three different concentrations, 5 µM, 15 µM, 25 µM were mixed with 5 µM nsp16-His_6 and 3 µM M^{pro} and incubated overnight. Due to the M^{pro} cleavage site nsp16 was obtained with natural termini. For the K_D determination, the ion transfer *m*/*z* optimization was set to "*low m*/*z*", capillary temperature of the mass spectrometer was set to 100 °C, in-source dissociation was set to 10 eV and HCD cell was set to 15 eV, providing optimized conditions for accurately representing molar proportions. Pressure readings for HV and UHV are shown in **Table 9**.

Untagged nsp7-11 (15 µM) was mixed with 5 µM nsp16-His_6. To exclude any artefacts caused by the His_6-tag, nsp16-His_6 was also incubated with a low concentration of M^{pro} (1:10). His_6-tag-free nsp16 was mixed with nsp7-11 at the same final concentration ratios (5 µM and

15 µM, respectively) prior to the measurements. Furthermore, the ion transfer *m*/*z* optimization was set to "*high m/z*" to investigate the potential binding between unprocessed nsp7-11 and nsp16, which would correspond to a complex exceeding 90 kDa.

To study binding interactions in MERS-CoV, nsp7-11 was mixed to 5 µM nsp16-His_6 in two different final concentrations, 5 µM and 15 µM. Nsp16-His_6 was incubated with M^{pro} at a ration of 10:1 to cleave the His_6-tag. The interaction between nsp7-11 (15 µM) and nsp16 (5 µM) was then tested. Interior capillary temperature of the mass spectrometer was set to 100 °C, in-source dissociation was set to 35 eV and HCD cell was set to 15 eV. Tandem MS was performed by incrementally increase the HCD voltage by 12.5 eV or 25 eV. Pressure readings for HV and UHV are shown in **Table 9**.

Table 9: Representative HV and UHV pressure values of the interaction binding studies.

experiment	strain	HV (mbar)	UHV (mbar)
K_D determination	SARS-CoV-2	3.32×10^{-9}	3.03×10^{-10}
Testing nsp7-11 and nsp16	SARS-CoV-2	3.30×10^{-9}	3.03×10^{-10}
Testing nsp7-11 and nsp16	MERS-CoV	3.27×10^{-9}	2.33×10^{-10}

5.4. Data Analysis

5.4.1. Spectra analysis

Peak assignment and mass analysis was performed using UniDec 7.0.2[210], MetaUniDec 7.0.2[269] and mMass 5.5.0[363]. The mMass software was used to extract the *m*/*z* values of the peak envelope and to calculate the average measured mass and the associated standard error for each protein species.

MetaUniDec was used with the standard settings unless otherwise indicated. As acquired data were sharp with little adducts no additional data processing parameters were used. UniDec Run parameters were used in standard mode, whereby some nearby points were smoothed and none artifacts were suppressed. Peak selection threshold was set to 0.001 and normalization was set to “Sum”, which normalizes as such that all peaks of a timepoint are 100%. The plots generated by MetaUniDec were saved as SVG files and edited in Adobe Illustrator.

5.4.2. SDS-PAGE

SDS-Page was performed with a 4%-12% gradient acrylamide Bis-tris gel with XT MES running buffer (Bio-Rad Laboratories). Both constructs nsp7-11C and nsp7-11N were mixed at 36 µM with 14 µM M^{pro} and incubated at 4°C. Aliquots were withdrawn at indicated time points and mixed with XT sample buffers to quench the reaction. Polyprotein nsp7-11 (54 µM) of the four CoVs SARS-CoV-2, SARS-CoV-1, MERS-CoV and HCoV-229E were mixed with 9 µM M^{pro} and incubated overnight at 4°C. SUMO-His_6-tagged nsp7-11, tag-cleaved nsp7-11 with authentic termini and processed nsp7-11 of the four CoVs were run on a 4%-12% gradient acrylamide Bis-tris gel with XT MES running buffer (Bio-Rad Laboratories).

To analyze protein purification using SDS-PAGE, samples were taken and mixed with either XT sample buffer or a custom gel-loading buffer. (250 mM Tris-HCl (pH 6,8), 10% SDS, 30% glycerol, 10 mM 2-mercaptoethanol, 0.02% bromophenol blue).

5.5. Bioinformatic tools

5.5.1. AlphaFold

Polyprotein sequences of nsp7-11C and nsp7-11N and sequences of all intermediates were submitted to AF2 standard run (20 cycles). All models were examined with UCSF ChimeraX and best models were picked for comparison[364]. Here, the best model for nsp7-11C was selected. pLDDT scores were displayed by using the B-factor column of the PDB-output file. Set color key thresholds in ChimeraX were 50 to 90.

Polyprotein sequences of nsp7-11 of the four CoVs were run with standard settings from AF3 server[251]. The best models were selected according to the overall and local confidence scores pTM and pLDDT. For the model selection, the confidence of the cleavage site areas was particularly decisive for the model selection. Thus, plotted pLDDT scores against the residues index were utilized. Furthermore, pLDDT scores were displayed by using the B-factor column of the output files. Regions with pLDDT scores higher than 70 are expected to be well predicted.

5.5.2. Multiple Sequence Alignment

Polyprotein sequences of nsp7-11 region from the four CoVs (**Table 5**) were submitted to Clustal Omega for alignment[365]. The job was submitted in default mode and loaded into Jalview for visualization[366]. Amino acids were colored on a gradient from white to blue: gaps are white, residues matching the consensus sequence were colored dark blue, and non-matching residues with a positive BLOSUM62 substitution score were colored light blue[367,368].

5.5.3. ConSurf Server

To obtain reliable estimates of evolutionary conservation, the SARS-CoV-2 nsp7-11 sequence was submitted to the server[369–372] using the ConSeq method[373]. Homologous sequences were identified in the UniRef90 database using the HMMER algorithm with an E-value cutoff of 0.0001. Redundant sequences were filtered using a CD-HIT cutoff of 95% maximum pairwise identity. A total of 150 homologs were selected for MSA using MAFFT. Conservation scores were calculated using the Bayesian method, and mapped onto the nsp7-11 region of SARS-CoV-2. Scores range from 1 (variable) to 9 (highly conserved).

5.6. Data analysis with Python

5.6.1. Data analysis of time-resolved polyprotein processing

Native mass spectra were investigated, and deconvolution was supported by UniDec[374] using version 7.0.2. Deconvoluted peaks and peak intervals were checked and *m*/*z* ranges were noted to feed into a home-made Python script. The analysis was performed using Python version 3.12.1, with the following package versions: NumPy (2.2.2), pandas (2.2.3), Matplotlib (3.10.0), and LMFIT (1.3.2). Every peak interval was checked before the AUC of the detected mass species was taken and assigned, which is here called intensity. The initial substrate includes five domains, resulting in five mature proteins. Thus, species intensities were normalized by using a multiplication factor corresponding to the domains or units depending on intermediate species or mature nsps. The multiplication array was adapted depending on the species that were detected. **Equation 13**-**Equation 18** show the multiplication arrays for all constructs.

Equation 13: Multiplication array for SARS-CoV-2 and SARS-CoV-1 nsp7-11 with authentic termini

$$m = 5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 3 \cdot I_{nsp9-11} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11}$$

Equation 14: Multiplication array for MERS-CoV nsp7-11 with authentic termini

$$m = 5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 3 \cdot I_{nsp9-11} + 2 \cdot I_{nsp10-11}$$

Equation 15: Multiplication array for HCoV-229E nsp7-11 with authentic termini.

$$m = 5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 2 \cdot I_{nsp8-9} + 4 \cdot I_{nsp8-9_2} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11} + 4 \cdot I_{nsp8-11}$$

In polyprotein processing of the His_6-tagged constructs nsp7-11N and nsp7-11C, nsp11 could not be detected. For nsp7-11C and nsp7-11N, this may be due to slightly higher HV and UHV pressures (cf. **Table 6**). Therefore, nsp11 domain was not considered in the multiplication arrays.

Equation 16: Multiplication array for SARS-CoV-2 nsp7-11N in the continuous polyprotein processing experiments.

$$m = 4 \cdot I_{nsp7-11N} + 4 \cdot I_{nsp7-10} + 2 \cdot I_{nsp7-8N} + 2 \cdot I_{nsp7-8N} + 2 \cdot I_{nsp9-11C} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp8_2} + 2 \cdot I_{nsp7_2} + I_{nsp7N} + I_{nsp8} + I_{nsp9} + I_{nsp10}$$

Equation 17: Multiplication array for SARS-CoV-2 nsp7-11C in the continuous polyprotein processing experiments.

$$m = 4 \cdot I_{nsp7-11C} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 2 \cdot I_{nsp9-11C} + 2 \cdot I_{nsp9-10} + I_{nsp10-11C} + 2 \cdot I_{nsp8_2} + 2 \cdot I_{nsp7_2} + I_{nsp7} + I_{nsp8} + I_{nsp9} + I_{nsp10}$$

Equation 18: Multiplication array for SARS-CoV-2 nsp7-11 with authentic termini in the continuous polyprotein processing experiments.

$$m = 5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 3 \cdot I_{nsp9-11} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11C} + 2 \cdot I_{nsp8_2} + 2 \cdot I_{nsp7_2} + I_{nsp7} + I_{nsp8} + I_{nsp9} + I_{nsp10} + I_{nsp11}$$

To correct for spray variation, the ratio of each individual species to the sum of all species was taken. As example, the ratio of untagged nsp7-11 from SARS-CoV-2 including the multiplication:

Equation 19

$$ratio_{nsp7-11} = \frac{5 \cdot I_{nsp7-11}}{5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 * I_{nsp7-8} + 3 \cdot I_{nsp9-11} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11}}$$

For the continuous polyprotein processing approach, the calculation of this ratio was as follows:

Equation 20

$$ratio_{nsp7-11} = \frac{5 \cdot I_{nsp7-11}}{5 \cdot I_{nsp7-11} + 4 \cdot I_{nsp7-10} + 3 \cdot I_{nsp7-9} + 2 \cdot I_{nsp7-8} + 2 \cdot I_{nsp9-11} + 2 \cdot I_{nsp9-10} + 2 \cdot I_{nsp10-11} + 2 \cdot I_{nsp8_2} + 2 \cdot I_{nsp7_2} + I_{nsp7} + I_{nsp8} + I_{nsp9} + I_{nsp10}}$$

Then the normalized intensities of the replicates were averaged. The standard error of the mean was also calculated. The fitted rates for the cleavage sites were calculated using the normalized intensities, with the species containing the intact cleavage site summed to the corresponding cleavage site. For SARS-CoV-2, SARS-CoV-1 and MERS-CoV this is done as described in **Equation 21**-**Equation 25**:

Equation 21: Species included into CS10/11 calculations for SARS-CoV-2, SARS-CoV-1 and MERS-CoV.

$$I_{10/11} = I_{nsp7-11} + I_{nsp9-11} + I_{nsp10-11}$$

Equation 22: Species included into CS9/10 calculations for SARS-CoV-2, SARS-CoV-1 and MERS-CoV.

$$I_{9/10} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp9-11} + I_{nsp9-10}$$

Equation 23: Species included into CS9/10 calculations for MERS-CoV.

$$I_{9/10} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp9-11}$$

Equation 24: Species included into CS8/9 calculations for SARS-CoV-2, SARS-CoV-1 and MERS-CoV.

$$I_{8/9} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp7-9}$$

Equation 25: Species included into CS7/8 calculations for SARS-CoV-2, SARS-CoV-1 and MERS-CoV.

$$I_{7/8} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp7-9} + I_{nsp7-8}$$

Since other mass species were detected in the polyprotein processing of nsp7-11 from HCoV-229E, these species were assigned to the cleavage sites accordingly:

Equation 26: Species included into CS10/11 calculations for HCoV-229E.

$$I_{10/11} = I_{nsp7-11} + I_{nsp8-11} + I_{nsp10-11}$$

Equation 27: Species included into CS9/10 calculations for HCoV-229E.

$$I_{9/10} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp8-11}$$

Equation 28: Species included into CS8/9 calculations for HCoV-229E.

$$I_{8/9} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp7-9} + I_{nsp8-9} + I_{nsp8-9\mathrm{d}} + I_{nsp8-11}$$

Equation 29: Species included into CS7/8 calculations for HCoV-229E.

$$I_{7/8} = I_{nsp7-11} + I_{nsp7-10} + I_{nsp7-9} + I_{nsp7-8}$$

The calculated time-dependent intensities for a given cleavage site were fitted to a first-order kinetics formula:

Equation 30: Exponential model used for fitting the data.

$$I_{fit}(t) = Ae^{-kt}$$

Rate constants *k* for each cleavage were extracted from the fit.

5.6.2. Quantification and plotting of nsp16 binding experiment data

Native MS spectra were viewed and analyzed using UniDec[374] and mMass[363]. Peak intensities were extracted and bar plots generated using basic components of the custom Python script (6.3). Peak selection and peak interval verification followed previously established protocols. Normalization was performed without applying the multiplication array. Domain correction was not required, as only unprocessed or fully processed polyprotein species were analyzed. Intensities were exported as CSV files and further processed in Microsoft Excel. The method for determining the dissociation constant (K_D) is detailed in the following section. Error bars represent the standard error calculated from triplicates.

5.6.3. K_D calculations

Affinities of protein-protein interactions were calculated based on the law of mass action. One binding pocket was considered for the calculation of the K_Ds of nsp16+10 complex, nsp10 + nsp16 ⇌ [nsp10+ nsp16]. The dissociation kinetics is described **Equation 31**.

Equation 31:

$$K_D = \frac{[nsp10] \cdot [nsp16]}{[nsp10 + nsp16]}$$

Nsp10 or nsp16 is the concentration of the protein without a ligand. Molar fractions can be calculated by using the signal intensities of the species and the known molar concentration of nsp10 ($[nsp10]_0$) and nsp16 ($[nsp16]_0$) that were introduced to the mass spectrometer. AUC was extracted for each peak, assigned to the corresponding species, and normalized to obtain relative signal intensity. Then, molar fractions were calculated using **Equation 32**-**Equation 34**.

Equation 32

$$[nsp16] = [nsp16]_0 \cdot I_{nsp16}$$

Equation 33

$$[nsp10] = [nsp10]_0 - [nsp16]_0 \cdot I_{nsp10+nsp16}$$

Equation 34

$$[nsp16 + nsp10] = [nsp16]_0 \cdot I_{nsp10+nsp16}$$

The Gaussian error propagation rule was used to determine the standard deviation for the K_D-values. K_D-values were calculated for each sample having different ratios. Since the replicate number was the same for each determined K_D, the values were simply averaged.

5.7. Data visualization

Protein structures were either retrieved from the Protein Data Bank (PBD)[255] at https://www.rcsb.org/ or modeled using AF2 or 3[250,251]. Protein structures were viewed and adjusted using ChimeraX 1.9 (*Resource for Biocomputing, Visualization, and Informatics at UC San Francisco*)[375,376].

MSA was viewed and visualized with Jalview[366].

Mass spectra were copied from the spectrum window of Thermo Xcalibur Qual Browser. Mass spectra extracts were saved as svg. or pdf. Using Unidec[210] or an in-house made software *MSänger*[377].

All figures were prepared or modified using Adobe Illustrator 2025.

5.8. Data Availability

The MS data generated in this study have been deposited in the PRIDE database: http://proteomecentral.proteomexchange.org/cgi/GetDataset?ID=PXD049009. PRIDE is a partner repository for the ProteomeXchange Consortium[378,379]. The corresponding extracted scans are also available in Zenodo under https://doi.org/10.5281/zenodo.15488266[380], and for an earlier version under https://doi.org/10.1101/2024.01.06.574466[84].

5.9. Code Availability

Customized python scripts were coded to analyze the data as described in 5.6. Python code is available in Zenodo: https://doi.org/10.5281/zenodo.15488266[380], and https://doi.org/10.1101/2024.01.06.574466[84].

6. Supplementary Material

6.1. Supplementary figures

P5	P4	P3	P2	P1	P1‘	P2‘	P3‘	P4‘	P5‘
R	A	T	L	Q |	A	I	A	S	E

Figure S1: Schechter and Berger nomenclature.
The nomenclature is illustrated using the sequence of cleavage site between nsp7 and nsp8.

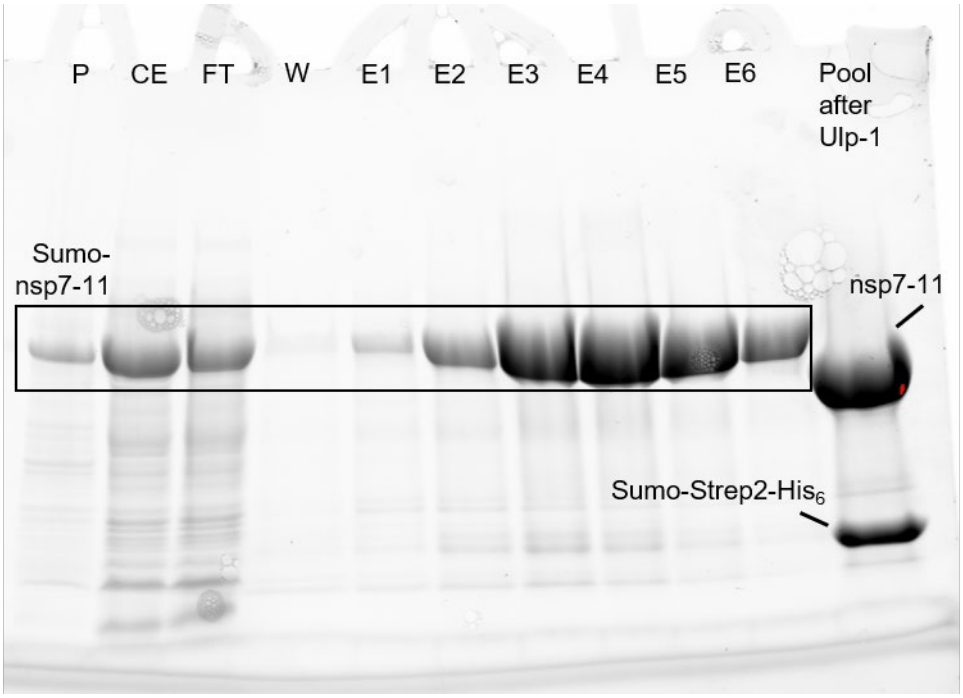


Figure S2: SDS-PAGE depicting steps in protein purification of nsp7-11 SARS-CoV-2.
Indicated letters show P = pellet, CE = crude extract or supernatant, FT = flow-through after binding to Ni-agarose resin, W = washing step with 50 mM imidazole (after third column volume wash step), E1-E6= elution one to six with 300 mM imidazole and 10 min incubation at 4°C. Last lane, shows pooled fraction after overnight digestion with Sumo-protease Ulp-1. SDS-PAGE shows the expressed Sumo-Strep-2-His_6-tag nsp7-11 construct of SARS-CoV-2 (~ 73.4 kDa) around 75 kDa. After overnight digestion the Sumo-Strep-2-His_6-tag is cleaved off, reducing the mass by approximately the tag ~14 kDa.

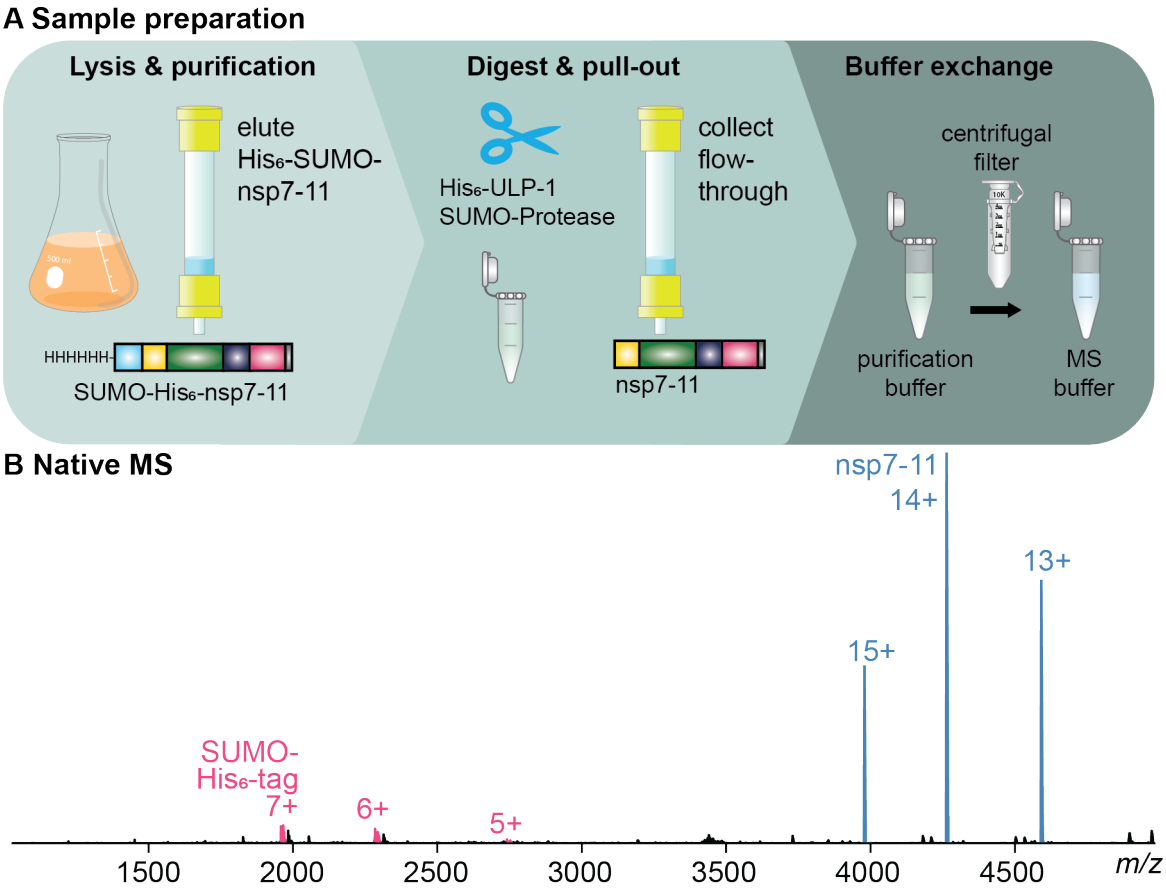


Figure S3: Sample preparation of SARS-CoV-2 nsp7-11 with pulling out ULP-1.
(**A**) A schematic overview of the sample preparation workflow is shown. SUMO-His$_6$-tagged nsp7-11 is first lysed and purified via affinity chromatography. Subsequently, ULP-1 protease is added to cleave the SUMO-His$_6$-tag, yielding nsp7-11 with authentic N- and C-termini. ULP-1 is then removed by a second round of affinity chromatography, during which nsp7-11 is collected in the flow-through and subsequently subjected to buffer exchange into a MS-compatible buffer surrogate. (**B**) Representative spectrum shows nsp7-11 with authentic termini produced using SUMO-protease ULP-1, which is completely pulled out. However, SUMO-His$_6$-tag is clearly visible and is not successfully pulled out.

