VALIDATION OF PEPTIDES AS PROTEIN SUBSTRATE MODELS FOR SPECIFICITY STUDIES OF CYSTEINE CATHEPSINS

Jure Loboda

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Prof. Dr. Dušan Turk, IPS and Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Prof. Ddr. Boris Turk, Chair, IPS and Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Dr. Janko Kos, Member, Jožef Stefan Institute and Faculty of Pharmacy, Ljubljana, Slovenia
Assist. Prof. Gregor Gunčar, Member, Faculty of Chemistry and Chemical Technology, Ljubljana, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Jure Loboda

VALIDATION OF PEPTIDES AS PROTEIN SUBSTRATE MODELS FOR SPECIFICITY STUDIES OF CYSTEINE CATHEPSINS

Doctoral Dissertation

VALIDACIJA PEPTIDOV KOT MODELOV PROTEINSKIH SUBSTRATOV ZA ŠTUDIJE SPECIFIČNOSTI CISTEINSKIH PROTEAZ

Doktorska disertacija

Supervisor: Prof. Dr. Dušan Turk

Ljubljana, Slovenia, November 2022

This work is dedicated to my loving family.

Acknowledgments

Prof. Dr. Dušan Turk, for his sincere help and mentorship during my PhD study.
Prof. Dr. Livija Tušar, for statistical analysis of proteomic data and her contribution with the interpretation and presentation of results of structural analysis.
Dr. Piotr Sosnowski, for optimizing crystallization conditions and for solving structures of crystal complexes 7Q8I, 7Q8Q, 7QO2, 7Q8K, 7Q8P, 7Q8G, 7Q8O, 7Q8J, 7QHK and 7QNS.
Dr. Katarina Karničar, Dr. Nataša Lindič, Dr. Aleksandra Usenik and Andreja Sekirnik for their help with the characterization of inhibitor compounds.
Prof. Ddr. Boris Turk, Prof. Dr. Marko Fonovič and Dr. Matej Vizovišek for their contribution with performing the proteomic analysis.
Prof. Dr. Francis Impens, Prof. Dr. Kris Gevaert and Hans Demol (Ghent, Belgium) for performing the proteomic analysis.
Dr. Robert Vidmar for guiding the work on MALDI-TOF, which was used for determining the peptide cleavage sites.
Dr. Gregor Kosec and Dr. Jaka Horvat (Acies Bio, d.o.o.) for their help with peptide purification.
Dr. Patrick YA Reinke, Prof. Dr. Sebastian Günther and Dr. Alke Meents (Hamburg, Germany) for leading the work on calpeptin project.
Dr. Elma Mons and Prof. Dr. Huib Ovaa (Leiden, Netherlands) for leading the work on cathepsin K inhibitor project.
Dr. Ajda Taler-Verčič, for introducing me to the laboratory work.
Dr. Matej Vizovišek and Ivica Štefe for their help and advices in the early stages of my PhD study.
Dr. Lea Udovč from K9 for kindly lyophilizing my peptide samples.

To all the members and formal members of the B12 group for their help, advices and many fruitful discussions during my PhD study: Prof. Dr. Dušan Turk, Dr. Nataša Lindič, Dr. Katarina Karničar, Marinka Horvat, Tjaša Peternel, Dr. Aleksandra Usenik, Dr. Ajda Taler-Verčič, Andreja Sekirnik, Prof. Dr. Livija Tušar, Matej Novak, Klemen Dretnik and Ana Kump.

To all other co-workers at the B1 department that helped me in any way.

To the examination board (Prof. Ddr. Boris Turk, Prof. Dr. Janko Kos and Assist. Prof. Gregor Gunčar).

To staff at the Jožef Stefan International Postgraduate School.

To CIPKEBIP and ARRS (Program P1-0048) for funding this work.

Abstract

Cysteine cathepsins are endosomal proteases that are involved in lysosomal protein degradation. Additionally, they engage in specific biological processes, like bone degradation (CatK), antigen presentation (CatS) and pro-hormone processing (CatV and CatL). Their enhanced activity was described in numerous pathological states, like cancer and viral activation, which made them promising drug targets. Yet their specificity remains elusive because methodology and tools applied could not match the complexity of their processing patterns. Recent statistical analysis of cleavage patterns from large-scale proteomics data enabled us to address this issue. Thirty peptides, representing a variety of all seven major clusters of CatV substrates, were selected and synthesized. Their interactions with the active site mutant of CatV were analyzed in crystal structures, and their cleavage patterns by the native CatL, K, and V were studied in solution.

Over 20 crystal structures of CatV-peptide complexes were determined. They were grouped into four binding patterns, based on their binding location or cleavage event. They interacted with the cathepsin surface at the sites S4 – S6'. Superimposition of structures showed that the residues at positions with non-normal distribution of residues, called heterogeneous positions, reflect the cathepsin specificity. Peptidyl residues at such positions bound to the rigid cathepsin regions. In contrast, the residues at positions with normal distribution of residues, called homogeneous positions, do not reflect cathepsin specificity. Peptidyl residues at such positions exploited the protease structural variability, sometimes also inducing it. By analyzing the structures of CatK, L, and S, from PDB database, we showed that the same conclusions can be applied to them, too. Moreover, we were able to pinpoint to the structural areas and the contributing residues responsible for specificity of CatL substrates at P3 and CatL, V, and F substrates at P1'. These areas can be targeted to enhance selective inhibition of cathepsin endopeptidases.

We compared cleavages of peptides and proteins, carrying the same primary sequence, and found that several were cleaved at different places when they were in the peptidyl form or as a part of the protein structure. The comparison of peptide complexes with the sole crystal complex of CatL and a protein substrate, published in the PDB database, suggested that substrate specificity of cathepsins cannot be explained by contributions of individual amino acids, but by the combined effect of the substrate as a whole. Hence, the knowledge gathered from the peptide processing does not necessarily apply to the processing of protein substrates and vice versa.

In addition, we determined the crystal complex of CatK and an alkyne-based inhibitor. The electron density indisputably confirmed the covalent bond between the catalytic Cys of CatK and the alkyne of inhibitor, showing that the alkyne warhead can be used as latent electrophile for targeting proteases. In addition, we showed that calpeptin and compounds alike, which inhibit M^{pro} protease of the SARS-CoV-2 virus and suppress viral activation in cell-based assays, strongly inhibit the human CatL, K, V, and to a lesser extent also CatB. This suggests that the viral suppression mediated by calpeptin may primarily be due to the cathepsin inhibition rather than inhibition of M^{pro}. To provide further support, we determined the crystal structure of CatV-calpeptin complex.

Povzetek

Cisteinski katepsini so pomembne endosomalne proteaze, primarno udeležene v razgradnjo proteinov. Nekateri izmed njih so dodatno vključeni v specifične biološke procese, kot so na primer razgradnja kostnega matriksa (katepsin K), predstavitev antigenih peptidov (katepsin S) in procesiranje pro-hormonov (katepsina L in V). Povečana aktivnost katepsinov je zaznana v nekaterih patoloških stanjih, kot sta na primer rak in aktivacija virusov. Njihova substratna specifičnost še ni povsem razjasnjena, saj metološki pristopi, ki so na voljo, ne sledijo kompleksnosti njihovega procesiranja. Tega problema smo se lotili s pomočjo statistične analize katepsinskih cepitev, pridobljenih iz obsežne proteomske študije. Izmed vseh cepitev smo izbrali in sintetizirali 30 peptidov, ki predstavljajo raznolikost vseh glavnih klastrov substratov katepsina V. Njihove interakcije s katepsinom V smo preučili v kristalnih kompleksih ter določili njihova cepitvena mesta z divjimi oblikami človeških katepsinov L, K, in V.

Določili smo preko 20 kristalnih struktur kompleksov med katepsinom V in posameznimi peptidi. Strukture smo razdelili v štiri skupine glede na področje vezave peptida oziroma glede na prisotnost cepitve. Interakcije med katepsinom in peptidi so se tvorile na mestih med S4 – S6'. S superimpozicijo struktur smo pokazali, da se preostanki na pozicijah z nenormalno porazdelitvijo preostankov – imenujemo jih heterogene pozicije - vežejo na katepsinska področja, ki so rigidna, in odražajo njihovo specifičnost. Nasprotno se preostanki na pozicijah z normalno porazdeljenimi preostanki – imenujemo jih homogene pozicije – vežejo na fleksibilna področja in nimajo vpliva na specifičnost. Podobno velja za katepsine K, L in S, kar smo ugotovili z analizo že objavljenih kristalnih struktur v bazi PDB. Pomembno je tudi to, da smo uspeli izpostaviti mesta na katepsinih in pripadajoče preostanke, ki ključno prispevajo k specifičnosti katepsina L na mestu P3 in katepsinov L, V in F na mestu P1'. Ta mesta se lahko izrabljajo pri načrtovanju novih specifičnih inhibitorjev katepsinskih endopeptidaz.

Cepitvena mesta peptidov in proteinov z enakim aminokoslinskih zaporedjem niso vselej sovpadala. Primerjava peptidnih kompleksov in edinega že objavljenega kompleksa med proteinom in katepsinom L nakazuje, da prepoznava substrata v aktivnem mestu katepsina ni odvisna zgolj od primarne strukture in afinitete aminokislin do aktivnega mesta, ampak tudi od njene strukture. Procesiranje peptidov zato ne odraža vedno procesiranja proteinov, in obratno.

Poleg tega smo določili kristalno strukturo kompleksa med katepsinom K in inhibitorjem z alkinsko reaktivno skupino. Elektronska gostota inhibitorja nedvoumno potrjuje tvorbo vezi med reaktivnim cisteinom katepsina K in alkinsko skupino inhibitorja, kar prikazuje, da se alkinska skupina lahko uporablja kot šibak in počasen elektrofil za specifično ciljanje tarčnih proteaz. Poleg tega smo pokazali, da so kalpeptin in njemu podobne spojine, ki inhibirajo proteazo M^{pro} virusa SARS-CoV-2 in preprečujejo njegovo aktivacijo v celičnih linijah, zelo močni inhibitorji človeških katepsinov L, K in V, ter tudi – malo manj močni – katepsina B. To nakazuje, da so morda glavna tarča pri zaviranju virusne aktivacije prav katepsini. Našo tezo smo dodatno podkrepili z določitvijo kristalne strukture kompleksa med katepsinom V in kalpeptinom.

Contents

List of Figures **xvii**

List of Tables **xix**

Abbreviations **xxi**

1 Introduction **1**
1.1 Proteases ... 1
1.1.1 Biological roles and regulation ... 1
1.1.2 Substrates of proteases ... 2
1.1.2.1 Protease profiling ... 2
1.1.3 Protease inhibitors ... 3
1.1.4 Cysteine cathepsins ... 4
1.1.4.1 Cysteine cathepsins as drug targets ... 4
1.1.4.1.1 Cysteine cathepsins as drug targets in SARS-CoV-2 ... 5
1.2 Studies of Protease Substrates and Inhibitors ... 6
1.2.1 Macromolecular crystallography ... 6
1.2.1.1 Protein crystallization ... 6
1.2.1.2 X-ray scattering from the crystal ... 7
1.2.1.3 Euler description, Fourier transform and structure factors ... 7
1.2.1.4 The Phase problem ... 8
1.2.1.4.1 Isomorphous replacement ... 8
1.2.1.4.2 Anomalous diffraction ... 8
1.2.1.4.3 Molecular replacement ... 9
1.2.1.4.4 Direct methods ... 9
1.2.1.5 Electron density map and model building ... 9
1.2.1.6 Model refinement ... 10
1.2.1.7 Model validation ... 10
1.3 Enzyme Kinetic Studies ... 10
1.3.1 Michaelis - Menten model ... 11
1.3.2 Enzyme activity in presence of inhibitor ... 11
1.3.2.1 Reversible inhibitors ... 11
1.3.2.2 Irreversible inhibition ... 12
1.3.2.3 Molecular reporters ... 13
1.4 Mass Spectrometry ... 13
1.4.1 Mass spectrometry for protease studies ... 13

2 The Aim of the Work **15**

3 Hypotheses **17**

4 Materials and methods **19**

4.1 Protein Expression 19
4.2 Protein Activation and Purification 19
4.2.1 Procathepsin K. 19
4.2.2 Procathepsin L, Procathepsin V and Procathepsins V C25S/A 20
4.2.2.1 Procathepsin V 20
4.2.2.2 Procathepsin L 21
4.2.2.3 Procathepsins V C25S/A 21
4.3 Peptide Purification 21
4.4 Protein Crystallization 22
4.4.1 Cathepsin K - alkyne inhibitor crystallization 22
4.4.2 Cathepsin V – calpeptin crystallization 22
4.4.3 Cathepsin V C25S/A - peptide soaking 23
4.4.4 Cathepsin V C25S/A - peptide co-crystallization 23
4.4.5 Data collection, structure determination and refinement 23
4.5 Inhibitory Assays 23
4.5.1 K_i determination of moderate inhibitors 24
4.5.2 K_i determination of tight binding inhibitors 24
4.5.3 Covalent inactivation test 24
4.6 Peptide Cleavage Analysis 25
4.6.1 Peptide digestion 25
4.6.2 Peptide separation on RP-HPLC 25
4.6.3 Mass spectrometry analysis 25

5 Results **27**
5.1 Structural and Biochemical Analysis of Peptide Binding to Cathepsin V 27
5.1.1 Peptide selection 27
5.1.2 Crystallization of cathepsin V-peptide complexes 28
5.1.3 Peptide digestion with cathepsins K, V, and L 34
5.2 Crystal Structure of Cathepsin K - Alkyne Inhibitor 36
5.2.1 Selectivity and reactivity of Odanacatib-like alkyne inhibitors 36
5.2.2 Crystallization of cathepsin K – alkyne-based inhibitor 36
5.2.3 Data collection and structure determination 38
5.3 Characterization of Calpeptin and Alike Compounds as Cathepsin Inhibitors 41
5.3.1 Enzyme inhibition assays 41
5.3.1.1 IC_{50} screen 41
5.3.1.2 K_i determination 42
5.3.2 Crystallization of cathepsin V-calpeptin complex 44
5.3.3 Data collection and structure determination 45

6 Discussion **49**
6.1 Structural Basis for Heterogeneous and Homogeneous Positions of Cathepsin Substrates 49
6.2 Peptides as Protein Substrate Model 54
6.3 Relevance for Drug Discovery Projects 56

7 Conclusions **59**

Appendix A Peptide Selection **61**
A.1 List of Synthesized Peptides 61
A.2 Major Clusters of Cathepsin Substrates 63

Appendix B Crystal Structures of Cathepsin V-Peptide Complexes 64

Appendix C Comparison of Protein and Peptide Cleavages 69
C.1 List of Protein and Peptide Cleavages and Predictions 69
C.2 Peptide Fragment Separation and Identification with RP-HPLC – MALDI-TOF 72
C.2.1 HPLC spectra with peak identification 72
C.2.2 Processing of peptides AYFKKVL and KVLATVTK from 5 sec-60 min. 80

Appendix D Author contributions 83

References 85

Bibliography 95

Biography 97

List of Figures

Figure 1: Crystallization phase diagram 6
Figure 2: SDS-PAGE of procathepsin K activation 20
Figure 3: H-bonding pattern between cathepsin V and its substrates 29
Figure 4: Binding geometry of peptides 30
Figure 5: Optimisation of cathepsin K inhibitors 36
Figure 6: Best crystals from the cathepsin K-alkyne inhibitor screening experiment 37
Figure 7: Crystal structure of cathepsin K-alkyne inhibitor 39
Figure 8: Chemical structures of calpeptin and alike compounds 41
Figure 9: IC_{50} screen of calpeptin and alike compounds 42
Figure 10: K_i determination of calpeptin and alike compounds 43
Figure 11: Crystal structure of cathepsin V-calpeptin 46
Figure 12. Binding specificity between Arg at P1' and D163 of cathepsin V 50
Figure 13: Flexible and rigid residues of cathepsins V and L and their substrate-binding areas 51
Figure 14: Flexible and rigid residues of cathepsins K and S from PDB database 52
Figure 15: The difference between peptide and protein binding to cathepsins 55
Figure 16: Comparison of calpeptin binding at S3 site of cathepsins L and V. 57

List of Tables

Table 1: Crystallographic table for cathepsin V-peptide complexes 28
Table 2: Summary of peptide binding to crystals of cathepsin V C25S/A 32
Table 3: Peptide and protein cleavage analysis 35
Table 4: Optimisation of cathepsin K-alkyne inhibitor crystallization conditions 37
Table 5: Crystallographic table for cathepsin K-alkyne inhibitor entry 6QBS. 40
Table 6: K_i values of calpeptin and compounds alike 44
Table 7: Crystallographic table for cathepsin V-calpeptin entry 7QGW. 47

Abbreviations

ACN	. . . acetonitrile
AMC	. . . 7-amino4-methylcoumarin
B-factor	. . . atomic displacement parameter
BMGY	. . . buffered glycerol-complex media
BMMY	. . . buffered methanol-complex media
$CaCL_2$	. . . calcium dichloride
CCP4	. . . Collaborative Computational Project Number 4
DABCYL	. . . 4-((4-(Dimethylamino)phenyl)azo)benzoic acid
DTT	. . . dithiothreitol
EDANS	. . . 5-((2-Aminoethyl)amino)naphthalene-1-sulfonic acid
EDTA	. . . 2,2′,2″,2‴-(Ethane-1,2-diyldinitrilo)tetraacetic acid
F_h	. . . structure factor
HCCA	. . . α-Cyano- 4-hydroxycinnamic acid
HEPES	. . . 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HIV	. . . human immunodeficiency virus
HPLC	. . . high performance liquid chromatography
k_{cat}	. . . the turnover number
k_{inact}	. . . the inactivation number
k_{on}	. . . the association rate
k_{off}	. . . the dissociation rate
K_2HPO_4	. . . dipotassium hydrogen phosphate
KH_2PO_4	. . . potassium dihydrogen phosphate
K_i	. . . the dissociation constant
K_m	. . . Michaelis-Menten constant
M^{pro}	. . . SARS-CoV-2 major protease
MAD	. . . multi-wavelength anomalous dispersion
MeOH	. . . methanol
MALDI	. . . matrix-assisted laser desorption/ionization
MALDI-TOF	. . . MALDI coupled with TOF
MHC	. . . major histocompatibility complex
MIR	. . . multiple isomorphous replacement
ML AK	. . . maximum-likelihood averaged kick map
MM	. . . Michaelis-Menten model
MMPs	. . . metalloproteinases
MMP-2	. . . metalloproteinase-2
MMP-9	. . . metalloproteinase-9
MPD	. . . 2-methyl-2,4-pentadiol
MS	. . . mass spectrometry
NaCl	. . . sodium chloride
NaOAc	. . . sodium acetate
$NH_4H_2PO_4$	. . . ammonium dihydrogen phosphate

Ni^{2+}-NTA	... nickel bound to nitrilotriacetic acid
ODN	... Odanacatib
OXT	... terminal oxygen atom
PDB	... protein data bank
PEG	... polyethylene glycol
PICS	... proteomic identification of protease cleavage sites
PURY	... pure parameters and topology files for crystallography
RCF	... relative centrifugal force
RFU	... relative fluorescence unit
RMSD	... root-mean-square deviation
RP-HPLC	... reverse phase high performance liquid chromatography
SAD	... single-wavelength anomalous dispersion
SDS-PAGE	... sodium dodecyl sulphate polyacrylamide gel electrophoresis
SIR	... single isomorphous replacement
S-protein	... Spike-protein
TFA	... trifluoroacetic acid
TOF	... time of flight mass spectrometer
TRIS	... 2-Amino-2-(hydroxymethyl)propane-1,3-diol
Z-FR-AMC	... N-carbobenzyloxy-Phe-Arg-7-amino4-methylcoumarin
Z-RR-AMC	... N-carbobenzyloxy-Arg-Arg-7-amino4-methylcoumarin

Chapter 1

Introduction

1.1 Proteases

Proteases cleave peptide (or amide) bonds of their substrates. Peptide bond is covalent bond between a carboxylic group of one amino acid residue and an amino group of the next residue in the peptide or protein molecule. Proteases are divided into endopeptidases, when they cleave in the middle of a protein or exopeptidases, when they cleave at their termini; more precisely, they are referred to as amino-peptidases or carboxy-peptidases when they perform cleavage at amino (N) or carboxy (C) terminal residue, respectively. Based on the type of the main catalytic residue, they are classified as serine, cysteine, threonine and aspartic proteases, or as metalloproteinases, when they use a metal ion for catalysis. The protease surface where the substrate binds and gets cleaved is called the active site cleft or simply the active site.

1.1.1 Biological roles and regulation

Proteases represent around 2 % of the human genome (Craik *et al.*, 2011). They were first recognized as catabolic (degrading) enzymes and their roles described in food digestion and intracellular protein turnover. In both of these processes, enzymes degrade their protein substrates gradually down to single amino-acids. In time, involvement of proteases in cell signaling was discovered and became an area of intensive research. In signaling, proteases do not degrade their substrates to smallest possible units, which is known as the unlimited proteolysis, but they cleave them at precise site(s), which is known as the limited proteolysis. Thereby the generated protein fragments either gain or lose a certain physiological function. Proteases control numerous key physiological processes, such as cell cycle progression, tissue remodeling, DNA replication, cell proliferation, blood coagulation, apoptosis, immune signaling, neurodegeneration and more (Bond, 2019; B. Turk, 2006).

In cells, protease activity is regulated at different levels. Protease gene transcription is either decreased or increased, based on the cell current needs. After their synthesis at the ribosomes, they are sent to specific locations in the cell, for example to endosomes or to cell membrane, where their actions are needed. They are synthesized in proenzyme form, which is inactive until the "pro" part is removed from the active protease sequence. Hence, their unwanted actions elsewhere in the cell are prevented. Cells also evolved a set of protein activators and inhibitors, which provide additional regulation by either enhancing or blocking proteolytic activity. Protease degradation is regulated by autolysis, degradation by other proteases, posttranslational modifications or receptor-mediated endocytosis (Twining, 1994). These regulatory mechanisms are crucial because abnormal protease activity, whether it is increased or decreased, has detrimental consequences to normal cell

or tissue homeostasis and usually leads to pathological states. Several of these were studied extensively, for example coagulopathies, cardiovascular and inflammatory diseases, osteoporosis, infectious diseases and cancer (López-Otín & Bond, 2008).

1.1.2 Substrates of proteases

Proteases actualize their roles by processing their substrates. Hence, to understand their function, it is necessary to identify their substrates and understand the effects of their processing (V. Turk et al., 2012). This knowledge helps to determine or predict protease biological roles. The ability of a protease to discern between its substrates and other peptidyl species, which it does not process, is called protease specificity. The pioneering study of the protease substrates dates back to 1967, when Schechter & Berger (Schechter & Berger, 1967) investigated the processing of poly-alanine peptides of different length by the enzyme papain. They found that the length of the peptide chain affects the speed of processing, and concluded that the length of seven Ala residues was optimal. According to their concept, each of the substrate residues binds to its own area on the enzyme. The separation of a residue from the cleavage site was called a position and was numbered, whereas the areas to which a certain residue bound was called subsite, which was numbered as well. They referred to the subsites, located N-terminally from the substrate binding point of view as non-primed subsites and those located C-terminal as primed subsites. The first non-primed subsite from the protease active site residue was named S1, the second S2, the third S3, and so on, and in a similar manner, the primed sites as S1', S2' and S3' and substrate residues that bind to these subsites were correspondingly named non-primed residues P1, P2, P3, and the primed residues P1', P2' and P3'.

1.1.2.1 Protease profiling

Proteases act on peptidyl species, such as peptides and proteins. Proteins are large polypeptides with defined 3-D structure, which shape their biological functions as well as their biochemical properties like size, charge distribution, solubility, stability, binding area for proteases and so on. Due to the advances in recombinant DNA technology, proteins can be nowadays produced and investigated in almost every biochemical laboratory. Nevertheless, their production and proper handling require human and time resources. In contrast, peptide production and handling is less laborious and they are hence routinely used in protease assays. High throughput screening of protease activity toward peptides with varying sequence in the form of commercially available peptide libraries, phage display technologies or digested cell lysates provide tools for initial characterization of virtually any protease. These methods are especially powerful for profiling proteases whose specificity is driven mainly by contribution of one subsite only, for example proteases like trypsin, endoproteinase Glu-C or caspases, due to their ability to readily expose which are favorable interactions at individual subsites (Choe et al., 2006; Gupta et al., 2010; Klingler & Hardt, 2012; O'Donoghue et al., 2012). The specificity of over 150 proteases have been explored in that ways (Drag & Salvesen, 2010). In contrast, unravelling specificity of proteases like metalloproteinases (MMPs) or cathepsins is more challenging because of their broader specificity (known as the ability to cleave different sequences), redundant processing (when the same sequence is cleaved by two or more similar proteases) and absence of one clear dominant position (Biniossek et al., 2011; Puzer et al., 2004; Ratnikov et al., 2021; Vidmar et al., 2017; Vizovišek et al., 2015). It is therefore remarkable that these proteases play roles also in specific biological processes. Hence, in order to understand their biological functions, an approach that could cope with their broad specificity profiles and redundancy needed to be applied.

1.1.3 Protease inhibitors

Proteases are important targets in drug discovery. It has been assessed that around 5 - 10 % of all drug targets, pursued by pharmaceutical industries, are proteases. Typically, drugs are small molecule inhibitors that bind to the protease active site, which prevents its association with substrates. This is not the case for allosteric inhibitors, which bind to distal protease areas and will not be discussed here any further. Their key properties are potency, which describes how strong is the association between the protease and the inhibitor, and selectivity, which describes inhibitor property to bind preferentially to the targeted protease. In this regard, the knowledge about protease processing patterns, its active site structure and interactions with the specific endogenous inhibitors help greatly in the design of potent and selective inhibitors. Inhibitors with low potency or low selectivity are more likely to cause side effects due to the increased chance of non-desirable or toxic interactions with other proteases or enzymes, which is referred to as the off-target side effects. In contrast, the side effects caused by the inhibition of a targeted protease are referred to as on-target side effects.

Before the drug is pursued, the biological roles of the targeted protease need to be well understood in order to predict and possibly avoid negative aspects of its inhibition. This includes detailed knowledge about the biological processes in which they are involved, pathways they regulate and identification of their substrate pool. Next, these roles need to be evaluated also in a disease state. Poor understanding of protease biology can have serious consequences in the late stages of drug development. One such example was the development of MMPs inhibitors as cancer therapeutics. The compounds caused severe side effects which were largely observed only in stage III of clinical trials. The compounds lacked selectivity because they inhibited several MMPs, leading to off-target side effects. In addition, it was not known at the time that MMPs are involved in immune regulation, thus leading to on-target side effects as well. These issues completely halted their further development. Nevertheless, proteases represent an excellent target and this is illustrated by several successful protease inhibitors which are indispensable in clinical use, for example the inhibitors of angiotensin-converting enzyme, widely used in treatment of hypertension and congestive heart failure; inhibitors of coagulation factors IIa (thrombin) and factor Xa that are used for prophylaxis of cardiovascular diseases, related to hyper-coagulative states and inhibitors of HIV protease which reduce the HIV viral load. (Coussens et al., 2002; Drag & Salvesen, 2010; Kaysser, 2019).

There are two main strategies of protease inhibition: the reversible and irreversible. Reversible inhibitors bind to the protease by non-covalent interactions such as hydrogen bonds, hydrophobic forces, ionic forces or by formation of reversible covalent bond. The association between protease and reversible inhibitor is not permanent, because it can be simply reversed, for example by dialysis. The reversible inhibitors are further classified as competitive, non-competitive or un-competitive inhibitors. In the competitive case, the inhibitor competes with substrate for binding; in the non-competitive case, the inhibitor binds to both the free enzyme and the enzyme - substrate complex and in the un-competitive case, the inhibitor binds only to the enzyme - substrate complex. Irreversible inhibitors have two steps of inhibition: the first reversible and the second irreversible. At the irreversible step, which permanently modifies the protease, they form a covalent bond with the catalytic protease residue. The functional group that chemically modifies the enzyme is called warhead. Irreversible inhibitors are considered less safe due to the warhead reactivity, which can engage in toxic, non-specific interactions with non-targeted residues on proteins or DNA and RNA molecules. In oppose, the beneficial consequences of irreversible covalent modification are prolonged drug effect and lower doses needed in therapy. Several irreversible inhibitors are drugs with excellent safety profiles, which shows

that irreversible modification remains a viable option of protease inhibition (Eisenthal et al., 2007; Strelow, 2017). Few examples of widely used and successful therapeutics are aspirin, penicillin, clopidogrel and omeprazole (Singh et al., 2011; Vita, 2021).

It has been shown recently that alkyne functional group, which was long considered inert, is capable of forming an irreversible covalent bond with the nucleophilic residue in the enzyme active site (Ekkebus et al., 2013). The reaction proceeds despite its slow rate because it is governed by the local forces in the catalytic site. The alkyne-based small molecule inhibitors should thereby retain selectivity against the targeted enzyme, provided that their molecular backbone remains unchanged, while their off-target properties should be diminished due to the inertness of the alkyne warhead elsewhere. The alkyne-based inhibitors thus present a novel option for targeting enzymes by the means of irreversible inactivation (Kim et al., 2021).

1.1.4 Cysteine cathepsins

Cysteine cathepsins are endosomal papain-like proteases. In humans there are 11 members of the family: cathepsins V, K, L, S, F, B, C, O, H, X and W. They are all endopeptidases, apart from cathepsins X and C, which are carboxymonopeptidase and aminodipeptidase, respectively, and cathepsins H and B which can act as both endopeptidases and aminopeptidase and carboxydipeptidase, respectively (V. Turk et al., 2012). They play redundant roles in protein digestion in late endosomal compartments and lysosomes. Sometimes they are involved in processing of proteins by limited proteolysis such as bone remodeling by cathepsin K (Garnero et al., 1998), processing of Ii by cathepsins S and V (Unanue et al., 2016), neuro prohormone processing by cathepsins V and L (Funkelstein et al., 2008, 2012), release of angiostatin-like fragments from plasminogen by cathepsin V (Puzer et al., 2008) and thyroid pro-hormone processing by various cathepsins (Di Jeso & Arvan, 2016). In the last decade it was established that cathepsins, at certain conditions, localize in other cellular compartments besides the endosomes where they process their substrates, for example in the extracellular space (Vizovišek et al., 2019), nucleus (Duncan et al., 2008) and cytosol (Yadati et al., 2020).

Specificity of cysteine cathepsin processing was studied by using peptidyl libraries or protein digestion samples as cathepsin substrates (Biniossek et al., 2011; Puzer et al., 2004; Vidmar et al., 2017; Vizovišek et al., 2015). These works showed that the S2 subsite is their main specificity determinant, which adapts well the bulkier hydrophobic residues or the aromatic residues. Apart from the S2, the subsites S1 and S1' also confer some selectivity to cathepsins, however it was not clear which residues are preferred at these sites because different groups reported different amino acid preferences. This discrepancy is likely due to the differences in the sample preparations or the amino acid sequence limitations of the samples. Crystal structures of cysteine cathepsins with small molecule substrate mimicking inhibitors confirmed that S2, S1 and S1' are the only well-defined subsites where the substrates can interact with the cathepsin surface with both main-chain and side-chain interactions. The exopeptidases contain additional structural features, such as occluding loop in cathepsin B, which limit the access of a substrate to the active site and hence endow them with the exopeptidase activity (V. Turk et al., 2012). Due to its presence, the cathepsin B contains additional specificity determinant, the preference for Gly at S3' subsite (Biniossek et al., 2011).

1.1.4.1 Cysteine cathepsins as drug targets

Cysteine cathepsins emerged as a promising drug targets once their specific physiological roles and implications in pathological states were established, such as immune and cardio-

vascular disorders, neurodegeneration, inflammation and cancer progression (V. Turk et al., 2004; Yadati et al., 2020). Specific inhibition of homologous proteases such as cysteine cathepsin endopeptidases is challenging because there are no distinctive structural features that can be used for selective targeting. However, many attempts have been made in this direction. The most studied was cathepsin K because of its distinctive processing of bone matrix, which made it a promising target for the treatment of osteoporosis. Several groups developed inhibitors with reasonably good specificity profiles, and Odanacatib (ODN) reached phase III of clinical trials (Gauthier et al., 2008; Lu et al., 2018). The specific inhibitors of cathepsin endopeptidases S (Fuchs et al., 2020), L (Dana & Pathak, 2020) and cathepsin exopeptidase B (Y.-Y. Li et al., 2017) have also been developed. In 2019, there were several cathepsin K and S inhibitors in clinical trials, one inhibitor of cathepsin B and none of cathepsins L and V (Cianni et al., 2019).

The selectivity of an inhibitor determined in a laboratory experiment does not always translate to the desired inhibition when it is administered to the patients. For example, the ODN is considered a highly selective compound among specific cathepsin inhibitors. Its IC_{50} against cathepsins K and S, the two that are inhibited most strongly, are 0.2 nM and 60 nM, respectively, so it is selective over cathepsin S by a factor of 300. However, when patients were administered 50 mg of ODN once a week, its plasma levels stayed above 100 nM for the whole week before the next dose was taken, and its high volume of distribution suggests its partitioning in body tissues (Stone et al., 2019; Zajic et al., 2016). At such conditions, the inhibition of cathepsin S could not be excluded. Thus, there is still a large need for potent and selective cysteine cathepsin inhibitors, which may one day become an important group of therapeutics for different disease states.

1.1.4.1.1 Cysteine cathepsins as drug targets in SARS-CoV-2

In an effort to discover novel drug candidates for treatment of SARS-CoV-2 infection, large-scale X-ray screening against viral protease M^{pro} was performed (Günther et al., 2021). Compounds in screening were drugs that already entered clinical trials or were already approved, so they are considered safe to humans. Calpeptin, an aldehyde-based reversible calpain inhibitor was among seven compounds that covalently attached to the catalytic Cys residue of the M^{pro} enzyme. It was observed that calpeptin is also a strong inhibitor of human cysteine cathepsins, which propagate viral entry by processing of Spike-protein (S-protein) in endolysosomes, the route alternative to these of transmembrane serine protease TMPRSS2, which process S-protein upon its binding to ACE2 receptor on the cell surface (Jackson et al., 2022; Juibari et al., 2022; Yang et al., 2022). Moreover, it has been shown that calpeptin suppresses viral activation in cell-based assay in concentrations below 100 nM, although it only inhibits M^{pro} enzyme in micromolar range (Mediouni et al., 2022). Hence, the actions of calpeptin could not be attributed solely to inhibition of M^{pro}, suggesting its actions against targets of host origin, namely cathepsin L, and possibly also of other cathepsins. Roles in the viral activation for cysteine cathepsins have been demonstrated for other viruses such as SARS-CoV (Bosch et al., 2008), Ebola (Gnirß et al., 2012), MERS-CoV and human papilloma type 16 virus (Scarcella et al., 2022).

The advantage of targeting human molecular species in SARS-CoV-2 infection is their negligible variability in the human population, whereas viral drug and vaccine targets are prone to mutations, which gives rise to resistant mutant strains. Hence, cysteine cathepsins can be regarded as a promising target for not just SARS-CoV-2, but also for treatment of viral infections in general.

1.2 Studies of Protease Substrates and Inhibitors

In the last few decades, several methods evolved that boost the investigation of protease associations with their substrates and inhibitors. These include macromolecular crystallography, enzyme kinetic studies and mass spectrometry techniques.

1.2.1 Macromolecular crystallography

Macromolecular crystallography is a method used to determine the 3-D atomic structure of the molecules, like proteins, in the crystalline form. Molecules are first crystallized. The smallest repeating unit of the crystal is called unit cell. The smallest part which can be used to reconstruct the unit cell by using the crystal symmetry operations is called asymmetric unit.

1.2.1.1 Protein crystallization

Protein crystallization is explained by the phase diagram, as shown in Figure 1. The state in which the protein concentration in the sample is higher than its solubility is called the supersaturation state. Supersaturation state is divided into metastable zone, precipitation zone and liable (or crystallization) zone. In the metastable zone, the supersaturation is too low, and the protein nuclei that form are unstable and usually quickly dissolve. In the precipitation zone, the supersaturation is too large and the protein molecules will precipitate rather than form crystals. The stable nuclei grow only in the liable zone.

