DISCIPLINE COUNCIL FOR CHEMICAL SCIENCES

mgr inż. Maria BZÓWKA

ANALYSIS OF MOLECULAR ASPECTS OF PROTEINS REGULATION CONSIDERING WATER MOLECULES AS A POTENTIAL MEDIATOR IN INTERMOLECULAR INTERACTIONS

DOCTORAL THESIS

Supervisor: dr hab. Artur GÓRA, prof. PŚ

Gliwice, 2023

If you do not hope, you will not win that which is not hoped for,
since it is unattainable and inaccessible

Heraclitus

I would like to express my sincere gratitude to my PhD Supervisor – Professor Artur Góra for His mentorship, guidance and inspiration to explore scientific issues.

My acknowledgement also goes to my Colleagues from the Tunneling Group for their participation in our joint research, as well as their kindness and support.

I am thankful to all Co-Authors of the scientific papers for the collaborative research efforts we have undertaken together.

Table of Contents

List of publications included in the doctoral thesis

[1] **Paper 1:** Mitusińska K., Raczyńska A., **Bzówka M.**, Bagrowska W., Góra A.
Applications of water molecules for analysis of macromolecule properties
Computational and Structural Biotechnology Journal (2020) 18, 355-365
https://doi.org/10.1016/j.csbj.2020.02.001
IF 2020: 7.271 Points from Polish Ministry of Science and Higher Education: 100

[2] **Paper 2:** Magdziarz T., Mitusińska K., **Bzówka M.**, Raczyńska A., Stańczak A., Banas M., Bagrowska W., Góra A.
AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective
Bioinformatics (2020) 36 (8), 2599-2601
https://doi.org/10.1093/bioinformatics/btz946
IF 2020: 6.937 Points from Polish Ministry of Science and Higher Education: 200

[3] **Paper 3: Bzówka M.*,** Mitusińska K.*, Raczyńska A., Samol A., Tuszyński J.A., Góra A.
Structural and Evolutionary Analysis Indicate That the SARS-CoV-2 Mpro Is a Challenging Target for Small-Molecule Inhibitor Design
International Journal of Molecular Sciences (2020) 21, 1-17
https://doi.org/10.3390/ijms21093099
IF 2020: 5.924 Points from Polish Ministry of Science and Higher Education: 140

[4] **Paper 4:** Fischer A.*, Sellner M.*, Mitusińska K.*, **Bzówka M.***, Lill M.A., Góra A., Smieško M.
Computational selectivity assessment of protease inhibitors against SARS-CoV-2
International Journal of Molecular Sciences (2021) 22 (4), 1-17
https://doi.org/10.3390/ijms22042065
IF 2021: 6.208 Points from Polish Ministry of Science and Higher Education: 140

[5] **Paper 5: Bzówka M.,** Mitusińska K., Hopko K., Góra A.
Computational insights into the known inhibitors of human soluble epoxide hydrolase
Drug Discovery Today (2021) 26 (8), 1914-1921
https://doi.org/10.1016/j.drudis.2021.05.017
IF 2021: 8.369 Points from Polish Ministry of Science and Higher Education: 200

[6] **Paper 6:** Mitusińska K., Wojsa P., **Bzówka M.**, Raczyńska A., Bagrowska W., Samol A., Kapica P., Góra A.
Structure-function relationship between soluble epoxide hydrolases structure and their tunnel network
Computational and Structural Biotechnology Journal (2022) 20, 193-205
https://doi.org/10.1016/j.csbj.2021.10.042
IF 2022: 6.155 Points from Polish Ministry of Science and Higher Education: 100
Together with Corrigendum to "Structure-function relationship between soluble epoxide hydrolases structure and their tunnel network" (vol 20, pg 193, 2022)
Computational and Structural Biotechnology Journal (2022) 20, 2198-2199

[7] **Paper 7:** Mitusińska K.*, **Bzówka M.***, Magdziarz T., Góra A.
Geometry-Based versus Small-Molecule Tracking Method for Tunnel Identification: Benefits and Pitfalls
Journal of Chemical Information and Modeling (2022) 62 (24), 6803-6811
https://doi.org/10.1021/acs.jcim.2c00985
IF 2022: 5.6 Points from Polish Ministry of Science and Higher Education: 100

[8] **Paper 8: Bzówka M*.**, Mitusińska K.*, Raczyńska A., Skalski T., Samol A., Bagrowska W., Magdziarz T., Góra A.
Evolution of tunnels in alpha/beta-hydrolase fold proteins - What can we learn from studying epoxide hydrolases?
PLoS Computational Biology (2022) 18 (5), 1-25
https://doi.org/10.1371/journal.pcbi.1010119
IF 2022: 4.3 Points from Polish Ministry of Science and Higher Education: 140

[9] **Paper 9: Bzówka M.**, Bagrowska W., Góra A.
Recent advances in studying Toll-like receptors with the use of computational methods
Journal of Chemical Information and Modeling (2023) 63 (12), 3669–3687
https://doi.org/10.1021/acs.jcim.3c00419
IF 2022: 5.6 Points from Polish Ministry of Science and Higher Education: 140

[10] **Preprint 1: Bzówka M.**, Szleper K., Stańczak A., Borowski T., Góra A. The proteolytic cleavage of TLR8 Z-loop by furin protease - molecular recognition, reaction mechanism and role of water molecules
Research Square Preprint (2023)
https://doi.org/10.21203/rs.3.rs-3590328/v1

*These authors contributed equally to the work

List of abbreviations

AQ – AQUA-DUCT

B factor (Debye-Waller factor) – temperature factor or atomic displacement parameter

bmEH – *Bacillus megaterium* epoxide hydrolase

CH65-EH – thermophilic enzyme collected from hot-springs in China

COVID-19 – COronaVIrus Disease of 2019

Cryo-EM – cryo-electron microscopy

CTD – C-terminal domain in human soluble epoxide hydrolase

DFT – density functional theory

DNA – deoxyribonucleic acid

EH – epoxide hydrolase

hsEH – *Homo sapiens* epoxide hydrolase / human soluble epoxide hydrolase

IC_{50} – half maximal inhibitory concentration

INT1 – first tetrahedral intermediate state of the TLR8-furin complex

INT2 – second intermediate state of the TLR8-furin (acyl-enzyme complex)

INT3 – subsequent tetrahedral intermediate state of the TLR8-furin complex

IR – infrared radiation

MAV – maximal accessible volume

MC – Monte Carlo

MD – molecular dynamics

MM/GBSA – Molecular Mechanics/Generalised Born Surface Area

Mpro – main protease

mRNA – messenger ribonucleic acid

msEH – *Mus musculus* epoxide hydrolase

NMR – nuclear magnetic resonance

PDB – protein data bank

PROD – product of the TLR8-furin complex

QM/MM – quantum mechanics/molecular mechanics

R factor – residues factor

RE – reactant of the TLR8-furin complex

RISM – reference interaction site model

RNA – ribonucleic acid

rRNA – ribosomal ribonucleic acid

SARS-CoV – severe acute respiratory syndrome coronavirus

SARS-CoV-2 – severe acute respiratory syndrome coronavirus 2

sEH – soluble epoxide hydrolase

SibeEH – thermophilic enzyme collected from hot-springs in Russia

StEH1 – *Solanum tuberosum* epoxide hydrolase

Tc/m – tunnel located at the boarder of cap and main domains in soluble epoxide hydrolase

Tg – tunnel located in a gorge between the Tc/m and Tm1 tunnels in soluble epoxide hydrolase

TIP3P – transferable intermolecular potential with 3 points

TLR – Toll-like receptor

Tm1 – tunnel located in the main domain of soluble epoxide hydrolase

TrEH – *Trichoderma resei* epoxide hydrolase

tRNA – transfer ribonucleic acid

UCHL1 - ubiquitin carboxyl-terminal hydrolase isozyme L1

VrEH2 – *Vigna radiata* epoxide hydrolase

X-ray – X-ray crystallography

Analysis of molecular aspects of proteins regulation considering water molecules as a potential mediator in intermolecular interactions

mgr inż. Maria BZÓWKA

Supervisor: dr hab. Artur GÓRA, prof. PŚ

Abstract

The doctoral dissertation presents the results of several works related to modelling the dynamics of water molecules in biological systems, resulting from the use of dedicated software and computational methods. Results presented in this thesis are related to the analysis of the various functions performed by water molecules in proteins. Specifically, they concern three areas of application: drug design, protein regulation and engineering, as well as studying enzymatic reaction.

As concerns the application of water molecules in drug design – I presented how, by using a combination of small-molecule tracking and local-distribution approaches, it is possible to describe the variations in the dynamics of the internal pockets within the macromolecules and identify novel potential sites for ligand binding. Such analyses were performed for different molecular targets, particularly for SARS-CoV-2 main protease (SARS-CoV-2 Mpro) and human soluble epoxide hydrolase (hsEH). Also, the above-mentioned approach was used during evaluating the potential risk of off-target for SARS-CoV-2 Mpro and a panel of various proteases. For the application in protein regulation and engineering – I showed that, by tracking of water molecules during molecular dynamics simulations, it is possible to describe in details tunnel networks and the transportation phenomena in proteins. Such an analysis was performed for enzymes from the soluble epoxide hydrolase subfamily (sEH). For these enzymes, it was possible to establish the relationship between their structure and their tunnel network but also conduct an evolutionary analysis of the identified tunnels. Also, the small-molecule tracking methods was compared to the geometry-based approach for detecting and analysing tunnels in proteins. As for the enzymatic reaction – I presented that the combination of small-molecule tracking and local-distribution approaches can reveal different roles of water molecules during

particular reaction cycle. This analysis was performed during the investigation of the proteolytic cleavage of the Z-loop in TLR8 by furin protease. In addition to the basic role of water, which is the catalytic function, additional roles have been proposed, in particular related to the stabilisation for certain intermolecular interactions or as a mediator, either during the transfer of a proton or in the dissociation process.

The results of the works included in the doctoral thesis were published in nine peer-reviewed journals and as one preprint sent for the revision.

Analiza molekularnych aspektów regulacji białek z uwzględnieniem cząsteczek wody jako potencjalnego mediatora w oddziaływaniach międzycząsteczkowych

mgr inż. Maria BZÓWKA

Promotor: dr hab. Artur GÓRA, prof. PŚ

Streszczenie

W rozprawie doktorskiej przedstawiono wyniki prac związanych z modelowaniem dynamiki cząsteczek wody w układach biologicznych, wynikających z zastosowania dedykowanego oprogramowania i metod obliczeniowych. Wyniki dotyczą analiz różnych funkcji pełnionych przez cząsteczki wody w białkach, w szczególności obejmują trzy obszary zastosowań: projektowanie leków, regulację i inżynierię białek, a także badania reakcji enzymatycznych.

W przypadku zastosowania cząsteczek wody w projektowaniu leków – przedstawiono w jaki sposób, wykorzystując połączenie śledzenia małych cząsteczek i analizę lokalnej dystrybucji rozpuszczalnika, możliwe jest opisanie zmian w dynamice kieszeni wewnętrznych w makrocząsteczkach oraz zidentyfikowanie nowych, potencjalnych miejsc wiążących ligandy. Analizy te przeprowadzono dla różnych celów molekularnych, między innymi dla głównej proteazy wirusa SARS-CoV-2 (SARS-CoV-2 Mpro) oraz dla ludzkiej rozpuszczalnej hydrolazy epoksydowej (hsEH). Powyższe podejście zastosowano również podczas oceny potencjalnego ryzyka niespecyficznego wiązania inhibitorów SARS-CoV-2 Mpro do różnych proteaz. W przypadku analizy cząsteczek wody w regulacji i inżynierii białek – pokazano, że śledząc cząsteczki wody podczas symulacji dynamiki molekularnej, można szczegółowo opisać sieci tuneli i zjawiska transportowe w białkach. Analiza ta została przeprowadzona dla enzymów z podrodziny rozpuszczalnych hydrolaz epoksydowych (sEH). W przypadku tych enzymów było możliwe ustalenie związku między ich strukturą a umiejscowieniem sieci tuneli, jak również przeprowadzenie analizy ewolucyjnej zidentyfikowanych tuneli. Dodatkowo wykonano analizę porównawczą metod służących do wykrywania i analizowania tuneli w białkach. Zestawiono ze sobą wyżej wymienione metody oparte na śledzeniu małych cząsteczek wraz z metodami wykorzystującymi podejście geometryczne. W przypadku analizy reakcji enzymatycznej – przedstawiono, że dzięki połączeniu metody śledzenia małych

cząsteczek i analizy lokalnej dystrybucji rozpuszczalnika jest możliwe opisanie różnych funkcji pełnionych przez cząsteczki wody w cyklu reakcji enzymatycznej. Analiza ta została przeprowadzona podczas badania cięcia proteolitycznego pętli, tzw. pętli Z, znajdującej się w receptorze Toll-podobnym TLR8 przez proteazę furynową. Oprócz potwierdzenia podstawowej roli wody w wyżej wymienionej reakcji, jaką jest jej funkcja katalityczna, przedstawiono hipotezę dotyczącą dodatkowych funkcji pełnionych przez cząsteczki wody, w szczególności związanych ze stabilizacją niektórych oddziaływań międzycząsteczkowych lub jako mediatora podczas przenoszenia protonów, lub w procesie dysocjacji kompleksu.