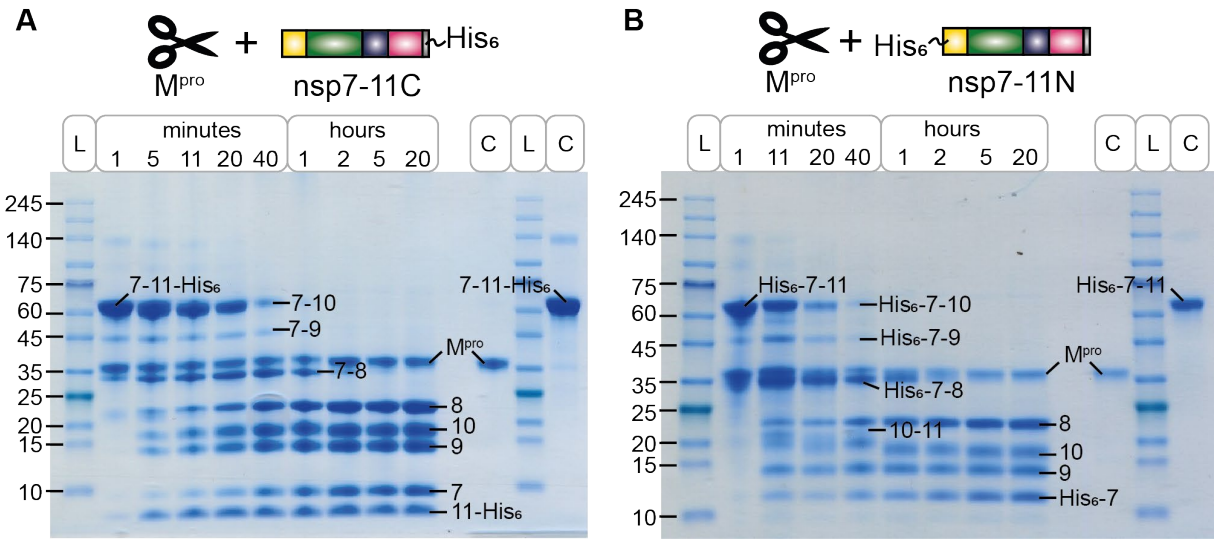


Figure S4: Monitoring polyprotein processing of SARS-CoV-2 in purification buffer.
SDS-PAGE of nsp7-11C (**A**) processing and nsp7-11N (**B**) processing showing protein marker ladder (L) and controls (C) of M^{pro} and nsp7-11C or nsp7-11N. Both constructs nsp7-11C and nsp7-11N were mixed at 36 µM with 14 µM M^{pro} and incubated at 4°C. Aliquots were withdrawn at indicated time points.

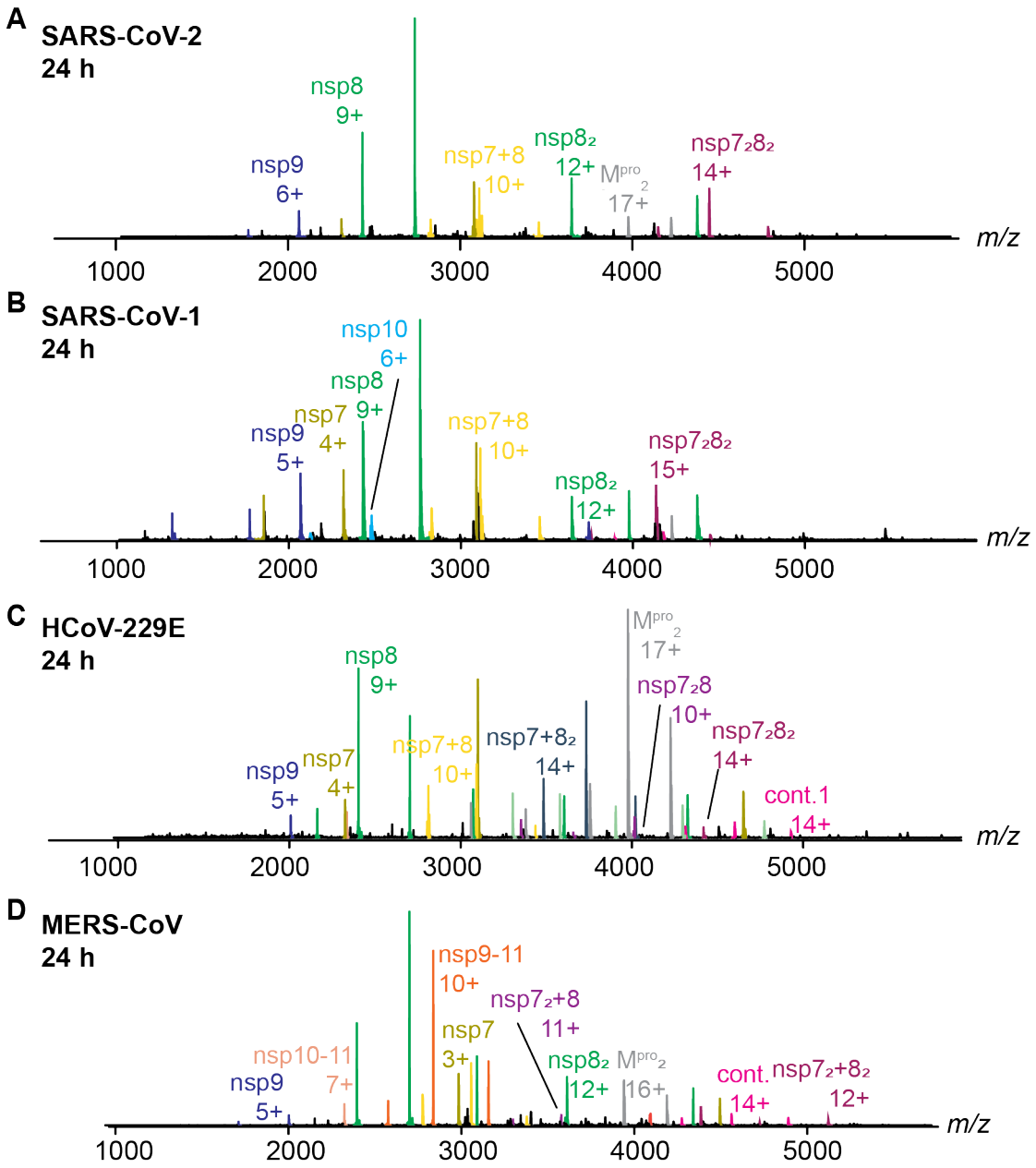


Figure S5: Complete polyprotein processing of nsp7-11 with authentic termini.
M^{pro} was added with a final concentration of ~3 µM to nsp7-11 of ~18 µM. Native MS spectra of four hCoVs: SARS-CoV-2 (**A**), SARS-COV-1 (**B**), HCoV-229E (**C**) and MERS-CoV (**D**) show complete or near complete processing of nsp7-11. All hCoVs show known heterotetramer complexes of $nsp7_2$+8_2 and the known heterotrimers $nsp7_2$+8 in MERS-CoV and HCoV-229E.

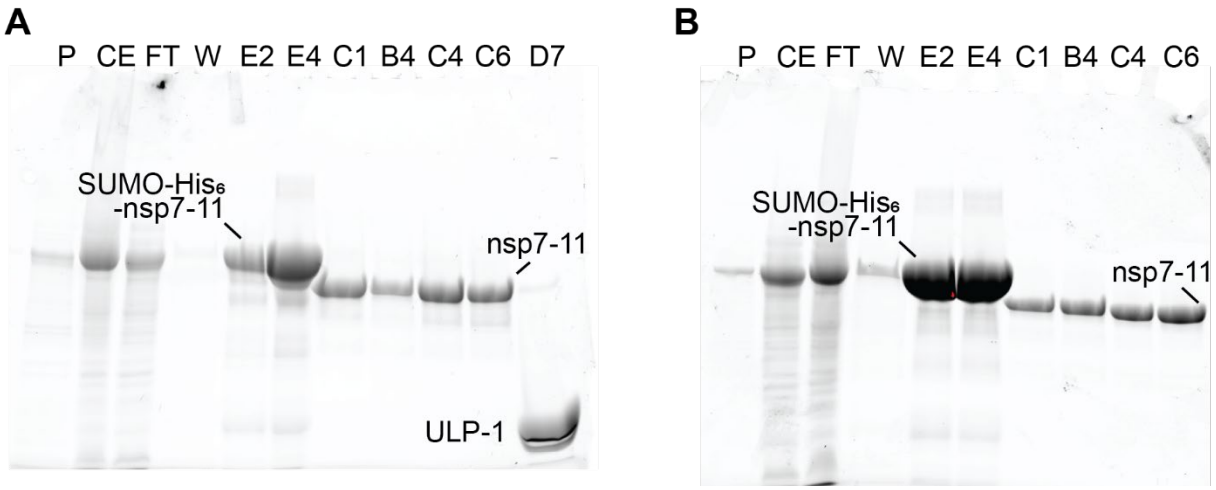


Figure S6: SDS-PAGE depicting steps in protein purification.
Protein purification of nsp7-11 HCoV-229E (**A**) and MERS-CoV (**B**) show a successful complete cleavage of the SUMO-His_6-tag. Indicated letters show P = pellet, CE = crude extract or supernatant, FT = flow-through after binding to Ni-agarose resin, W = washing step with 50 mM imidazole (after third column volume wash step), E1/E4= elution fractions with 300 mM imidazole and 10 min incubation at 4°C. B4 to C6 show SEC fractions after overnight digestion with ULP-1 the SUMO-protease that is cleaving of the SUMO-His_6-tag. SEC fraction are eluted after 9.6 ml (B4), after 11.2 ml (C1), after 12.4 ml (C4), after 13.2 ml (C6) and after 18 ml (D7), respectively.

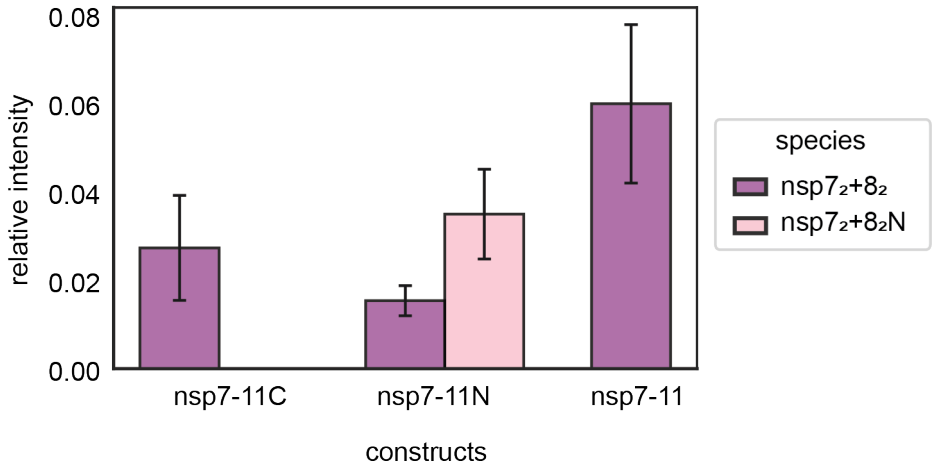


Figure S7: Quantitative comparison of heterotetrametric $nsp7_2+8_2$
Relative intensity of $nsp7_2+8_2$ of the three constructs nsp7-11C, nsp7-11N and nsp7-11 with authentic termini showing no significantly reduced complex formation within the three constructs.

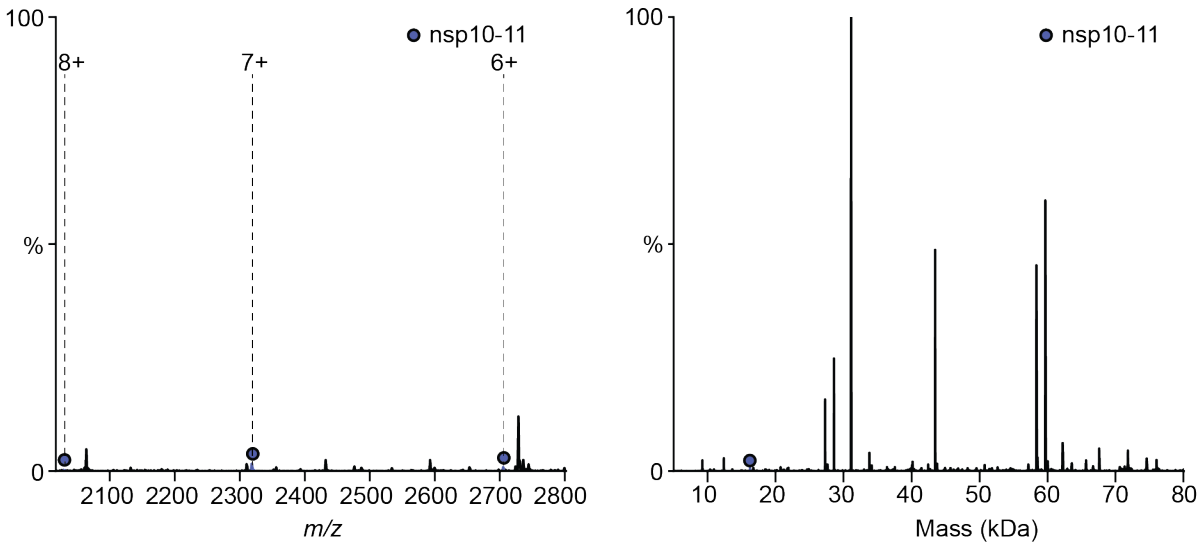


Figure S8: Unidec detection of low intensity species nsp10-11.
(**A**) nsp10-11 peaks detected by UniDec are shown with three charge states in native mass spectrum. (**B**) The deconvoluted nsp10-11 signal is displayed in the mass spectrum, highlighting its low intensity relative to other detected species.

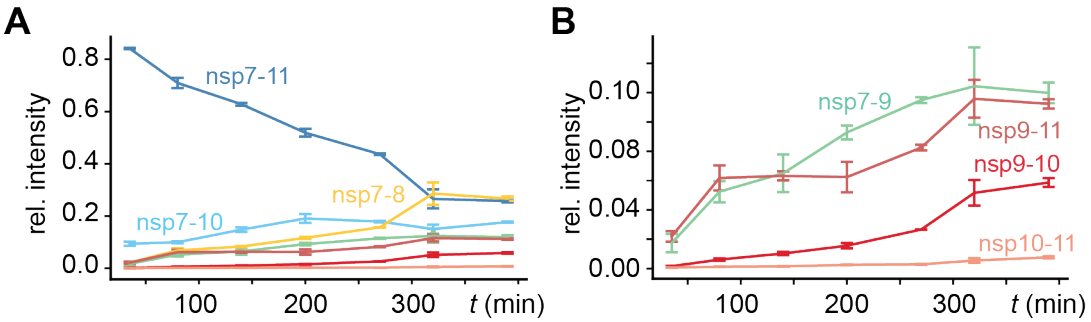


Figure S9: Discontinuous tag-free nsp7-11 polyprotein processing of SARS-CoV-2.
(**A**) shows relative intensities of all detected intermediates over the time. (**B**) shows the intermediate species below a relative intensity of 15% over the time. Error bars are standard error.

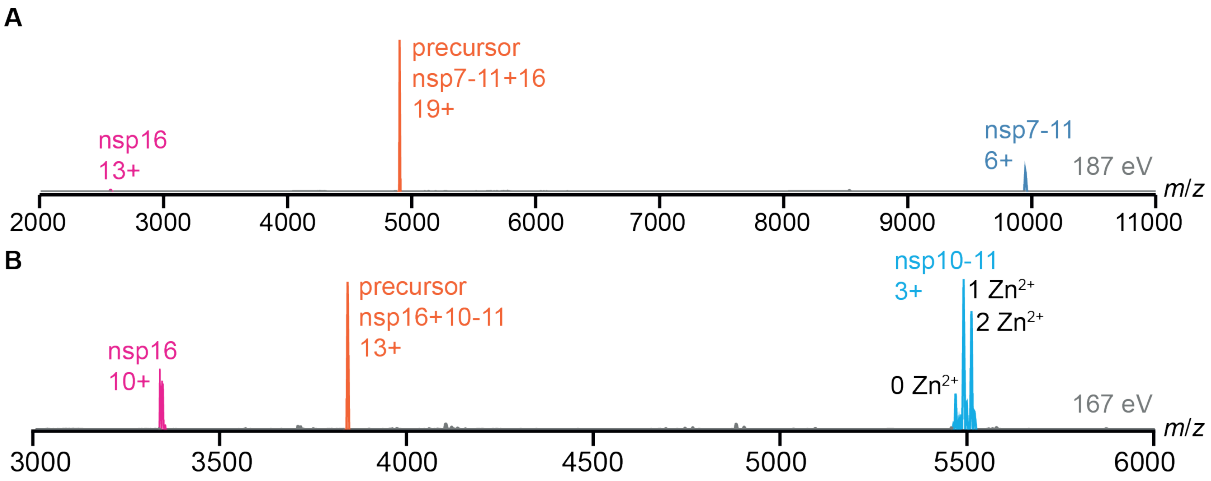
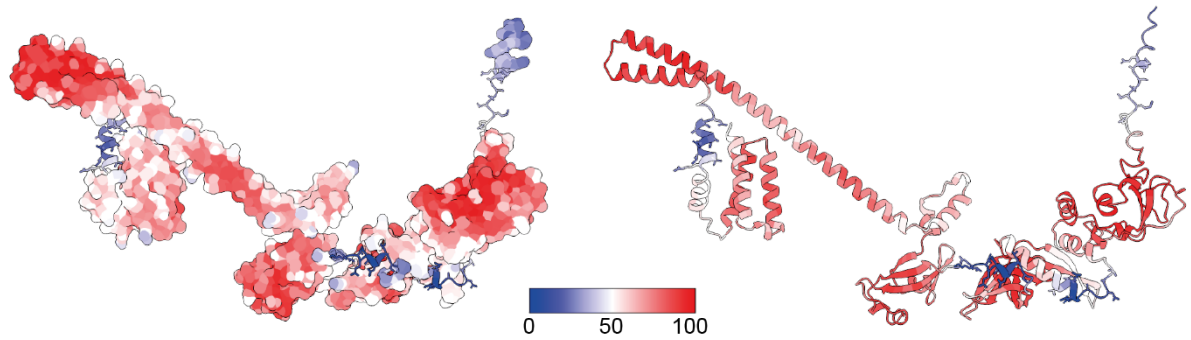


Figure S10: MS^2 of nsp7-11+16 and nsp16+10-11 complexes in MERS-CoV.
(**A**) Precursor peak *m*/*z* 4896 is the chimeric complex of MERS-CoV nsp7-11+SARS-CoV-2 nsp16. Dissociation of the complex is shown at 187 eV. (**B**) In processed MERS-CoV nsp7-11, nsp10-11 interacts with SARS-CoV-2 nsp16. Precursor peak *m*/*z* 3836, 13+ of nsp16+10-11 (orange) is dissociating at 167 eV. Nsp10 can bind up to two Zn^{2+} giving rise to the observed fine structure as indicated. Some Zn^{2+} remains with nsp16 upon nsp10 dissociation resulting in a double peak

A Representative nsp7-11 model of SARS-CoV-2 colored by pLDDT values



B Local cleavage sites confidence scores SARS-CoV-2

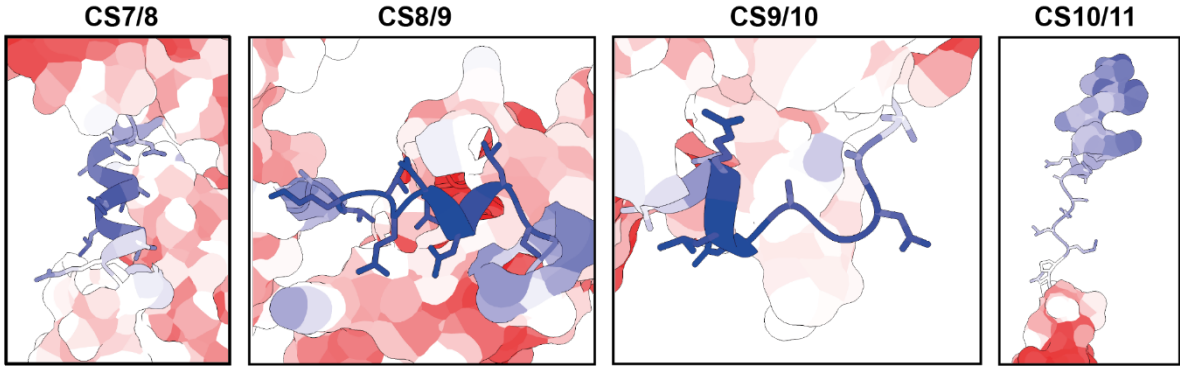


Figure S11: Local confidence scores of SARS-CoV-2 AF3 model presented on the structure.
pLDDT scores are depicted showing the overall folding (**A**) and the close-ups from the cleavage sites (**B**). The predicted template modeling (pTM) score is 0.33 and has thereby low confidence that the prediction of the folding would resemble the true strucutre. pTM scores above 0.5 are probably close to the real structure.

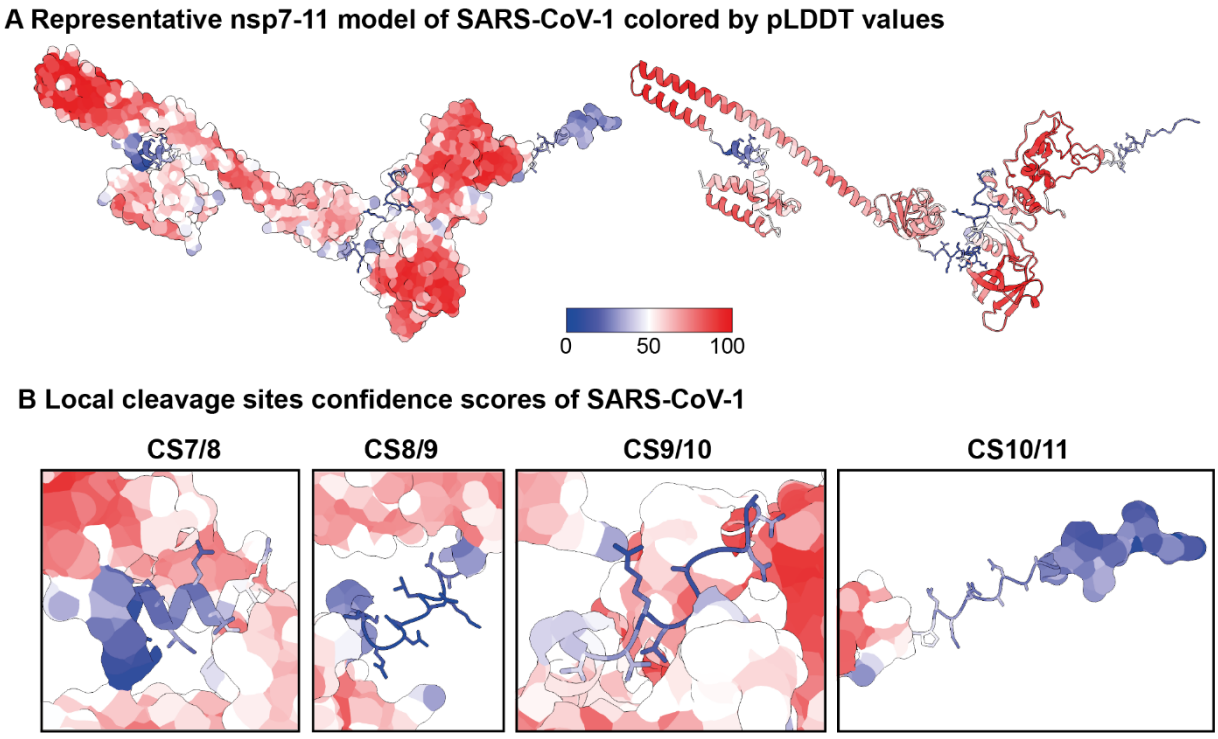


Figure S12: Local confidence scores of SARS-CoV-1 AF3 model presented on the structure.
Overview of model folding is illustrated with and without surface whereby cleavage sites are always shown as ribbon with residue sidechains (**A**). Zoom in on the cleavage site regions are shown in (B). pTM-score is low with 0.32 asserting low accuracy of the entire structure.

