UiT

THE ARCTIC
UNIVERSITY
OF NORWAY

Faculty of Health Sciences

Molecular interaction studies of initial electrostatic attraction between trypsin and the human PAR-2 receptor

Molecular Pharmacology and Toxicology research group, Department of medical biology

—

Tonje Håtveit Kristoffersen

Master's thesis in Biomedicine, January 2020

Table of Contents

Acknowledgements

Abstract

Workers in livestock and fish cultivation are at increased risk of occupational airway damage caused by proteases. Proteases, such as trypsin, activate PAR-2 which in turn triggers an inflammatory response, potentially causing airway damage over time. There has been some speculation that PAR-2 receptors easier attract trypsin form species where this enzyme has a more negative electrostatic charge. A molecular modelling approach was used to assess the initial binding of the activating peptide segment of PAR-2 to trypsin from multiple animal species.

Homology modelling was used to predict the structures of Pacific sardine trypsin, yellowtail trypsin and red king crab trypsin, as well as to construct the N-terminal peptide segment of PAR-2. Protein-protein docking was performed to predict initial surface interactions between the PAR-2 peptide segment and trypsin. The binding interaction was mapped, and the interacting amino acids were compared across the species, as well as the charge of the protein binding surfaces.

The study indicates that there is, at least, a stronger initial interaction between the N-terminal peptide segment of PAR-2 and trypsin with a stronger negative charge.

Abbreviations

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimensional
BLAST	Basic Local Alignment Search Tool
C-terminal	Carboxyl terminal
Ca^{2+}	Calcium
DUD-E	Directory of useful decoys: enhanced
ER	Enrichment factors
eV	Electronvolt
GPCR	G-protein coupled receptor
H-bond	Hydrogen bond
HCl	Hydrochloric acid
kDa	Kilodalton
N-terminal	Amino terminal
NMR	Nuclear magnetic resonance
PAR	Protease-activated receptor
PDB	Protein Data Bank
pKa	Acid dissociation constant
RMSD	Root-mean-square deviation
SAVES	Structural Analysis and Verification Server
UniProt	Universal Protein Resource
$Å^2$	Square Ångstroms

1 Introduction

1.1 The role of proteases in airway inflammation

The airways are a continuous part of an organisms epithelium, whose primary role is gas exchange between an organism and its external environment, while also working as a barrier between the internal organs and the external environment[1, 2]. The airways are primarily composed of nose, oropharynx, larynx, trachea, bronchi, bronchioles and lungs, where the latter further divides into alveoli where gas exchange occurs[2].

Proteolytic enzymes, or proteases, catalyse the cleavage of peptide bonds by hydrolysis, which is the addition of a water molecule where one hydrogen is added to one of the resulting products and the other hydrogen and the oxygen to the other resulting product[3].

Workers in agriculture, including the cultivation of livestock, are at increased risk of inflammatory airway diseases, including occupational asthma, rhinosinusitis, and chronic obstructive pulmonary disease[4, 5]. Studies have indicated that the presence of proteases in the airways trigger a reaction by T helper 2 cells and release of related inflammatory cytokines[4, 6], whose excessive activation can lead to allergic disease and fibrosis[7]. Proteinase activated receptors (PARs), in particular PAR-2, have been found to play an integral part of these protease-induced immune reactions[4-6, 8], and is further discussed in section 1.3.

The research article *Differences in PAR-2 activating potential by king crab (Paralithodes camtschaticus), salmon (Salmo salar), and bovine (Bos taurus) trypsin* by Larsen *et. al.*, 2013 published in *The Journal of Pharmacology and Experimental Therapeutics* discusses the role of electrostatic potential and divergent amino acid residues of the trypsin binding site and their role in PAR-2 binding interactions, using increased risk of occupational airway damage in workers cultivating fish and crustaceans as a basis. The article suggests that differences in electrostatic potential of trypsin across species may warrant variations in assay substrate binding, with trypsin derived from fish generally showing higher enzymatic activity. The article emphasizes that the assayed trypsin may different behaviour upon PAR-2 binding rather than assay substrate binding[9].

1.2 Trypsin

Enzymes are proteins that greatly enhance the reaction rate of chemical compounds, for example between a peptide bond and water molecule as mentioned in the section 1.1, creating new products[3].

Trypsin is a digestive enzyme synthesised in the pancreas and subsequently released into the small intestine where it hydrolyses peptide bonds of dietary protein. Trypsin belongs to the S1 family of the PA superfamily of serine proteases[10, 11], which get their name from the nucleophilic serine residue of a catalytic residue triad that attacks the peptide-substrate carbonyl to form an acyl-enzyme complex. Most members of the PA superfamily, including trypsin, have a substrate specificity for peptides with a positively charged residue, usually arginine or lysine, at the P1 position of the peptide[12].

The three-dimensional (3D) structure of trypsin is generally well-conserved across different species, although the amino acid sequence can vary significantly between species, but the position of key amino acid residues involved in the catalytic reaction are generally the same[10].

Trypsin is synthesised as its inactive precursor, or proenzyme, trypsinogen in the exocrine (acinar) cells of the pancreas and stored in granules awaiting secretion into the intestinal lumen[3, 10, 13]. Synthesizing and storing the enzyme as a proenzyme and packaging it in granules helps protect the pancreatic cells from unwanted proteolytic activity and subsequent damage from the active enzyme[3, 13].

The digestion of dietary protein starts in the stomach by the enzyme pepsin and hydrochloric acid (HCl), secreted by the parietal cells and chief cells of the gastric glands of the mucosa, respectively. The acidity of the stomach (pH 2-5) provided largely by HCl works as a denaturing agent that unfolds globular proteins to make the peptide bonds more accessible to proteolytic activity by pepsin, which has a broad specificity for peptide bonds, and subsequent proteolytic enzymes[3, 14]. Upon gastric emptying into the duodenum (upper part of small intestine), the partially digested dietary protein and amino acids stimulate the release of bicarbonate and several pancreatic digestive enzymes, including trypsinogen. The bicarbonate raises neutralizes the gastric acid and raises the pH to approximately 7, which creates a more optimal environment for these pancreatic enzymes[3]. When trypsinogen is released into the intestinal lumen, an amino-terminal (N-terminal) peptide sequence is cleaved off at a lysine residue by enteropeptidase, which is a protease secreted by cells lining the

intestine[3, 13, 14]. In addition to being activated by enteropeptidase, trypsin itself also contributes to the cleavage of trypsinogen and its subsequent activation, as well as activating a number of other pancreatic proenzymes in the same manner[13, 14]. The activation cleavage of the N-terminal peptide of trypsinogen to trypsin triggers a conformational change that exposes the active site of the enzyme, rendering it ready for proteolytic action[3]. Trypsin contributes to the digestion of dietary protein with a specificity for cleaving internal bonds of the amino acids arginine and lysine, breaking protein down to shorter peptides that are further broken down into shorter peptides or free amino acids that are subsequently transported through the epithelial cells lining the small intestine and absorbed into the blood stream[3, 13].

The active site of trypsin is defined by its conserved catalytic triad made up by the amino acids serine, histidine and aspartic acid in specific positions[10, 15]. Upon peptide (substrate) binding the catalysis is initiated when the hydroxyl group of the serine acts as a nucleophile (electron donor) and attacks the carbonyl carbon of the substrate residue, a lysine or an arginine, while the histidine acts as a general base that increases the nucleophilic property of the serine by having one of the nitrogens of its imidazole group acting as a proton acceptor, and the result is an intermediate acyl-enzyme complex. The aspartic acid is believed to help stabilize the catalytic triad by forming a hydrogen bond from its carboxyl group to the other nitrogen of the imidazole group, polarizing the histidine and allowing it to act as a proton acceptor in the acylation reaction, in effect making the catalytic triad a charge relay system[10, 15, 16]. After the formation of the acyl-enzyme complex, a deacylation occurs by a similar reaction in reverse, but in place of the serine, a water molecule from the solvent acts as the attacking nucleophile, and the peptide bond is finally cleaved by hydrolysis[15, 16]. Additional residues in the trypsin binding site contribute to substrate binding without being directly involved in the catalysis, namely an aspartic acid and two glycine residues that help facilitate lysine or aspartic acid recognition. The non-catalytic aspartate binds the substrate residue primarily through electrostatic interaction between the positively charged substrate residue and the negatively charged aspartate, while the glycine residues are positioned on opposite sides of the binding pocket and interact with the hydrocarbon chain of the bound substrate residue[10, 16]. In addition to the residues that help bind substrate, the amide hydrogen of a glycine two residues upstream from the catalytic serine help stabilize the substrate residue during the transitional state of the catalysis[15, 16].

1.3 Proteinase activated receptor 2

Membrane receptors are proteins embedded in a membrane structure either inside or on the surface of a cell, that bind ligands which in turn initiate transmission of signals to other parts of the cell to either induce activity by other proteins or induce changes in gene expression[17].

Proteinase activated receptors (PARs) are a family belonging to the seven-transmembrane G-protein activated receptors (GPCRs) superfamily, and consists of four members: PAR 1-4[8, 18]. These receptors are embedded in the plasma membrane and have a characteristic structure consisting of a single polypeptide chain roughly 400 amino acids long, with seven alpha helices that each span the plasma membrane, an extracellular N-terminal domain, and an intracellular carboxyl terminal (C-terminal) domain[18-20]. As their name suggests, PARs are activated by proteases, specifically serine proteases, through cleavage of a short peptide sequence positioned at the N-terminal, unmasking a new N-terminal that functions as a tethered ligand that binds to the body of the receptor activating G proteins and subsequently stimulating mobilization of IP_3 and mobilization of Ca^{2+}[8, 18, 19]. This mechanism of activation is illustrated in Figure 1, using PAR-2 as a model. The activation of PARs is irreversible, and activated receptors shut off by internalization and degradation in lysosomes[18, 19].
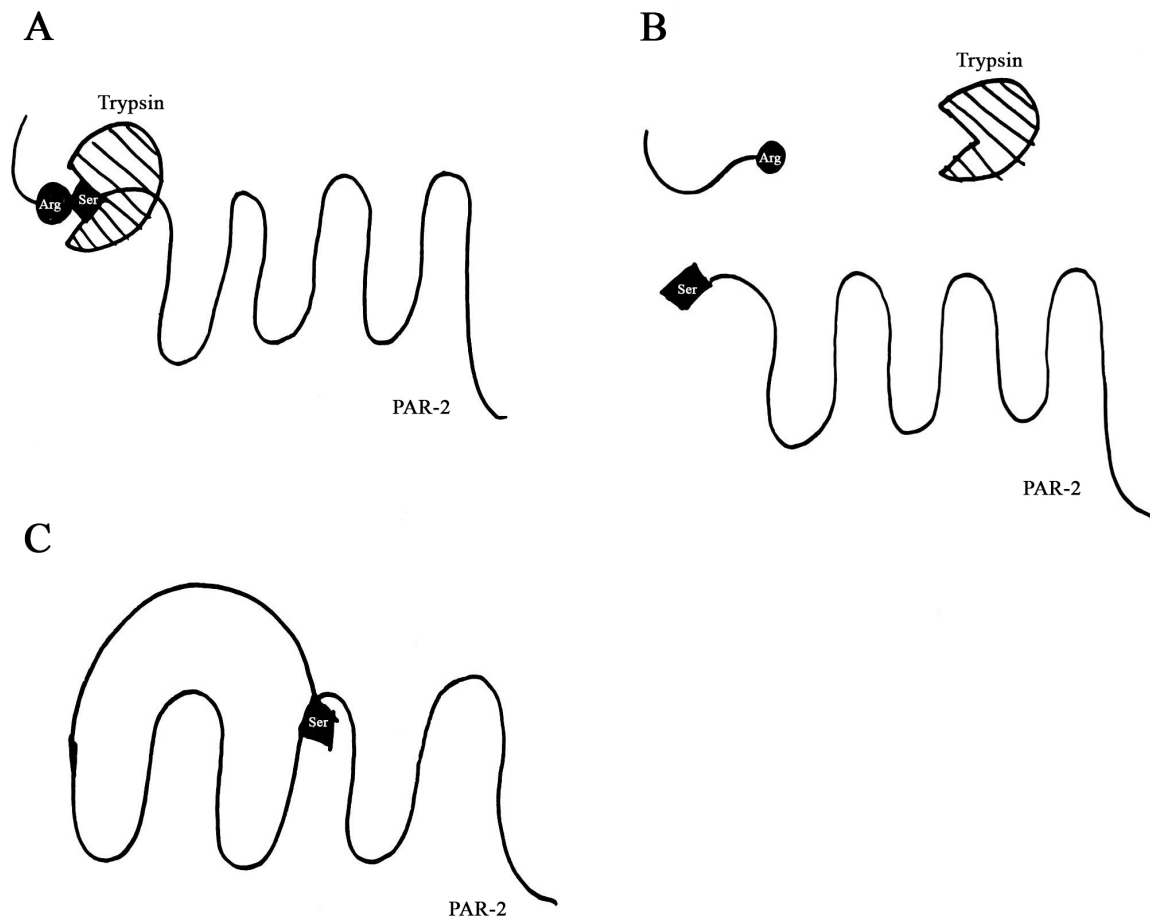
Figure 1 – Unmasking of the tethered ligand of PAR-2 by trypsin. The figure illustrates the mechanisms involved in activating the PAR-2 receptor based on[19]. Figure 1A shows trypsin cleaving the peptide bond between the arginine and serine residues, unmasking the tethered ligand (1B) at the new N-terminal, that subsequently binds to the body of the receptor(1C).

Thrombin is a known activator of PAR-1, PAR-3 and PAR-4, while PAR-1 and PAR-3 can also be activated by the enzyme factor Xa, and cathepsin G is a known activator of PAR-4. The general function of PAR activation is believed to be recruitment of leukocytes and platelets in inflammatory responses and hemostasis[19]. The PARs are expressed in a vast variety of tissues throughout the mammalian body[8, 19].

PAR-2 is expressed in a variety of tissues, including the nervous system, endothelium, and the epithelial cells of the skin, gastrointestinal tract, and respiratory system[8, 21]; but here we will focus on the latter. Known activators of PAR-2 include trypsin, tryptase, factor Xa, tissue factor-Factor VIIa complex, membrane-type serine protease 1, and a number of exogenous activators including proteases form dust mite allergens, fungi and cockroaches[5, 18, 19]. Activation of PAR-2 in airway epithelium is believed to trigger inflammatory responses

through release of cytokines that recruit inflammatory cells that cause increase in vascular permeability, leukocyte infiltration and airway hyperreactivity[8, 21, 22]. PAR-2 activation has been shown to play a dual part the smooth muscle of the airways, causing both relaxation and constriction of the bronchi. Relaxation of the bronchi has a protective function through decreased total airway resistance, while bronchial constriction causes an increase in total airway resistance. The nature of smooth muscle response to PAR-2 activation has been speculated to differ between specific tissue or region of the airways, as well as differing between species[8, 21].

Human PAR-2 is 397 amino acids long, and proteolytic cleavage between an arginine and serine at positions 36 and 37, respectively, unmasks its tethered ligand with the sequence SLIGKV that subsequently binds to the body of the receptor[8, 18]. Humans only have one copy of the PAR-2 gene, located in the 13th cluster of the long arm (q) on chromosome 5[18].

1.4 Molecular modelling

Molecular modelling is a theoretical, computer-based (*in silico*) approach to studying the 3D structure and molecular interactions of proteins, macromolecules or other chemical compounds. Molecular modelling applies established principles in physics, chemistry and biology and experimental data to mathematically predict and describe the molecular energy of inter- and intramolecular interactions of 3D structures, including their relative positioning, bonds, bond lengths and angles, attractive and repulsive forces, and geometry. Such interactions can be described using molecular mechanics, quantum mechanics or a combination of both. Molecular mechanics uses the information and data of established structures to predict a theoretical model of a molecule. Molecular mechanics do not take in to account certain chemical properties, and thus is more appropriate to use when predicting larger molecular structures such as proteins. Quantum mechanics take additional properties, such as the behaviour of electrons, into account when predicting molecular structures, but require much more powerful computational effort than the molecular mechanics method[23]. The energy functions that calculate the total energy and help find the minimum energy of a model are called 'force fields' and are divided into three categories: physics-based, knowledge-based or hybrid potentials. Physics-based force fields try to accurately calculate the actual physical potential energy of a protein conformation. Knowledge-based force fields use conformational knowledge of already resolved structures to calculate the energy of a protein conformation. Hybrid potentials are a combination of physics-based and knowledge-

based potentials[24]. 3D models can be resolved either through an experimental approach, like X-ray crystallography or Nuclear Magnetic Resonance (NMR) spectroscopy, or theoretically predicted based on similar and related structures using computational methods in the process of homology modelling[24, 25].

Molecular modelling has a wide area of application, for instance to study the structure-activity relationships in drug design, but here the focus will be on homology modelling and predicting protein interactions through docking.

1.4.1 Homology modelling

Homology modelling is the process of theoretically predicting the 3D structure of a protein with a known amino acid sequence by applying the principle that evolutionary related proteins have similar structures with highly conserved folding and have already been resolved through experimental means[26, 27]. The process of homology modelling consists of five main steps: (1) identification of template(s), (2) sequence and structure alignment of target and template, (3) model construction, (4) model refinement, and (5) model validation[24, 27]. These steps might need repeating until a satisfying model is achieved. Figure 2 summarizes the flow of the homology modelling process.