Wyniki prac przedstawione w rozprawie doktorskiej zostały opublikowane w dziewięciu recenzowanych czasopismach naukowych oraz w formie jednego preprintu przesłanego do recenzji.

1. Introduction

1.1. Importance of the subject

The origin and evolution of life is a long-standing subject that has been studied for hundreds, even thousands of years. Over the years, philosophers and scientists have been working intensively, trying to get to the point when and where life emerged, as well as how it evolved. Several theories have been proposed which offer possible explanations for how life may have come up on Earth. Already in ancient times thinkers and philosophers like Aristotle proposed the Spontaneous Generation theory that suggests that life could arise spontaneously from non-living material if such a material contained 'vital heat' – *pneuma* [11]. That was, however, disproved through later scientific experiments. The most famous study refuting the Spontaneous Generation theory was an experiment conducted by Louis Pasteur. He disproved it by boiling beef broth in a special swan-neck flask that deters contamination. When the broth was not exposed to air, it remained sterile and free of microorganisms. When the flask neck was broken and the air was allowed to reach the broth, the fluid became cloudy with microbial contamination. Based on the results Pasteur postulated that life only comes from life [12]. Another theory developed over the centuries and popularised at the beginning of the 20th century by Svante Arrhenius was the Panspermia theory. This theory proposes that life exists throughout the universe and could be transported to Earth through comets or meteorites [13]. In the 1920s, Aleksandr Oparin and John Haldane proposed the Primordial Soup theory which suggests that life emerged from a mixture of organic molecules in the early Earth's oceans (the soup) through the process named abiogenesis. Under specific conditions such as the presence of different energy sources, e.g. lightning or ultraviolet radiation as well as a suitable environment with the necessary building blocks, simple organic compounds could have formed. Over time, these compounds interacted, leading to the formation of more complex (macro)molecules [14]. In 1952, the Miller-Urey experiment supported this theory by demonstrating that organic molecules, including amino acids, could be produced under conditions resembling early Earth's atmosphere. The authors hypothesised that these organic molecules could have served as precursors to life, giving rise to self-replicating molecules and the emergence of cellular life forms [15]. The results of the Miller-Urey experiment motivated other scientists to carry out studies in which life-related molecules were shown

to be abiotically synthesised from different precursors and based on electrical, thermal, chemical, and photochemical energy [16–19]. In the 1960s and 1970s, two other theories were proposed. First, Alexander Clairns-Smith proposed the Clay theory which suggests that minerals, mainly clay, played a role in the assembly and organisation of organic molecules [20,21]. Then, Jack Corliss shared the Hydrothermal Vent hypothesis that assumes that life may have originated around underwater hydrothermal vents where favourable chemical conditions existed [22,23]. Probably the latest of the 'big' theories was the RNA World hypothesis, proposed in the 1980s by Walter Gilbert. It suggests that early life relied on self-replicating RNA molecules which acted as both genetic material and catalysts [24].

Over the years, these theories have evolved, and new results continue to emerge as scientific understanding progresses. Examples of such noteworthy studies are analyses carried out on archived samples of Miller's original experiments. These studies involved the use of either a reducing gas mixture and a specialised apparatus to simulate a water-rich volcanic eruption with lightning or the inclusion of H_2S as a component in the reducing gas mixture. As a result, over 40 different amino acids and amines were synthesised, which demonstrate the formation of biological compounds under possible cosmo-geochemical conditions [19]. In recent years, computational methods have also been used to get insight into the origins of life [25–28]. As technology develops rapidly, it seems that its role in understanding the evolution of biochemical processes may indeed be invaluable. Nevertheless, there is still a lot to discover.

We know famous quotes "The only constant in life is change" or "Everything flows" ("*Panta Rhei"*) by Heraclitus. These quotes emphasise the dynamic nature of life which is characterised by ongoing adaptation, diversification, and transformation. Also, they highlight the ever-changing nature of life's evolutionary processes. From the biomolecular point of view, we can link these quotes with the flow of genetic information and biological traits across generations. The genetic material is passed from parents to the offspring and through various mechanisms, e.g. mutation, recombination, and natural selection, variations arise and propagate. Most of the genes encode amino acid sequences that further form unique proteins which serve a wide range of critical functions. Proteins are created through a process called protein synthesis. This process starts with the transcription of a gene, where the DNA sequence is transcribed into messenger RNA (mRNA) in the nucleus. Transcription is a three-step process that begins with initiation when an enzyme called RNA polymerase recognises the promoter and starts unwinding the DNA. Then, elongation follows and RNA molecules complementary

to the template DNA strand are synthesised. Finally, sequences called terminators signal that the RNA transcript is complete, which leads to the release of the newly formed mRNA molecule. In eukaryotes, the new mRNA is not directly ready for the next stage of protein synthesis. It requires more processing before it leaves the nucleus as mature mRNA. The processing may include removing the non-coding regions (introns) and splicing back together the coding regions (exons), editing, capping and polyadenylation. Such a processed mRNA moves to the cytoplasm, where it binds to a ribosome, an organelle composed of ribosomal RNA (rRNA) and proteins. Then, translation, the crucial part of protein synthesis, begins with the binding of specific transfer RNA (tRNA) molecules to the start codon of the mRNA. Each tRNA carries a specific amino acid, which is activated by attaching it to the corresponding mRNA. The ribosome reads the mRNA codons and catalyses the formation of peptide bonds between the amino acids, elongating the growing polypeptide chain. Translation continues until a stop codon is reached, causing its termination. With that, the ribosome releases the completed polypeptide chain, which further undergoes protein folding and possible post-translational modifications [29]. There are numerous modifications that proteins can go through but, according to recent studies, the five most frequently found are as follows: phosphorylation, acetylation, ubiquitination, succinylation, and methylation [30].

Proteins are involved in almost every aspect of cellular processes, highlighting their indispensable role in maintaining the structure, function, and regulation of whole organisms. They provide structural support to cells and tissues, facilitate the transport of molecules, and participate in cell signalling pathways. Also, they act as enzymes, catalysing biochemical reactions, function as hormones to regulate physiological processes, contribute to the immune response by producing antibodies, and offer defence and protection against a broad range of pathogens. Finally, they control gene expression and regulation. Proteins undergo evolutionary diversifications, with genetic variations giving rise to new protein sequences, structures, and functions over time. They further exhibit functional adaptation, adjusting to environmental conditions and selective pressures to maintain their optimal functionality and activity. Proteins also show structural flexibility, transitioning between various conformations to effectively carry out their biological functions. Moreover, they are engaged in a complex and dynamic network of interactions [29]. All of the above-mentioned examples can be considered part of the evolution of life. Therefore, it is of great interest to the scientific community to investigate the molecular basis of these processes.

Like other processes in the world, evolution must have been embedded in some environment. There are several compelling arguments supporting the hypothesis that life began to evolve in an aqueous milieu. Water is an abundant and ubiquitous substance found throughout the planet Earth. It is present in large reservoirs such as oceans, seas, and lakes, as well as in rivers and underground. The immense diversity of life in those aquatic environments also supports the idea that water was and is crucial for the evolution of life. Aquatic ecosystems provide a wide range of habitats that have fostered the development and proliferation of diverse species. Also, the oldest evidence of life on Earth, which can be found in the form of fossilised microorganisms, indicates that life existed in ancient water environments [31]. These findings suggest that water provided a suitable habitat for the evolution and survival of early life forms. The evolutionary history revealed numerous connections between organisms that indicate a common origin in water. The earliest evidence is the evolution of single-celled organisms into multicellular aquatic life forms [32,33]. Also, plants made one of the earliest transitions, with green algae evolving into terrestrial plants [34]. The transition of animals from water to land is exemplified by the evolution of amphibians from fish ancestors, with tetrapods (including amphibians, reptiles, birds, and mammals) evolving various adaptations, e.g. limbs and lungs. Genetic analyses reveal close relations between the genetic material of tetrapods and lobe-finned fish, highlighting common ancestry. Anatomical parallels additionally reinforce this evolutionary connection [35,36].

Other convincing arguments that water is the key to the evolution of life can be found across the fields of biochemistry and biophysics. Water molecules serve as a crucial solvent and medium for biochemical and metabolic reactions. As Albert Szent-Györgyi, a Hungarian biochemist stated: “Water is life’s *mater* and *matrix*, mother and medium. There is no life without water” [37]. Most cells contain approximately 65% water by volume and 70% by mass, and living organisms have developed intricate mechanisms for appropriating and conserving water [38]. Water’s properties make it an excellent solvent for a broad range of both organic and inorganic compounds. Many biological molecules, e.g. amino acids, nucleotides, and carbohydrates (sugars) dissolve in water, enabling the reactions to occur and form the basis of various biochemical processes. The majority of enzymatic reactions occur in aqueous environments, and water helps transport nutrients, ions, and waste products throughout the body. It facilitates cellular functions, supports metabolic processes and maintains homeostasis. It can be said that it is the driving force that is necessary for the functioning of our organism.

Water is one of the most basic chemical compounds - it consists only of one oxygen atom and two hydrogens. Yet, it possesses several important characteristics contributing to its great role in biological systems. Water is a polar molecule, with a bent molecular structure that gives rise to its dipole moment and the ability to form hydrogen bonds not only with other water molecules but also with a variety of different (macro)molecules. These hydrogen bonds are essential for various biological processes. Moreover, the hydrogen bonds enable water to exhibit both cohesive and adhesive properties, resulting in high surface tension, capillary action, and the capacity to dissolve a wide range of substances. Since water molecules are associated with each other quite tightly, water exhibits relatively high values for melting and boiling points. Furthermore, water's high heat capacity allows it to stabilise temperatures and buffer their fluctuations [39–41]. These properties contribute to the maintenance of stable conditions for life.

It is universally accepted that water molecules contribute to governing the structure, stability, dynamics, and function of biomolecules, especially proteins. Water molecules not only contribute chemically to the catalytic function of proteins, but also play a physical role in the folding process [42–45]. The adoption of a properly folded structure is usually crucial for protein to perform its biological functions. Specifically, hydrophobic interactions play a central role in folding, while hydrogen bonds make a large contribution to protein stability [46–48]. This is supported by the fact that water molecules tend to exclude the non-polar side chains of amino acids from their surroundings. This leads the hydrophobic side chains to collapse and form a tightly packed core within the protein structure. Even more than 80% of the non-polar residues may be found in the core of a protein [44]. Besides hydrogen bonds and hydrophobic interactions, there are also other forces contributing to protein stability – disulphide bonds, ionic bonds, van der Waals forces [48].

Water is essential for biomolecules to maintain their stability through various mechanisms. For instance, it forms a hydration shell (layer) around proteins, with ordered water molecules interacting specifically with the protein surface through hydrogen bonds that stabilise the protein structure. Such proper stabilisation is often a key to maintaining protein activity [49–51]. Water molecules also ensure the distribution of electrostatic interactions which maintains a cohesive environment around the macromolecules [40]. It has been shown that thermodynamic changes in an aqueous environment combined with alterations of ordering the water molecules can lead to protein denaturation or/and unfolding [52–54]. In general, any disruption of the hydrogen bonding interactions can destabilise the protein

structure and result in loss of function. Therefore, ensuring stable conditions is crucial. Various studies have confirmed that water molecules from the hydration shell exhibit dynamic properties and may exchange positions either with solvent molecules from the bulk or with water molecules found within the protein core [49–51]. The dynamics of water molecules positioned around and within the structure vary, depending on their location and specific conditions. In the bulk, molecules can move relatively freely, while within the protein's hydration shell, water molecules are more organised, leading to a slower exchange time. It is quite common to find water molecules trapped within internal cavities and pockets of proteins. These internal water molecules can directly interact with the protein's backbone and side chains, and they may even form clusters within hydrophobic cavities. The resistance time of such buried water molecules is much longer compared to those in the first hydration shell. Nevertheless, water molecules from the interior can escape into the bulk and be replaced by those from the hydration shell. If these water molecules are not exchanged, but they are rather bound, they play essential roles in interactions with protein side chains that would otherwise not occur. Water molecules in such fixed positions can be considered integral components of the macromolecule's tertiary structure. Besides, it has been demonstrated that internal water molecules can participate in many processes, such as proton transfer reactions, catalysis, redox processes, and ligand binding [55]. Regarding the binding events, water molecules may mediate protein-ligand complex formation through hydrogen bonds, which also influences the overall stability [56,57]. On the other hand, the displacement of particular water molecules from the binding site during ligand association may impact the affinity and control the enthalpy-entropy partitioning. In some cases, the displacement of such thermodynamically unfavourable water molecules can substantially enhance the ligand's affinity [58,59]. The influence of water molecules on allostery has also been discussed [60].