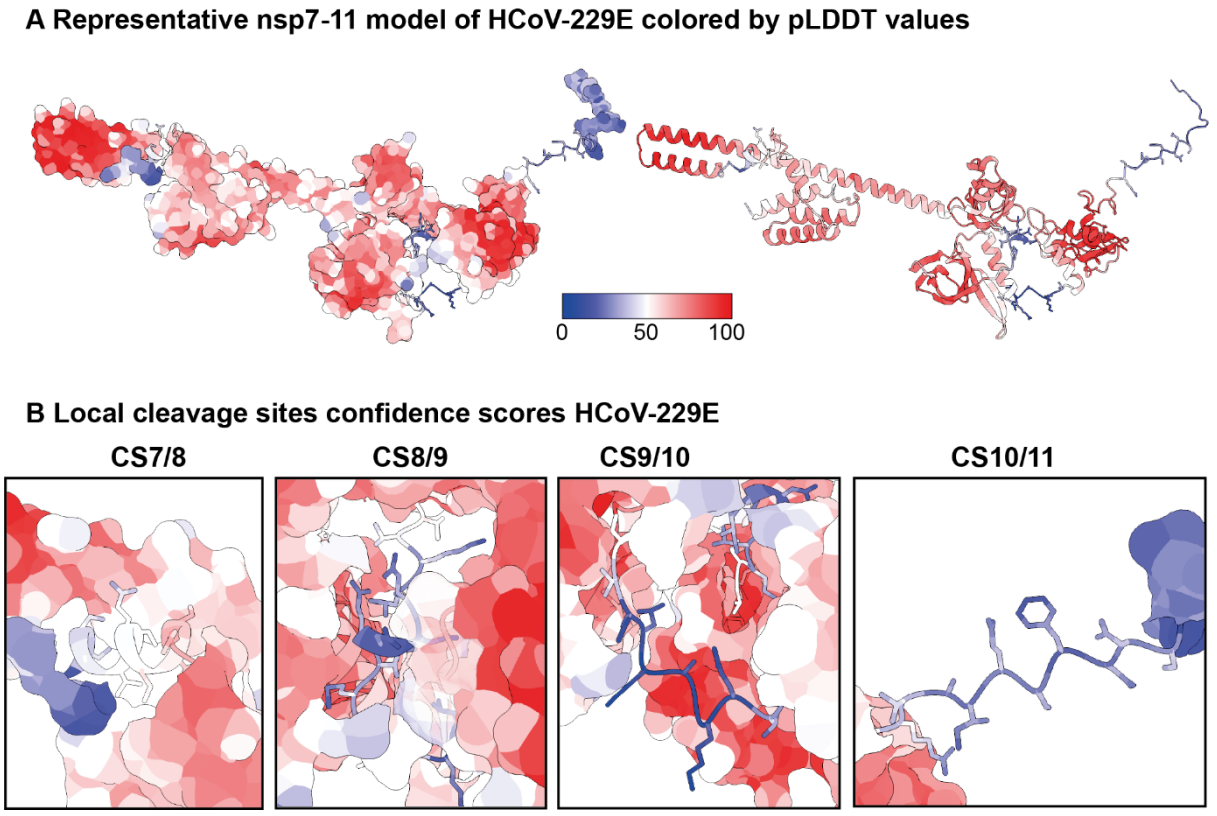


Figure S13: Local confidence scores of HCoV-229E AF3 model presented on the structure.
Overview of model folding is illustrated with and without surface whereby cleavage sites are always shown as ribbon with residue sidechains (**A**). Zoom in on the cleavage site regions are shown in (**B**). pTM-score is low with 0.39 asserting low accuracy of the entire structure.

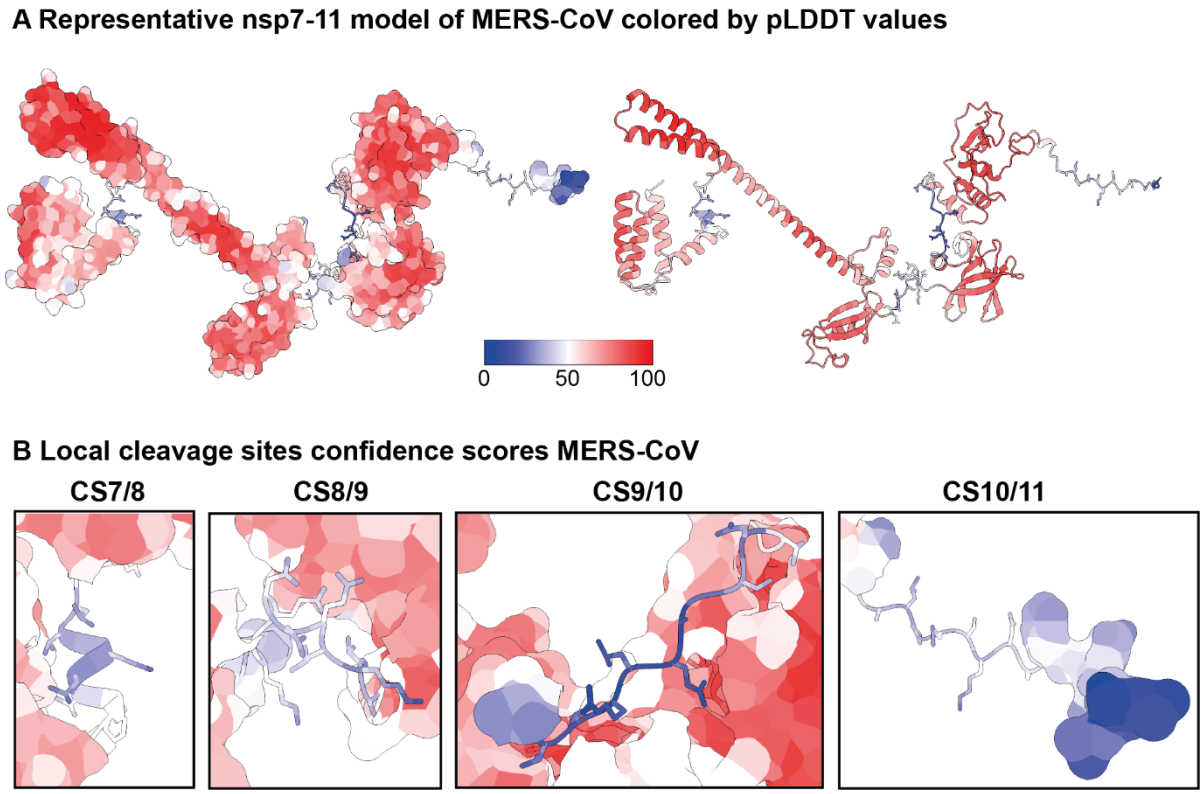


Figure S14: Local confidence scores of MERS-CoV AF3 model presented on the structure.
Overview of model folding is illustrated with and without surface whereby cleavage sites are always shown as ribbon with residue sidechains (**A**). Zoom in on the cleavage site regions are shown in (**B**). pTM-score is low with 0.37 asserting low accuracy of the entire structure.

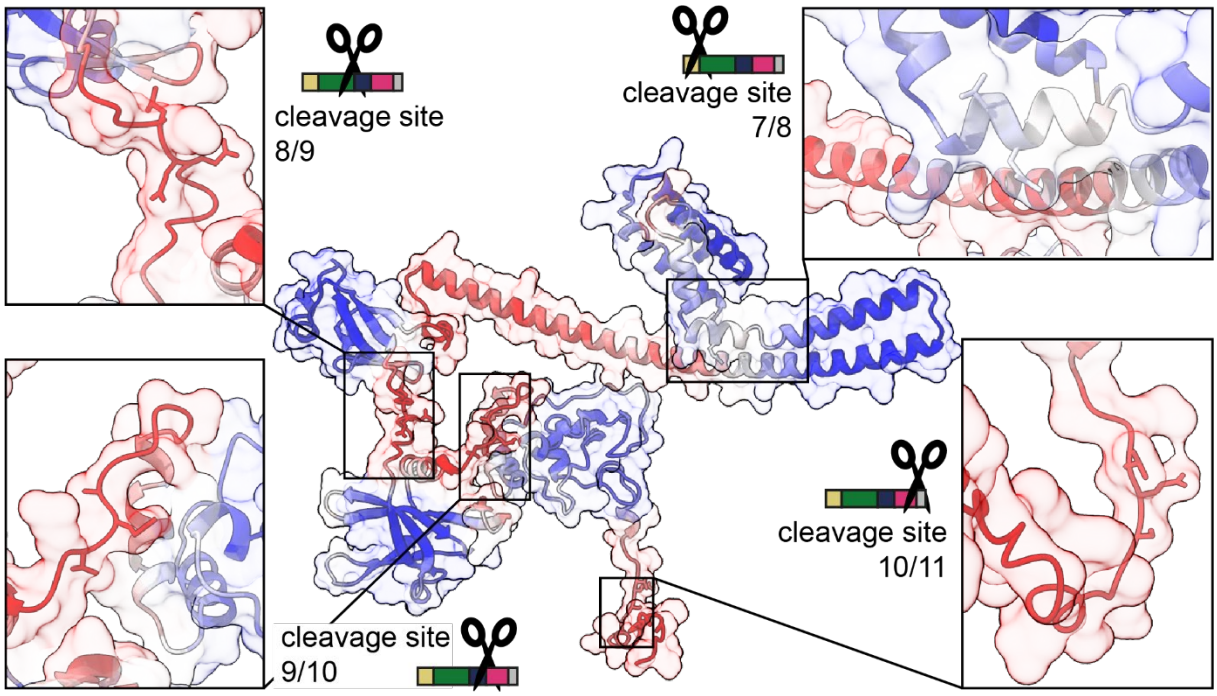


Figure S15: Local confidence scores of SARS-CoV-2 AF2 model presented on the structure.
SARS-CoV-2 nsp7-11 prediction model was colored by *B*-factor by using palette blue, white, red with ranges set from 50 to 90.

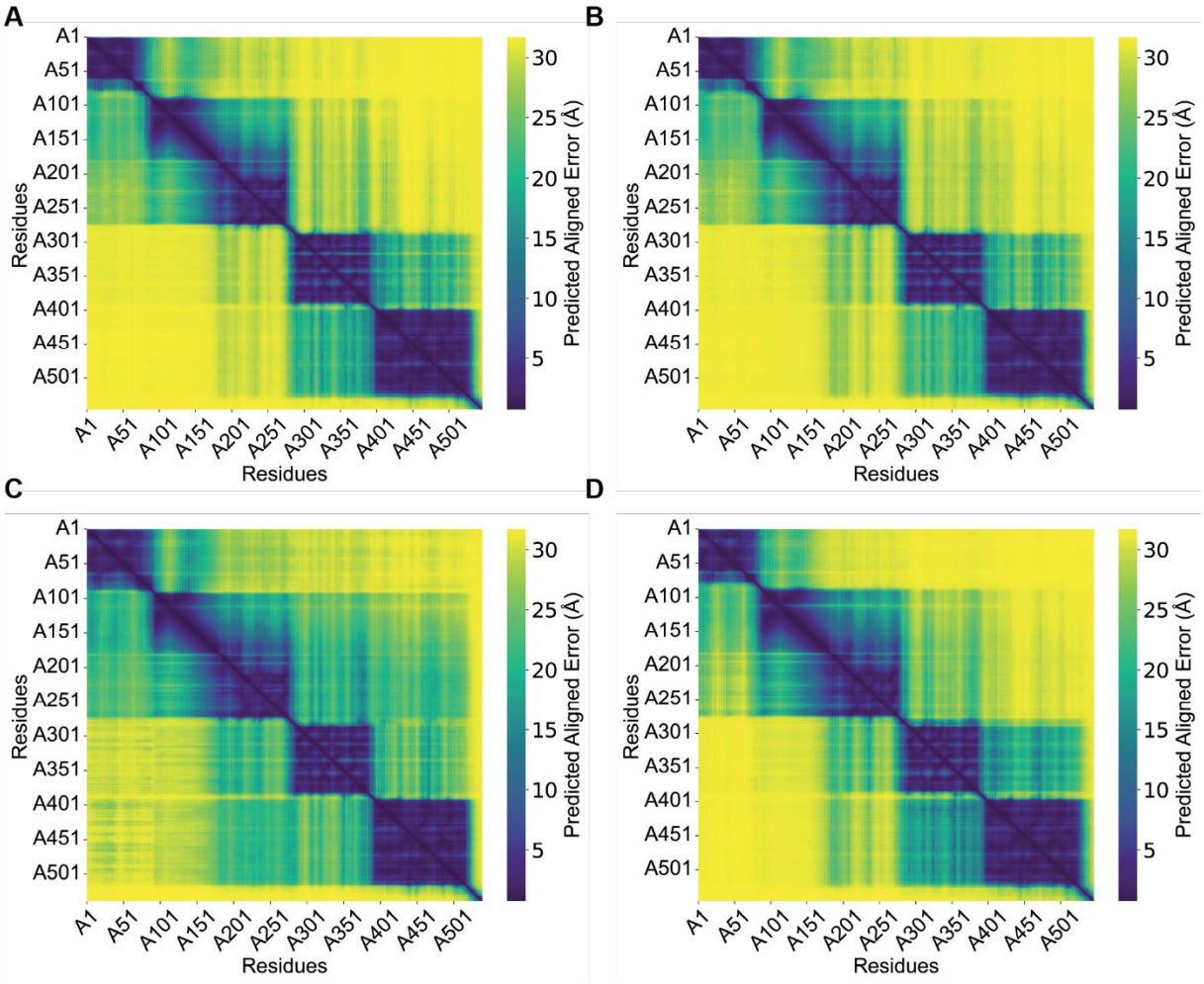


Figure S16: Global confidence plots for the selected models.
Predicted aligned errors (PAE) in Å are depicted as heatmaps showing higher errors (yellowish-green) around the cleavage site regions and lower errors (blueish) for the indivdual nsp domains. The four panels show PAE plots for SARS-CoV-2 (**A**), SARS-CoV-1 (**B**), HCoV-229E (**C**) and MERS-CoV (**D**).

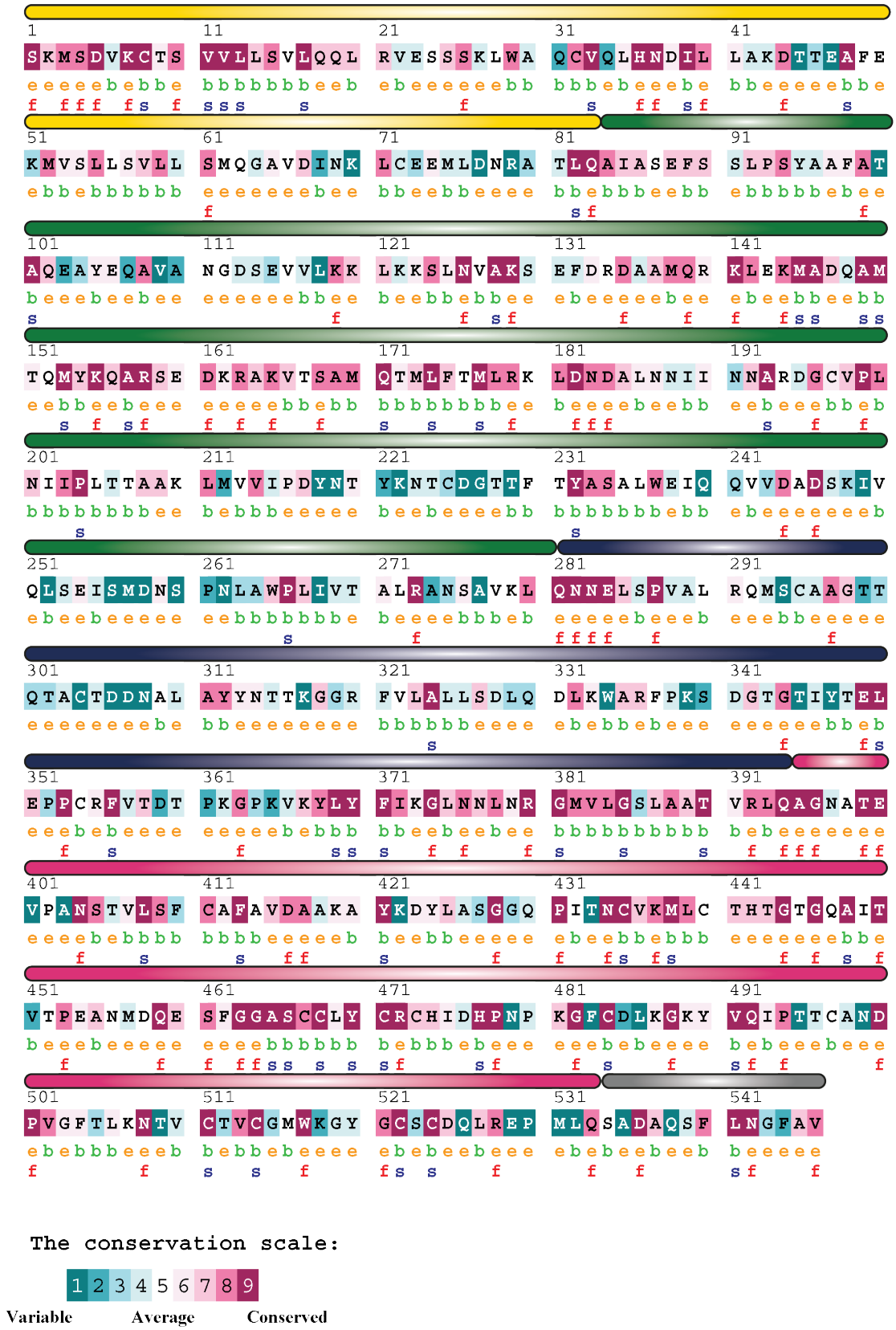


Figure S17: ConSurf Server output for SARS-CoV-2 nsp7-11 reveals patterns of residue conservation. The region is annotated with colored bars: nsp7 (yellow), nsp8 (forest green), nsp9 (purple), nsp10 (magenta), and nsp11 (gray). ConSurf calculates evolutionary conservation scores per residue using a deep multiple sequence alignment (MSA). Here, an MSA of 150 homologous sequences was generated using MAFFT (Multiple Alignment using Fast Fourier Transform), and conservation scores were computed with a Bayesian approach. Notably, most cleavage sites are highly conserved, with scores of 8 or 9, except for CS10/11, which shows specific variability at the P2 position.

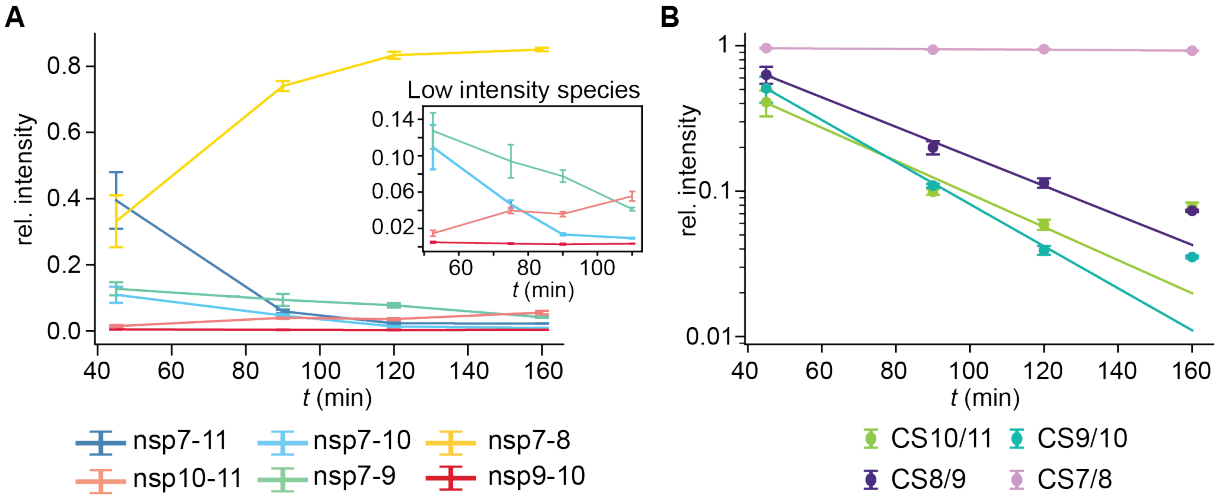


Figure S18: Discontinuous polyprotein processing of frozen SARS-CoV-1 nsp7-11.
Processing conditions were the same as in discontinuous processing with nsp7-11 directly used after purification with 18 µM of nsp7-11 were mixed with 3 µM of M^{pro}. (**A**) Relative intensity of intermediate species is plotted over time and low intensity species are shown in a smaller panel on the right. (**B**) Data points were fitted with an exponential model and kinetic rates were extracted. Errors bars are standard errors.

6.2. Supplementary tables

Table S1: High and low intensity species in two representative measurements.
Examples of two selected low intensity species and the highest intensity species are shown.

m/*z*	species	intensity	measurement
3979	nsp7-11	3.64×10^6	SARS-CoV-2, 120 min
2318	nsp10-11	8.22×10^3	SARS-CoV-2,120 min
2704.9	nsp10-11	5.96×10^3	SARS-CoV-2, 120 min
1348.6	nsp11	1.41×10^2	SARS-CoV-2, 120 min
3983	nsp7-11	1.41×10^7	SARS-CoV-1, 60 min
2322	nsp10-11	6.66×10^4	SARS-CoV-1, 60 min
2709	nsp10-11	2.59×10^4	SARS-CoV-1, 60 min
3247.7	nsp10-11	2.47×10^4	SARS-CoV-1 60 min
1854	nsp7	2.00×10^3	SARS-CoV-1, 60 min

Table S2: Measured mass species of nsp7-11C and nsp7-11N processing.
Measured masses were averaged and standard error is given.

Name	Theoretical Mass (in Da)	Measured mass (in Da)
nsp7	9,239.8	9,239.0 ± 0.1
nsp7N	10,649.3	10,648.6 ± 1
nsp9	12,378.2	12,378.3 ± 0.2
nsp10+2 Zn^{2+}	14,919.9	14,915.7 ± 0.2
nsp10-11+2 Zn^{2+}	17,509.7	17,506 ± 1
$nsp7_2$	18,479.6	18,479.8 ± 0.2
nsp8	21,881.1	21,880.4 ± 0.1
nsp9-10+2 Zn^{2+}	27,280.1	27,276.6 ± 0.1
nsp9-11+2 Zn^{2+}	29,869.8	29,866.2 ± 0.2
nsp7-8	31,102.9	31,103 ± 1
nsp7+8	31,120.9	31,120 ± 1
nsp7-8N	32,512.3	32,512 ± 1
nsp7N+8	32,530.4	32,530 ± 1.2
M^{pro}	33,668.5	33,796 ± 1
nsp7-9	43,463.1	43,462 ± 1
$nsp8_2$	43,762.2	43,762 ± 2
nsp7-10C+2 Zn^{2+}	58,365.0	58,360 ± 2
nsp7-10N+2 Zn^{2+}	59,774.7	59,774 ± 3
nsp7-11C+2 Zn^{2+}	60,954.7	60,950 ± 4
nsp7-11N+2 Zn^{2+}	61,082.9	61,085 ± 1
$nsp7_2+8_2$	62,241.8	62,244 ± 5
nsp7N+7C +8_2	63,651.3	63,655 ± 3
$nsp7N_2+8_2$	65,024.6	65,061 ± 3
$M^{pro}{}_2$	67,337.0	67,598 ± 1

Table S3: Measured masses of SARS-CoV-2 nsps and their complexes.
Masses were determined from representative native mass spectra. Uncertainty was computed as standard error. Masses are reported in Daltons (Da).

Virus strain	Protein	Theoretical Mass (Da)	Measured Mass (Da)
SARS-CoV-2	nsp11	1325.6	1325.654 ± 0.001
SARS-CoV-2	nsp11+Na^+	1349.4	1347.63 ± 0.01
SARS-CoV-2	nsp7	9239.8	9239.1 ± 0.2
SARS-CoV-2	nsp9	12378.2	12377.6 ± 0.2
SARS-CoV-2	nsp10+2 Zn^{2+}	14789.9	14916.3 ± 0.2
SARS-CoV-2	nsp10-11+2 Zn^{2+}	16228.3	16224 ± 1
SARS-CoV-2	nsp8	21881.1	21880.7 ± 0.2
SARS-CoV-2	nsp9-10+2 Zn^{2+}	27280.1	27275.9 ± 0.2
SARS-CoV-2	nsp9-11+2 Zn^{2+}	28588.5	28584.7 ± 0.5
SARS-CoV-2	nsp7-8	31102.9	31102.3 ± 0.3
SARS-CoV-2	nsp7+8	31120.9	31102.7 ± 0.3
SARS-CoV-2	nsp7-9	43463.1	43460 ± 2
SARS-CoV-2	$nsp8_2$	43762.2	43761 ± 2
SARS-CoV-2	nsp7-10+2 Zn^{2+}	58235	58366 ± 1
SARS-CoV-2	nsp7-11+2 Zn^{2+}	59543.4	59677 ± 3
SARS-CoV-2	$nsp7_2$+8_2	62241.6	62244 ± 2

Table S4: Measured masses of SARS-CoV-1 nsps and their complexes determined by native MS.
Masses were determined from representative mass spectra. Uncertainty was computed as standard error. Masses are reported in Daltons (Da).