A template is the structure-sequence to be used as a base (template) for constructing a model of a desired target (amino acid sequence to be modelled). Identification of a suitable template is usually achieved by searching a database, usually the Basic Local Alignment Search Tool (BLAST)[28], for protein sequences that have already been resolved through experimental means and have high identity to the target sequence. The target structure is usually imported from another database, usually the Protein Data Bank (PDB)[29, 30].

A higher degree of identity and query coverage between a target and template often predicts a more accurate model. BLAST applies an alignment algorithm (for example ClustalW or T-coffee) that aligns the amino acid sequences of target and template and which results in a score that indicates the degree of identical amino acid residues between the two sequences[24, 27]. An identity of 30% is usually adequate for sequences longer than 100 amino acid residues[26]. Sequences with lower identity and/or coverage may need additional methods of alignment or even manual correction, while sequences with a high degree of identity and coverage often do not require manual correction[24, 26].

The model construction step can be further divided into three key steps: (1) backbone generation, (2) loop modelling and (3) side-chain modelling. The backbone of the model is simply generated using the backbone coordinates of the template structure in a process called rigid-body assembly, and if the amino acid side-chains are identical for corresponding residues, these are also simply copied from the template structure[24, 26]. Another method for backbone generation is segment matching, which uses both the atomic positions of the template backbone and a reference database of similar segments as to guide the model construction[24, 27]. A third option is to model by satisfaction of spatial restraints of the initial alignment and generates a model by minimizing the violations of these restraints[24]. Loop modelling may be necessary if the alignment contains gaps between target and template and is essentially the process of reconstructing the affected area, which often occurs in less conserved loop regions of proteins[24, 26]. Loop regions of a protein tend to be more dynamic since they lack proper secondary structure rigidity, thus there can be more than one correct conformation of such segments. Loop modelling can be carried out using knowledge-

based or energy-based force fields[26]. Side chain modelling is the process of assigning side chains that are not identical to the template onto the constructed backbone, and are placed using, at least partially, a knowledge-based preferred protein sequence[26, 27].

Model refinement is in essence the final step of the construction and boils down to energy minimization using force fields, Monte Carlo sampling and/or molecular dynamics on the entire structure. However, this is a delicate process since it requires an accurate structure prediction to result from the previous steps. If significant errors in loop or side-chain modelling have occurred, the structure may deviate further from an ideal conformation, and in addition too much energy minimization might also yield this result. Thus, an iterative process of model alignment, construction, refinement and validation is often necessary[24, 26, 27]. Monte Carlo sampling is algorithm-based and allows for focus on those regions of the structure most likely to contain errors to be minimized in an all-atom force field and can improve upon backbone and side-chain conformations, but depends on sequence similarity between target and template[27].

Model validation software can help localize and identify problems in a constructed model, as well as to assess the overall quality. It is especially important that the quality of the active site of a protein structure is high. The Structural Analysis and Verification Server (SAVES)[31] is a tool for assessing both local and global structural quality, including stereochemistry and the relationship between the amino acid sequence and 3D structure of the model, while docking and scoring can help assess the quality of an active site by discriminating between known active compounds with affinity to the active site from decoy compounds that probably do not have a high binding affinity[24, 27].

1.4.2 Docking and scoring

Docking of active ligands with known affinity to homology models can serve as an additional step in the model validation process, by assessing the quality of the active site of constructed models. Docking of ligands to the active site of a protein predicts how a ligand will bind and interact with the amino acid residues of a protein's binding site, including its conformation(s) and free energy. Docking algorithms search for the best possible poses and roughly estimate the binding of a ligand to a defined protein binding site, by searching the conformational space and sampling the many possible ways a ligand can bind by applying a scoring function[25]. Docking algorithms usually take ligand flexibility into account, but usually treats the docking target as rigid. Although many algorithms also have the option of receptor

flexibility of the binding site to be taken into account, this requires a lot of time and computational power due to the addition of significant degrees of freedom[25]. Even without taking receptor flexibility into account, docking of ligands to a receptor is an exhaustive process, especially when docking thousands or millions of compounds. Thus, meta-heuristic methods are applied to conduct prior filtering[32]. In some docking algorithms, for example GLIDE, greedy scoring and filtering steps are applied to dismiss compounds that have unfavourable binding to a target before taking compounds with an estimated more favourable affinity to the receptor further for more meticulous docking and scoring, which usually is more computationally challenging[33]. Docking and scoring of ligands have a broader spectre of use than just assessment of homology model quality, and is also used in drug discovery and design[25, 32].

Scoring functions also estimates the binding affinity between ligand and receptor in its most favourable pose(s), assigning the best score to the ligand with lowest free energy and strongest binding interactions[25, 32, 33]. Such scoring functions take several factors into account, for example Van der Waals energy, Coulomb energy, hydrophobic interactions, hydrogen-bonds, metal-binding, polar interactions, and other rewards and penalties for favourable and unfavourable interactions[33].

Docking and scoring can be used as statistical analysis in quality assessment of binding sites of homology models, by implicating a set of ligands with known binding affinity (positives) to the target, called active compounds, along with a much larger set of inactive ligands (false positives), called decoys. The ranking of active compounds and decoys can be used in measuring enrichment factors (ER) and area under the receiver curve (ROC). ER is a measure of the portion of actives are found within a given top fraction N% of the ranked ligands relative to a hypothetical equal distribution. ROC-plots are made based on a resulting list of ranked ligands, moving up the Y-axis per actives encounter and along the X-axis per decoy encountered, resulting in a curve. The area under the curve reveals weather the actives bind more favourably to the receptor, or if the distribution of ranked actives and decoys is random. Early rankings of actives thus play crucial part in calculating the area under the curve, which is given a value between 0 and 1, with a high score indicating early ranking of actives, and a score of 0.5 indicating random distribution of ranked actives and inactives[34].

1.4.3 Protein-protein docking

Protein-protein docking is the process of docking the desired area of two proteins against each other at the lowest free energy conformation, by either template-based or direct docking. Template-based docking is usually reserved for structures that share more than 30% sequence identity that often interact in the same way, thus providing an interface for docking of homologous proteins. Direct docking uses the basis of thermodynamics to find the structure target at the lowest free energy within the conformational space, requiring a free energy evaluation model and minimization algorithm. ClusPro[35] preforms direct docking of two proteins in three steps: (1) a rigid-body docking sampling of billions of conformations, (2) a clustering of the 1000 lowest energy conformations based on root-mean-square deviation (RMSD) that finds the largest clusters of models that are the most likely models of the complex, and (3) refinement of the structures using energy minimization. The rigid-body docking uses PIPER, which places one protein on a rigid grid and the other on a movable grid, and interaction energy is given based on energies at each grid point. The rigid-body docking does not take into account the flexibility of either protein. Approximately 30 highly populated clusters of low-energy structures with similar docking conformations are selected as predictions of the protein-protein complex and returned as the most observed poses[35].

Such protein-protein docking does not necessarily return the native structure of the complex as the highest-ranking pose, but rather the most observed poses in energetically favourable conformations[35].

2 Aim of study

Several species have resolved structures for trypsin, but Pacific sardine, yellowtail and red king crab do not have resolved structures. These were picked for homology modelling due to their big impact in the fishery business, where workers may be at risk for airway inflammation and damage. Nor is there a resolved structure for the N-terminal peptide segment of PAR-2, which is the segment cleaved off by proteases to activate PARs.

The project was essentially composed of two major steps:

1) Construction of 3D homology models of the target trypsin and PAR-2 segment, followed by structure refinement, validation and quality assessment, including docking and scoring of actives and decoys.
2) Protein-protein docking of trypsin structures to PAR-2 peptide segment, and subsequent interaction and binding site studies, including assessment of surface charge.

3 Methods and materials

3.1 Materials

3.1.1 Databases

UniProt

The Universal Protein Resource (UniProt)[36] is an online resource database containing manually curated and reviewed protein sequence and annotation data, including information about structure and function of individual proteins as well as data about complete proteomes. The data is obtained through experimental techniques and large-scale sequencing of protein and proteomes[36]. UniProt is made up of three databases, UniProt Knowledgebase (UniProtKB), UniProt Reference Clusters and UniProt Archive; UniProt is a collaboration between by European Bioinformatics Institute, a part of European Molecular Biology Laboratory (EMBL-EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR)[36].

MEROPS

MEROPS[37] is an online database of peptidases and peptidase inhibitors, containing sequences and sequence information, classification and nomenclature, such as substrate binding sites and catalytic residues, classification into sequence homolog clusters, related sequences are clustered into families, and related tertiary structures are clustered into clans[37]. MEROPS is provided by EMBL-EBI[37].

BLAST

The Basic Local Alignment Search Tool (BLAST)[28] is an online sequence similarity search program for comparing and aligning nucleotide or protein query sequences with nucleotide or protein sequences in various selected online databases, or for comparing and aligning two or more nucleotide or protein sequences, while also providing additional statistical information about such alignments[28]. BLAST is provided by National Centre for Biotechnology Information (NCBI)[28].

PDB

The Protein Data Bank (PDB)[29, 30] is an online resource database containing information about the 3D structure and spatial arrangement of proteins, nucleotides and complex assemblies. The structural and spatial information is obtained through methods such as X-ray crystallography, NMR spectroscopy, cryo-electron microscopy and theoretical modelling[29, 30]. PDB is run by Research Collaboratory for Structural Bioinformatics (RCSB)[29, 30].

DUD-E

The Directory of Useful Decoys: Enhanced (DUD-E)[38] is an online database containing proteins, their known active compounds and decoys for each structure, and provides decoys for enrichment calculations generated using similarity fingerprints to minimize the topological similarity between ligands and decoys. DUD-E is available to generate decoys for any target based on a list of known actives[38].

SAVES

The Structural Analysis and Verification Server (SAVES)[31] is an online metaserver that runs multiple programs for checking and validating protein structures during and after model refinement, including Verify3D[39], which determines the compatibility of an atomic model (3D) with its own amino acid sequence (1D), ERRAT, which checks the overall quality factor for non-bonded interactions and which residues that are creating problems in the model, and PROCHECK, which offers detailed Ramachandran plots that assess the stereochemical quality of the 3D protein structure[31].

3.1.2 Software

Schrödinger software release 2019-1 and 2019-3

The Schrödinger software package offers *in silico* simulation and analysis of chemical compounds and their properties on an atomic scale, allowing discovery and optimization of structures and compounds ahead of synthesis and assays, using specialized tools with high predictive power. This is accomplished by applying principles of classical and quantum physics and next-generation machine learning techniques in algorithms and calculations. The Schrödinger software package is used in drug discovery, pharmaceuticals, predictive modelling, biotechnology, and a variety of materials research areas.

Maestro

Maestro[40] is the graphical user interface of a large portion of the Schrödinger software packages, offering tools for visualizing, building, editing and analysing structures and compounds. The interface offers options for organizing and storing such entries, as well as setting up, running and monitoring related jobs[40].

Prime

Prime[41] is a program accessible through the Maestro interface of the Schrödinger software package that offers protein structure prediction and homology modelling based on Comparative Modelling and Fold Recognition. It takes the user through a step-by-step process of submitting a target amino acid sequence, identifying and selecting a suitable template structure, alignment of target sequence and template(s), and model construction. The resulting 3D structure can be further refined after construction, including loop refinement, side-chain prediction and minimization[41].

Protein Preparation Wizard

The Protein Preparation Wizard[42] is a program used in the Maestro interface of the Schrödinger software package that offers an automatic process to prepare a protein to a form suitable for further use in other Schrödinger software programs. It is primarily used on raw crystal structures to add missing information on connectivity, such as bond orders and formal charge of atoms. The Protein Preparation Wizard offers three steps to fixing the structure, importing and basic structure fixes, modifying and deleting unwanted co-crystallized structures, compounds, molecules and other het-groups, and refinement by optimization of hydrogen bond groups' orientation and minimizing the protein structure[42].

LigPrep

LigPrep[43] is a program in the Schrödinger software package available through the Maestro interface, that is used to convert large numbers of 2D or 3D structures into energetically favourable 3D structures with correct chiralities ready for further use by other Schrödinger software programs. LigPrep offers the option of generating one or multiple variations a structure could provide, including ionization states, tautomers, stereochemistries, and ring conformations[43].

Glide

Glide[44] is a program accessible through the Maestro interface of the Schrödinger software package, that is used to dock ligands to proteins and assign a score based on how energetically favourable the interaction is. Glide searches for favourable interactions between one or more ligand molecules and the defined active site of a receptor molecule, by positioning and orienting a ligand relative to the receptor in search of the most favourable pose. In Glide, the selected ligands are run through a filtering algorithm that evaluates the ligand-receptor interaction that eliminates compounds that are too unfavourable while the remaining ligands are assigned an energy-minimization score and ranked accordingly[44].

BioLuminate

BioLuminate[45] is a program accessible through the Maestro interface of the Schrödinger software package, that offers tools for protein modelling, protein analysis, and protein-protein docking. BioLuminate uses Piper in protein-protein dockings and is performed as a rigid-body docking with no subsequent energy minimization. The protein-protein docking clusters the initial docking results and returns one structure pose as a cluster representative.

3.2 Methods

3.2.1 Homology modelling of trypsin

There is an increased risk of occupational airway damage in workers cultivating livestock, and the target sequences were chosen on the basis of this. Crystal structures of pig (*Sus scrofa*) and Atlantic salmon (*Salmo salar*) trypsin have already been resolved and are available via PDB. However, structures for Pacific sardine (*Sardinops caeruleus*), yellowtail (*Seriola quinqueradiata*) (also known as Japanese amberjack) and red king crab (*Paralithodes camtschaticus*) have not yet been resolved, so homology modelling was chosen as an approach to determine the theoretical 3D structure of their respective trypsin molecules. Homology modelling was preformed using the Schrödinger Maestro interface and Prime software package.

Sequence and structure alignment

Target sequences for the peptidase units of trypsin molecules were retrieved from UniProtKB[36] and MEROPS[37], belonging to Pacific sardine (https://www.ebi.ac.uk/merops/cgi-bin/speccards?sp=sp026656;type=peptidase), yellowtail (https://www.ebi.ac.uk/merops/cgi-bin/speccards?sp=sp004054;type=peptidase) and red king crab (https://www.uniprot.org/blast/?about=Q8WR10[30-266]&key=Domain).

The sequences were imported to the structure prediction wizard. An in-program BLAST Homology Search was used find appropriate homologous structures from PDB to be used as templates, and templates were selected based on sequence similarity and identity to the respective target sequence (table 1 in results section 4.1), PDB ID: 1HJ8 (chain A) for Pacific sardine and yellowtail, and PDB ID: 2F91 (chain A) for red king crab.

Alignment of target and template was conducted in-program using the ClustalW alignment option, which is suitable in cases of high sequence identity and thus appropriate based on the output from the BLAST Homology Search. No manual editing of the sequences was carried out in the construction of these homology models.

Model construction and refinement

The homology models were constructed using the energy-based method for model building, which constructs and refines residues that are not identical to template residues based on their energy.

Generated models were prepared and refined in multiple steps using the protein preparation wizard to add missing side chains, cap termini, optimize H-bond assignments at pH 7.2, and minimize the energy of structure, and loop refinements were run on shorter segments of less than 11 amino acids in length.

Quality assessment and validation of the generated structures was conducted using the online meta server SAVES, and the model structures for each target were chosen based on Verify3D score, ERRAT score and Ramachandran scores.

Docking and scoring

Docking of ligands with known affinity for trypsin was performed as an additional quality assessment of the generated trypsin homology models.

Ligands with known affinity for bovine (*Bos taurus*) trypsin was used for docking, since similarities in affinity is to be expected across species. A set of active ligands and decoys was obtained from DUD-E. The active ligands in the set were obtained from ChEMBL[46, 47], which is an online database of bioactive drug-like small molecules (ligands), containing information about their 2D structures, bioactivities, known biological targets, structure-activity relationships, and additional calculated properties[46, 47]. The obtained set of active ligands and decoys was imported into the Maestro interface and prepared with LigPrep[43], which prepares molecular structure files into chemically correct, energy minimized 3D structures ready for docking. The ligands were prepared at the physiological pH, with the parameter set to pH 7.2 +/- 0.2.

When docking ligands to a 3D model a docking grid must be defined at the expected binding site. Using Glide in the Maestro interface, the grid was defined by choosing the amino acid residues of the catalytic triad as the grid centre, and the grid size was set to 20Å in all directions to accommodate for the larger ligands.

The docking of ligands was also performed using Glide, by specifying the generated receptor grid and active ligands and decoys to be docked. The docking was conducted using standard Glide docking settings with flexible ligand sampling. When performing the docking, Glide calculates and assigns a score to each compound sampled, called GlideScore, using an algorithm that recognizes favourable interactions such as hydrophobic or hydrogen-bonding,

and penalizes unfavourable interactions such as steric clashes and electrostatic mismatches[44].

The Enrichment Calculation task in Maestro was used to assess the enrichment of the active ligands in the docking of active ligands and decoys. The output provides statistical values for common metrics used to determine the significance of the score assigned to active ligands during docking[34].