The dynamics of proteins in solution is a broad topic and includes a wide range of processes, such as backbone and side-chain fluctuations, interdomain motions, as well as global rotational and translational diffusion. Exploring these mechanisms not only deepens our understanding of protein functioning but also provides a more extensive insight into biological processes as a whole. Numerous examples from the literature illustrate this field. The first articles, in which processes related to proteins' dynamics were investigated, emerged approximately in the second half of the previous century. This was related to the availability of methods, both experimental and computational, enabling such research. With the development of techniques, the number of studies being conducted increases substantially.

Here, I wanted to highlight only a few review articles that broadly approached the topic of protein dynamics and functioning, emphasising the role of water molecules.

The first one is the work of Bellissent-Funel *et al.* [45], which I have already referred to in this thesis. The authors gathered details not only about the dynamics of water at a protein interface and in the interior of globular protein but also regarding the coupling of protein/water dynamics and modelling protein hydration. Further, they described the relationship between water structure and protein function, taking into account the hydrogen bond networking at biological surfaces and protein hydration in the gas phase. Also, they investigated the effects of pressure and temperature in protein unfolding. Finally, they reported the role of water in a (bio)molecular crowded environment, as well as in membrane channels. Another interesting review was published by Grimaldo *et al.* [61]. The authors outlined four classes of dynamical processes occurring in proteins, from the largest supramolecular length scale to the smallest atomic length scale, all of which are linked and can contribute to the protein function. They were as follows: diffusion of the entire protein, fluctuations of protein domains, localised and confined diffusive relaxations, and vibrational dynamics. Around the same time, Maurer and Oostenbrink reviewed the significance of water mostly in protein hydration and ligand recognition [62]. Their primary emphasis was on elucidating how water molecules influenced various aspects, including providing mechanical support to proteins, facilitating thermal coupling, acting as a dielectric screen, facilitating mass and charge transport, and competing with ligands for binding site occupancy. Additionally, an engaging Perspective article by Spyrakis *et al.* was published in which the various roles of water in the protein matrix, especially in the context of drug discovery were described [63]. The authors described in detail the role of thermodynamics of water in biological systems. They proposed a set of terminology that describes water molecules as being ‘hot’ and ‘cold’ which they have defined as being easy and difficult to displace in the analysed systems. Such a dual role of water molecules underlines the complexity of water’s involvement in biomolecular interactions, where their location, arrangement, and interactions can either promote or hinder molecular associations.

Considering the huge amount of studies referenced in the articles mentioned above, it is undeniable that water plays a pivotal role in the investigation of biological processes. As can be also concluded, while investigating the role of water molecules in biological systems, and in general the dynamics and function of biomolecules, one of the key elements is to select appropriate methods and techniques for that. A short overview of the most important methods and techniques used to study biomolecular systems that take into account the role of water molecules is provided in the next chapters.

1.2. Studying the role of water molecules in biomolecular systems

Several approaches can be used while studying the importance of water molecules in biomolecular systems. First of all, water molecules can be investigated experimentally using various techniques. For instance, X-ray crystallography (X-ray) reveals the positions of stable water molecules by analysing the electron density maps obtained from the crystallised biomolecule. Another way is to use the nuclear magnetic resonance spectroscopy (NMR) which allows studying the interactions and dynamics of water molecules by analysing their NMR signals. Also, infrared radiation (IR) spectroscopy can analyse the vibrational modes of water molecules and get insight into their hydrogen bonding interactions and structural dynamics. A different technique is cryo-electron microscopy (Cryo-EM) which provides insights into the positioning and organisation of water molecules around biomolecules through the visualisation of high-resolution structures preserved in vitreous water. There are also neutron scattering techniques, such as neutron diffraction and small-angle neutron scattering which can supply information about the positions and interactions of water molecules in biomolecular systems, benefiting from the sensitivity of neutrons to hydrogen atoms. Details about more advanced techniques which could be used in the context of studying the dynamics of proteins in solution were described in the review by Grimaldo *et al.* [61].

The standard experimental techniques may have several shortcomings while using them for studying the role of water in biomolecular systems. First, water artefacts can be a challenge when attempting to accurately determine the structure and interactions of dynamic water molecules, especially in X-ray crystallography. In output files (PDB files) resulting from crystallography, only a small fraction of water that was present in the crystal is included in the atomic model. Only a few of these water molecules are found in the protein core. Most of the water molecules cover the protein surface, by forming interactions with both hydrogen-donor and hydrogen-acceptor protein atoms and with other water molecules (and occasionally with other molecules) [64]. However, it is difficult to confirm that they accurately reflect e.g. the hydration shell surrounding the protein surface. In their study, Gnesi and Carugo examined nearly 10 000 protein crystal structures and showed that the number of detectable water molecules depends on the following variables: crystallographic resolution, residual (R) factor, percentage of solvent in the crystal, average B factor (Debye-Waller factor) of the protein atoms, percentage of amino acid residues in loops, average solvent-accessible surface areas of the amino acid residues, grand average of hydropathy

of the protein(s) in the asymmetric unit and the normalised number of heteroatoms other than water molecules [65]. Additionally, the high-energy X-ray beams used in crystallography can cause radiation damage, ionising or dissociating water molecules and potentially altering the protein structure. The crystallisation process might also result in the loss or rearrangement of the hydration shell around the protein [66]. Crystal structure obtained in this process is usually a purified, recombinant or synthetic version of a native protein. Both the purification and crystallisation conditions which are often connected with a manipulation of such factors as pH, temperature, precipitant and other additives concentration may cause that the obtained protein deviates from its native state [67]. For protein, sample preparation steps might be a bottleneck, regardless of the technique used.

On the other hand, in NMR spectroscopy, water molecules may pose challenges due to their high concentration, which results in strong signals that can interfere with and overshadow these from the protein. Besides, water molecules can exchange with labile protons in proteins, affecting the signal intensity. This exchange, along with water-protein interactions, can lead to line broadening in NMR spectra, and decreasing resolution [68]. Interactions of water molecules with side chain residues of proteins are also temperature-sensitive, so temperature fluctuations can affect NMR spectra [69,70]. Water influences also the viscosity and diffusion properties of the solution, which can impact relaxation times and distort the results of spectroscopy experiments [71].

In Cryo-EM experiments, the rapid freezing of the sample aims to vitrify water and prevent ice crystal formation. However, the occasional imperfections in vitrification can still lead to ice crystals that deform the protein structure and obscure fine details. Moreover, water molecules can create varying contrast in images due to their differing densities, making them difficult to interpret. The Cryo-EM technique involves imaging in a vacuum, and while freezing protects the sample, there can still be a sublimation of water which may cause structural alterations. Also, due to the inherent flexibility of macromolecules, the water molecules in the hydration shell may not always be uniform across the sample, leading to heterogeneity that makes it difficult to generate a proper three dimensional reconstruction [72,73].

In general, for these mentioned experimental techniques, constraints specific to the size, complexity, and nature of the analysed systems can impact the accurate investigation of the roles of the water within the biomolecular complexes. Additionally, the timescale of the experiments is quite limited, thus, it may hinder a comprehensive understanding of water dynamics and interactions. Also, one of the challenges is linked to the fact that one never is sure if all the water molecules have been identified in the analysed system. In the case of a protein,

if the sequence is available, it is relatively straightforward to verify whether the obtained structure is complete or not and, if necessary, model the missing elements. In the case of water molecules, such information is not easily available. Nevertheless, there are quite a lot of computational tools trying to deal with such a problem. Those tools have been reviewed in an article that I co-authored and which will be discussed later in this thesis.

Taking into account the time scale of processes occurring in biological systems and the willingness to describe the dynamic phenomena occurring in them as precisely as possible, the use of computational (*in silico*) methods becomes a natural choice. One of the widely used methods are molecular dynamics (MD) simulations. During molecular dynamics, the movement and interactions of biomolecules and water molecules are simulated over time, providing insights into their dynamics and behaviour. MD simulations involve solving Newton's equation of motion to simulate the physical movements of atoms over time through deterministic algorithms. The behaviour of biomolecules is modelled based on the use of appropriate force fields which approximate the potential energy of a system as a function of the atomic positions using classical mechanics. The potential energy of the system consists of various components, e.g. bond stretching, angle bending, dihedral angles, and non-bonded interactions, allowing for detailed analysis and understanding of various aspects of biomolecule behaviour. Also, applying force fields allows for the incorporation of empirical data and the known behaviour of biomolecules, making the simulations more realistic [74].
Regarding the solvation effects, there are two common ways to include them in MD simulations: the so-called explicit solvent, and implicit (or continuum) solvent models. The term 'model' is used to describe the set of force field parameters (bond lengths, angles, force constants for flexible models, atomic partial charges, and Lennard-Jones potentials that describe van der Waals interaction and balance the attraction and repulsion forces) necessary to perform MD simulations of water [75,76]. Historically, the implicit solvent model was the first quantitative water model compatible with the atomistic level of description. It treats solvent as a continuum with dielectric and non-polar properties of water. There are two basic types of implicit solvent methods: based on accessible surface areas and based on the continuum electrostatics. One of the goals of introducing these models was to simplify the simulations and reduce computational costs. Thus, implicit solvent models are particularly useful for large-scale simulations or cases where the detailed interactions between the solute and individual solvent molecules are not the primary focus of the study [77]. Explicit models represent each water molecule individually with varying numbers of interaction

sites and parameters. This approach aims to provide a more realistic representation of solute-solvent interactions by simulating the actual molecular dynamics of the solvent molecules surrounding the solute. Most MD simulations are carried out with the solute surrendered by a periodic box of explicit water molecules. Typically, solvent accounts for over 80% of particles in the simulated system, therefore, water-water interactions dominate the overall computational costs of such MD simulations. The most common criteria used to categorise explicit solvent models concerns the number of interaction sites, geometry of the water molecule, polarizability, and flexibility. Generally, these models are empirical models aimed at reproducing many bulk properties in a particular phase. Some of them reproduce protein hydration energies well, while others are better in the prediction of water structure but not that good for hydration free energy [75,76,78]. As for today, none of the water models can accurately reproduce all of its key properties. Therefore, the choice between the models depends strongly on finding the right balance between the level of accuracy and detail required for a particular study, as well as on the computational resources.

Similarly to MD simulations, Monte Carlo (MC) simulations sample different configurations of the analysed system and might be helpful while studying the behaviour of biomolecules and the impact of water molecules. However, MD and MC simulations differ in several aspects. MC simulations are not time-dependent and involve generating random configurations from phase space based on statistical probabilities through a probabilistic algorithm. They can jump to distant points in configuration space, making them efficient for systems with rugged energy landscapes [79]. Nevertheless, water molecules in MC simulations can be modelled using similar approaches as in MD simulations. The typical way is to use a representation of water molecules through explicit solvent models. In MC simulations, random changes can be made to the positions and orientations of water molecules, and the energy change associated with each move is calculated based on the force field used. The action is then accepted or rejected based on a probability criterion, often the Metropolis criterion, which depends on the energy change and temperature [80]. MC simulations can also employ implicit solvent, and this choice (as in the case of MD simulations) depends on the level of details needed for the specific study. There is also a whole range of quantum chemistry methods for studying water at the molecular level. Particularly, they incorporate density functional theory (DFT), Hartree-Fock and post-Hartree-Fock methods, semi-empirical methods, or *ab initio* molecular dynamics. Here, only a brief explanation about the DFT is provided, since they are one of the most widely used.