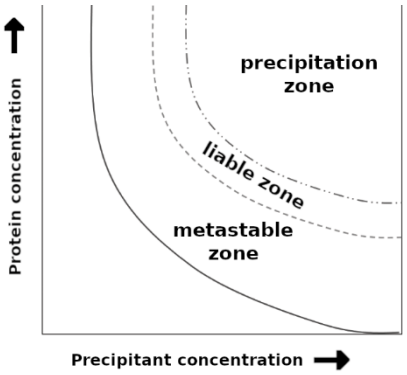


Figure 1: Crystallization phase diagram

After the nuclei are formed, the crystal growth will take place, until the supersaturation state is exhausted and the equilibrium between crystal and liquid system is reached (Rupp, 2015). The conditions in which protein crystallizes depend heavily on the composition of the precipitate solution, protein concentration and the temperature. The optimal conditions for crystal growth are determined experimentally, by testing many different varieties of these variables. In practice, the concentrated protein solution is mixed with the crystallization solution (or precipitate solution or simply precipitate) in a droplet and left sealed against a much bigger volume of precipitate (reservoir). The evaporation of water from the droplet to the reservoir is driven by higher vapor pressure in the droplet. Hence, the protein concentration in droplet slowly grows toward the supersaturation state (Cheraghian Radi et al., 2021; Forsythe et al., 2002).

1.2.1.2 X-ray scattering from the crystal

The X-rays are suitable for crystal structure determination because they are scattered from the electrons so that their energy (frequency) remains the same, the effect known as the elastic scattering. Their wavelength is around 1 Å, which is in the same range as the distances between atoms in the proteins, hence the scattered waves interfere with each other. When the scattered waves interfere constructively (which means that they add up in phase), their signal will be enhanced across the millions of millions of protein molecules in the crystal. The Braggs model states that constructive interference between scattered waves occurs from evenly distanced "planes" so that the Bragg's equation is fulfilled:

$$n\lambda = 2\,\mathrm{d}\,sin\,\theta \tag{1.1}$$

where n is an integer, λ is the X-ray wavelength, d is the distance between the planes and θ is the value of the incident and diffracted angle (Bragg & Bragg, 1915). The signal from the scattering planes is amplified across the crystal only at certain directions, which are specified by Miller indices (h, k, l). In reality, protein atoms do not lie on planes, but the vectors of the scattered waves from the atoms, which interfere constructively at the direction of Miller indices, appear in the same way. The signal from each diffracting plane, called the reflection, is collected at the detector. The position and identity of the reflections are calculated once the unit cell parameters (axes a, b, c, and their angles) and its symmetry (point group or space group) are obtained. The collection of the reflections on one image, which is obtained with one exposure event, is called the diffraction pattern, and the set of all images is called the dataset (Neil W. Ashcroft, 1976).

1.2.1.3 Euler description, Fourier transform and structure factors

Mathematically, the scattered waves are described as vectors in complex plane with the Euler's formula:

$$e^{i\alpha} = cos\alpha + isin\alpha \tag{1.2}$$

where the unknowns are the amplitude of the wave and its phase. In crystallographic terms, the scattered waves are related to structure factors. Each reflection (h, k, l) holds the information about the amplitude of the associated structure factor ($F_{h,\,k,\,l}$ or shortly F_h), which is proportional to the intensity of the reflection. The underlying mathematical operation that treats the reflections as the representation of the crystal electron density in the diffracting space is linked by Fourier transform:

$$F(h) = \int_V \rho(x) \times e^{2\pi i h x}\, dV \tag{1.3}$$

where F(h) is the structure factor, ρ(x) is the electron density at vector x (x stands shortly for fractional coordinates x, y, z of the unit cell a, b, c), the vector product hx represents the contribution of scattering object at the vector x to the structure factor F_h and the V represents the unit cell volume. The Fourier transform is inverse operation, so it can also be stated as:

$$\rho(x) = \frac{1}{V}\sum_h F(h) \times e^{-2\pi i h x} \tag{1.4}$$

Hence, if the integral in the first equation goes over the whole unit cell, then all the atoms in the unit cell contribute to the value of each structure factor. Vice versa, the second equation states that the electron density at every position is obtained by the summation of every structure factor

Considering that electron density is calculated by the summation of each structure factor, the goal in the diffraction experiment is to collect as many reflections as possible and extract the amplitudes of each structure factor from the intensity of the reflections. This is achieved by rotating the crystal (and hence its diffracting space) so that each reflection eventually fulfils the Bragg's conditions. During the experiment, the crystal is exposed to X-rays after every rotation event, which is usually in the range of 0.2 - 1 degree. The higher the symmetry of the crystal and hence its diffraction pattern, the lesser the portion of the diffraction space that needs to be sampled in order to collect all the reflections. The imperfections in the crystal lattice and thermal motion of atoms in the crystal (both described with atomic displacement parameter, called the B-factor) impose a limit to where the diffraction reaches. This limit is known as the resolution (covered in Rupp, 2009).

1.2.1.4 The Phase problem

In the X-ray experiment, only amplitudes of the structure factors are measured, whereas the phase information is lost. This is known as the phase problem. There are different ways to obtain phases of the structure factors. These include single or multiple isomorphous replacement (SIR or MIR), single or multi-wavelength anomalous diffraction (SAD or MAD) and molecular replacement. In structures with very high resolution the direct phasing methods can also be used. In phasing, the goal is not to determine precise values of the phases, but to find the initial set of phases for calculating the electron density map that is good enough to enable modelling of protein residues into it and thereby proceed with structure determination.

1.2.1.4.1 Isomorphous replacement

The isomorphous replacement is a technique where heavy atom is introduced into a protein structure. Position of heavy atoms in the unit cell can be determined from the difference Patterson map, which is an auto-correlation function of the electron density with its inverse and uses only structure factor amplitudes to determine the relative distances between atoms in the unit cell. The Patterson map of the protein atoms is too crowded; however, the introduced heavy atom scatterers stand out in the map and can thus be assigned their relative location in the unit cell. By applying symmetry of the unit cell, their absolute location in the unit cell (and hence their phases) can be determined. Their phases, together with measured amplitudes of the native crystal (without heavy atoms) and crystal with heavy atoms, are used to calculate phases of all structure factors (covered in Rupp, 2009).

1.2.1.4.2 Anomalous diffraction

When the X-rays are absorbed by the atoms and cause electronic transitions in their orbitals, they lose their centrosymmetry as the scatterers. This breaks the scattering symmetry between Friedel pairs (reflections with opposite Miller indexes, F_h and F_{-h}), whose amplitudes are equal and their phases opposite in the absence of anomalous signal. The atoms absorb X-rays only in a specific region of the wavelength spectrum. The experiment is performed by choosing the wavelength that corresponds to the absorption peak of the atom whose anomalous diffraction is being measured (SAD) or combining this data with the data from other wavelengths where anomalous scattering is weaker but still present (MAD). For protein structure determination, the most important anomalous

scatterers are metals, selenium atoms (for recombinant proteins with incorporated selenomethionines) and sulfur atoms. The position of the anomalous scatterer and thereon protein phases are determined in a similar way as with isomorphous replacement (covered in Rupp, 2009).

1.2.1.4.3 Molecular replacement

The molecular replacement uses phases of known protein structure (called the model), which is similar to the structure of the investigated protein. To extract phases, the model molecule first needs to be orientated in the same way and placed in the same cell as the investigated protein, which is done with the help of the Patterson transformation. The similarity of model structure with the investigated protein can be guessed from their sequence homology. As a rule of thumb, between 30 - 40 % homology is usually enough to obtain the initial set of phases. Molecular replacement is the most frequently used technique in phasing due to its simplicity and the continuous supply of solved protein structures in the Protein Data Bank (PDB) (Vagin & Teplyakov, 1997).

1.2.1.4.4 Direct methods

Direct methods are not used on their own for phasing purposes because they do not carry enough information, but they can be used to provide additional data in combination with other phasing techniques. Direct methods exploit the high probability of phase correlation between certain triplets or quartets of structure factors that exhibit high diffraction peaks (high intensity). Mathematically, they are described with the tangent formula and Sayre equation. In real space, this can be viewed as high probability that atoms are concentrated at intersections of their diffracting planes (Hai-Fu, 1998).

1.2.1.5 Electron density map and model building

When the phases are obtained, they are used to calculate the electron density map using the Fourier transform. When the map is not clear, then the phases need to be first refined and then map re-calculated. When the electron density takes a recognizable shape of the protein structure, the protein molecule can be built into the electron density map, which is called the model building. Model building leans on recognizable features of the proteins, such as their main chain direction, shape of each amino acid residue, secondary structural elements (α-helixes and β-sheets), solvent regions and primary sequence of the investigated protein to locate their corresponding electron density in the map. The electron density map is calculated with model phases and experimentally determined amplitudes, in which the amplitude term is modified so that it gives the best representation of the true electron density map:

$$2mFo - DFc \tag{1.5}$$

where Fo and Fc are the observed and calculated amplitudes, respectively (Fc is calculated from the model coordinates), m is figure of merit and represents the estimation of the phase error and D is the Luzzati factor, which represents the estimated model errors (Pannu & Read, 1996). Usually, another map is calculated, called the difference map:

$$mFo - DFc \tag{1.6}$$

which exposes areas where the model is not consistent with the data. Model building is done manually or automatically, using crystallographic software platforms CCP4 (Winn et al., 2011), Phenix (Liebschner et al., 2019) or MAIN software (D. Turk, 2013).

1.2.1.6 Model refinement

When the model is built, its atoms are manipulated in order to improve their fit to the electron density map. This includes modification of their coordinates and modification of their B-factors. The phases of the corrected model are then used to calculate new electron density map, and the cycle is then again repeated several times. This is called the model refinement. In refinement, all available knowledge about the crystallographic terms (amplitudes, B-factors) and stereo-chemistry of the proteins is used to build the most reliable model. The chemical terms, which are simply called the restraints, are bond lengths, bond angles, planarity, Van Der Wals and electrostatic forces, and improper and proper dihedral angles. Refinement software like Refmac (Murshudov et al., 2011), phenix.refine (Afonine et al., 2012) or MAIN (D. Turk, 2013) use crystallographic terms and restraints in the so-called target function, which search for the lowest energy of the system by changing the model parameters (atomic coordinates and their B-factors). The general fit of the model to the experimental data is evaluated by the R-factor:

$$R = \left| \frac{Fo - Fc}{Fo} \right| \tag{1.7}$$

where Fo and Fc are measured and calculated amplitudes, respectively. The smallest the R-factor, the better the agreement between the model and the data. When it is not clear how to correct the model further and the R-factor is low enough and stops changing after each refinement cycle, the structure is likely finished (Murshudov et al., 2011; Pražnikar & Turk, 2014).

1.2.1.7 Model validation

The last step in structure determination is its validation. In general, validation reports deviation of all the chemistry and crystallographic terms from the expected values, and they should be checked at each step of the structure determination. These include bond lengths and angles, close contacts, clashes, rotameric outliers and B-factors. When certain term deviates, then its weight in the target function may need to be adjusted. The most powerful validation tools are however the terms that were not used during the refinement. Historically, the most important is the R-free factor, which is calculated the same as the common R-factor but uses only 5 % of the reflections, which were set aside and not used in model building and refinement. This is called the cross-validation. Hence, if the model really agrees with the data, the R-free value should be close to the R-factor value (Brünger, 1997). A newer approach is the calculation of R-kick, in which all the reflections are used in structure determination and in cross-validation. The introduced small changes (kicks) of the atomic coordinates break the dependence of the model to the restraints imposed during the refinement, so the model is better evaluated against the data (Pražnikar & Turk, 2014). An important validation tool are the main chain dihedral angles phi and psi, which mostly exist in certain favorable regions. These regions are presented in the Ramachandran plot (Kleywegt, 2000; Richardson et al., 2013).

1.3 Enzyme Kinetic Studies

Enzyme kinetics is a study of rates of chemical reaction between an enzyme and its substrates. Parameters obtained in the experiment are valid only at the conditions tested, because enzyme activity depends on the reaction conditions, which are given by buffer composition and concentration of its components. These include pH value, ionic strength,

salt concentration, additives like metal ions, detergents or chelators and temperature. The principal formula for the enzyme catalysis is given as:

$$E + S \leftrightarrow ES \leftrightarrow EP \leftrightarrow E + P \tag{1.8}$$

where E is the free (un-bound) enzyme, S is the free substrate, P is the product and ES and EP are enzyme-substrate and enzyme-product complexes, respectively. The arrows indicate the direction of the reaction (left toward the substrates and right towards the products). For each arrow, a rate constant k exists, which tells how quickly the underlying reaction occurs (Srinivasan, 2021). Because the rate constants cannot be measured or calculated, various attempts were made to derive mathematical models with variables that can be measured in an experiment.

1.3.1 Michaelis - Menten model

The most widely applicable and used is the Michaelis - Menten (MM) model, which can be stated as:

$$v_0 = \frac{v_{max} \times [S]}{[S] + K_m} \tag{1.9}$$

where v_0 is the initial reaction velocity, v_{max} is the maximal reaction velocity when all enzyme is in the form of ES complex, [S] is the substrate concentration and K_m is the Michaelis - Menten constant; the brackets stand for concentration of the component that is listed within the bracket (Johnson & Goody, 2011). MM model requires that the measurements are performed under steady state conditions, which assume that the product formation and substrate depletion do not affect the speed of the reaction. In practice, these conditions are met at the initial, linear portion of the reaction progression curve, from where the v_0 values are calculated. The v_{max} and K_m values can be obtained from the plotted values of v_0 against different substrate concentration, where v_{max} is the asymptote at the velocity axis and K_m is the concentration of substrate where the reaction speed equals one half of the v_{max}. An important parameter is also the rate constant k_{cat}, which describes the reaction step ES $\rightarrow$ EP. It is also called the turnover number and can be directly calculated as:

$$k_{cat} = \frac{v_{max}}{E_{total}} \tag{1.10}$$

where E_{total} is the total enzyme concentration in the assay. Because K_m is a measure of affinity between the substrate and enzyme and k_{cat} is a measure of efficiency of a product formation, the substrates of a given protease are widely compared based on their k_{cat} / K_m ratio, also known as the specificity constant (Srinivasan, 2021).

1.3.2 Enzyme activity in presence of inhibitor

1.3.2.1 Reversible inhibitors

The general formula for reversible inhibition is written as:

$$E + I \leftrightarrow EI \tag{1.11}$$

where E is the free (un-bound) enzyme, I is the free inhibitor and EI is the enzyme - inhibitor complex. The rate constants that describe the association and disassociation of

the EI complex are known as k_{on} and k_{off}, respectively, and their ratio is known as the dissociation constant, K_i:

$$\frac{k_{off}}{k_{on}} = \frac{[E] \times [I]}{[EI]} = K_i \tag{1.12}$$

Inhibitors with higher affinities (recognized by lower K_i values) efficiently block enzyme activity at lower concentrations. In general, inhibitors with K_i values in micromolar range (µM) are considered weak, in nanomolar range (nM) strong and in picomolar range (pM) very strong. Another important and widely used parameter for classifying inhibitors is the IC_{50} value, which is defined as the concentration of inhibitor that reduces enzyme activity in a given assay by one half. For drug design purposes, the inhibition at least in nM range is desirable, although there are drugs that act in µM range, for example etoposide (Kingma et al., 1999).

Usually, the inhibitor properties are measured in an assay where initial reaction velocities, v_o, are measured in the presence of different inhibitor and substrate concentrations. Based on the reaction conditions, two types of inhibition can be observed: classic and tight-binding. Classic inhibition is observed under the experimental conditions where [I] and K_i values are (much) higher than that of the [E] value. In these cases, the observed K_i value approaches the value of IC_{50}. The type of inhibitor (competitive, non-competitive or un-competitive) can be established from the MM plot obtained in the presence of several different inhibitor concentrations: in competitive case, the v_{max} remains the same and the K_m value appears higher; in non-competitive case, v_{max} appears lower and the K_m value remains the same; and in un-competitive case, both v_{max} and K_m values appear lower. If inhibitor type cannot be established, the mixed model can be used which includes competitive, non-competitive and un-competitive inhibition as special cases.

Tight-binding is observed when K_i value is equal to or smaller than that of the enzyme concentration. Hence, the inhibitor concentrations used in the assay are also in the same concentration range. At these conditions, the inhibitor is readily bound to the enzyme, and the free inhibitor significantly affects the equilibria. Here, the true K_i values can be (much) smaller than the observed IC_{50} values. Tight-binding is accounted for in the Morrison equation for reversible inhibition (Morrison, 1969).

1.3.2.2 Irreversible inhibition

General representation of irreversible inhibition is:

$$E + I \leftrightarrow EI \rightarrow EI^* \tag{1.13}$$

where E is the free (un-bound) enzyme, I is the free inhibitor, EI is the reversible enzyme - inhibitor complex and EI* is the irreversible enzyme - inhibitor complex. The rate constant that describes the irreversible step EI → EI* is known as k_{inact}, which describes the maximal rate of enzyme inactivation. To account for binding potency at the reversible step and efficiency of covalent bond formation at the second step, the ratio k_{inact}/K_i is used to present overall efficiency of covalent inhibitors. Covalent inhibitors can be recognized and differed from reversible inhibitors by time-dependent inactivation of the protease activity (Strelow, 2017).

1.3.2.3 Molecular reporters

Substrate and inhibitor assays require that enzyme activity is monitored in real time. One option is to attach special molecular tags or reporters with spectroscopic properties to the molecular backbone of the substrates. Once they are cleaved away by the protease, their absorbance or fluorescence (the emission of the absorbed light) changes and the change is measured with a spectrometer. The strength of their signal is proportional to the number of substrate molecules converted to products, which is proportional to the speed of the underlying reaction. Few examples of spectroscopic tags include 4-nitrophenol (once cleaved, absorbs light at 400 nm), 7-amino-4-methylcoumarine (AMC; once cleaved, it is excited at 360 nm and emits light at 440 nm) or fluorescence resonance energy transfer pairs like DABCYL and EDANS (once cleaved, EDANS will fluoresce because DABCYL will no longer quench its emitted light).

1.4 Mass Spectrometry

In proteomics, mass spectrometry (MS) is used to identify molecular species in the investigated sample by measuring their mass-to-charge (m/z) ratios. MS has a broad range of applications, from investigation of single proteins to analysis of complex biological samples, like protein expression profiles in cell lysates. It is performed in a dedicated instrument, which consists of three main units: the ionizer, analyzer and the detector. The ionizer converts molecules into gas-phased ions, which can then enter into the mass analyzer. Most frequently used techniques for ionization are electrospray ionization (ESI) and matrix-assisted laser desorption/ionization (MALDI). The mass analyzer separates ions based on their m/z ratios and sort them for the detection. The separation is usually achieved with quadrupoles, time of flight or ion trap techniques. Two analyzers can be combined in one configuration, known as tandem MS (MS/MS), in which ions from the first analyzer are selected and fragmented inside the mass spectrometer to smaller ions, which are then measured. This configuration has an improved accuracy and sensibility. Detector converts ion currents into measurable signals (Han et al., 2008). Protein identity is inferred from the mass spectra of recorded peptides (also called peptide mass fingerprinting), with the help of protein databases like MASCOT (Perkins et al., 1999) or SEQUEST (Pruitt et al., 2012). An integral part of MS is also preparation of the sample, which needs to be enriched with molecules of interest in order to record as many relevant fragments as possible and reduce signals from other molecules in the sample, which contribute to the noise.

1.4.1 Mass spectrometry for protease studies

MS is a powerful technique for studying the association of proteases with substrates or inhibitors due to its ability to identify protease substrates, to determine their exact cleavage sites and to confirm if an inhibitor is attached to the protease. Due to its unique informative capacity, it was used to determine specificity profiles of numerous proteases (Gupta et al., 2010; S. Y. Luo et al., 2019; B. E. Turk et al., 2001; Vidmar et al., 2017; Vizovišek et al., 2015). In classic derivation of MS, known as the “bottom up” approach, protein species in the sample are digested to peptides with well characterized protease, like trypsin, chymotrypsin or Glu-C. Digestion is crucial in order to break proteins into smaller fragments, which are more susceptible to ionization. These peptides are called tryptic peptides (the term "tryptic" comes from treatment of sample with trypsin). Then, protease of interest is added to the sample to perform its own cleavages on the digested proteins and (poly)peptides in the sample. Usually, tryptic peptides are chemically labelled in a way

that allows separation of their cleavage sites from cleavages of the investigated protease in subsequent purification step, which is usually achieved by high-performance liquid chromatography (HPLC) techniques. Different chromatography techniques can be applied, based on chemical properties of the incorporated tags, for example reverse-phase HPLC, which is suitable for separation of peptides with hydrophobic tags and ion-exchange HPLC for separation of peptides with charged tags. Using this method, a large number of cleavage sites can be detected, which reveals positional preferences of the investigated protease. The downside is that sequencing of the protein and peptide species in the sample is not absolute, due to losses in the process of sample preparation or measurement, hence certain amount of cleavage sites is not recovered (Biniossek et al., 2011; Gupta et al., 2010; Han et al., 2008; Lapek et al., 2019; H. Luo et al., 2019; Staes et al., 2008; Vidmar et al., 2017; Vizovišek et al., 2015).

In the opposite approach, called the "top down" proteomics, the intact proteins are analyzed. The analysis is usually performed in MS/MS configuration in order to recover full peptide fingerprint of the analyzed protein and hence to determine all cleavage sites (Catherman et al., 2014). It is also appropriate for the detection and determination of the exact site and type of enzyme modifications, including covalent modification of the protease by the inhibitor. It is however less suited for complex samples and high-throughput purposes.

Chapter 2

The Aim of the Work

In a recent work, the proteomic study performed at native conditions delivered over 30,000 cleavages of cysteine cathepsins V, L, S, K, F, and B (done by Kris Gevaert group). Cleavage data was used as an input for statistical analysis, developed for this purpose. It revealed the substrate positions with non-normal distribution of residues that contributed to specificity, called heterogeneous positions, in contrast to positions with normal distribution of residues carrying no specific information, called homogeneous positions. The cleaved sequences were then clustered using heterogeneous substrate positions. They revealed prevailing residues or prevailing type of residues for most clusters (done by Dr. Livija Tušar).

In this work, cathepsins specificity determinants were studied by crystal structure determination of cathepsin V with a selected set of peptides that represented the variety of all cathepsin V substrate clusters. In addition, these peptides were cleaved by native cathepsins V, K, and L, and their cleavages compared with the cleavages observed in protein samples. As a part of our collaboration with other research groups, we assisted in the evaluation of novel cathepsin inhibitors by crystallizing cathepsin K with alkyne-based inhibitor and by characterizing calpeptin and alike compounds as dual inhibitors of SARS-CoV-2 major protease (M^{pro}) and human cathepsins.

Chapter 3

Hypotheses

I. Using a minimal set of representative peptides, we can explain their processing by cathepsin V.
II. Using a minimal set of representative peptides from each substrate cluster, we can expose the structural features that govern specificity of cathepsin V.
III. Peptides are not always a reliable model of protein substrates.
IV. Alkyne functional group can be used as a latent electrophile for the inhibition of cathepsin K.
V. Calpeptin and alike compounds inhibit human cathepsins at concentrations below those required for inhibition of SARS-CoV-2 M^{pro} protease.

Chapter 4

Materials and methods

4.1 Protein Expression

Recombinant cathepsins K, L, and V were expressed in Pichia pastoris strain GS115. Gene for recombinant human procathepsin K was obtained from Deutsche Ressourcenzentrum für Genomforschung and cloned into pPIC9 vector between EcoRI/NotI restriction sites without modifications. Genes for human procathepsins V and L were mutated at sites 179 (Asn to Gln) and 110 (Thr to Ala), respectively (active cathepsin numbering) to prevent cathepsin glycosylation. For crystallization purposes, catalytic residue of cathepsin V was mutated at site 25 (Cys to Ser or Ala), hereafter called cathepsin V C25S/A. Both cathepsin L and cathepsin V constructs had six His residues attached to their N-terminal for purification purposes.

Recombinant cathepsins were expressed according to the Invitrogen Pichia Expression kit (Invitrogen, K1710-01). pPIC9 vectors, containing cathepsin genes, were introduced into Pichia pastoris by electroporation. Pichia were grown for 24 - 48 hours at 30 °C in BMGY medium (1 % yeast extract, 2 % peptone, 100 mM KH_2PO_4/K_2HPO_4, pH 6, 13.4 g/L yeast nitrogen base, 0.4 mg/L biotin and 1 % glycerol). After Pichia reached an exponential phase of growth, indicated by the value of optical density measured at 600 nm between 2 and 6, it was centrifuged and resuspended in BMMY media (0.5 % methanol instead of glycerol) to start the protein expression. The best performing colonies were selected in small- scale screening of up to twenty Pichia colonies in 10 - 15 mL of expression media in 50 mL bioreactors. The expression took 4 days at 30 °C with feeding interval of 1 % MeOH per day. To estimate their protein yields, the media was either put through nickel affinity resins (Ni^{2+}-NTA) in order to enrich the cathepsins with incorporated His-tags or proteins were precipitated from the media using trifluoroacetic acid (TFA). The samples were then analysed with SDS-PAGE and the amount of expressed cathepsins between the colonies assessed with visual inspection. The best performing colony was selected for large-scale expression, which was performed in a 10 - 20 2-L erlenmeyer flask, containing 400 - 500 mL of BMMY medium. Procathepsin K expression was boosted by the addition of antifoam 204 (Sigma, A8311) at final concentration 0.01 % (v/v), which improved the yield by approximately 2.5-fold. After 4 days, supernatant was collected, concentrated to approximately 300 mL and used for later stages of protein isolation.

4.2 Protein Activation and Purification

4.2.1 Procathepsin K.

Concentrated procathepsin media from the expression was diluted at 1:1 ratio with 20 mM HEPES, pH 7.1. The ion-exchanger SP-sepharose FF (GE Healthcare, 17-079-01) was added to the sample and the sample was left shaking overnight at 6 °C. Procathepsin was eluted from the exchanger with 400 mM NaCl. For procathepsin activation, the sample was diluted at 1:1 ratio with activation buffer (100 mM NaOAc, pH 4, 10 mM DTT and 40 μg/mL pepsin (Sigma, P8667)) and left incubating at 37 °C until all procathepsin was activated. The optimal activation time was determined by taking reaction aliquots at different time points and analyzed by SDS-PAGE (Figure 2). For procathepsin K, the optimal time was 45 min. To stop the activation, the pH of the sample was raised to 5.5 with 1 M TRIS (pH 8.5). Activated cathepsin was then purified on another IEX chromatography, performed on Äkta Express system (GE Healthcare), using MONO S 5/50 column (GE Healthcare, 17-5168-01). Cathepsin eluted from the column with 1 M NaCl. To protect the cathepsin from autolysis, the 10-fold molar excess of reversible inhibitor MMTS was added to the sample. Protein was then desalted on HiTrap 5 mL column (GE Healthcare) to the final buffer (50 mM NaOAc, pH 5.5, 50 mM NaCl). Active cathepsin concentration was determined by titration with potent cathepsin inhibitor E-64 based on previously described protocols (Borišek et al., 2015; Rozman-Pungerčar et al., 2003). Activated and purified cathepsin was stored at -80 °C until further use.

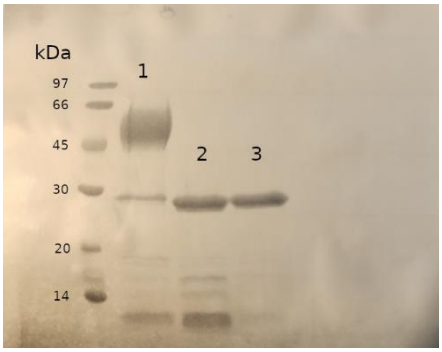


Figure 2: SDS-PAGE of procathepsin K activation. Band between 45 and 66 kDa corresponds to procathepsin K. Bands at 30 kDa correspond to activated cathepsin K. LMW protein ladder (Lot. Num. 17-0446-01, Cytiva) was used as a molecular weight marker. Lane 1, sample after 5 min activation. Lane 2, sample after 30 min activation. Lane 3, sample after the end of activation.

4.2.2 Procathepsin L, Procathepsin V and Procathepsins V C25S/A

Concentrated procathepsin media from the expression was dialysed three times against 30 mM TRIS, pH 7.5, 400 mM NaCl. Procathepsins were isolated by Ni^{2+}-NTA, performed on Äkta Express system, using 5 mL HisTrap column (GE Healthcare, 17-5247-01). Procathepsins were eluted from the column with 300 mM imidazole. Immediately after, the sample was applied to gel filtration chromatography, using HiPrep 26/60 Sephacryl S-200 column (GE Healthcare, 17119501) in order to remove imidazole from the sample. Procathepsin sample was concentrated and stored at -80 °C until further use.

4.2.2.1 Procathepsin V

Procathepsin V was activated by auto-activation. The optimal activation time was determined in the same way as for procathepsin K (described in 4.2.1). The procathepsin

sample was diluted at 1 : 1 ratio with activation buffer (100 mM NaOAc, pH 4, 10 mM DTT) and left at 37 °C for 2 hours, . The activation was stopped by raising the pH to 5.5 by exchanging the buffer in MilliporeSigma Amicon stirred cells (Thermo Fisher Scientific). Cathepsin was then applied to SP-sepharose FF (GE Healthcare, 17-079-01), from where it was eluted with 400 mM NaCl. Cathepsin was then blocked with MMTS and titrated with E-64, as described under cathepsin K section (4.2.1).

4.2.2.2 Procathepsin L

Procathepsin L was activated by auto-activation. The optimal incubation time was determined in the same way as for procathepsin K (described in 4.2.1). The procathepsin sample was diluted at 1 : 1 ratio with activation buffer (100 mM NaOAc, pH 4.5, 10 mM DTT, 50 mM NaCl) and left first at 37 °C for 1.5 hours and then at room temperature for additional 2.5 hours. The sample was then dialysed overnight to 50 mM NaOAc buffer, pH 5.5. Next day, it was purified by IEX chromatography on Äkta Express system using MONO Q 5/50 column, from where it eluted with 400 mM NaCl. Cathepsin was then blocked with MMTS and titrated with E-64, as described under cathepsin K section (4.2.1).

4.2.2.3 Procathepsins V C25S/A

Activation and purification of cathepsin V mutants were based on a previously published procedure (Sosnowski, Piotr, 2016). Procathepsin sample was dialyzed to activation buffer (100 mM NaOAc, pH 5, 1 mM EDTA, 5 mM DTT). The activation of procathepsin V mutants was initiated by the addition of 5 % (n/n) of activated cathepsin L. The mixture was left overnight at room temperature to completely digest the “pro” part of the procathepsin V C25S/A molecule. The activation was stopped by blocking cathepsin L activity the next day by the addition of inhibitor E-64 to the sample. The sample was then applied to SP-sepharose FF (GE Healthcare, 17-079-01) from where it was eluted with 400 mM NaCl, and then dialyzed overnight to crystallization buffer (20 mM NaOAC, pH 4.5, 10 mM NaCl, 1 mM DTT and 5 % glycerol). Cathepsin was then concentrated to 40 mg / mL with Amicon Ultra devices (cut-off 10 kDa) and used immediately for crystallization, or stored at -80 °C until further use.

4.3 Peptide Purification

Peptide powders from crude synthesis were obtained from the group of Kris Gevaert (Ghent, Belgium). Peptides were purified with RP-HPLC, using semi-preparative Nucleodur C18 column (Macherey-Nagel), mobile phases A (0.1 % TFA in Milli-Q water) and B (0.1 % TFA in acetonitrile, ACN) and gradient separation (0.5 - 2 % ACN/ min). Pure peptides were identified by strong absorption peaks at 214 nm (absorption of peptide bonds) or 280 nm (absorption of tyrosine and tryptophan rings) and they were captured as they eluted from the column. Joined fractions of peptides from different runs were parceled in several tubes due to different requirements in crystallization and cleavage assays afterwards. Sample tubes were then put on speedvac in order to remove ACN, lyophilized and stored at -80 °C until further use.

4.4 Protein Crystallization

4.4.1 Cathepsin K - alkyne inhibitor crystallization

The alkyne inhibitor of cathepsin K, the inhibitor 7 (200 μM) and cathepsin K (20 μM) in total volume of 8 mL were incubated for 10 hours at 37 °C in the buffer, consisting of 50 mM NaOAc, pH 5.5, 50 mM NaCl and 10 mM DTT. The formation of covalent bond between inhibitor and the reversible enzyme-inhibitor complex follows the first order kinetics:

$$c = c_0 \mathrm{e}^{-k_{inact}t}$$

where C_0 and C are concentrations of reversible inhibitor complex at times zero and t, respectively. Time needed to transform half of the reactant molecules to product is known as $t_{1/2}$ and can be calculated based on the previous equation as:

$$t_{1/2} = \frac{0.693}{k_{inact}}$$

The value of 0.011 was taken from the k_{inact} of inhibitor 4, so the approximated $t_{1/2}$ value for inhibitor 7 was one hour. As a rule of thumb, time equivalent of five $t_{1/2}$ is needed to complete the reaction, but we left the incubation stand for twice as much time. Hence, after 10 hours of incubation, the cathepsin K-inhibitor 7 complex was concentrated to 15 mg / mL and stored at -80 ºC until crystallization trials. The complex was screened against several commercial crystallization screens, namely JCSG I-IV (Molecular Dimensions), INDEX and AmSO4 (Hampton Research) by mixing 0.2 μL drops of the complex and screen solutions, performed on Gryphon crystallization robot (Art Robbins Instruments). After a few cycles of crystallization optimisation, in which protein concentration (6 - 15 mg / mL) as well as salt and PEG type and concentrations were varied, the best diffracting crystal grew from 0.2 M $CaCL_2$ and 20 % PEG 3350 and at protein concentration of 15 mg / mL. After the crystals were harvested, they were soaked for 10 seconds in 35 % PEG 3350 and 0.2 M CaCL2 for cryo-protection and stored in liquid nitrogen until they were measured at the Elettra synchrotron (Trieste, Italy).

4.4.2 Cathepsin V – calpeptin crystallization

12 mL of procathepsin V at approximate concentration of 1.7 mg / mL was auto-activated by diluting it at 1: 1 ratio with activation buffer (100 mM NaOAc, pH 4 and 10 mM DTT). The activation took two hours at 37 °C. Afterwards, the calpeptin was added to the sample in 5-fold molar excess relative to cathepsin V. After 1 hour of incubation, the cathepsin - calpeptin complex was purified by ion-exchange chromatography, using SP-sepharose FF resins. The complex was eluted with buffer consisting of 50 mM NaOAc, pH 5.5 and 400 mM NaCl. The sample was then desalted and concentrated with Amicon Ultra devices (cut-off 10 kDa) to approximately 40 mg / mL. The complex was then dialysed against 20 mM NaOAc, pH 4.5, 100 mM NaCl, 5 % glycerol and 1 mM DTT in D-Tube Dialyzer Mini 6 – 8 kDa cutoff (Millipore). Additional 1.7 mM Calpeptin was added to the sample towards the end of dialysis to replace any inhibitor molecules that might be dialysed from the cathepsin V in the process. The complex was then centrifuged at 14.000 RCF, supernatant collected and further concentrated to approximately 35 mg / mL. Crystals were grown in Intelli-Plate 24-4 (MiTeGen) at 6 °C by mixing equal volumes of protein solution and crystallization solution, which has been optimised and published previously for cathepsin

V mutants (77 % of MPD, 23 % of 60 mM Tris, pH 8) (Sosnowski, Piotr, 2016). The crystals were harvested and stored in liquid nitrogen until they were measured at the Elettra synchrotron (Trieste, Italy).

4.4.3 Cathepsin V C25S/A - peptide soaking

Cathepsin V C25S/A crystals were grown in Intelli-Plate 24-4 (MiTeGen) at 6 °C by mixing equal volumes of protein solution and crystallization solution (77 % of MPD, 23 % of TRIS buffer, pH 8). Peptide stocks from lyophilised peptide powders were made in 60 mM Tris, pH 8, in the concentration range from 20 mM to 90 mM and supplemented with DMSO, depending on the solubility of each peptide. Peptides in TRIS solution were mixed with MPD so that the final TRIS and MPD concentrations were equal to the crystal precipitate solution. One or two microliters of peptide solution was then added on top of crystals. Crystals were harvested after approximately 1 hour, 8 hours and 24 hours of soaking and stored in liquid nitrogen until they were measured at synchrotrons BESSY (Berlin, Germany) or Elettra (Trieste, Italy).

4.4.4 Cathepsin V C25S/A - peptide co-crystallization

Tris buffer component (60 mM Tris, pH 8) in crystallization solution was replaced with peptide stocks, made in the same buffer. Crystals were grown in Intelli-Plate 24-4 (MiTeGen) at 6 °C by mixing equal volumes of protein solution and crystallization solution, which has been optimized and published previously for cathepsin V mutants (77 % of MPD, 23 % of 60 mM Tris, pH 8) (Sosnowski, Piotr, 2016). The crystals were harvested and stored in liquid nitrogen until they were measured at synchrotrons BESSY (Berlin, Germany) or Elettra (Trieste, Italy).