3.2.2 PAR-2 peptide segment construction

The crystal structure of the human PAR-2 receptor have been resolved by X-ray diffraction, however, the resolved structures lack the loop-region where cleavage by trypsin occurs, including the propeptide (region that is cleaved off upon PAR-2 activation) as well the tethered ligand sequence that is unmasked upon cleavage and subsequently activates the receptor. Thus, homology modelling was used to add this crucial region to the resolved structure. Homology modelling was performed using Prime in the Maestro interface.

Before constructing the PAR-2 peptide segment, consensus models of available human PAR-1 and PAR-4 structures were generated using homology modelling in order to cover as much of the peptide segment as possible. Peptide sequences for PAR-1 and PAR-4 were retrieved from UniProtKB[36]. Position 31-64 was used for PAR-1 (https://www.uniprot.org/blast/?about=P25116[22-41]&key=Propeptide&id=PRO_0000012740), and position 39-56 was used for PAR4 (https://www.uniprot.org/blast/?about=Q96RI0[18-47]&key=Propeptide&id=PRO_0000012762). Templates used for PAR-1 were retrieved from PDB entries 1NR0 (chain R), 1NRN (chain R), 1NRP (chain R), 1NRQ (chain R), 3LU9 (chain C), 3HKI (chain C), 3HKJ (chain C), and 3BEF (chain C). Templates used for PAR-4 were retrieved from PDB entries 3QDZ (chain E) and 2ZPK (chain P).

Sequence and structure alignment

The target sequence for the human PAR-2 propeptide, tethered ligand and adjacent peptide segment was retrieved from UniProtKB[36] (https://www.uniprot.org/blast/?about=P55085[26-36]&key=Propeptide&id=PRO_0000012750), and constitutes the amino acid sequence from position 26 to position 58. The corresponding peptide sequence of this region has been

resolved for PAR-1 and PAR-4 and were thus used as templates to construct this peptide region of the PAR-2 receptor from the constructed peptide consensus models

The target sequence and selected templates were imported into the structure prediction wizard, and due to low sequence similarity, the peptide sequences of PAR-1 and PAR-4 were manually aligned at the cleavage site of their respective propeptides to the corresponding sequence of PAR-2 in an attempt to predict its 3D structure. To construct a complete PAR-2 structure, a template of the remaining PAR-2 was also imported into the structure prediction wizard.

Segment construction (and preparation)

The model was constructed using the consensus model option as a method, which is an option when building a model based on multiple templates and constructs a model based on the consensus between the templates at each residue position.

Generated models were prepared and refined using the protein preparation wizard to add missing side chains, cap termini, optimize H-bond assignments at pH 7.2, and minimize the energy of structure.

The validity of the structure was assessed by visual comparison between the PAR-2 peptide segment and the corresponding PAR-1 and PAR-4 templates.

3.2.3 Protein-protein docking and interactions

Protein-protein docking

To predict the initial interaction between the selected trypsin structures and human PAR-2, protein-protein docking was performed. The trypsin structures were defined as receptors and PAR-2 was defined as the ligand. Constraints were defined in order to increase the likelihood of interaction between the propeptide cleavage site of PAR-2 and the catalytic triad of the trypsin models, and restraints between the same residues were defined in order to reject poses where these residues were too far apart. Restraints were set to the default minimum distance of 2Å between the relevant residues of the structures, and maximum distance was set to 10Å. In addition to defining residues for constraints and restraints, default parameters were used for the protein-protein docking, with number of ligand rotations set to 70 000 and maximum number of poses to return set to 32.

Protein interaction analysis

To assess the interactions between the docked PAR-2 and trypsin structures, the Protein Interaction Analysis tool from BioLuminate was used in the Maestro interface, which provides a spreadsheet list of residues from both proteins that interact or are close to residues on the other protein.

Map binding sites

To assess what parts of the trypsin binding surface that interact or are in proximity to the residues of PAR-2 upon docking, the spreadsheets from the protein interaction analysis step was used to map and align areas of the binding interaction surface of the trypsin models. The trypsin sequences of pig, Atlantic salmon, Pacific sardine, yellowtail and red king crab was done using ClustalOmega[48].

Surface analysis

To evaluate the electrostatic compatibility between the binding pocket of the different trypsin structures and the peptide segment of human PAR-2, surface analysis of the proteins was performed using the Protein Surface Analyser tool from BioLuminate in the Maestro interface.

4 Results

4.1 Alignment and homology modelling of trypsin

After retrieving the trypsin target sequences from MEROPS and UniProtKB and running them through a BLAST Homology Search, the structures listed in Table 1 were chosen as homology model templates. In all, three models were constructed.

Table 1 – Templates chosen for the construction of homology models. The table presents the PDB IDs of the resolved structures chosen for use as templates in generating homology models, their crustal structure resolution and what organism they derive from.

Target species	PDB ID template	Crystal structure resolution	Model organism
Pacific sardine (*S. caeruleus*)	1HJ8, chain A	1 Å	Atlantic salmon (*S. salar*)
Yellowtail (*S. quinqueradiata*)	1HJ8, chain A	1 Å	Atlantic salmon (*S. salar*)
Red king crab (*P. camtschaticus*)	2F91, chain A	1.2 Å	Danube crayfish (*A. leptodactylus*)

One template structure was chosen for each of the three trypsin target sequences (table 1). The same template structure was chosen for the trypsin from Pacific sardine and yellowtail: trypsin from Atlantic salmon (PDB ID: 1HJ8, chain A); while trypsin from Danube crayfish (PDB ID: 2F91, chain A) was chosen as a template structure for red king crab. Table 2 presents the homology data between target and template sequences, including the portion of identical amino acid residues, residues with similar chemical properties, and gaps in the alignments.

Table 2 – Homology data resulting from sequence alignment of target and template sequences. The table presents the degree of homology determined by the ClustalW algorithm in Schrödinger's Prime Structure Prediction Wizard between target species and the chosen templates PDB IDs amino acid sequences, 'Identities' represents the percentage of aligned amino acids that are identical between target and template. 'Positives' represents the percentage of aligned amino acids that have similar properties between target and template. 'Gaps' represent the percentage of gaps across the sequences for target and template.

Target	Template	Identities	Positives	Gaps
Sardine	1HJ8 (chain A)	88%	93%	<0.5%
Yellowtail	1HJ8 (chain A)	84%	90%	0%
Red king crab	2F91 (chain A)	65%	78%	0%

The alignment result data (table 2) shows an 88% identical amino acid sequence between the trypsin of Pacific sardine and Atlantic salmon, an additional 5% positive alignments indicating similar chemical properties between the amino acids. Figure 3 displays the sequence alignment between Pacific sardine and Atlantic salmon trypsin and shows a single gap in the Pacific sardine sequence at position 130, but accounts for less than 0.5% of the alignment. The remaining 7% of the alignment are made up of residues with mismatched chemical properties. The target sequence of trypsin from Pacific sardine has a length of 220 amino acids.
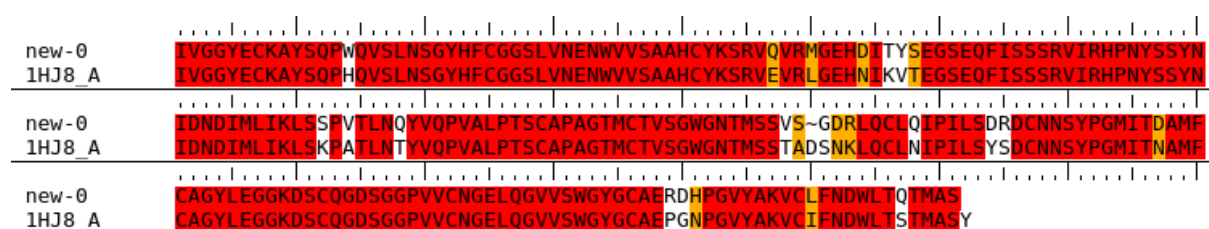
Figure 3 – Sequence alignment of sardine trypsin and salmon trypsin. The figure presents the sequence alignment between the trypsin of Pacific sardine ('new-0') and Atlantic salmon (1HJ8_A) according to the ClustalW algorithm used in Schrödinger's Prime Structure Prediction Wizard, including the identical residues marked in red, the positive matches marked in orange, while white residues indicate mismatches or gap/inserts between the sequences.

Alignment data (table 2) between yellowtail and Atlantic salmon trypsin shows that the amino acid sequences are 84% identical, an additional 6% are amino acids with similar properties, and the remaining 10% of the alignment are mismatched residues. Figure 4 displays the

sequence alignment, with indicated identities and positives, between yellowtail and Atlantic salmon trypsin. The target sequence of trypsin from yellowtail has a length of 222 amino acids.
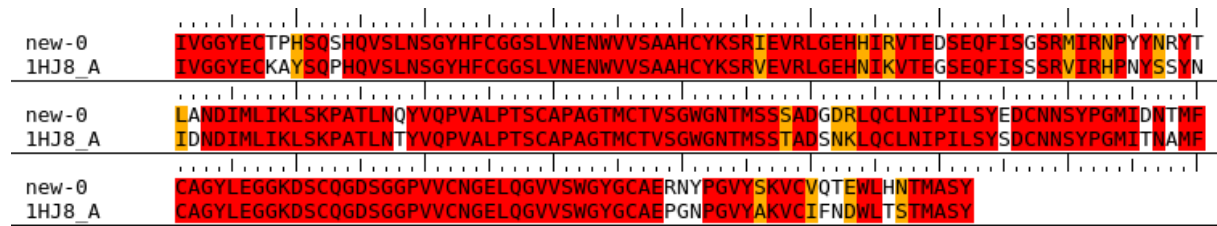
Figure 4 – Sequence alignment of yellowtail trypsin and salmon trypsin. The figure presents the sequence alignment between the trypsin of yellowtail ('new-0') and Atlantic salmon (1HJ8_A) according to the ClustalW algorithm used in Schrödinger's Prime Structure Prediction Wizard, including the identical residues marked in red, the positive matches marked in orange, while white residues indicate mismatches between the sequences.

Red king crab and Danube crayfish trypsin alignment data (table 2) shows a 65% sequence identity, an additional 13% of amino acids with similar properties, and the remaining 22% of the alignment accounts for mismatched residues. The sequence alignment between red king crab and Danube crayfish trypsin is displayed in Figure 5, with indicated identical and positive matches. The target sequence of trypsin from red king crab has a length of 237 amino acids.
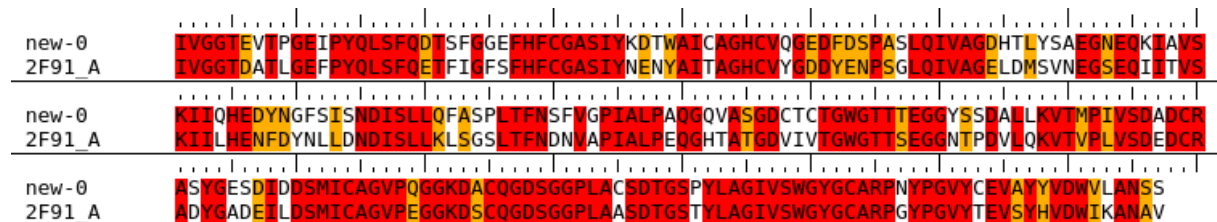
Figure 5 – Sequence alignment of red king crab trypsin and Danube crayfish trypsin. The figure presents the sequence alignment between the trypsin of red king crab ('new-0') and Danube crayfish trypsin (1HJ8_A) according to the ClustalW algorithm used in Schrödinger's Prime Structure Prediction Wizard, including the identical residues marked in red, the positive matches marked in orange, while white residues indicate mismatches between the sequences.

Structure quality assessment of the three constructed trypsin homology models was carried out using the online SAVES metaserver. The Verify3D and ERRAT scores for each homology model are presented in Table 3. The Verify3D score reveals the percentage of amino acids residues that has a 3D-1D score higher than 0.2, showing compatibility between the 3D folding of the model and the 1D amino acid sequence, and indicate a passable compatibility of all three homology models (table 3). The ERRAT score is the percentage of

amino acid residues in the protein that falls below a rejection limit for the quality factor of non-bonded atomic interactions with a cut off at 95.0000, indicating that the yellowtail and red king crab trypsin do not fall below the rejection limit, with a score of 94.8718 and 91.5094, respectively (table 3).

Table 3 – Verify3D and ERRAT scores for each trypsin homology model. Verify3D scores reveal the percentage of amino acids in the homology models that have a 3D-1D score higher than 0.2. ERRAT scores is reveal the percentage of amino acid residues that fall below a rejection limit for the quality factor of non-bonded atomic interactions.

Homology model	Verify3D score	ERRAT score
Sardine	97.27%	99.5000
Yellowtail	98.64%	94.8718
Red king crab	91.53%	91.5094

The ERRAT assessment for the homology model of sardine trypsin is displayed in Figure 6, which shows a single amino acid that is over the 95% warning zone of rejection limit for non-bonded atomic interactions, at position 102.
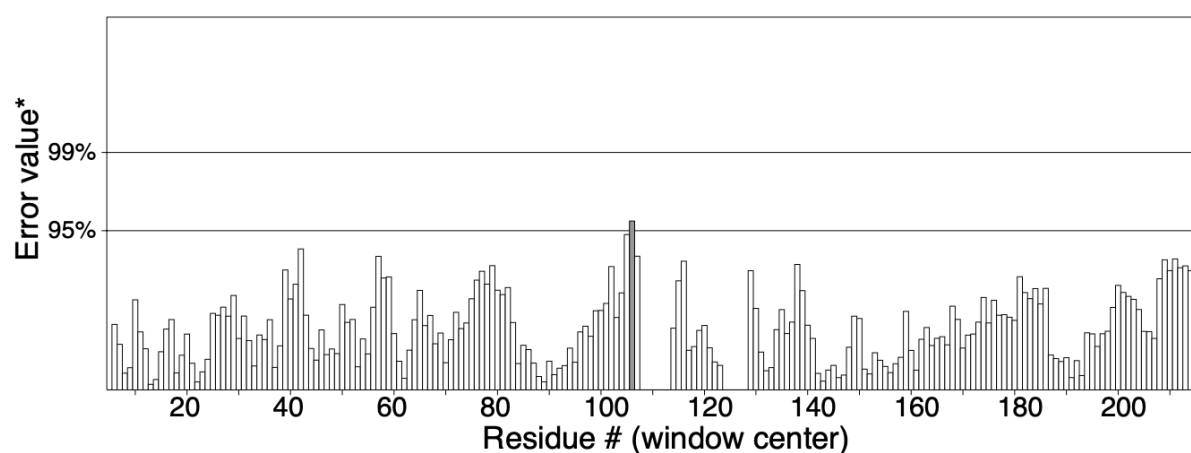
Figure 6 – ERRAT assessment of homology model of sardine trypsin. The figure presents the amino acids of sardine trypsin and where the cut-off for the 95% warning zone and 99% error zone for the quality factor of non-bonded atomic interactions lies.

The ERRAT assessment for the homology model of yellowtail trypsin can be seen in Figure 7, which displays one amino acid within the 99% error zone and a handful of amino acids within the 95% warning zone of the rejection limit for non-bonded atomic interactions. Most notably residues at positions 35-38 and 56-63 fall below the rejection limit.
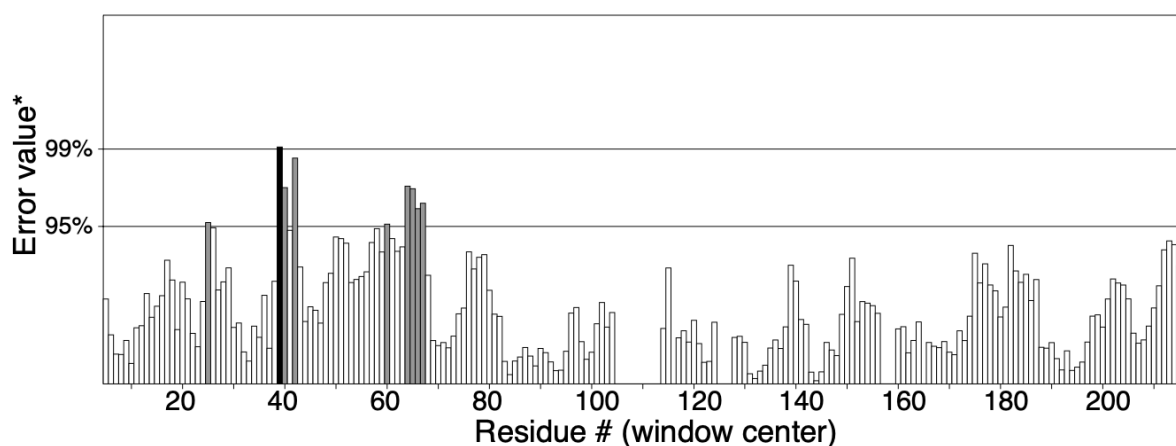
Figure 7 - ERRAT assessment of homology model of yellowtail trypsin. *The figure presents the amino acids of yellowtail trypsin and where the cut-off for the 95% warning zone and 99% error zone for the quality factor of non-bonded atomic interactions lies.*

The ERRAT assessment for the red king crab trypsin homology model is displayed in figure 6, showing a handful of amino acid residues in both the 99% error zone and the 95% warning zone of the rejection limit for non-bonded atomic interactions, most notably in positions 83-87 and 162-169.
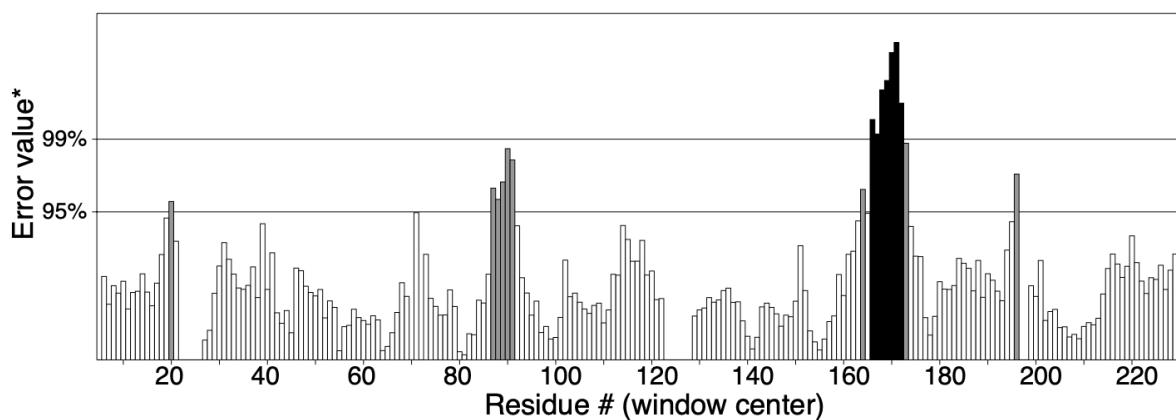
Figure 8 - ERRAT assessment of homology model of red king crab trypsin. *The figure presents the amino acids of red king crab trypsin and where the cut-off for the 95% warning zone and 99% error zone for the quality factor of non-bonded atomic interactions lies.*

Ramachandran plots (via PROCHECK) were used to assess the quality of main-chain torsion angles φ and ψ and sorts the amino acid residues into regions based on stereochemical favourability into most favoured regions, additional allowed regions, generously allowed regions and disallowed regions. Table 4 presents that a large portion of the amino acids in the constructed trypsin homology models belong in the most favoured regions and smaller

portion of the amino acids of all three models belong in the additional allowed regions. The aspartic acid at position 60 of the yellowtail trypsin model is the only amino acid residue in the generously allowed regions, while none of the structures had any residues in the disallowed regions.