In DFT Density Functional Theory, properties of biomolecules and water molecules are described through electron density, which is critical for understanding chemical bonding,

reactivity, and electronic properties of the system. With the DFT method, it is possible to predict the equilibrium geometry, essential for the structure and function of biomolecules. DFT accounts for intermolecular interactions like hydrogen bonding, van der Waals, and electrostatic interactions, and computes vibrational frequencies and modes for insights into the dynamics and thermodynamics. Moreover, DFT may be used to examine potential energy surfaces associated with chemical reactions and conformational changes, and when coupled with water molecules, it enables the study of solvation effects. In DFT, water molecules are modelled at the quantum mechanical level, where the electronic structure of water is considered. Each water molecule is treated explicitly with its electrons and nuclei represented, and electronic correlation effects, electron density, and other quantum properties are calculated [81]. This makes DFT computationally intensive, especially for larger systems. This is also why DFT is rather used to study smaller systems, where single water molecules or small clusters of water molecules are involved. When analysing systems such as proteins, hybrid Quantum Mechanics/Molecular Mechanics (QM/MM) are often used. The behaviour of both macromolecule and water molecules is modelled by partitioning the system into a QM region and a MM region. The QM region typically includes the active site of the biomolecule or the other part where electronic changes are significant, while the MM region involves the remaining part of the biomolecule. This approach combines the advantages of the QM aimed at accuracy and MM aimed at speed [82].

Certainly, the computational methods are not without flaws. Using the *in silico* methods, one needs to keep in mind that approximations are inherent and may introduce errors and constraints to the analysis. Sometimes the computational cost can be significant, requiring substantial resources and thereby limiting the scale and complexity of the system that can be analysed. Also, the lack of experimental validation may undermine the reliability of computational results, requiring caution while interpreting the results. Initial conditions and sensitivity to parameters also need to be taken into account, as slight variations can yield different outcomes. Regardless of whether an experimental or computational method is used, awareness of the research question and ensuring that the chosen technique aligns with it is fundamental to conducting effective and meaningful research.

1.3.Motivation

The unprecedented importance of the role of water in various biological processes was my main motivation to tackle this topic during my doctoral studies. In particular, I became interested in investigating the role of water molecules in proteins. In the *Introduction*, I mentioned that both experimental and computational methods can be used for this type of research but the choice relies on the scientific problem to be solved. Since my interests were mainly linked with investigating the dynamic processes and regulation of proteins at the level, for which experimental methods often lack the required precision or cannot be used, the application of computational methods was the rational choice. During my research, I focused particularly on proteins due to their importance as molecular targets in drug design, their utility in rational biomolecule engineering, and their role in the signalling pathways.

Regarding drug design, proteins are often molecular targets in various diseases. Many disorders are caused by aberrant protein activity or dysfunction. Targeting specific proteins can lead to the development of effective treatments. Nevertheless, other strategies can be also applied. For instance, other proteins in the same pathway, either upstream or downstream, can be targeted to restore normal pathway function. Also, targeting regulatory proteins or those with similar functions can work well. By understanding the structure and function of proteins involved in diseases, regardless if directly or indirectly, researchers can design potential drugs in a better way, for example, to selectively bind to these biomolecules, modulate their activity, and restore normal cellular function.

As concerns the protein engineering, by understanding the structure-function relationship of proteins, scientists can design and engineer these macromolecules to enhance their properties or create entirely new functionalities. Rational engineering allows for the targeted optimisation of protein properties, such as enzymatic activity, stability, substrate specificity, and binding affinity. It has great potential in diverse fields, enabling the development of novel therapies, diagnostic tools, industrial applications, and many others.

About the signalling pathways, proteins are often involved in their regulation. They can act as molecular switches, amplifiers, and effectors by detecting, transmitting, and modulating signals. Proteins use various mechanisms to control the signalling, ranging from conformational changes, allosteric regulation, and protein-protein interactions through enzymatic reactions to post-translational modifications. Therefore, understanding the basis of these mechanisms

is crucial, especially if disturbances translate into potential disorders in organism functioning. Taking into consideration all the information, it became evident to me that I would like to incorporate the investigation of the roles of water in biomolecules in order to better understand the mechanisms behind drug design, protein engineering, and protein regulation in the analysed systems.

1.4. Information about the conditions of the computational experiments

The vast majority of the analyses presented in this thesis were based on the results obtained from molecular dynamics simulations. In all the studies, MD simulations were carried out using the AMBER package (versions 14, 18, and 22, respectively) [83–85]. To achieve a quite detailed and accurate representation of solvent dynamics and interactions with the analysed proteins, the transferable intermolecular potential with 3 points (TIP3P) explicit solvent model [86] was used throughout all simulations. I am aware that some of the properties of the water, e.g. density, diffusion coefficient, the heat of vaporisation, melting point or dielectric constant could be underestimated by the choice of the TIP3P water model [75,76]. However, this model is computationally efficient, allowing for carrying out longer simulations and analysing larger systems, which was crucial in the case of systems studied. As a force field for all the analysed proteins, the ff14SB force field, implemented in the AMBER package, was used [87]. First of all, this force field is widely utilised in MD simulations due to its enhanced accuracy, especially in representing backbone and side-chain torsion angles, which is vital for proper protein folding and dynamics. Its balanced treatment of polar and non-polar interactions captures the intricate balance governing protein structure and function. Besides, according to the AMBER reference manual, the ff14SB force field is intended for use with the TIP3P water model. Finally, the choice of running all the simulations with the explicit solvent model instead of the implicit solvent was caused by the fact that the software (AQUA-DUCT; AQ) was primarily used for the post-processing of the results from MD simulations. This software was developed to track and analyse particular water molecules within the system. Since an implicit solvent model incorporates the solvent effects into the force field as a mean-field approximation and the solvent is represented by a continuous medium with average properties, it would not be possible to obtain detailed information about the analysed systems.

2. Aims of the doctoral thesis

Given the importance of the topic, four main objectives were defined.
They were as follows:

1. To survey available computational tools that incorporate water molecules for studying macromolecules' properties and to participate in the development of a new version of software enabling the structural and functional analysis of macromolecules from the 'intramolecular voids' perspective.
2. To demonstrate applications of water and (co)solvent molecules in computational drug design-related studies.
3. To explore applications of water molecules in protein engineering and macromolecule regulation.
4. To characterise the role of water molecules in proteolytic cleavage enzymatic reaction, a process related to the regulation of a signalling pathway.

To fulfil the first objective, I participated in the preparation of the review article (**Paper 1**) that gathered and summarised information about various software and tools that can be incorporated for studying the macromolecules' properties using water molecules.
More importantly, I was part of the team (with members of Tunneling Group) which worked on and developed the new version of the AQUA-DUCT software (AQ 1.0), dedicated to performing the structural and functional analyses of macromolecules using a novel, intramolecular voids perspective (**Paper 2**). The vast majority of analyses carried out in subsequent studies were made possible by the new functionalities introduced in the AQ 1.0 software.

To accomplish the second goal, I was involved in studying and describing the applications of water and (co)solvent molecules in computational drug design-related studies for the following molecular targets:

- the main protease of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2 Mpro) (**Paper 3**),
- the SARS-CoV-2 Mpro and a panel of selected proteases (**Paper 4**),
- the human soluble epoxide hydrolase (hsEH) (**Paper 5**).

To achieve the third objective, I participated in studying the soluble epoxide hydrolase (sEH) family and describing the structure-function relationship between sEH structure and their tunnel network (**Paper 6**). I was also involved in an investigation of the evolution of tunnels (**Paper 8**) identified in the selected members of the sEH family. Additionally, I was involved in comparing geometry-based and small-molecule tracking methods for tunnel identification (**Paper 7**).

Regarding the fourth goal, firstly, I outlined the problem with studying the proteolytic cleavage reaction in Toll-like receptors (TLRs) (**Paper 9**). Then, I participated in characterising this process in the TLR8. The aim of this study was not only to propose the putative reaction mechanism of the proteolytic cleavage of the Z-loop in TLR8 but also to explore the role of water molecules in the course of this reaction. The results of the later study have been published as a preprint (**Preprint 1**) which currently is under consideration for publication in a peer-reviewed journal.

3. Summary of the doctoral research

In the following sections, I have presented the most important results and conclusions of works carried out during the doctoral studies. Detailed information about the research methodology can be found in particular articles. As most of the studies conducted during my doctoral research were part of bigger projects, in which also other people were involved, I decided to use the term 'we' when describing the results. All the results (with one exception) were already published in peer-reviewed journals. For each article, I provided a detailed description of my contribution. Descriptions of the individual contributions of other co-authors are provided together with the papers, as the attachment.

3.1. Analysis of macromolecule structure, properties, and functions with the use of water molecules

As already stated in the *Introduction*, water molecules are of special importance in maintaining biomolecules' structure, stability, dynamics and functions. Throughout the years, the growing awareness of these facts has translated into the development of different software and tools that incorporate water molecules into the analysis of the behaviour of various types of macromolecules.

In **Paper 1** entitled "Applications of water molecules for analysis of macromolecule properties", together with colleagues from the Tunneling Group, we reviewed available computational methods that employ water molecules to analyse the properties and structural dynamics of macromolecules. The article was divided into three subsequent parts. The first part was focused on describing software for analysing protein hydration, while the second section was dedicated to providing an overview of tools for detecting water sites and analysing ligand-binding events. Finally, the third part was aimed at reviewing software for tunnel detection and various transportation phenomena (**Figure 1**).

My contribution was in gathering information and providing descriptions regarding the applicability and functionality of software reviewed in the second and third parts of the article. Overall, I participated in organising data, writing the manuscript, reviewing and editing the final version, and providing answers to the reviewers' comments.

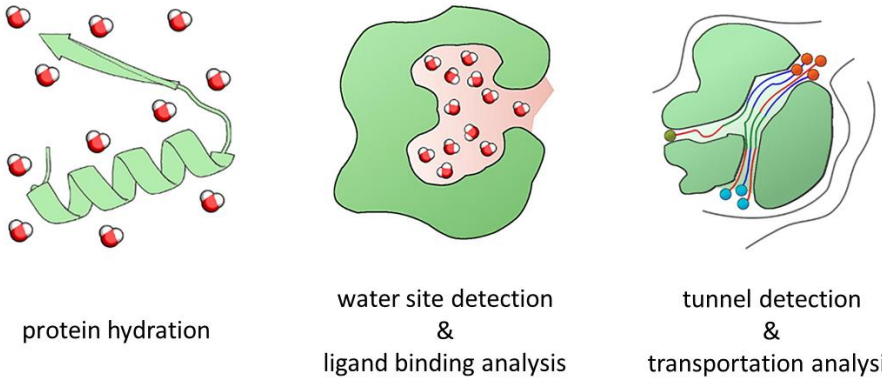

Figure 1 Different applications of water molecules used for the analysis of the macromolecule properties. From the left: protein hydration, water site detection and ligand binding analysis, as well as tunnel detection and transportation analysis. Figure adapted from **Paper 1** [1] with some modifications.

In the first part of the review, we focused on the role of water molecules within the internal cavities of biomolecules. We identified challenges in detecting water molecules, especially when they are buried inside the protein core and when experimental techniques cannot reliably reflect their number and positions. Given the importance of the exchange between internal water molecules and those from the bulk for the proper functioning of biomolecules, we highlighted the necessity to accurately fill the internal cavities with water molecules. We reviewed three groups of methods for predicting the positions of water molecules within biomolecule structure: i) docking-based, which are fast and provide accurate positioning of water molecules as determined by crystallography; ii) reference interaction site model, RISM-based, which calculate solvent distribution and are more accurate for complex systems; and iii) similarity-based, which detect conserved water molecules inside cavities by superimposing structures. Even though we noted that those methods have some limitations in accounting for flexibility and conformational changes in the target structures, we highlighted that they can provide quite an accurate modelling of water molecules, also in low-quality structures, making them a better starting point for various analyses, especially for studies requiring running simulations.

In the second part, we concentrated on detecting hydration sites important for ligand binding. Water molecules, which mediate interactions between proteins and other molecules, are often found in high-density regions known as water (or hydration) sites. The balance between hydration and dehydration affects protein-ligand binding, involving both entropic and enthalpic factors. We characterised various experimental- and knowledge-based approaches for assessing potential water sites, highlighting their use for both static and simulation-based inputs.

Specifically, we explained how the experimental-based approach uses information on water molecules' location within the crystal structures, along with information on B-factors, to construct a special grid and calculate the energetics of water probes inside the macromolecule's binding site. Additionally, we described how the knowledge-based approach employs the inhomogeneous fluid solvation theory to assess the role of structural water molecules by calculating their contribution to the thermodynamics of protein solvation. We emphasised that both these approaches may significantly aid in drug design by offering more precise results compared to methods that overlook the contribution of water molecules.