4.4.5 Data collection, structure determination and refinement

Diffraction data of crystals were collected at synchrotrons Bessy, Berlin or Elettra, Trieste. Processing of diffraction data was done with XDS software (Kabsch, 2010). The MTZ reflection files were generated with XDS software or with Pointless, Aimless and Ctruncate programs inside the CCP4 program suite (Winn et al., 2011). The phases for cathepsin K were obtained from the PDB database structure 2FTD. The molecular replacement was carried out with the Molrep program (Vagin & Teplyakov, 1997) in the CCP4 program suite. The unit cell of cathepsin V-peptide complexes was the same as reported previously (Sosnowski, Piotr, 2016). The refinement of structures was done in MAIN software, using maximum-likelihood (ML) free-kick target function (Pražnikar & Turk, 2014; D. Turk, 2013). The peptides were modelled to the omitted maximum-likelihood averaged kick maps (ML AK) (Pražnikar et al., 2009). The geometric restraints for the inhibitors were generated in PURY (Andrejašič et al., 2008). Pictures were generated with RASTER 3D software (Merritt & Bacon, 1997).

4.5 Inhibitory Assays

Inhibitors calpeptin, S-calpeptin, N-calpeptin and GC-376 were tested for their inhibitory properties against cysteine cathepsins V, L, K, and B. All measurements were performed in the same buffer (50 mM NaOAc, pH 5.5, 50 mM NaCl and 5 mM DTT), with fluorogenic substrates Z-FR-AMC (cathepsins V and K) or Z-RR-AMC (cathepsins L and B) at 37 ^{o}C. Inhibitory assays against SARS-CoV-2 M^{pro} were performed in buffer 20 mM HEPES,

pH 7.3, 1 mM EDTA, 100 mM NaCl and 1 mM DTT, with fluorogenic substrates QS1 (Rut et al. 2020) or Acetyl-VKLQ-AMC (CL Biochem, Shanghai) at 37 ºC. All measurements were performed in 96-well black flat bottom microplates (Greiner, Germany) using Tecan INFINITE M1000 pro plate reader (Tecan, Switzerland) with excitation and emission wavelengths of 370 and 460 nm, respectively.

4.5.1 K_i determination of moderate inhibitors

K_i for inhibitors acting in nM or µM range were determined using mixed model inhibition formula in GraphPad Prism software. For cathepsin B and calpeptin-like inhibitors, the 10x stock solutions of seven inhibitor concentrations from 1.56 µM - 100 µM and 5 substrate concentrations from 62.5 - 1000 µM were prepared by 2-fold dilution series. 10 µM of each inhibitor stock and one control sample without inhibitor was added to 80 uL of cathepsin solution (final concentration of cathepsin B in the assay was 10 nM). Then, 10 µL of substrate stocks were added to the samples, so that each inhibitor concentration was measured at each substrate concentration. Measurements were performed immediately after the substrate was added to the sample. For cathepsin L and N-calpeptin, the 10x inhibitor stock concentrations were made from 15.6 - 1000 µM and substrate stocks from 31.3 - 1000 µM. For M^{pro} and calpeptin and S-Calpeptin, the working concentration of M^{pro} was 100 nM, the inhibitor concentrations were from 0.4 - 97 µM and substrate concentrations from 400 - 800 µM (Acetyl-VKLQ-AMC) or 100 – 400 µM (QS1). For M^{pro} and GC-376, the working concentration of M^{pro} was 100 nM and the inhibitor concentrations were from 12.5 - 800 nM. These measurements were performed in duplicates or triplicates.

4.5.2 K_i determination of tight binding inhibitors

K_i of inhibitors acting in pM range were determined with Morrison equation in GraphPad Prism software. 10x stock solutions of 15 inhibitor concentrations from 0.01 nM - 375 nM were prepared by 1.8-fold dilution series. 10 µM of each inhibitor stock and one control sample without inhibitor was added to 80 µL of cathepsin solution (final cathepsin concentration in the assay was 266 pM). 10 µM of substrate stock solution was then added to the sample just before the measurement. Final concentration of substrate in the sample was 24 and 25 µM Z-FR-AMC (cathepsin K and V, respectively) and 40 µM Z-RR-AMC (cathepsin L). The Km values of substrates were restrained to values as obtained from the MM plots in previous experiments: for Z-FR-AMC and cathepsins K and V they were 12 and 25 µM, respectively, and for Z-RR-AMC and cathepsin L they were 20 µM. The measurements were performed in duplicates or triplicates.

4.5.3 Covalent inactivation test

Cathepsin L (10 nM) was incubated with N-calpeptin (concentration range 1 - 27 µM) for 135 min at 37 °C. Sample aliquots were taken after 10 min, 45 min and at the end of incubation and measured for activity. Relative inhibition did not change over time, thus concluding that under the conditions tested, the inhibition is reversible.

4.6 Peptide Cleavage Analysis

4.6.1 Peptide digestion

Peptide stocks from lyophilized peptide powders were made either in 60 mM Tris, pH 8 or in 30 mM NaOAc and NaCl, pH 5.5, in the concentration range from 20 mM to 90 mM and supplemented with DMSO, depending on the solubility of the peptides. Peptide digestion with cathepsins was carried out in 30 mM NaOAc, 30 mM NaCl, pH 5.5 and 5 mM DTT, at 37 °C for 2 hours. Cathepsin concentration in the assay was 1 µM (cathepsins K and L) or 2 µM (cathepsin V), whereas the amount of peptide used in the assay was estimated empirically based on the peptide signal on RP-HPLC. After incubation, the sample was centrifuged for 10 min at 10.000 RCF and stored at -20 °C until further use. Activation of peptides, whose digestion was monitored in time, was stopped by submerging aliquots of the sample (30 µL) in boiling water at 5 sec, 30 sec, 2 min, 6 min, 20 min and 60 min of incubation time.

4.6.2 Peptide separation on RP-HPLC

Peptide fragments were separated on RP-HPLC system (Waters), using analytical Nucleodur C18 column (Macherey-Nagel), mobile phases A and B (degassed Milli-Q water and ACN supplemented with 0.1 % TFA, respectively) and gradient elution from 0.5 - 2 % ACN / min. Fragments of cathepsin-treated peptides were identified by absorption peaks at 214 nm (absorption of peptide bonds) or 280 nm (absorption of tyrosine and tryptophan rings) and captured as they eluted from the column. Samples containing peptide fragments were stored at -80 °C until they were measured with mass spectrometry.

4.6.3 Mass spectrometry analysis

Mass spectrometry was carried on Ultrafl eXtreme III MALDI-TOF/TOF mass spectrometer (Bruker, Billerica, MA, USA). For matrix, HCCA (1.4 mg / mL) was used. It was prepared by mixing 85% acetonitrile, 15% water, 0.1% TFA and 1 mM $NH_4H_2PO_4$. Captured fragments from peptide digestion (1 µL) were mixed with HCCA matrix (1 µL) on ground steel plate and left to dry at room temperature. Positive ions were measured in the range from 0 - 3500 Da, with parameters ion source 1, 25 kV; ion source 2, 22.30 kV; lens, 7.5 kV; reflector, 26.4 kV; reflector 2, 13.3 kV; pulsed ion extraction, 60 ns; and reflector detector voltage, 2230 V. Calibration was carried out externally by bradykinin (1-7), angiotensine I and angiotensine II and internally by 4-HCCA. Spectra were acquired, processed, and calibrated using FlexControl 3.0 and FlexAnalysis software (Bruker).

Chapter 5

Results

5.1 Structural and Biochemical Analysis of Peptide Binding to Cathepsin V

The results described in this section are part of a publication (Tušar & Loboda & Impens, 2023).

5.1.1 Peptide selection

Large scale proteomic data of cathepsin cleavages, performed under native conditions, delivered more than 30,000 cleavages of cathepsins V, K, L, S, F and B (carried out by the group of Kris Gevaert, Ghent, Belgium). The subsequent statistical analysis revealed positions with non-normal Gaussian distribution of residues which were important for substrate specificity, called heterogeneous positions, and those with normal distribution that carried no specificity information, called homogeneous positions (carried out by Livija Tušar, Ljubljana, Slovenia). Grouping of sequences with common characteristics at heterogeneous positions yielded major clusters, whose representatives distinctively share one or more common features (Figure A-1). Thirty peptides, cleaved by cathepsin V, were chosen for the crystallization and biochemical experiments. Cathepsin V was chosen as a model cathepsin due to the abundance of expressed material and good diffracting properties of its crystals. Peptide selection was based on the following criteria:

1 They had to represent the diversity of all seven major cathepsin V substrate clusters.
2 Sequences were from shared and unique cleavage sites (shared: cleavages performed by more than one cathepsin; unique: cleavages performed by only one cathepsin).
3 Sequences were from cleavage areas and positional cleavages (cleavage area: cleavages appeared next to each other; positional cleavage: the only cleavage in the neighborhood).
4 Sequences were from six to ten amino acid residues long.

Five sequences, not matching any of the protein sequences, were included in the synthesis. Most peptides had protected termini which mimic polypeptide chain of the protein substrates. To assess the potential role of charged termini on the binding, ten sequences were synthesized without termini protection. The list of synthesized peptides is given in Table A-1.

5.1.2 Crystallization of cathepsin V-peptide complexes

Two active site mutants of cathepsin V were prepared for crystallization purposes: C25S and C25A. Out of 41 peptides, 21 were built into the electron density maps of the crystal structures. These 21 peptides yielded 28 unique geometries, due to differences in binding of some peptides between the two cathepsin V molecules in the asymmetric unit, called A and B molecules. Of those, 26 were bound in the active site of cathepsin V in a substrate-like manner. Several structures revealed cleaved peptide fragments, indicating that both of the cathepsin V mutants retained some of its activity under crystallization conditions. This was not unexpected because catalytic activity of cathepsin L mutant in crystals has been observed previously (Adams-Cioaba et al., 2011; Sosnowski & Turk, 2016). Table 1 summarizes data collection and refinement statistics. All crystal structures are shown in Appendix B

Table 1: Crystallographic table for cathepsin V-peptide complexes. Shown are the range of values for all 21 structures.

Data collection statistics (last shell)	
Resolution limit (Å)	2.1 - 1.3
Space group	P 43 21 2
Multiplicity	7 – 25 (4 – 22)
Completeness (%)	> 99 (> 80)
Mean I/sigma(I)	10 – 32 (0.5 – 2.6)
Wilson B-factor	15 - 36
R-merge	0.05 – 0.20
CC1/2	0.995 - 1
Refinement statistics (last shell)	
R-work	0.16 – 0.21 (0.22 – 0.33)
R-kick	0.18 – 0.24 (0.26 – 0.34)
RMS (bonds)	0.013 – 0.023
RMS (angles)	1.7 – 2.2
Ramachandran favored (%)	95 - 99
Ramachandran allowed (%)	2 - 5
Ramachandran outliers (%)	0 - 1
Average B-factor	22 - 46

Our structures contained peptides bound in non-primed and primed substrate-binding sites. All peptides exhibited equivalent binding at subsites S2 – S2'. At S3, S4, S3', and S4', the binding of most peptides still followed the same direction, whereas beyond S4 and S4', the electron density maps of most structures worsened significantly, and the noisy maps indicated that there were no clearly defined binding areas. At S1, N and O atoms of the P1 residue formed H-bond to O atom of D163 and ND atom of Q19, respectively, and the carboxylic end of cleaved peptides or amide protective groups of protected peptides interacted with the NE atom of H164. At S2, the carbonyl O of the P2 residue formed an H-bond either to the N atom of G68 or N atom of Y26, and the N atom of the P2 residue formed an H-bond with the O atom of G68. Additionally, the amino group of Lys residues at S2 interacted with the carbonyl O atom of cathepsin V L162 and peptide P4 residues. At S3, no main-chain interactions occurred, but the side chains of longer residues at P3

interacted with the side chains of Q63 and N66. Two peptide fragments, VACK and TAHE, were the exceptions because their main chain ran into the S3 binding area. At S1', the carbonyl O of the P1' residue formed an H-bond with the NE atom of W190. At S2', and the O of the P2' residue formed an H-bond with the NE atom of Q145. At S3', the carbonyl O of the P3' residue formed an H-bond with the NE atom of Q21. At S4', N atom of P4' formed an H-bond with the OE atom of Q145 (Figure 3). The first notable exception in the primed site binding was the peptide RLSAKP with deviation at P4', where the side chain of Ala was placed, and at P6', where Pro formed electrostatic interactions with its carboxylic terminal to the amino group of K20. The second exception was the peptide AVAEKQ, which formed an internal H-bond between the O atom of the P3' residue and the amino group of P6'. It also formed additional interactions with neighboring molecules in the crystal.

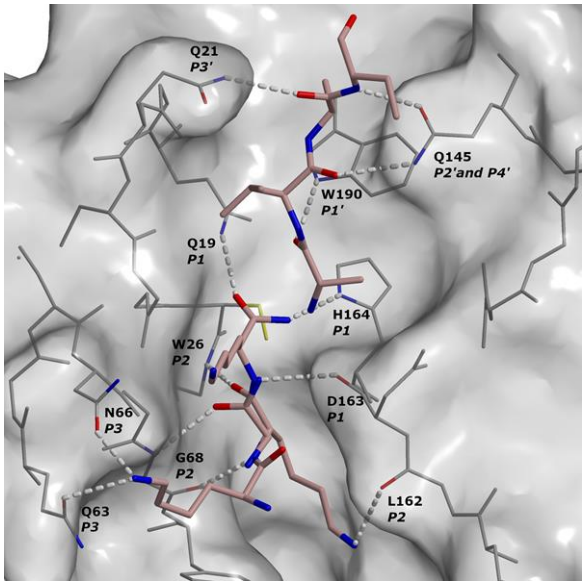


Figure 3: H-bonding pattern between cathepsin V and its substrates. Peptide fragments KKK (at P3–P1) and AVAE (at P1'–P4') are shown with stick model on the surface of a semi-transparent cathepsin V structure. Cathepsin V residues that constitute the active site cleft are shown with stick model in grey. Interacting oxygen and nitrogen atoms are shown in red and blue. Carbon atoms of peptides are shown in rose. H-bonds are presented whit dashed lines. At position P2, two main chain conformations are shown. In one conformation, the H-bond is formed between the O atom of P2 residue and the N atom of W26 residue, and in the other, the O atom of P2 residue forms H-bond with the N atom of G68 residue. Cathepsin residues that participated in peptide H-bonding are marked with sequence IDs and the interacting peptide position. The mutant residue S25 is highlighted in yellow.

We grouped peptide binding geometries into four binding patterns, based on the location of peptide binding or cleavage event: I. Peptides were cleaved and only their N-terminal fragments remained bound in the non-primed binding sites, II. The uncleaved peptides were bound to the non-primed binding sites only, III. The uncleaved peptides were bound to the primed binding sites only, and IV. Peptides were cleaved and both fragments remained bound to the non-primed and primed binding sites (Table 2). Crystal structures of peptides that belong to the same pattern of binding are superimposed in Figure 4.

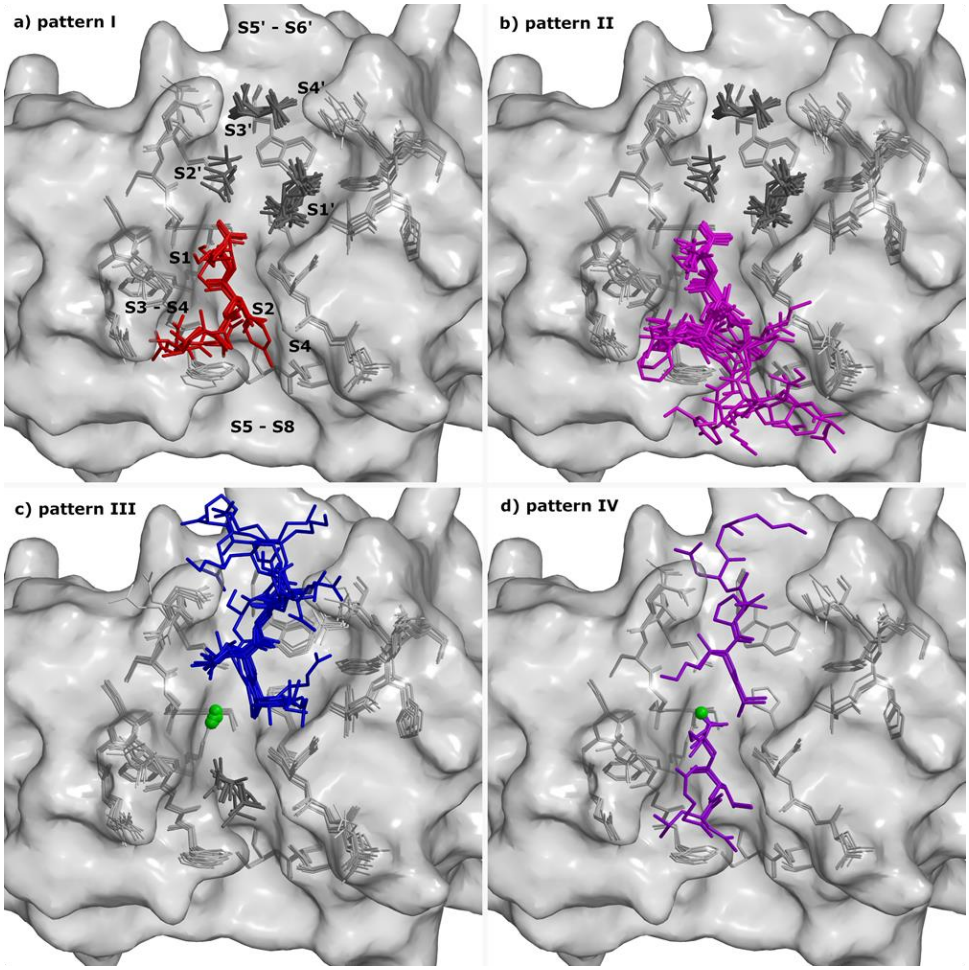


Figure 4: Binding geometry of peptides. Crystal structures of all complexes of cathepsin V were superimposed. Bound peptides are shown with stick model on the surface of a semi-transparent cathepsin V structure. Cathepsin V residues that constitute the active site cleft are shown with stick model in grey. Active site labels are shown only in pannel a for better clarity of the figure. Peptides are colored according to the patterns I (red, pannel a), II (purple, pannel b), III (blue, pannel c), and IV (violet, pannel d). Chlorine ions are shown as green balls. MPD molecules are shown as dark grey sticks.

The pattern I group of peptides consisted of structures of six peptides that were cleaved, and only their N-terminal fragments remained bound in the non-primed sites S4–S1. Electron density maps of all but one peptide enabled an unambiguous interpretation of the modelled residues from S3–S1, whereas the electron density map for the peptide fragment VACK was weaker, suggesting that the main chain of Ala and Val binds to the S3 binding site. The MPD molecules were bound to the primed site region. The peptide LLKVAL was cleaved and bound to non-primed sites only when co-crystallized, whereas soaking yielded binding in the primed binding area (Figure 4, panel a; Figure B-1).

The pattern II group consisted of structures of nine protected peptides and one non-protected peptide that all bound uncleaved into the non-primed sites. Electron density maps of the peptides enabled the unambiguous interpretation of residues from P3–P1. All peptides bound to cathepsin in the same manner, except for fragment TAHE, which bound to cathepsin-like fragment VACK (described in the pattern I group). The nitrogen of the amide protective group at the peptide C-terminus was bound approximately 3 Å away

from the ND atom of catalytic H164. The amide likely shared a hydrogen bond with the deprotonated His residue. At P4 and beyond, the features of the electron density maps became weaker and less precise; however, it was possible to model the peptides KPKKKTK, RLSAKP, GNYKEAKK, and EVCKKKK up to P8 in the ML AK omit maps (Pražnikar et al. 2009). MPD molecules were bound to the primed sites of all structures (Figure 4, panel b; Figure B-2).

The pattern III group consisted of eight non-protected peptides that all bound uncleaved into the primed sites. Electron density maps in the region from P1'–P3' and partly P4' of molecule A enabled an unambiguous interpretation. In molecule B, as mentioned previously, the electron density maps enabled the unambiguous interpretation of two further residues, P5' and P6', of the peptides AVAEKQ and RLSAKP. On the non-primed side, the MPD molecule occupied the S2 site, and the CL^- anion occupied the same position as the carbonyl oxygens or amide protective group of P1 residues in groups I and II, respectively. Its negative charge appeared to mimic the absent negatively charged SG atom of the reactive site cysteine. The positively charged amino group of the N-terminal residues of the peptides interacted with the negatively charged CL^- and ND atom of H164 at approximately 3 Å (Figure 4, panel c; Figure B-3).

The pattern IV group included structures of two peptides, LLKAVAEKQ and RLSAKP, which were both bound along the active site cleft of cathepsin V. LLKAVAEKQ was designed as a hybrid containing the LLK-fragment from the LLKVAL peptide, which was cleaved and remained bound to cathepsin V on the non-primed side, and AVAEKQ, which bound non-cleaved to the primed side. Overall, their electron densities were weak, and the fragments were refined with partial occupancies. As expected, the fragment LLK bound to the non-primed subsites S3–S1, whereas the fragment AVAEK bound to the primed subsites S1'–S5', as resolved by the ML AK omit map (Pražnikar et al., 2009). Despite the continuous electron density at the cleavage site, the distance of 2.4 Å between the C atom of Lys at P1 and the N atom of Ala at P1' was too wide to support a covalent bond between the fragments. However, in the middle there was sufficient space and density to attach the OXT atom to the Lys residue. In the structure of RLSAKP, the fragment RLS bound to the non-primed subsites S3–S1 and the fragment AKP to the primed subsites S1 – S3'. The distance of 2.6 Å between C of Ser at P1 and N of Ala at P1' and the continuous electron density between them resembled the LLKAVAEKQ structure. In both structures, the MPD molecules competed with peptide binding at subsites S2 and S1' – S3' and CL^- ions at the S1 site (Figure 4, panel d; Figure B-4).

Table 2: Summary of peptide binding to crystals of cathepsin V C25S/A. Peptide sequences are written out in the Sequence column. Peptide's clusters are written out in the Cluster column. Peptide residues modeled to ML AK omit maps are written in blue, under the columns that denote peptide positions from P8–P6'. The "A" and "B" in the Molecule column denote the cathepsin V molecule in the assymetric unit with the observed peptide binding. "Y" and "N" in the Protection column stand for "yes" and "no" and mark whether the peptide had their C- and N-terminals protected with amidation and acetylation, respectively. Method column marks the crystallization technique applied for peptide–cathepsin V complex formation: "s" for soaking, "c" for co-crystallization, and "s/c" if both techniques yield the same binding. Four patterns of peptide binding to cathepsin V were observed; table is divided into four parts: I, binding of cleaved peptide fragments to the non-primed site; II, binding shifted to the non-primed site; III, binding shifted to the primed site; and IV, binding of partially cleaved peptides across the active site. Peptides exhibiting multiple binding patterns with respect to position in the asymmetric unit or crystallization method used are highlighted in the sequence column. Peptides with the same sequence but different termini are not treated as equivalent peptides. MPD is 2-methyl-2,4-pentanediol. Chlorine anion is marked as CL^-.

Sequence	Cluster	P8	P7	P6	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'	Molecule	Protection	Method
Pattern I. Binding of cleaved peptide fragments at the non-primed site																		
VACKSSQP	2					**V**	**A**	**C**	**K**	MPD						A, B	Y	s
VYEKKP	5						**V**	**Y**	**E**	MPD						A, B	N	s
GAKSAA	2						**G**	**A**	**K**	MPD						A	N	s
LLSGKE	1						**L**	**L**	**S**	MPD						B	N	s
LLKVAL	2						**L**	**L**	**K**	MPD						A, B	N	c
LLKAVAEKQ	2						**L**	**L**	**K**	MPD						B	Y	s/c
Pattern II. Binding shifted to the non-primed site																		
EVCKKKK	3		**E**	**V**	**C**	**K**	**K**	**K**	**K**	MPD						A	Y	s
AYFKKVL	5		A	Y	F	K	**K**	**V**	**L**	MPD						B	Y	s
RLSAKP	1			R	**L**	**S**	**A**	**K**	**P**	MPD						B	Y	s/c
TRESEDLE	6	T	R	E	S	E	**D**	**L**	**E**	MPD						A, B	Y	s
GNYKEAKK	2	**G**	**N**	**Y**	**K**	**E**	**A**	**K**	**K**	MPD						A	Y	s
KPKKKTK	7		K	P	K	**K**	**K**	**T**	**K**	MPD						B	Y	s
GAKSAA	2			G	A	K	**S**	**A**	**A**	MPD						A, B	Y	s/c
KKYDAFLA	6	K	K	Y	D	A	**F**	**L**	**A**	MPD						A, B	Y	s
VPCGTAHE	6	V	P	C	G	**T**	**A**	**H**	**E**	MPD						A, B	Y	s
QLRQQE	1			Q	L	R	**Q**	**Q**	**E**	MPD						B	N	s
Pattern III. Binding shifted to the primed site																		
LLSGKE	1						MPD	CL^-	**L**	**L**	**S**	G	K	E	A	N	s	
QLRQQE	1						MPD	CL^-	**Q**	**L**	**R**	**Q**	Q	E	A	N	s	
RLSAKP	1						MPD	CL^-	**R**	**L**	**S**	**A**	**K**	**P**	B	N	s	
GAKSAA	2						MPD	CL^-	**G**	**A**	**K**	**S**	A	A	A	N	s	
IILKEK	3						MPD	CL^-	**I**	**I**	**L**	K	E	K	A	N	s	
LLKVAL	2						MPD	CL^-	**L**	**L**	**K**	**V**	A	L	A	N	s	
AVAEKQ	4						MPD	CL^-	**A**	**V**	**A**	**E**	**K**	**Q**	B	N	s	
ALAASS	1						MPD	CL^-	**A**	**L**	**A**	**A**	S	S	A	N	s	
Pattern IV. Binding of cleaved peptide fragments across the active site																		
RLSAKP	1						**R**	**L**	**S**	**A**	**K**	**P**				A	Y	s/c
LLKAVAEKQ	2						**L**	**L**	**K**	**A**	**V**	**A**	**E**	**K**	Q	A	Y	s/c

5.1.3 Peptide digestion with cathepsins K, V, and L

After treatment with native cathepsins V, L, and K, we determined the cleavage sites of all 27 peptides that were selected from protein substrate cleavages of cathepsin V. These are peptides p1 – p30, with exception of peptides p1 and p2, which could not be dissolved, and peptide p15, which was spend in the structural assay. In addition, we also treated the peptide p35, whose sequence doesn't originate from the protein substrate cleavages (Table A.1). In total, 150 cleavages were observed. Next we compared peptide cleavages to cleavages of their protein counterparts (Table C.1). Table 3 shows that 42% of all peptide cleavages and 69% of all protein cleavages were identical, whereas the remaining cleavages were observed only with one type of substrate (58% of total peptide and 31% of total protein cleavages). Most cleavages were shared (performed by more than one cathepsin), whereas a few were unique to only one cathepsin (Table 3, a). Statistical comparison of the patterns of cleaved peptides and their protein counterparts showed that there was no significant difference between their cleavage patterns, demonstrating that the selected sequences indeed represented a variety of protein cleavage samples of all seven cathepsin V clusters, despite the fact that peptides were cleaved in more places than their protein counterparts (146 peptide and 90 protein cleavages among the selected sequences; Table 3, b; Table C-1).

Of all 27 peptides treated with three different cathepsins, only 11 were cleaved in peptides and proteins at the same position by at least one cathepsin. Other peptides contained additional cleavages that were observed with only one substrate type. In contrast, sequences EVC↓KKKK, IIL↓KEK, and TRES↓EDLE had only one observed protein cleavage site, indicating very restrictive processing, whereas in peptides they were cleaved at two sites by all three cathepsins (IIL↓K↓EK), at two sites by cathepsins K and L (EVC↓K↓KKK), and at three sites by cathepsins V and L (TR↓E↓S↓EDLE). In addition, several peptide sequences, cleaved by all cathepsins, were not cleaved in proteins by cathepsin K, L, or both, whereas they cleaved each sequence at least at one site in the peptidyl form. Interestingly, four of these sequences (AWKKEA, SIYEVDKQ, KKYDAFLA, and GNYKEAKK) appeared as weak substrates of cathepsin K in the peptidyl form and were only partially processed by cathepsin K during the incubation period. We also observed multiple fragments that had in their sequences embedded protein cleavage sites, which were evidently not cleaved when present in peptides. These were ESEDLE, ATVT, and KPK fragments with intact protein cleavage sites TRES↓EDLE, KVLAT↓VTK, KP↓KKKTK (cathepsin K), fragments NPKGN and AKP with cleavage sites EIDLRNPKG↓N and RLSA↓KP (cathepsin V), and KSVT with cleavage sites ACMK↓SVTE (cathepsins V and K) and ACMKSV↓TE (cathepsin K) (Table C-1 and Figure C.2.1). This data shows that recognition of several sequences in proteins and peptides was not the same.

Interestingly, we discovered that cathepsins cleaved peptides along their entire length, including the terminal residues and their protective groups, four of which had their C-terminal residues removed by all cathepsins, whereas cathepsins K and V cleaved the amino-terminal residue of one peptide each. This suggested exopeptidase-like activity of cathepsins toward peptides. To gain further insight, we followed the cleavage of peptides AYFKKVL and KVLATVTK from 5 seconds to 60 minutes. The analysis confirmed the carboxypeptidase-like activity of cathepsins V and L, but not K, which is evident from gradual processing of fragments AYFK and KVLA to AYF and KVL, respectively (Figure C.2.2).

Table 3: Peptide and protein cleavage analysis. **a)** Unique and shared cleavages of peptides. Unique cleavages were performed by only one cathepsin, whereas shared cleavages were performed by two or three cathepsins. Percentages in brackets refer to the total observed cleavages. **b)** Comparison of peptide and protein cleavages. Cleavage sites identical among peptides and proteins were separated from cleavages that were observed with one type of substrate only (listed in the rows "Peptides only" and "Proteins only"). Percentages in brackets refer to the total observed peptide or protein cleavages. Cleavages of peptide p35 were exluded from the comparison because it was not derived from the protein substrate cleavages.

a) Cleavages in number	**CatK**	**CatL**	**CatV**	**All**
Total peptide cleavages	52	51	47	150
Unique	11 (21 %)	4 (8 %)	3 (6 %)	18 (12 %)
Shared	41 (79 %)	47 (92 %)	44 (94 %)	132 (88 %)
b) Comparison of cleavages	**CatK**	**CatL**	**CatV**	**All**
Total cleavages (peptides, proteins)	51, 29	49, 23	46, 38	146, 90
Peptides only	34 (67 %)	30 (61 %)	20 (43 %)	84 (58 %)
Proteins only	12 (41 %)	4 (17 %)	12 (32 %)	28 (31 %)
Identical cleavages (peptides, proteins)	17 (33 %, 59 %)	19 (39 %, 38 %)	26 (57 %, 68 %)	62 (42 %, 69 %)

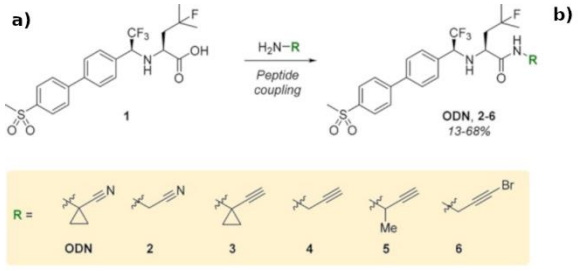
5.2 Crystal Structure of Cathepsin K - Alkyne Inhibitor

The results described in this section are part of a publication (Mons et al., 2019).

5.2.1 Selectivity and reactivity of Odanacatib-like alkyne inhibitors

In this work, the nitrile warhead of ODN was replaced with alkyne functional group. Skeleton optimization and compound characterization was carried out by the group of Huib Ovaa (Leiden, Netherlands). Five different derivatives based on ODN skeleton were prepared (Figure 5, a). While inhibitors 3 - 5 were purely alkyne-based, the inhibitor 6 was activated with the addition of electron-withdrawing bromine ion to the alkyne group and the inhibitor 2 remained nitrile-based. The inhibitors were first incubated with cysteine molecules to assess their indiscriminate thiol reactivity. Only inhibitors 2 and 6 formed adducts with cysteine molecules, whereas inhibitors 3 - 5 were inactive.

The selectivity of alkyne inhibitors 3 - 5, but not 6, towards cathepsin K was retained, whereas their potency dropped by approximately 10^2 - 10^3 (inhibitors 4 and 5) or 10^5 (inhibitor 3) fold relative to nitrile inhibitors ODN and inhibitor 2 (Figure 5, c). The reactivity of inhibitors was compared based on their k_{inact} values. Surprisingly, the addition of bromine to the alkyne moiety of inhibitor 6 did not improve its reactivity over inhibitors 4 and 5 (Figure 5, b). This suggested that formation of covalent bond between alkyne group and cysteine of cathepsin K was governed by local forces in the protease active site.



b)

Irreversible covalent

E + I $\rightleftharpoons$ E···I $\rightarrow$ E-I (k_{on}, k_{off}, k_{inact})

Non-covalent — Covalent

Compound	k_{inact} (min^{-1})	K_i (μM)	k_{inact} / K_i
Inhibitor 4	0.011 ± 0.00073	0.21 ± 0.047	53 x 10^3 ± 12 x 10^3
Inhibitor 5	0.047 ± 0.0037	3.3 ± 0.60	15 x 10^3 ± 2.9 x 10^3
Inhibitor 6	0.019 ± 0.0027	0.19 ± 0.083	99 x 10^3 ± 45 x 10^3
E-64	0.081 ± 0.0052	0.011 ± 0.00016	71 x 10^5 ± 11 x 10^5

Reversible covalent

E + I $\rightleftharpoons$ E···I $\rightleftharpoons$ E-I (k_{on}, k_{off}, k_5, k_6)

Non-covalent — Covalent

Compound	k_5 (min^{-1})	k_6 (min^{-1})	K_i^{app} (nM)
ODN	0.12 ± 0.026	0.033 ± 0.0077	3.8 ± 2.7

c)

Cathepsin (mol / L) / compound	Cat K	Cat L	Cat S	Cat V	Cat B
ODN	5.6 x 10^{-10}	5.8 x 10^{-6}	2.4 x 10^{-8}	6.0 x 10^{-7}	6.3 x 10^{-8}
Inhibitor 2	5.7 x 10^{-10}	> 10^{-6}	1.8 x 10^{-8}	9.1 x 10^{-7}	2.1 x 10^{-8}
Inhibitor 3	2.6 x 10^{-5}	> 10^{-4}	> 10^{-4}	> 10^{-4}	> 10^{-4}
Inhibitor 4	2.9 x 10^{-7}	> 10^{-4}	1.1 x 10^{-5}	2.4 x 10^{-5}	9.4 x 10^{-6}
Inhibitor 5	3.5 x 10^{-7}	> 10^{-4}	1.6 x 10^{-5}	4.6 x 10^{-5}	4.0 x 10^{-5}
Inhibitor 6	4.7 x 10^{-8}	1.0 x 10^{-7}	5.5 x 10^{-8}	1.6 x 10^{-8}	9.9 x 10^{-8}

Figure 5: Optimisation of cathepsin K inhibitors. **a)** Synthesis of alkyne inhibitors, based on Odanacatib backbone. **b)** In vitro kinetic evaluation of irreversible inhibitors (upper part) and reversible inhibitor (lower part). **c)** In vitro determined IC_{50} values against selected human cathepsins (in mol / L) (Mons et al., 2019).

5.2.2 Crystallization of cathepsin K – alkyne-based inhibitor

We made an attempt to solve at least one structure of cathepsin K with bound inhibitor. Inhibitors 4 and 5 were prioritized because of their pure alkyne character. Their poor solubility was a limiting factor, so another derivative was prepared (inhibitor 7) by substitution of fluorine atom with hydrogen on the L-leucine building block of inhibitor 4, which indeed improved the compound solubility. Cathepsin K (20 μM) was incubated with

inhibitor 7 (200 µM) at 37 ºC for 10 hours, which was enough to block virtually all cathepsin K molecules (see Section 4.4.1).

The complex was first screened against several commercially available screens. Needles grew from several conditions, containing $CaCL_2$ and different types and concentrations of polyethylene glycol (PEG). Two most promising conditions, namely JCSG-III Num. 30 (0,2 M CaCl2, 0,1 M HEPES pH 7.5, 28 % PEG-400) and JCSG-III Num. 77 (0.2 M CaCl2, 20 % PEG-3350) were used as a starting point for further optimization (Figure 6, Table 4).