Table 4 – PROCHECK Ramachandran scores for the constructed homology models. The table presents the portions of residues for each constructed trypsin homology model that lie within the stereochemistry quality regions of the Ramachandran plot, divided into 'most favoured', 'additional allowed', 'generously allowed', and 'disallowed' regions.

Homology model	Most favoured regions	Additional allowed regions	Generously allowed regions	Disallowed regions
Sardine	86.6%	13.4%	0.0%	0.0%
Yellowtail	88.7%	10.8%	0.5%	0.0%
Red king crab	84.5%	15.5%	0.0%	0.0%

4.1.1 Docking and scoring of trypsin

Docking and scoring of known active ligands and decoys of bovine trypsin were conducted to assess and compare the quality of the two resolved structures of pig trypsin and Atlantic salmon trypsin to the constructed homology models of Pacific sardine, yellowtail and red king crab.

Table 5 presents the ROC-score, which reveals whether the active ligands bind more favourably than decoys, and enrichment factor (EF), how many more actives than decoys are within a certain cut off of raked compounds, at the top 1%, 2%, 5%, 10% and 20%. The table shows that all models have a ROC-score that indicates that the active ligands have a significant binding affinity to the models, with salmon having the highest ROC-score of 0.90 and sardine the lowest of 0.85. The ER also indicates that the active compounds bind more favourably than the decoys at the strictest cut offs (1%, 2%, 5%) and declining towards more random distribution at the more generous cut offs (10%, 20%), although still not being completely random. The ROC-score of salmon is reflected in its EF cut offs, with an EF of 43 in the 1% and 25 in the 2% before it starts to even out with the others at the 10% cut off. The

other four lower EF than salmon in all cut offs, although still quite significant, and are quite even across the species.

Table 5 – Enrichment calculations for docking and scoring of trypsin models. The table presents the calculated metrics for ROC and EF. The ROC values are calculated by plotting the ranked actives and decoys against each other along each their axis and calculating the area under the curve drawn from this plotting. The EF values represent how many more actives were found in the top N% of the rankings relative to a hypothetical equal distribution.

Trypsin model	ROC	EF 1%	EF 2%	EF 5%	EF 10%	EF 20%
Pig	0.87	24	15	9.2	5.9	3.7
Salmon	0.90	43	25	12	6.6	3.9
Sardine	0.85	24	15	7.9	4.7	3.3
Yellowtail	0.86	23	13	7.8	5.2	3.5
Red king crab	0.87	25	15	8.9	5.6	3.6

Figure 9 presents the ROC-plots of the enrichment calculations of the actives and decoys docking of the five trypsin models. Salmon (figure 9B) shows a steeper curve at the highest scoring docking results, indicating that they mostly consist of highly ranking actives, while sardine (figure 9C) shows a slower incline than the others, indicating a few more ranking decoys than the rest, while the ratio of actives and decoys docked to pig (figure 9A), yellowtail (figure 9D) and red king crab (figure 9C) trypsin are relatively similar to each other.
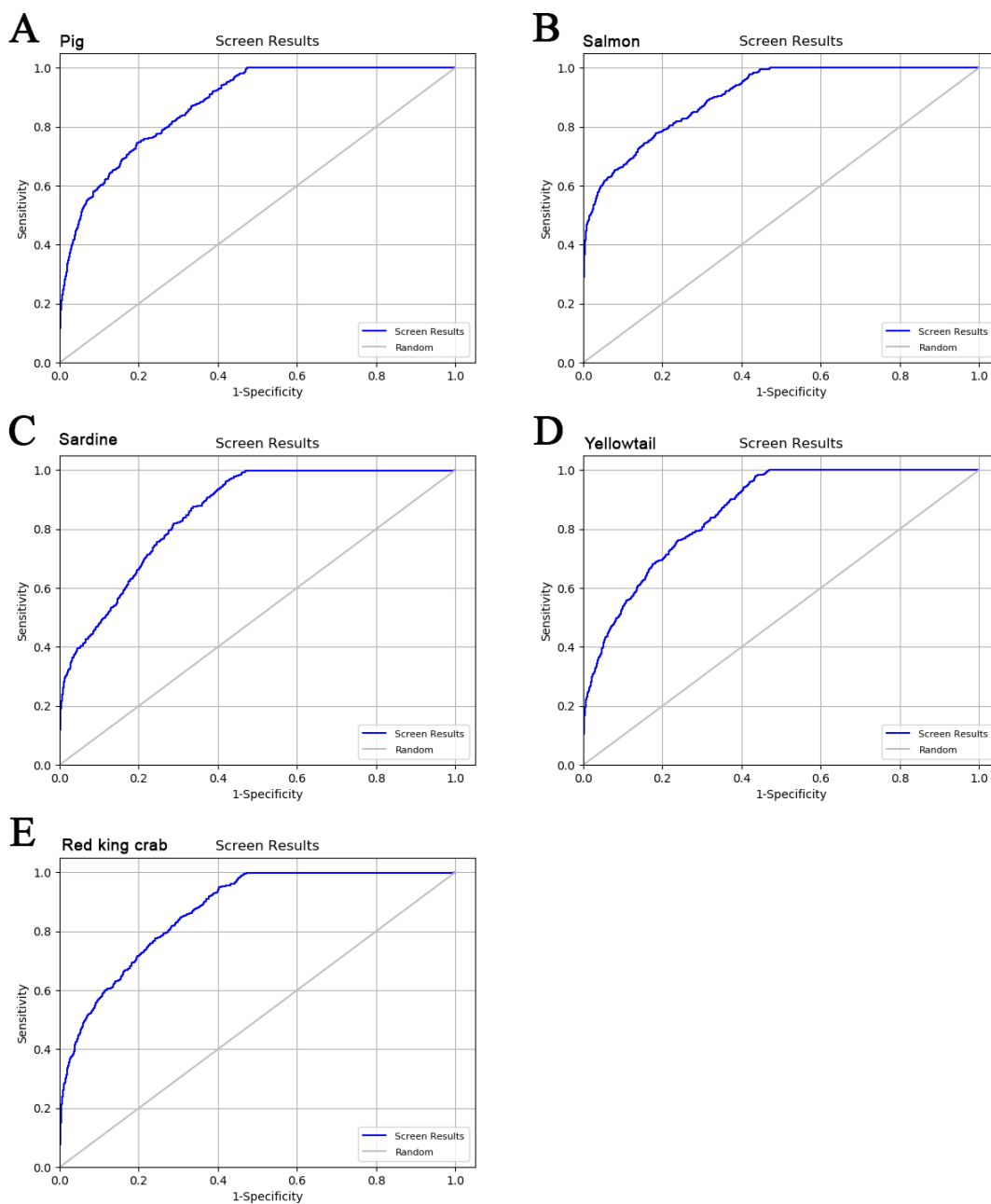
Figure 9 – ROC plots resulting from the enrichment calculations after docking and scoring known active ligands and decoys to the trypsin models. *The ROC-plot plots the ranking of actives and decoys against each other. Actives are ranked along the Y-axis (sensitivity) and decoys along the X-axis (1-specificity). The blue line is the plotted curve resulting from each ranked active compound. 9A is the ROC-plot for the docking of pig trypsin, 9B is the plot for salmon trypsin, 9C is sardine trypsin, 9D is yellowtail trypsin, and 9E is the ROC-plot for red king crab trypsin.*

4.2 Alignment and homology modelling of peptide segment of PAR-2

Homology modelling was used to construct a human PAR-2 model containing the N-terminal peptide segment where the cleavage site for unmasking the tethered ligand that subsequently activates the receptor is located, using constructed consensus models of corresponding N-terminal peptide segments from human PAR-1 and PAR-4 as templates.

Figure 10 presents the alignment of PAR-2 to PAR-1 and PAR-4 with identical residues and positives highlighted and shows low homology between the PARs. The already resolved human PAR-2 structure starts at residue position 34 (FSVD…) in the alignment in figure 10, while the cleavage site for all three structures are between positions 11 and 12 (RS). The PAR-1 sequence shows slightly more residue matches and larger sequence coverage than PAR-4 sequence.

Figure 10 – Sequence alignment of PAR-2, PAR-1 and PAR-4. The figure presents the sequence alignment of PAR-2 ('new-0') to the consensus homology models of PAR-1 ('consensus-PAR-1') and PAR-4 ('consensus-PAR-4'). The identical residues are marked in red, the positive matches marked in orange, while white residues indicate mismatches between the sequences.

Figure 11 presents the folding of the N-terminal peptide segments of all three PAR-models with the cleavage site residues displayed. The folding of the PAR-2 N-terminal peptide (figure 11C) shows similar folding to the PAR-1 structure (figure 11A).
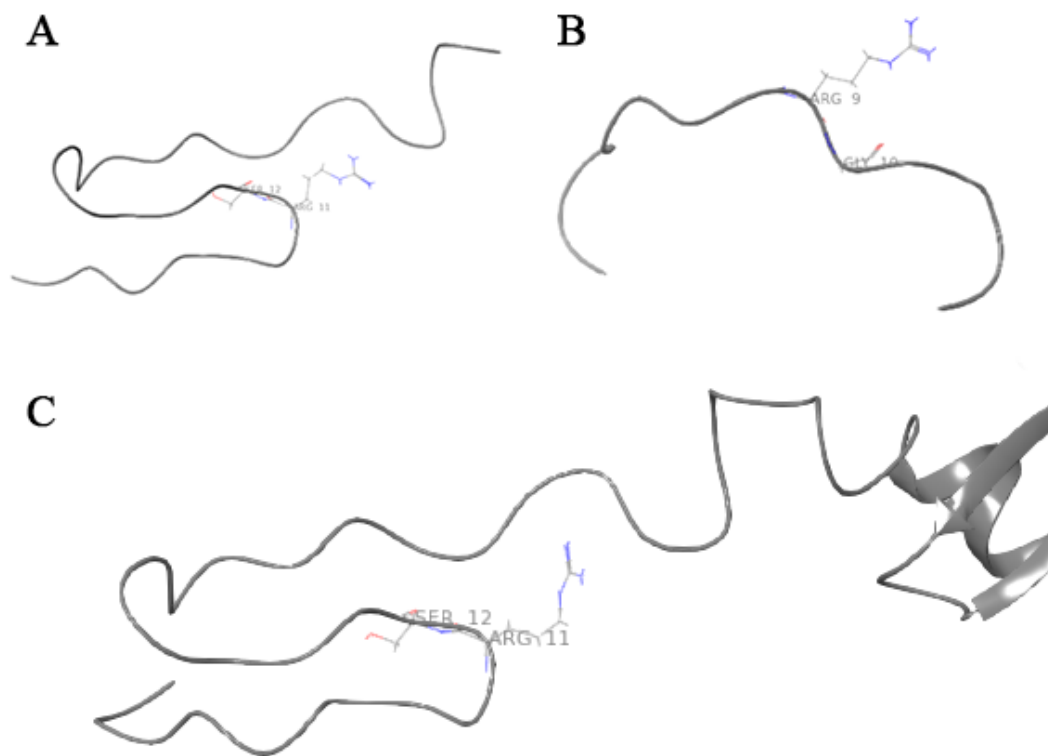
Figure 11 – N-terminal peptide segments of PARs. *The figure presents the constructed peptide segments of the N-terminal peptide segments of PAR-1 (11A), PAR-4 (11B), and PAR-3 (11C). The arginine and serine residues where proteolytic cleavage occurs are marked.*

4.3 Protein-protein docking and interactions of PAR-2 and trypsin models

Docking of the constructed N-terminal peptide segment of PAR-2 to the various trypsin structures was performed in Schrödinger's Maestro interface using the Protein-Protein Docking tool from BioLuminate software package, with the intention of predicting the initial interaction between the two proteins.

Tables 6-10 presents the docking results of the three most observed poses of the PAR-2 docking to pig, Atlantic salmon, Pacific sardine, yellowtail and red king crab trypsin, respectively. The docking results present PIPER pose energy which represent the interaction energy between the two proteins derived from repulsive and attractive interactions based on van der Waal interaction energy, and electrostatic energy. PIPER cluster size is a ranking

system that reveals how many times similar docking poses were observed during the docking job and is the recommended ranking method for protein-protein docking results[35].

Table 6 – Protein-protein docking output for PAR-2 docking to pig trypsin. *The table presents the three most observed posed, ranked by cluster size, and their piper pose energy.*

Pose	PIPER pose energy	PIPER cluster size
1	-795.237	153
2	-680.316	118
3	-657.817	100

Table 7 - Protein-protein docking output for PAR-2 docking to salmon trypsin. *The table presents the three most observed posed, ranked by cluster size, and their piper pose energy.*

Pose	PIPER pose energy	PIPER cluster size
1	-768.314	180
2	-539.955	127
3	-682.646	115

Table 8 - Protein-protein docking output for PAR-2 docking to sardine trypsin. *The table presents the three most observed posed, ranked by cluster size, and their piper pose energy.*

Pose	PIPER pose energy	PIPER cluster size
1	-810.674	165
2	-763.537	133
3	-629.261	126

Table 9 - Protein-protein docking output for PAR-2 docking to yellowtail trypsin. *The table presents the three most observed posed, ranked by cluster size, and their piper pose energy.*

Pose	PIPER pose energy	PIPER cluster size
1	-740.659	161
2	-608.187	134
3	-574.830	113

Table 10 - Protein-protein docking output for PAR-2 docking to red king crab trypsin. *The table presents the three most observed posed, ranked by cluster size, and their piper pose energy.*

Pose	PIPER pose energy	PIPER cluster size
1	-643.491	116
2	-661.109	107
3	-775.659	100

4.3.1 Protein-protein interactions

To study the interactions resulting from the protein-protein docking, the Protein Interaction Analysis tool from Schrödinger's BioLuminate software package was used. Tables x-x presents the number of hydrogen bonds (H-bonds), salt bridges, pi stackings and van der Waal clashes involved in each protein-protein interaction from the top three poses resulting from each protein-protein docking, as well as the specific amino acid resides involved in each H-bond, salt bridge and pi stacking interaction. H-bonds are electrostatic interactions between a covalently bound hydrogen and a hydrogen bond acceptor with an available electron pair, salt bridges are a combination of H-bonding and ionic bonding between a negatively charged atom and a positively charged atom, Pi stacking is an interaction between two aromatic rings within close proximity to each other. Van der Waal clashes are an overlap of the van der Waals radii of two atoms by a specific cut-off.

Only the interactions from the N-terminal peptide segment of PAR-2 to the respective trypsin structure is taken into account, and that interactions from other regions of PAR-2 was observed in some of the docking poses.

Table 11 shows a summary of the interactions for the top three poses of the N-terminal PAR-2 docking to trypsin derived from pig, while tables 12-14 presents the amino acid residues involved in each specific interaction (not including the van der Waal clashes) of the top three poses, respectively. Pose 1 had two H-bond interactions (tables 11, 12), Pose 2 had 4 H-bond interactions and one Pi stacking (tables 11, 13), and Pose 3 had seven H-bond interactions (tables 11, 14).