Virus strain	Protein	Theoretical Mass (Da)	Measured Mass (Da)
SARS-CoV-1	nsp7	9267.8	9267.3 ± 0.2
SARS-CoV-1	nsp9	12401.2	12400.8 ± 0.3
SARS-CoV-1	nsp10+2 Zn^{2+}	14844	14970.4 ± 0.3
SARS-CoV-1	nsp10-11+2 Zn^{2+}	16125.4	16233.1 ± 0.5
SARS-CoV-1	$nsp7_2$	18535.7	18535 ± 1
SARS-CoV-1	nsp8	21866	21865.7 ± 0.3
SARS-CoV-1	nsp9-10+2 Zn^{2+}	27227.1	27353 ± 3
SARS-CoV-1	nsp9-11+2 Zn^{2+}	28508.5	28635.3 ± 0.8
SARS-CoV-1	nsp7-8	31115.9	31115.7 ± 0.2
SARS-CoV-1	nsp7-9	43499	43497 ± 2
SARS-CoV-1	$nsp8_2$	43732.1	43731 ± 1
SARS-CoV-1	nsp7-10+2 Zn^{2+}	58325	58454 ± 2
SARS-CoV-1	nsp7-11+2 Zn^{2+}	59606.3	59753 ± 4
SARS-CoV-1	$nsp7_2$+8_2	62267.7	62264 ± 4

Table S5: Measured masses of MERS-CoV nsps and their complexes.
Masses were determined from representative native mass spectra. Uncertainty was computed as standard error. Masses are reported in Daltons (Da).

Virus strain	Protein	Theoretical Mass (Da)	Measured Mass (Da)
MERS-CoV	nsp11	1522.7	1522.80 ± 0.01
MERS-CoV	nsp11+Na^+	1545.7	1543.79 ± 0.01
MERS-CoV	nsp7	9063.4	9063.2 ± 0.7
MERS-CoV	nsp9	12236.2	12235.2 ± 0.4
MERS-CoV	nsp10+2 Zn^{2+}	15020.9	15018 ± 2
MERS-CoV	nsp10-11+2 Zn^{2+}	16525.6	16521.8 ± 0.3
MERS-CoV	nsp8	21886.2	21885.4 ± 0.2
MERS-CoV	nsp9-11+2 Zn^{2+}	28743.8	28741 ± 1
MERS-CoV	nsp7-8	30931.5	30931.1 ± 0.2
MERS-CoV	$nsp7_2 8$	40012.9	40011 ± 1
MERS-CoV	nsp7-9	43149.7	43149 ± 1
MERS-CoV	$nsp8_2$	43772.4	43770.2 ± 0.5
MERS-CoV	nsp7-11+2 Zn^{2+}	59527.3	59654 ± 1
MERS-CoV	$nsp7_2 8_2$	61899.4	61898 ± 4

Table S6: Measured masses of HCoV-229E nsps and their complexes determined by native MS.
Masses were determined from representative mass spectra. Uncertainty was computed as standard error. Masses are reported in Daltons (Da).

Virus strain	Protein	Theoretical Mass (Da)	Measured Mass
HCoV-229E	nsp11	1841.0	1841.92 ± 0.03
HCoV-229E	nsp7	9300.6	9299.8 ± 0.8
HCoV-229E	nsp9	12045.9	12045.5 ± 0.2
HCoV-229E	nsp10+2 Zn^{2+}	14525.4	14521 ± 1
HCoV-229E	nsp10-11+2 Zn^{2+}	16348.3	16345 ± 6
HCoV-229E	$nsp7_2$	18601.3	18600.5 ± 0.3
HCoV-229E	nsp8	21623.1	21622.9 ± 0.3
HCoV-229E	$nsp8_2$	24091.7	43248 ± 1
HCoV-229E	nsp7-8	30905.7	30908 ± 3
HCoV-229E	nsp8-9	33651	33651 ± 1
HCoV-229E	$nsp7_2 8$	40224.4	40224 ± 4
HCoV-229E	nsp7-9	42933.6	42932 ± 2
HCoV-229E	nsp8-11+2 Zn^{2+}	49981.3	49992 ± 16
HCoV-229E	nsp7-10+2 Zn^{2+}	57440.9	57440.9 ± 0.4
HCoV-229E	nsp7-11+2 Zn^{2+}	59263.9	59256 ± 25
HCoV-229E	$nsp8\text{-}9_2$	67302	67296 ±10

Table S7: Measured mass of SARS-CoV-2 and MERS-CoV nsp7-11 and their complexes with SARS-CoV-2 nsp16 determined by native MS.

Masses were determined from representative mass spectra. Uncertainty was computed as standard error. Masses are reported in Daltons (Da).

Virus strain	Protein	Theoretical Mass (Da)	Measured Mass ± Uncertainty (Da)
SARS-CoV-2	nsp7	9239.8	9239.25 ± 0.23
SARS-CoV-2	nsp9	12378.2	12377.55 ± 0.12
SARS-CoV-2	nsp10+2 Zn^{2+}	14920.7	14918 ± 4
SARS-CoV-2	$nsp7_2$	18479.6	18478 ± 1
SARS-CoV-2	nsp8	21881.1	21881.04 ± 0.12
SARS-CoV-2	$nsp9_2$	24756.4	24755.3 ± 0.6
SARS-CoV-2	nsp7-8/nsp7+8	31120.9	31111 ± 10
SARS-CoV-2	nsp16	33323.3	33323.27 ± 0.14
SARS-CoV-2	nsp16-His_6	34644.8	34774 ± 1
SARS-CoV-2	nsp16+10+2 Zn^{2+}	48224.0	48236 ± 1
SARS-CoV-2	nsp7-11+2 Zn^{2+}	59674.2	59674 ± 3
SARS-CoV-2	$nsp7_2$+8_2	62241.6	62251 ± 8
SARS-CoV-2	nsp7-11+2 Zn^{2+}+16	92997.5	93008 ± 6
MERS-CoV	nsp9	12236.2	12235.93 ± 0.24
MERS-CoV	nsp10-11+2 Zn^{2+}	16526.4	16522 ± 1
MERS-CoV	$nsp7_2$	18126.8	18126.01 ± 0.34
MERS-CoV	nsp8	21886.2	21886.22 ± 0.13
MERS-CoV	nsp7+8	30949.6	30949.87 ± 0.43
MERS-CoV	nsp16	33323.3	33322 ± 1
MERS-CoV	nsp16-His_6	34644.8	34775 ± 1
MERS-CoV	$nsp7_2$8	40012.9	40014 ± 1
MERS-CoV	nsp16+10-11+2 Zn^{2+}	49849.7	49848 ± 2
MERS-CoV	nsp7-11+2 Zn^{2+}	59658.0	59658 ± 4
MERS-CoV	$nsp7_2 8_2$	61899.4	61898 ± 7
MERS-CoV	nsp7	9063.4	9062.88 ± 0.20
MERS-CoV	nsp7-11+2 Zn^{2+}+16	92981.4	93001 ± 22
MERS-CoV	nsp7-11+2 Zn^{2+}+16-His_6	94302.8	94449 ± 7

Table S8: Local confidence score of SARS-CoV-2 nsp7-11 AF2 prediction model.

Cleavage area is here defined as 5 residues before and after the corresponding cleavage sites LQ/A/S/N. Confidence scores from AF2 for the cleavage area were averaged.

Cleavage area	Residues	Average confidence score
CS7/8	77-89	73.5
CS8/9	275-287	34.9
CS9/10	388-400	39.6
CS10/11	527-539	40.9

6.3. Custom Python script

```
# %%
import numpy as np
import pandas as pd
import glob
import matplotlib.pyplot as plt
import sys
from numpy import log as ln
from lmfit import Model
import os

def exp_decay(x,a,k):
    return a*np.exp(-k*x)

def get_lengths_of_sublists(list_of_lists):
    # Use list comprehension to get the lengths of each sublist
    return [len(sublist) for sublist in list_of_lists]

def make_list_species(peaks, species_list):
    """
    Create a list of species with their charge states.

    Args:
    - peaks: DataFrame containing peak information
    - species_list: List of species names

    Returns:
    - list: List of species with charge states
    """
    list = []
    for i in range(len(species_list)):
        charge_states_species = peaks.query("Name=='"+str(species_list[i])+"'")['charge
states'].values
        for j in range(len(charge_states_species)):
            list = np.append(list, str(species_list[i]+'_'+str(charge_states_species[j])))
    return list


# %%
plt.close('all')
fgc = 0  # Figure counter

# path to save plots
path_to_save = 'C:/Users/kischa/Nextcloud/Corona/revision_polyprotein_processing/Sumo-
nsp711_SARSCOV2//20240912_sars2_processing/20240912_sars2_corrected_peaks_figure/'
path = 'C:/Users/kischa/Nextcloud/Corona/revision_polyprotein_processing/Sumo-
nsp711_SARSCOV2/'
```

```
files_l = glob.glob(path + '20240912_sars2_processing/*.txt')
peaks_l = pd.read_excel(path + 'sumo_nsp711_SARSCOV2_masslist.xlsx', sheet_name='Tabelle1')

peaks_l['charge states'] = peaks_l['charge states'].astype(int)
species_list_all_l = peaks_l.Name.tolist()

# Create species peak lists
species_peak_list_l = make_list_species(peaks_l, species_list_all_l)
```

6.3.1. Extraction of intensities and peak overlap correction

```
# %% [markdown]
# # Intensity Proportionality Approach to estimate intensity of overlapping peak nsp711 15+
and Mpro dimer 17+

# %% [markdown]
# # lgda function edit and uses "corrected" intensities for nsp711 and mpro

# %%
def lgda(files, species_peak_list, df_l, fgc, peaks_l, species_list):
    """

    Args:
    - files: List of file paths
    - species_peak_list: List of species with charge states
    - df: DataFrame for storing data
    - fgc: Not used in the function
    - peaks: DataFrame containing peak information
    - species_list: List of species names


    Returns:
    - df: Updated DataFrame with data
    - fgc: Not used in the function
    - time: Time values
    - df_final_result: Final DataFrame with LGDA results
    """
    # Loop over files
    # Sum the intensities in the gates
    for i in range(len(files)):
        df_spec = pd.read_csv(str(files[i]), skiprows=7, delimiter='\t')
        for j in range(len(species_peak_list)):
            str1 = species_peak_list[j]
            str2 = '_'
            where = str1.find(str2)
            # Extract gate information from peaks DataFrame
            gate_low = peaks_l.query("Name=='"+str(species_peak_list[j][:where])+"' and
`charge states`=="+str(species_peak_list[j][(where+1):]))['m/z min'].values[0]
            gate_high = peaks_l.query("Name=='"+str(species_peak_list[j][:where])+"' and
`charge states`=="+str(species_peak_list[j][(where+1):]))['m/z max'].values[0]
```

```
            # Sum the intensities within the gate
            sum_intensity = df_spec.query(f"{gate_low} < Mass <
{gate_high}")['Intensity'].sum()
            min_t = int(files[i][-12:-9])  # Time
            digest = files[i][-5:-4]  # Experiment cycle

            # Create a temporary DataFrame to store the data
            df_temp = pd.DataFrame({'species': str(species_peak_list[j][:where]), 'charge
state': str(species_peak_list[j][(where+1):]),
                                    'digest': digest, 'time': min_t, 'intensity':
sum_intensity}, index=[0])
            df_l = pd.concat([df_l, df_temp])

    df_l = df_l.sort_values(by=['digest', 'time', 'charge state'])
    df_l = df_l.reset_index(drop=True)
    df_l['charge state'] = df_l['charge state'].astype(float, errors='raise')

    # Step 1: Filter adjacent peaks
    df_l = df_l.drop_duplicates()
    df_711 = df_l[(df_l['species'] == "nsp7-11") & (df_l['charge state'] == 14)] # Filter
the dataframe for nsp7-11 with charge state 14
    df_mprod = df_l[(df_l['species'] == "Mprod") & (df_l['charge state'] == 16)] # Filter
the dataframe for Mprod with charge state 16

    # Step 2: Merge data on 'time' and 'digest' to align intensities of the adjacent peaks
    df_tint = pd.merge(df_711, df_mprod, on=['time', 'digest'], suffixes=('_nsp7_11',
'_Mprod')) # This results in a DataFrame with combined adjacent peaks per time and digest
    # Step 3: Calculate the sum of intensities for adjacent peaks
    df_tint['intensity_total'] = df_tint['intensity_nsp7_11'] + df_tint['intensity_Mprod']


    # Step 4: Calculate the intensity ratios for each species at each time and digest
    df_tint['ratio_nsp7_11'] = df_tint['intensity_nsp7_11'] / df_tint['intensity_total'] #
This gives the proportional contribution of each adjacent peak
    df_tint['ratio_Mprod'] = df_tint['intensity_Mprod'] / df_tint['intensity_total']

    # Step 5: Identify the overlapping peak (nsp7-11 15+ and Mprod 17+) in testdf
    # Filter the overlapping peak data in testdf
    overlap_df = df_l[(df_l['species'] == "nsp7-11") & (df_l['charge state'] == 15)]
    overlap_df = overlap_df.merge(df_tint[['time', 'digest', 'ratio_nsp7_11',
'ratio_Mprod']], on=['time', 'digest'], how='left')

    # Step 6: Calculate corrected intensities for the overlapping peak
    overlap_df['intensity_nsp7_11_corrected'] = overlap_df['intensity'] *
overlap_df['ratio_nsp7_11']
    overlap_df['intensity_Mprod_corrected'] = overlap_df['intensity'] *
overlap_df['ratio_Mprod']
```

```
    # Step 7: Update the original dataframe with the corrected intensities for nsp7-11 15+
    # Ensure that all relevant columns have consistent data types
    df_l['charge state'] = df_l['charge state'].astype(float)
    df_l['time'] = df_l['time'].astype(float)
    overlap_df['charge state'] = overlap_df['charge state'].astype(float)
    overlap_df['time'] = overlap_df['time'].astype(float)

    # Update the intensity of nsp7-11 15+ in the main DataFrame and Mprod 17+
    for idx, row in overlap_df.iterrows():
    # Update the intensity for nsp7-11 15+
        df_l.loc[
        (df_l['species'] == "nsp7-11") &
        (df_l['charge state'] == 15) &
        (df_l['time'] == row['time']) &
        (df_l['digest'] == row['digest']),
        'intensity'
    ] = row['intensity_nsp7_11_corrected']

    # Create a new row dictionary for Mprod 16+ with corrected intensity
        df_l.loc[
            (df_l['species'] == "Mprod") &
            (df_l['charge state'] == 17) &
            (df_l['time'] == row['time']) &
            (df_l['digest'] == row['digest']),
            'intensity'
        ] = row['intensity_Mprod_corrected']

    # Plot Individual Runs of selected species
    df_final_result = pd.DataFrame(columns=['species', 'repetiton', 'time', 'intensity'])
    repetition = ['A', 'B', 'C']

    for j in range(len(species_list)):
        for i in range(len(repetition)):
            # Query the data based on species and repetition
            time = df_l.query("species == '"+str(species_list[j])+"' and digest ==
'"+str(repetition[i])+"'")["time"].unique()
            intensity = df_l.query("species == '"+str(species_list[j])+"' and digest ==
'"+str(repetition[i])+"'").groupby(['time'])['intensity'].mean()

            # Create a temporary DataFrame to store the data
            df_temp = pd.DataFrame({'species': str(species_list[j]), 'repetiton':
repetition[i], 'time': time, 'intensity': intensity})
            df_final_result = pd.concat([df_final_result, df_temp])

    df_final_result = df_final_result.reset_index(drop=True)

    return df_l, fgc, time, df_final_result

df_l = pd.DataFrame(columns=['species', 'charge state', 'digest', 'time', 'intensity'])
```

```
df_l, fgc_l, time_l, df_fr_l = lgda(files_l, species_peak_list_l, df_l, fgc, peaks_l,
species_list_all_l)

# %%
peaks_l.Name.unique()
array(['nsp10-11', 'nsp7-10', 'nsp7-11', 'nsp7-8', 'nsp7-8t', 'nsp7-9',
       'nsp7d', 'nsp7m', 'nsp8d', 'nsp8m', 'nsp9-10', 'nsp9-11', 'nsp9m',
       'nsp11', 'nsp10', 'Mprod'], dtype=object)
```

6.3.2. Domain correction

```
# %%
# domain correction using a multiplication array

mutliplication_ar = np.array([2,4,5,2,0,3,0,0,0,0,2,3,0,0,0,0])
v_l = np.stack((peaks_l.Name.unique(),mutliplication_ar))

df_fr_l['multiplication_factor'] = ''


for i in range (df_fr_l.shape[0]):
    for j in range (v_l.shape[1]):
        if df_fr_l.species.iloc[i] == v_l[0,j]: df_fr_l['multiplication_factor'].iloc[i] =
v_l[1,j]

# normalization of df considering species domains
df_fr_l['m_intensity'] = df_fr_l.intensity*df_fr_l.multiplication_factor
df_fr_l['m_intensity'] = pd.to_numeric(df_fr_l['m_intensity'], errors='coerce')
```

6.3.3. Assignment of substrate species and normalization

```
# assigning substrate species to their cleavage sites
sp_list = ['nsp7-11','nsp7-10', 'nsp7-9', 'nsp7-8', 'nsp9-10','nsp10-11', 'nsp9-11']

species_list_10_11 = ['nsp7-11','nsp10-11','nsp9-11']
species_list_9_10 = ['nsp7-11','nsp7-10','nsp9-10','nsp9-11']
species_list_8_9 = ['nsp7-11','nsp7-10','nsp7-9']
species_list_7_8 = ['nsp7-11','nsp7-10','nsp7-9','nsp7-8']

cs_species_all = [species_list_10_11,species_list_9_10,species_list_8_9,species_list_7_8]
cs_name_list = ['cs_10_11','cs_9_10','cs_8_9','cs_7_8']

# %%
def data_analysis(df_fr_l, species_list, normalization=None, error='standard error'):
    """
    Function for data analysis.

    Parameters:
    - df_fr (DataFrame): DataFrame containing the data.
```

```
    - species_list (list): List of species names.
    - normalization (str, optional): Normalization method. Default is None.
    - error (str, optional): Error calculation method. Default is 'standard error'.

    Returns:
    - ar_mean (array): Array containing the mean values for each species.
    - ar_err (array): Array containing the error values for each species.
    """
    len1 = len(df_fr_l.query("species == '" + str(species_list[0]) + "' and repetiton == '"
+
                             str(df_fr_l.repetiton.unique()[0]) +
"'").groupby(['time'])['m_intensity'].mean())
    len2 = len(df_fr_l.repetiton.unique())
    len3 = len(species_list)
    ar1 = np.zeros((len1, len2, len3))

    # Populate ar1 array with mean intensity values
    for j in range(len(df_fr_l.repetiton.unique())):
        for i in range(len(species_list)):
            ar1[:, j, i] = df_fr_l.query("species == '" + str(species_list[i]) + "' and
repetiton == '" +
                                         str(df_fr_l.repetiton.unique()[j]) +
"'").groupby(['time'])['m_intensity'].mean()
            #print(ar1)
            sf = df_fr_l.query("repetiton ==
'"+str(df_fr_l.repetiton.unique()[j])+"'").groupby(['time',
'repetiton'])['m_intensity'].sum()
            ar1[:, j, i] = ar1[:, j, i]/sf # Normalization: All species considered are 100%
                                           #at each time point per replicate

    ar_mean = np.mean(ar1, axis=1)  # Calculate the mean values

    ar_err = np.std(ar1, axis=1) / np.sqrt(ar1.shape[1])  # Calculate the error as the
standard error

    return ar_mean, ar_err

# Perform data analysis for different combinations of normalization and error methods

ar_mean_l_n1_err1, ar_err_l_n1_err1 = data_analysis(df_fr_l, sp_list)

# %%
def data_analysis2(df_fr_l, species_list,cs_species_all):
    """
    Function for data analysis.

    Parameters:
    - df_fr (DataFrame): DataFrame containing the data.
    - species_list (list): List of species names.
```

```
    - normalization (str, optional): Normalization method. Default is None.
    - error (str, optional): Error calculation method. Default is 'standard error'.

    Returns:
    - ar_mean (array): Array containing the mean values for each species.
    - ar_err (array): Array containing the error values for each species.
    """
    df_fr_l = df_fr_l.drop_duplicates()
    len1 = len(df_fr_l.query("species == '" + str(species_list[0]) + "' and repetiton == '"
+
                             str(df_fr_l.repetiton.unique()[0]) +
"'").groupby(['time'])['m_intensity'].mean())
    len2 = len(df_fr_l.repetiton.unique())
    len3 = len(species_list)
    ar1 = np.zeros((len1, len2, len3))

    # Populate ar1 array with mean intensity values
    for j in range(len(df_fr_l.repetiton.unique())):
        for i in range(len(species_list)):
            ar1[:, j, i] = df_fr_l.query("species == '" + str(species_list[i]) + "' and
repetiton == '" +
                                         str(df_fr_l.repetiton.unique()[j]) +
"'").groupby(['time'])['m_intensity'].mean()

            sf = df_fr_l.query("repetiton ==
'"+str(df_fr_l.repetiton.unique()[j])+"'").groupby(['time',
'repetiton'])['m_intensity'].sum()
            ar1[:, j, i] = ar1[:, j, i]/sf

    ar_mean = np.mean(ar1, axis=1)  # Calculate the mean values

    ar_err = np.std(ar1, axis=1) / np.sqrt(ar1.shape[1])  # Calculate the error as the
standard error


    len0 = len(cs_species_all)

    len1 = len(df_fr_l.query("species == '" + str(species_list[0]) + "' and repetiton == '"
+
                             str(df_fr_l.repetiton.unique()[0]) +
"'").groupby(['time'])['m_intensity'].mean())
    len2 = len(df_fr_l.repetiton.unique())
    len3 = max(get_lengths_of_sublists(cs_species_all))

    cs_ar = np.zeros((len0,len1,len2,len3))


    for l in range(len0):
        len_m =  get_lengths_of_sublists(cs_species_all)[l]
```

```
        for j in range(len(df_fr_l.repetiton.unique())):
            for i in range(len_m):
                cs_ar[l,:, j, i] = df_fr_l.query("species == '" + str(cs_species_all[l][i])
+ "' and repetiton == '" +
                                        str(df_fr_l.repetiton.unique()[j]) +
"'").groupby(['time'])['m_intensity'].mean()
                sf = df_fr_l.query("repetiton ==
'"+str(df_fr_l.repetiton.unique()[j])+"'").groupby(['time',
'repetiton'])['m_intensity'].sum()
                cs_ar[l,:, j, i] = cs_ar[l,:, j, i]/sf


    cs_ar = np.sum(cs_ar,axis=-1)
    cs_mean = np.mean(cs_ar, axis=-1)  # Calculate the mean values
    cs_err = np.std(cs_ar, axis=-1) / np.sqrt(ar1.shape[1])  # Calculate the error as the
standard error

    return ar_mean, ar_err, cs_mean, cs_err

ar_mean_l, ar_err_l, cs_mean_l, cs_err_l = data_analysis2(df_fr_l, sp_list, cs_species_all)
```