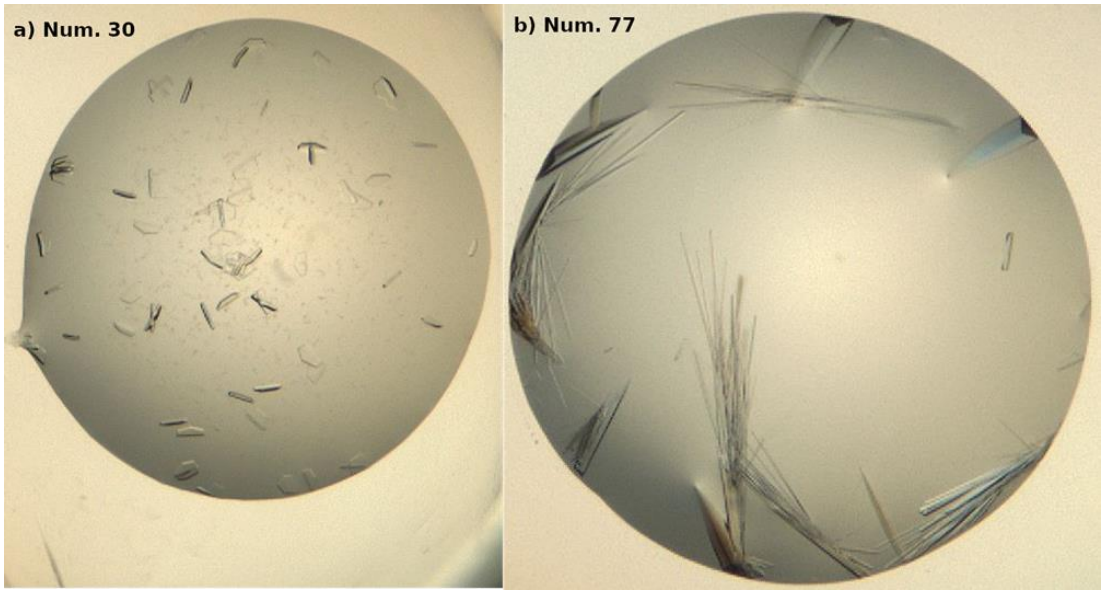


Figure 6: Best crystals from the cathepsin K-alkyne inhibitor screening experiment. Crystals grew from JCSG-III screen in conditions Num. 30 (left) and Num. 77 (right). Pictures were taken after 12 days.

Table 4: Optimisation of cathepsin K-alkyne inhibitor crystallization conditions. The variables of the condition Num. 30 were CaCl2 (0.1 M, 0.2 M or 0.5 M), PEG-400 (25 % or 35 %) and protein complex concentration (6 or 8 mg / mL). Few 3-D crystals grew in condition Num. 77, hence its composition was fixed and instead volumes of the protein complex solution (denoted as "p") and crystallization solution (denoted as "c") were varied. Asterisk (*) marks conditions which yield well diffracting crystals in the optimisation. Two asterisks (**) mark condition which yields the best diffracting crystal.

Num. 30			Num. 77	
0.1 M $CaCl_2$ 25 % Peg-400 8 mg / mL	0.2 M $CaCl_2$ 25 % Peg-400 8 mg / mL	0.5 M $CaCl_2$ 25 % Peg-400 8 mg / mL	*0.2 M $CaCl_2$ 20 % Peg-3350 0.5 µL (p) + 0.5 µL (c)	*0.2 M $CaCl_2$ 20 % Peg-3350 1 µL (p) + 0.7 µL (c)
0.1 M $CaCl_2$ 35 % Peg-400 8 mg / mL	0.2 M $CaCl_2$ 35 % Peg-400 8 mg / mL	0.5 M $CaCl_2$ 35 % Peg-400 8 mg / mL	*0.2 M $CaCl_2$ 20 % Peg-3350 0.5 µL (p) + 0.8 µL (c)	
0.1 M $CaCl_2$ 25 % Peg-400 6 mg / mL	0.2 M $CaCl_2$ 25 % Peg-400 6 mg / mL	0.5 M $CaCl_2$ 25 % Peg-400 6 mg / mL	**0.2 M $CaCl_2$ 20 % Peg-3350 0.5 µL (p) + 1 µL (c)	
0.1 M $CaCl_2$ 35 % Peg-400 6 mg / mL	0.2 M $CaCl_2$ 35 % Peg-400 6 mg / mL	0.5 M $CaCl_2$ 35 % Peg-400 6 mg / mL	0.2 M $CaCl_2$ 20 % Peg-3350 1 µL (p) + 0.5 µL (c)	

5.2.3 Data collection and structure determination

Crystals of cathepsin K-alkyne inhibitor complexes diffracted from 2.3 – 1.7 Å in several different space groups: $P2_12_12_1$, P222, P6 and $P6_122$. The best diffracting crystal was chosen for data collection. Data was processed at the spot with XDS software (Kabsch, 2010). Crystal phases were obtained by molecular replacement in Molrep (Vagin & Teplyakov, 1997), using cathepsin K molecule of the PDB entry 2FTD as a starting model. Structure was refined in MAIN software (D. Turk, 2013). First round of refinement was performed only with cathepsin coordinates. The continuous electron density in the difference map in the active site cleft of cathepsin suggested that inhibitor was attached to the catalytic residue Cys 25. The inhibitor, which was generated in PURY (Andrejašič et al., 2008), and solvent molecules were then added to the model and the model was further refined. The refinement staggered at the value of R-work around 0.206 (R-kick 0.230). Data collection and refinement statistics are provided in Table 5. The structure was deposited to PDB server and was given the entry code 6QBS.

The crystal asymmetric unit is composed of two molecules of cathepsin K-inhibitor complexes with the root-mean-square deviation (RMSD) of their CA atoms 0.33 Å (Figure 7). Their chains were resolved across their entire sequence. Crystal packing is stabilized by calcium ion at the interface of both cathepsin K molecules. Both active sites are occupied with the inhibitor 7 in the same way. The SG atom of Cys 25 and C2 atom of inhibitor 7 form a covalent bond which is evident from their electron density maps. Binding of inhibitor is further stabilized by two H-bonds, formed between amide part of the inhibitor and carbonyl O of the N161 and amide N of the G66 residues. These two bonds are equivalent to bonds that are formed between cathepsins and their substrates, where carbonyl O of the N161 and amide N of the G66 residues form H-bond to N of the P1 and carbonyl O of the P2 residue, respectively. The leucine-like moiety occupies the S2 binding pocket in a substrate-like manner, where it is stabilized by hydrophobic interactions. The 3-fluoro-methyl group faces toward the solvent. The biphenyl group binds between the amino groups of N60-D61 and G65-G66 on the one side and Y67 of the neighboring cathepsin molecule on the other side. The atomic B-factors of inhibitor atoms (with the exception of sulphonyl group, which are higher) are comparable to that of the cathepsin K residues around the active site (mean value 14.9 $Å^2$) which shows that all cathepsin molecules reacted with the inhibitor.

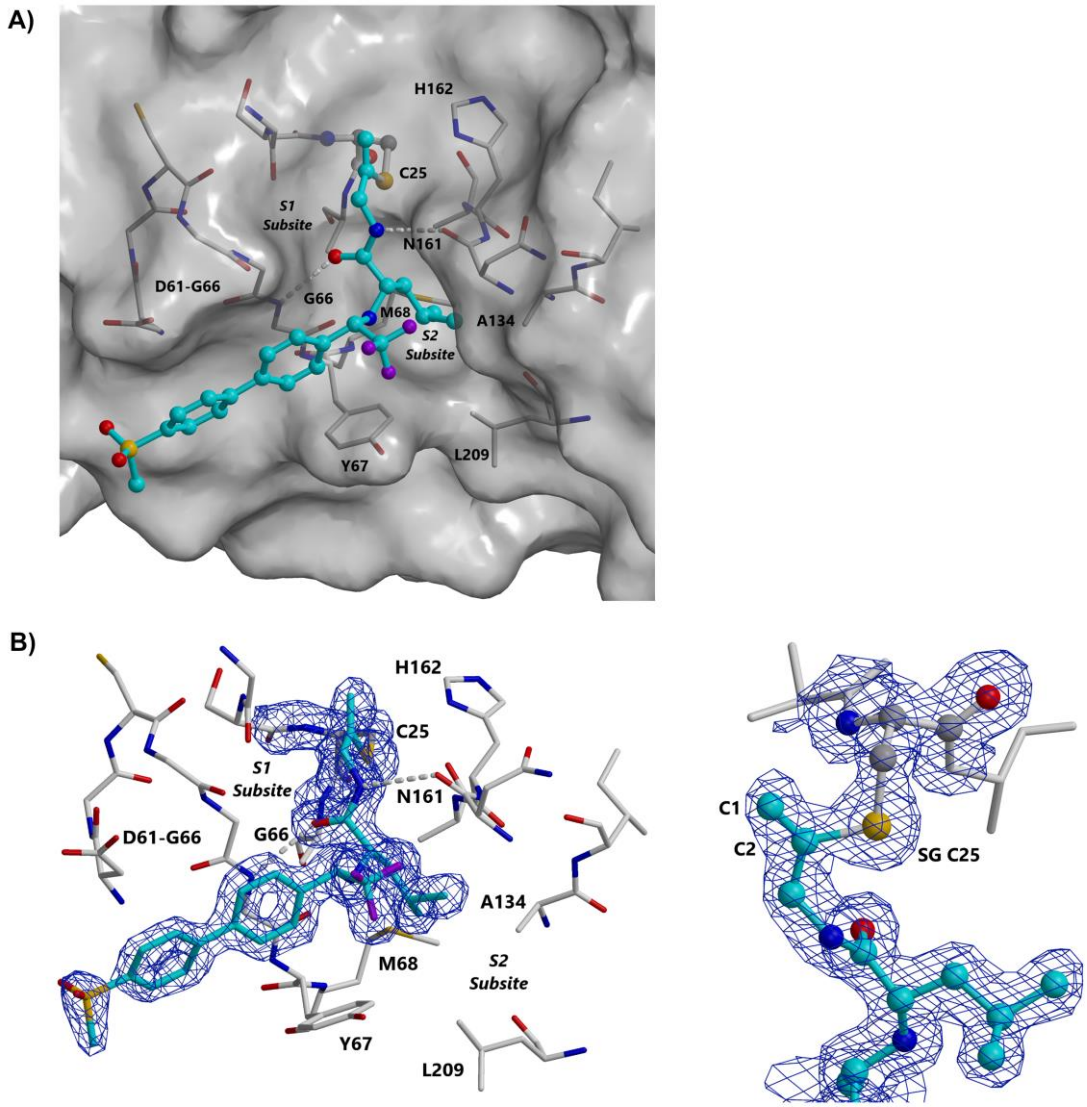


Figure 7: Crystal structure of cathepsin K-alkyne inhibitor. Nitrogen, oxygen, fluorine, and sulphur atoms are shown in blue, red, violet, and yellow, respectively, whereas carbon atoms of inhibitor 7 and cathepsin K are shown in cyan and grey, respectively. **a)** Binding of inhibitor 7 to cathepsin K. Inhibitor is shown with ball-and-stick model, relevant residues in cathepsin K are shown with stick model. Covalent bonds of inhibitor 7 are shown as cyan sticks, whereas those of cathepsin K are shown as white sticks. Cathepsin K is wrapped in white transparent surface. **b)** Maximum-likelihood free-kick electron density map ($2Fo\text{-}Fc$) around inhibitor 7 and Cys 25. Blue mesh represents electron density map contoured at 1.3 σ. Relevant cathepsin K residues are shown with stick model. Inhibitor is shown in stick model (left) or ball-and-stick model (right) (Mons et al., 2019).

Table 5: Crystallographic table for cathepsin K-alkyne inhibitor entry 6QBS.

Data collection statistics (last shell)	
Resolution range (Å)	32.63 - 1.703 (1.763 - 1.703)
Space group	P 61 2 2
Unit cell	75.345 75.345 340.195 90 90 120
Total reflections	2125267 (192495)
Unique reflections	64020 (6254)
Multiplicity	33.2 (30.8)
Completeness (%)	99.94 (99.71)
Mean I/sigma(I)	17.23 (2.29)
Wilson B-factor	19.79
R-merge	0.1519 (0.9156)
CC1/2	0.999 (0.908)
Refinement statistics (last shell)	
Reflections used in refinement	64018 (3091)
Reflections used for R-kick	64018 (3091)
R-work	0.209 (0.337)
R-kick	0.233 (0.362)
RMS (bonds)	0.020
RMS (angles)	1.98
Ramachandran favored (%)	96.71
Ramachandran allowed (%)	3.29
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	3.47
Average B-factor	19.35

5.3 Characterization of Calpeptin and Alike Compounds as Cathepsin Inhibitors

The results described in this section are part of a publication (Reinke et al, 2023; submitted).

5.3.1 Enzyme inhibition assays

We determined inhibitory properties of calpeptin, a potential anti-viral SARS-CoV-2 agent, against human cathepsins K, V, L and B. Three additional compounds were tested: N-calpeptin and S-calpeptin, both derivatives of calpeptin, and GC-376, a compound very similar to calpeptin by structural and functional means (Figure 8). Inhibitors were also tested against viral M^{pro} protease, prepared in our laboratory, to compare our data with the literature. To further establish the role of other cathepsins in SARS-CoV-2 infection, we set out to determine the crystal structure of cathepsin V-calpeptin complex.

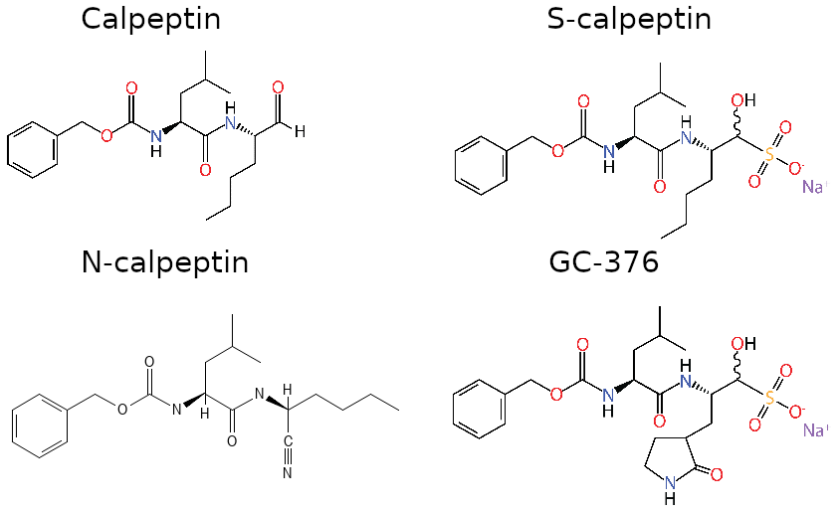


Figure 8: Chemical structures of calpeptin and alike compounds.

5.3.1.1 IC_{50} screen

In initial screening, the inhibitors were tested against 10 nM cathepsins L (calpeptin, S-calpeptin and N-calpeptin) and V (calpeptin) in concentration range from 1 nM - 250 µM. Inhibitor GC-376 was not tested because it was not available at the time. The IC_{50} of calpeptin and S-calpeptin approached the lower limit of the assay (5 nM) for both cathepsins tested, whereas N-calpeptin appeared as approximately 10^3-fold weaker inhibitor, with the IC_{50} value of 6400 nM (Figure 9). These results suggested that calpeptin suppression of SARS-CoV-2 infection could be mediated through the inhibition of cathepsin L and possibly also other cathepsins.

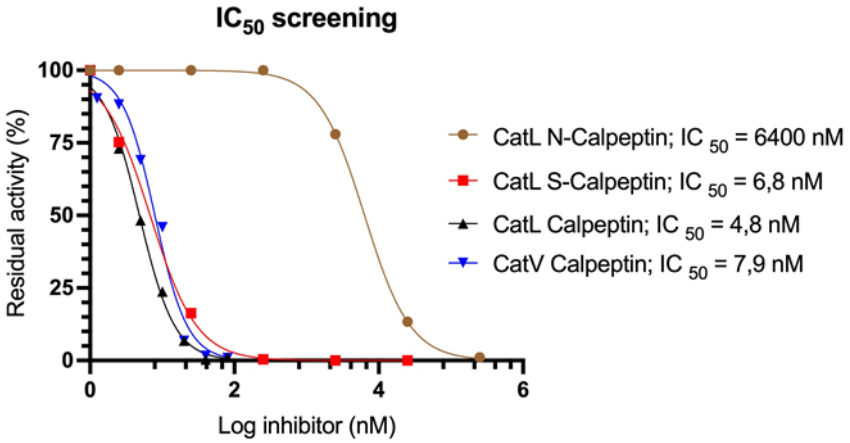


Figure 9: IC_{50} screen of calpeptin and alike compounds. Residual cathepsin activity (in %) was plotted against logarithmic inhibitor concentration. Brown: cathepsin L and N-calpeptin; red: cathepsin L and S-calpeptin; black: cathepsin L and calpeptin; blue: cathepsin V and calpeptin. The IC_{50} values, written on the right, were derived as the inhibitor concentration where 50 % of residual cathepsin activity was retained. Data were fitted with non-linear regression, using variable slope model. Chart was prepared in GraphPad Prism software.

5.3.1.2 K_i determination

Results from the IC_{50} screening experiment indicated that calpeptin and S-calpeptin, but not N-calpeptin, inhibit cathepsins at low nanomolar or picomolar range. Hence, to determine their K_i values, we lowered the concentration of cathepsins in the subsequent assays to 0.26 nM, which was to our experience the lowest cathepsin concentration in the assay used that still yielded sensible signal. Cathepsins were incubated with 15 different inhibitor concentrations, made with 1.8-dilution series in the concentration range from 0.01 - 37.5 nM. Because inhibition occurred at concentrations below that of the cathepsins, the tight binding had to be acknowledged. Relative cathepsin activity (in %) was plotted against each inhibitor concentration used in the assay and fitted with non-linear regression to the Morrison equation in GraphPad Prism software (Figure 10, upper row;). The exception was cathepsin B, which was inhibited only at higher inhibitor concentrations (above 10 nM) and thus its K_i determined as described below.

K_i values of weaker complexes, namely those of cathepsin B and M^{pro} (Inhibitors calpeptin, S-calpeptin and GC-376) and cathepsin L (inhibitor N-calpeptin), were determined by mixing the enzymes with several different concentrations of inhibitor, spanning across their area of inhibition, and their velocities determined with several different substrate concentrations. Enzyme velocity (in relative fluorescence units per second; RFU / sec) was plotted against substrate concentration for each inhibitor concentration used in the assay and fitted with non-linear regression to the mixed-model inhibition formula in GraphPad Prism software. Because N-calpeptin was a much weaker cathepsin L inhibitor than the rest, it was not tested against other cathepsins (Figure 10, middle and bottom rows; Table 6)

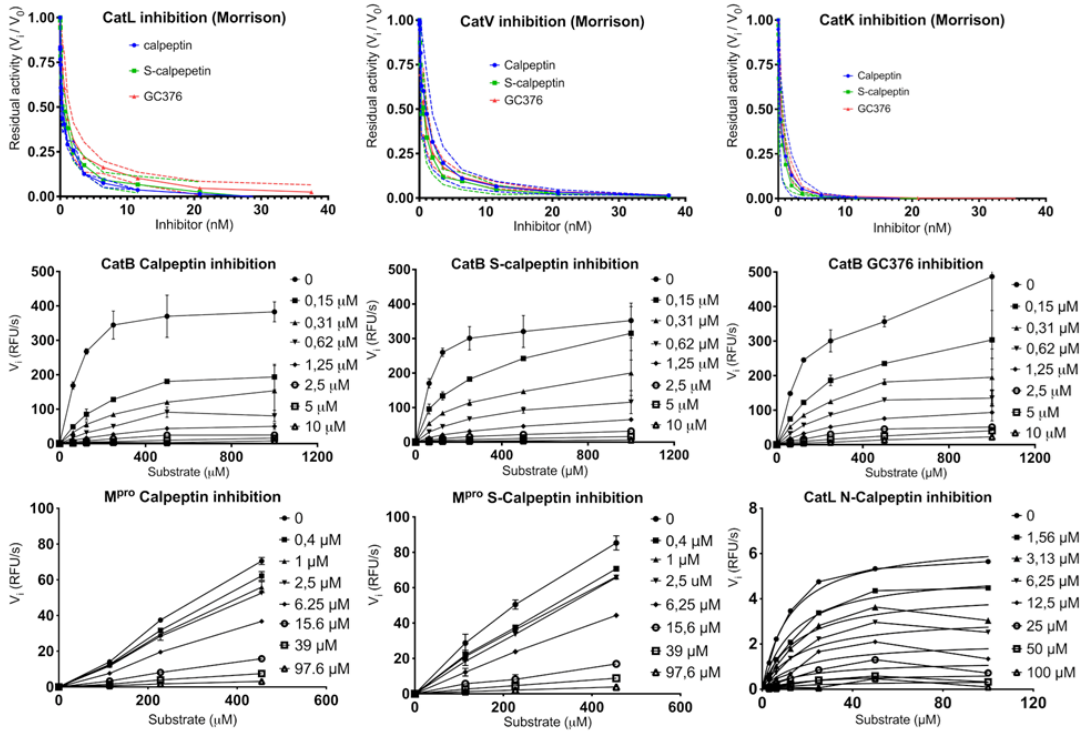


Figure 10: K_i determination of calpeptin and alike compounds. The upper row shows charts of cathepsins L (left), V (middle) and K (right) residual activity plotted against different inhibitor concentrations in the range from 0.01 – 37.5 nM in blue (calpeptin), green (S-calpeptin) and red (GC-376). The measure uncertainty is presented with dotted lines. The middle row shows charts of cathepsin B reaction velocities (in relative fluorescence units per second; RFU / sec) measured at several substrate concentrations for different concentrations of inhibitors calpeptin (left chart), S-calpeptin (middle chart) and GC-376 (right chart). The error boxes are shown around the measured points. The bottom row shows inhibition of M^{pro} by calpeptin (left chart), S-calpeptin (middle chart) and cathepsin L by N-calpeptin (right chart). Data were fitted with non-linear regression and inhibitor K_i values determined using Morrison equation (upper row) or mixed-model inhibition formula (middle and bottom rows). K_i values, determined from these data, are listed in Table 6.

Table 6: K_i values of calpeptin and compounds alike. The K_i for inhibitor GC-376 and M^{pro} was below the sensibility of the assay. In this case, asterisk (*) shows the residual M^{pro} activity measured at 100 nM inhibitor concentration.

K_i ± std. error / 95% confidence interval (pM, unless written otherwise)					
	CatV	CatK	CatL	CatB	M^{pro}
Calpeptin	361 ± 47 / 268 – 454	61 ± 12 / 37 – 85	131 ± 21 / 90 – 172	41 ± 7 nM / 27 – 54 nM	5,6 ± 1.2 µM/ 3,0 – 9,2 µM
S-Calpeptin	169 ± 18 / 133 – 204	50 ± 12/ 28 – 73	148 ± 19 / 111 – 185	70 ± 15 nM / 39 – 100 nM	4.8 ± 1.5 µM / 1,1 – 7,9 µM
GC-376	242 ± 23 / 196 – 288	91 ± 11 / 70 – 112	259 ± 27 / 204 – 314	163 ± 50 nM / 63 – 262 nM	< 100 nM 10 % *
N-Calpeptin	Not performed	Not performed	3,5 ± 1 µM 1,4 – 5,4 µM	Not performed	Not performed

The inhibition assay showed that calpeptin, S-calpeptin and GC-376 inhibit cathepsin V, L and K in picomolar range. The compounds also inhibit cathepsin B but at approximately 10^2 – 10^3-fold higher inhibitor concentrations. Our results confirmed that calpeptin and S-calpeptin inhibit M^{pro} enzyme only at low micromolar range, at approximately 10^4-fold higher inhibitor concentrations than that required for inhibition of cathepsins V, L, and K. We confirmed that compound GC-376 is a better inhibitor of M^{pro} than calpeptin and S-calpeptin by at least 15-fold, however the sensibility of the assay was too low to precisely determine its inhibition constant.

5.3.2 Crystallization of cathepsin V-calpeptin complex

The calpeptin is aldehyde-based inhibitor, so it forms reversible hemithioacetal bond with the cathepsin Cys residue. Due to reversible nature of the chemical bond, the inhibition formula can be written as:

$$K_i = \frac{[E] \times [I]}{[EI]} \tag{5.1}$$

or

$$\frac{[I]}{K_i} = \frac{[E\bar{I}]}{[E]} \tag{5.2}$$

where [E], [I] and [EI] stand for free enzyme, free inhibitor and enzyme-inhibitor concentrations, respectively. To ensure that all cathepsin molecules are bound with inhibitor, the [I] needs to be in large access relative to K_i value and [I] also needs to be in access relative to [E]. Hence, the cathepsin V-calpeptin complex was formed by the addition of 40 µM calpeptin, the value much above its K_i, to the 8 µM solution of freshly activated cathepsin V. The complex was then purified on ion-exchange chromatography, concentrated to approximately 40 mg / mL (around 1.7 mM) and dialyzed to crystallization buffer (20 mM NaOAc, pH 4.5, 100 mM NaCl, 5 % glycerol and 1 mM DTT). Shortly before the dialysis ended, calpeptin in the final concentration of 1.7 mM was added to the sample in order to substitute inhibitor molecules which might have been removed from the

cathepsin during purification and dialysis. The cathepsin V-calpeptin complex was then centrifuged and crystallized at the same conditions as cathepsin V C25S/A mutant (77 % MPD, 23 % of 60 mM TRIS, pH 8). The crystals grew in a few days. Crystals were harvested after they reached their final size and transferred to the synchrotron Elettra, where their diffraction data was collected.

5.3.3 Data collection and structure determination

The complex crystallized in space group $P4_32_12$. Data was collected from the best diffracting crystal, which diffracted up to 1.3 Å. Crystal phases were obtained from the cathepsin part of the cathepsin V-peptide structure which had the most similar unit cell parameters. After initial refinement, the continuous electron density in the difference map in the active site cleft of cathepsin suggested that the inhibitor was attached to the catalytic Cys 25 residue. The inhibitor, generated in PURY, and solvent molecules were then added to the model and the model was further refined. The refinement staggered at the value of R-work around 0.172 (R-kick 0.196). Data collection and refinement statistics are provided in Table 7. The structure was deposited to PDB server and was given the entry code 7QGW.

The crystal asymmetric unit is composed of two molecules of cathepsin V-inhibitor complexes with the RMSD of their CA atoms 0.30 Å (Figure 11). Their chains were resolved across their entire sequence. Calpeptin covalently modified both active sites of cathepsin V. As expected, the covalent bond was formed between SG atom of Cys 25 and aldehyde warhead of calpeptin. Its binding was further stabilized by three H-bonds, formed between amide part of the inhibitor and carbonyl O of the D163 and amide N of the G68 residues and between nitrogen of the carbamate part of the inhibitor and O of the G68 residue. The alkyl and leucine groups occupied the S1 and S2 subsites, respectively, in a substrate-like manner. The benzyl part of the inhibitor bound in the S3 binding area between residues F69, R72 and Q63. The average B-factor of an inhibitor was 20, which is comparable to the B-factors of cathepsin residues in the inhibitor vicinity, indicating that calpeptin covalently modified all cathepsin molecules.

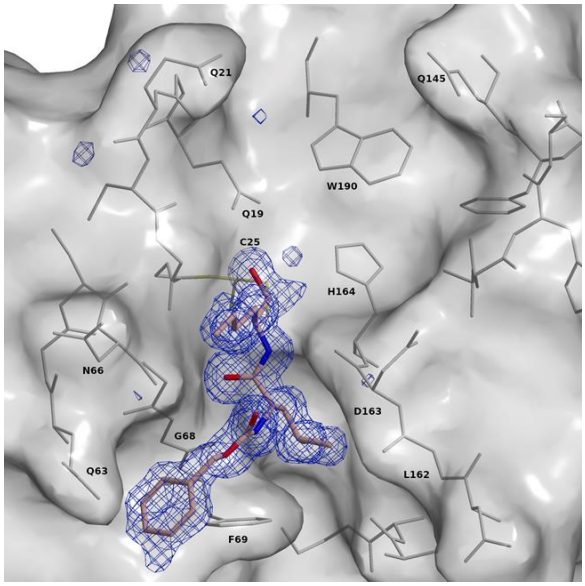


Figure 11: Crystal structure of cathepsin V-calpeptin. Calpeptin molecule is shown with a stick model on the surface of a semi-transparent cathepsin V structure. Cathepsin V residues that constitute the active site cleft are shown with a stick model in grey. Oxygen, nitrogen and carbon atoms of calpeptin are shown with red, blue and rose, respectively. Electron density of calpeptin was calculated with maximum-likelihood averaged kick omit map (*Fo-Fc*) and is shown with a blue mesh, contoured at 4.5 σ. Cathepsin key residues are written out.

Table 7: Crystallographic table for cathepsin V-calpeptin entry 7QGW.

Data collection statistics (last shell)	
Resolution range (Å)	47.12 - 1.303 (1.35 - 1.303)
Space group	P 43 21 2
Unit cell	94.242 94.242 126.956 90 90 90
Total reflections	3293905 (217958)
Unique reflections	139006 (13653)
Multiplicity	23.7 (16.0)
Completeness (%)	99.91 (99.16)
Mean I/sigma(I)	32.94 (2.56)
Wilson B-factor	15.66
R-merge	0.0548 (1.023)
CC1/2	1 (0.837)
Refinement statistics (last shell)	
Reflections used in refinement	138983 (6775)
Reflections used for R-kick	138983 (6775)
R-work	0.1718 (0.2454)
R-kick	0.1956(0.2775)
RMS(bonds)	0.017
RMS(angles)	1.83
Ramachandran favored (%)	97.26
Ramachandran allowed (%)	2.74
Ramachandran outliers (%)	0.00
Rotamer outliers (%)	2.14
Average B-factor	25.89

Chapter 6

Discussion

6.1 Structural Basis for Heterogeneous and Homogeneous Positions of Cathepsin Substrates

Analysis of crystal structures of substrate-mimicking inhibitors in complexes with papain-like cysteine proteases established that substrate binding is facilitated by a conserved hydrogen bonding network between the main chains of the peptidyl substrate and of cathepsin residues G68 at the S2 binding site, the N-terminal amino group of catalytic residues C25 and Q19 side chain as anchors in the S1 subsite, and the side chain of W190 at the S1' binding subsite of cathepsins. These anchors are equivalent in all cysteine cathepsins and thus provide a conserved docking surface to the main chain backbone of peptidyl substrates from P2–P1' in all known cysteine cathepsins. The rest of the interactions, which refer to the side chain inside and outside this stretch, as well as the main chain binding outside this stretch, are not conserved (D. Turk et al., 1998). This analysis expands this view from two decades earlier. Statistical analysis of proteome cleavages enabled us to pinpoint the positions on the substrate chain that exhibited heterogeneous and homogeneous residue compositions. The question seeking answer here is how this behavior reflects the structural features of enzymes, whose interplay with substrates results in each protease cleaving different substrate sequences, and sometimes sharing the same cleavage sites. In our analysis only endopeptidases were considered: Cathepsin B was not considered because of its evolutionary (by sequence, structure) and biochemical (endopeptidase and carboxydipeptidase) distances from endopeptidase cathepsins V, L, K, F, and S (Zhou et al., 2015). The structural features of cathepsin F were also not analyzed because there was only one structure available in the PDB at the time of writing.

In the cathepsin V peptidyl complexes presented here, the conserved docking surface containing the subsites from S2–S1' is rigid, whereas the side chains of cathepsin V residues outside this region exhibit more than one conformation. This was observed in at least one of the two structures in the asymmetric unit, which is associated with a unique contact with the peptide residue. The heterogeneity of the P2 position is evidently a consequence of the prevalence of hydrophobic residues in the three clusters of cathepsin V (V1, V3, and V4). A similar prevalence of hydrophobic residues at P2 was also observed in several clusters of cathepsin K, L, and S (Figure A-1). On the non-primed side of the cysteine cathepsin structures, only the S2 subsite is in the shape of a pocket surrounded by residues, which define the S2 binding surface. The deviations from the pure hydrophobic character of the P2 side chains are evident in the structures of peptides that belong to a group of pattern II, which contains several Lys residues as well as His, Thr and Gln. Partial non-specific behavior at the P2 position is likely cathepsin structure independent because the

observed position of the Lys side chain NZ atom in the GNYKEAKK peptide structure is stabilized by a hydrogen bond formed with the main chain carbonyl of the cathepsin V L162 residue. The S1 subsite seems to bear no specific structural reason for Lys specificity, apart from the P1 residue side chain pointing toward the solution and thus excluding large and bulky hydrophobic residues, which can also be observed in the clusters (Figure A.2). Indeed, we observed that S1 can chemically bind different types of residues, such as Lys, Glu, Ala, Ser, Leu, and Pro (Table 2).

On the primed side of the active site cleft, the P1' composition of substrates of cathepsins V, L, and F was found to be heterogeneous in contrast cathepsins K and S, which were found to be homogeneous. The dominant residue at the heterogeneous P1' position was Lys (Figure A-1). The structural basis for this behavior in the S1' binding subsites of cathepsins V, L, and F appears to be the rigid aspartic residue D163 that interacted with the guanidinium group of Arg at P1' in the RLSAKP complex (Figure 12). In cathepsin K and S, the equipositioned residues are N161 and N163, respectively; therefore, their interaction with the positively charged Lys is not as strong nor as specific, thus rendering these positions homogeneous (Figure 13, Figure 14).

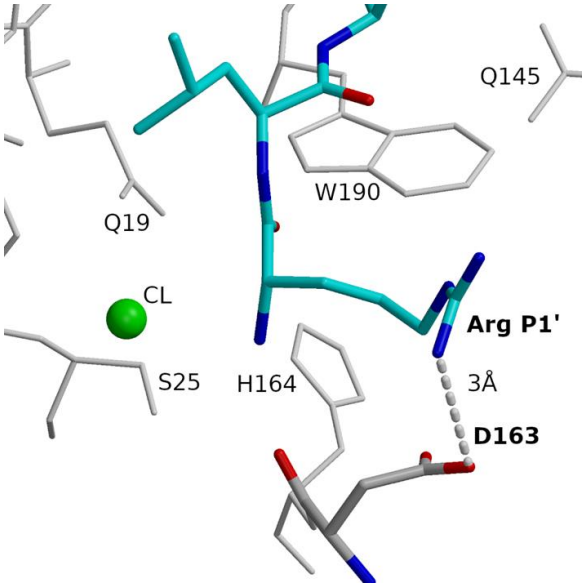


Figure 12. Binding specificity between Arg at P1' and D163 of cathepsin V. Arg and Leu of peptide RLSAKP (non-protected), bound at subsites S1' and S2', are shown in the bond model in blue (nitrogen), red (oxygen), and cyan (carbon). Aspartate at position 163 is shown in the bond model in blue (nitrogen), red (oxygen), and grey (carbon). Hydrogen bond is shown with a dashed line. Neighboring residues and the chlorine ion are also provided. Figure was prepared using MAIN (Dušan Turk 2013) and rendered using Raster 3D (Merritt and Bacon 1997).

The homogeneity of the cathepsin V substrate positions outside the P2 – P1' range suggests the structural basis of the adaptability of the underlying structure and capability of the peptide ligands to find appropriate anchors. The flexible side chains shown in blue (Figure 13, a) provide a versatile binding surface capable of adapting to the binding of different ligand sequences. The ambivalence of the Gln and Asn side chains, which provide hydrogen donors and acceptors, is well suited for adaption to various substrates. Binding adaptability is provided by residues Q63 and N66 at S3 and S4, N161, which assisted in the binding of two Lys residues at S4 and F69, whose side chains in the three structures replaced the Cl-ion at the bottom of the S2 pocket, and on the primed side by residues Q145 at S2' and

S4' and Q21 at S3'. Furthermore, the binding geometries of the peptides in the regions outside the P2 – P1' range diverged to the extent that the S4 and S4' surfaces were found in two separate areas on the left and right of the active site cleft (Figure 13, c).