*Table 11 – **Summary of specific interactions in the top three poses of PAR-2 docking to pig trypsin**. The table shows the number of specific interactions observed in a pose, including hydrogen bonds, salt bridges, pi stacking and Van der Waal clashes.*

Pose	H-bonds	Salt bridges	Pi stacking	Van der Waal clashes
1	2	0	0	4
2	4	0	1	2
3	7	0	0	1

*Table 12 – **Specific residue interactions in PAR-2 docking to pig trypsin**. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Pig trypsin residue	Interaction
1	Ser8	His57	H-bond
1	Lys9	Gln192	H-bond

Table 13 - Specific residue interactions in PAR-2 docking to pig trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Pig trypsin residue	Interaction
2	Arg11	His57	H-bond
2	Thr24	Gly96	H-bond
2	Lys26	Thr90	H-bond
2	Glu31	Thr90	H-bond
2	Phe24	His91	Pi stacking

Table 14 - Specific residue interactions in PAR-2 docking to pig trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Pig trypsin residue	Interaction
3	Arg6	Trp141	2x H-bond
3	Lys9	Gly216	H-bond
3	Gly10	Ser214	H-bond
3	Arg11	Tyr217	H-bond
3	Gly15	Lys60	H-bond
3	Lys26	Tyr217	H-bond

Table 15 presents a summary of all interactions observed in the top three poses from the docking of the N-terminal PAR-2 peptide to trypsin from Atlantic salmon, while tables 16-18 presents the residues involved in the specific interactions (not including van der Waal clashes). Pose 1 had five H-bond and one salt bridge interaction (tables 15, 16), Pose 2 had two H-bond and two salt bridge interactions (tables 15, 17), and Pose 3 had four H-bond and one salt bridge interaction (tables 15, 18).

Table 15 - Summary of specific interactions in the top three poses of PAR-2 docking to salmon trypsin. *The table shows the number of specific interactions observed in a pose, including hydrogen bonds, salt bridges, pi stacking and Van der Waal clashes.*

Pose	H-bonds	Salt bridges	Pi stacking	Van der Waal clashes
1	5	1	0	2
2	2	2	0	4
3	4	1	0	2

Table 16 - Specific residue interactions in PAR-2 docking to salmon trypsin. *The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Salmon trypsin residue	Interaction
1	Ser8	Gln192	H-bond
1	Gly15	Gln192	H-bond
1	Lys16	Ser152	H-bond
1	Lys26	Ser96	H-bond
1	Glu31	Arg90	H-bond, Salt bridge

Table 17 - Specific residue interactions in PAR-2 docking to salmon trypsin. *The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Salmon trypsin residue	Interaction
2	Arg11	Glu221	Salt bridge
2	Gly15	Tyr97	H-bond
2	Lys26	Glu221	Salt bridge
2	Glu38	Ser147	H-bond

Table 18 - Specific residue interactions in PAR-2 docking to salmon trypsin*. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Salmon trypsin residue	Interaction
3	Arg6	Trp141	H-bond
3	Arg6	Gln192	H-bond
3	Ser8	Gly193	H-bond
3	Lys16	Glu49	Salt bridge
3	Lys16	Ser61	H-bond

Table 19 presents the summary of specific interactions between the N-terminal PAR-2 peptide and the homology model of Pacific sardine trypsin of the top three poses from the protein-protein docking, while tables 20-22 show the specific amino acid residues involved in each interaction (not including van der Waal clashes). Pose 1 had six H-bond interactions (tables 19, 20), Pose 2 had five H-bond interactions and two salt bridges (tables 19, 21), and Pose 3 had five H-bond interactions and one salt bridge (tables 19, 22).

Table 19 - Summary of specific interactions in the top three poses of PAR-2 docking to sardine trypsin*. The table shows the number of specific interactions observed in a pose, including hydrogen bonds, salt bridges, pi stacking and Van der Waal clashes.*

Pose	H-bonds	Salt bridges	Pi stacking	Van der Waal clashes
1	6	0	0	3
2	5	2	0	5
3	5	1	0	2

Table 20 - Specific residue interactions in PAR-2 docking to sardine trypsin. *The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Sardine trypsin residue	Interaction
1	Arg6	Tyr22	H-bond
1	Arg6	Trp121	H-bond
1	Arg6	Gly122	H-bond
1	Ser8	Ser175	H-bond
1	Ser12	His40	H-bond
1	Gly15	Lys43	H-bond

Table 21 - Specific residue interactions in PAR-2 docking to sardine trypsin. *The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Sardine trypsin residue	Interaction
2	Arg6	Glu197	H-bond, Salt bridge
2	Gly10	Ser190	H-bond
2	Lys16	Asp131	H-bond, Salt bridge
2	Gly27	Arg72	2x H-bond

Pose	PAR-2 residue	Sardine trypsin residue	Interaction
3	Arg6	Met125	H-bond
3	Lys9	Gln172	H-bond
3	Lys9	Glu197	Salt bridge
3	Gly10	Gly194	H-bond
3	Lys16	Tyr22	H-bond
3	Lys26	Tyr79	H-bond

Table 23 presents a summary of the interactions from the top three poses of the protein-protein docking between the N-terminal PAR-2 peptide and the homology model of yellowtail trypsin, while tables 24-26 shows the specific amino acid residues involved in the specific interactions (not including van der Waal clashes). Pose 1 had four H-bond interactions and one salt bridge (tables 23, 24), Pose 2 had seven H-bond interactions and one salt bridge (tables 23, 25), and Pose 3 had eight H-bond interactions and one salt bridge (tables 23, 26).

Table 23 - Summary of specific interactions in the top three poses of PAR-2 docking to yellowtail trypsin. *The table shows the number of specific interactions observed in a pose, including hydrogen bonds, salt bridges, pi stacking and Van der Waal clashes.*

Pose	H-bonds	Salt bridges	Pi stacking	Van der Waal clashes
1	4	1	0	3
2	7	1	0	2
3	8	1	0	4

Table 24 - Specific residue interactions in PAR-2 docking to yellowtail trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Yellowtail trypsin residue	Interaction
1	Arg6	Glu198	H-bond
1	Ser8	Gly195	H-bond
1	Arg11	His40	H-bond
1	Glu31	Tyr42	H-bond, Salt bridge

Table 25 - Specific residue interactions in PAR-2 docking to yellowtail trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Yellowtail trypsin residue	Interaction
2	Arg11	Glu198	H-bond, Salt bridge
2	Ser12	Gly193	H-bond
2	Gly15	Tyr79	H-bond
2	Lys16	Thr80	H-bond
2	Gly19	Arg78	H-bond
2	Lys26	Glu198	H-bond
2	Glu38	Ser127	H-bond

Table 26 - Specific residue interactions in PAR-2 docking to yellowtail trypsin. *The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.*

Pose	PAR-2 residue	Yellowtail trypsin residue	Interaction
3	Arg6	Ser127	H-bond
3	Lys9	Glu198	H-bond, Salt bridge
3	Gly10	Ser176	H-bond
3	Leu13	Gln173	H-bond
3	Val33	Asn77	H-bond
3	Ser35	Tyr79	2x H-bonds

Table 27 presents a summary of the protein-protein docking interactions between the N-terminal PAR-2 peptide and the homology model of red king crab trypsin, while tables 28-30 presents the specific amino acid residues involved in each interaction (not including van der Waal clashes). Pose 1 had seven H-bond interactions and three salt bridges (tables 27, 28), Pose 2 had four H-bond interactions and three salt bridges (tables 27, 29), and Pose 3 had eight H-bond interactions and three salt bridges (tables 27, 30).

Table 27 - Summary of specific interactions in the top three poses of PAR-2 docking to red king crab trypsin. *The table shows the number of specific interactions observed in a pose, including hydrogen bonds, salt bridges, pi stacking and Van der Waal clashes.*

Pose	H-bonds	Salt bridges	Pi stacking	Van der Waal clashes
1	7	4	1	8
2	4	3	0	3
3	8	3	0	7

Table 28 - Specific residue interactions in PAR-2 docking to red king crab trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Red king crab trypsin residue	Interaction
1	Lys9	Asp167	H-bond, Salt bridge
1	Arg11	Asp169	Salt bridge
1	Gly15	Gln186	H-bond
1	Lys16	Asp20	H-bond, Salt bridge
1	His22	Phe91	Pi stacking
1	Lys26	Ser166	H-bond
1	Lys26	Asp169	H-bond, Salt bridge

Table 29 - Specific residue interactions in PAR-2 docking to red king crab trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Red king crab trypsin residue	Interaction
2	Arg6	His28	H-bond
2	Ser8	Ser189	H-bond
2	Arg11	Asp167	H-bond, Salt bridge
2	Gly15	Gln48	H-bond
2	Lys16	Glu50	Salt bridge
2	Lys26	Asp167	Salt bridge

Table 30 - Specific residue interactions in PAR-2 docking to red king crab trypsin. The table presents the two specific amino acid resides involved in the interaction and what kind of interaction it is.

Pose	PAR-2 residue	Red king crab trypsin residue	Interaction
3	Asn5	Gly90	H-bond
3	Arg6	Glu86	2x H-bond, Salt bridge
3	Ser8	Gln48	H-bond
3	Lys9	Glu50	Salt bridge
3	Gly10	Gly187	H-bond
3	Lys16	Tyr163	H-bond
3	Lys16	Asp167	H-bond, Salt bridge
3	Val36	Tyr68	H-bond

4.3.2 Mapping protein binding site

To assess which residues are involved in the initial binding interaction, the top three poses from each protein-protein docking result were studied.

Table 31 presents the amino acid residues of the PAR-2 receptor peptide segment observed in 7 or more of the 15 poses. The table presents five amino acid residues with positive charge, two residues with negative charge, seven uncharged polar residues, five hydrophobic residues, and four residues with other properties involved in the binding site.

Table 31 – Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. *The table presents amino acids observed 7 or more times across the binding poses in the protein-protein docking interactions. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, and black residues indicate neutral and miscellaneous residues. Please note that the residues listed do not correspond to the adjacent column but are listed that way for convenience.*

Residue and position	Residue and position, contd.
Thr4	Thr24
Asn5	Gly25
Arg6	Lys26
Ser7	Gly27
Ser8	Val28
Lys9	
Gly10	Glu31
Arg11	
Ser12	Val33
Leu13	Phe34
Ile14	Ser35
Gly15	
Lys16	Glu38

Tables 32-39 presents amino acid residues predicted to be involved in the binding site of the five trypsin structures. The residues were selected from regions that were observed to be part of the binding site across species in at least one of the top three poses of each N-terminal PAR-2 peptide and trypsin protein-protein docking job, and that corresponded according to a ClustalOmega multiple sequence alignment of the five trypsin sequences, presented in Appendix A.

Please note that the numbering of amino acid residues differs between the species, the pig and Atlantic salmon had pre-numbered residues from the PDB crystal structure that take into

account the propeptide not featured in the active trypsin structures, while also having some inconsistent numbering; and the homology models are numbered according to the part of the trypsin sequence used as a target, starting at 1. The presented residues are colour coded according to their properties, and residues that were not observed in the binding interactions of a particular species but were observed in others are also listed but not considered.

Table 32 presents the first region of interest, which seems generally well conserved with the exception of the first two-three residues of red king crab trypsin, which most notably has a glycine where the others has a uncharged polar serine (top row), a negatively charged glutamic acid where the others have a glycine (second row), and pig trypsin has a glycine residue where the others have a hydrophobic residue.

Table 32 – Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, and cursive grey residues represent residues that were not observed in any of the poses. Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Ser37	Ser37	Ser20	*Ser20*	Gly25
Gly38	*Gly38*	Gly21	Gly21	Glu26
Ser39	Tyr39	Tyr22	Tyr22	Phe27
His40	His40	His23	His23	His28
Phe41	Phe41	Phe24	Phe24	Phe29
Cys42	Cys42	Cys25	Cys25	Cys30

Table 33 presents the second region of interest, which seems to be well conserved between pig, Atlantic salmon, Pacific sardine and yellowtail trypsin, while red king crab seems less conserved compared to the others. The table also shows that red king crab has a region of insertions compared to the four other species, containing two negatively charged residue and a

hydrophobic phenylalanine residue (row 7-9). Red king crab also has a polar uncharged glutamine residue where the others have a positively charged lysine residue (row 5), and a glycine where the others have an uncharged polar serine residue (row 6).

Please note that there are more residues located between the red king crab residues listed in the last and second to last row of the table but were not of interest since they were not observed as being part of the binding interactions.

Table 33 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, cursive grey residues represent residues that were not observed in any of the poses, and dashes indicate gaps in sequence relative to other sequence(s). Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Ala56	Ala56	Ala39	Ala39	*Gly44*
His57	His57	His40	His40	His45
Cys58	Cys58	Cys41	Cys41	Cys46
Tyr59	Tyr59	Tyr42	Tyr42	Val47
Lys60	Lys60	Lys43	Lys43	Gln48
Ser61	Ser61	Ser44	Ser44	Gly49
-	-	-	-	Glu50
-	-	-	-	Asp51
-	-	-	-	Phe52
Arg62	Arg62	*Arg45*	*Arg45*	*Ser57*

Table 34 presents the third region of interest, which seems less conserved across the species. Most notably, red king crab and pig trypsin have polar uncharged residues where the other three species have a positively charged arginine residue (row 4), yellowtail trypsin has a polar

uncharged asparagine where the others have a positively charged histidine (row 5), red king crab has a negatively charged glutamic acid where the others have a proline residue (row 6), yellowtail has a hydrophobic tyrosine residue where pig and salmon have uncharged polar residues and the residues of sardine and red king crab are not involved in the binding site (row 7), yellowtail has a positively charged arginine where sardine and salmon trypsin have an uncharged polar serine residue and red king crab and pig trypsin have a glycine residue (row 10), pig trypsin has a polar uncharged asparagine residue where the others have hydrophobic residues (row 11), and salmon and sardine have negatively charged aspartic acid residues where yellowtail has a hydrophobic alanine residue and red king crab has a polar uncharged serine residue (row 12). The bottom row lists the aspartic acid involved in the catalytic triad of trypsin.

Table 34 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, and cursive grey residues represent residues that were not observed in any of the poses. Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Lys87	Arg87	Arg69	Arg69	*Lys81*
Ile88	Val88	Val70	Met70	*Ile82*
Ile89	Ile89	Ile71	Ile71	*Ile83*
Thr90	Arg90	Arg72	Arg72	Gln84
His91	His91	His73	Asn73	His85
Pro92	Pro92	Pro74	Pro74	Glu86
Asn93	Asn93	*Asn75*	Tyr75	*Asp87*
Phe94	Tyr94	Tyr76	Tyr76	Tyr88
Asn95	Ser95	Ser77	Asn77	Asn89

Gly96	Ser96	Ser78	Arg78	Gly90
Asn97	Tyr97	Tyr79	Tyr79	Phe91
Thr98	Asn98	Asn80	Thr80	Ser92
Leu99	Ile99	Ile81	Leu81	Ile93
Asp100	Asp100	Asp82	Ala82	Ser94
Asn101	*Asn101*	*Asn83*	Asn83	Asn95
Asp102	Asp102	Asp84	*Asp84*	*Asp96*

Table 35 shows the fourth region of interest, which show that the first five surface amino acids are well conserved between the species, followed by approximately eight less conserved surface amino acids which includes some gaps between the species. The most notable variations are the polar uncharged serine residue of salmon, sardine and yellowtail which is a negatively charged glutamic acid in the corresponding position of red king crab (row 6), the subsequent glycine of pig and red king crab correspond to a polar uncharged threonine in salmon and serine in yellowtail (row 8), sardine has a gap that corresponds to a polar uncharged serine in pig an hydrophobic alanine and tyrosine in salmon and red king crab respectively (row 9), sardine has an insertion of an uncharged polar serine residue that corresponds to a gap in the other four species (row 10), the hydrophobic tyrosine residue in pig corresponds to a negatively charged aspartic acid residue in salmon and yellowtail and a polar uncharged serine residue in sardine and red king crab (row 11), pig trypsin has a proline residue that corresponds to a glycine residue in sardine and yellowtail and a polar uncharged serine residue in salmon and red king crab (row 12), and finally sardine, yellowtail and red king crab has a negatively charged aspartic acid residue that corresponds to a polar uncharged serine residue in pig trypsin and a polar uncharged asparagine residue in salmon trypsin.

Table 35 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. *The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, cursive grey residues represent residues that were not observed in any of the poses, and dashes indicate gaps in sequence relative to other sequence(s). Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.*

Pig	Salmon	Sardine	Yellowtail	Red king crab
Trp141	Trp141	Trp121	Trp121	Trp133
Gly142	Gly142	Gly122	Gly122	Gly134
Asn143	Asn143	Asn123	Asn123	Thr135
Thr144	Thr144	Thr124	Thr124	*Thr136*
Lys145	Met145	Met125	A Met125	*Thr136*
Ser146	Ser146	Ser126	Ser126	Glu138
Ser146	Ser147	Ser127	Ser127	*Gly139*
Gly148	Thr148	*Val128*	Ser128	Gly140
Ser149	Ala149	-	*Ala129*	Tyr141
Ser150	-	-	-	-
Tyr151	Asp150	Ser129	Asp130	Ser142
Pro152	Ser152	Gly130	Gly131	Ser143
Ser153	Asn153	Asp131	Asp132	Asp144

Table 36 presents the fifth region of interest, which show less conservation of surface residues in trypsin across the species, particularly for red king crab. Pig trypsin has a positively charged lysine residue that corresponds to a positively charged arginine residue in red king crab and polar uncharged asparagine in salmon and sardine (row 1), a polar uncharged serine residue in pig trypsin that corresponds to a polar uncharged asparagine residue in both salmon and sardine and a hydrophobic alanine in red king crab (row 2), followed by two fully

conserved residues across the species (row 3 and 4), a glycine residue in red king crab that corresponds to a proline residue in the four other species (row 5), red king crab has a negatively charged glutamic acid that corresponds to glycine in the four other species (row 6), red king crab also has an insertion of a polar uncharged serine residue that corresponds to alignment gaps in the four other species (row 7), a polar uncharged glutamine residue in pig trypsin corresponds to a hydrophobic methionine residue in salmon, sardine and yellowtail trypsin and to a negatively charged aspartic acid in red king crab (row 8), followed by an isoleucine that is conserved across all five species (row 9), a polar uncharged threonine residue in pig, salmon and sardine trypsin that corresponds to a negatively charged aspartic acid in both yellowtail and red king crab trypsin (row 10), sardine and red king crab trypsin has a negatively charged aspartic acid residue that corresponds to a polar uncharged asparagine residue in yellowtail (row 11), and finally a threonine residue in yellowtail that corresponds to a polar uncharged serine residue in red king crab trypsin (row 12).