6.3.4. Plotting of species intensities over time

```
# %%
##### two different lists of colors - one for species to match the colors in the paper -
the other for cs
colors_sp = ['#4783B4', '#72caf5', '#7fe0a1', '#ffcb3a', '#db1e2b', '#f09980', '#cc615f',
'#24435c', '#164e29']
colors_cs = ['#74C439', '#24B3AA','#462778','#d2a1ca']
xlbl= 't (min)'

# %%
def plot_data_long(ar_mean, ar_err, df_l, sp_list, colors_mix, xlbl):
    path_to_plot3 = path_to_save

    # Set the font properties globally
    plt.rcParams['font.family'] = 'Arial'
    plt.rcParams['font.size'] = 12
        # Set the thickness of axes
    plt.rcParams['axes.linewidth'] = 1.0


    fig, ax1 = plt.subplots(figsize=(6, 4), dpi=300)

    low_intensity_found = False  # Flag to check if low-intensity species are found

    for i in range(len(sp_list)):
        # Plot all species on the main plot
```

```
        ax1.errorbar(df_l.time.unique(), ar_mean[:, i], ar_err[:, i], color=colors_mix[i],
label=sp_list[i],
                     capsize=5, lw=2)

        if np.max(ar_mean[:, i]) < 0.2 * np.max(ar_mean):
            # If it's a low-intensity species, set the flag to True
            low_intensity_found = True

    ax1.set_xlabel(xlbl)
    ax1.set_ylabel('rel. intensity')

    #ax1.legend(frameon=False, loc='upper center', bbox_to_anchor=(0.5, -0.15), ncol=4)
    ax1.legend(frameon=False, loc='lower center', ncol=4)
    plt.savefig(path_to_plot3 +'20240912_species_int_nsp711_SARS2.svg')

    if not low_intensity_found:
        # If no low-intensity species were found, display a message in the interpreter
        print("No Low-Intensity Species Found.")
    else:
        # Create a new figure and axis for the low-intensity species
        fig_low_intensity, ax_low_intensity = plt.subplots(figsize=(6, 4), dpi=300)
        ax_low_intensity.set_xlabel(xlbl)  # Set the x-axis label for the low-intensity
plot
        ax_low_intensity.set_ylabel('rel. intensity')  # Set the y-axis label for the low-
intensity plot

        for i in range(len(sp_list)):
            if np.max(ar_mean[:, i]) < 0.2 * np.max(ar_mean):
         # If it's a low-intensity species, plot it on the low-intensity plot
                ax_low_intensity.errorbar(df_l.time.unique(), ar_mean[:, i], ar_err[:, i],
color=colors_mix[i], label=sp_list[i],
                                          capsize=5, lw=2)

               ax_low_intensity.legend(frameon=False, loc='lower center', ncol=4)  # Add a
legend to the low-intensity plot
               plt.savefig(path_to_plot3 + '20240912_species_int_nsp711_SARS2_low.svg')
    plt.show()  # Display either the low-intensity plot or nothing

# plot data
plot_data_long(ar_mean_l, ar_err_l, df_l, sp_list, colors_sp, xlbl)
```

6.3.5. Data fitting and plotting of cleavage site kinetics

```
# %%
def plot_data_cs(ar_mean, ar_err, df_fr, cs_species_all,cs_name_list, colors, xlbl):
    """
    Function to plot the data.

    Parameters:
```

```
    - ar_mean (array): Array containing the mean values for each species.
    - ar_err (array): Array containing the error values for each species.
    - df_fr (DataFrame): DataFrame containing the data.
    - sp_list (list): List of species names.
    - colors (list): List of colors for each species.
    - normalization (str): Normalization value.
    - error (str): Error value.
    - xlbl (str): Label for the x-axis.
    """

    fig, ax1 = plt.subplots(figsize=(6, 4), dpi = 300)  # Create a new figure and axis

    ins = 0  # Variable to track if an inset axis is created

    for i in range(len(cs_species_all)):
        # plt.figure()


        if np.max(ar_mean[i, :]) < 0.2 * np.max(ar_mean):

            if ins == 0:
                ins = 1
                # Create an inset axis for small values
                left, bottom, width, height = [0.523, 0.6, 0.23, 0.25]
                ax2 = fig.add_axes([left, bottom, width, height])

            # Plot the data with error bars on the appropriate axis
            ax2.errorbar(df_fr.time.unique(), ar_mean[i, :], ar_err[i, :], color=colors[i],
label=cs_name_list[i],
                         capsize=5, fmt = 'o',marker = None)
        else:
            ax1.errorbar(df_fr.time.unique(), ar_mean[i, :], ar_err[i, :], color=colors[i],
label=cs_name_list[i],
                         capsize=5,fmt = 'o',marker = None)

        gmodel = Model(exp_decay)
        result =gmodel.fit( ar_mean[i, :],x=df_fr.time.unique(),k=-0.01,a=7e7)
        a = result.best_values['a']
        k = result.best_values['k']
        k_error = result.params['k'].stderr # get the standard error for k
        print('Fitting parameters: species:', cs_name_list[i],' a=', a, ' k=', k,
'k_error=', k_error)
        ax1.plot(df_fr.time.unique(),result.best_fit,color=colors[i])

    ax1.set_xlabel(xlbl)  # Set the x-axis label for the main plot
    ax1.set_ylabel('rel. intensity')  # Set the y-axis label for the main plot
    ax1.legend(frameon=False, loc = 'lower center', ncol =4)  # Add a legend to the main
plot
```

```
    if ins == 1:
        ax2.set_xlabel(xlbl, fontsize=7)  # Set the x-axis label for the inset plot
        ax2.set_ylabel('rel. intensity', fontsize=7)  # Set the y-axis label for the inset
plot
        ax2.tick_params(labelsize=7)  # Set the tick label size for the inset plot
        ax2.yaxis.get_offset_text().set_fontsize(7)  # Set the font size of the offset text
for the inset plot
        ax2.legend(frameon=False, prop={'size': 6})  # Add a legend to the inset plot

# %%
plot_data_cs(cs_mean_l, cs_err_l, df_l,cs_species_all,cs_name_list, colors_cs, 't (min)')
plt.savefig( path_to_save + '20240912_cs_fit_nsp711_SARS2.svg')

plot_data_cs(cs_mean_l, cs_err_l, df_l,cs_species_all,cs_name_list, colors_cs, 't (min)')
plt.yscale('log')
ytick_values = [10**i for i in range(-1, 1)]  # define y-scale
plt.yticks(ytick_values, [str(val) for val in ytick_values]) # Set the y-tick values and
labels
plt.savefig(path_to_save + '20240912_cs_fit_log_nsp711_SARS2.svg')
```

6.4. Hazardous substances according to the GHS

Substance (CAS)	Hazard pictogram	Hazard statements	Precautionary statements
2-mercaptoethanol (60-24-2)		H301 + H331 Toxic if swallowed or if inhaled. H310 Fatal in contact with skin. H315 Causes skin irritation. H317 May cause an allergic skin reaction. H318 Causes serious eye damage. H361d Suspected of damaging the unborn child. H373 May cause damage to organs (Liver, Heart) through prolonged or repeated exposure if swallowed. H410 Very toxic to aquatic life with long lasting effects.	P273 Avoid release to the environment. P280 Wear protective gloves/protective clothing/eye protection/face protection. P301 + P310 IF SWALLOWED: Immediately call a POISON CENTER/doctor. P302 + P352 + P310 IF ON SKIN: Wash with plenty of water. Immediately call a POISON CENTER/doctor. P304 + P340 + P311 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Call a POISON CENTER/doctor. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.
3-(Morpholin-4-yl)propane-1-sulfonic acid		H315 Causes skin irritation. H319	P261 Avoid breathing dust/fume/gas/mist/vapours/spray P264

		Causes serious eye irritation H335 May cause respiratory irritation	Wash skin thoroughly after handling P271 Use only outdoors or in a well-ventilated area. P280 Wear protective gloves/protective clothing/eye protection/face protection. P302+P352 IF ON SKIN: Wash with sopa and water P304+P340 IF INHALED: Remove victim to fresh air and keep at rest in position comfortable for the breathing P305+P351+P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing. P312 Call a POISON Center or doctor/physician if you feel unwell P332+P313 If skin irritation occurs. Get medical advice/attention P362 Take off contaminated clothes P403+P233 Store in a well-ventilated place. Keep container tightly closed P405 Store locked up
acetic acid (64-19-7)		H226 Flammable liquid and vapor. H314 Causes severe skin burns and eye damage	P210 Keep away from heat, hot surfaces, sparks, open flames, and other ignition sources. No smoking. P280 Wear protective gloves/protective clothing/eye protection/face protection. P303+P361+P353 IF ON SKIN (or hair): Take off immediately all contaminated clothing. Rinse skin with water. P305+P351+P338+P310 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue

			rinsing. Immediately call a poison center/doctor.
ammonium acetate (631-61-8)	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
ammonium hydroxide solution (1336-21-6)		H314 Causes severe skin burns and eye damage. H335 May cause respiratory irritation. H410 Very toxic to aquatic life with long lasting effects.	P261 Avoid breathing dust/fume/gas/mist/vapors/spray. P271 Use only outdoors or in a well-ventilated area. P273 Avoid release to the environment. P280 Wear protective gloves/protective clothing/eye protection/face protection. P303+P361+P353 IF ON SKIN (or hair): Take off immediately all contaminated clothing. Rinse skin with water. P305+P351+P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.
ampicillin		H317 May cause an allergic skin reaction. H334 May cause allergy or asthma symptoms or breathing difficulties if inhaled	P261 Avoid breathing dust. P272 Contaminated work clothing should not be allowed out of the workplace. P280 Wear protective gloves. P284 Wear respiratory protection. P302 + P352 IF ON SKIN: Wash with plenty of water. P333 + P313 If skin irritation or rash occurs: Get medical advice/ attention
anhydrotetracycline		H302 Harmful if swallowed. H360D May damage the unborn child	P202 Do not handle until all safety precautions have been read and understood. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection. P308 + P313 IF exposed or concerned: Get medical advice/ attention. P405 Store locked up
Blauer Jonas	This substance or mixture has not been classified as hazardous according to the Globally Harmonized System (GHS) of Classification and Labeling of Chemicals.		

bromophenol blue (115-39-9)	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
cesium iodide (7789-17-5)		H361fd Suspected of damaging fertility. Suspected of damaging the unborn child. H400 Very toxic to aquatic life.	P201 Obtain special instructions before use. P202 Do not handle until all safety precautions have been read and understood. P273 Avoid release to the environment. P280 Wear protective gloves/protective clothing/eye protection/face protection. P308 + P313 IF exposed or concerned: Get medical advice/attention. P391 Collect spillage.
cOmplete™ EDTA-free Protease Inhibitor Cocktail		H314 Causes severe skin burns and eye damage	P260 Do not breathe dust. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection/ hearing protection. Response: P301 + P330 + P331 IF SWALLOWED: Rinse mouth. Do NOT induce vomiting. P303 + P361 + P353 IF ON SKIN (or hair): Take off immedi- ately all contaminated clothing. Rinse skin with water. P304 + P340 + P310 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Immediately call a POISON CENTER/ doctor. P305 + P351 + P338 + P310 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if pre- sent and easy to do. Continue rinsing. Immediately call a POISON CENTER/ doctor.
chloramphenicol		H318 Causes serious eye damage. H351 Suspected of causing cancer.	P202 Do not handle until all safety precautions have been read and understood. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection.

		H361fd Suspected of damaging fertility. Suspected of damaging the unborn child	P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing. P308 + P313 IF exposed or concerned: Get medical advice/ attention. P405 Store locked up. P501 Dispose of contents/ container to an approved waste disposal plant.
dithiothreitol, DTT (3483-12-3)		H302 Harmful if swallowed. H315 Causes skin irritation. H318 Causes serious eye damage.	P264 Wash skin thoroughly after handling. P270 Do not eat, drink or smoke when using this product. P280 Wear protective gloves/eye protection/face protection. P301 + P312 IF SWALLOWED: Call a POISON CENTER/doctor if you feel unwell. P302 + P352 IF ON SKIN: Wash with plenty of water. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.
ethanol		H225 Highly flammable liquid and vapor. H319 Causes serious eye irritation.	P210 Keep away from heat, hot surfaces, sparks, open flames and other ignition sources. No smoking. P233 Keep container tightly closed. P240 Ground and bond container and receiving equipment. P241 Use explosion-proof electrical/ ventilating/ lighting/ equipment. P242 Use non-sparking tools. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing

ethylenediaminetetraacetic acid, EDTA (60-00-4)		H319 Causes serious eye irritation.	P264 Wash skin thoroughly after handling. P280 Wear eye protection/face protection. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing. P337 + P313 If eye irritation persists: Get medical advice/attention.
glycerol (56-81-5)	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
guanidine hydrochloride (50-01-1)		H302 + H332 Harmful if swallowed or if inhaled. H315 Causes skin irritation. H319 Causes serious eye irritation	P261 Avoid breathing dust. P264 Wash skin thoroughly after handling. P301 + P312 IF SWALLOWED: Call a POISON CENTER/doctor if you feel unwell. P302 + P352 IF ON SKIN: Wash with plenty of water. P304 + P340 + P312 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Call a POISON CENTER/doctor if you feel unwell. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.
hydrochloric acid		H225 Highly flammable liquid and vapor. H319 Causes serious eye irritation.	P234 Keep only in original packaging. P261 Avoid breathing mist or vapors. P271 Use only outdoors or in a well-ventilated area. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection. P303 + P361 + P353 IF ON SKIN (or hair): Take off immediately all contaminated

			clothing. Rinse skin with water. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing
Imidazole		H302 Harmful if swallowed. H314 Causes severe skin burns and eye damage. H360D May damage the unborn child	P260 Do not breathe dust. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection. P301 + P312 IF SWALLOWED: Call a POISON CENTER/ doctor if you feel unwell. P303 + P361 + P353 IF ON SKIN (or hair): Take off immediately all contaminated clothing. Rinse skin with water. P304 + P340 + P310 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Immediately call a POISON CENTER/ doctor. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing
isopropyl-β-D-1-thiogalactopyranosid, IPTG (367-93-1)		H319 Causes serious eye irritation. H351 Suspected of causing cancer.	P281 Use personal protective equipment as required. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing
kanamycin monosulfat		H360 May damage fertility or the unborn child.	P201 Obtain special instructions before use. P202 Do not handle until all safety precautions have been read and understood. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection. P308 + P313 IF exposed or concerned: Get medical advice/ attention. P405 Store locked up.

			P501 Dispose of contents/ container to an approved waste disposal plant.
phenylmethylsulfonyl fluoride, PMSF (329-98-6)		H301 Toxic if swallowed. H314 Causes severe skin burns and eye damage.	P260 Do not breathe dusts or mists. P270 Do not eat, drink or smoke when using this product. P280 Wear protective gloves/protective clothing/eye protection/face protection. P303 + P361 + P353 IF ON SKIN (or hair): Take off immediately all contaminated clothing. Rinse skin with water. P304 + P340 + P310 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Immediately call a POISON CENTER/doctor. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing
phosohate buffered saline	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
PureCube 100 Ni-INDIGO Agarose resin	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
sodium chloride, NaCl (7647-14-5)	Not a hazardous substance or mixture according to Regulation (EC) No 1272/2008.		
sodium dodecyl sulfate, SDS (151-21-3)		H228 Flammable solid. H302 + H332 Harmful if swallowed or if inhaled. H315 Causes skin irritation. H318 Causes serious eye damage.	P210 Keep away from heat, hot surfaces, sparks, open flames and other ignition sources. No smoking. P273 Avoid release to the environment. P280 Wear protective gloves/protective clothing/eye protection/face protection. P301 + P312 IF SWALLOWED: Call a POISON CENTER/doctor if you feel unwell. P304 + P340 + P312

		H335 May cause respiratory irritation. H412 Harmful to aquatic life with long lasting effects.	IF INHALED: Remove person to fresh air and keep comfortable for breathing. Call a POISON CENTER/doctor if you feel unwell. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.
sodium hydroxide		H290 May be corrosive to metals. H314 Causes severe skin burns and eye damage	P234 Keep only in original packaging. P260 Do not breathe dust. P280 Wear protective gloves/ protective clothing/ eye protection/ face protection. P303 + P361 + P353 IF ON SKIN (or hair): Take off immediately all contaminated clothing. Rinse skin with water. P304 + P340 + P310 IF INHALED: Remove person to fresh air and keep comfortable for breathing. Immediately call a POISON CENTER/ doctor. P305 + P351 + P338 IF IN EYES: Rinse cautiously with water for several minutes. Remove contact lenses, if present and easy to do. Continue rinsing.

7. References

1. Chen, K.-T. *et al.* SARS in Taiwan: an overview and lessons learned. *International Journal of Infectious Diseases* **9**, 77–85 (2005).
2. Cui, Y. *et al.* Air pollution and case fatality of SARS in the People's Republic of China: an ecologic study. *Environ Health* **2**, 15 (2003).
3. Drosten, C. *et al.* Identification of a Novel Coronavirus in Patients with Severe Acute Respiratory Syndrome. *N Engl J Med* **348**, 1967–1976 (2003).
4. Van Der Hoek, L. Human Coronaviruses: What Do They Cause? *Antiviral Therapy* **12**, 651–658 (2007).
5. Wang, M. *et al.* SARS-CoV Infection in a Restaurant from Palm Civet. *Emerg. Infect. Dis.* **11**, 1860–1865 (2005).
6. Zaki, A. M., Van Boheemen, S., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Isolation of a Novel Coronavirus from a Man with Pneumonia in Saudi Arabia. *N Engl J Med* **367**, 1814–1820 (2012).
7. Baharoon, S. & Memish, Z. A. MERS-CoV as an emerging respiratory illness: A review of prevention methods. *Travel Medicine and Infectious Disease* **32**, 101520 (2019).
8. Sabir, J. S. M. *et al.* Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. *Science* **351**, 81–84 (2016).
9. Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe* **27**, 325–328 (2020).
10. Zhou, P. *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270–273 (2020).
11. Organization, W. H. & others. WHO Director-General's opening remarks at the media briefing on COVID-19. *(No Title)* (2020).
12. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses *et al.* The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. *Nat Microbiol* **5**, 536–544 (2020).
13. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. *Nat Med* **26**, 450–452 (2020).
14. Alwine, J. C., Casadevall, A., Enquist, L. W., Goodrum, F. D. & Imperiale, M. J. A Critical Analysis of the Evidence for the SARS-CoV-2 Origin Hypotheses. *mBio* **14**, e00583-23 (2023).
15. Almazán, F. *et al.* Coronavirus reverse genetic systems: Infectious clones and replicons. *Virus Research* **189**, 262–270 (2014).
16. Cui, J., Li, F. & Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat Rev Microbiol* **17**, 181–192 (2019).
17. Banerjee, A., Doxey, A. C., Mossman, K. & Irving, A. T. Unraveling the Zoonotic Origin and Transmission of SARS-CoV-2. *Trends in Ecology & Evolution* **36**, 180–184 (2021).
18. Huang, X.-Y. *et al.* A pangolin-origin SARS-CoV-2-related coronavirus: infectivity, pathogenicity, and cross-protection by preexisting immunity. *Cell Discov* **9**, 59 (2023).
19. Corman, V. M., Muth, D., Niemeyer, D. & Drosten, C. Hosts and Sources of Endemic Human Coronaviruses. in *Advances in Virus Research* vol. 100 163–188 (Elsevier, 2018).
20. Van Der Hoek, L. *et al.* Croup Is Associated with the Novel Coronavirus NL63. *PLoS Med* **2**, e240 (2005).
21. The Pediatric Infectious Diseases Society Annual Awards, 2010. *Pediatric Infectious Disease Journal* **29**, 789–793 (2010).
22. Bailey, E. S., Fieldhouse, J. K., Choi, J. Y. & Gray, G. C. A Mini Review of the Zoonotic Threat Potential of Influenza Viruses, Coronaviruses, Adenoviruses, and Enteroviruses. *Front. Public Health* **6**, 104 (2018).
23. Zmasek, C. M., Lefkowitz, E. J., Niewiadomska, A. & Scheuermann, R. H. Genomic evolution of the Coronaviridae family. *Virology* **570**, 123–133 (2022).
24. Chang, C.-K. *et al.* Multiple Nucleic Acid Binding Sites and Intrinsic Disorder of Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein: Implications for Ribonucleocapsid Protein Packaging. *Journal of Virology* **83**, 2255–2264 (2009).
25. Masters, P. S. The Molecular Biology of Coronaviruses. in *Advances in Virus Research* vol. 66 193–292 (Academic Press, 2006).
26. Fung, T. S. & Liu, D. X. Human Coronavirus: Host-Pathogen Interaction. *Annual Review of Microbiology* **73**, 529–557 (2019).
27. Rak, A., Isakova-Sivak, I. & Rudenko, L. Overview of Nucleocapsid-Targeting Vaccines against COVID-19. *Vaccines* **11**, 1810 (2023).
28. Liu, D. X., Fung, T. S., Chong, K. K.-L., Shukla, A. & Hilgenfeld, R. Accessory proteins of SARS-CoV and other coronaviruses. *Antiviral Res* **109**, 97–109 (2014).

29. Fang, P., Fang, L., Zhang, H., Xia, S. & Xiao, S. Functions of Coronavirus Accessory Proteins: Overview of the State of the Art. *Viruses* **13**, 1139 (2021).
30. Stertz, S. *et al.* The intracellular sites of early replication and budding of SARS-coronavirus. *Virology* **361**, 304–315 (2007).
31. Pietilä, M. K., Hellström, K. & Ahola, T. Alphavirus polymerase and RNA replication. *Virus Research* **234**, 44–57 (2017).
32. Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J. & Snijder, E. J. Nidovirales: Evolving the largest RNA virus genome. *Virus Research* **117**, 17–37 (2006).
33. Zimmern, D. Evolution of RNA Viruses. in *RNA Genetics* (CRC Press, 1988).
34. Brierley, I., Digard, P. & Inglis, S. C. Characterization of an efficient coronavirus ribosomal frameshifting signal: Requirement for an RNA pseudoknot. *Cell* **57**, 537–547 (1989).
35. Plant, E. P. *et al.* A Three-Stemmed mRNA Pseudoknot in the SARS Coronavirus Frameshift Signal. *PLOS Biology* **3**, e172 (2005).
36. Irigoyen, N. *et al.* High-Resolution Analysis of Coronavirus Gene Expression by RNA Sequencing and Ribosome Profiling. *PLOS Pathogens* **12**, e1005473 (2016).
37. Baranov, P. V. *et al.* Programmed ribosomal frameshifting in decoding the SARS-CoV genome. *Virology* **332**, 498–510 (2005).
38. Marczinke, B., Hagervall, T. & Brierley, I. The Q-base of asparaginyl-tRNA is dispensable for efficient -1 ribosomal frameshifting in eukaryotes. *J Mol Biol* **295**, 179–191 (2000).
39. Herold, J. & Siddell, S. G. An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. *Nucleic Acids Res* **21**, 5838–5842 (1993).
40. Bredenbeek, P. J. *et al.* The primary structure and expression of the second open reading frame of the polymerase gene of the coronavirus MHV-A59; a highly conserved polymerase is expressed by an efficient ribosomal frameshifting mechanism. *Nucleic Acids Res* **18**, 1825–1832 (1990).
41. Snijder, E. J., Decroly, E. & Ziebuhr, J. Chapter Three - The Nonstructural Proteins Directing Coronavirus RNA Synthesis and Processing. in *Advances in Virus Research* (ed. Ziebuhr, J.) vol. 96 59–126 (Academic Press, 2016).
42. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R. & Hilgenfeld, R. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science* **300**, 1763–1767 (2003).
43. Zhao, Y. *et al.* Structural basis for replicase polyprotein cleavage and substrate specificity of main protease from SARS-CoV-2. *Proceedings of the National Academy of Sciences* **119**, e2117142119 (2022).
44. Harcourt, B. H. *et al.* Identification of Severe Acute Respiratory Syndrome Coronavirus Replicase Products and Characterization of Papain-Like Protease Activity. *Journal of Virology* **78**, 13600–13612 (2004).
45. Wojdyla, J. A. *et al.* Papain-Like Protease 1 from Transmissible Gastroenteritis Virus: Crystal Structure and Enzymatic Activity toward Viral and Cellular Substrates. *Journal of Virology* **84**, 10063–10073 (2010).
46. Klemm, T. *et al.* Mechanism and inhibition of the papain-like protease, PLpro, of SARS-CoV-2. *The EMBO Journal* **39**, e106275 (2020).
47. Li, M. *et al.* Structure of the multiple functional domains from coronavirus nonstructural protein 3. *Emerging Microbes & Infections* **10**, 66–80 (2021).
48. Chen, Y. *et al.* X-ray Structural and Functional Studies of the Three Tandemly Linked Domains of Non-structural Protein 3 (nsp3) from Murine Hepatitis Virus Reveal Conserved Functions*. *Journal of Biological Chemistry* **290**, 25293–25306 (2015).
49. Herold, J., Siddell, S. G. & Gorbalenya, A. E. A Human RNA Viral Cysteine Proteinase That Depends upon a Unique Zn2+-binding Finger Connecting the Two Domains of a Papain-like Fold *. *Journal of Biological Chemistry* **274**, 14918–14925 (1999).
50. Anand, K. *et al.* Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain. *The EMBO Journal* **21**, 3213–3224 (2002).
51. Schechter, I. & Berger, A. On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications* **27**, 157–162 (1967).
52. Zhu, L. *et al.* Peptide aldehyde inhibitors challenge the substrate specificity of the SARS-coronavirus main protease. *Antiviral Research* **92**, 204–212 (2011).
53. Rut, W. *et al.* SARS-CoV-2 Mpro inhibitors and activity-based probes for patient-sample imaging. *Nat Chem Biol* **17**, 222–228 (2021).
54. Noske, G. D. *et al.* A Crystallographic Snapshot of SARS-CoV-2 Main Protease Maturation Process. *J Mol Biol* **433**, 167118 (2021).
55. Chen, S., Jonas, F., Shen, C. & Hilgenfeld, R. Liberation of SARS-CoV main protease from the viral polyprotein: N-terminal autocleavage does not depend on the mature dimerization mode. *Protein Cell* **1**, 59–74 (2010).