In the third part of the study, we focused on the role of water molecules from the perspective of tunnel detection and transportation phenomena. In general, voids like cavities, tunnels, channels and pores are vital for the functioning of biomolecules. Tunnels, for instance, facilitate processes like substrate entry, product egress, and the transportation of ions and water. Since these voids form a dynamic network within macromolecules, their proper detection and description might be quite challenging. The initial tools for tunnel detection employed geometry-based approaches, enabling the identification of empty spaces in protein structures, often using Voronoi diagrams. However, these approaches may have several limitations. For instance, geometry-based methods struggle to capture the flexibility and asymmetry of tunnels and do not effectively account for transportation phenomena such as water flow, especially if the results are based only on a single crystal structure. These limitations are particularly significant when considering an enzyme's activity and selectivity. To address these challenges, other approaches that incorporate the dynamics and role of water molecules, have been introduced. For instance, the Visual Abstractions of Solvent Pathlines method allowed the visualisation of solvent pathways as Bézier curves, while the solvent flux method introduced a solvent concentration gradient and modified water molecule velocities for the analysis of rare events. Yet, neither method has found wide applications. Another tool, trj_cavity, became more popular, since it is capable of identifying both cavities and tunnels, providing time-dependent calculations of their volume and solvent capacity. Nevertheless, it has not been well-suited for tracing ligands or solvent molecules. To bridge the gap between tunnel detection and advanced water flux investigation tools, software such as Watergate and AQUA-DUCT were developed. Specifically, they primarily aim at tracking water molecules (and other small molecules), clustering their trajectories, providing statistical analysis, and offering visualisations. Here, I also want to highlight that, learning from the limitations of existing methods, my colleagues and I in the Tunneling Group have been

developing a novel version of the AQUA-DUCT software, AQ 1.0. Our goal was to enable more advanced analyses, not only regarding tunnel detection and ligand tracking but also concerning cavities and pockets characterisation and water sites identification. More details are provided when discussing **Paper 2**. Overall, we can postulate that using water molecules as probes for tunnel detection offers a more comprehensive understanding of macromolecules compared to geometry-based methods. Tools that utilise water molecules may help in the identification of crucial residues for enzyme activity, selectivity, gating mechanisms, and even the evolution of cavity shapes during simulations, making them valuable for understanding macromolecule properties, facilitating drug design, and contributing to rational protein engineering. Even though it seems that methods for tunnel identification based on the tracking of water molecules tracking may have more applications, geometry-based methods are still more widely used.

Summing up, the increasing interest in studying water-mediated interactions has led to the development of various tools that utilise water molecules to analyse macromolecules. The progress in this area will strongly rely on the collaboration between theoretical biochemists, biophysicists and experimentalists.

In **Paper 2** entitled "AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective", we presented the new functionalities of the AQUA-DUCT software. In particular, we focused on two types of analyses that one can perform using the presented tool - advanced small-molecule tracking and local-distribution analysis. I took part in the preparation and development of the new version of the software, mainly by testing new features that were added to drivers (called *valve* and *pond*) that use *aqueduct* module (the core of the AQ software) to perform relevant analyses. I was working with the main developer of the software on the optimisation of the *pond* driver which enables the local-distribution analysis and facilitates the detection of pockets and hot-spots within the macromolecule's structure. In addition, I was testing distinct modes of analysis (*time-window* and *sandwich*) that have been implemented in the new version of AQ. I was working on the following case studies (presented in the Supporting Information of **Paper 2**): human soluble epoxide hydrolase, potato soluble epoxide hydrolase, and haloalkane dehalogenase LinB. I was involved in writing the manuscript, preparation of figures and also in reviewing and editing the final version, as well as in providing answers to the reviewers' comments.

The first version of the AQUA-DUCT software, released in 2017, was primarily aimed at detecting tunnels and investigating water flux within the macromolecule structure during MD simulations. As already mentioned, in AQUA-DUCT 1.0, we went beyond identification of tunnels. We focused on the identification of structurally important residues and/or regions of macromolecules, analysis of the evolution of the internal voids (pockets) and hot-spots dynamics, as well as an approximation of free energy profiles of the transportation pathways. In the new version, we put more emphasis on tracking a broader range of particles, including not only water molecules but also other co-solvent molecules, ions and various ligands. By adopting this novel approach of investigating macromolecules from the intramolecular voids perspective using specific 'chemical probes', we were able to successfully characterise functionally relevant compartments while overcoming the difficulties posed by commonly used geometry-based methods.

To better understand the results presented in further studies, a brief explanation of how the software works is necessary (more details are provided in Supplementary Information for **Paper 2**). Generally, the utilisation of AQUA-DUCT software involves running a series of stages. The basic workflow regarding ligand tracking does not differ significantly between the first and subsequent version of AQ and has been encoded in the *valve* driver, which is the flow analysis module. Initially, a list of molecules to be traced is created through the screening of MD trajectory, wherein the identification is made based on their entrance into the user-defined *object* region of the macromolecule, which is typically the active site or another cavity. Subsequently, the *scope* region, encompassing usually the whole macromolecule, is defined as the area for tracing purposes. Following this, coordinates are computed, facilitating the mapping of traceable molecules across frames of the MD simulation. Then, distinct paths are generated for each individual traced molecule (solvent/ligand), considering its entry (ingress) and exit (egress) occurrences from both the *scope* and *object* regions. Each occurrence is treated as an individual phenomenon and is thus represented by a separate path, while the molecule's identification and the number of occurrences are retained. If the separate path has a beginning and an end (either one or both) at the boundaries of the *scope*, it is considered an inlet, i.e. the point that marks where the traceable molecule enters or leaves the *scope*. Groups of inlets can be used to find clusters that are supposed to point to the end of tunnels in the macromolecule. The next stage entails conducting various statistical analyses on the produced paths to extract relevant information and metrics. Finally, the resulting trajectories and analysis outcomes are visualised, thereby providing a comprehensive depiction of the transportation events observed during MD simulation (**Figure 2**).

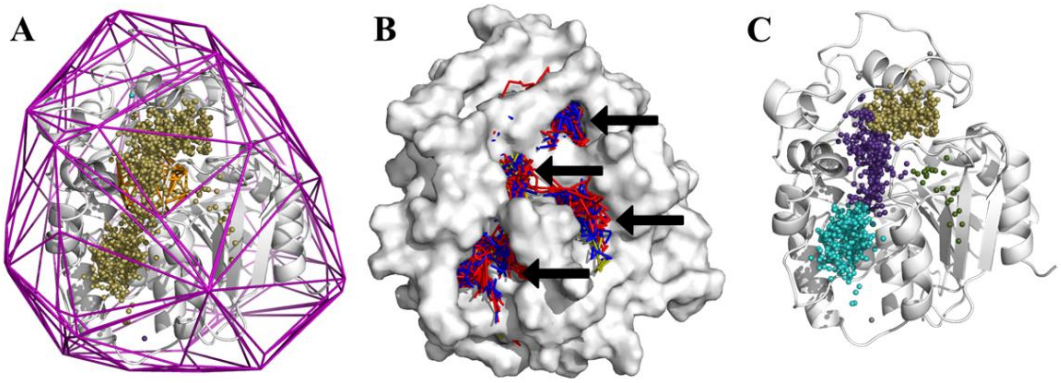

Figure 2 An example of small-molecule tracking analysis performed with AQUA-DUCT 1.0 software. **(A)** Visualisation of water molecules (inlets) that entered and left the protein. Inlets are shown as yellow spheres, while scope and object regions are shown as purple and orange shapes, respectively. **(B)** Visualisation of paths (shown as lines) of tracked water molecules within the protein together with the identified entries to the active site pocket (marked with arrows). Paths are colour-coded as follows: red – parts of paths that entered the protein, blue – parts of paths that left the protein. Additionally, the green colour indicates parts of paths that remained in the object region, and the yellow colour indicates parts of paths that left the object, explored the interior of the protein and went back to the object (these two types of the paths not visible in **Figure 2B**). **(C)** Results of the clustering of inlets. The division of inlets into clusters corresponds to the identified entries to the active site pocket. Figure adapted from [88] with some modifications.

Most of the new functionalities of the AQ software have been encoded in the *pond* driver, which is the module dedicated to conduct distribution and energy analyses. *Pond* performs analyses of results obtained from the *valve*, including pocket analysis, hot-spots detection, and energy profile calculations. Pockets are calculated by analysis of paths and by the construction of a regular grid and computing the density of traced molecules for each grid cell. Grid cells with non-zero density are used for pocket detection which can be further partitioned into areas of a different overall distribution of traced molecules. Further analysis of the distribution of densities in the grid cells allows for finding points of the highest local density. Those points are considered to be hot-spots, i.e. regions of particular importance where traced molecules might be either attracted by favourable interactions with the surrounding residues or trapped for a considerably long time. Such an approach to finding the hot-spots is similar to the grid-based method for predicting the water sites which are of special interest for ligand binding (**Figure 3**). Besides, *pond* drivers can estimate the energy profile of user-defined paths by using the computed density of traced molecules. Such an estimation of free energy is done according to Boltzmann's inversion and can provide relevant information e.g. about the energetical barrier that a particular molecule needs to overcome. All these new functionalities can be used in various applications, starting from drug design (e.g. by providing

relevant descriptions of hydrophilic/hydrophobic regions in the protein core), through protein engineering (e.g. by identification of functionally important residues like gates), to chemical reactions (e.g. by finding the catalytic water molecules or describing the transportation of the substrate/product).

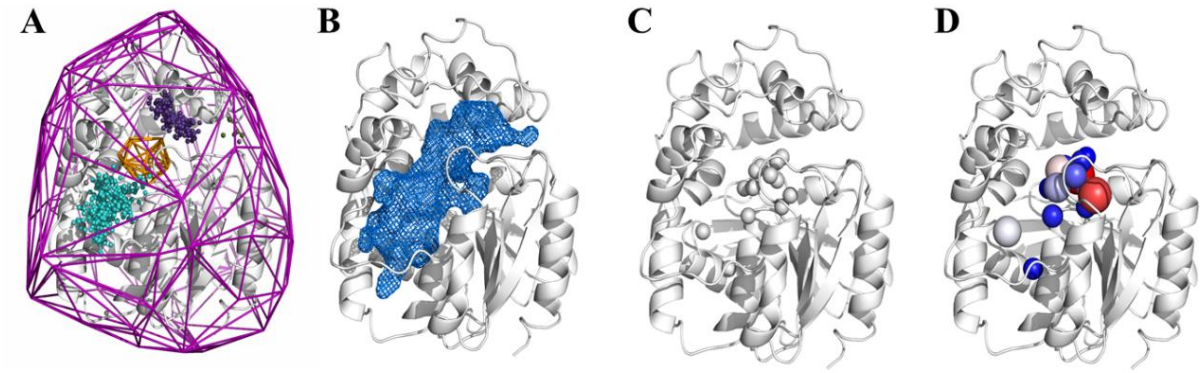

Figure 3 An example of local-distribution analysis performed with AQUA-DUCT 1.0 software. **(A)** Visualisation of clusters of inlets. **(B)** Visualisation of the pockets identified within protein's interior. Pockets are shown as blue mesh. **(C)** Visualisation of regions with a higher density of water molecules (hot-spots). Hot-spots are shows as grey spheres. **(D)** Visualisation of hot-spots reflecting their increasing density – size (from smallest to largest) and colour of spheres (from blue to red). Figure adapted from [88] with some modifications.

Case studies from the Supporting Information of **Paper 2**, in which I was directly involved, included:

i) presentation of the difference between the flow of water in a structure of haloalkane dehalogenase LinB with introduced cysteine bridge in oxidative and reductive conditions,

ii) detection of hot-spots in potato epoxide hydrolase with different co-solvents and detection of key amino acids important for catalytic activity, binding cavity shape, gating residues, and potentially important residues in the main tunnel in the same enzyme,

iii) investigation of the time evolution of the solvent-accessible volume in human soluble epoxide hydrolase in a mixture of water with different co-solvents.

Additionally, together with the members of the Tunneling Group, we prepared several tutorials ranging from a basic analysis of solvent trajectory used for identification of the entries/exits to and from the protein core to a complex one, like identification of the key hot-spots in co-solvent (mixed-solvent) MD simulations that can be used as an insight for macromolecule description, analysis, re-engineering, and for drug design.

These tutorials were gathered together and published in the Living Journal of Molecular Science as an article entitled "AQUA-DUCT: Analysis of Molecular Dynamics Simulations of Macromolecules with the Use of Molecular Probes [Article v1.0]" [88]. This article, however, was not included as part of this doctoral thesis, since it was mostly aimed at educational purposes for new users of the AQ software.