Table 36 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, cursive grey residues represent residues that were not observed in any of the poses, and dashes indicate gaps in sequence relative to other sequence(s). Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Lys169	Asn169	Asn147	*Asn149*	Arg160
Ser170	Asn170	Asn148	*Asn149*	Ala161
Ser171	Ser171	Ser149	*Ser150*	Ser162
Tyr172	Tyr172	Tyr150	*Tyr151*	Tyr163
Pro173	Pro173	Pro151	Pro152	Gly164
Gly174	Gly174	Gly152	Gly153	Glu165
-	-	-	-	Ser166
Gln175	Met175	Met153	Met154	Asp167

Ile176	Ile176	*Ile154*	*Ile155*	Ile168
Thr177	Thr177	Thr155	Asp156	Asp169
Gly178	*Asn178*	Asp156	Asn157	Asp170
Asn179	*Ala179*	*Ala179*	Thr158	Ser171

Table 37 presents the seventh region of interest, which shows that all the amino acid residues that are observed exposed at the surface are conserved across the five species. Note that the serine residue involved in the catalytic triad is present this segment (row 8), as well as the aspartic acid involved in substrate binding (row 3).

*Table 37 - **Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models**. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, and cursive grey residues represent residues that were not observed in any of the poses. Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.*

Pig	**Salmon**	**Sardine**	**Yellowtail**	**Red king crab**
Gly187	Gly188A	*Gly167*	Gly168	*Gly181*
Lys188	Lys188	*Lys168*	Lys169	*Lys182*
Asp189	*Asp189*	*Asp169*	*Asp170*	*Asp183*
Ser190	*Ser190*	Ser170	Ser171	*Ala184*
Cys191	Cys191	Cys171	Cys172	*Cys185*
Gln192	Gln192	Gln172	Gln173	Gln186
Gly193	Gly193	Gly173	Gly174	Gly187
Asp194	*Asp194*	*Asp174*	*Asp175*	*Asp188*
Ser195	Ser195	Ser175	Ser176	Ser189

Table 38 presents the seventh region of interest, which also shows a high degree of conservation, but with some less conserved residues in the bottom half of the table. Residues

in rows 1-8 and 14 are completely conserved across the species, while a positively charged arginine residue in red king crab trypsin corresponds to a negatively charged glutamic acid in salmon, sardine and yellowtail (row 9), the proline in salmon corresponds to a positively charged arginine in yellowtail (row 10), and a positively charged lysine in pig trypsin corresponds to a polar uncharged asparagine residue in salmon, a positively charged histidine in sardine, and a hydrophobic tyrosine residue in yellowtail and red king crab trypsin (row 12).

Table 38 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, black residues indicate neutral and miscellaneous residues, and cursive grey residues represent residues that were not observed in any of the poses. Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Val213	Val213	Val189	Val189	*Val207*
Ser214	Ser214	Ser190	Ser191	Ser208
Trp215	Trp215	Trp191	Trp192	Trp209
Gly216	Gly216	Gly192	Gly193	Gly210
Tyr217	Tyr217	Tyr193	Tyr194	Tyr211
Gly219	Gly219	Gly194	Gly195	Gly212
Cys220	Cys220	Cys195	Cys196	Cys213
Ala221A	Ala221A	Ala196	Ala197	*Ala212*
Gln221	Glu221	Glu197	Glu198	Arg215
Lys222	Pro222	*Arg198*	Arg199	*Pro216*
Asn223	*Gly223*	*Asp199*	*Asn200*	*Asn217*
Lys224	Asn224	His200	Tyr201	Tyr218

Pro225	*Pro225*	*Pro201*	*Pro202*	*Pro218*
Gly226	*Gly226*	*Gly202*	Gly203	*Gly219*

Table 39 shows three other but significantly smaller regions where amino acids from more than one species were exposed at the surface of the binding interaction site and have a lower degree of conservation. Rows 1-2 show one region where the first amino acid residue is hydrophobic across all species, but is a leucine residue in red king crab trypsin whereas an isoleucine in pig, salmon and sardine trypsin, followed by a negatively charged aspartic acid in pig trypsin that corresponds to a polar uncharged threonine in sardine, a positively charged arginine in yellowtail and a hydrophobic tyrosine in red king crab. Row 4 shows a negatively charged aspartic acid in red king crab trypsin that corresponds to a hydrophobic tyrosine in salmon and yellowtail trypsin. Finally, row 6 shows a hydrophobic phenylalanine in salmon trypsin that corresponds to a hydrophobic tyrosine in red king crab trypsin.

Table 39 - Amino acid residues observed in the binding of PAR-2 peptide segment to trypsin models. The table present residues found in the protein-protein docking poses from all five trypsin models and their properties that correspond across species according to alignment. Purple residues represent polar, uncharged residues, blue residues represent positively charged residues, green residues represent hydrophobic residues, red residues indicate negatively charged residues, and cursive grey residues represent residues that were not observed in any of the poses. Please note that the numbering assigned to the constructed homology models (sardine, yellowtail and red king crab) are not "official" numberings, but rather the number assigned to the residue upon model construction based on their amino acid position in the target structure.

Pig	Salmon	Sardine	Yellowtail	Red king crab
Ile73	Ile73	Ile55	*Ile55*	Leu67
Asp74	*Lys74*	Thr56	Arg56	Tyr68
Asp165	Tyr165	*Asp143*	Tyr144	Asp156
Tyr234	Phe234	*Phe210*	*Gln211*	Tyr228

4.3.3 Surface analysis of protein structures

To get an idea of how well the trypsin from the different species attracts to human PAR-2, the surfaces of the trypsin structures were assessed in Maestro using the Protein Surface Analyser tool from the BioLuminate package. Tables 40 and 41 presents the overall properties of the five individual structures. Table 40 presents the sum of positively and negatively charged surface area, and the sum of donor and acceptor surface areas, all in square Ångstroms ($Å^2$). The table results show that all trypsin structures have a larger positive surface area than negative surface area, although this difference appears to be largest in the pig trypsin structure and decreases in order of the table to red king crab, where the difference is smaller. The acceptor surface area is larger than the donor surface area in all five trypsin structures, but the difference is largest in red king crab, followed by Pacific sardine, Atlantic salmon and yellowtail, and smallest in pig trypsin.

Table 40 – Properties of the surface area of trypsin models. The table presents the sum positive surface area, sum negative surface area, sum donor surface area and sum acceptor surface area in square Ångstroms ($Å^2$) for all trypsin models.

Trypsin model	Sum positive surface area ($Å^2$)	Sum negative surface area ($Å^2$)	Sum donor surface area ($Å^2$)	Sum acceptor surface area ($Å^2$)
Pig	4873.43	3047.14	1876.14	2017.64
Salmon	4788.70	3187.31	1724.86	2081.01
Sardine	4916.54	3345.07	1699.86	2165.69
Yellowtail	4839.52	3308.80	1790.85	2099.92
Red king crab	5006.95	3586.07	1399.56	2411.35

Table 41 presents the structures hydrophobic moment, molecular weight in kilodalton (kDa) and charge in electronvolt (eV). The table shows that the hydrophobic moment, a value derived from hydrophobic measures of surface amino acid residues, is largest in pig trypsin and red king crab, followed by Atlantic salmon, Pacific sardine and yellowtail. The table also shows that the molecular weight is lowest in trypsin derived from pig and increases in order of the table, to the red king crab trypsin homology model which has the highest molecular

weight. The table also shows that red king crab trypsin has the most negative charge, that trypsin from Pacific sardine, yellowtail and Atlantic salmon have a slight negative charge, and that trypsin derived from pig has a slight positive charge.

*Table 41 – **Surface properties of trypsin models***. *The table presents the hydrophobic moment, molecular weight in kilodalton (kDa), and charge in electronvolt (eV) for all trypsin models.*

Trypsin model	Hydrophobic moment	Molecular weight (kDa)	Charge (eV)
Pig	638.55	23.52	4.00
Salmon	574.39	23.86	-3.00
Sardine	484.93	23.91	-7.00
Yellowtail	453.25	24.07	-5.00
Red king crab	600.36	24.75	-20.00

The Protein Surface Analyser tool also sorts the surface residues into 'patches' that are regions of the surface that display certain property types, either positively charged, negatively charged or hydrophobic, their size in $Å^2$, sum of potential energies from all points on the patch assigned as a score, the intensity of the patch defined as score over size, and the residues contributing to the energy score of the patch.

Figure 12 presents the side of the surface of the five trypsin models where the substrate binding site is located, with the serine, histidine and aspartic acid residues of the catalytic triad and the aspartic acid involved in substrate binding marked, and the binding site roughly indicated by a circle, as well as displaying the colour coded patches involved in the binding site. The pig trypsin (figure 12A) shows a more hydrophobic an positively charged binding site surface, Atlantic salmon trypsin (figure 12B) shows a more positively charged binding site surface, Pacific sardine trypsin (figure 12C) shows a mostly neutral binding site surface with some negative and hydrophobic patches, yellowtail trypsin (figure 12D) shows a more positively charged binding site surface, and red king crab trypsin (figure 12E) shows a more negatively charged and hydrophobic binding site surface.
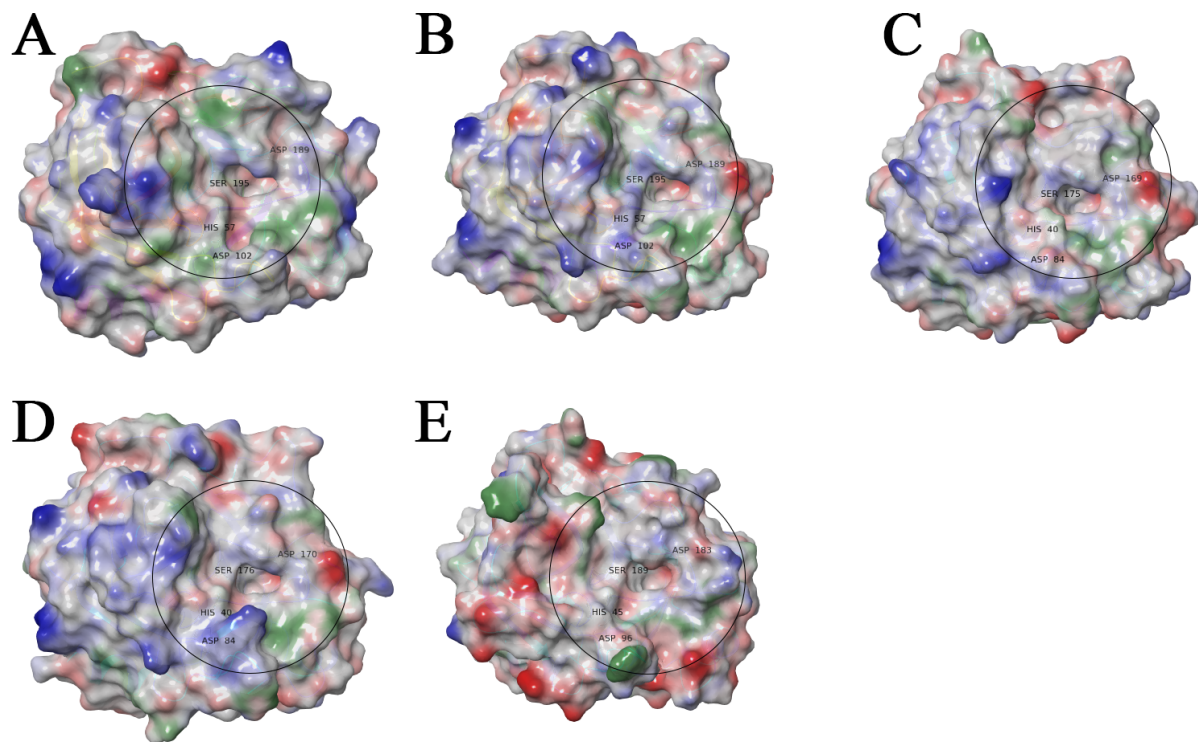
Figure 12 – Surface patches of the binding sites of trypsin models. The figure presents the surface patches and charge of the trypsin models of pig (12A), salmon (12B), sardine (12C), yellowtail (12D), red king crab (12E). Green patches indicate hydrophobic areas, blue patches indicate positive charge, red patches indicate negative charge, while grey areas indicate more neutral areas or overlapping patches. The approximate location of surface binding areas is indicated by a black circle, and the aspartic acid, histidine, and serine of the catalytic triad and the aspartic acid involved in ligand binding are indicated.

Figure 13 presents the side of the surface of the constructed PAR-2 peptide where the serine and arginine where the activation cleavage occurs, as well as presenting the colour coded surface patches, showing a hydrophobic and positively charged patch where the cleavage occurs.
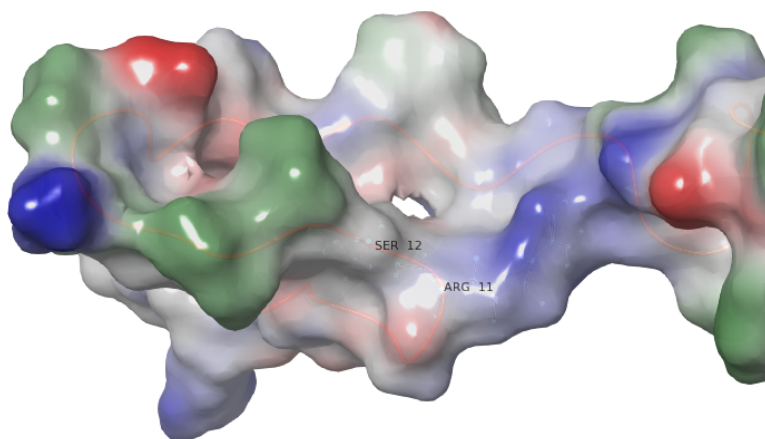
Figure 13 – Surface patches of the interacting side of the N-terminal PAR-2 peptide segment. *The figure presents the surface patches and charge of the N-terminal PAR-2 peptide segment. Green patches indicate hydrophobic areas, blue patches indicate positive charge, red patches indicate negative charge, while grey areas indicate more neutral areas or overlapping patches. The serine and arginine residues where the peptide bond cleavage occurs are indicated.*

Tables 42-46 presents the surface patches with amino acid residues present in the binding site of the trypsin structures (listed in tables 32-39) with a score higher than 150.

Table 42 presents four patches with predicted involvement in the binding site of pig trypsin; the two highest scoring patches display positive charge, while the other two display negative charge.

Table 42 – Patches involved in the surface area binding site of pig trypsin. *The table presents the patches containing amino acid residues that are part of the surface binding area of pig trypsin with a score higher than 150, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms (Å²), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.*

Patch no.	Type	Size (Å²)	Score	Intensity (score/size)	Residues
16	pos	378.4	435.785	1.152	**Arg62**, **Lys60**, Arg66, **Ser61**, Asn84, Asn84, Asn34, **Ser37**, Phe82, Gln64, **Gly38**, Ile63
1	pos	322.5	291.448	0.904	Lys230, **Lys169**, Ser166, Gly178, Ser164, Cys128, Ala132, **Gly174**, Ser127, Asp165, **Thr177**, Ala130
26	neg	210.0	192.548	0.917	**His57**, **Asp189**, **Ser190**, **Ser214**, **Gly219**, **Tyr59**, **Asp102**, **Trp215**, **Lys224**, **Gly216**, **Cys220**, **Ser195**, **Phe41**, **Gly226**, **Cys42**
36	neg	145.3	177.511	1.222	Asp74, **Ile73**, **Ser153**, Tyr151, Val75, Pro152

Table 43 presents the patches with residues predicted to be involved in the binding site of Atlantic salmon trypsin. Table 43 presents five patches, two positively charged patches, two negatively charged patches and one patch displaying hydrophobic properties, with the highest scoring patch having positive charge.

Table 43 - Patches involved in the surface area binding site of salmon trypsin. The table presents the patches containing amino acid residues that are part of the surface binding area of salmon trypsin with a score higher than 150, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms (Å^2), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.