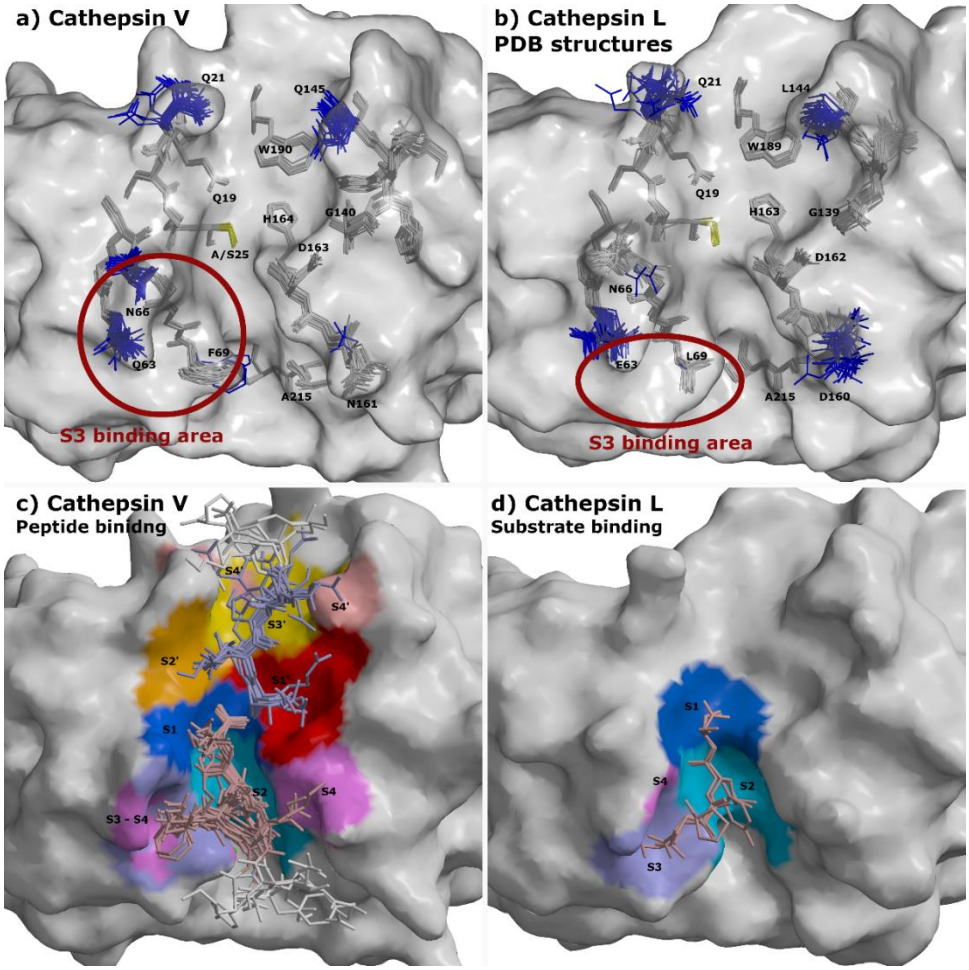


Figure 13: Flexible and rigid residues of cathepsins V and L and their substrate-binding areas. **a)** Superimposed cathepsin V complexes. Flexible cathepsin V residues that provided a versatile binding area for peptide binding are shown with a stick model in blue on the surface of a semi-transparent cathepsin V structure. Rigid residues are shown in grey. Key residues are written out. Red circle depicts the S3 binding area. Catalytic residues at site 25 are shown in yellow. **b)** Superimposed structures of cathepsins L from PDB database with the equivalent labeling (PDB entries 1CJL, 1CS8, 1ICF, 1MHW, 2NQD, 2XU1, 2XU3, 2XU4, 2XU5, 2YJ2, 2YJ8, 2YJ9, 2YJB, 3BC3, 3H89, 3H8B, 3H8C, 3HHA, 3HWN, 3K24, 3KSE, 3OF8, 3OF9, 4AXL, 4AXM, 5F02, 5MAE, 5MQY, 6EZP, 6EZX, 6F06, 6JD0, and 6JD8). The structure of C25A mutant with $SO4_2^-$ ion in the active site (3IV2) is not included due to the distorted active site. Red elipse depicts the S3 binding area. Catalytic residues at site 25 are shown in yellow. **c)** Superimposed cathepsin V-peptide complexes. Peptides are shown with a stick model on the surface of a cathepsin V structure. Binding areas of peptides at positions from P4-P4' are shown in color spectra from blue to magenta at the non-primed side and from red to rose at the primed side. Peptide residues from P1-P4 and from P1'-P4' are shown in pale pink and pale blue, respectively, whereas the residues beyond P4 and P4' are shown in white. **d)** Processed peptide and protein substrates of cathepsin L structures (3K24 and 5I4H,

respectively) at the non-primed side. Their binding areas are presented with the same coloring annotation as in pannel c.

In contrast to other cathepsins, the P3 position of cathepsin L is heterogeneous. At P3, the L2 cluster exhibited an almost purely hydrophobic profile with a small contribution of aromatic residues (Figure A.2). The cathepsin L residue L69 forms part of the S3 surface area (Figure 13, b and d) and appears to be locked in position, in contrast to F69 in cathepsin V, which appears in two conformations (Figure 13, a). In cathepsin L, this feature provides space for the P3 residue to bind along L69 and below E63. In contrast, the binding surface of most P3 residues in cathepsin V complexes is provided by the flexible Q63 and N66 side chains, which can provide either hydrogen bond donor or acceptor groups or not when turned away (red circle in Figure 13, a). Consequently, cathepsin V substrates were able to adopt two conformations (Figure 13, c) in contrast to cathepsin L, where the main chain trace of substrates was in one conformation only (Figure 13, d). Therefore, the structural background of the heterogeneous position of cathepsin L can be explained by the features of the S3 binding area (elliptical red circle in Figure 13, b) positioned below its flexible residues E63 and N66. The homogeneity of P3 for cathepsin S and K substrates can likely be explained by the similarity of cathepsin K Y67 and cathepsin S F70 to the equipositioned cathepsin V F69 and the flexibility of cathepsin K D61 and cathepsin S K64 residues (Figure 14).

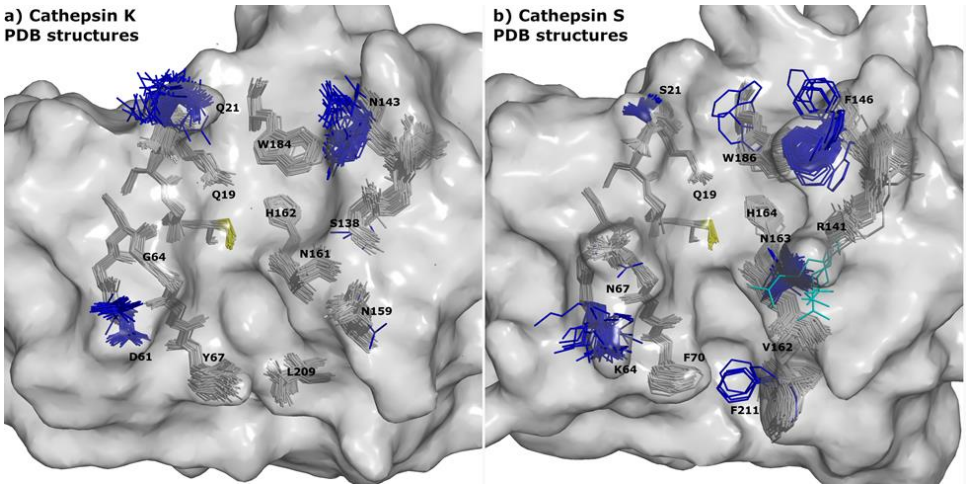


Figure 14: Flexible and rigid residues of cathepsins K and S from PDB database. Superimposed structures of cathepsins K and S from PDB database. Flexible cathepsin residues are shown with a stick model in blue on the surface of a semi-transparent cathepsin structure. Rigid residues are shown in grey. Key residues are written out. Catalytic residues at site 25 are shown in yellow. **a)** Cathepsin K (PDB entries 1BGO, 1NLJ, 1ATK, 1AU0, 1AU2, 1AU3, 1AU4, 1AYU, 1AYV, 1AYW, 1BY8, 1MEM, 1NL6, 1Q6K, 1SNK, 1TU6, 1U9V, 1U9W, 1U9X, 1YK7, 1YK8, 1YT7, 2ATO, 2AUX, 2AUZ, 2BDL, 2F7D, 2FTD, 2R6N, 3C9E, 3H7D, 3KW9, 3KWB, 3KWZ, 3KX1, 3O0U, 3O1G, 3OVZ, 4DMX, 4DMY, 4N79, 5N8W, 4X6H, 4X6I, 4X6J, 4YV8, 4YVA, 5J94, 5JA7, 5JH3, 5TDI, 5TUN, 5Z5O, 6ASH, 6HGY, 6PXF, 6QBS, 6QL8, 6QLM, 6QLW, 6QLX, 6QM0, 7NXL, 7NXM, and 7PCK. **b)** Cathepsin S (PDB entries 2HXZ, 2F1G, 2FT2, 3OVX, 2R9N, 2R9M, 1MS6, 2HHN, 4P6G, 2OP3, 4P6E, 6YYN, 2HH5, 2G7Y, 2FQ9, 6YYR, 6YYP, 2H7J, 2FRQ, 3N3G, 2R9O, 2FRA, 6YYO, 3N4C, 2FUD, 1NPZ, 1NQC, 5QC0, 5QCH, 5QC4, 5QC2, 2C0Y, 5QCG, 5QCE, 5QCI, 5QCC, 5QCA, 5QC7, 5QBV, 5QC5, 5QBZ, 5QBX, 5QC9, 5QC3, 5QC1, 3IEJ, 5QCF, 5QCJ, 5QCD, 5QCB, 5QBW,

5QC8, 5QC6, 5QBU, 2G6D, 5QBY, 1GLO, and 2FYE). Only three residues of flexible R141 are presented in cyan (entries 2FUD and 5QCA) for clarity of the figure.

Collectively, the selectivity of the substrate-binding subsites in cysteine cathepsins analyzed is not absolute; however, it renders the binding positions statistically heterogeneous and thereby exposes the preferential binders. For the heterogeneous positions, we were able to pinpoint the residues at the cathepsin surface, which with specific interactions directly contributed to heterogeneity of the binding subsites, and also to the structural differences among cysteine cathepsins, which endow them with similar and different subsite specificities. Therefore, there are structural restraints that drive local selectivity of interactions, resembling the lock and key mechanism (Fischer, 1894), and there is flexibility that provides a basis for the promiscuity of local interactions, resembling the induced fit (Koshland, 1958) and conformational selection models (Gáspár & Csermely, 2012).

6.2 Peptides as Protein Substrate Model

The structural and cleavage analysis combined enabled us to compare cathepsin recognition of the same sequences when presented in protein and peptidyl form. Regarding the crystal structures, only peptides of groups I and IV (8 examples) bound to cathepsin V are consistent with the observed protein cleavage sites, whereas the peptides of groups II and III deviated from the expected binding (18 examples). Apparently, their binding was biased by the amide protective group at the carboxylic end of peptides that belong to the group of pattern II and by the absence of a negative charge at the reactive site Cys, which led to shifted binding of peptides that belong to a group of pattern III (Table 2, Figure 4). Moreover, the IILKEK peptide bound in a non-substrate-like manner to the non-primed side of molecule B, and in the structure of peptide VYEKKP, there was a third peptide molecule bound at the interface of the two cathepsin V molecules (Figure B-5). Besides, five peptides bound differently to A and B molecules of cathepsin V and one peptide bound differently when it was co-crystallized or when soaked. Furthermore, the span of average B-factors of peptides at positions P3 – P3', where binding was conserved, was in the range of 20 – 75, whereas the cathepsin residues in their vicinity had B-average values around 20 – 25. Together, this shows that binding of peptides to cathepsin V was rather weak.

Observations of the interactions between cysteine cathepsins and protein substrates are scarce, but the crystal structure of the catalytic site mutant C25S of cathepsin L provides an excellent comparison with the peptide complexes presented here (Sosnowski & Turk, 2016). In this structure, one cathepsin L molecule cleaved the neighboring molecule in the crystal such that the residues of the non-primed side remained bound to the active site cleft. In contrast to the other structures of cathepsin L in complex with the peptide (Adams-Cioaba et al., 2011) and the cathepsin V complexes presented here, the substrate chain of cathepsin L enters the active site cleft from above, forming a helical turn at P4 and continuing along the S3, S2, and S1 subsites to the reactive site (Figure 15). This example illustrates how the proteins and peptides may approach protease active site in a different manner. In agreement with this, we exposed several examples of distinct processing of the same sequences when they were presented in protein or peptidyl form, by determining their cleavage sites with cathepsins K, V, and L (Appendix C).

To summarize, the extent to which the same sequences are recognized differently appears to at least be affected by the fit of substrate primary sequence itself to the protease active site, resulting in stronger interactions, which are otherwise quite subtle, as well as by other factors like the 3-D arrangement of the sequence which is being cleaved and its neighborhood, which likely affects its approach to the active site.

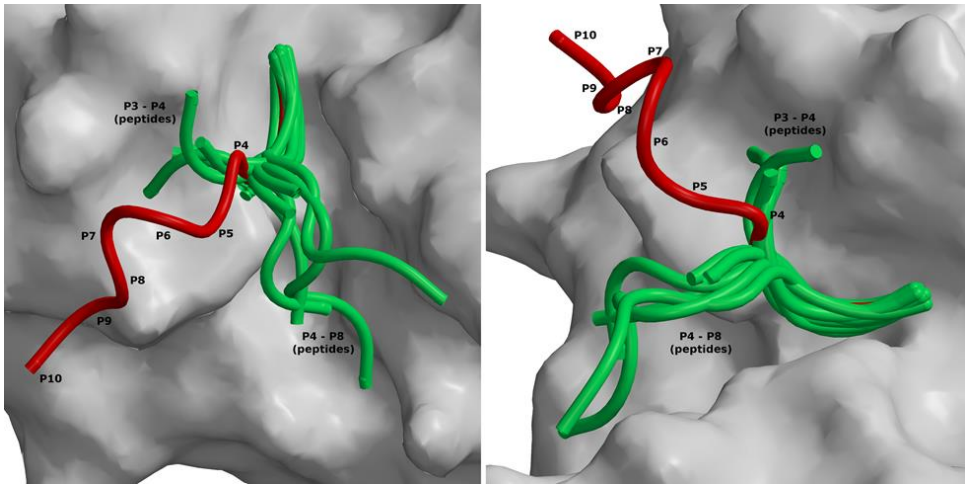


Figure 15: The difference between peptide and protein binding to cathepsins. The main chains of peptides bound to the non-primed side of cathepsin V complexes are presented with green ribbons. Protein substrate fold from cathepsin L structure 5I4H, from sequence E96:S105 which bound in the non-primed side at positions from P1–P10, is presented with red ribbons. Labels are provided from P4 onward, where protein and peptide folds begin to diverge. Image is presented from two viewpoints.

6.3 Relevance for Drug Discovery Projects

In this study we exposed crucial structural areas, together with contributing residues, which participate in substrate binding along the whole active site cleft, from S4 – S6'. Some of those have already been targeted in previous work, either on purpose or unknowingly, and their interactions have been described in several crystal complexes. For example, the L209 of cathepsin K which narrows its S2 pocket and Y67 and D61 which participate in binding of P3 fragment (DesJarlais et al., 1998; Gauthier et al., 2008; C. S. Li et al., 2006; McGrath et al., 1997; Robichaud et al., 2004; Yamashita et al., 2006); F211 of cathepsin S, whose rotation creates opening of its S2 pocket, as well as the K64 and R141 which might be responsible for accommodating the positions P3 and P1, respectively, (Alper et al., 2006; Gauthier et al., 2007; Liu et al., 2005; Tully et al., 2006); the S2 pocket of cathepsin L which prefers fragments like Tyr, as well as the cathepsin L residues Y72 and E63 at S3 site and hydrophobic patch on the D-domain of primed side formed by residues A138, G139, Y189 and L144 (Chowdhury et al., 2008; Dana et al., 2014; Marquis et al., 2005; Shenoy & Sivaraman, 2011). So, on the one hand, our results are validated by this data, and on the other, they expose additional key regions and residues that can be exploited in the development of more potent and more selective inhibitors.

In this regard, we observed the difference in binding of P3 fragment of calpeptin inhibitor to the cathepsins V and L (Figure 16) which is in agreement with heterogeneous distribution of cathepsin substrates at P3 position (Figure 13, b and d; Figure A-1). The phenyl ring of the benzyloxycarbonyl group at S3 of cathepsin L complex binds along the L69 into the hydrophobic patch, whereas in the cathepsin V complex, the F69 directs its binding toward residues Q63 and N66 in a similar manner that it directed binding of peptides to cathepsin V (Figure 13, c).

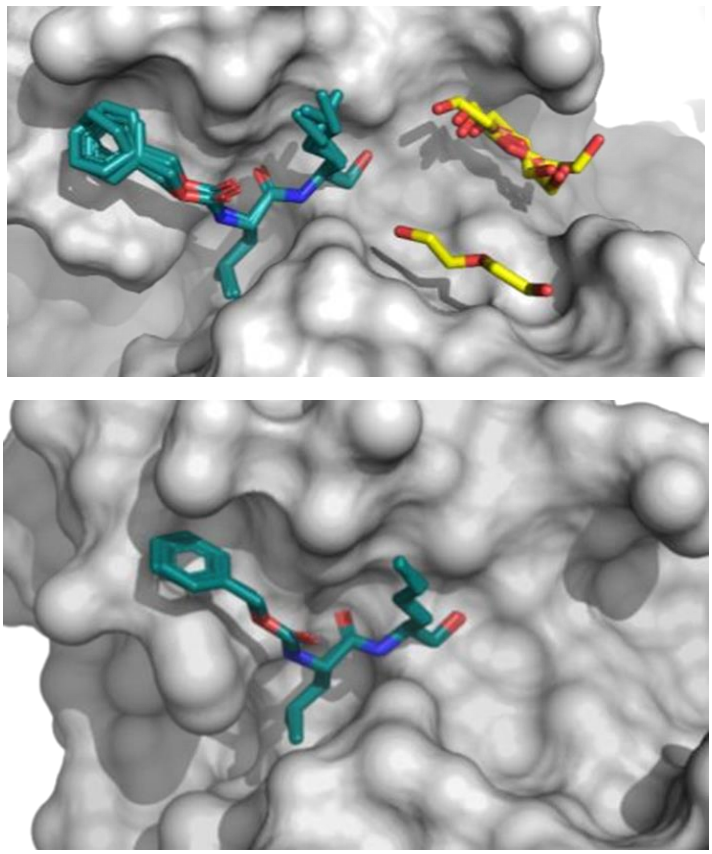


Figure 16: Comparison of calpeptin binding at S3 site of cathepsins L and V. The figure shows crystal structures of cathepsin L – S-calpeptin complex (upper image, PDB entry 7QKC) and cathepsin V-calpeptin complex (bottom image, PDB entry 7QGW). Calpeptin is presented with a stick model on top of the cathepsin surface. Oxygen, nitrogen and carbon atoms of calpeptin are shown in red, blue and green, respectively. PEG molecules from the solvent in the primed side of cathepsin L-calpeptin complex are shown in a yellow-red combination.

Chapter 7

Conclusions

The combined approach of proteomic, statistical and structural analysis enabled us to elucidate the elusive specificity of cathepsin endopeptidases. The key was to obtain enough cathepsin cleavage data which enabled us to differentiate between heterogeneous and homogeneous substrate positions, and to cluster them on heterogeneous positions, which yield a well-defined and separated subset of a few major representative clusters for each cathepsin. This enabled us to choose a limited set of sequences, representing the variety of all seven cathepsin V major clusters, and crystallized them in complexes with cathepsin V. The crystal structures revealed that the heterogeneous and homogeneous positions of peptide substrates bind to rigid and flexible surface areas on the cathepsins, respectively, and revealed structural features that cause specific behavior on heterogeneous positions P3 (cathepsin L) and P1' (cathepsins V, L, and F) (hypothesis I and II confirmed).

In the crystal structures, peptides bound to cathepsin V in four different ways. Peptides in groups I and IV (8 examples) bound in accordance with the observed cleavages of protein substrates, whereas peptides binding in groups II and III (18 examples) were shifted, hence, their binding did not correspond to the assigned clusters. Thus, we could not analyze the binding specificities of peptides that belong to the same cluster.

Comparison of peptide binding observed in this structural study to the binding of processed protein substrate suggests that the approach of the sequence that is being cleaved to the active site is affected by the protein structure. In addition, we exposed several examples of distinct processing of the same sequences by wild-type cathepsins when they were presented in the peptidyl or in the protein substrate form. Hence, the peptides are not always a reliable model for studies of protein interactions (hypothesis III confirmed).

The characterization of cathepsin inhibitors provided new insight that may facilitate the development of novel therapeutics. The alkyne-based inhibitors of cathepsin K exhibited no indiscriminate thiol reactivity, yet they were able to selectively modify the catalytic Cys residue in the active site of cathepsin K. This showed that alkynes can be used as a latent electrophilic group for specific enzyme targeting. In addition, we showed that calpeptin and compounds alike strongly inhibit human cathepsins K, V, and L, which suggested that the suppression of SARS-CoV-2 activation by calpeptin may be mediated through cathepsin inhibition (hypothesis IV and V confirmed).

To summarize, this work provides new insight and understanding of the complex biology of cathepsins. This opens new ways in design of novel selective cathepsin inhibitors. Moreover, the methodology that was used to decipher the elusive cathepsin specificity applied in this study is not limited to cathepsins, but can be used to characterize any given protease. In addition, we proposed that cathepsins are involved in activation of SARS-CoV-2 and established that an alkyne functional group can be used as a safer alternative to other electrophilic warheads in the design of irreversible small molecule inhibitors.

Appendix A

Peptide Selection

A.1 List of Synthesized Peptides

Table A-1: List of synthesized peptides. Peptides were from 6–11 residues long. Their sequence is inserted in the columns at positions from P5–P6'. The cleavage site is between P1–P1'. The cleavage site cluster for cathepsin V is presented under the cluster column (for example, V1 for cathepsin V cluster 1). Residues that are shaded represent dominant residues in the corresponding cathepsin V cluster. Cleavage area: several neighboring cleavages. Positional cleavage: one cleavage site in the area. Indices 1 and 2 refer to one or two separated cleavage areas in the originated protein, respectively, where cathepsin acted. Indices 3 and 4 mark where there were only one or two cleavage sites in the originated protein, respectively. Termini of most peptides were protected with N-acetylation and C-amidation (marked "Y" in protection column). Some peptides were also synthesized without protection (marked "N") or with and without protection (marked "Y/N"). Peptides have UniProt codes (www.uniprot.org) of their corresponding proteins. *These are proteins with Uniprot codes Q6S8J3, P0CG38, P0CG39 and Q9BYX7. ** The amount of peptide was sufficient to carry out structural analysis only. *** These are proteins with Uniprot codes Q6S8J3, P60709, P63261, A5A3E0, P0CG38 and P0CG39 **** Peptide sequences were not obtained from the protein cleavages.

Number	P5	P4	P3	P2	P1	P1'	P2'	P3'	P4'	P5'	P6'	Cluster	Cleavage type	Protection	UniProt code
p1		T	C	L	C	Q	V	P	Q			**V1**	Positional	Y	P49588
p2		I	L	L	T	E	A	P	L			**V1**	Area	Y	multiple*
p3		K	D	L	L	H	P	S	P			**V1**	1 Area	Y	P42677
p4	E	I	D	L	R	N	P	K	G	N		**V1**,	Area	Y	P27695
p5		Q	L	L	V	A	C	K	V	K		**V1**,	Positional	Y	Q9Y490
p6		K	V	L	A	T	V	T	K			**V1**,	Area	Y	Q02878
p7			R	L	S	A	K	P				**V1**,	Area	Y/N	Q15651
p8			L	L	S	G	K	E				**V1**,	Area	Y/N	A6NHL2
p9			Q	L	R	Q	Q	E				**V1**,	4 Positional	Y/N	O43818
p10		G	N	Y	K	E	A	K	K			**V2**	Positional	Y	P42704

p11		V	L	L	K	V	A	A	S			**V2**,	3 Positional	Y	Q53FA7
p12		A	C	M	K	S	V	T	E			**V2**,	Area	Y	P63104
p13		V	A	C	K	S	S	Q	P			**V2**,	Positional	Y	P46013
p14			G	A	K	S	A	A				**V2**,	Area	Y/N	Q8NC51
p15**			G	V	T	K	A	A				**V3**,	Area	Y	P27797
p16			G	M	C	K	A	G				**V3**	Area	Y	multiple***
p17			K	I	A	K	T	H				**V3**	Positional	Y	O75533
p18			E	V	C	K	K	K	K			**V3**	3 Positional	Y	Q92772
p19			I	I	L	K	E	K				**V3**	3 Positional	Y/N	P07199
p20		R	G	I	R	E	A	A	K			**V4**	Positional	Y	P25398
p21		K	R	F	Q	N	V	A	K			**V4**	Area	Y	P14625
p22			A	Y	F	K	K	V	L			**V5**	Area	Y	P25205
p23			V	Y	E	K	K	P				**V5**	Area	Y/N	P46777
p24		S	I	Y	E	V	D	K	Q			**V6**	Positional	Y	Q92747
p25		T	R	E	S	E	D	L	E			**V6**	3 Positional	Y	Q8N5V2
p26		V	P	C	G	T	A	H	E			**V6**	Positional	Y	O43823
p27		K	K	Y	D	A	F	L	A			**V6**	Positional	Y	P62906
p28			A	W	K	K	E	A				**V7**	4 Positional	Y	Q9C0B0
p29			P	V	K	K	K	A	K			**V7**	Area	Y	P16402
p30			K	P	K	K	K	T	K			**V7**	Area	Y	Q6NWY9
p31			L	L	K	V	A	L						N	****
p32			A	V	A	E	K	Q						N	****
p33			A	L	A	A	S	S						N	****
p34			A	V	R	A	R	L						N	****
p35			L	L	K	A	V	A	E	K	Q			Y	****

A.2 Major Clusters of Cathepsin Substrates

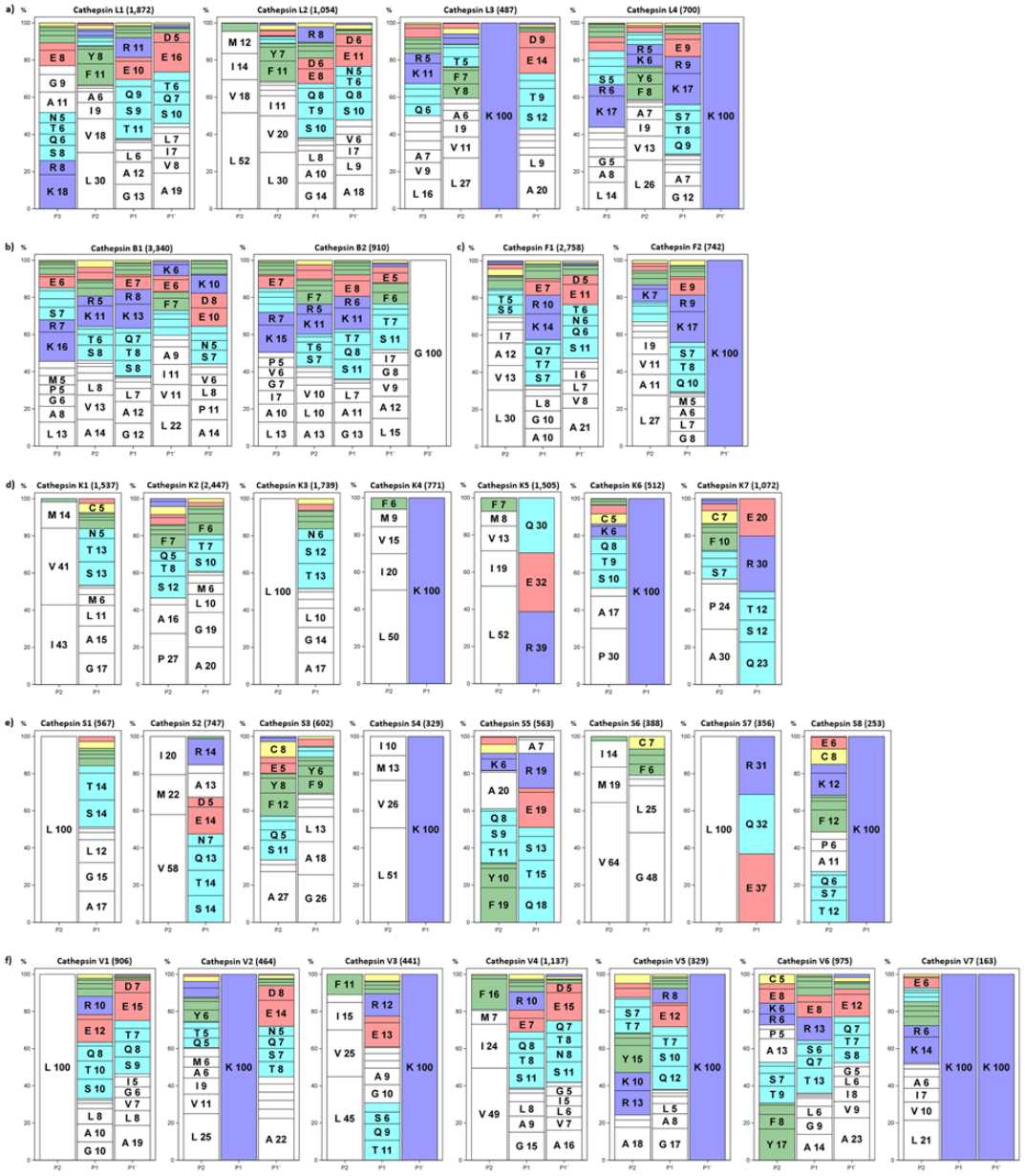


Figure A-1: Major clusters of cathepsin substrates. Upper row, cathepsin L; second row from the top, cathepsins B (left) and F (right); third row from the top, cathepsin K; fourth row from the top, cathepsin S; bottom row, cathepsin V. Each frame describes one cluster. The description of the cluster is based on heterogeneous positions (on x-axis) and share of a certain aminoacid or a group of aminoacids at an individual heterogeneous position on y-axis (in %). Their shares are proportional to the the height of the belonging rectangle. Aminoacids are identified by one letter code. The background color represents the type of amino acid residue: hydrophobic, white; hydrophilic, cyan; basic, blue; acidic, red; aromatic, green; cysteine and methionine, yellow. The figures were generated using SAS for Windows (SAS Institute).

Appendix B

Crystal Structures of Cathepsin V-Peptide Complexes

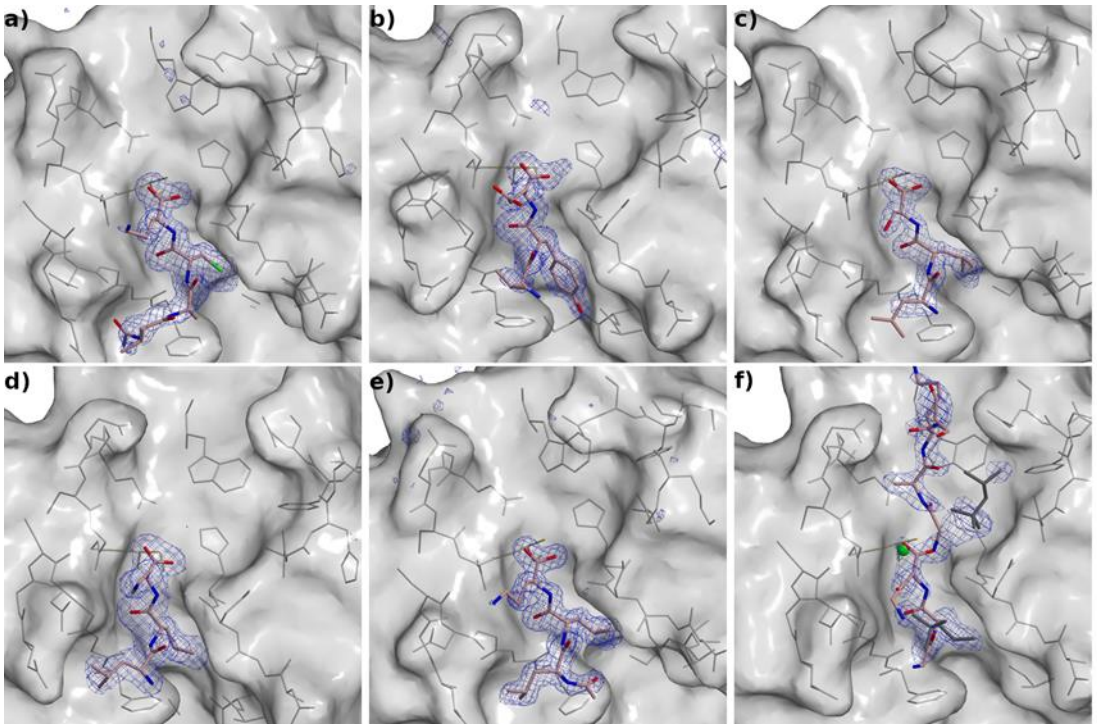


Figure B-1: Electron density maps of peptides in the group of pattern I. Peptides are shown with stick model on the surface of a semi-transparent cathepsin V structure. Peptide oxygen, nitrogen and carbon atoms are shown in red, blue and pale pink, respectively. a) Fragment VACK of peptide VACKSSQP. b) Fragment VYE of peptide VYEKKP. c) Fragment LLS of peptide LLSGKE. d) Fragment LLK of peptide LLKVAL. e) Fragment LLK of peptide LLKAVAEKQ. f) Peptide GAK of peptide GAKSAA is shown in the non-primed site (primed site is occupied by another peptide GAKSAA, which belongs to a group of pattern IV). Electron densities were constructed using maximum-likelihood averaged kick omit maps (*Fo-Fc*) (Pražnikar et al., 2009) and are contoured at 4.5 σ. The masks are presented around peptides. Figures in the panel were prepared using MAIN (D. Turk, 2013) and rendered using Raster 3D (Merritt & Bacon, 1997).

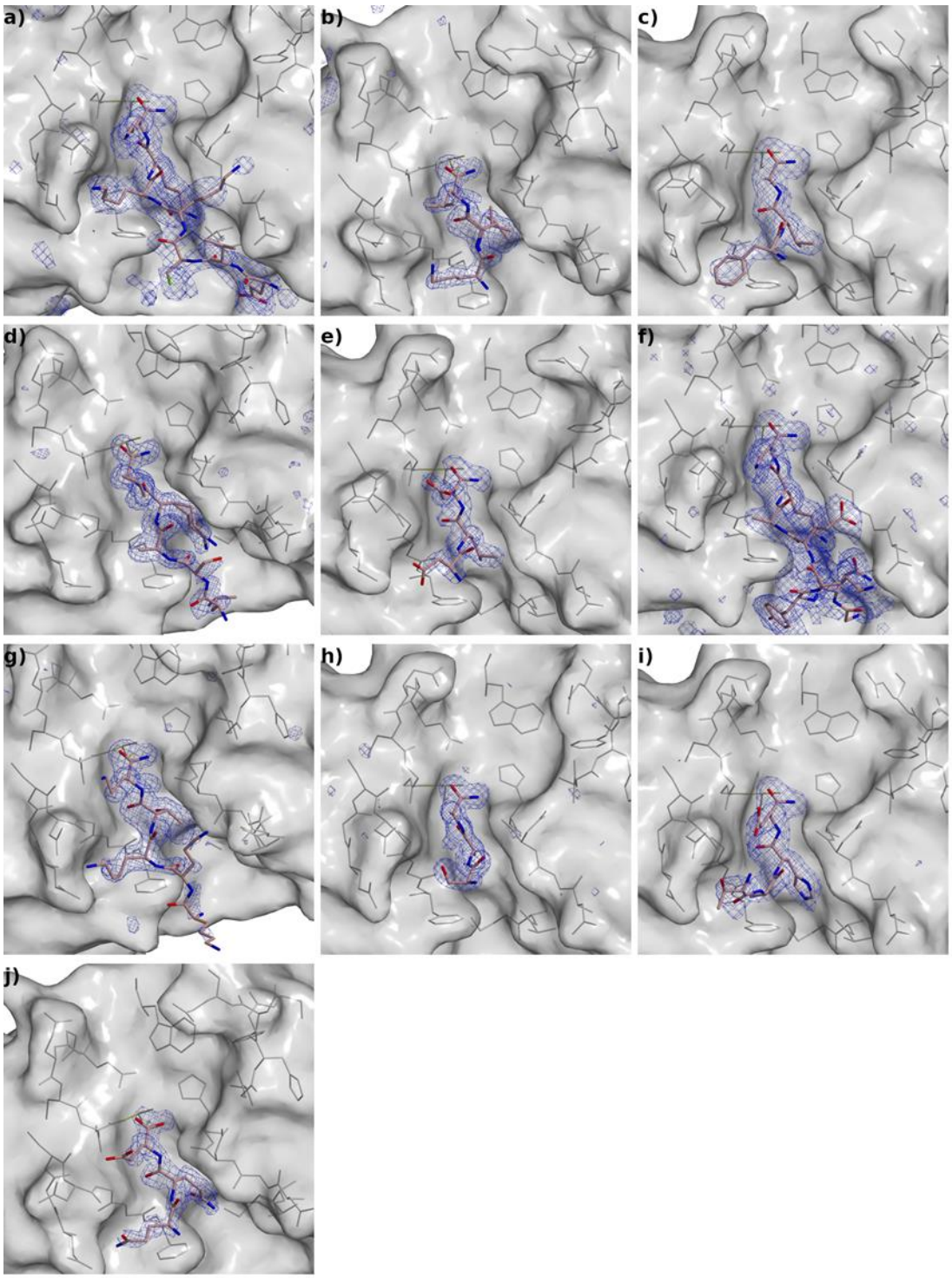


Figure B-2: Electron density maps of peptides in the group of pattern II. Peptides are shown with a stick model on the surface of a semi-transparent cathepsin V structure. Peptide oxygen, nitrogen and carbon atoms are shown in red, blue and pale pink, respectively. a) Peptide EVCKKKK. b) Fragment KVL of peptide AYFKKVL. c) Fragment FLA of peptide KKYDAFLA. d) Fragment LSAKP of peptide RLSAKP (protected). e) Fragment DLE of peptide TRESEDLE. f) Peptide GNYKEAKK. g) Fragment KKKTK peptide KPKKKTK. h) Fragment SAA of peptide GAKSAA. i) Fragment TAHE of peptide VPCGTAHE. j) Fragment QQE of peptide QLRQQE.