56. Li, C. *et al.* Maturation mechanism of severe acute respiratory syndrome (SARS) coronavirus 3C-like proteinase. *J Biol Chem* **285**, 28134–28140 (2010).
57. Hsu, M.-F. *et al.* Mechanism of the Maturation Process of SARS-CoV 3CL Protease*[boxs]. *Journal of Biological Chemistry* **280**, 31257–31266 (2005).
58. Lin, C.-W. *et al.* Characterization of trans- and cis-cleavage activity of the SARS coronavirus 3CLpro protease: basis for the in vitro screening of anti-SARS drugs. *FEBS Letters* **574**, 131–137 (2004).
59. Cao, D. *et al.* The SARS-CoV-2 papain-like protease suppresses type I interferon responses by deubiquitinating STING. *Sci Signal* **16**, eadd0082 (2023).
60. van Huizen, M. *et al.* The Main Protease of Middle East Respiratory Syndrome Coronavirus Induces Cleavage of Mitochondrial Antiviral Signaling Protein to Antagonize the Innate Immune Response. *Viruses* **16**, 256 (2024).
61. Martiáñez-Vendrell, X., van Kasteren, P. B., Myeni, S. K. & Kikkert, M. HCoV-229E Mpro Suppresses RLR-Mediated Innate Immune Signalling Through Cleavage of NEMO and Through Other Mechanisms. *International Journal of Molecular Sciences* **26**, 1197 (2025).
62. Luo, S. Y. *et al.* Identification of Human Host Substrates of the SARS-CoV-2 Mpro and PLpro Using Subtiligase N-Terminomics. *ACS Infect Dis* **9**, 749–761 (2023).
63. Woo, P. C. Y., Huang, Y., Lau, S. K. P. & Yuen, K.-Y. Coronavirus Genomics and Bioinformatics Analysis. *Viruses* **2**, 1804–1820 (2010).
64. Mishchenko, E. L. & V.A. Ivanisenko, V. A. Replication-transcription complex of coronaviruses: functions of individual viral non-structural subunits, properties and architecture of their complexes. *Vavilovskii Zhurnal Genet Selektsii* **26**, 121–127 (2022).
65. Schubert, K. *et al.* Universal features of Nsp1-mediated translational shutdown by coronaviruses. *Mol Cell* **83**, 3546-3557.e8 (2023).
66. Sergio, M. C., Ricciardi, S., Guarino, A. M., Giaquinto, L. & De Matteis, M. A. Membrane remodeling and trafficking piloted by SARS-CoV-2. *Trends in Cell Biology* **34**, 785–800 (2024).
67. Bills, C., Xie, X. & Shi, P.-Y. The multiple roles of nsp6 in the molecular pathogenesis of SARS-CoV-2. *Antiviral Res* **213**, 105590 (2023).
68. Subissi, L. *et al.* One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proceedings of the National Academy of Sciences* **111**, E3900–E3909 (2014).
69. Krichel, B. *et al.* Hallmarks of *Alpha-* and *Betacoronavirus* non-structural protein 7+8 complexes. *Sci. Adv.* **7**, eabf1004 (2021).
70. Lehmann, K. C. *et al.* Discovery of an essential nucleotidylating activity associated with a newly delineated conserved domain in the RNA polymerase-containing protein of all nidoviruses. *Nucleic Acids Research* **43**, 8416–8434 (2015).
71. Dwivedy, A. *et al.* Characterization of the NiRAN domain from RNA-dependent RNA polymerase provides insights into a potential therapeutic target against SARS-CoV-2. *PLoS Comput Biol* **17**, e1009384 (2021).
72. Slanina, H. *et al.* Coronavirus replication–transcription complex: Vital and selective NMPylation of a conserved site in nsp9 by the NiRAN-RdRp subunit. *Proceedings of the National Academy of Sciences* **118**, e2022310118 (2021).
73. Posthuma, C. C., te Velthuis, A. J. W. & Snijder, E. J. Nidovirus RNA polymerases: Complex enzymes handling exceptional RNA genomes. *Virus Research* **234**, 58–73 (2017).
74. Park, G. J. *et al.* The mechanism of RNA capping by SARS-CoV-2. *Res Sq* rs.3.rs-1336910 (2022) doi:10.21203/rs.3.rs-1336910/v1.
75. Lundrigan, E., Toudic, C., Pennock, E. & Pezacki, J. P. SARS-CoV-2 Protein Nsp9 Is Involved in Viral Evasion through Interactions with Innate Immune Pathways. *ACS Omega* **9**, 26428–26438 (2024).
76. Walker, A. P. *et al.* The SARS-CoV-2 RNA polymerase is a viral RNA capping enzyme. *Nucleic Acids Research* **49**, 13019–13030 (2021).
77. Sevajol, M., Subissi, L., Decroly, E., Canard, B. & Imbert, I. Insights into RNA synthesis, capping, and proofreading mechanisms of SARS-coronavirus. *Virus Research* **194**, 90–99 (2014).
78. Snijder, E. J. *et al.* Unique and Conserved Features of Genome and Proteome of SARS-coronavirus, an Early Split-off From the Coronavirus Group 2 Lineage. *Journal of Molecular Biology* **331**, 991–1004 (2003).
79. Feder, M., Pas, J., Wyrwicz, L. S. & Bujnicki, J. M. Molecular phylogenetics of the RrmJ/fibrillarin superfamily of ribose 2′-*O*-methyltransferases. *Gene* **302**, 129–138 (2003).
80. Neuman, B. W. *et al.* Giant RNA genomes: Roles of host, translation elongation, genome architecture, and proteome in nidoviruses. *Proceedings of the National Academy of Sciences* **122**, e2413675122 (2025).
81. Glasner, D. R. & Daugherty, M. D. Unchecked growth: Pushing the limits on RNA virus genome size in the absence of known proofreading. *Proceedings of the National Academy of Sciences* **121**, e2414223121 (2024).

82. Bouraï, M. *et al.* Mapping of Chikungunya Virus Interactions with Host Proteins Identified nsP2 as a Highly Connected Viral Component. *Journal of Virology* **86**, 3121–3134 (2012).
83. Yadav, R. *et al.* Biochemical and structural insights into SARS-CoV-2 polyprotein processing by Mpro. *Sci. Adv.* **8**, eadd2191 (2022).
84. Schamoni-Kast, K. *et al.* The kinetics of SARS-CoV-2 nsp7-11 polyprotein processing and impact on complexation with nsp16. 2024.01.06.574466 Preprint at https://doi.org/10.1101/2024.01.06.574466 (2024).
85. Kenward, C., Vuckovic, M., Paetzel, M. & Strynadka, N. C. J. Kinetic comparison of all eleven viral polyprotein cleavage site processing events by SARS-CoV-2 main protease using a linked protein FRET platform. *Journal of Biological Chemistry* **300**, (2024).
86. Ziebuhr, J. & Siddell, S. G. Processing of the Human Coronavirus 229E Replicase Polyproteins by the Virus-Encoded 3C-Like Proteinase: Identification of Proteolytic Products and Cleavage Sites Common to pp1a and pp1ab. *J Virol* **73**, 177–185 (1999).
87. Prediction and biochemical analysis of putative cleavage sites of the 3C-like protease of Middle East respiratory syndrome coronavirus. *Virus Research* **208**, 56–65 (2015).
88. Herold, J., Raabe, T., Schelle-Prinz, B. & Siddell, S. G. Nucleotide Sequence of the Human Coronavirus 229E RNA Polymerase Locus. *Virology* **195**, 680–691 (1993).
89. Herald, J. & Siddell, S. G. An 'elaborated' pseudoknot is required for high frequency frameshifting during translation of HCV 229E polymerase mRNA. *Nucleic Acids Research* **21**, 5838–5842 (1993).
90. Hartenian, E. *et al.* The molecular virology of coronaviruses. *Journal of Biological Chemistry* **295**, 12910–12934 (2020).
91. Xu, H. Y., Lim, K. P., Shen, S. & Liu, D. X. Further Identification and Characterization of Novel Intermediate and Mature Cleavage Products Released from the ORF 1b Region of the Avian Coronavirus Infectious Bronchitis Virus 1a/1b Polyprotein. *Virology* **288**, 212–222 (2001).
92. Smith, L. M. & Kelleher, N. L. Proteoform: a single term describing protein complexity. *Nat Methods* **10**, 186–187 (2013).
93. Schamoni-Kast, K. & Uetrecht, C. From Science to Fiction – Where are We Connecting in Vivo and in Vitro Results in Polyprotein Processing? SSRN Scholarly Paper at https://doi.org/10.2139/ssrn.5144731 (2025).
94. Denison, M. R. *et al.* Intracellular processing of the N-terminal ORF 1a proteins of the coronavirus MHV-A59 requires multiple proteolytic events. *Virology* **189**, 274–284 (1992).
95. Lim, K. P., Ng, L. F. P. & Liu, D. X. Identification of a Novel Cleavage Activity of the First Papain-Like Proteinase Domain Encoded by Open Reading Frame 1a of the Coronavirus Avian Infectious Bronchitis Virus and Characterization of the Cleavage Products. *Journal of Virology* **74**, 1674–1685 (2000).
96. Liu, D. X., Xu, H. Y. & Brown, T. D. Proteolytic processing of the coronavirus infectious bronchitis virus 1a polyprotein: identification of a 10-kilodalton polypeptide and determination of its cleavage sites. *J Virol* **71**, 1814–1820 (1997).
97. Heusipp, G., Grötzinger, C., Herold, J., Siddell, S. G. & Ziebuhr, J. Identification and subcellular localization of a 41 kDa, polyprotein 1ab processing product in human coronavirus 229E-infected cells. *Journal of General Virology* **78**, 2789–2794 (1997).
98. Schiller, J. J., Kanjanahaluethai, A. & Baker, S. C. Processing of the Coronavirus MHV-JHM Polymerase Polyprotein: Identification of Precursors and Proteolytic Products Spanning 400 Kilodaltons of ORF1a. *Virology* **242**, 288–302 (1998).
99. Piñón, J. D. *et al.* Efficient Autoproteolytic Processing of the MHV-A59 3C-like Proteinase from the Flanking Hydrophobic Domains Requires Membranes. *Virology* **230**, 309–322 (1997).
100. Baker, S. C. *et al.* Identification of a domain required for autoproteolytic cleavage of murine coronavirus gene A polyprotein. *Journal of Virology* **63**, 3693–3699 (1989).
101. Bost, A. G., Carnahan, R. H., Lu, X. T. & Denison, M. R. Four Proteins Processed from the Replicase Gene Polyprotein of Mouse Hepatitis Virus Colocalize in the Cell Periphery and Adjacent to Sites of Virion Assembly. *Journal of Virology* **74**, 3379–3387 (2000).
102. Herold, J., Gorbalenya, A. E., Thiel, V., Schelle, B. & Siddell, S. G. Proteolytic Processing at the Amino Terminus of Human Coronavirus 229E Gene 1-Encoded Polyproteins: Identification of a Papain-Like Proteinase and Its Substrate. *Journal of Virology* **72**, 910–918 (1998).
103. Prentice, E., McAuliffe, J., Lu, X., Subbarao, K. & Denison, M. R. Identification and Characterization of Severe Acute Respiratory Syndrome Coronavirus Replicase Proteins. *Journal of Virology* **78**, 9977–9986 (2004).
104. Lee, H.-J. *et al.* The complete sequence (22 kilobases) of murine coronavirus gene 1 encoding the putative proteases and RNA polymerase. *Virology* **180**, 567–582 (1991).
105. Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. Coronavirus genome: prediction of putative functional domains in the non-structural polyprotein by comparative amino acid sequence analysis. *Nucleic Acids Research* **17**, 4847–4861 (1989).

106. Tibbles, K. W., Brierley, I., Cavanagh, D. & Brown, T. D. Characterization in vitro of an autocatalytic processing activity associated with the predicted 3C-like proteinase domain of the coronavirus avian infectious bronchitis virus. *Journal of Virology* **70**, 1923–1930 (1996).
107. Lu, Y., Lu, X. & Denison, M. R. Identification and characterization of a serine-like proteinase of the murine coronavirus MHV-A59. *Journal of Virology* **69**, 3554–3559 (1995).
108. Zmasek, C. M., Lefkowitz, E. J., Niewiadomska, A. & Scheuermann, R. H. Genomic evolution of the *Coronaviridae* family. *Virology* **570**, 123–133 (2022).
109. Lauber, C. *et al.* Deep mining of the Sequence Read Archive reveals major genetic innovations in coronaviruses and other nidoviruses of aquatic vertebrates. *PLOS Pathogens* **20**, e1012163 (2024).
110. Bonilla, P. J., Hughes, S. A. & Weiss, S. R. Characterization of a second cleavage site and demonstration of activity in trans by the papain-like proteinase of the murine coronavirus mouse hepatitis virus strain A59. *Journal of Virology* **71**, 900–909 (1997).
111. Fan, K. *et al.* Biosynthesis, Purification, and Substrate Specificity of Severe Acute Respiratory Syndrome Coronavirus 3C-like Proteinase. *Journal of Biological Chemistry* **279**, 1637–1642 (2004).
112. Deming, D. J., Graham, R. L., Denison, M. R. & Baric, R. S. Processing of Open Reading Frame 1a Replicase Proteins nsp7 to nsp10 in Murine Hepatitis Virus Strain A59 Replication. *J Virol* **81**, 10280–10291 (2007).
113. Fan, K. *et al.* The substrate specificity of SARS coronavirus 3C-like proteinase. *Biochemical and Biophysical Research Communications* **329**, 934–940 (2005).
114. Krichel, B., Falke, S., Hilgenfeld, R., Redecke, L. & Uetrecht, C. Processing of the SARS-CoV pp1a/ab nsp7–10 region. *Biochemical Journal* **477**, 1009–1019 (2020).
115. Hegyi, A. & Ziebuhr, J. Conservation of substrate specificities among coronavirus main proteases. *Journal of General Virology* **83**, 595–599 (2002).
116. Kao, R. Y. *et al.* Characterization of SARS-CoV main protease and identification of biologically active small molecule inhibitors using a continuous fluorescence-based assay. *FEBS Letters* **576**, 325–330 (2004).
117. Blanchard, J. E. *et al.* High-Throughput Screening Identifies Inhibitors of the SARS Coronavirus Main Proteinase. *Chemistry & Biology* **11**, 1445–1453 (2004).
118. May, J., Korba, B., Medvedev, A. & Viswanathan, P. Enzyme kinetics of the human norovirus protease control virus polyprotein processing order. *Virology* **444**, 218–224 (2013).
119. Solowiej, J. *et al.* Steady-State and Pre-Steady-State Kinetic Evaluation of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) 3CLpro Cysteine Protease: Development of an Ion-Pair Model for Catalysis. *Biochemistry* **47**, 2617–2630 (2008).
120. Zhu, L. *et al.* Peptide aldehyde inhibitors challenge the substrate specificity of the SARS-coronavirus main protease. *Antiviral Research* **92**, 204–212 (2011).
121. Grum-Tokars, V., Ratia, K., Begaye, A., Baker, S. C. & Mesecar, A. D. Evaluating the 3C-like protease activity of SARS-Coronavirus: Recommendations for standardized assays for drug discovery. *Virus Research* **133**, 63–73 (2008).
122. Shin, G. *et al.* Structural and functional insights into alphavirus polyprotein processing and pathogenesis. *Proceedings of the National Academy of Sciences* **109**, 16534–16539 (2012).
123. Lee, J. *et al.* X-ray crystallographic characterization of the SARS-CoV-2 main protease polyprotein cleavage sites essential for viral processing and maturation. *Nat Commun* **13**, 5196 (2022).
124. Rappsilber, J. The beginning of a beautiful friendship: Cross-linking/mass spectrometry and modelling of proteins and multi-protein complexes. *Journal of Structural Biology* **173**, 530–540 (2011).
125. Swaim, C. L., Smith, J. B. & Smith, D. L. Unexpected products from the reaction of the synthetic cross-linker 3,3′-dithiobis(sulfosuccinimidyl propionate), DTSSP with peptides. *J Am Soc Mass Spectrom* **15**, 736–749 (2004).
126. Leavell, M. D., Novak, P., Behrens, C. R., Schoeniger, J. S. & Kruppa, G. H. Strategy for selective chemical cross-linking of tyrosine and lysine residues. *J Am Soc Mass Spectrom* **15**, 1604–1611 (2004).
127. Mendoza, V. L. & Vachet, R. W. Probing protein structure by amino acid-specific covalent labeling and mass spectrometry. *Mass Spectrom Rev* **28**, 785–815 (2009).
128. Puchades, C. *et al.* Epitope mapping of diverse influenza Hemagglutinin drug candidates using HDX-MS. *Sci Rep* **9**, 4735 (2019).
129. Shukla, A. K. *et al.* Visualization of arrestin recruitment by a G-protein-coupled receptor. *Nature* **512**, 218–222 (2014).
130. Wales, T. E. & Engen, J. R. Hydrogen exchange mass spectrometry for the analysis of protein dynamics. *Mass Spectrom Rev* **25**, 158–170 (2006).
131. Zhang, Q. *et al.* Epitope Mapping of a 95 kDa Antigen in Complex with Antibody by Solution-Phase Amide Backbone Hydrogen/Deuterium Exchange Monitored by Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Anal. Chem.* **83**, 7129–7136 (2011).

132. Renaud, J.-P. *et al.* Biophysics in drug discovery: impact, challenges and opportunities. *Nat Rev Drug Discov* **15**, 679–698 (2016).
133. Simanjuntak, Y., Schamoni-Kast, K., Grün, A., Uetrecht, C. & Scaturro, P. Top-Down and Bottom-Up Proteomics Methods to Study RNA Virus Biology. *Viruses* **13**, 668 (2021).
134. Yang, J. & Zhang, Y. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Research* **43**, W174–W181 (2015).
135. Ziebuhr, J. Coronavirus Replicative Proteins. in *Nidoviruses* 65–81 (John Wiley & Sons, Ltd, 2007). doi:10.1128/9781555815790.ch5.
136. Habersetzer, J. *et al. In vitro* translation of virally-encoded replication polyproteins to recapitulate polyprotein maturation processes. *Protein Expression and Purification* **175**, 105694 (2020).
137. Shoemaker, G. K. *et al.* Norwalk Virus Assembly and Stability Monitored by Mass Spectrometry *. *Molecular & Cellular Proteomics* **9**, 1742–1751 (2010).
138. Chorev, D. S. *et al.* Protein assemblies ejected directly from native membranes yield complexes for mass spectrometry. *Science* **362**, 829–834 (2018).
139. Garcia, N. K. *et al.* Structural dynamics reveal subtype-specific activation and inhibition of influenza virus hemagglutinin. *Journal of Biological Chemistry* **299**, (2023).
140. Bogdanow, B., Katsimani, N., Liu, F. & Selbach, M. Combining Metabolic Pulse Labeling and Quantitative Proteomics to Monitor Protein Synthesis Upon Viral Infection. in *Virus-Host Interactions* (ed. Aquino De Muro, M.) vol. 2610 149–165 (Springer US, New York, NY, 2023).
141. Sänger, L. *et al.* RNA to Rule Them All: Critical Steps in Lassa Virus Ribonucleoparticle Assembly and Recruitment. *J. Am. Chem. Soc.* **145**, 27958–27974 (2023).
142. Emmott, E., Sweeney, T. R. & Goodfellow, I. A Cell-based Fluorescence Resonance Energy Transfer (FRET) Sensor Reveals Inter- and Intragenogroup Variations in Norovirus Protease Activity and Polyprotein Cleavage *. *Journal of Biological Chemistry* **290**, 27841–27853 (2015).
143. Emmott, E. *et al.* Polyprotein processing and intermolecular interactions within the viral replication complex spatially and temporally control norovirus protease activity. *Journal of Biological Chemistry* **294**, 4259–4271 (2019).
144. Khater, I. M., Nabi, I. R. & Hamarneh, G. A Review of Super-Resolution Single-Molecule Localization Microscopy Cluster Analysis and Quantification Methods. *PATTER* **1**, (2020).
145. Baharom, F. *et al.* Visualization of early influenza A virus trafficking in human dendritic cells using STED microscopy. *PLOS ONE* **12**, e0177920 (2017).
146. Eiring, P. *et al.* Coronaviruses Use ACE2 Monomers as Entry-Receptors. *Angewandte Chemie International Edition* **62**, e202300821 (2023).
147. Tugarinov, V., Muhandiram, R., Ayed, A. & Kay, L. E. Four-Dimensional NMR Spectroscopy of a 723-Residue Protein: Chemical Shift Assignments and Secondary Structure of Malate Synthase G. *J. Am. Chem. Soc.* **124**, 10025–10035 (2002).
148. Lyumkis, D. *et al.* Cryo-EM Structure of a Fully Glycosylated Soluble Cleaved HIV-1 Envelope Trimer. *Science* **342**, 1484–1490 (2013).
149. Saibil, H. R. Cryo-EM in molecular and cellular biology. *Molecular Cell* **82**, 274–284 (2022).
150. Wrapp, D. *et al.* Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science* **367**, 1260–1263 (2020).
151. Vollmer, B. *et al.* The prefusion structure of herpes simplex virus glycoprotein B. *Science Advances* **6**, eabc1726 (2020).
152. Narwal, M., Armache, J.-P., Edwards, T. J. & Murakami, K. S. SARS-CoV-2 polyprotein substrate regulates the stepwise Mpro cleavage reaction. *Journal of Biological Chemistry* **299**, 104697 (2023).
153. Perry, J. K. *et al.* An atomistic model of the coronavirus replication-transcription complex as a hexamer assembled around nsp15. *Journal of Biological Chemistry* **297**, 101218 (2021).
154. Gao, Y. *et al.* Structure of the RNA-dependent RNA polymerase from COVID-19 virus. *Science* **368**, 779–782 (2020).
155. Yan, L. *et al.* Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis. *Cell* **184**, 184-193.e10 (2021).
156. Chen, J. *et al.* Ensemble cryo-EM reveals conformational states of the nsp13 helicase in the SARS-CoV-2 helicase replication–transcription complex. *Nat Struct Mol Biol* **29**, 250–260 (2022).
157. Matsuda, A. *et al.* Despite the odds: formation of the SARS-CoV-2 methylation complex. *Nucleic Acids Research* **52**, 6441–6458 (2024).
158. Klein, S. *et al.* SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography. *Nat Commun* **11**, 5885 (2020).