Patch no.	Type	Size (Å^2)	Score	Intensity (score/size)	Residues
1	pos	945.2	950.839	1.006	Lys110, **Arg90**, **Arg62**, **Arg87**, **Lys60**, Lys107, Arg66, **Gln192**, **His57**, **Tyr39**, Ser84, Tyr245, **Ser96**, **Ser37**, **Ser61**, **Tyr59**, **Tyr94**, Ser244, Asn34, Phe82, **Hie40**, Ser86, **Tyr97**, **Phe41**, Asn143, Gly38, **Ser195**, Ile89, **Gly219**, **Gly193**, Ser109
28	neg	143.6	193.957	1.351	Asp236, **Ser240**, Asn235, Thr239, Ser244, **Thr241**, Ala243
42	hyd	294.6	184.446	0.626	**Tyr217**, **Tyr39**, **Ile99**, **Met175**, **Trp215**, **Cys42**, **Phe41**, Lys60, **His57**, **Tyr172**
18	neg	189.0	173.892	0.920	Asp189, Ser190, **Gly219**, **Gly216**, **Tyr97**, **Ile99**, **His57**, **Trp215**, **Ser214**, Ser96, **Cys220**, **Asn224**, Gly226, Tyr217
12	pos	147.2	158.242	1.075	**Lys188**, Asn159, Met135, Leu158, Tyr20

Table 44 presents seven patches containing amino acid residues predicted to be involved in the binding site of Pacific sardine; two positively charged patches, three negatively charged patches, and two patches with hydrophobic properties. The highest scoring patch displays positive charge, while the three subsequent patches display negative charge.

Table 44 - Patches involved in the surface area binding site of sardine trypsin. The table presents the patches containing amino acid residues that are part of the surface binding area of sardine trypsin with a score higher than 150, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms (Å2), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.

Patch no.	Type	Size (Å²)	Score	Intensity (score/size)	Residues
4	pos	473.5	478.206	1.010	Arg45, Arg69, Lys89, Gln47, Ser66, Arg49, Asn33, Phe64, Asn19, Ser68, **Gly21**, Trp34, Ser220, Ile71, Ser91, **Tyr22**
32	neg	457.0	467.975	1.024	**Asp131**, Glu59, Glu62, Tyr57, **Thr56**, **Tyr22**, Asp54, Ser58, Gly60, Ser61, Glu52, **Trp121**, Hie53, **Gly122**, **Gly130**, Gly51, **Hie23**, Gln15, Phe64, Pro13, Tyr99
31	neg	281.3	343.737	1.222	Asp199, Glu197, **Glu165**, **His200**, **Ser149**, **Asn148**, Gly166, Leu164
35	neg	251.5	238.905	0.950	Asp174, **Ser127**, Gln134, **Met125**, **Ser126**, Gly4, Val2, Gly3, **Ser129**, **Gly122**, Asp169, **Asn123**, Ile1, Gly120, **Thr124**, Val128
52	hyd	281.4	172.914	0.614	Ile71, **Arg72**, **Pro74**, Ala110, Ile140, Ala157, Phe159, Gln186, Cys208, Leu209, **Trp213**, **Thr217**
50	hyd	267.0	161.791	0.606	**Tyr193**, **Tyr79**, Met153, **Ile81**, Hie40, **Trp191**, **His200**, Tyr150

1	pos	150.1	154.152	1.027	**Arg72**, **Ser78**, Tyr42, **Tyr79**, **Tyr76**

Table 45 presents eight patches containing residues predicted to be involved in the binding site of trypsin from yellowtail. Three patches show negative charge, four patches show positive charge, and one patch display hydrophobic properties. The highest scoring patch displays negative charge, and the three subsequent highest scoring patches display positive charge.

Table 45 - Patches involved in the surface area binding site of yellowtail trypsin. The table presents the patches containing amino acid residues that are part of the surface binding area of yellowtail trypsin with a score higher than 150, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms (Å^2), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.

Patch no.	Type	Size (Å^2)	Score	Intensity (score/size)	Residues
35	neg	460.7	532.592	1.156	Glu6, Cys7, Thr8, Gln12, Ser13, Gln15, Gly51, Glh52, Hie53, **Arg56**, Val57, Thr58, Glu59, Asp60, Ser61, Glu62, Phe64, Tyr99, Arg133, Leu134,
2	pos	373.3	393.204	1.053	Lys92, Arg45, **Lys43**, Ser66, **Ser44**, Ser20, **Tyr22**, **Phe24**, Ser68, Phe64, **Gly21**, Ser91
1	pos	271.8	278.628	1.025	**Arg78**, **Arg72**, **Tyr79**, **Tyr76**, **Tyr42**
7	pos	166.5	262.380	1.576	**Arg69**, Lys89, Asn33, Trp34, **Ile71**, **Met70**
27	neg	190.0	203.487	1.071	**Asn123**, Asn148, Asn149, Ser150, **Pro152**, **Gln173**, Gly195, Cys196, Ala197, **Glu198**, Asn200

49	hyd	308.3	192.489	0.624	**Tyr22**, **His40**, **Thr80**, **Leu81**, Tyr151, Met154, **Trp192**, **Tyr194**, **Tyr201**
12	pos	192.6	187.903	0.976	Ser108, Cys109, Ala110, Pro111, Ala112, Ser143, **Tyr144**, Cys147, **Asn157**, **Thr158**, Lys207, Val210
23	neg	154.1	187.070	1.214	**Asp156**, **Gly153**, T**hr80**, Ile155, **Tyr144**, Tyr79, **Pro152**, **Met154**

Table 46 presents six patches containing amino acid residues predicted to be involved in the binding site of trypsin from red king crab; five negatively charged patches, with the lowest scoring patch displaying hydrophobic properties.

Table 46 - Patches involved in the surface area binding site of red king crab trypsin. The table presents the patches containing amino acid residues that are part of the surface binding area of red king crab trypsin with a score higher than 150, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms (Å²), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.

Patch no.	Type	Size (Å²)	Score	Intensity (score/size)	Residues
29	neg	681.9	839.743	1.231	**His45**, **Gln48**, **Gly90**, **Phe91**, **Ser92**, **Ser94**, Ser155, Ala157, Asp158, **Arg160**, **Ala161**, **Ser162**, **Tyr163**, **Gly164**, **Glu165**, **Ser166**, **Asp167**, **Ile168**, **Asp169**, **Ser171**, Met172, Asp183, Ala184, Cys185, **Gln186**, **Gly187**, **Ser189**, Val207, **Ser208**, **Trp209**, **Gly212**, **Cys213**, Asn217, Pro219, Gly220, Val221, Tyr222
33	neg	527.8	652.979	1.237	Gln19, **Asp20**, **Thr21**, **Ser22**, **Glu26**, **Phe27**, **Phe29**, **His45**, **Cys46**, **Val47**, **Gln48**, **Gly49**, **Glu50**, **Asp51**,

					Phe52, Asp53, Ser54, Pro55, Ala56, Ser57, Leu58, Gln59, Lys76, Val79, Ser80, Ile82
34	neg	410.9	419.425	1.021	Asp144, Glu71, **Tyr68**, **Leu67**, Glu74, Asp64, Ser142, Asn73, Ala70, Gly72, Thr66, Ser69, Gly63, Pro9
23	neg	318.1	412.045	1.295	Asp170, Asp230, Glu224, **Asp156**, Tyr227, Gln120, Ile173, Ile153, Gly121, Ala124, Leu117, Val123, Gln122, Pro118
26	neg	251.2	341.585	1.360	**Glu86**, Asp87, Asn235, Ala234, **Tyr88**, **Gln84**, Leu233, Ile83
46	hyd	363.3	243.436	0.670	**Phe27**, **Hie28**, **Phe29**, **His45**, Tyr163, **Ser189**, Pro216

Table 47 presents surface patches of the PAR-2 peptide segment with a score over 100. Three of the patches have positive charge, two patches have negative charge and one displays hydrophobic properties. The two patches with the highest score and biggest size have positive charge.

Table 47 - Patches involved in the surface area binding area of the N-terminal PAR-2 peptide segment. The table presents the patches containing amino acid residues that are part of the surface binding area of pig trypsin with a score higher than 100, and includes the patch number derived from the surface analyser tool, type of charge, size in square Ångstroms ($Å^2$), an assigned score derived from the sum of potential energies, patch intensity expressed as score over size, and the amino acid residues contributing to the patch.

Patch no.	Type	Size ($Å^2$)	Score	Intensity (score/size)	Residues
22	pos	454.6	515.232	1.133	**Lys9**, **Arg11**, **Lys26**, **Thr24**, Val30, **Gly10**, **Gly27**, Gly25, Val23, **Glu31**, Val28
24	pos	310.6	294.111	0.947	**Arg6**, **Asn5**, **Thr4**, Thr20, Gly3, **Ser7**, Gln2, Gly19, **Ser8**
58	neg	238.4	267.950	1.124	Asp18, Val17, **Leu13**, **Ile14**, Thr20, **Lys16**, **Asn5**, Ser21, His22, **Ser12**, Gly19, **Ser8**
74	hyd	227.9	162.149	0.711	**Ile14**, **Leu13**, Val17, **Lys16**
25	pos	112.5	155.563	1.383	**Lys16**, Val17
57	neg	114.3	128.615	1.125	**Glu31**, **Val33**, Val28, **Phe34**, Thr32, Val30

5 Discussion

In order to be able to assess the electrostatic affinity of PAR-2 to trypsin, homology models of trypsin from three species: Pacific sardine, yellowtail and red king crab were constructed, as well as the peptide segment of PAR-2 where the tethered ligand is located, since it has yet to be resolved. Homology modelling is the process of computationally predicting the 3D structure of protein structures using the structures of homologous proteins that has already been resolved. Protein-protein docking and interaction studies were conducted to assess the initial binding interaction between the PAR-2 peptide segment and the selected trypsin models.

As mentioned in the introduction, a target and template structure of 100 amino acids in length should have roughly 30% sequence identity between them[26], which the alignment data for the trypsin sequences presented in table 2 exceeds, indicating that the templates selected, and presented in table 1, are sufficient for constructing homology models. Figures 3-5 also present close to perfect query coverage between the trypsin targets and templates, with the exception of on small gap/insertion at position 130 of the alignment between the Pacific sardine target sequence and the Atlantic salmon template, which also is a good indicator for an accurate homology model. The SAVES metaserver and enrichment calculations of docking and scoring results was used to evaluate the quality of the generated homology models. The Verify3D scores for the homology models of Pacific sardine, yellowtail and red king crab trypsin presented in table 3 are all within acceptable scores for the 1D-3D compatibility according to the output on the SAVES metaserver[31]. The ERRAT score for the homology models presented in table 3 indicate that the homology model structure for sardine trypsin fall below the rejection limit for the quality factor of non-bonded atomic interactions. However, the homology models for yellowtail trypsin and red king crab trypsin do not fall below this limit, according to the SAVES metaserver[31], with scores of 94.8718 and 91.5094, respectively. As presented in figure 7, an area from amino acid residue position 56 to 63 of the yellowtail homology model is within the warning zone limit for quality factor for non-bonded interactions, as well as an area of residues stretching from position 35 to 38 is within the error zone limit. The residues in positions 35-38 appear to be buried within the protein structure, and even though they are relatively close to the histidine of the catalytic triad at position 40, they should not affect the initial binding interaction of the PAR-2 receptor to the yellowtail trypsin, nor were they observed in any selected poses of docked active ligands to

the structure. The residues in position 56-63 at the yellowtail trypsin homology model are not anywhere close to the catalytic triad of the structure, the arginine residue does seem to be within proximity to at least one of the poses of PAR-2 docking to yellowtail trypsin (table 39), but does not appear to form any direct contacts with the PAR-2 peptide, and the residues are located in the periphery of the initial binding interaction site observed in this study. As presented in figure 8, an area from amino acid positions 83-87 in the red king crab trypsin model are within the warning zone for limit for the quality factor of non-bonded interactions, and a stretch of residues at positions 162-169 fall within the error zone of the quality limit. The residues at positions 83-87 are exposed to the surface and appears to be both present in the initial binding interaction site between red king crab and the PAR-2 peptide (table 34), and the glutamic acid residue at position 86 also appears to directly interact with the PAR-2 peptide in the third most observed pose of the protein-protein docking between the PAR-2 peptide and red king crab trypsin, forming two H-bond bridges and a salt bridge to an arginine residue on PAR-2. The Ramachandran results presented in table 4, show one amino acid in the generously allowed regions in the homology model for yellowtail trypsin, which is an aspartic acid residue at position 60, but it does not appear in the observed binding interaction site. According to the Ramachandran results output from PROCHECK in SAVES, good quality models with a resolution of at least 2Å should score above 90% in the most favourable region[31], which the results in table 4 do not reflect. The Pacific sardine trypsin appear to have the best structure quality of the models, while both yellowtail and especially red king crab seem to have a few problems even on the surface binding site, and some of these proteins seem to actively bind to the PAR-2 receptor peptide. However, these were still the highest quality structures produced. These quality problems may be due to the high degree of loop regions observed in the trypsin structure, which can be difficult to predict a good conformation due to their less rigid nature, as discussed in the introduction[48].

Docking and scoring of active ligands with known affinity to bovine trypsin and decoys was used as an additional quality check of the binding site of the trypsin structures. Figure 9 presents the ROC-plots that rank actives and decoys against each other to assess whether actives truly bind better to the protein than the decoys, and the resulting ROC-score, along with enrichment factors, are presented in table 5, and show that all five structures have decent affinity to the active compounds, but salmon has significantly higher scores than the rest of the structures. This might simply be because the structure is of better quality, due to sardine and yellowtail having quite similar amino acid sequences to the Atlantic salmon, but still

achieving the same scores as pig and red king crab trypsin. Nevertheless, all five structures have significant affinity for the active compounds in the docking process, indicating that they have the binding properties of trypsin.

The N-terminal peptide sequence of PAR-2 was constructed using homology modelling and consensus models of PAR-1 and PAR-4 as template structures due to the structural conservation seen across GPCRs. As figure 10 presents, the sequence identity between PAR-2 and the templates is too low to build a decent homology model from, but the sequences were still used as structural templates for the construction of the homology model. As mentioned previously, loop regions are less conserved and are harder to model due to their less rigid nature, thus the resulting PAR-2 model is probably not very great. Due to being attached to a GPCR, isolated quality assessment of this short peptide segment was reduced to comparing it to the PAR-1 and PAR-4 segments. Figure 11 indicates that its structure is mostly, if not entirely, derived from the PAR-1 peptide segment, likely due to slightly higher sequence identity. Additionally, the folding of the constructed PAR-2 N-terminal segment makes the peptide bond between the serine and arginine residues unavailable to some degree.

Protein-protein docking was conducted in order to study the binding interactions of the PAR-2 peptide to trypsin models. The protein-protein docking in BioLuminate only uses rigid-body docking, meaning that it does not take in to account the dynamics and flexibility related to the interaction between two proteins, thus the observed interactions between proteins are more likely to be that of an initial binding interaction. Tables 11, 15, 19, 23 and 27 present the summary of binding interactions found in the top three poses of each PAR-2-trypsin docking for pig, salmon, sardine, yellowtail and red king crab, respectively. The most obvious variation between the five species is the number of salt bridges present in the PAR-2-trypsin, with residues on the surface of pig trypsin forming no salt bridges to residues on PAR-2, salmon forming 1-2 salt bridges per pose, sardine forming 0, 2, and 1 salt bridges across the poses, yellowtail forming 1 salt bridge per pose, and red king crab forming 3-4 salt bridges per pose. Red king crab also has more H-bonds and Van der Waal clashes than the other structures, 19 and 18, respectively (table 27), followed by yellowtail with 19 H-bonds and 10 Van der Waal clashes (table 23), sardine with 16 H-bonds and 10 Van der Waal clashes (table 19), salmon with 11 H-bonds and 8 Van der Waal clashes (table 15), and finally pig trypsin with 13 H-bonds and 7 Van der Waal clashes in the binding interactions with the PAR-2 peptide segment (table 11). These numbers might indicate a slightly closer, or at least stronger, binding of the trypsin structures with more bonds and clashes than the structures

with fewer bonds and clashes. Salt bridges are electrostatic interactions and might indicate stronger electrostatic attractions between PAR-2 and the trypsin structures with more salt bridges present in the protein-protein docking results.

Mapping of binding sites can be crucial in understanding the differences in the initial binding interactions between the trypsin structures and the PAR-2 peptide segment. The amino acid residues resulting from the top three poses of each trypsin structure docking to the PAR-2 peptide segment were used to determine which amino acid residues were present in the surface binding area of the structures as well as studying which of the amino acid residues involved in the binding interaction vary between the trypsin structures. Table 31 presents the amino acid residues observed in the binding of the PAR-2 peptide segment to trypsin structures. Tables 32-39 present segments of amino acids from the trypsin structures involved in the surface binding area involved in the initial binding of the PAR-2 peptide segment. Table 33 (row 7) shows an insert of a negatively charged glutamic acid residue for red king crab that forms a salt bridge both to a lysine in PAR-2 in poses 2 and 3 (tables 29 and 30, respectively) that might be advantageous for initial binding. Table 34 (row 4) shows a positively charged arginine in salmon that form a salt bridge and an H-bond to the PAR-2 peptide segment (table 16), which corresponds to another arginine in sardine trypsin that forms double H-bonds to a residue on PAR-2 (table 21), and a polar uncharged threonine in pig trypsin that forms an H-bond to the PAR-2 peptide segment (table 13), with the corresponding arginine and glutamine of yellowtail and red king crab do not form bonds with the peptide in the poses. Table 34 (row 6) also shows a negatively charged glutamic acid in red king crab that corresponds to hydrophobic proline residues in the four other species, and the glutamic acid forms a double hydrogen bond and a salt bridge with PAR-2. Table 36 (row 8) shows a negatively charged aspartic acid in red king crab that for a salt bridge and an H-bond to PAR-2 (table 30), which corresponds to a hydrophobic methionine in salmon, sardine and yellowtail trypsin, and a polar uncharged glutamine in pig trypsin. Table 38 (row 9) display a positively charged arginine in red king crab that does not interact with the PAR-2 peptide segment, but corresponds to negatively charged glutamic acids that form salt bridges and H-bonds to the PAR-2 segment in salmon (table 17), sardine (table 21 and 22), and yellowtail (tables 24-26). The latter two residues (table 36, row 8 and table 38, row 9) have been suggested by Larsen, *et.al.,* 2013[9] to might have an effect on the binding of substrate to the binding pocket, and base on the observations form the protein-protein docking, that the PAR-2 peptide segment seems more prone to bonded interactions, in particular creating salt

bridges, with negatively charged residues in these particular positions. However, table 34, row 4 shows that the PAR-2 segment can be prone to bonding to positively charged amino acids.