In the following sections, I have presented various applications of water (and other co-solvents) molecules in studies related to drug design, as well as in the fields of protein regulation and engineering, and enzymatic reaction. The development of the AQUA-DUCT 1.0 software and its advanced features has been crucial in the subsequent research. Without it, conducting such studies would have been considerably more challenging.

3.2. Applications of water and (co)solvent molecules in computational drug design-related studies

The turn of the year 2019 brought disturbing information about the outbreak of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and the rapidly spreading COVID-19 disease. Since the beginning of 2020, scientists from all over the world have joined the fight against the novel type of coronavirus.
The Tunneling Group team was also involved, primarily, in working on one of the molecular targets of the SARS-CoV-2 virus - the main protease, SARS-CoV-2 Mpro. This protein was chosen because of its significance in the viral life cycle and the availability of its crystal structure. An important aspect is also the fact that we had a new, ready-to-use version of the AQUA-DUCT software, with which we were able to perform the majority of the analyses.

The results of the first analyses were gathered in **Paper 3** entitled "SARS-CoV-2 Mpro as a challenging molecular target for small-molecule inhibitor design". The publication concluded that targeting SARS-CoV-2 Mpro as a therapeutic target is not straightforward and that there are indications that this strategy may not be effective. In the publication, we reported on the findings of MD simulations (including both classical and mixed-solvent) which were enriched by the evolutionary and stability analyses. We made a comparison with a highly similar SARS-CoV Mpro.

In **Paper 3**, I share the first authorship with Karolina Mitusińska. I was responsible for running all mixed-solvent MD simulations and their further analysis with AQUA-DUCT 1.0 software, including the analysis of the intramolecular voids and hot-spots. Also, I was involved in conducting the same analysis (with AQUA-DUCT 1.0 software) for classical MD simulations. Together, we made a comparison of the crystal structures of Mpros from SARS-CoV-2 and SARS-CoV and performed an analysis of the plasticity of their binding cavities. I was also involved in the analysis of the potential mutability of SARS-CoV-2 Mpro. Additionally, I compared the space occupied by covalently bound fragments in the active site cavity (reported that time by Diamond Light Source Group) with an accessible volume computed by AQUA-DUCT. Overall, I was involved in writing the manuscript, preparation of figures and also in reviewing and editing the final version, as well as in providing answers to the reviewers' comments.

Genome annotation analyses revealed the similarity between SARS-CoV-2 and SARS-CoV, particularly in their main proteases, which share over 90% sequence similarity. Structurally, both enzymes (in a monomeric form) consist of three domains and resemble cysteine proteases, although their active site lacks the third catalytic residue. Instead, their active site comprises a catalytic dyad, H41 and C145, and a stable water molecule that forms hydrogen bond interactions with surrounding residues, which corresponds to the position of a third catalytic member. The difference between Mpros is only 12 amino acids, almost all of which are located far from the binding site of enzymes. Only one residue, S46 in SARS-CoV-2 and A46 in SARS-CoV is located in the proximity of the entrance to the binding cavity, more specifically on a C44-P52 loop which is bordering the active site cavity. Usually, such small structural changes are not expected to significantly affect the binding of small molecules. However, as we proved in **Paper 3**, the described flexibility and plasticity of the binding site of Mpro may pose challenges for inhibitor design.

In our work, we used classical MD simulation with water molecules as molecular probes to provide a detailed picture of the main proteases' interior dynamics. We examined four Mpro structures: i) SARS-CoV-2 Mpro apo structure (PDB ID: 6y2e), ii) SARS-CoV-2 Mpro with an N3 inhibitor (PDB ID: 6lu7), iii) SARS-CoV Mpro apo structure (PDB ID: 1q2w), and iv) SARS-CoV Mpro with the same N3 inhibitor (PDB ID: 2amq).

We employed the small-molecule tracking approach to assess the accessibility of the active site pocket, and a local-distribution approach to provide information about the solvent distribution within the proteins' interior. The AQ 1.0 software was used to analyse the flow of water

molecules through the binding cavity in each analysed Mpro and calculate the maximal accessible volume (MAV) of this site. Surprisingly, despite their high similarity, the MAV of the binding cavities differed significantly between SARS-CoV-2 and SARS-CoV Mpros, with the first having a smaller maximal volume (**Figure 4A** and **B**). Also, we noticed that both proteins reduced their MAV upon inhibitor binding. Further, we investigated the flexibility of the main proteases' binding cavities by analysing the movements of loops surrounding their entrances. We observed that the C44-P52 loop in SARS-CoV Mpro was more flexible than the corresponding loop in SARS-CoV-2 Mpro, while the adjacent loops in the structures co-crystallised with inhibitors showed mild flexibility. The observed flexibility suggests that the presence of an inhibitor may stabilise the loops surrounding the active site. Moreover, the flexibility of the SARS-CoV Mpro binding cavity poses challenges for traditional virtual screening approaches and drug design. In general, we found out that the binding cavity's flexibility and alterations in ligand binding may pose obstacles to drug design efforts. To identify potential binding/interacting sites in Mpro structures, I performed the mixed-solvent MD simulations with six different co-solvents: acetonitrile, benzene, dimethyl sulfoxide, methanol, phenol, and urea. The co-solvents represented various chemical properties and functional groups, allowing for the identification of regions where they were trapped or caged within the protein structure. The use of mixed-solvent MD simulations has become more and more popular in drug design research since they can improve the accuracy of potential binding site detection by considering diverse chemical environments. These simulations are for the exploration of alternative binding sites, assessment of binding affinity, and study of solvent effect on protein-drug interactions [89,90]. Here, two types of hot-spots, local and global, were identified, providing complementary information about potential binding sites (water and co-solvent sites) within the active site region and the whole protein, respectively. The distribution of global hot-spots from different co-solvents revealed specific interactions with particular regions of the analysed proteins. The active site region and the region involved in the Mpro dimerisation showed the highest density of hot-spots (**Figure 4D**). Insight from the local hot-spots distribution once again underlined the differences in binding sites' plasticity (**Figure 4C**). Interestingly, the hot-spots of SARS-CoV Mpro were also found close to the C44-P52 loop, which (as was previously mentioned) may play a role in regulating access to the active site. It may also imply that a potent inhibitor of SARS-CoV and/or SARS-CoV-2 Mpro(s) needs to be able to open its way to reach the active site to bind within this cavity.

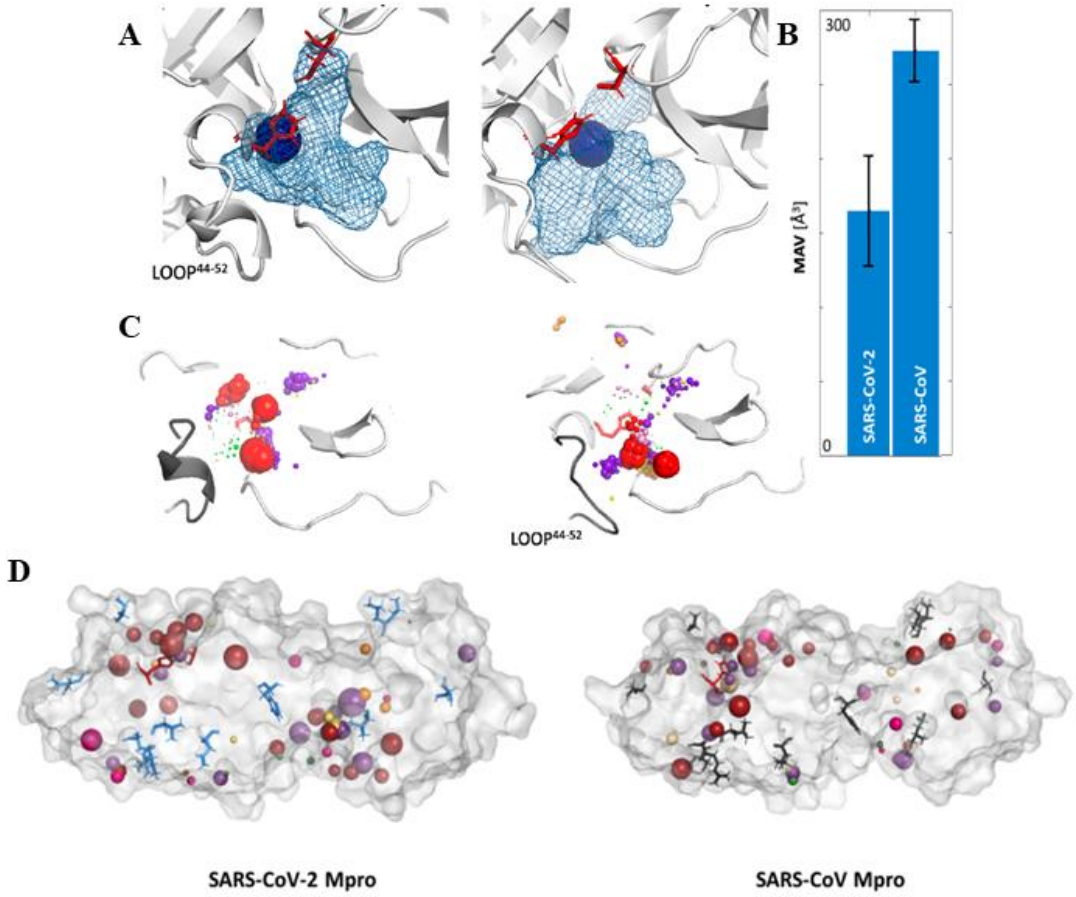

Figure 4 Differences in binding cavities and within the entire structures for SARS-CoV-2 and SARS-CoV main proteases. **(A)** and **(B)** Maximal accessible volumes for the solvent molecules (shown as blue mesh) in the binding cavities of apo structures of SARS-CoV2 and SARS-CoV Mpros together with the position of the identified water hot-spots with the highest density (blue sphere). **(C)** Localisation of co-solvent hot-spots identified in the binding cavity of both enzymes. **(D)** Localisation of co-solvent hot-spots identified within the entire structures of both enzymes. Hot-spots of individual co-solvent are represented as spheres, and their size reflects the hot-spots density. Co-solvent hot-spots are colour-coded as follows: acetonitrile - orange, benzene - red, dimethyl sulfoxide - green, methanol - yellow, phenol - pink, and urea – purple. Amino acids that differ between the main proteases of SARS-CoV-2 and SARS-CoV are marked as blue and black sticks, respectively. Figure adapted from **Paper 3** [3] with some modifications.

The last part involved the analysis of the potential mutability of the SARS-CoV-2 Mpro. To do that, we performed correlated mutation analysis of multiple sequence alignment, together with the analysis of the contribution of the identified differences between the SARS-CoV-2 and SARS-CoV Mpros to protein stability, as well as the prediction of possible mutations caused by the substitution of a single nucleotide in the mRNA sequence of SARS-CoV-2 Mpro. The analyses confirmed that within viral Mpros, evolutionary-correlated residues are located throughout the structures and supported our hypothesis that mutations, even distant from

the active site, can impact Mpro binding properties. The analysis of the potential mutability of the C44-P52 region indicated that further mutation may occur during the evolution, which in consequence could alter the affinity between Mpro and its ligands. Moreover, we identified potential mutations in the binding cavity and catalytic dyad that suggest the possibility of drug resistance development, highlighting the challenge of targeting SARS-CoV-2 Mpro as a viable molecular target for coronavirus treatment.

After the publication of the results described above, scientists from the Computational Pharmacy research group, based in the Department of Pharmaceutical Sciences, University of Basel, contacted us to propose carrying out joint research related to the SARS-CoV-2 Mpro. Soon, we established a successful collaboration between both institutions which also resulted in the publishing of several research articles. On behalf of the Tunneling Group, I was responsible for coordinating the cooperation with the Computational Pharmacy research group. In this doctoral thesis, I included one of the publications (**Paper 4**) entitled "Computational Selectivity Assessment of Protease Inhibitors against SARS-CoV-2".
In this article, we used molecular dynamics simulations protocols (classical and mixed-solvent), together with molecular docking, and toxicity profiling to assess the selectivity of more than 30 reported non-covalent inhibitors of SARS-CoV-2 Mpro against eight different proteases and 16 anti-targets. The main aim of the study was to assess the potential risk of off-target binding, which can lead to drug-induced toxicity and safety issues.