Electron densities were constructed using maximum-likelihood averaged kick omit maps (*Fo-Fc*) (Pražnikar et al., 2009) and are contoured at 4.5 σ. Figures in the panel were prepared using MAIN (D. Turk, 2013) and rendered using Raster 3D (Merritt & Bacon, 1997).

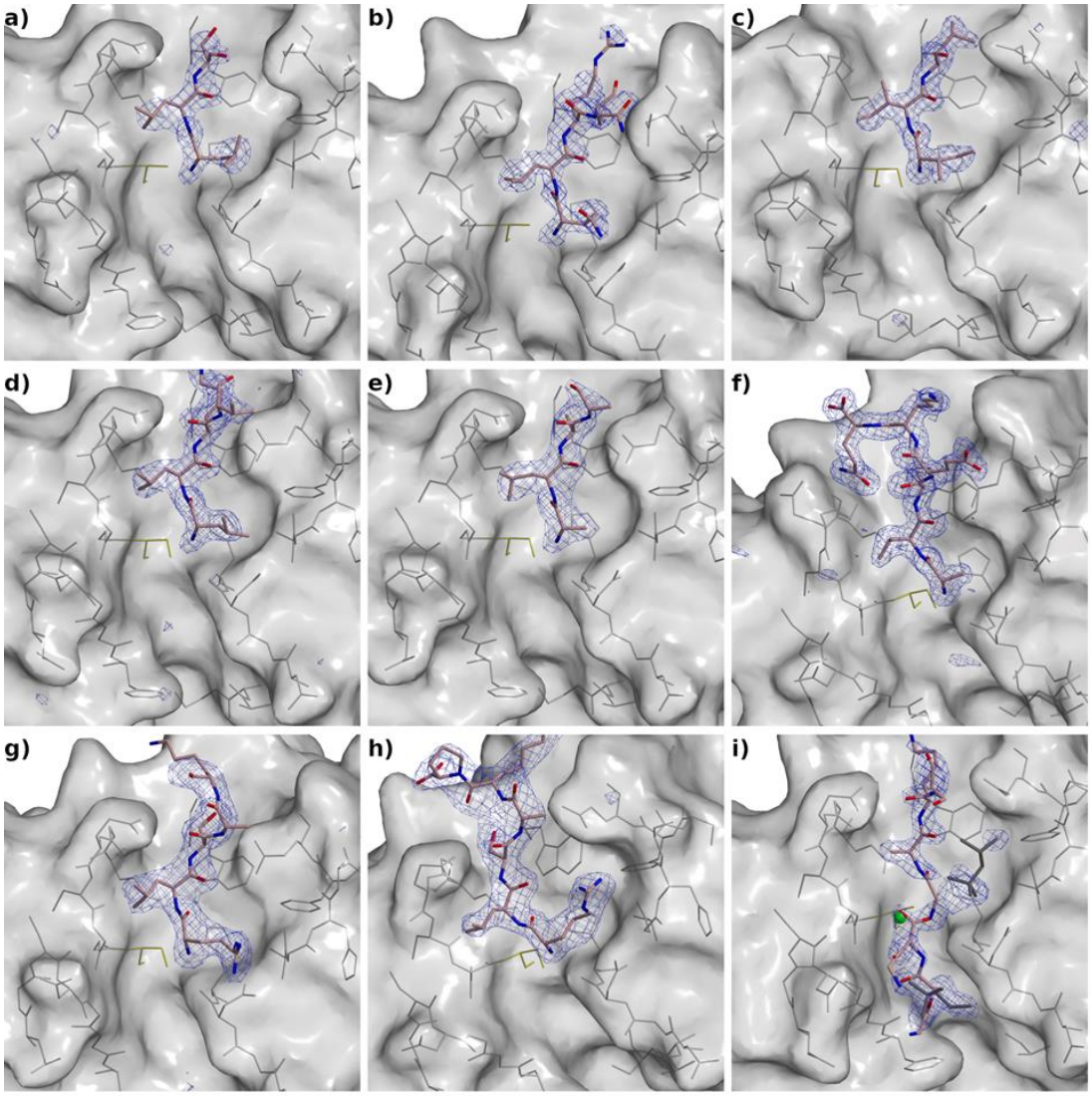


Figure B-3: Electron density maps of peptides in the group of pattern III. Peptides are shown with a stick model on the surface of a semi-transparent cathepsin V structure. Peptide oxygen, nitrogen and carbon atoms are shown in red, blue and pale pink, respectively. a) Fragment LLS of peptide LLSGKE. b) Fragment QLRQ of peptide QLRQQE. c) Fragment IIL of peptide IILKEK. d) Fragment LLKV of peptide LLKVAL. e) Fragment ALAA of peptide ALAASS. f) Peptide AVAEKQ. g) Fragment RLSAK of peptide RLSAKP (non-protected; molecule A). h) Peptide RLSAKP (molecule B). i) Fragment GAKS of peptide GAKSAA is shown in the primed site. Non-primed site is occupied by peptide GAKSAA, belonging to a group of pattern I. Electron densities were constructed using maximum-likelihood averaged kick omit maps (*Fo-Fc*) (Pražnikar et al., 2009) and are contoured at 4.5 σ. The masks are presented around peptides. Figures in the panel were prepared using MAIN (D. Turk, 2013) and rendered using Raster 3D (Merritt & Bacon, 1997).

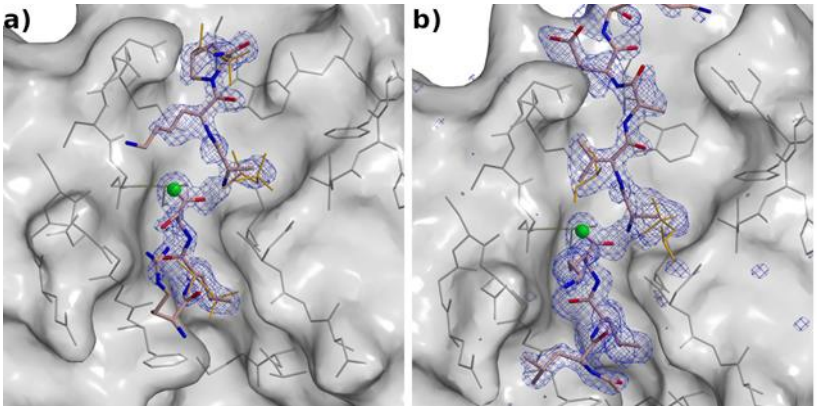


Figure B-4: Electron density maps of peptides in the group of pattern IV. Peptides are shown with a stick model on the surface of a semi-transparent cathepsin V structure. Peptide oxygen, nitrogen and carbon atoms are shown in red, blue and pale pink, respectively. a) Fragments RLS and AKP of peptide RLSAKP. b) Fragments LLK and AVAEKQ of peptide LLKAVAEKQ. Electron densities were constructed using maximum-likelihood averaged kick omit maps (*Fo-Fc*) (Pražnikar et al., 2009) and are contoured at 4.5 σ. The masks are presented around peptides. Figures in the panel were prepared using MAIN (D. Turk, 2013) and rendered using Raster 3D (Merritt & Bacon, 1997).

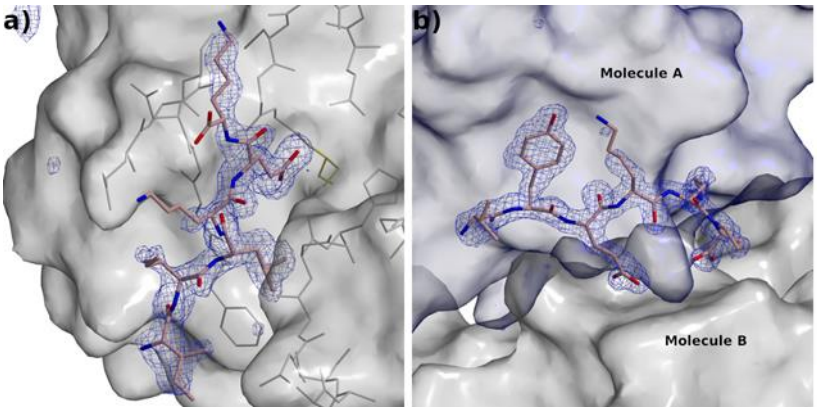


Figure B-5: Electron density maps of peptides bound otherwise. Peptides are shown with a stick model on the surface of a semi-transparent cathepsin V structure. Peptide oxygen, nitrogen and carbon atoms are shown in red, blue and pale pink, respectively. a) Peptide IILKEK bound to cathepsin V like an inhibitor. b) Peptide VYEKKP bound at the surface of two cathepsin V molecules. Electron densities were constructed using maximum-likelihood averaged kick omit maps (*Fo-Fc*) (Pražnikar et al., 2009) and are contoured at 4.5 σ. The masks are presented around peptides. Figures in the panel were prepared using MAIN (D. Turk, 2013) and rendered using Raster 3D (Merritt & Bacon, 1997).

Appendix C

Comparison of Protein and Peptide Cleavages

C.1 List of Protein and Peptide Cleavages and Predictions

Table C-1: The 1st column of the table (UniProt/Type) contains UniProt code of the protein origin (www.uniprot.org) and cleavage type. There are two types of cleavages: positional cleavages (one cleavage in the sequence) and cleavage area (several cleavages in the sequence). The 2nd column specifies whether the cleavages in the sequence in the next three columns were obtained from protein or peptide analysis or from support vector machine (SVM)-based predictions (Tusar, Loboda, Impens, 2023). Cleavage sites are marked with arrows (↓). Asterisks (*) at the end of peptide sequences mark peptides with an unique peptide cleavage. Protein sequences marked with (§) had no observed protein cleavage sites. The sequences marked with (&) were not selected from the cleaved protein sequences.

UniProt /Type	Substrate form	Cathepsin K	Cathepsin L	Cathepsin V
P42677 Area[1]	Peptide	K D L L↓H P S P	K D L↓L↓H P S P*	K D L L↓H P S P
	Protein	K D L L H P S P§	K D L L H P S P§	K D L L↓H P S P
	Prediction	K D L L H P S P	K D L L H P S P	K D L L H P S P
Q6NWY9 Area[2]	Peptide	K P K↓K K T K	K P K↓K↓K T K	K P K↓K↓K T K
	Protein	K P↓K K K T K	K P K↓K K T K	K P K↓K K T K
	Prediction	K P K↓K↓K T K	K P↓K↓K↓K T K	K P K↓K↓K T K
P46777 Area	Peptide	V Y↓E K K P	V Y↓E↓K K P	V Y↓E↓K K P
	Protein	V Y↓E K K P	V Y E↓K K P	V Y E↓K K P
	Prediction	V Y↓E K K P	V Y↓E↓K K P	V Y↓E↓K K P
Q8NC51 Area	Peptide	G A K↓S A A	G A K↓S A A	G A K↓S A A
	Protein	G↓A K↓S A↓A	G A K↓S A A	G A K↓S A A
	Prediction	G↓A K↓S A↓A	G↓A K S A↓A	G↓A K S A↓A
P27695 Area	Peptide	E I D L R↓N P K↓G N	E I D L↓R↓N P K↓G N*	E I D↓L R↓N P K G N*
	Protein	E I D L R N P K↓G N	E I D L R↓N P K↓G N	E I D L R↓N P K G↓N
	Prediction	E I D L R↓N P K↓G N	E I D L R↓N P K G N	E I D L R↓N P K↓G↓N
P16402 Area	Peptide	P V K↓K K A K	P V K↓K K A K	P V K↓K K A K
	Protein	P V K↓K K A K	P V K K K A K§	P V K↓K↓K↓A K
	Prediction	P V K↓K↓K↓A K	P V K↓K↓K↓A K	P V K↓K↓K↓A K
Q15651	Peptide	R L S↓A K P	R L S↓A K P	R L↓S↓A K P*

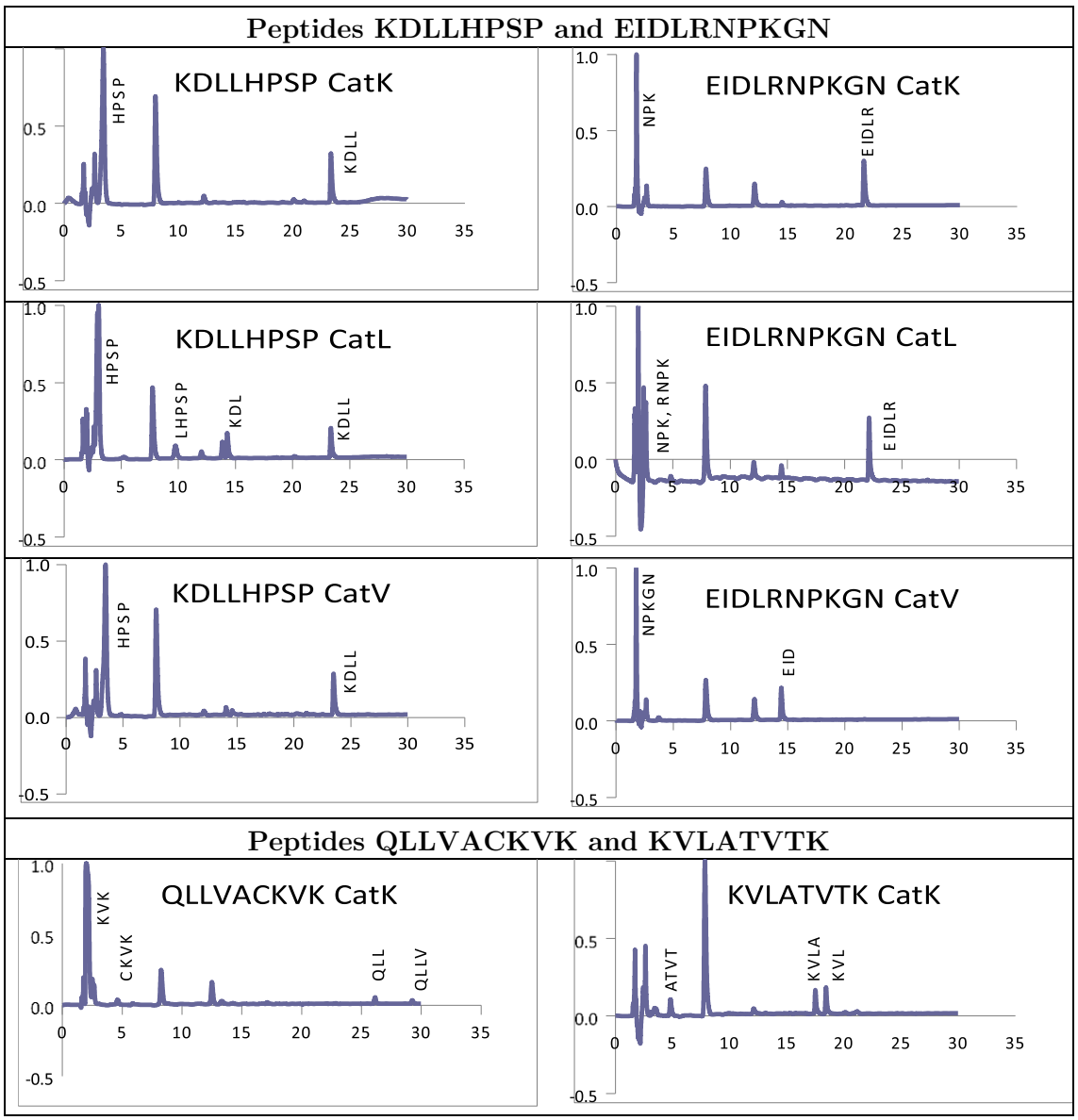
Area	Protein	R L S↓A K P	R L S↓A K P	R L S↓A↓K P
	Prediction	R↓L S↓A K P	R↓L S↓A↓K P	R↓L↓S↓A K P
A6NHL2 Area	Peptide	L L↓S↓G K E	L L↓S↓G K E	L L↓S↓G K E
	Protein	L L S↓G↓K↓E	L L S↓G↓K E	L L S↓G↓K E
	Prediction	L L↓S↓G↓K↓E	L L↓S↓G↓K↓E	L L↓S↓G↓K E
Q6S8J3, P60709, P63261, A5A3E0, P0CG38, P0CG39 Area	Peptide	G M C↓K A G	G M C↓K↓A G*	G M C↓K A G
	Protein	G M C↓K↓A G	G M C K↓A G	G M C↓K↓A G
	Prediction	G M C↓K↓A G	G M C K↓A G	G M C↓K↓A G
P14625 Area	Peptide	K↓R↓F Q↓N↓V A↓K*	K R↓F↓Q↓N V A↓K	K↓R↓F↓Q↓N V A↓K
	Protein	K R F Q↓N V A↓K	K R F Q N V A↓K	K R F Q↓N V A↓K
	Prediction	K R F Q↓N↓V A↓K	K R F Q↓N V A↓K	K R F↓Q↓N↓V A↓K
P63104 Area	Peptide	A C M↓K S V T↓E	A C M↓K S V T↓E	A C M↓K S V T↓E
	Protein	A C M K↓S V↓T↓E	A C M↓K S V T↓E	A C M↓K↓S V T↓E
	Prediction	A C M↓K↓S V T↓E	A C M↓K↓S V T↓E	A C M↓K↓S V T↓E
P25205 Area	Peptide	A Y F K↓K V L	A Y F↓K↓K V L	A Y F↓K K V L
	Protein	A↓Y F K↓K V L	A Y F K↓K V L	A Y F↓K↓K V L
	Prediction	A↓Y F K↓K V L	A↓Y F↓K↓K V L	A Y F↓K↓K V L
Q02878 Area	Peptide	K V L↓A↓T V T↓K	K V L↓A T V T↓K	K V L↓A T V T↓K
	Protein	K V L A↓T↓V T↓K	K V L A T V T K§	K V L A↓T V T↓K
	Prediction	K V L↓A↓T V T↓K	K V L↓A↓T↓V T↓K	K V L↓A↓T↓V T↓K
Q92772 Positional[3]	Peptide	E V C↓K↓K K K	E V C↓K↓K K K	E V C↓K K K K
	Protein	E V C↓K K K K	E V C↓K K K K	E V C↓K K K K
	Prediction	E V C↓K K K K	E V C↓K↓K K K	E V C↓K↓K K K
P07199 Positional[3]	Peptide	I I L↓K↓E K	I I L↓K↓E K	I I L↓K↓E K
	Protein	I I L↓K E K	I I L↓K E K	I I L↓K E K
	Prediction	I I L↓K↓E K	I I L↓K↓E K	I I L↓K↓E K
Q8N5V2 Positional[3]	Peptide	T R↓E S E D L E	T R↓E↓S↓E D L E	T R↓E↓S↓E D L E
	Protein	T R E S↓E D L E	T R E S↓E D L E	T R E S↓E D L E
	Prediction	T R↓E S E D L E	T R E S E D L E	T R E S E D L E
Q53FA7 Positional[3]	Peptide	V L ↓L K↓V A A S	V L↓L K V A A S	V L↓L K V A A S
	Protein	V L L K V A A S§	V L L K↓V A A S	V L L K↓V A A S
	Prediction	V L↓L K↓V A↓A S	V L↓L K↓V A↓A S	V L L K↓V A↓A S
Q9C0B0 Positional[4]	Peptide	A W↓K↓K↓E A*	A W K↓K E A	A W K↓K E A
	Protein	A W K K E A§	A W K↓K E A	A W K↓K E A
	Prediction	A W K↓K↓E A	A W K↓K E A	A W↓K↓K E A
O43818 Positional[4]	Peptide	Q L↓R↓Q Q E	Q L↓R↓Q Q E	Q L↓R↓Q Q E
	Protein	Q L R↓Q Q E	Q L R↓Q Q E	Q L R↓Q Q E
	Prediction	Q L R↓Q Q E	Q L R↓Q Q E	Q L R↓Q Q E

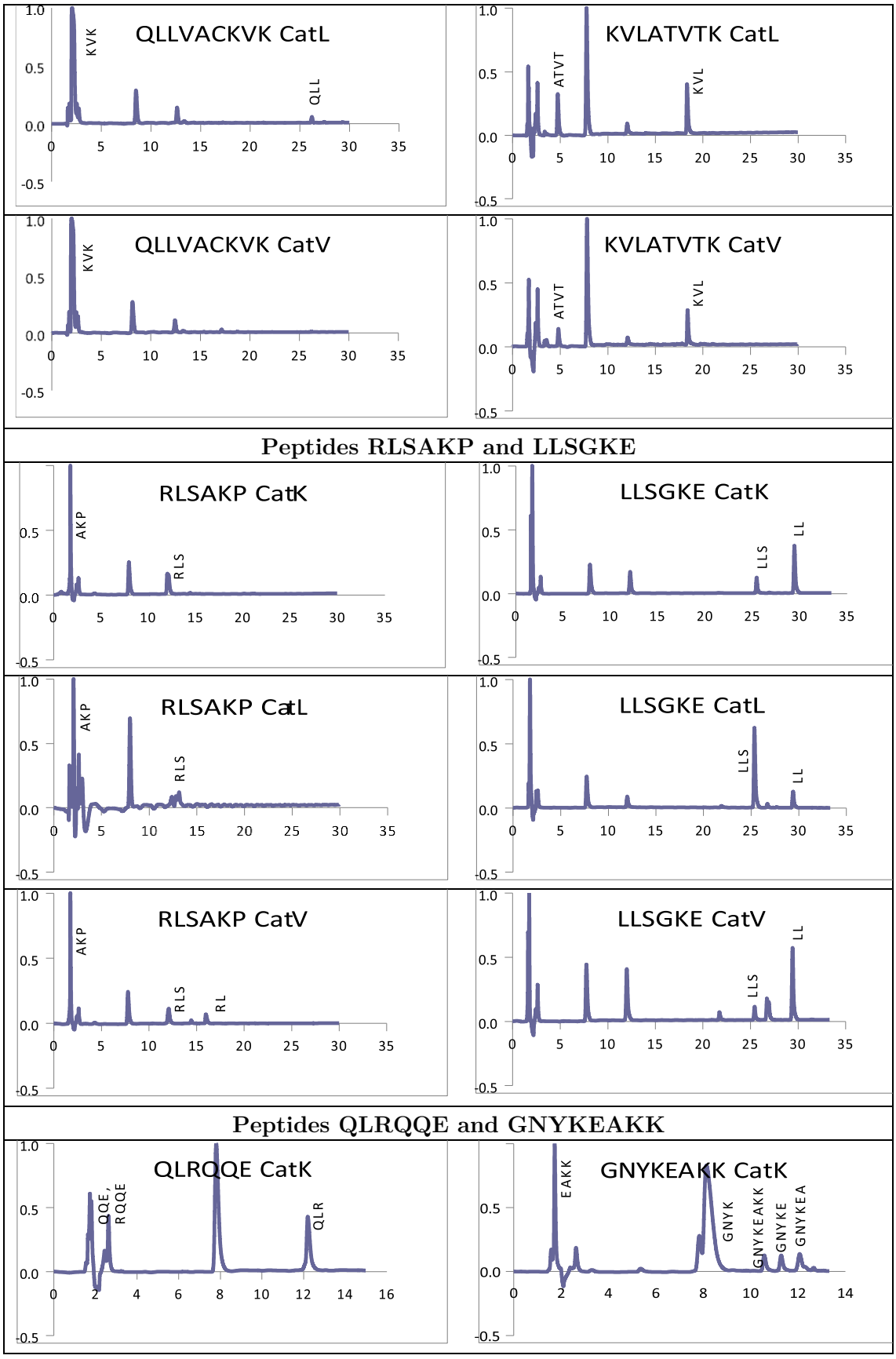
O43823 positional	Peptide	V P C↓G T A H E*	V P C G↓T A H E	V P C G↓T A H E
	Protein	V P C G T A H E§	V P C G T A H E§	V P C G↓T A H E
	Prediction	V P C G↓T A H E	V P C G T A H E	V P C G T A H E
O75533 Positional	Peptide	K I A↓K T H	K I A↓K T H	K I A↓K T H
	Protein	K I A K T H§	K I A K T H§	K I A↓K T H
	Prediction	K I A↓K T H	K I A↓K T H	K I A↓K T H
P46013 Positional	Peptide	V A↓C K S S Q P	V A↓C K S S Q P	V A↓C K S S Q P
	Protein	V A C K↓S S Q P	V A C K↓S S Q P	V A C K↓S S Q P
	Prediction	V A↓C K↓S S Q P	V A↓C K↓S S Q P	V A↓C K↓S S Q P
P42704 Positional	Peptide	G N Y K↓E↓A↓K K*	G N Y K↓E A K K	G N Y K↓E A K K
	Protein	G N Y K E A K K§	G N Y K E A K K§	G N Y K↓E A K K
	Prediction	G N Y K↓E A K↓K	G N Y K↓E A K K	G N Y K↓E↓A K↓K
Q9Y490 Positional	Peptide	Q L L↓V↓A↓C↓K V K*	Q L L↓V A C↓K V K	Q L L V A C↓K V K
	Protein	Q L L V A C K V K§	Q L L V↓A C K V K	Q L L V↓A C K V K
	Prediction	Q L L↓V A C↓K↓V K	Q L L↓V↓A↓C↓K↓V K	Q L L↓V A↓C↓K↓V K
P25398 Positional	Peptide	R G↓I R↓E A A K	R G↓I R↓E A A K	R G↓I R↓E A A K
	Protein	R G I R↓E A A K	R G I R↓E A A K	R G I R↓E A A K
	Prediction	R G I R↓E A A K	R G I R↓E↓A A K	R G I R↓E↓A A K
Q92747 Positional	Peptide	S I Y↓E V D↓K Q*	S I Y↓E↓V D K Q	S I Y↓E↓V D K Q
	Protein	S I Y E V D K Q§	S I Y E V D K Q§	S I Y E↓V D K Q
	Prediction	S I Y↓E V D K Q	S I Y↓E↓V D K Q	S I Y↓E↓V D↓K Q
P62906 Positional	Peptide	K K Y D↓A F↓L↓A*	K K Y D↓A F L↓A	K K Y↓D↓A F L↓A*
	Protein	K K Y D A F L A§	K K Y D↓A F L A	K K Y D↓A F L A
	Prediction	K K Y D A F↓L↓A	K K Y D↓A F L↓A	K K Y D↓A F L↓A
	Peptide	L L↓K A V A E K Q	L L↓K↓A V A E K Q	L L↓K A V A E K Q
	Protein[&]	No data	No data	No data
	Prediction	L L↓K↓A V A E K Q§	L L↓K↓A V A E K Q§	L L↓K↓A V A E K Q§

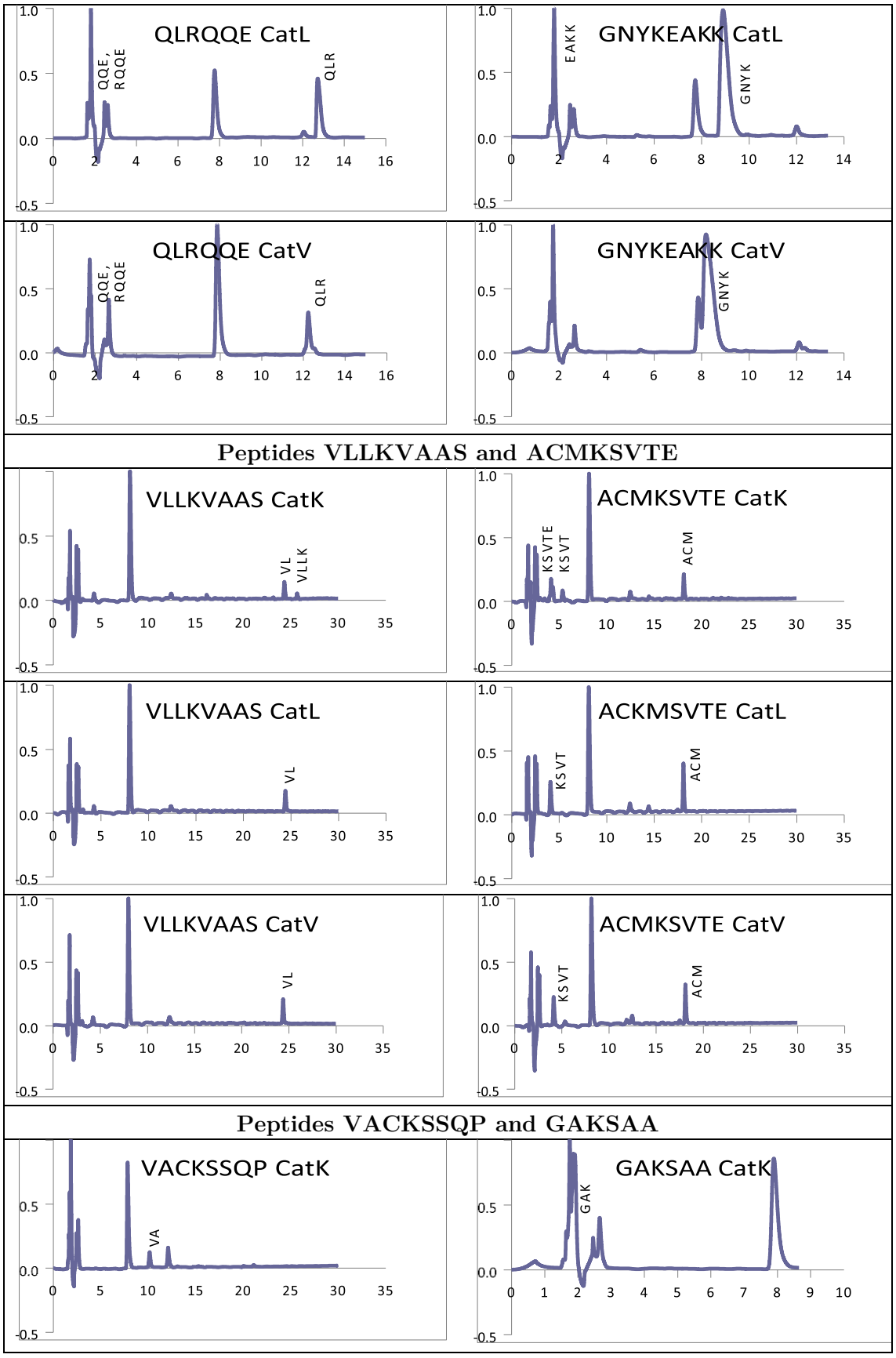
C.2 Peptide Fragment Separation and Identification with RP-HPLC – MALDI-TOF

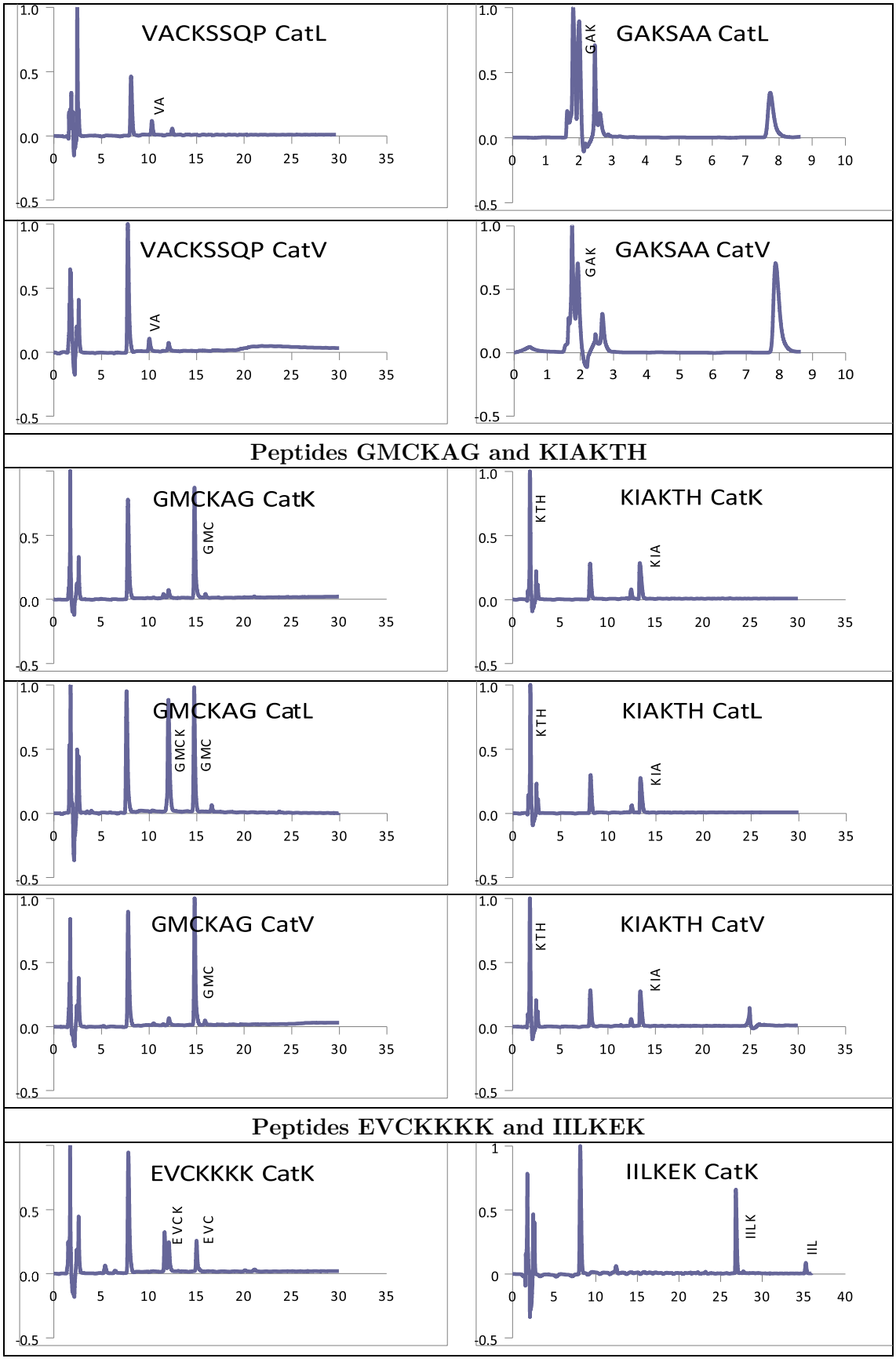
C.2.1 HPLC spectra with peak identification

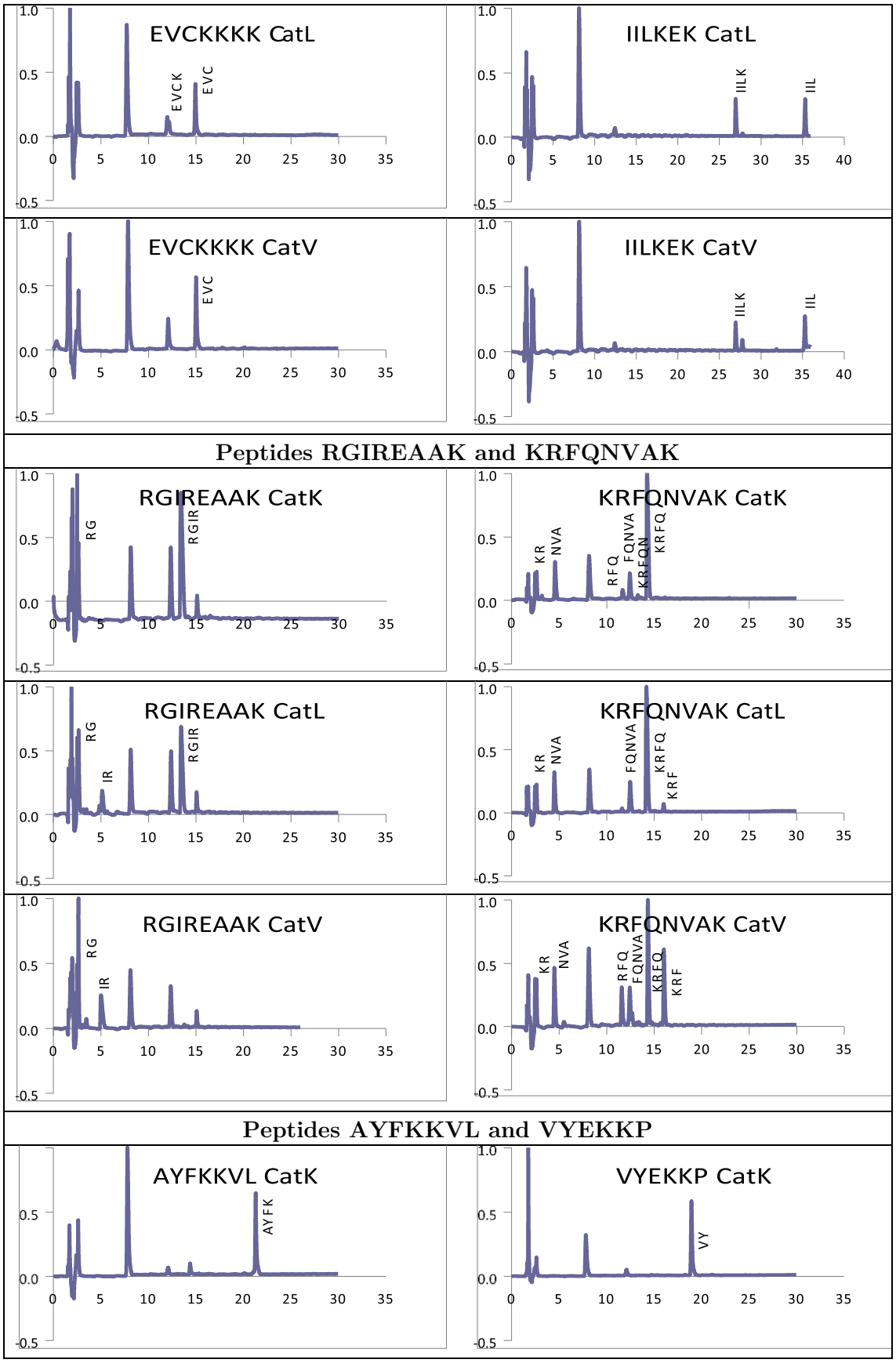
Peptides treated with cathepsins were put on RP-HPLC. Peaks representing the peptide fragments were captured. Separation was monitored at 214 nm. The y-axis shows a normalized signal of absorbance and the x-axis shows separation time in minutes. Sequences of captured peptide fragments, as obtained from MALDI-TOF analysis, are written on top or next to their corresponding HPLC signals. Characteristic signals at approximately 8 and 12 min correspond to buffer component DTT and cathepsin, respectively. The last peptide in this table (sequence LLKAVAEKQ) was not selected from protein cleavages.