Tables 40-41 present the total surface charge properties of the trypsin models. Table 40 shows that all the trypsin models have a larger sum positive surface area than negative surface area, even though table 41 presents that all structures except pig trypsin have a negative charge. This is likely due to the use of physiological pH in molecular modelling, causing positively charged histidine residues to lose their charge due to a pKa of approximately 6[3], and the calculation accounting for them as part of the positive surface area. Table 41 shows that trypsin derived from pig has a small net positive charge, while salmon, sardine and yellowtail have a low negative charge, and red king crab has a larger negative charge. Tables 42-46 presents the patches that contain amino acid residues involved the surface binding area mapped from the protein-protein docking. The patches indicate that red king crab has a very negatively charged binding area, while the other four trypsin structures has both positive and negative patches presented, particularly pig and salmon have some large positively charged patches in their binding site. The surface patch data of the N-terminal PAR-2 can be seen in figure 13 and in table 47, which seem to indicate that the binding of the PAR-2 peptide segment has a slight positive charge, but definitely has both positive and negative patches present around the its binding area.

The idea that the surface charge of trypsin affects the binding and subsequent cleavage of PAR-2 was briefly mentioned in the introduction. The results from this study certainly points toward a larger initial electrostatic attraction between the significantly negatively charged red king crab trypsin and the N-terminal peptide segment of PAR-2 considering the higher degree of bonding between the two proteins compared to the slightly positively charged pig trypsin, and showing a little better bonding with the slightly negatively charged trypsin models of salmon, sardine and yellowtail. As mentioned in the introduction, Larsen *et.al.*, 2013[9] have suggested that electrostatic interactions can be important in binding, cleavage and activation of the PAR-2 receptor.

Overall, it seems that the constructed N-terminal peptide segment PAR-2 have better initial electrostatic binding to the more negatively charged king crab trypsin than the slightly positively charged pig trypsin, with the Atlantic salmon, Pacific sardine and yellowtail having an electrostatic affinity in between these two.

6 Conclusion

Homology modelling was used to construct trypsin models of Pacific sardine, yellowtail and red king crab using templates derived from the same enzyme with a high degree of homology. The models showed a good 1D-3D relationship, but had some slight, but not major, problems with non-bonded atomic interactions and stereochemistry, likely due to the high degree of loop regions and long loop regions making it hard to predict a high-quality conformation. Docking and scoring was conducted to assess the binding site quality and compare it to the two other models used further in the study, trypsin from pig and Atlantic salmon. The homology models were successful in distinguishing between ligands with known affinity to bovine trypsin from a much larger set of decoys. The homology modelling of the PAR-2 N-terminal peptide segment was another challenge, since the templates that were used for construction did not satisfy the degree of sequence identity necessary to predict a good homology model, and instead corresponding segments from structurally related PAR-1 and PAR-4 were used to construct a model of the PAR-2 peptide segment. The N-terminal peptide sequence of PAR-2 was docked on to all the chosen trypsin models as a rigid body, in order to assess initial binding reactions between the two entities. Interaction analysis reviled that red king crab had a higher number of interactions than the other trypsin models, followed by yellowtail, sardine and salmon, and pig trypsin in that order. Mapping and alignment of trypsin binding sites and surface analysis of the trypsin models confirmed that red king crab has a significantly more negative surface charge than the other molecules, and some of the differentiating amino acid residues with negative charge in red king crab were involved in forming bonds to the docked PAR-2 peptide segment.

The studies do indicate that the N-terminal peptide segment of PAR-2 have a higher degree of electrostatic attraction towards more negatively charged trypsin, but they only assess initial binding and electrostatic attraction in a static, simulated environment, and thus further studies in using assays and molecular dynamics should be applied before jumping to conclusions.

Works cited

1.   Kim, Y.H. and S.-H. Lee, *TGF-β/SMAD4 mediated UCP2 downregulation contributes to Aspergillus protease-induced inflammation in primary bronchial epithelial cells.* Redox Biology, 2018. **18**: p. 104-113.

2.   Haddad, M. and S. Sharma, *Physiology, Lung.* 2019, StatPearls Publishing [Internet]: Treasure Island, FL.

3.   Nelson, D.L. and M.M. Cox, *Lehninger Principles of Biochemistry.* 6th ed. 2013, New York: W. H. Freeman.

4.   Florsheim, E., et al., *Integrated innate mechanisms involved in airway allergic inflammation to the serine protease subtilisin.* Journal of immunology (Baltimore, Md. : 1950), 2015. **194**(10): p. 4621-4630.

5.   Romberger, D.J., et al., *Proteases in agricultural dust induce lung inflammation through PAR-1 and PAR-2 activation.* American journal of physiology. Lung cellular and molecular physiology, 2015. **309**(4): p. L388-L399.

6.   Florencio, A.C., et al., *Effects of the serine protease inhibitor rBmTI-A in an experimental mouse model of chronic allergic pulmonary inflammation.* Scientific reports, 2019. **9**(1): p. 12624-12624.

7.   Gieseck, R.L., M.S. Wilson, and T.A. Wynn, *Type 2 immunity in tissue repair and fibrosis.* Nature Reviews Immunology, 2018. **18**(1): p. 62-76.

8.   Kawabata, A. and N. Kawao, *Physiology and Pathophysiology of Proteinase-Activated Receptors (PARs): PARs in the Respiratory System: Cellular Signaling and Physiological/Pathological Roles.* Journal of Pharmacological Sciences, 2005. **97**(1): p. 20-24.

9.   Larsen, A.K., et al., *Differences in PAR-2 activating potential by king crab (Paralithodes camtschaticus), salmon (Salmo salar), and bovine (Bos taurus) trypsin.* BMC research notes, 2013. **6**: p. 281-281.

10.   Baird, T.T. and C.S. Craik, *Chapter 575 - Trypsin*, in *Handbook of Proteolytic Enzymes*, N.D. Rawlings and G. Salvesen, Editors. 2013, Academic Press. p. 2594-2600.

11.   Hedstrom, L., *Serine Protease Mechanism and Specificity.* Chemical Reviews, 2002. **102**(12): p. 4501-4524.

12.   Di Cera, E., *Serine proteases.* IUBMB Life, 2009. **61**(5): p. 510-515.

13.   Williams, J.A., *Trypsin*, in *Encyclopedia of Gastroenterology*, L.R. Johnson, Editor. 2004, Elsevier: New York. p. 533-534.

14.   Bhutta, Z.A. and K. Sadiq, *Protein Digestion and Bioavailability*, in *Encyclopedia of Human Nutrition*, B. Caballero, Editor. 2013, Academic Press: Waltham. p. 116-122.

15.   Polgár, L., *The catalytic triad of serine peptidases.* Cellular and Molecular Life Sciences CMLS, 2005. **62**(19): p. 2161-2172.

16.   Baird, T.T., *Trypsin*, in *Brenner's Encyclopedia of Genetics (Second Edition)*, S. Maloy and K. Hughes, Editors. 2013, Academic Press: San Diego. p. 216-219.

17. Cooper, G.M., *Functions of Cell Surface Receptors*. 2nd ed. The Cell: A Molecular Approach. 2000, Sunderland, MA: Sinauer Associates.

18. Houle, S. and M.D. Hollenberg, *PAR2 Proteinase-Activated Receptor*, in *xPharm: The Comprehensive Pharmacology Reference*, S.J. Enna and D.B. Bylund, Editors. 2007, Elsevier: New York. p. 1-13.

19. Coughlin, S.R., *Chapter 26 - Protease-Activated Receptors*, in *Handbook of Cell Signaling (Second Edition)*, R.A. Bradshaw and E.A. Dennis, Editors. 2010, Academic Press: San Diego. p. 171-175.

20. Zhang, P., L. Covic, and A. Kuliopulos, *Chapter 13 - Protease-Activated Receptors*, in *Platelets (Third Edition)*, A.D. Michelson, Editor. 2013, Academic Press. p. 249-259.

21. Ossovskaya, V.S. and N.W. Bunnett, *Protease-Activated Receptors: Contribution to Physiology and Disease.* Physiological Reviews, 2004. **84**(2): p. 579-621.

22. Schmidlin, F., et al., *Protease-Activated Receptor 2 Mediates Eosinophil Infiltration and Hyperreactivity in Allergic Inflammation of the Airway.* The Journal of Immunology, 2002. **169**(9): p. 5315.

23. Ramachandran, K.I., D. Gopakumar, and K. Namboori, *Computational Chemistry and Molecular Modeling: Principles and Applications*. 2008, Springer Berlin Heidelberg: Berlin, Heidelberg.

24. Orry, A.J.W. and R. Abagyan, *Homology Modelling: Methods and Protocols*. Springer Protocols. 2012, San Diego, CA: Humana Press.

25. Leelananda, S.P. and S. Lindert, *Computational methods in drug discovery.* Beilstein journal of organic chemistry, 2016. **12**: p. 2694-2718.

26. Krieger, E., S.B. Nabuurs, and G. Vriend, *Homology Modeling*, in *Structural Bioinformatics*. 2005, John Wiley & Sons, Inc. p. 509-523.

27. Vyas, V.K., et al., *Homology modeling a fast tool for drug discovery: current perspectives.* Indian journal of pharmaceutical sciences, 2012. **74**(1): p. 1-17.

28. Johnson, M., et al., *NCBI BLAST: a better web interface.* Nucleic acids research, 2008. **36**(Web Server issue): p. W5-W9.

29. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Research, 2000. **28**(1): p. 235-242.

30. Burley, S.K., et al., *RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy.* Nucleic Acids Research, 2018. **47**(D1): p. D464-D474.

31. UCLA-DOE. *SAVES v5.0* [cited 2019; Available from: https://servicesn.mbi.ucla.edu/SAVES/.

32. Glaab, E., *Building a virtual ligand screening pipeline using free software: a survey.* Briefings in bioinformatics, 2016. **17**(2): p. 352-366.

33. Friesner, R.A., et al., *Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy.* Journal of Medicinal Chemistry, 2004. **47**(7): p. 1739-1749.

34. Truchon, J.-F. and C.I. Bayly, *Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem.* Journal of Chemical Information and Modeling, 2007. **47**(2): p. 488-508.

35. Kozakov, D., et al., *The ClusPro web server for protein–protein docking.* Nature Protocols, 2017. **12**(2): p. 255-278.

36. The UniProt, C., *UniProt: a worldwide hub of protein knowledge.* Nucleic Acids Research, 2018. **47**(D1): p. D506-D515.

37. Rawlings, N.D., et al., *The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database.* Nucleic Acids Research, 2017. **46**(D1): p. D624-D632.

38. Mysinger, M.M., et al., *Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking.* Journal of Medicinal Chemistry, 2012. **55**(14): p. 6582-6594.

39. Lüthy, R., J.U. Bowie, and D. Eisenberg, *Assessment of protein models with three-dimensional profiles.* Nature, 1992. **356**(6364): p. 83-85.

40. *Schrödinger Release 2019-3: Maestro.* 2019, Schrödinger, LLC: New York, NY.

41. *Schrödinger Release 2019-3: Prime.* 2019, Schrödinger, LLC: New York, NY.

42. *Schrödinger Release 2019-3: Protein Preparation Wizard.* 2019, Schrödinger, LLC: New York, NY.

43. *Schrödinger Release 2019-3: LigPrep.* 2019, Schrödinger, LLC: New York, NY.

44. *Schrödinger Release 2019-3: Glide.* 2019, Schrödinger, LLC: New York, NY.

45. *Schrödinger Release 2019-3: BioLuminate.* 2019, Schrödinger, LLC: New York, NY.

46. Davies, M., et al., *ChEMBL web services: streamlining access to drug discovery data and utilities.* Nucleic Acids Research, 2015. **43**(W1): p. W612-W620.

47. Gaulton, A., et al., *The ChEMBL database in 2017.* Nucleic Acids Research, 2016. **45**(D1): p. D945-D954.

48. Madeira, F., et al., *The EMBL-EBI search and sequence analysis tools APIs in 2019.* Nucleic acids research, 2019. **47**(W1): p. W636-W641.

Appendix A – Alignment of trypsin sequences

A multiple sequence alignment of the 5 trypsin models in the study, used to align the binding site assessment.

Results for job clustalo-I20191215-225803-0460-22236288-p1m

```
CLUSTAL O(1.2.4) multiple sequence alignment


tr|Q8WR10_PARCM|30-266      IVGGTEVTPGEIPYQLSFQDTSFGGEFHFCGASIYKDTWAICAGHCVQGEDFDSPASLQI      60
sp|P00761_PIG|9-229         IVGGYTCAANSIPYQVSL-----NSGSHFCGGSLINSQWVVSAAHCYKS-------RIQV      48
MER0052997_yellowtail       IVGGYECTPHSQSHQVSL-----NSGYHFCGGSLVNENWVVSAAHCYKS-------RIEV      48
sp|P35031_SALSA|21-240      IVGGYECKAYSQTHQVSL-----NSGYHFCGGSLVNENWVVSAAHCYKS-------RVEV      48
MER0279703_sard             IVGGYECKAYSQPWQVSL-----NSGYHFCGGSLVNENWVVSAAHCYKS-------RVQV      48
                            ****    .   *:*:       ..  ****.*: :. *.:.*.** :.        :::


tr|Q8WR10_PARCM|30-266      VAGDHTLYSAEGNEQKIAVSKIIQHEDYNGFSISNDISLLQFASPLTFNSFVGPIALPAQ      120
sp|P00761_PIG|9-229         RLGEHNIDVLEGNEQFINAAKIITHPNFNGNTLDNDIMLIKLSSPATLNSRVATVSLPRS      108
MER0052997_yellowtail       RLGEHHIRVTEDSEQFISGSRMIRNPYYNRYTLANDIMLIKLSKPATLNQYVQPVALPTS      108
sp|P35031_SALSA|21-240      RLGEHNIKVTEGSEQFISSSRVIRHPNYSSYNIDNDIMLIKLSKPATLNTYVQPVALPTS      108
MER0279703_sard             RMGEHDITYSEGSEQFISSSRVIRHPNYSSYNIDNDIMLIKLSSPVTLNQYVQPVALPTS      108
                              *:* :   *..** *  ::.* :  :.  .: *** *:::..* *.*  *  ::** .


tr|Q8WR10_PARCM|30-266      GQVASGDCTCTGWGTTTEGGY-SSDALLKVTMPIVSDADCRASYGESDIDDSMICAGVPQ      179
sp|P00761_PIG|9-229         CAAAGTECLISGWGNTKSSGSSYPSLLQCLKAPVLSDSSCKSSYPG-QITGNMICVGFLE      167
MER0052997_yellowtail       CAPAGTMCTVSGWGNTMSSSA-DGDRLQCLNIPILSYEDCNNSYPG-MIDNTMFCAGYLE      166
sp|P35031_SALSA|21-240      CAPAGTMCTVSGWGNTMSSTA-DSNKLQCLNIPILSYSDCNNSYPG-MITNAMFCAGYLE      166
MER0279703_sard             CAPAGTMCTVSGWGNTMSSV--SGDRLQCLQIPILSDRDCNNSYPG-MITDAMFCAGYLE      165
                            *.  *  :***.* ..     . *  : *::*  .*. **    * . *:*.*   :


tr|Q8WR10_PARCM|30-266      GGKDACQGDSGGPLACSDTGSPYLAGIVSWGYGCARPNYPGVYCEVAYYVDWVLANSS--      237
sp|P00761_PIG|9-229         GGKDSCQGDSGGPVVCNG----QLQGIVSWGYGCAQKNKPGVYTKVCNYVNWIQQTIA--      221
MER0052997_yellowtail       GGKDSCQGDSGGPVVCNG----ELQGVVSWGYGCAERNYPGVYSKVCVQTEWLHNTMASY      222
sp|P35031_SALSA|21-240      GGKDSCQGDSGGPVVCNG----ELQGVVSWGYGCAEPGNPGVYAKVCIFNDWLTSTMA--      220
MER0279703_sard             GGKDSCQGDSGGPVVCNG----ELQGVVSWGYGCAERDHPGVYAKVCLFNDWLTQTMAS-      220
                            ****:********:.*..     * *:********. . **** :*.   :*: . . :
```