In **Paper 4**, I share equal authorship with André Fischer, Manuel Sellner and Karolina Mitusińska. We were all involved in the selection of the panel of proteases for carrying out the research. Then, we divided the tasks, performing them according to our competencies and taking into account the possibility of access to dedicated software. Together with Karolina Mitusińska, I was involved in the analysis of the similarity of the selected proteins, especially in the context of their active site. Then, we performed all the classical MD simulations together with the water molecules tracking and hydration sites identification. Additionally, I was involved in the analysis of the co-solvent sites, identified based on the mixed-solvent MD simulations carried out by colleagues from the University of Basel. They also performed molecular docking and toxicity profiling. Altogether, we assessed the selectivity of analysed proteases, taking into account different perspectives. I was involved in writing the manuscript, preparation of figures and also

in reviewing and editing the final version, as well as in providing answers to the reviewers' comments.

We selected the set of proteases based on their catalytic residues, sequence and structural similarity, availability of crystallographic structures, as well as their pharmacological and physiological relevance. This panel included the following proteases: SARS-CoV-2 Mpro, SARS-CoV Mpro, caspase-3, factor Xa, cathepsin G, ubiquitin carboxyl-terminal hydrolase isozyme L1 (UCHL1), prostasin, thrombin, and chymase. Although the selected proteases (except for SARS-CoV Mpro) had different overall folds and low sequence similarity to SARS-CoV-2 Mpro, their active site cavities exhibited a substantial degree of similarity, both regarding their catalytic residues, but also the composition of the active site cavities. Even though some proteases had different electrostatics potentials of the binding cavities, their overall similarity was remarkably high. We conducted further characterisation and comparison of selected proteases by analysing their hydration sites and small-molecule binding hot-spots using different molecular probes - water, acetonitrile, isopropanol, and pyridine. Similarly to the previous publication, our objective was to characterise potential binding sites with relevance to structure-based drug design, particularly in fine-tuning the selectivity profile of protease inhibitors. In the case of SARS-CoV-2 Mpro, we identified two hydration sites near the catalytic residue H41, accompanied by a small-molecule hot-spot for acetonitrile. We did not find any hydration site near the catalytic residue C141, although pyridine and isopropanol hot-spots were observed in that vicinity. For SARS-CoV Mpro, we identified more hydration sites, potentially due to the increased flexibility compared to SARS-CoV-2 Mpro, which we already mentioned in **Paper 3**. Notably, we observed differences in small-molecule hot-spots between the two main proteases. Caspase-3 demonstrated two unique hydration sites, while UCHL1 displayed three within their active site cavities. For those enzymes, we also spotted the multiple organic probes gathered within the unique region, not observed in any other proteases. Despite sharing the same catalytic mechanism, caspase-3 and UCHL1 exhibited some differences regarding the hydration and small-molecule sites, compared to the viral Mpros. Factor Xa displayed two conserved hydration sites corresponding to those in prostasin and chymase. The absence of one of these sites in the viral Mpros suggests its potential involvement in ligand specificity. Co-solvent densities among factor Xa, cathepsin G, thrombin, and chymase showed high similarity, with additional densities for acetonitrile and pyridine. Comparing these enzymes to SARS-CoV-2 Mpro, we observed preferences for acetonitrile in specific regions. Cathepsin G and thrombin shared one hydration site located

near their catalytic serine, as in the case of other serine proteases, but lacked the common site observed in the viral main proteases, indicating the possibility of selective binding through the displacement of this water molecule. Thrombin and chymase exhibited the highest number of water hot-spots among the enzymes studied. Notably, chymase displayed high accessibility to co-solvents. At this stage, it was difficult to make strong conclusions about which proteases might pose the greatest challenge in possible off-targeting since the differences in both hydration and small-molecule were quite substantial. Therefore, we used a set of experimentally verified ligands for all nine proteases to assess the selectivity of reported (back then) SARS-CoV-2 Mpro inhibitors. Cross-docking enabled comparing the docking scores of those inhibitors with the native ligands of each protease. Cathepsin G showed the highest overlap with the SARS-CoV-2 Mpro inhibitors, suggesting a risk for off-target inhibition. Additionally, chymase, UCHL1, and SARS-CoV Mpro, also exhibited significant score overlaps. Interestingly, when ligands of respective targets were docked to SARS-CoV-2 Mpro, the prostasin inhibitors displayed better docking scores compared to native inhibitors of SARS-CoV-2 Mpro. Also, the distribution of scores indicated similar (but more subtle) trends for ligands of thrombin, factor Xa, chymase, and caspase-3. To further assess the potential toxicity and undesired effects of SARS-CoV-2 Mpro inhibitors, colleagues from the Computational Pharmacy research group developed the VirtualToxLab. This tool was aimed to evaluate the interactions of inhibitors with 16 anti-targets related to endocrine disruption, cardiac adverse effects, and metabolism. Among the analysed compounds, four (nelfinavir, lopinavir, pimozide, and baicalein) were identified to have the highest binding affinity toward more than half of the analysed proteases. These compounds, together with stereoisomers of bepridil were also predicted to interact with a large group of known anti-targets. Overall, we performed a comprehensive analysis of the selectivity of a set of proteases from different perspectives, including sequence and active site similarity, the location of hydration and small-molecule hot-spots, as well as molecular docking and toxicological profiling. Even though the different metrics were not always consistent for a single target, we could indicate the proteases with the higher risk of potential off-targeting. Taking into consideration all the analysed factors, we predicted the highest potential for off-target binding of SARS-CoV-2 Mpro inhibitors for factor Xa, SARS-CoV Mpro and cathepsin G, (**Figure 5)** but also in some aspects for chymase, and UCHL1. We estimated a lower probability for prostasin, and thrombin, and the lowest for caspase-3.

With this article, we wanted to underline the need to include research on selectivity assessment from the early stages of drug design, since it would potentially provide better chances for success while designing novel inhibitors towards various diseases. However, we still need to discuss which metrics (and to what extent) should be taken into account in the computer-aided selectivity assessment.

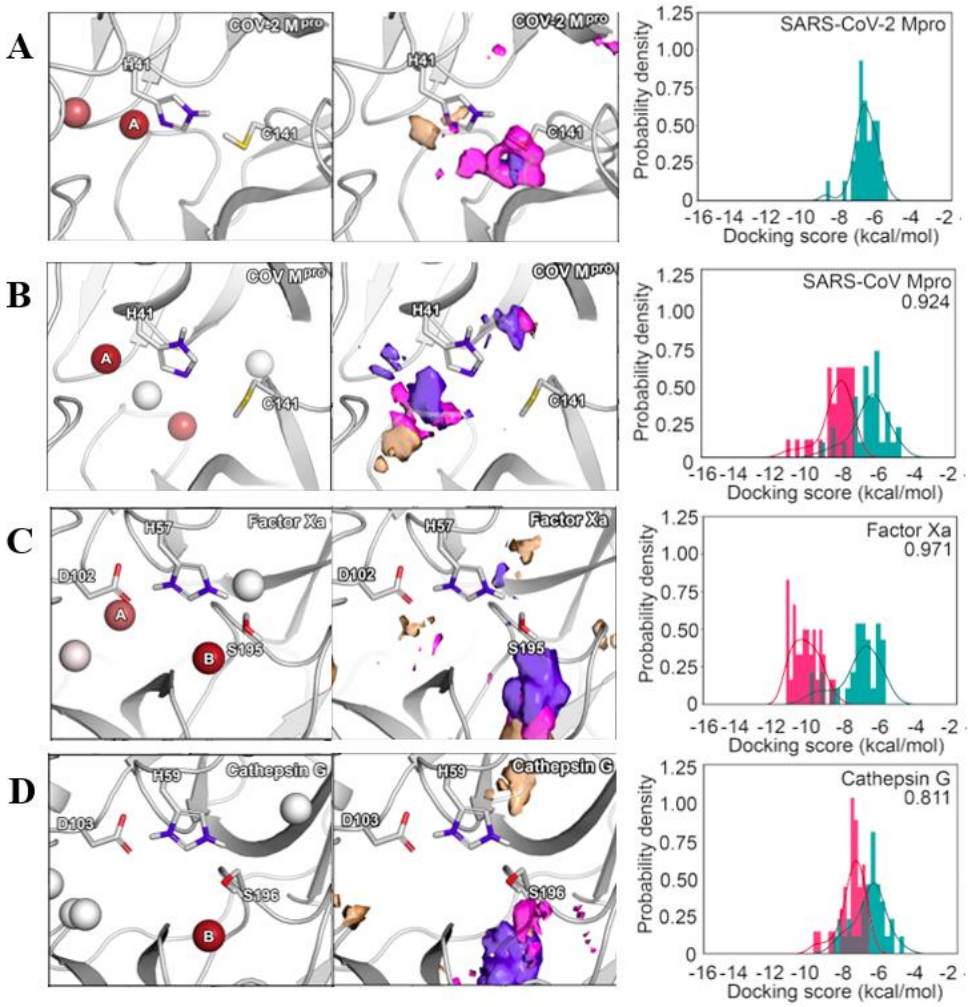

Figure 5 Selectivity form different perspectives for selected proteases: **(A)** SARS-CoV-2 Mpro, **(B)** SARS-CoV Mpro, **(C)** factor Xa and **(D)** cathepsin G. Each panel consists of three parts. Left side: Localisation of the water hot-spots (shown as spheres) identified in the selected proteases in relation to their catalytic residues. Hot-spots are coloured according to the density values (low density – white, high density – red). Two most important hot-spots are marked A and B. Middle: Localisation of the small-molecule hot-spots identified in the selected proteases in relation to their catalytic residues. Blue densities represent isopropanol, pink densities pyridine, and orange acetonitrile. Right side: Score distribution of the SARS-CoV-2 Mpro inhibitors docked to the selected proteases. SARS-CoV-2 Mpro inhibitors are shown in green, the known actives for the remaining targets are shown in red. Figure adapted from **Paper 4** [4] with some modifications.

Another example of the application of water molecules in drug design-related studies was presented in **Paper 5** entitled "Computational insights into the known inhibitors of human soluble epoxide hydrolase". Human soluble epoxide hydrolase (hsEH) is a bifunctional homodimeric enzyme that comprises two independently folded domains: an N-terminal domain (NTD) and a C-terminal domain (CTD). The CTD domain is responsible for the enzyme's hydrolase activity and converts epoxyeicosatrienoic acids known for their anti-inflammatory properties. Given that during the reaction, non-bioactive molecules are created, inhibition of this enzyme would be beneficial. Already a few decades ago, hsEH was proposed as a molecular target in many diseases and disorders, including cardiovascular, metabolic and neurodegenerative. Nevertheless, despite significant efforts to develop inhibition strategies and propose potential inhibitor structures, none have been successfully applied in the clinic, mostly due to poor solubility.

In **Paper 5**, we analysed all deposited hsEH–ligand complexes to gain insight into the binding of inhibitors and to provide feedback on the possibilities of how to improve the drug design process. In addition to studying the possible interactions, we analysed the architecture of the hsEH hydrolase domain, mostly by incorporating the analysis of solvent transportation and internal voids' dynamics. In this article, I am the first author and I carried out all the analyses for known inhibitors co-crystallised with hsEH CTD. I performed the analysis of the interactions, including clustering of binding residues and inhibitors, I ran MD simulations and accomplished the identification of tunnels and inner voids within the hsEH structure. I was involved in writing the manuscript, preparation of figures, in reviewing and editing the final version, as well as in providing answers to the Reviewers' comments.

In the article, we showed that most of the conducted studies indicate the necessity of ensuring direct interactions with side chain residues in the hsEH active site when designing inhibition for this enzyme. We confirmed that the interaction network between the co-crystallised inhibitors and these residues is a highly conserved feature across the reported hsEH inhibitors. In addition to the active site residues, we also showed that the remaining hydrophobic interior of the binding cavity, as well as the hydrophobic surface, were also targeted by the crystallised inhibitors. In general, the presence of a hydrophobic interior implies that a substantial part of the inhibitor structure needs to be hydrophobic, which, in consequence, reduces its solubility. Since the active site of hsEH is buried inside the protein's core, in all analysed crystal structures, the proposed inhibitors are located inside the interior of the enzyme. The fact that the active site

of hsEH is buried quite deep within the protein's core adds certain constraints for inhibitor design, however, it can also provide new opportunities. For instance, inhibitors can target not only the active site itself but also the network of tunnels providing access to it. The literature indicates that utilising such an approach might be very promising [91].