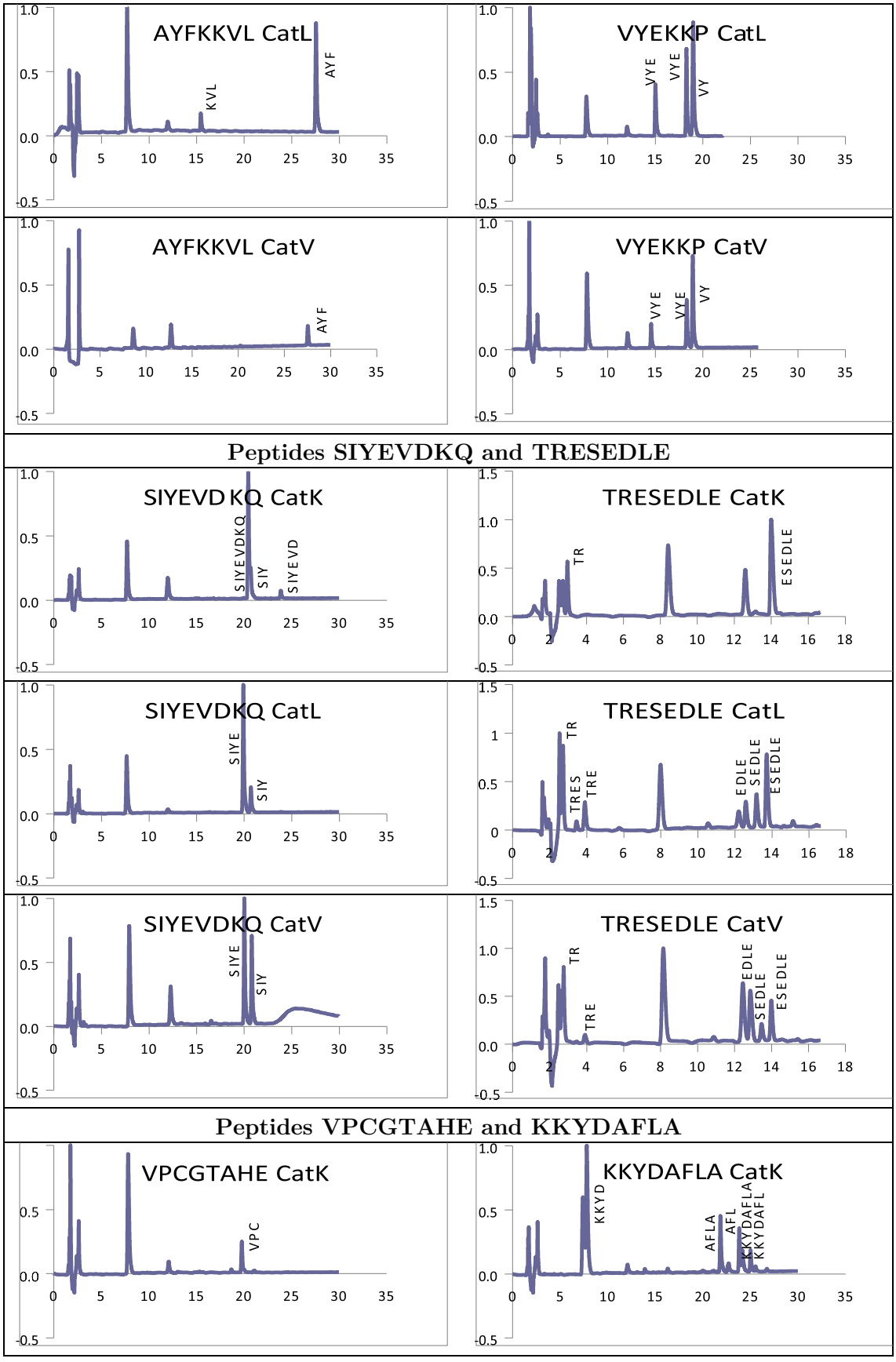


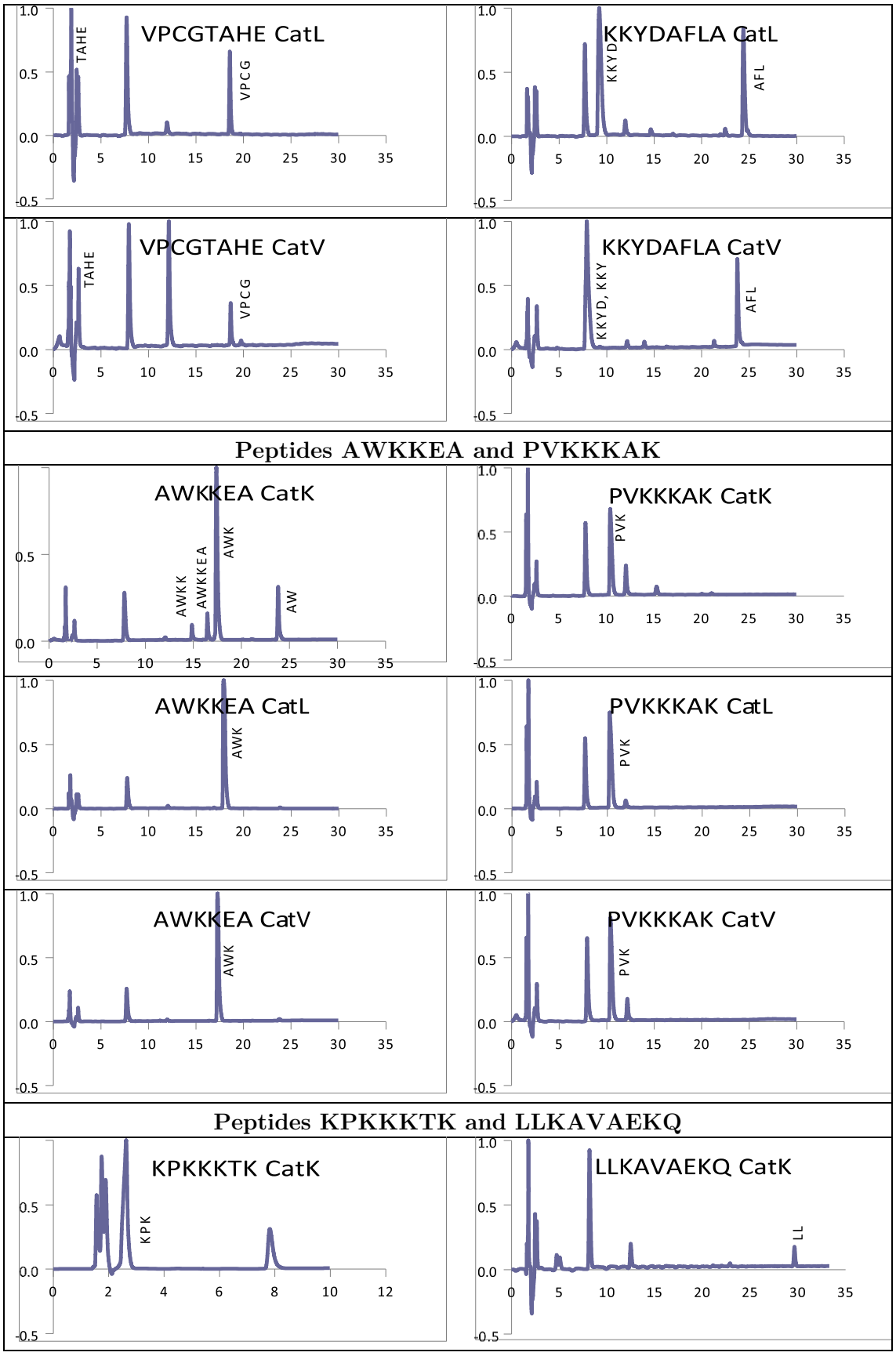

QLLVACKVK CatL
KVK
QLL
KVLATVTK CatL
ATVT
KVL
QLLVACKVK CatV
KVK
KVLATVTK CatV
ATVT
KVL
Peptides RLSAKP and LLSGKE
RLSAKP CatK
AKP
RLS
LLSGKE CatK
LLS
LL
RLSAKP CatL
AKP
RLS
LLSGKE CatL
LLS
LL
RLSAKP CatV
AKP
RLS
RL
LLSGKE CatV
LLS
LL
Peptides QLRQQE and GNYKEAKK
QLRQQE CatK
QQE, RQQE
QLR
GNYKEAKK CatK
EAKK
GNYK
GNYKEAKK
GNYKE
GNYKEA

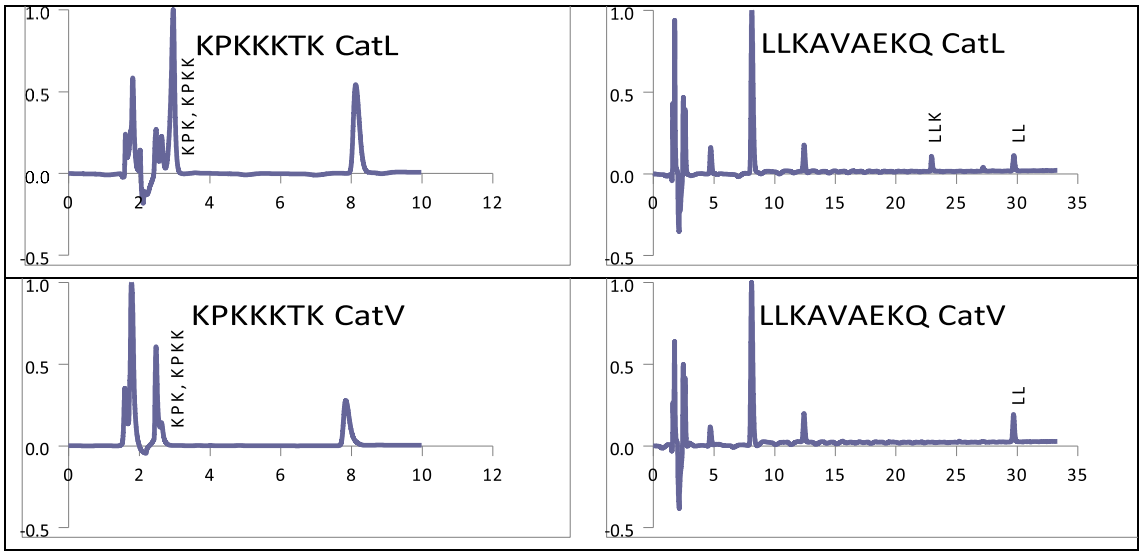

QLRQQE CatL
QQE, RQQE
QLR
GNYKEAKK CatL
EAKK
GNYK
QLRQQE CatV
QQE, RQQE
QLR
GNYKEAKK CatV
GNYK
Peptides VLLKVAAS and ACMKSVTE
VLLKVAAS CatK
VL
VLLK
ACMKSVTE CatK
KSVTE
KSVT
ACM
VLLKVAAS CatL
VL
ACKMSVTE CatL
KSVT
ACM
VLLKVAAS CatV
VL
ACMKSVTE CatV
KSVT
ACM
Peptides VACKSSQP and GAKSAA
VACKSSQP CatK
VA
GAKSAA CatK
GAK


VACKSSQP CatL
VA
GAKSAA CatL
GAK
VACKSSQP CatV
VA
GAKSAA CatV
GAK
Peptides GMCKAG and KIAKTH
GMCKAG CatK
GMC
KIAKTH CatK
KTH
KIA
GMCKAG CatL
GMCK
GMC
KIAKTH CatL
KTH
KIA
GMCKAG CatV
GMC
KIAKTH CatV
KTH
KIA
Peptides EVCKKKK and IILKEK
EVCKKKK CatK
EVCK
EVC
IILKEK CatK
IILK
IIL


EVCKKKK CatL
EVCK
EVC
IILKEK CatL
IILK
IIL
EVCKKKK CatV
EVC
IILKEK CatV
IILK
IIL
Peptides RGIREAAK and KRFQNVAK
RGIREAAK CatK
RG
RGIR
KRFQNVAK CatK
KR
NVA
RFQ
FQNVA
KRFQN
KRFQ
RGIREAAK CatL
RG
IR
RGIR
KRFQNVAK CatL
KR
NVA
FQNVA
KRFQ
KRF
RGIREAAK CatV
RG
IR
KRFQNVAK CatV
KR
NVA
RFQ
FQNVA
KRFQ
KRF
Peptides AYFKKVL and VYEKKP
AYFKKVL CatK
AYFK
VYEKKP CatK
VY

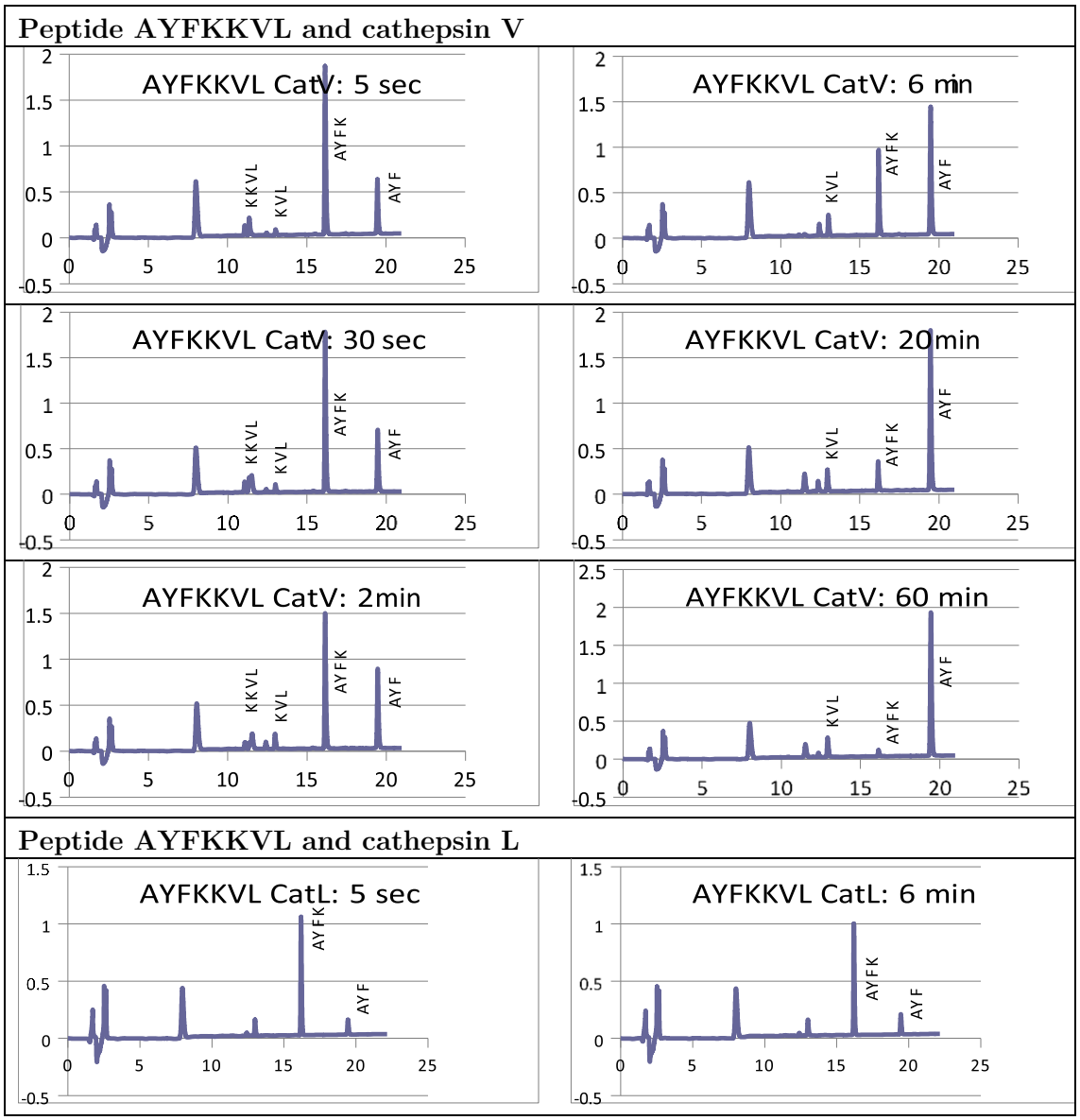

AYFKKVL CatL
KVL
AYF
VYEKKP CatL
VYE
VYE
VY
AYFKKVL CatV
AYF
VYEKKP CatV
VYE
VYE
VY
Peptides SIYEVDKQ and TRESEDLE
SIYEVDKQ CatK
SIYEVDKQ
SIY
SIYEVD
TRESEDLE CatK
TR
ESEDLE
SIYEVDKQ CatL
SIYE
SIY
TRESEDLE CatL
TR
TRES
TRE
EDLE
SEDLE
ESEDLE
SIYEVDKQ CatV
SIYE
SIY
TRESEDLE CatV
TR
TRE
EDLE
SEDLE
ESEDLE
Peptides VPCGTAHE and KKYDAFLA
VPCGTAHE CatK
VPC
KKYDAFLA CatK
KKYD
AFLA
AFL
KKYDAFLA
KKYDAFL

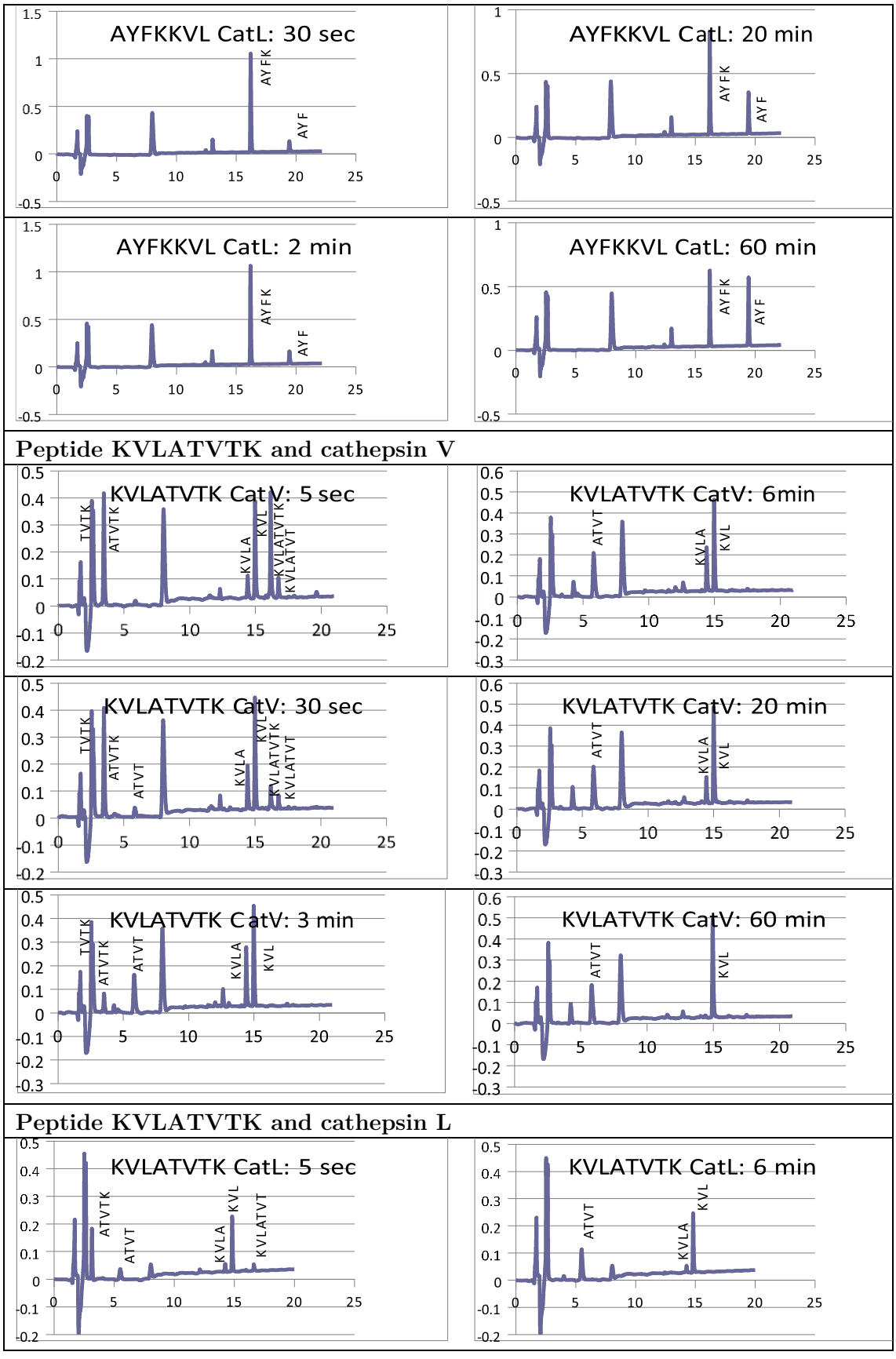

VPCGTAHE CatL
TAHE
VPCG
KKYDAFLA CatL
KKYD
AFL
VPCGTAHE CatV
TAHE
VPCG
KKYDAFLA CatV
KKYD, KKY
AFL
Peptides AWKKEA and PVKKKAK
AWKKEA CatK
AWKK
AWKKEA
AWK
AW
PVKKKAK CatK
PVK
AWKKEA CatL
AWK
PVKKKAK CatL
PVK
AWKKEA CatV
AWK
PVKKKAK CatV
PVK
Peptides KPKKKTK and LLKAVAEKQ
KPKKKTK CatK
KPK
LLKAVAEKQ CatK
LL

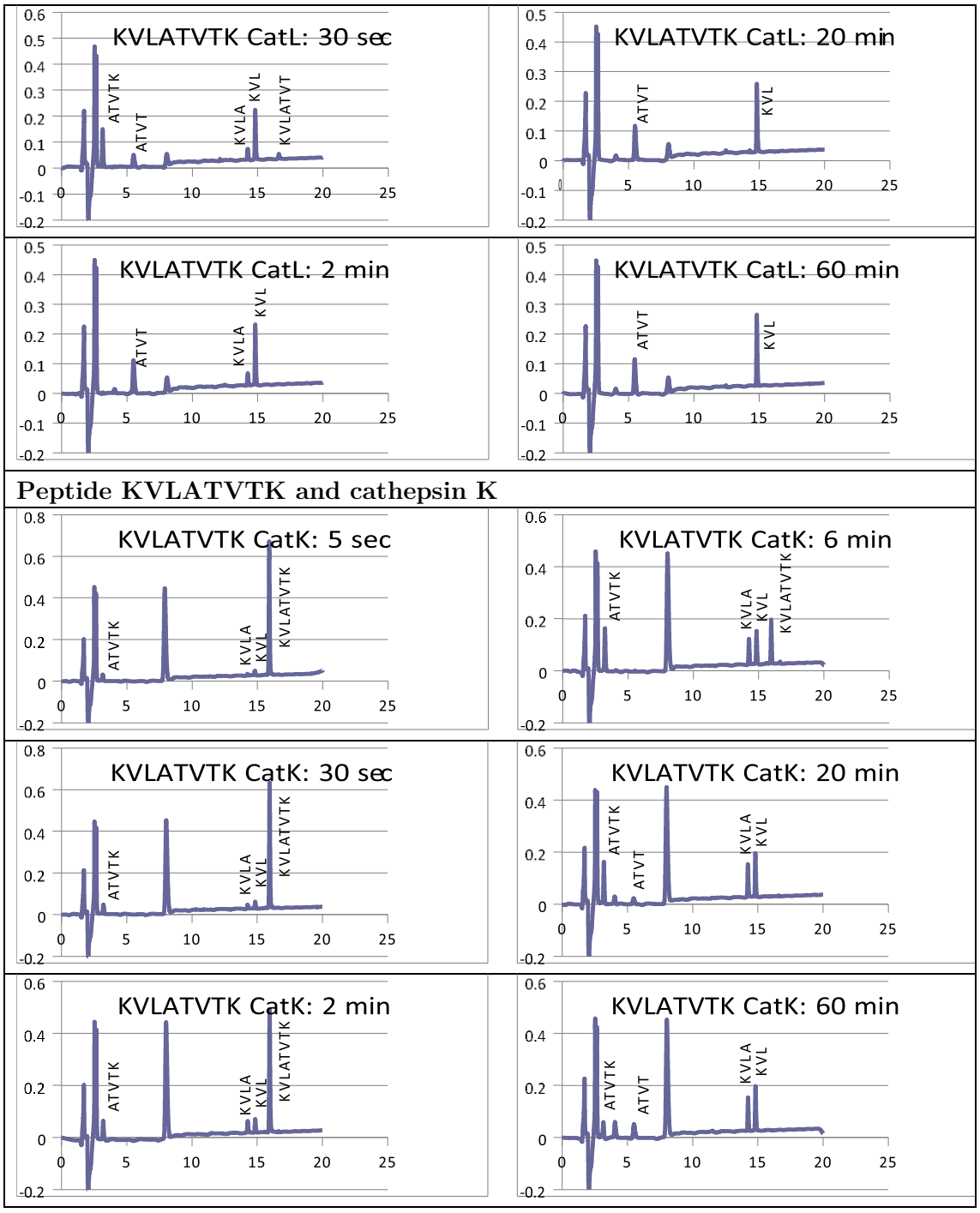

KPKKKTK CatL
KPK, KPKK
LLKAVAEKQ CatL
LLK
LL
KPKKKTK CatV
KPK, KPKK
LLKAVAEKQ CatV
LL

C.2.2 Processing of peptides AYFKKVL and KVLATVTK from 5 sec-60 min.

Aliquots were taken after 5 sec, 30 sec, 2 min, 6 min, 20 min, and 60 min of incubation time with cathepsins V and L (both peptides) and K (peptide KVLATVTK). Separation was monitored at 214 nm. The y-axis shows the signal of absorbance and the x-axis shows the separation time in minutes (signal at y-axis is not normalized in order to quantitatively compare signals at different time points). Sequences of captured peptide fragments, as obtained from MALDI-TOF analysis, are written on top or next to their corresponding HPLC signals. Characteristic signals at approximately 8 and 12 min correspond to buffer component DTT and cathepsin, respectively.




AYFKKVL CatL: 30 sec
AYFK
AYF
AYFKKVL CatL: 20 min
AYFK
AYF
AYFKKVL CatL: 2 min
AYFK
AYF
AYFKKVL CatL: 60 min
AYFK
AYF
Peptide KVLATVTK and cathepsin V
KVLATVTK CatV: 5 sec
TVTK
ATVTK
KVLA
KVL
KVLATVTK
KVLATVT
KVLATVTK CatV: 6min
ATVT
KVLA
KVL
KVLATVTK CatV: 30 sec
TVTK
ATVTK
ATVT
KVLA
KVL
KVLATVTK
KVLATVT
KVLATVTK CatV: 20 min
ATVT
KVLA
KVL
KVLATVTK CatV: 3 min
TVTK
ATVTK
ATVT
KVLA
KVL
KVLATVTK CatV: 60 min
ATVT
KVL
Peptide KVLATVTK and cathepsin L
KVLATVTK CatL: 5 sec
ATVTK
ATVT
KVLA
KVL
KVLATVT
KVLATVTK CatL: 6 min
ATVT
KVLA
KVL


KVLATVTK CatL: 30 sec
ATVTK
ATVT
KVLA
KVL
KVLATVT
KVLATVTK CatL: 20 min
ATVT
KVL
KVLATVTK CatL: 2 min
ATVT
KVLA
KVL
KVLATVTK CatL: 60 min
ATVT
KVL
Peptide KVLATVTK and cathepsin K
KVLATVTK CatK: 5 sec
ATVTK
KVLA
KVL
KVLATVTK
KVLATVTK CatK: 6 min
ATVTK
KVLA
KVL
KVLATVTK
KVLATVTK CatK: 30 sec
ATVTK
KVLA
KVL
KVLATVTK
KVLATVTK CatK: 20 min
ATVTK
ATVT
KVLA
KVL
KVLATVTK CatK: 2 min
ATVTK
KVLA
KVL
KVLATVTK
KVLATVTK CatK: 60 min
ATVTK
ATVT
KVLA
KVL

■

Appendix D

Author contributions

The paper "**Proteomic data and structure analysis combined reveal interplay of structural rigidity and flexibility on selectivity of cysteine cathepsins**" (Tušar & Loboda & Impens *et al.*, 2023, Communications Biology) presents the major work of this PhD thesis. The results are presented in section 5.1 and its discussion in sections 6.1 and 6.2. The author contributed to this paper by preparing and crystallizing cathepsin V-peptide complexes, and by solving, refining and analyzing 3-D structures of the following complexes: 7Q9H, 7Q8H, 7QFH, 7Q9C, 7Q8D, 7Q8F, 7Q8M, 7QHJ, 7Q8N and 7Q8L, and by refining and analyzing all cathepsin V-peptide complexes; by purifying all of the protected peptides from p1 – p30, as well as the peptide p35, on semi-preparative RP-HPLC; by expressing and purifying human cathepsins K, V, and L, and mutant cathepsin V C25A; by performing peptide cleavage analysis with wild type cathepsins, by isolating peptide fragments by RP-HPLC and by determination of their primary sequences on MALDI-TOF spectrometer; by preparing the following figures: Figure 3, Figure 4, Figures 12-15 and all Figures in Appendixes B and C.

The paper "**The Alkyne Moiety as a Latent Electrophile in Irreversible Covalent Small Molecule Inhibitors of Cathepsin K**" (Mons *et al.*, 2019, Journal of the American Chemical Society) is presented in section 5.2. The author contributed to this paper by preparing, crystallizing and solving the 3-D structure of the crystal complex between cathepsin K and inhibitor 7 (PDB entry 6QBS); by preparing Figures 6 and 7.

The paper "**X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease**" (Günther *et al.*, 2021, Science) was part of the project for discovering inhibitors of SARS-CoV-2 M^{pro} protease. The author contributed to this paper by performing and analyzing the inhibitory assays between inhibitors HEAT, tolperisone, pelitinib and calpeptin on SARS-CoV-2 M^{pro} protease and on human cathepsins K, V, and L. The results of the most promising compound from this screen, the calpeptin, are presented in section 5.3.

The paper "**The Sulfonated Calpeptin is a promising drug candidate against SARS-CoV-2 infections**" (Reinke *et al.*, 2023, submitted) is under revision at the time of writing this PhD thesis. Its results are presented in section 5.3 and its discussion in section 6.3. The author contributed to this paper by expressing and purifying cathepsins K, V, and L; by performing and analyzing the inhibitory assays between cathepsins and M^{pro} and inhibitors calpeptin, S-calpeptin, N-calpeptin and GC-376 and by determining the IC_{50} and K_i values in GraphPad Prism software; by preparing, crystallizing and solving the 3-D structure of the crystal complex between cathepsin V and calpeptin (PDB entry 7QGW); by preparing Figures 9-11.