159. Bost, A. G., Carnahan, R. H., Lu, X. T. & Denison, M. R. Four Proteins Processed from the Replicase Gene Polyprotein of Mouse Hepatitis Virus Colocalize in the Cell Periphery and Adjacent to Sites of Virion Assembly. *J Virol* **74**, 3379–3387 (2000).
160. Gosert, R., Kanjanahaluethai, A., Egger, D., Bienz, K. & Baker, S. C. RNA Replication of Mouse Hepatitis Virus Takes Place at Double-Membrane Vesicles. *Journal of Virology* **76**, 3697–3708 (2002).
161. Knoops, K. *et al.* SARS-Coronavirus Replication Is Supported by a Reticulovesicular Network of Modified Endoplasmic Reticulum. *PLOS Biology* **6**, e226 (2008).
162. Oudshoorn, D. *et al.* Expression and Cleavage of Middle East Respiratory Syndrome Coronavirus nsp3-4 Polyprotein Induce the Formation of Double-Membrane Vesicles That Mimic Those Associated with Coronaviral RNA Replication. *mBio* **8**, 10.1128/mbio.01658-17 (2017).
163. Wolff, G. *et al.* A molecular pore spans the double membrane of the coronavirus replication organelle. *Science* **369**, 1395–1398 (2020).
164. Stukalov, A. *et al.* Multilevel proteomics reveals host perturbations by SARS-CoV-2 and SARS-CoV. *Nature* **594**, 246–252 (2021).
165. Aggarwal, S., Acharjee, A., Mukherjee, A., Baker, M. S. & Srivastava, S. Role of Multiomics Data to Understand Host–Pathogen Interactions in COVID-19 Pathogenesis. *J. Proteome Res.* **20**, 1107–1132 (2021).
166. Emmott, E. *et al.* Norovirus-Mediated Modification of the Translational Landscape via Virus and Host-Induced Cleavage of Translation Initiation Factors *. *Molecular & Cellular Proteomics* **16**, S215–S229 (2017).
167. Kaushal, P. & Lee, C. N-terminomics – its past and recent advancements. *Journal of Proteomics* **233**, 104089 (2021).
168. Wilken, L., Rimmelzwaan, G. F. & Elbahesh, H. The Raf kinase inhibitors Dabrafenib and Regorafenib impair Zika virus replication via distinct mechanisms. *J Virol* **98**, e0061824 (2024).
169. Kim, D.-K. *et al.* A proteome-scale map of the SARS-CoV-2–human contactome. *Nat Biotechnol* **41**, 140–149 (2023).
170. Steinbach, M. K., Leipert, J., Matzanke, T. & Tholey, A. Digital Microfluidics for Sample Preparation in Low-Input Proteomics. *Small Methods* **9**, 2400495 (2025).
171. Reubsaet, L. & Halvorsen, T. G. Advancements in clinical approaches, analytical methods, and smart sampling for LC–MS-based protein determination from dried matrix spots. *Journal of Separation Science* **47**, 2400061 (2024).
172. Schmidt, N. *et al.* The SARS-CoV-2 RNA–protein interactome in infected human cells. *Nat Microbiol* **6**, 339–353 (2021).
173. Peng, Y. *et al.* Top-down Proteomics Reveals Concerted Reductions in Myofilament and Z-disc Protein Phosphorylation after Acute Myocardial Infarction. *Mol Cell Proteomics* **13**, 2752–2764 (2014).
174. Melby, J. A. *et al.* High sensitivity top–down proteomics captures single muscle cell heterogeneity in large proteoforms. *Proceedings of the National Academy of Sciences* **120**, e2222081120 (2023).
175. S. Roberts, D. *et al.* Distinct core glycan and O -glycoform utilization of SARS-CoV-2 Omicron variant Spike protein RBD revealed by top-down mass spectrometry. *Chemical Science* **13**, 10944–10949 (2022).
176. Leney, A. C. & Heck, A. J. R. Native Mass Spectrometry: What is in the Name? *J. Am. Soc. Mass Spectrom.* **28**, 5–13 (2017).
177. Gan, J. *et al.* Native Mass Spectrometry of Recombinant Proteins from Crude Cell Lysates. *Anal. Chem.* **89**, 4398–4404 (2017).
178. Konermann, L. Addressing a Common Misconception: Ammonium Acetate as Neutral pH “Buffer” for Native Electrospray Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **28**, 1827–1835 (2017).
179. Susa, A. C., Xia, Z. & Williams, E. R. Small Emitter Tips for Native Mass Spectrometry of Proteins and Protein Complexes from Nonvolatile Buffers That Mimic the Intracellular Environment. *Anal. Chem.* **89**, 3116–3122 (2017).
180. Hopper, J. T. S. *et al.* Detergent-free mass spectrometry of membrane protein complexes. *Nat Methods* **10**, 1206–1208 (2013).
181. Laganowsky, A., Reading, E., Hopper, J. T. S. & Robinson, C. V. Mass spectrometry of intact membrane protein complexes. *Nat Protoc* **8**, 639–651 (2013).
182. Barrera, N. P., Di Bartolo, N., Booth, P. J. & Robinson, C. V. Micelles protect membrane complexes from solution to vacuum. *Science* **321**, 243–246 (2008).
183. Hedges, J. B., Vahidi, S., Yue, X. & Konermann, L. Effects of Ammonium Bicarbonate on the Electrospray Mass Spectra of Proteins: Evidence for Bubble-Induced Unfolding. *Anal. Chem.* **85**, 6469–6476 (2013).
184. Marchand, A. & Gabelica, V. Native Electrospray Mass Spectrometry of DNA G-Quadruplexes in Potassium Solution. *J. Am. Soc. Mass Spectrom.* **25**, 1146–1154 (2014).
185. Pagel, K., Hyung, S.-J., Ruotolo, B. T. & Robinson, C. V. Alternate Dissociation Pathways Identified in Charge-Reduced Protein Complex Ions. *Anal. Chem.* **82**, 5363–5372 (2010).

186. Zhou, M., Dagan, S. & Wysocki, V. H. Protein Subunits Released by Surface Collisions of Noncovalent Complexes: Nativelike Compact Structures Revealed by Ion Mobility Mass Spectrometry. *Angew Chem Int Ed* **51**, 4336–4339 (2012).
187. Lemaire, D., Marie, G., Serani, L. & Laprévote, O. Stabilization of Gas-Phase Noncovalent Macromolecular Complexes in Electrospray Mass Spectrometry Using Aqueous Triethylammonium Bicarbonate Buffer. *Anal. Chem.* **73**, 1699–1706 (2001).
188. Zhuang, X., Gavriilidou, A. F. M. & Zenobi, R. Influence of Alkylammonium Acetate Buffers on Protein–Ligand Noncovalent Interactions Using Native Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **28**, 341–346 (2017).
189. Fenn, J. B. Electrospray Wings for Molecular Elephants (Nobel Lecture). *Angewandte Chemie International Edition* **42**, 3871–3894 (2003).
190. Heck, A. J. R. & van den Heuvel, R. H. H. Investigation of intact protein complexes by mass spectrometry. *Mass Spectrometry Reviews* **23**, 368–389 (2004).
191. Karas, Michael. & Hillenkamp, Franz. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Anal. Chem.* **60**, 2299–2301 (1988).
192. Karas, M., Bahr, U. & Dülcks, T. Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. *Fresenius J Anal Chem* **366**, 669–676 (2000).
193. Konermann, L., Ahadi, E., Rodriguez, A. D. & Vahidi, S. Unraveling the Mechanism of Electrospray Ionization. *Anal. Chem.* **85**, 2–9 (2013).
194. McAllister, R. G., Metwally, H., Sun, Y. & Konermann, L. Release of Native-like Gaseous Proteins from Electrospray Droplets via the Charged Residue Mechanism: Insights from Molecular Dynamics Simulations. *J. Am. Chem. Soc.* **137**, 12667–12676 (2015).
195. F.R.S, L. R. XX. On the equilibrium of liquid conducting masses charged with electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* (1882) doi:10.1080/14786448208628425.
196. Cech, N. B. & Enke, C. G. Practical implications of some recent studies in electrospray ionization fundamentals. *Mass Spectrometry Reviews* **20**, 362–387 (2001).
197. Meyer, T., Gabelica, V., Grubmüller, H. & Orozco, M. Proteins in the gas phase. *WIREs Computational Molecular Science* **3**, 408–425 (2013).
198. Loo, J. A. Studying noncovalent protein complexes by electrospray ionization mass spectrometry. *Mass Spectrometry Reviews* **16**, 1–23 (1997).
199. Light-Wahl, K. J., Schwartz, B. L. & Smith, R. D. Observation of the noncovalent quaternary associations of proteins by electrospray ionization mass spectrometry. *Journal of the American Chemical Society* **116**, 5271–5278 (1994).
200. Schwartz, B. L., Light-Wahl, K. J. & Smith, R. D. Observation of noncovalent complexes to the avidin tetramer by electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry* **5**, 201–204 (1994).
201. Skinner, O. S., McLafferty, F. W. & Breuker, K. How Ubiquitin Unfolds after Transfer into the Gas Phase. *J. Am. Soc. Mass Spectrom.* **23**, 1011–1014 (2012).
202. Wyttenbach, T. & Bowers, M. T. Structural Stability from Solution to the Gas Phase: Native Solution Structure of Ubiquitin Survives Analysis in a Solvent-Free Ion Mobility–Mass Spectrometry Environment. *J. Phys. Chem. B* **115**, 12266–12275 (2011).
203. Hogan, C. J., Ruotolo, B. T., Robinson, C. V. & Fernandez De La Mora, J. Tandem Differential Mobility Analysis-Mass Spectrometry Reveals Partial Gas-Phase Collapse of the GroEL Complex. *J. Phys. Chem. B* **115**, 3614–3621 (2011).
204. Shi, L. *et al.* Characterizing Intermediates Along the Transition from Polyproline I to Polyproline II Using Ion Mobility Spectrometry-Mass Spectrometry. *J. Am. Chem. Soc.* **136**, 12702–12711 (2014).
205. Breuker, K. & McLafferty, F. W. Stepwise evolution of protein native structure with electrospray into the gas phase, 10^{-12} to 10^{2} s. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 18145–18152 (2008).
206. Fernandez De La Mora, J. Electrospray ionization of large multiply charged species proceeds via Dole's charged residue mechanism. *Analytica Chimica Acta* **406**, 93–104 (2000).
207. Kaltashov, I. A. & Mohimen, A. Estimates of Protein Surface Areas in Solution by Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **77**, 5370–5379 (2005).
208. Dobo, A. & Kaltashov, I. A. Detection of Multiple Protein Conformational Ensembles in Solution via Deconvolution of Charge-State Distributions in ESI MS. *Anal. Chem.* **73**, 4763–4773 (2001).
209. Gross, J. H. *Mass Spectrometry*. (Springer International Publishing, Cham, 2017). doi:10.1007/978-3-319-54398-7.
210. Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).

211. Marty, M. T. A Universal Score for Deconvolution of Intact Protein and Native Electrospray Mass Spectra. *Anal. Chem.* **92**, 4395–4401 (2020).
212. Rosati, S. *et al.* Qualitative and Semiquantitative Analysis of Composite Mixtures of Antibodies by Native Mass Spectrometry. *Anal. Chem.* **84**, 7227–7232 (2012).
213. Lorenzen, K. & Duijn, E. van. Native Mass Spectrometry as a Tool in Structural Biology. *Current Protocols in Protein Science* **62**, 17.12.1-17.12.17 (2010).
214. Barylyuk, K., Gülbakan, B., Xie, X. & Zenobi, R. DNA Oligonucleotides: A Model System with Tunable Binding Strength to Study Monomer–Dimer Equilibria with Electrospray Ionization-Mass Spectrometry. *Anal. Chem.* **85**, 11902–11912 (2013).
215. Bennett, J. L., Nguyen, G. T. H. & Donald, W. A. Protein–Small Molecule Interactions in Native Mass Spectrometry. *Chem. Rev.* **122**, 7327–7385 (2022).
216. Glish, G. L. & Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nat Rev Drug Discov* **2**, 140–150 (2003).
217. Faull, K. F. *et al.* Chapter 1 An Introduction to the Basic Principles and Concepts of Mass Spectrometry. in *Comprehensive Analytical Chemistry* (ed. Whitelegge, J. P.) vol. 52 1–46 (Elsevier, 2008).
218. Li, C. *et al.* Towards Higher Sensitivity of Mass Spectrometry: A Perspective From the Mass Analyzers. *Front. Chem.* **9**, (2021).
219. Boesl, U. Time-of-flight mass spectrometry: Introduction to the basics: TIME-OF-FLIGHT MASS SPECTROMETRY. *Mass Spec Rev* **36**, 86–109 (2017).
220. Xian, F., Hendrickson, C. L. & Marshall, A. G. High Resolution Mass Spectrometry. *Anal. Chem.* **84**, 708–719 (2012).
221. Hu, Q. *et al.* The Orbitrap: a new mass spectrometer. *J. Mass Spectrom.* **40**, 430–443 (2005).
222. Greaves, J. & Roboz, J. *Mass Spectrometry for the Novice*. (CRC Press, 2013). doi:10.1201/b15436.
223. Nolting, D., Malek, R. & Makarov, A. Ion traps in modern mass spectrometry. *Mass Spectrometry Reviews* **38**, 150–168 (2019).
224. Makarov, A. Electrostatic Axially Harmonic Orbital Trapping: A High-Performance Technique of Mass Analysis. *Anal. Chem.* **72**, 1156–1162 (2000).
225. Zubarev, R. A. & Makarov, A. Orbitrap Mass Spectrometry. *Anal. Chem.* **85**, 5288–5296 (2013).
226. Makarov, A., Denisov, E., Lange, O. & Horning, S. Dynamic range of mass accuracy in LTQ orbitrap hybrid mass spectrometer. *J. Am. Soc. Mass Spectrom.* **17**, 977–982 (2006).
227. Wörner, T. P. *et al.* Frequency chasing of individual megadalton ions in an Orbitrap analyser improves precision of analysis in single-molecule mass spectrometry. *Nat. Chem.* **14**, 515–522 (2022).
228. Wörner, T. P. *et al.* Resolving heterogeneous macromolecular assemblies by Orbitrap-based single-particle charge detection mass spectrometry. *Nat Methods* **17**, 395–398 (2020).
229. Fort, K. L. *et al.* Expanding the structural analysis capabilities on an Orbitrap-based mass spectrometer for large macromolecular complexes. *Analyst* **143**, 100–105 (2018).
230. Van De Waterbeemd, M. *et al.* High-fidelity mass analysis unveils heterogeneity in intact ribosomal particles. *Nat Methods* **14**, 283–286 (2017).
231. Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. R. High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nat Methods* **9**, 1084–1086 (2012).
232. Aquilina, J. A., Benesch, J. L. P., Bateman, O. A., Slingsby, C. & Robinson, C. V. Polydispersity of a mammalian chaperone: Mass spectrometry reveals the population of oligomers in αB-crystallin. *Proceedings of the National Academy of Sciences* **100**, 10611–10616 (2003).
233. Olsen, J. V. *et al.* Higher-energy C-trap dissociation for peptide modification analysis. *Nat Methods* **4**, 709–712 (2007).
234. Brodbelt, J. S. Ion Activation Methods for Peptides and Proteins. *Anal. Chem.* **88**, 30–51 (2016).
235. Whittle, P. J. & Blundell, T. L. Protein structure--based drug design. *Annu Rev Biophys Biomol Struct* **23**, 349–375 (1994).
236. Pavlopoulou, A. & Michalopoulos, I. State-of-the-art bioinformatics protein structure prediction tools (Review). *Int J Mol Med* **28**, 295–310 (2011).
237. Torda, A. E. Protein Threading. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 921–938 (Humana Press, Totowa, NJ, 2005). doi:10.1385/1-59259-890-0:921.
238. Hardin, C., Pogorelov, T. V. & Luthey-Schulten, Z. *Ab initio* protein structure prediction. *Current Opinion in Structural Biology* **12**, 176–181 (2002).
239. Powell, H. R., Islam, S. A., David, A. & Sternberg, M. J. E. Phyre2.2: A Community Resource for Template-based Protein Structure Prediction. *Journal of Molecular Biology* 168960 (2025) doi:10.1016/j.jmb.2025.168960.
240. Zhou, X. *et al.* I-TASSER-MTD: a deep-learning-based platform for multi-domain protein structure and function prediction. *Nat Protoc* **17**, 2326–2353 (2022).

241. Roy, A., Kucukural, A. & Zhang, Y. I-TASSER: a unified platform for automated protein structure and function prediction. *Nat Protoc* **5**, 725–738 (2010).
242. Du, Z. *et al.* The trRosetta server for fast and accurate protein structure prediction. *Nat Protoc* **16**, 5634–5651 (2021).
243. Baek, M. *et al.* Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
244. Wang, L. *et al.* Overview of AlphaFold2 and breakthroughs in overcoming its limitations. *Computers in Biology and Medicine* **176**, 108620 (2024).
245. Wu, R. *et al.* High-resolution de novo structure prediction from primary sequence. 2022.07.21.500999 Preprint at https://doi.org/10.1101/2022.07.21.500999 (2022).
246. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
247. Callaway, E. Major AlphaFold upgrade offers boost for drug discovery. *Nature* **629**, 509–510 (2024).
248. Varadi, M. *et al.* AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).
249. Porta-Pardo, E., Ruiz-Serra, V., Valentini, S. & Valencia, A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol* **18**, e1009818 (2022).
250. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
251. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
252. Roy, R. & Al-Hashimi, H. M. AlphaFold3 takes a step toward decoding molecular behavior and biological computation. *Nat Struct Mol Biol* **31**, 997–1000 (2024).
253. Akdel, M. *et al.* A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol* **29**, 1056–1067 (2022).
254. Marcu, Ş.-B., Tăbîrcă, S. & Tangney, M. An Overview of Alphafold's Breakthrough. *Front. Artif. Intell.* **5**, (2022).
255. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
256. Litvin, U. *et al.* Viro3D: a comprehensive database of virus protein structure predictions. 2024.12.19.629443 Preprint at https://doi.org/10.1101/2024.12.19.629443 (2024).
257. Jahirul Islam, Md., Nawal Islam, N., Siddik Alom, Md., Kabir, M. & Halim, M. A. A review on structural, non-structural, and accessory proteins of SARS-CoV-2: Highlighting drug target sites. *Immunobiology* **228**, 152302 (2023).
258. Krichel, B., Falke, S., Hilgenfeld, R., Redecke, L. & Uetrecht, C. Processing of the SARS-CoV pp1a/ab nsp7–10 region. *Biochemical Journal* **477**, 1009–1019 (2020).
259. Xue, X. *et al.* Production of Authentic SARS-CoV Mpro with Enhanced Activity: Application as a Novel Tag-cleavage Endopeptidase for Protein Overproduction. *Journal of Molecular Biology* **366**, 965–975 (2007).
260. Rambo, R. P. Considerations for Sample Preparation Using Size-Exclusion Chromatography for Home and Synchrotron Sources. in *Biological Small Angle Scattering: Techniques, Strategies and Tips* (eds. Chaudhuri, B., Muñoz, I. G., Qian, S. & Urban, V. S.) 31–45 (Springer, Singapore, 2017). doi:10.1007/978-981-10-6038-0_3.
261. Konermann, L. & Douglas, D. J. Acid-Induced Unfolding of Cytochrome c at Different Methanol Concentrations: Electrospray Ionization Mass Spectrometry Specifically Monitors Changes in the Tertiary Structure. *Biochemistry* **36**, 12296–12302 (1997).
262. Kaltashov, I. A. & Abzalimov, R. R. Do Ionic Charges in ESI MS Provide Useful Information on Macromolecular Structure? *Journal of the American Society for Mass Spectrometry* **19**, 1239–1246 (2008).
263. Kaltashov, I. A. & Mohimen, A. Estimates of Protein Surface Areas in Solution by Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **77**, 5370–5379 (2005).
264. Iavarone, A. T. & Williams, E. R. Mechanism of Charging and Supercharging Molecules in Electrospray Ionization. *J. Am. Chem. Soc.* **125**, 2319–2327 (2003).
265. Pablos, I. *et al.* Mechanistic insights into COVID-19 by global analysis of the SARS-CoV-2 3CLpro substrate degradome. *Cell Reports* **37**, (2021).
266. Cesar Ramos de Jesus, H. *et al.* Optimization of quenched fluorescent peptide substrates of SARS-CoV-2 3CLpro main protease (Mpro) from proteomic identification of P6—P6' active site specificity. *J Virol* **98**, e00049-24.
267. Sugasawa, N. *et al.* Formation of c- and z-ions due to preferential cleavage at the NC bond of Xxx-Asp/Asn residues in negative-ion CID of peptides. *International Journal of Mass Spectrometry* **383–384**, 38–43 (2015).
268. Bashyal, A., Sanders, J. D., Holden, D. D. & Brodbelt, J. S. Top-Down Analysis of Proteins in Low Charge States. *J Am Soc Mass Spectrom* **30**, 704–717 (2019).