References

Adams-Cioaba, M. A., Krupa, J. C., Xu, C., Mort, J. S., & Min, J. (2011). Structural basis for the recognition and cleavage of histone H3 by cathepsin L. *Nature Communications*, *2*(1), 197. https://doi.org/10.1038/ncomms1204

Afonine, P. V., Grosse-Kunstleve, R. W., Echols, N., Headd, J. J., Moriarty, N. W., Mustyakimov, M., Terwilliger, T. C., Urzhumtsev, A., Zwart, P. H., & Adams, P. D. (2012). Towards automated crystallographic structure refinement with phenix.refine. *Acta Crystallographica. Section D, Biological Crystallography*, *68*(Pt 4), 352–367. https://doi.org/10.1107/S0907444912001308

Alper, P. B., Liu, H., Chatterjee, A. K., Nguyen, K. T., Tully, D. C., Tumanut, C., Li, J., Harris, J. L., Tuntland, T., Chang, J., Gordon, P., Hollenbeck, T., & Karanewsky, D. S. (2006). Arylaminoethyl amides as noncovalent inhibitors of cathepsin S. Part 2: Optimization of P1 and N-aryl. *Bioorganic & Medicinal Chemistry Letters*, *16*(6), 1486–1490. https://doi.org/10.1016/j.bmcl.2005.12.056

Andrejašič, M., Pražnikar, J., & Turk, D. (2008). PURY: A database of geometric restraints of hetero compounds for refinement in complexes with macromolecular structures. *Acta Crystallographica Section D: Biological Crystallography*, *64*(11), 1093–1109. https://doi.org/10.1107/S0907444908027388

Biniossek, M. L., Nägler, D. K., Becker-Pauly, C., & Schilling, O. (2011). Proteomic identification of protease cleavage sites characterizes prime and non-prime specificity of cysteine cathepsins B, L, and S. *Journal of Proteome Research*, *10*(12), 5363–5373. https://doi.org/10.1021/pr200621z

Bond, J. S. (2019). Proteases: History, discovery, and roles in health and disease. *Journal of Biological Chemistry*, *294*(5), 1643–1651. https://doi.org/10.1074/jbc.TM118.004156

Borišek, J., Vizovišek, M., Sosnowski, P., Turk, B., Turk, D., Mohar, B., & Novič, M. (2015). Development of N-(Functionalized benzoyl)-homocycloleucyl-glycinonitriles as Potent Cathepsin K Inhibitors. *Journal of Medicinal Chemistry*, *58*(17), 6928–6937. https://doi.org/10.1021/acs.jmedchem.5b00746

Bosch, B. J., Bartelink, W., & Rottier, P. J. M. (2008). Cathepsin L Functionally Cleaves the Severe Acute Respiratory Syndrome Coronavirus Class I Fusion Protein Upstream of Rather than Adjacent to the Fusion Peptide. *Journal of Virology*, *82*(17), 8887–8890. https://doi.org/10.1128/JVI.00415-08

Bragg, W. H., & Bragg, W. L. (1915). *X rays and crystal structure*. London : G. Bell. http://archive.org/details/xrayscrystalstru00braguoft

Brünger, A. T. (1997). [19] Free R value: Cross-validation in crystallography. In *Methods in Enzymology* (Vol. 277, pp. 366–396). Academic Press. https://doi.org/10.1016/S0076-6879(97)77021-6

Catherman, A. D., Skinner, O. S., & Kelleher, N. L. (2014). Top Down Proteomics: Facts and Perspectives. *Biochemical and Biophysical Research Communications*, *445*(4), 683–693. https://doi.org/10.1016/j.bbrc.2014.02.041

Cheraghian Radi, H., Hajipour-Verdom, B., & Molaabasi, F. (2021). Macromolecular crystallization: Basics and advanced methodologies. *Journal of the Iranian Chemical Society*, *18*(3), 543–565. https://doi.org/10.1007/s13738-020-02058-y

Choe, Y., Leonetti, F., Greenbaum, D. C., Lecaille, F., Bogyo, M., Brömme, D., Ellman, J. A., & Craik, C. S. (2006). Substrate profiling of cysteine proteases using a combinatorial peptide library identifies functionally unique specificities. *The Journal of Biological Chemistry*, *281*(18), 12824–12832. https://doi.org/10.1074/jbc.M513331200

Chowdhury, S. F., Joseph, L., Kumar, S., Shenoy, R. T., Bhat, S., Ziomek, E., Ménard, R., Sivaraman, J., & Purisima, E. O. (2008). Exploring Inhibitor Binding at the S′ Subsites of Cathepsin L. *Journal of Medicinal Chemistry*, *51*(5), 1361–1368. https://doi.org/10.1021/jm701190v

Cianni, L., Feldmann, C. W., Gilberg, E., Gütschow, M., Juliano, L., Leitão, A., Bajorath, J., & Montanari, C. A. (2019). Can Cysteine Protease Cross-Class Inhibitors Achieve Selectivity? *Journal of Medicinal Chemistry*, *62*(23), 10497–10525. https://doi.org/10.1021/acs.jmedchem.9b00683

Coussens, L. M., Fingleton, B., & Matrisian, L. M. (2002). Matrix Metalloproteinase Inhibitors and Cancer—Trials and Tribulations. *Science*, *295*(5564), 2387–2392. https://doi.org/10.1126/science.1067100

Craik, C. S., Page, M. J., & Madison, E. L. (2011). Proteases as therapeutics. *The Biochemical Journal*, *435*(1), 1–16. https://doi.org/10.1042/BJ20100965

Dana, D., De, S., Rathod, P., Davalos, A. R., Novoa, D. A., Paroly, S., Torres, V. M., Afzal, N., Lankalapalli, R. S., Rotenberg, S. A., Chang, E. J., Subramaniam, G., & Kumar, S. (2014). Development of a highly potent, selective, and cell-active Inhibitor of cysteine cathepsin L–A hybrid design approach. *Chemical Communications*, *50*(74), 10875–10878. https://doi.org/10.1039/C4CC04037F

Dana, D., & Pathak, S. K. (2020). A Review of Small Molecule Inhibitors and Functional Probes of Human Cathepsin L. *Molecules*, *25*(3). https://doi.org/10.3390/molecules25030698

DesJarlais, R. L., Yamashita, D. S., Oh, H.-J., Uzinskas, I. N., Erhard, K. F., Allen, A. C., Haltiwanger, R. C., Zhao, B., Smith, W. W., Abdel-Meguid, S. S., D'Alessio, K., Janson, C. A., McQueney, M. S., Tomaszek, T. A., Levy, M. A., & Veber, D. F. (1998). Use of X-ray Co-crystal Structures and Molecular Modeling To Design Potent and Selective Non-peptide Inhibitors of Cathepsin K. *Journal of the American Chemical Society*, *120*(35), 9114–9115. https://doi.org/10.1021/ja981171v

Di Jeso, B., & Arvan, P. (2016). Thyroglobulin From Molecular and Cellular Biology to Clinical Endocrinology. *Endocrine Reviews*, *37*(1), 2–36. https://doi.org/10.1210/er.2015-1090

Drag, M., & Salvesen, G. S. (2010). Emerging principles in protease-based drug discovery. *Nature Reviews. Drug Discovery*, *9*(9), 690–701. https://doi.org/10.1038/nrd3053

Duncan, E. M., Muratore-Schroeder, T. L., Cook, R. G., Garcia, B. A., Shabanowitz, J., Hunt, D. F., & Allis, C. D. (2008). Cathepsin L proteolytically processes histone H3 during mouse embryonic stem cell differentiation. *Cell*, *135*(2), 284–294. https://doi.org/10.1016/j.cell.2008.09.055

Eisenthal, R., Danson, M. J., & Hough, D. W. (2007). Catalytic efficiency and kcat/KM: A useful comparator? *Trends in Biotechnology*, *25*(6), 247–249. https://doi.org/10.1016/j.tibtech.2007.03.010

Ekkebus, R., van Kasteren, S. I., Kulathu, Y., Scholten, A., Berlin, I., Geurink, P. P., de Jong, A., Goerdayal, S., Neefjes, J., Heck, A. J. R., Komander, D., & Ovaa, H. (2013). On Terminal Alkynes That Can React with Active-Site Cysteine Nucleophiles in

Proteases. *Journal of the American Chemical Society*, *135*(8), 2867–2870. https://doi.org/10.1021/ja309802n

Fischer, E. (1894). Einfluss der Configuration auf die Wirkung der Enzyme. *Berichte Der Deutschen Chemischen Gesellschaft*, *27*(3), 2985–2993. https://doi.org/10.1002/cber.18940270364

Forsythe, E. L., Maxwell, D. L., Pusey, M., & IUCr. (2002, September 26). *Vapor diffusion, nucleation rates and the reservoir to crystallization volume ratio* (Text No. 10). Acta Crystallographica Section D: Biological Crystallography; International Union of Crystallography. https://doi.org/10.1107/S0907444902014208

Fuchs, N., Meta, M., Schuppan, D., Nuhn, L., & Schirmeister, T. (2020). Novel Opportunities for Cathepsin S Inhibitors in Cancer Immunotherapy by Nanocarrier-Mediated Delivery. *Cells*, *9*(9), 2021. https://doi.org/10.3390/cells9092021

Funkelstein, L., Lu, W. D., Koch, B., Mosier, C., Toneff, T., Taupenot, L., O'Connor, D. T., Reinheckel, T., Peters, C., & Hook, V. (2012). Human cathepsin V protease participates in production of enkephalin and NPY neuropeptide neurotransmitters. *The Journal of Biological Chemistry*, *287*(19), 15232–15241. https://doi.org/10.1074/jbc.M111.310607

Funkelstein, L., Toneff, T., Mosier, C., Hwang, S.-R., Beuschlein, F., Lichtenauer, U. D., Reinheckel, T., Peters, C., & Hook, V. (2008). Major Role of Cathepsin L for Producing the Peptide Hormones ACTH, β-Endorphin, and α-MSH, Illustrated by Protease Gene Knockout and Expression. *The Journal of Biological Chemistry*, *283*(51), 35652–35659. https://doi.org/10.1074/jbc.M709010200

Garnero, P., Borel, O., Byrjalsen, I., Ferreras, M., Drake, F. H., McQueney, M. S., Foged, N. T., Delmas, P. D., & Delaissé, J.-M. (1998). The Collagenolytic Activity of Cathepsin K Is Unique among Mammalian Proteinases. *Journal of Biological Chemistry*, *273*(48), 32347–32352. https://doi.org/10.1074/jbc.273.48.32347

Gáspár, M. E., & Csermely, P. (2012). Rigidity and flexibility of biological networks. *Briefings in Functional Genomics*, *11*(6), 443–456. https://doi.org/10.1093/bfgp/els023

Gauthier, J. Y., Black, W. C., Courchesne, I., Cromlish, W., Desmarais, S., Houle, R., Lamontagne, S., Li, C. S., Massé, F., McKay, D. J., Ouellet, M., Robichaud, J., Truchon, J.-F., Truong, V.-L., Wang, Q., & Percival, M. D. (2007). The identification of potent, selective, and bioavailable cathepsin S inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *17*(17), 4929–4933. https://doi.org/10.1016/j.bmcl.2007.06.023

Gauthier, J. Y., Chauret, N., Cromlish, W., Desmarais, S., Duong, L. T., Falgueyret, J.-P., Kimmel, D. B., Lamontagne, S., Léger, S., LeRiche, T., Li, C. S., Massé, F., McKay, D. J., Nicoll-Griffith, D. A., Oballa, R. M., Palmer, J. T., Percival, M. D., Riendeau, D., Robichaud, J., … Black, W. C. (2008). The discovery of odanacatib (MK-0822), a selective inhibitor of cathepsin K. *Bioorganic & Medicinal Chemistry Letters*, *18*(3), 923–928. https://doi.org/10.1016/j.bmcl.2007.12.047

Gnirß, K., Kühl, A., Karsten, C., Glowacka, I., Bertram, S., Kaup, F., Hofmann, H., & Pöhlmann, S. (2012). Cathepsins B and L activate Ebola but not Marburg virus glycoproteins for efficient entry into cell lines and macrophages independent of TMPRSS2 expression. *Virology*, *424*(1), 3–10. https://doi.org/10.1016/j.virol.2011.11.031

Günther, S., Reinke, P. Y. A., Fernández-García, Y., Lieske, J., Lane, T. J., Ginn, H. M., Koua, F. H. M., Ehrt, C., Ewert, W., Oberthuer, D., Yefanov, O., Meier, S., Lorenzen, K., Krichel, B., Kopicki, J.-D., Gelisio, L., Brehm, W., Dunkel, I., Seychell, B., … Meents, A. (2021). X-ray screening identifies active site and allosteric inhibitors of

SARS-CoV-2 main protease. *Science (New York, N.Y.)*, *372*(6542), 642–646. https://doi.org/10.1126/science.abf7945

Gupta, N., Hixson, K. K., Culley, D. E., Smith, R. D., & Pevzner, P. A. (2010). Analyzing protease specificity and detecting in vivo proteolytic events using tandem mass spectrometry. *Proteomics*, *10*(15), 2833–2844. https://doi.org/10.1002/pmic.200900821

Hai-Fu, F. (1998). Sayre Equation, Tangent Formula and SAYTAN. In S. Fortier (Ed.), *Direct Methods for Solving Macromolecular Structures* (pp. 79–85). Springer Netherlands. https://doi.org/10.1007/978-94-015-9093-8_8

Han, X., Aslanian, A., & Yates, J. R. (2008). Mass Spectrometry for Proteomics. *Current Opinion in Chemical Biology*, *12*(5), 483–490. https://doi.org/10.1016/j.cbpa.2008.07.024

Jackson, C. B., Farzan, M., Chen, B., & Choe, H. (2022). Mechanisms of SARS-CoV-2 entry into cells. *Nature Reviews Molecular Cell Biology*, *23*(1), Article 1. https://doi.org/10.1038/s41580-021-00418-x

Johnson, K. A., & Goody, R. S. (2011). The Original Michaelis Constant: Translation of the 1913 Michaelis-Menten Paper. *Biochemistry*, *50*(39), 8264–8269. https://doi.org/10.1021/bi201284u

Juibari, A. D., Rezadoost, M. H., & Soleimani, M. (2022). The key role of Calpain in COVID-19 as a therapeutic strategy. *Inflammopharmacology*, *30*(5), 1479–1491. https://doi.org/10.1007/s10787-022-01002-1

Kabsch, W. (2010). XDS. *Acta Crystallographica Section D: Biological Crystallography*, *66*(2), 125–132. https://doi.org/10.1107/S0907444909047337

Kaysser, L. (2019). Built to bind: Biosynthetic strategies for the formation of small-molecule protease inhibitors. *Natural Product Reports*, *36*(12), 1654–1686. https://doi.org/10.1039/C8NP00095F

Kim, H., Hwang, Y. S., Kim, M., & Park, S. B. (2021). Recent advances in the development of covalent inhibitors. *RSC Medicinal Chemistry*, *12*(7), 1037–1045. https://doi.org/10.1039/D1MD00068C

Kingma, P. S., Burden, D. A., & Osheroff, N. (1999). Binding of etoposide to topoisomerase II in the absence of DNA: Decreased affinity as a mechanism of drug resistance. *Biochemistry*, *38*(12), 3457–3461. https://doi.org/10.1021/bi982855i

Kleywegt, G. J. (2000). Validation of protein crystal structures. *Acta Crystallographica. Section D, Biological Crystallography*, *56*(Pt 3), 249–265. https://doi.org/10.1107/s0907444999016364

Klingler, D., & Hardt, M. (2012). Profiling protease activities with dynamic proteomics workflows. *Proteomics*, *12*(4–5), 587–596. https://doi.org/10.1002/pmic.201100399

Koshland, D. E. (1958). Application of a Theory of Enzyme Specificity to Protein Synthesis. *Proceedings of the National Academy of Sciences*, *44*(2), 98–104. https://doi.org/10.1073/pnas.44.2.98

Lapek, J. D., Jiang, Z., Wozniak, J. M., Arutyunova, E., Wang, S. C., Lemieux, M. J., Gonzalez, D. J., & O'Donoghue, A. J. (2019). Quantitative Multiplex Substrate Profiling of Peptidases by Mass Spectrometry. *Molecular & Cellular Proteomics : MCP*, *18*(5), 968–981. https://doi.org/10.1074/mcp.TIR118.001099

Li, C. S., Deschenes, D., Desmarais, S., Falgueyret, J.-P., Gauthier, J. Y., Kimmel, Donald. B., Léger, S., Massé, F., McGrath, M. E., McKay, D. J., Percival, M. D., Riendeau, D., Rodan, S. B., Thérien, M., Truong, V.-L., Wesolowski, G., Zamboni, R., & Black, W. C. (2006). Identification of a potent and selective non-basic cathepsin K inhibitor. *Bioorganic & Medicinal Chemistry Letters*, *16*(7), 1985–1989. https://doi.org/10.1016/j.bmcl.2005.12.071

Li, Y.-Y., Fang, J., & Ao, G.-Z. (2017). Cathepsin B and L inhibitors: A patent review (2010 - present). *Expert Opinion on Therapeutic Patents*, *27*(6), 643–656. https://doi.org/10.1080/13543776.2017.1272572

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., … Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: Recent developments in Phenix. *Acta Crystallographica. Section D, Structural Biology*, *75*(Pt 10), 861–877. https://doi.org/10.1107/S2059798319011471

Liu, H., Tully, D. C., Epple, R., Bursulaya, B., Li, J., Harris, J. L., Williams, J. A., Russo, R., Tumanut, C., Roberts, M. J., Alper, P. B., He, Y., & Karanewsky, D. S. (2005). Design and synthesis of arylaminoethyl amides as noncovalent inhibitors of cathepsin S. Part 1. *Bioorganic & Medicinal Chemistry Letters*, *15*(22), 4979–4984. https://doi.org/10.1016/j.bmcl.2005.08.017

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

Lu, J., Wang, M., Wang, Z., Fu, Z., Lu, A., & Zhang, G. (2018). Advances in the discovery of cathepsin K inhibitors on bone resorption. *Journal of Enzyme Inhibition and Medicinal Chemistry*, *33*(1), 890–904. https://doi.org/10.1080/14756366.2018.1465417

Luo, H., Tie, L., Cao, M., Hunter, A. K., Pabst, T. M., Du, J., Field, R., Li, Y., & Wang, W. K. (2019). Cathepsin L Causes Proteolytic Cleavage of Chinese-Hamster-Ovary Cell Expressed Proteins During Processing and Storage: Identification, Characterization, and Mitigation. *Biotechnology Progress*, *35*(1), e2732. https://doi.org/10.1002/btpr.2732

Luo, S. Y., Araya, L. E., & Julien, O. (2019). Protease Substrate Identification Using N-terminomics. *ACS Chemical Biology*, *14*(11), 2361–2371. https://doi.org/10.1021/acschembio.9b00398

Marquis, R. W., James, I., Zeng, J., Trout, R. E. L., Thompson, S., Rahman, A., Yamashita, D. S., Xie, R., Ru, Y., Gress, C. J., Blake, S., Lark, M. A., Hwang, S.-M., Tomaszek, T., Offen, P., Head, M. S., Cummings, M. D., & Veber, D. F. (2005). Azepanone-Based Inhibitors of Human Cathepsin L. *Journal of Medicinal Chemistry*, *48*(22), 6870–6878. https://doi.org/10.1021/jm0502079

McGrath, M. E., Klaus, J. L., Barnes, M. G., & Brömme, D. (1997). Crystal structure of human cathepsin K complexed with a potent inhibitor. *Nature Structural Biology*, *4*(2), Article 2. https://doi.org/10.1038/nsb0297-105

Mediouni, S., Mou, H., Otsuka, Y., Jablonski, J. A., Adcock, R. S., Batra, L., Chung, D.-H., Rood, C., de Vera, I. M. S., Rahaim Jr., R., Ullah, S., Yu, X., Getmanenko, Y. A., Kennedy, N. M., Wang, C., Nguyen, T.-T., Hull, M., Chen, E., Bannister, T. D., … Spicer, T. P. (2022). Identification of potent small molecule inhibitors of SARS-CoV-2 entry. *SLAS Discovery*, *27*(1), 8–19. https://doi.org/10.1016/j.slasd.2021.10.012

Merritt, E. A., & Bacon, D. J. (1997). Raster3D: Photorealistic molecular graphics. *Methods in Enzymology*, *277*, 505–524. https://doi.org/10.1016/s0076-6879(97)77028-9

Mons, E., Jansen, I. D. C., Loboda, J., van Doodewaerd, B. R., Hermans, J., Verdoes, M., van Boeckel, C. A. A., van Veelen, P. A., Turk, B., Turk, D., & Ovaa, H. (2019). The Alkyne Moiety as a Latent Electrophile in Irreversible Covalent Small Molecule

Inhibitors of Cathepsin K. *Journal of the American Chemical Society*, *141*(8), 3507–3514. https://doi.org/10.1021/jacs.8b11027

Morrison, J. F. (1969). Kinetics of the reversible inhibition of enzyme-catalysed reactions by tight-binding inhibitors. *Biochimica et Biophysica Acta (BBA) - Enzymology*, *185*(2), 269–286. https://doi.org/10.1016/0005-2744(69)90420-3

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F., & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, *67*(4), Article 4. https://doi.org/10.1107/S0907444911001314

Neil W. Ashcroft. (1976). *Solid state physics*. Holt, Rinehart and Winston. http://archive.org/details/solidstatephysic00ashc

O'Donoghue, A. J., Eroy-Reveles, A. A., Knudsen, G. M., Ingram, J., Zhou, M., Statnekov, J. B., Greninger, A. L., Hostetter, D. R., Qu, G., Maltby, D. A., Anderson, M. O., DeRisi, J. L., McKerrow, J. H., Burlingame, A. L., & Craik, C. S. (2012). Global Identification of Peptidase Specificity by Multiplex Substrate Profiling. *Nature Methods*, *9*(11), 1095–1100. https://doi.org/10.1038/nmeth.2182

Pannu, N. S., & Read, R. J. (1996). Improved Structure Refinement Through Maximum Likelihood. *Acta Crystallographica Section A*, *52*(5), 659–668. https://doi.org/10.1107/S0108767396004370

Perkins, D. N., Pappin, D. J., Creasy, D. M., & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18), 3551–3567. https://doi.org/10.1002/(SICI)1522-2683(19991201)20:18<3551::AID-ELPS3551>3.0.CO;2-2

Pražnikar, J., Afonine, P. V., Gunčar, G., Adams, P. D., & Turk, D. (2009). Averaged kick maps: Less noise, more signal…and probably less bias. *Acta Crystallographica Section D Biological Crystallography*, *65*(9), 921–931. https://doi.org/10.1107/S0907444909021933

Pražnikar, J., & Turk, D. (2014). Free kick instead of cross-validation in maximum-likelihood refinement of macromolecular crystal structures. *Acta Crystallographica Section D: Biological Crystallography*, *70*(12), 3124–3134. https://doi.org/10.1107/S1399004714021336

Pruitt, K., Brown, G., Tatusova, T., & Maglott, D. (2012). The Reference Sequence (RefSeq) Database. In *The NCBI Handbook [Internet]*. National Center for Biotechnology Information (US). https://www.ncbi.nlm.nih.gov/books/NBK21091/

Puzer, L., Barros, N. M. T., Paschoalin, T., Hirata, I. Y., Tanaka, A. S., Oliveira, M. C., Brömme, D., & Carmona, A. K. (2008). Cathepsin V, but not cathepsins L, B and K, may release angiostatin-like fragments from plasminogen. *Biological Chemistry*, *389*(2). https://doi.org/10.1515/BC.2008.020

Puzer, L., Cotrin, S. S., Alves, M. F. M., Egborge, T., Araújo, M. S., Juliano, M. A., Juliano, L., Brömme, D., & Carmona, A. K. (2004). Comparative substrate specificity analysis of recombinant human cathepsin V and cathepsin L. *Archives of Biochemistry and Biophysics*, *430*(2), 274–283. https://doi.org/10.1016/j.abb.2004.07.006

Ratnikov, B. I., Cieplak, P., Remacle, A. G., Nguyen, E., & Smith, J. W. (2021). Quantitative profiling of protease specificity. *PLoS Computational Biology*, *17*(2), e1008101. https://doi.org/10.1371/journal.pcbi.1008101

Richardson, J. S., Prisant, M. G., & Richardson, D. C. (2013). Crystallographic Model Validation: From Diagnosis to Healing. *Current Opinion in Structural Biology*, *23*(5), 707–714. https://doi.org/10.1016/j.sbi.2013.06.004

Robichaud, J., Bayly, C., Oballa, R., Prasit, P., Mellon, C., Falgueyret, J.-P., David Percival, M., Wesolowski, G., & Rodan, S. B. (2004). Rational design of potent and selective NH-linked aryl/heteroaryl cathepsin K inhibitors. *Bioorganic & Medicinal Chemistry Letters*, *14*(16), 4291–4295. https://doi.org/10.1016/j.bmcl.2004.05.087

Rozman-Pungerčar, J., Kopitar-Jerala, N., Bogyo, M., Turk, D., Vasiljeva, O., Štefe, I., Vandenabeele, P., Brömme, D., Puizdar, V., Fonović, M., Trstenjak-Prebanda, M., Dolenc, I., Turk, V., & Turk, B. (2003). Inhibition of papain-like cysteine proteases and legumain by caspase-specific inhibitors: When reaction mechanism is more important than specificity. *Cell Death & Differentiation*, *10*(8), Article 8. https://doi.org/10.1038/sj.cdd.4401247

Rupp, B. (2009). *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology* (1st edition). Garland Science.

Rupp, B. (2015). Origin and use of crystallization phase diagrams. *Acta Crystallographica. Section F, Structural Biology Communications*, *71*(Pt 3), 247–260. https://doi.org/10.1107/S2053230X1500374X

Scarcella, M., d'Angelo, D., Ciampa, M., Tafuri, S., Avallone, L., Pavone, L. M., & De Pasquale, V. (2022). The Key Role of Lysosomal Protease Cathepsins in Viral Infections. *International Journal of Molecular Sciences*, *23*(16), 9089. https://doi.org/10.3390/ijms23169089

Schechter, I., & Berger, A. (1967). On the size of the active site in proteases. I. Papain. *Biochemical and Biophysical Research Communications*, *27*(2), 157–162. https://doi.org/10.1016/S0006-291X(67)80055-X

Shenoy, R. T., & Sivaraman, J. (2011). Structural basis for reversible and irreversible inhibition of human cathepsin L by their respective dipeptidyl glyoxal and diazomethylketone inhibitors. *Journal of Structural Biology*, *173*(1), 14–19. https://doi.org/10.1016/j.jsb.2010.09.007

Singh, J., Petter, R. C., Baillie, T. A., & Whitty, A. (2011). The resurgence of covalent drugs. *Nature Reviews Drug Discovery*, *10*(4), Article 4. https://doi.org/10.1038/nrd3410

Sosnowski, P., & Turk, D. (2016). Caught in the act: The crystal structure of cleaved cathepsin L bound to the active site of Cathepsin L. *FEBS Letters*, *590*(8), 1253–1261. https://doi.org/10.1002/1873-3468.12140

Sosnowski, Piotr. (2016). *Structural insight in the substrate specificity of cathepsins.*

Srinivasan, B. (2021). A guide to the Michaelis–Menten equation: Steady state and beyond. *The FEBS Journal*, *n/a*(n/a). https://doi.org/10.1111/febs.16124

Staes, A., Damme, P. V., Helsens, K., Demol, H., Vandekerckhove, J., & Gevaert, K. (2008). Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC). *PROTEOMICS*, *8*(7), 1362–1370. https://doi.org/10.1002/pmic.200700950

Stone, J. A., McCrea, J. B., Witter, R., Zajic, S., & Stoch, S. A. (2019). Clinical and translational pharmacology of the cathepsin K inhibitor odanacatib studied for osteoporosis. *British Journal of Clinical Pharmacology*, *85*(6), 1072–1083. https://doi.org/10.1111/bcp.13869

Strelow, J. M. (2017). A Perspective on the Kinetics of Covalent and Irreversible Inhibition. *SLAS DISCOVERY: Advancing the Science of Drug Discovery*, *22*(1), 3–20. https://doi.org/10.1177/1087057116671509

Tully, D. C., Liu, H., Alper, P. B., Chatterjee, A. K., Epple, R., Roberts, M. J., Williams, J. A., Nguyen, K. T., Woodmansee, D. H., Tumanut, C., Li, J., Spraggon, G., Chang, J., Tuntland, T., Harris, J. L., & Karanewsky, D. S. (2006). Synthesis and evaluation of arylaminoethyl amides as noncovalent inhibitors of cathepsin S. Part 3:

Heterocyclic P3. *Bioorganic & Medicinal Chemistry Letters*, *16*(7), 1975–1980. https://doi.org/10.1016/j.bmcl.2005.12.095

Turk, B. (2006). Targeting proteases: Successes, failures and future prospects. *Nature Reviews. Drug Discovery*, *5*(9), 785–799. https://doi.org/10.1038/nrd2092

Turk, B. E., Huang, L. L., Piro, E. T., & Cantley, L. C. (2001). Determination of protease cleavage site motifs using mixture-based oriented peptide libraries. *Nature Biotechnology*, *19*(7), Article 7. https://doi.org/10.1038/90273

Turk, D. (2013). MAIN software for density averaging, model building, structure refinement and validation. *Acta Crystallographica Section D: Biological Crystallography*, *69*(Pt 8), 1342–1357. https://doi.org/10.1107/S0907444913008408

Turk, D., Gunčar, G., Podobnik, M., & Turk, B. (1998). Revised Definition of Substrate Binding Sites of Papain-Like Cysteine Proteases. *Biological Chemistry*, *379*, 137–147. https://doi.org/10.1515/bchm.1998.379.2.137

Turk, V., Kos, J., & Turk, B. (2004). Cysteine cathepsins (proteases)—On the main stage of cancer? *Cancer Cell*, *5*(5), 409–410. https://doi.org/10.1016/s1535-6108(04)00117-5

Turk, V., Stoka, V., Vasiljeva, O., Renko, M., Sun, T., Turk, B., & Turk, D. (2012). Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, *1824*(1), 68–88. https://doi.org/10.1016/j.bbapap.2011.10.002

Twining, S. S. (1994). Regulation of Proteolytic Activity in Tissues. *Critical Reviews in Biochemistry and Molecular Biology*, *29*(5), 315–383. https://doi.org/10.3109/10409239409083484

Unanue, E. R., Turk, V., & Neefjes, J. (2016). Variations in MHC Class II Antigen Processing and Presentation in Health and Disease. *Annual Review of Immunology*, *34*(1), 265–297. https://doi.org/10.1146/annurev-immunol-041015-055420

Vagin, A., & Teplyakov, A. (1997). MOLREP: An Automated Program for Molecular Replacement. *Journal of Applied Crystallography*, *30*(6), Article 6. https://doi.org/10.1107/S0021889897006766

Vidmar, R., Vizovišek, M., Turk, D., Turk, B., & Fonović, M. (2017). Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *The EMBO Journal*, *36*(16), 2455–2465. https://doi.org/10.15252/embj.201796750

Vita, E. D. (2021). 10 years into the resurgence of covalent drugs. *Future Medicinal Chemistry*, *13*(2), 193–210. https://doi.org/10.4155/fmc-2020-0236

Vizovišek, M., Fonović, M., & Turk, B. (2019). Cysteine cathepsins in extracellular matrix remodeling: Extracellular matrix degradation and beyond. *Matrix Biology*, *75–76*, 141–159. https://doi.org/10.1016/j.matbio.2018.01.024

Vizovišek, M., Vidmar, R., Van Quickelberghe, E., Impens, F., Andjelković, U., Sobotič, B., Stoka, V., Gevaert, K., Turk, B., & Fonović, M. (2015). Fast profiling of protease specificity reveals similar substrate specificities for cathepsins K, L and S. *Proteomics*, *15*(14), 2479–2490. https://doi.org/10.1002/pmic.201400460

Winn, M. D., Ballard, C. C., Cowtan, K. D., Dodson, E. J., Emsley, P., Evans, P. R., Keegan, R. M., Krissinel, E. B., Leslie, A. G. W., McCoy, A., McNicholas, S. J., Murshudov, G. N., Pannu, N. S., Potterton, E. A., Powell, H. R., Read, R. J., Vagin, A., & Wilson, K. S. (2011). Overview of the CCP4 suite and current developments. *Acta Crystallographica Section D: Biological Crystallography*, *67*(Pt 4), 235–242. https://doi.org/10.1107/S0907444910045749

Yadati, T., Houben, T., Bitorina, A., & Shiri-Sverdlov, R. (2020). The Ins and Outs of Cathepsins: Physiological Function and Role in Disease Management. *Cells*, *9*, 1679. https://doi.org/10.3390/cells9071679

Yamashita, D. S., Marquis, R. W., Xie, R., Nidamarthy, S. D., Oh, H.-J., Jeong, J. U., Erhard, K. F., Ward, K. W., Roethke, T. J., Smith, B. R., Cheng, H.-Y., Geng, X., Lin, F., Offen, P. H., Wang, B., Nevins, N., Head, M. S., Haltiwanger, R. C., Narducci Sarjeant, A. A., … Veber, D. F. (2006). Structure Activity Relationships of 5-, 6-, and 7-Methyl-Substituted Azepan-3-one Cathepsin K Inhibitors. *Journal of Medicinal Chemistry*, *49*(5), 1597–1612. https://doi.org/10.1021/jm050915u

Yang, W.-L., Li, Q., Sun, J., Huat Tan, S., Tang, Y.-H., Zhao, M.-M., Li, Y.-Y., Cao, X., Zhao, J.-C., & Yang, J.-K. (2022). Potential drug discovery for COVID-19 treatment targeting Cathepsin L using a deep learning-based strategy. *Computational and Structural Biotechnology Journal*, *20*, 2442–2454. https://doi.org/10.1016/j.csbj.2022.05.023

Zajic, S., Rossenu, S., Hreniuk, D., Kesisoglou, F., McCrea, J., Liu, F., Sun, L., Witter, R., Gauthier, D., Helmy, R., Joss, D., Ni, T., Stoltz, R., Stone, J., & Stoch, S. A. (2016). The Absolute Bioavailability and Effect of Food on the Pharmacokinetics of Odanacatib: A Stable-Label i.v./Oral Study in Healthy Postmenopausal Women. *Drug Metabolism and Disposition*, *44*(9), 1450–1458. https://doi.org/10.1124/dmd.116.069906

Zhou, J., Zhang, Y.-Y., Li, Q.-Y., & Cai, Z.-H. (2015). Evolutionary History of Cathepsin L (L-like) Family Genes in Vertebrates. *International Journal of Biological Sciences*, *11*(9), 1016–1025. https://doi.org/10.7150/ijbs.11751

Bibliography

Publications Related to the Thesis

Tušar, L.*, **Loboda, J.*,** Impens, F.*, Sosnowski, P., Van Quickelberghe, E., Vidmar, R., Demol, H., Sedeyn, K., Saelens, X., Vizovišek, M., Mihelič, M., Fonović, M., Horvat, J., Kosec, G., Turk, B., Gevaert, K., Turk, D. Proteomic data and structure analysis combined reveal interplay of structural rigidity and flexibility on selectivity of cysteine cathepsins. Accepted in Communications Biology. **Authors equally contributed.*

Mons, E., Jansen, I. D. C., **Loboda, J.**, van Doodewaerd, B. R., Hermans, J., Verdoes, M., van Boeckel, C. A. A., van Veelen, P. A., Turk, B., Turk, D., & Ovaa, H. (2019). The Alkyne Moiety as a Latent Electrophile in Irreversible Covalent Small Molecule Inhibitors of Cathepsin K. Journal of the American Chemical Society, 141(8), 3507–3514. https://doi.org/10.1021/jacs.8b11027

Günther, S., Reinke, P. Y. A., Fernández-García, Y., Lieske, J., Lane, T. J., Ginn, H. M., Koua, F. H. M., Ehrt, C., Ewert, W., Oberthuer, D., Yefanov, O., Meier, S., Lorenzen, K., Krichel, B., Kopicki, J.-D., Gelisio, L., Brehm, W., Dunkel, I., Seychell, B., … **Loboda, J.**, … Meents, A. (2021). X-ray screening identifies active site and allosteric inhibitors of SARS-CoV-2 main protease. Science (New York, N.Y.), 372(6542), 642–646. https://doi.org/10.1126/science.abf7945

Reinke, P.Y.A., De Souza, E.E., Günther, S., Falke, S., Lieske, J., Ewert, W., **Loboda, J.**, Hermann, A., Karničar, K., Usenik, A., Lindič, N., Sekirnik, S., Botosso, V.F., Kapronezai, J., De Araújo, M.W., Silva-Pereira, T.T., Filho, A.F.S., Tavares, M.S., Giaretta, P.R., Mashhour, A.R., Hauser, M., Lach, M., Böhler, H., Beck, T., Cox, R., Chapman, H.N., Betzel, C., Hinrichs, W., Ebert, G., De Sá Guimarães, A.M., Turk, D., Wrenger, D., Meents, A. Sulfonated Calpeptin is a promising drug candidate against SARS-CoV-2 infections. *Submitted.*

Other Publications

Lindič, N., **Loboda, J.**, Usenik, A., Vidmar, R., & Turk, D. (2020). The Structure of Clostridioides difficile SecA2 ATPase Exposes Regions Responsible for Differential Target Recognition of the SecA1 and SecA2-Dependent Systems. International Journal of Molecular Sciences, 21(17), 6153. https://doi.org/10.3390/ijms21176153

. . .

Biography

Jure Loboda, the author of this thesis, was born on April 19, 1991 in Ljubljana. He completed his secondary education at the Gymnasium Rudolf Maister in Kamnik in 2010. After that, from 2010 – 2016, he studied at the Faculty of Pharmacy in University of Ljubljana, Slovenia. In his final year of study, he received a prize for excellent grades. He worked as a student researcher at the Institute of Biochemistry at the University of Ljubljana, Slovenia, under the supervision of Doc. Dr. Metka Lenassi. The group focused on vesicles that microglia, cells of the central nervous system, release in the extracellular space under different in vitro conditions. His tasks were mainly cell culturing, isolation of secreted vesicles and their proteins from extracellular media and western blot analysis. In May 2016, he successfully defended his master's thesis, titled "Protein and Morphological Analysis of Vesicles, secreted by Human Microglia," which contained the results of his research. In the same year, he was selected to a PhD position at the Jožef Stefan Institute at the department of Biochemistry and Molecular and Structural Biology, in the group for Structural Biology, led by Prof. Dr. Dušan Turk. His research aims was to better understand biological roles of cysteine cathepsins, major endosomal proteases, by gaining new insights into their substrate specificity, as well as characterising novel cathepsin inhibitors. Additionaly, he worked on crystal structure determination of various Clostridia *difficile* cell wall proteins and with characterisation of novel inhibitors of SARS-CoV-2 main protease (M^{pro}). His tasks included molecular cloning, protein expression, purification and crystallization, protein structure determination, operative knowledge of RP-HPLC, characterisation of substrates and inhibitors of various proteases by determination of their K_m, IC_{50} and K_i values, and sample preparation for mass spectrometry. During his work, he attended several international conferences, where he presented his work. He is one of the first authors of one original scientific paper, and one of the co-authors on three other papers. He is also a co-author on one paper, which is in review, and co-author on few other papers which are in the preparation at the time of writing.