269. Reid, D. J. *et al.* MetaUniDec: High-throughput Deconvolution of Native Mass Spectra. *J Am Soc Mass Spectrom* **30**, 118–127 (2019).
270. Barth, M. & Schmidt, C. Native mass spectrometry-A valuable tool in structural biology. *J Mass Spectrom* **55**, e4578 (2020).
271. Mann, M., Meng, C. K. & Fenn, J. B. Interpreting mass spectra of multiply charged ions. *Analytical Chemistry* **61**, 1702–1708 (1989).
272. Cornish-Bowden, A. *Fundamentals of Enzyme Kinetics*. (John Wiley & Sons, 2013).
273. Fryčák, P. & Schug, K. A. On-Line Dynamic Titration: Determination of Dissociation Constants for Noncovalent Complexes Using Gaussian Concentration Profiles by Electrospray Ionization Mass Spectrometry. *Anal. Chem.* **79**, 5407–5413 (2007).
274. Root, K., Wittwer, Y., Barylyuk, K., Anders, U. & Zenobi, R. Insight into Signal Response of Protein Ions in Native ESI-MS from the Analysis of Model Mixtures of Covalently Linked Protein Oligomers. *J. Am. Soc. Mass Spectrom.* **28**, 1863–1875 (2017).
275. Newville, M. *et al.* LMFIT: Non-Linear Least-Squares Minimization and Curve-Fitting for Python. Zenodo https://doi.org/10.5281/ZENODO.598352 (2025).
276. Tamara, S., den Boer, M. A. & Heck, A. J. R. High-Resolution Native Mass Spectrometry. *Chem. Rev.* **122**, 7269–7326 (2022).
277. Lu, J. *et al.* Improved Peak Detection and Deconvolution of Native Electrospray Mass Spectra from Large Protein Complexes. *J Am Soc Mass Spectrom* **26**, 2141–2151 (2015).
278. Nosti, A. J., Barrio, L. C., Calderón Celis, F., Soldado, A. & Encinar, J. R. Absolute quantification of proteins using element mass spectrometry and generic standards. *Journal of Proteomics* **256**, 104499 (2022).
279. Root, K., Wittwer, Y., Barylyuk, K., Anders, U. & Zenobi, R. Insight into Signal Response of Protein Ions in Native ESI-MS from the Analysis of Model Mixtures of Covalently Linked Protein Oligomers. *J. Am. Soc. Mass Spectrom.* **28**, 1863–1875 (2017).
280. Kebarle, P. & Verkerk, U. H. Electrospray: From ions in solution to ions in the gas phase, what we know now. *Mass Spectrometry Reviews* **28**, 898–917 (2009).
281. Bisswanger, H. *Enzyme Kinetics: Principles and Methods*. (John Wiley & Sons, 2017).
282. Miczi, M. *et al.* Identification of Host Cellular Protein Substrates of SARS-COV-2 Main Protease. *IJMS* **21**, 9523 (2020).
283. Bouvet, M. *et al.* Coronavirus Nsp10, a Critical Co-factor for Activation of Multiple Replicative Enzymes. *Journal of Biological Chemistry* **289**, 25783–25796 (2014).
284. Sawicki, S. G. *et al.* Functional and Genetic Analysis of Coronavirus Replicase-Transcriptase Proteins. *PLoS Pathog* **1**, e39 (2005).
285. Romano, M., Ruggiero, A., Squeglia, F., Maga, G. & Berisio, R. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* **9**, 1267 (2020).
286. Thibert, S. M. *et al.* Native Mass Spectrometry Dissects the Structural Dynamics of an Allosteric Heterodimer of SARS-CoV-2 Nonstructural Proteins. *Journal of the American Society for Mass Spectrometry* **35**, 912 (2024).
287. Zhou, M. *et al.* Higher-order structural characterisation of native proteins and complexes by top-down mass spectrometry. *Chemical Science* **11**, 12918–12936 (2020).
288. Pan, P. & McLuckey, S. A. The Effect of Small Cations on the Positive Electrospray Responses of Proteins at Low pH. *Anal. Chem.* **75**, 5468–5474 (2003).
289. Hernández, H. & Robinson, C. V. Determining the stoichiometry and interactions of macromolecular assemblies from mass spectrometry. *Nat Protoc* **2**, 715–726 (2007).
290. Radivoyevitch, T. Mass action models versus the Hill model: An analysis of tetrameric human thymidine kinase 1 positive cooperativity. *Biology Direct* **4**, 49 (2009).
291. Mitchell, A. L. *et al.* MGnify: the microbiome analysis resource in 2020. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gkz1035.
292. Mirdita, M. *et al.* ColabFold: Making Protein folding accessible to all. *Nature Methods* (2022) doi:10.1038/s41592-022-01488-1.
293. Mirdita, M., Steinegger, M. & S"oding, J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics* **35**, 2856–2858 (2019).
294. Mirdita, M. *et al.* Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.* **45**, D170–D176 (2017).
295. Biochemical and structural insights into SARS-CoV-2 polyprotein processing by Mpro | Science Advances. https://www.science.org/doi/10.1126/sciadv.add2191.
296. Garcia-Alai, M. M. *et al.* Epsin and Sla2 form assemblies through phospholipid interfaces. *Nat Commun* **9**, 328 (2018).

297. Krokidis, M. G. *et al.* AlphaFold3: An Overview of Applications and Performance Insights. *International Journal of Molecular Sciences* **26**, 3671 (2025).
298. Wee, J. & Wei, G.-W. Evaluation of AlphaFold 3's Protein-Protein Complexes for Predicting Binding Free Energy Changes upon Mutation. *J Chem Inf Model* **64**, 6676–6683 (2024).
299. Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* **450**, 964–972 (2007).
300. Guo, H.-B. *et al.* AlphaFold2 models indicate that protein sequence determines both structure and dynamics. *Sci Rep* **12**, 10696 (2022).
301. Zhang, S., Hou, Y. & Huang, Y. Inception of AlphaFold 3: Shining light from structure prediction to de novo design of biomolecules. *MedComm – Biomaterials and Applications* **3**, e102 (2024).
302. Chakravarty, D., Lee, M. & Porter, L. L. Proteins with alternative folds reveal blind spots in AlphaFold-based protein structure prediction. *Current Opinion in Structural Biology* **90**, 102973 (2025).
303. The UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**, D523–D531 (2023).
304. Xue, X. *et al.* Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. *J Virol* **82**, 2515–2527 (2008).
305. Shen, Y., Joachimiak, A., Rich Rosner, M. & Tang, W.-J. Structures of human insulin-degrading enzyme reveal a new substrate recognition mechanism. *Nature* **443**, 870–874 (2006).
306. Robertson, A. L. *et al.* Protein unfolding is essential for cleavage within the α-helix of a model protein substrate by the serine protease, thrombin. *Biochimie* **122**, 227–234 (2016).
307. Zhao, Y. *et al.* Structural basis for replicase polyprotein cleavage and substrate specificity of main protease from SARS-CoV-2. *Proc Natl Acad Sci U S A* **119**, e2117142119 (2022).
308. Lee, J. *et al.* X-ray crystallographic characterization of the SARS-CoV-2 main protease polyprotein cleavage sites essential for viral processing and maturation. *Nature Communications* **13**, 5196 (2022).
309. Rosas-Lemus, M. *et al.* High-resolution structures of the SARS-CoV-2 2′- *O* -methyltransferase reveal strategies for structure-based inhibitor design. *Sci. Signal.* **13**, eabe1202 (2020).
310. Koudelka, T. *et al.* N-Terminomics for the Identification of In Vitro Substrates and Cleavage Site Specificity of the SARS-CoV-2 Main Protease. *Proteomics* **21**, 2000246 (2021).
311. Chuck, C.-P., Chow, H.-F., Wan, D. C.-C. & Wong, K.-B. Profiling of Substrate Specificities of 3C-Like Proteases from Group 1, 2a, 2b, and 3 Coronaviruses. *PLoS ONE* **6**, e27228 (2011).
312. Ziebuhr, J. & Siddell, S. G. Processing of the human coronavirus 229E replicase polyproteins by the virus-encoded 3C-like proteinase: Identification of proteolytic products and cleavage sites common to pp1a and pp1ab. *Journal of Virology* **73**, 177–185 (1999).
313. Dai, W. *et al.* Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **368**, 1331–1335 (2020).
314. Slanina, H. *et al.* Coronavirus replication–transcription complex: Vital and selective NMPylation of a conserved site in nsp9 by the NiRAN-RdRp subunit. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2022310118 (2021).
315. Yaghi, R. M., Andrews, C. L., Wylie, D. C. & Iverson, B. L. High-Resolution Substrate Specificity Profiling of SARS-CoV-2 Mpro; Comparison to SARS-CoV Mpro. *ACS Chem. Biol.* **19**, 1474–1483 (2024).
316. Hegyi, A. & Ziebuhr, J. Conservation of substrate specificities among coronavirus main proteases. *J Gen Virol* **83**, 595–9 (2002).
317. Chuck, C.-P. *et al.* Profiling of substrate specificity of SARS-CoV 3CLpro. *PloS one* **5**, e13197 (2010).
318. Dai, W. *et al.* Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **368**, 1331–1335 (2020).
319. Bonanno, J. B. *et al.* X-Ray Crystal Structure of the SARS Coronavirus Main Protease: 1q2w. https://doi.org/10.2210/pdb1q2w/pdb (2003).
320. Lovell, S. *et al.* 1.55 A resolution structure of MERS 3CL protease in complex with inhibitor GC813: 5wkk. https://doi.org/10.2210/pdb5wkk/pdb (2018).
321. Lee, C. C. & Wang, A. H.-J. complex structure of CoV 229E 3CL protease with EPDTC: 2zu2. https://doi.org/10.2210/pdb2zu2/pdb (2009).
322. Stobart, C. C. *et al.* Chimeric Exchange of Coronavirus nsp5 Proteases (3CLpro) Identifies Common and Divergent Regulatory Determinants of Protease Activity. *J Virol* **87**, 12611–12618 (2013).
323. Kräusslich, H.-G., Nicklin, M. J. H., Lee, C.-K. & Wimmer, E. Polyprotein processing in picornavirus replication. *Biochimie* **70**, 119–130 (1988).
324. Matsuda, A. *et al.* Despite the odds: formation of the SARS-CoV-2 methylation complex. *Nucleic Acids Research* **52**, 6441–6458 (2024).
325. Yan, L. *et al.* Cryo-EM Structure of an Extended SARS-CoV-2 Replication and Transcription Complex Reveals an Intermediate State in Cap Synthesis. *Cell* **184**, 184-193.e10 (2021).
326. Den Boon, J. A., Diaz, A. & Ahlquist, P. Cytoplasmic Viral Replication Complexes. *Cell Host & Microbe* **8**, 77–85 (2010).

327. Lulla, V. *et al.* Timeliness of Proteolytic Events Is Prerequisite for Efficient Functioning of the Alphaviral Replicase. *Journal of Virology* **92**, (2018).
328. Lulla, V., Karo-Astover, L., Rausalu, K., Merits, A. & Lulla, A. Presentation Overrides Specificity: Probing the Plasticity of Alphaviral Proteolytic Activity through Mutational Analysis. *Journal of Virology* **87**, 10207–10220 (2013).
329. Mechanistic insights into COVID-19 by global analysis of the SARS-CoV-2 3CLpro substrate degradome: Cell Reports. https://www.cell.com/cell-reports/fulltext/S2211-1247(21)01362-0.
330. Alhadrami, H. A. *et al.* Peptide substrate screening for the diagnosis of SARS-CoV-2 using fluorescence resonance energy transfer (FRET) assay. *Microchim Acta* **188**, 137 (2021).
331. Ziebuhr, J. & Siddell, S. G. Processing of the Human Coronavirus 229E Replicase Polyproteins by the Virus-Encoded 3C-Like Proteinase: Identification of Proteolytic Products and Cleavage Sites Common to pp1a and pp1ab. *J Virol* **73**, 177–185 (1999).
332. Wu, A. *et al.* Prediction and biochemical analysis of putative cleavage sites of the 3C-like protease of Middle East respiratory syndrome coronavirus. *Virus Research* **208**, 56–65 (2015).
333. Zhang, L. *et al.* Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science* **368**, 409–412 (2020).
334. Heck, A. J. R. Native mass spectrometry: a bridge between interactomics and structural biology. *Nat Methods* **5**, 927–933 (2008).
335. Krichel, B. In vitro processing and complex formation of coronavirus polyprotein NSP7-10 region. (Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky, 2019).
336. Konermann, L. Addressing a Common Misconception: Ammonium Acetate as Neutral pH “Buffer” for Native Electrospray Mass Spectrometry. *Journal of The American Society for Mass Spectrometry* **28**, 1827–1835 (2017).
337. Shoemaker, G. K., Soya, N., Palcic, M. M. & Klassen, J. S. Temperature-dependent cooperativity in donor-acceptor substrate binding to the human blood group glycosyltransferases. *Glycobiology* **18**, 587–592 (2008).
338. Rapid online buffer exchange for screening of proteins, protein complexes and cell lysates by native mass spectrometry | Nature Protocols. https://www.nature.com/articles/s41596-019-0281-0.
339. Hecht, E. S., Scigelova, M., Eliuk, S. & Makarov, A. Fundamentals and Advances of Orbitrap Mass Spectrometry. in *Encyclopedia of Analytical Chemistry* 1–40 (John Wiley & Sons, Ltd, 2019). doi:10.1002/9780470027318.a9309.pub2.
340. Zhang, S., Van Pelt, C. K. & Wilson, D. B. Quantitative Determination of Noncovalent Binding Interactions Using Automated Nanoelectrospray Mass Spectrometry. *Anal. Chem.* **75**, 3010–3018 (2003).
341. Gavriilidou, A. F. M., Sokratous, K., Yen, H.-Y. & De Colibus, L. High-Throughput Native Mass Spectrometry Screening in Drug Discovery. *Front Mol Biosci* **9**, 837901 (2022).
342. Grün, A. F. R. *et al.* Fast tracking native mass spectrometry: Skipping over buffer exchange. 2025.02.22.639503 Preprint at https://doi.org/10.1101/2025.02.22.639503 (2025).
343. Kakavandi, S. *et al.* Structural and non-structural proteins in SARS-CoV-2: potential aspects to COVID-19 treatment or prevention of progression of related diseases. *Cell Communication and Signaling* **21**, 110 (2023).
344. Paim, F. C. *et al.* Epidemiology of Deltacoronaviruses (δ-CoV) and Gammacoronaviruses (γ-CoV) in Wild Birds in the United States. *Viruses* **11**, 897 (2019).
345. Guo, J., He, J., Liang, Z., Huang, S. & Wen, F. Birds as reservoirs: unraveling the global spread of Gamma- and Deltacoronaviruses. *mBio* **15**, e02324-24 (2024).
346. Hoferle, P. J., Anderson, T. K. & Kirchdoerfer, R. N. A genus-specific nsp12 region impacts polymerase assembly in Alphacoronavirus and Gammacoronavirus. *J Biol Chem* **300**, 107802 (2024).
347. Gan, J. *et al.* Native Mass Spectrometry of Recombinant Proteins from Crude Cell Lysates. *Anal. Chem.* **89**, 4398–4404 (2017).
348. Vimer, S., Ben-Nissan, G. & Sharon, M. Direct characterization of overproduced proteins by native mass spectrometry. *Nat Protoc* **15**, 236–265 (2020).
349. Melby, J. A. *et al.* Novel Strategies to Address the Challenges in Top-Down Proteomics. *J. Am. Soc. Mass Spectrom.* **32**, 1278–1294 (2021).
350. CDC. Laboratory Biosafety Guidelines for working with SARS-CoV-2. *COVID-19* https://www.cdc.gov/covid/php/lab/index.html (2025).
351. Soh, T. K. *et al.* A validated protocol to UV-inactivate SARS-CoV-2 and herpesvirus-infected cells. *PLOS ONE* **18**, e0274065 (2023).
352. Camarillo, J. M. *et al.* Coupling Fluorescence-Activated Cell Sorting and Targeted Analysis of Histone Modification Profiles in Primary Human Leukocytes. *J Am Soc Mass Spectrom* **30**, 2526–2534 (2019).

353. Schwenzer, A.-K., Kruse, L., Jooß, K. & Neusüß, C. Capillary electrophoresis-mass spectrometry for protein analyses under native conditions: Current progress and perspectives. *PROTEOMICS* **24**, 2300135 (2024).
354. Zhai, Z. *et al.* Characterization of Complex Proteoform Mixtures by Online Nanoflow Ion-Exchange Chromatography-Native Mass Spectrometry. *Anal. Chem.* **96**, 8880–8885 (2024).
355. Chea, E. E. & Jones, L. M. Modifications generated by fast photochemical oxidation of proteins reflect the native conformations of proteins. *Protein Science* **27**, 1047–1056 (2018).
356. Smyrnakis, A. *et al.* Characterization of an Omnitrap-Orbitrap Platform Equipped with Infrared Multiphoton Dissociation, Ultraviolet Photodissociation, and Electron Capture Dissociation for the Analysis of Peptides and Proteins. *Anal. Chem.* **95**, 12039–12046 (2023).
357. Papanastasiou, D. *et al.* The Omnitrap Platform: A Versatile Segmented Linear Ion Trap for Multidimensional Multiple-Stage Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **33**, 1990–2007 (2022).
358. Papanastasiou, D. *et al.* The Omnitrap Platform: A Versatile Segmented Linear Ion Trap for Multidimensional Multiple-Stage Tandem Mass Spectrometry. *J. Am. Soc. Mass Spectrom.* **33**, 1990–2007 (2022).
359. Esser, T. K. *et al.* Cryo-EM of soft-landed β-galactosidase: Gas-phase and native structures are remarkably similar. *Science Advances* **10**, eadl4628 (2024).
360. Kierspel, T. *et al.* Coherent diffractive imaging of proteins and viral capsids: simulating MS SPIDOC. *Anal Bioanal Chem* **415**, 4209–4220 (2023).
361. Kung, J. C. K. *et al.* X-ray Spectroscopy Meets Native Mass Spectrometry: Probing Gas-phase Protein Complexes. *Phys. Chem. Chem. Phys.* (2025) doi:10.1039/D5CP00604J.
362. Chen, J., Malone, B., Campbell, E. A. & Darst, S. A. SARS-CoV-2 replication-transcription complex bound to nsp13 helicase - nsp13(1)-RTC: 7re2. https://doi.org/10.2210/pdb7re2/pdb (2021).
363. Strohalm, M., Kavan, D., Novák, P., Volný, M. & Havlíček, V. mMass 3: A Cross-Platform Software Environment for Precise Analysis of Mass Spectrometric Data. *Anal. Chem.* **82**, 4648–4651 (2010).
364. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Science* **27**, 14–25 (2018).
365. Madeira, F. *et al.* The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Res* **52**, W521–W525 (2024).
366. Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics* **25**, 1189–1191 (2009).
367. Procter, J. B. *et al.* Alignment of Biological Sequences with Jalview. *Methods Mol Biol* **2231**, 203–224 (2021).
368. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences* **89**, 10915–10919 (1992).
369. Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344–W350 (2016).
370. Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Research* **38**, W529–W533 (2010).
371. Celniker, G. *et al.* ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Israel Journal of Chemistry* **53**, 199–206 (2013).
372. Yariv, B. *et al.* Using evolutionary data to make sense of macromolecules with a “face-lifted” ConSurf. *Protein Science* **32**, e4582 (2023).
373. Berezin, C. *et al.* ConSeq: the identification of functionally and structurally important residues in protein sequences. *Bioinformatics* **20**, 1322–1324 (2004).
374. Marty, M. T. *et al.* Bayesian Deconvolution of Mass and Ion Mobility Spectra: From Binary Interactions to Polydisperse Ensembles. *Anal. Chem.* **87**, 4370–4376 (2015).
375. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**, 14–25 (2018).
376. Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70–82 (2021).
377. Sänger, L. Characterization of interactions between Lassa virus nucleoprotein, matrix protein and RNA essential for RNP assembly and recruitment. (Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky, 2024).
378. Deutsch, E. W. *et al.* The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res* **45**, D1100–D1106 (2017).
379. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Research* **47**, D442–D450 (2019).
380. Schamoni-Kast, K. *et al.* The kinetics of nsp7-11 polyprotein processing and impact on complexation with nsp16 among human coronaviruses. Zenodo https://doi.org/10.5281/ZENODO.15488266 (2025).

8. Acknowledgements

This ‘native MS’-story began with an internship during my bachelor's degree and I have not been able to get rid of this technology to this day. The mass spec field is too broad and fascinating. Thank you, **Charlotte** for wanting me as an intern back then and for taking me on again later as a PhD student. Like so many other doctoral theses, mine didn't go according to the plan. But I learned a lot in the process, thank you for that.

I'll be honest, it wasn't just the technology that led me back to this group for my doctoral thesis, but also my colleagues. A funfact: As a clueless person back then, I thought: ‘well, coronaviruses aren't that exciting, but however...’; and suddenly we were in the middle of the pandemic. So maybe it wasn't the worst idea to start the PhD in a familiar place. In the end, it felt like doing two doctors: One at LIV and the other one at CSSB. My thanks are therefore expressed in chronological order.

For the start at LIV many thanks to **Nicho** who brought me back to wet lab mode after writing my master thesis at the center for bioinformatics. Thank you, **Lennart**, for being my HDX partner in crime and being a great companion at IMSC2022. Of course, we would never have learned HDX without a great teacher! Thank you, **Jasmin**!

I will always be grateful for the opportunity that the LIV Mobility Scholarship gave me. I had such an exciting time in the Jones Lab in Baltimore in 2021; meeting all the special people and learning FPOP-MS. Many thanks to **Lisa, Raqie, Danté, Riley** and **Chris.**

Thanks to my first students **Sibel & Tobi** who I supervised together with Boris. It is really a pity that nsp12 production did not work out again for a long time after the move to CSSB, but now we are back on track. Thank you, **Ronja** that we went together through these many intricacies of the mass spec move and the reconstruction of the new lab.

There is no great beamtime, without awesome colleagues. Thank you, **Thomas, Tomislav** and **Jocky**. It’s always nice to come by and chat with you about kids, cats and Canada, respectively. Thank you, **Simon**, for showing me how to work with Git. Thank you **Lo**, for always asking how it goes and really meaning it. Thank you **Sanne**, my favorite Norwegian student. I am sorry that *E. coli*-nsp12 turned out to be your worst enemy. It was never up to you. Thanks to **Jürgen** for your expertise and interesting early morning thoughts. Thank you **Aqila & Assel** for bringing the nsp crew 2.0. to life. The future is bright.

Thank you, **Alice**, for the cheekiest sayings during lunch breaks. Thank you, **Marius**, whenever I needed an opinion, I knew who to ask. Thank you, **Lars,** for the best dad jokes. Thank you, **Jonas,** for always being helpful and being the only person interested in football, like me. Thanks, **Jenny**, for reading my AF-part and for cheering me up with your dirty laugh. Thank

you, **AG77**, for making the office a funny and loud, but very comfortable place. Thanks, **Boris**, for mentoring. Working with you was funny, chaotic and fruitful, and thanks for the deep talks about childhood in Bronto cave. Thanks, **Janine**, for being the pickiest proofreader of all times, for being always helpful and open to all kinds of questions. I always appreciated you honest and direct personality. Keep this and your kind heart.

An alle Freunde und Familie, die mitgefiebert und gelitten haben. Danke dafür! Für euer offenes Ohr und eure Ratschläge. Danke an: **Jenny, Uschi, Kerri, Isa, Oda, Michi, Silja, Nadine, Sarah und Eva.**

Danke, **Mama und Papa**, für eure Unterstützung. Ihr hattet nie einen Plan für mich und habt mich einfach immer machen lassen. Ihr habt mich bestärkt und in allem bedingungslos unterstützt. DANKE! Das ist das Sicherungsseil, das einen immer höher Klettern lässt. Danke, **Pia, Marina** und **Philipp**. Wir sind alle so grundverschieden und die Distanz macht es nicht leichter, Zeit miteinander zu verbringen. Wenn wir es tun, dann ist es wie früher. Danke dafür, dass ihr so eine wundervolle Familie seid. Ein besonderer Dank an meine Zwillingsschwester, für den wertvollen Blick von außen, für den klaren Blick und die Unterstützung von Herzen. Zu guter Letzt: Meine eigene kleine Familie, **Ina** und **Matteo**. Der doppelte Boden. Danke Ina, dass du mit mir durch diese Höhen und Tiefen der Doktorzeit gegangen bist. Vor allem dieses letzte Jahr war so intensiv und anstrengend. Du wusstest immer, was ich gerade brauchte, danke fürs Zuhören, Motivieren, Trösten und Aushalten. Danke für deine Unterstützung und deine Liebe. Matteo, ich würde lügen, wenn ich sagen würde, dass du alles einfacher gemacht hast. Du hast aber, alles schöner und wertvoller gemacht. Vor allem, konntest du zuverlässig jedes noch so erfolglose Experiment vergessen machen und mich daran erinnern, was im Leben wirklich zählt. Danke dafür. Bleib so verrückt, wie du bist.

9. Declaration of Authorship| Eidesstattliche Versicherung

„Hiermit versichere ich an Eides statt, die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt zu haben. Sofern im Zuge der Erstellung der vorliegenden Dissertationsschrift generative Künstliche Intelligenz (gKI) basierte elektronische Hilfsmittel verwendet wurden, versichere ich, dass meine eigene Leistung im Vordergrund stand und dass *eine vollständige Dokumentation aller verwendeten Hilfsmittel gemäß der Guten wissenschaftlichen Praxis vorliegt. Ich trage die Verantwortung für eventuell durch die gKI generierte fehlerhafte oder verzerrte Inhalte, fehlerhafte Referenzen, Verstöße gegen das Datenschutz- und Urheberrecht oder Plagiate.“*

21.10.2025 K. Schanz-Kost

Datum, Unterschrift

"I hereby declare and affirm that this doctoral dissertation is my own work and that I have not used any aids and sources other than those indicated. If electronic resources based on generative artificial intelligence (gAI) were used in the course of writing this dissertation, I confirm that my own work was the main and value-adding contribution and that complete documentation of all resources used is available in accordance with good scientific practice. I am responsible for any erroneous or distorted content, incorrect references, violations of data protection and copyright law or plagiarism that may have been generated by the gAI.."

21.10.2025 K. Schanz-Kost

Datum, Unterschrift