Regarding the potential application of such an approach for human soluble epoxide hydrolase - prior to our study, the literature indicated that the active site of hsEH CTD is connected to the environment through a cavity ("L-shaped" tunnel), which consists of two branches, a long one and a short one. However, our MD-simulation-based analysis provided information about more ingress/egress locations (four tunnels in total), suitable for providing access for molecules penetrating the hsEH. Two tunnels were in line with the commonly used long and short branches. We named them Tm1, and Tc/m, respectively. The Tm1 tunnel, located in the main domain of hsEH CTD, was used by the majority of the identified water molecules and was permanently open. The Tc/m tunnel, found at the border of the cap and main domains, was used by about one-fifth of the water molecules. From two additional tunnels, the first one was found between Tc/m and Tm1. We named it Tg to underline that it was a gorge between two main tunnels. The other tunnel was located deep in the main domain, separated by two loops and we named it Tm2. Although these two tunnels were rather rarely used by water molecules, they represent additional entrances to the protein. Also, the performed analysis of the inner voids revealed that the internal cavity of hsEH is substantially larger than that observed in the crystal structure (**Figure 6A**). Taking this into account, together with the results from the interaction analysis, we indicated that, in the case of hsEH, the interactions with active site residues and their surrounding are not vital for successful inhibitor design. We proposed that besides occupying large hydrophobic moiety, small inhibitors could be positioned on the border of the buried and surface-exposed residues and might benefit from residues donating functional groups, which are essential for increasing the solubility of the compounds. Targeting such regions while designing novel inhibitors could overcome existing limitations regarding e.g. the low solubility of inhibitors. Additionally, analysis of all deposited hsEH–inhibitor complexes indicated that the inhibitors do not fully occupy the available internal cavity (pocket) and that there is still some unused space that could host novel inhibitors.

We also confirmed the possibility of targeting the potential novel binding sites through running additional mixed-solvent MD simulations (with acetonitrile, dimethyl sulfoxide, methanol, phenol, and urea as co-solvents alongside water molecules) and carrying out local-distribution analysis. The results, while not included in **Paper 5**, were detailed in the preprint [92]. We indicted three potential binding sites and named them as P1, P2, and P3.

The P1 region was detected by a hot-spot from acetonitrile molecules and is mostly surrounded by short and aliphatic residues, including L417, V422, A425, G427, L428, and S418. Targeting this site could potentially disturb the mechanism of cap domain opening. The second unique region, P2, was identified by a large-density hot-spot for water molecules and is surrounded by M441, C522, G523, W525, and D529. Positioning the novel inhibitors within this site could potentially block the gorge between the cap and main domains and facilitate more interactions with identified polar groups, thereby enhancing the ligands' properties. The last unique region, P3, was detected by both water and acetonitrile hot-spots. These hot-spots are surrounded by M339, W342, Y343, F362, and I363 (**Figure 6B**). The presence of hot-spots composed of more than one polar chemical probe suggests the potential of hosting additional functional groups that could improve both the solubility and selectivity of designed ligands.

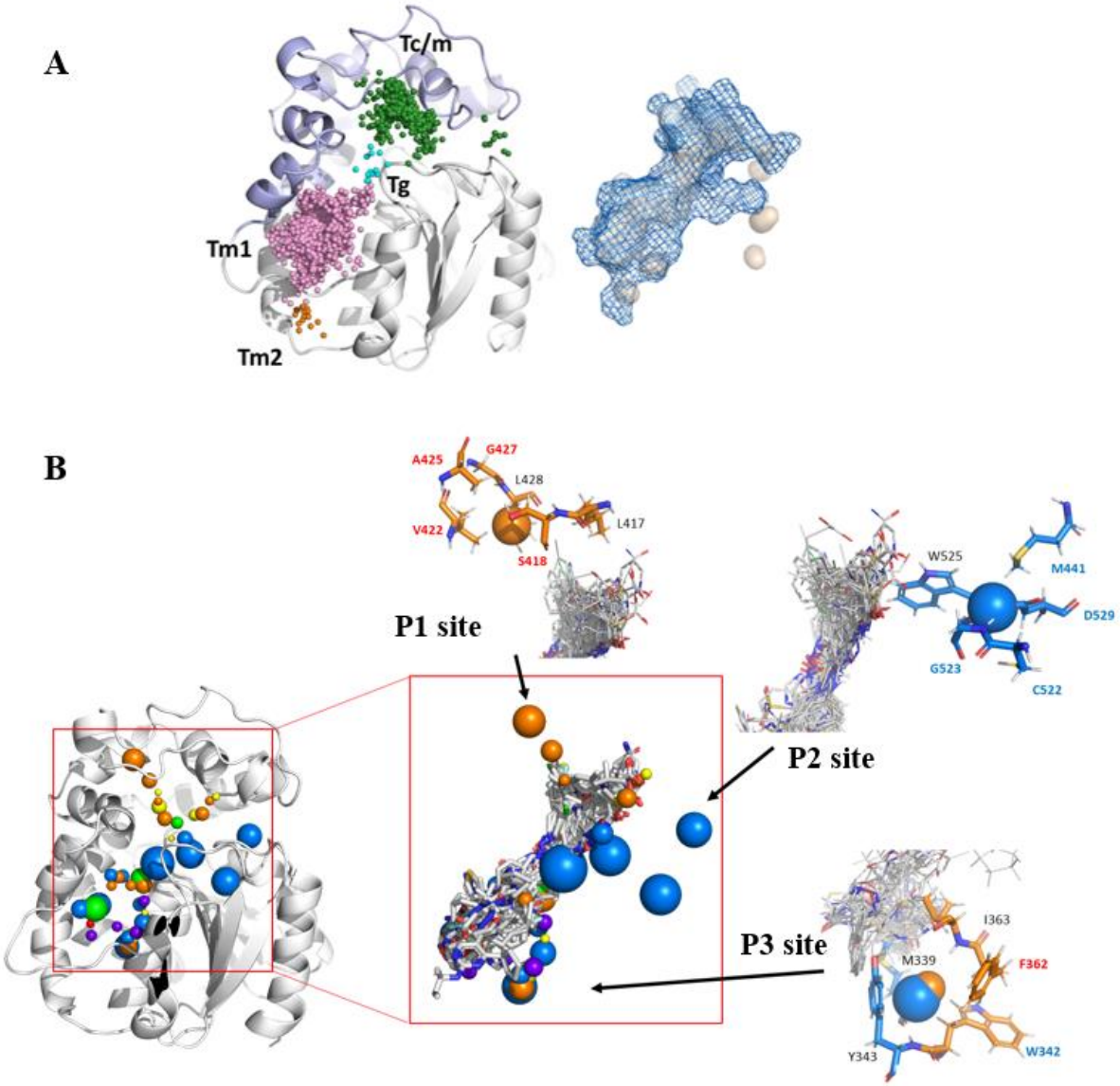

Figure 6 Results of the combination of small-molecule tracking and local-distribution analysis for human soluble epoxide hydrolase (hsEH). **(A)** Left side: Localisation of clusters of inlets indicates different entries to the active site cavity. Colours correspond to the identified tunnels (Tc/m – green, Tm1 – pink, Tg – cyan, and Tm2 – orange). Right side: Comparison of the water-accessible pockets in the crystal structure (beige surface) and during MD simulation (blue mesh). Figure adapted from **Paper 5** [5] with some modifications. **(B)** Localisation of water and co-solvent hot-spots identified in the hsEH, together with the superposition of inhibitors co-crystallised with hsEH (available from Protein Data Bank). Hot-spots are represented as spheres, and their size reflects the hot-spot density. Hot-spots are colour-coded as follows: water – blue, acetonitrile – orange, dimethyl sulfoxide – green, methanol – yellow, phenol – red, urea – purple. Also, the close-up on the surroundings of the novel binding sites (P1, P2, and P3) is presented. Figure adapted from [92] with some modifications.

Using the information about the potential novel binding sites, I proposed a strategy for finding next-generation inhibitors targeting non-obvious cavities in the hydrolase domain of hsEH. First, I prepared the libraries of compounds. I employed all the compounds deposited in the DrugBank (referring to the concept of drug repurposing) but also compounds deposited in the MolPort database, which showed a particular degree of similarity with the known inhibitors, to perform multiple virtual screening and molecular docking calculations. These calculations were aimed at targeting the three identified sites using various software (AutoDock Vina [93] and LeadFinder [94]). For each site, I identified a list of compounds that exhibited low binding energies. The obtained results were further post-processed to obtain a consensus on the ranking of the compounds. Also, the docking poses were compared by calculating and sorting by their root-mean-square deviation of atomic positions, as well as by visual inspection. I performed all the *in silico* studies in cooperation with the Structural Bioinformatics and High Performance Computing (BIO-HPC) Research Group from the Catholic University of Murcia in Spain (I was supervised by Horacio Pérez-Sánchez with technical support provided by Barcelona Supercomputing Center).

The selected compounds were experimentally verified by measuring their half maximal inhibitory concentration (IC_{50}). This experimental validation was performed by our collaborators – Christophe Morriseau and Bruce Hammock from the University of California, Davis, CA, in the United States of America. Indeed, the experimental results confirmed that some of the compounds that were found during the computational screening procedure showed satisfactory inhibitory effects on tested cell lines. As the patent application is currently in preparation, the chemical structure and the exact IC_{50} values cannot be provided. The most promising compounds were further evaluated to assess if they are either competitive, non-competitive or mixed inhibitors. Also, to get better insight, I conducted the MD simulation and subsequent Molecular Mechanics/Generalised Born Surface Area (MM/GBSA) calculations to compute the binding free energy of protein-ligand complexes. In all cases, the obtained binding free energies suggest the strong favourable binding.

Summing up, carrying out small-molecule tracking and local-distribution analyses provided the basis for proposing an improved strategy for searching for novel inhibitors targeting human soluble epoxide hydrolase. I am confident that this approach might be further developed and applied to many other molecular targets as well.

3.3. Applications of water molecules in protein regulation and engineering

Working in the Tunneling Group, we have conducted extensive research focused on the epoxide hydrolase (EH) family, especially the soluble epoxide hydrolase subfamily. One of the reasons for choosing enzymes from this subfamily was that they are proteins with a buried active site, connected to the environment *via* a network of tunnels. As previously mentioned, studying the role of tunnels in proteins is very interesting and needed. This is not only because they enable and regulate substrate entrance and product release, as well as, are used for solvent or ions transportation, but also because they ensure specific conditions within the active site region for the chemical reaction. In general, tunnels are equipped with different regulatory mechanisms such as molecular gates that can control the access to the active site but also synchronise chemical reactions or even protect the active site from various adverse events. This makes studying the tunnels in macromolecules even more valuable.
Due to the quite a diverse network of tunnels in members of the sEH subfamily, those enzymes turned out to be a very good case study to characterise their tunnel network and draw conclusions that may be useful, e.g. in the context of protein regulation and engineering, also for other families.

In **Paper 6** entitled "Structure-function relationship between soluble epoxide hydrolase structure and their tunnel network" (and further Corrigendum to this article, **P6'**), we examined the available structures of soluble epoxide hydrolases and conducted a comprehensive analysis of their tunnel network. We specifically focused on functional tunnels, which we defined as pathways within the protein that allow the movement of water molecules to and from the active site. By using water molecules as probes, we investigated the internal voids of sEHs and gained a deeper understanding of their architecture. This analysis allowed us to elucidate the structural characteristics of the tunnel network in sEHs.
In this article, I was mainly involved in analyses of the flow of water molecules for all the analysed structures, particularly the determination of the appropriate parameters for detecting tunnels and their subsequent characteristics. I also participated in a general comparison of the sEH subfamily members by analysing their structural features. I was involved in data organisation, and partially in writing and reviewing the manuscript, as well as responding to the reviewers' comments.

In **Paper 6**, we examined the structural features of the selected members of the soluble epoxide hydrolase subfamily. These members represented different clades and were as follows: from mammals - *Mus musculus* (msEH), *Homo sapiens* (hsEH), from fungi - *Trichoderma resei* (TrEH), from plants - *Solanum tuberosum* (StEH1), *Vigna radiata* (VrEH2), from bacteria - *Bacillus megaterium* (bmEH). Also, two thermophilic enzymes collected from hot-springs in Russia (SibeEH), and China (CH65-EH) from unknown organisms were added to the analysis. In our study, we described various structural compartments, such as the active site, cap and main domains, as well as cap-loop, NC-loop, and back-loop, which are characteristic features of the soluble epoxide hydrolase subfamily. We showed that the representative structures of sEHs differ while taking into account the number of amino acids making up particular structural compartments, which directly translates into the size of the overall structure. Considering the overall size of the structures, we ranked them in the following order, starting from the smallest: bmEH, SibeEH, CH65-EH, msEH, hsEH, VrEH2, StEH1, and TrEH. We observed variations in the number of amino acids within the main and cap domains, and also we noted that the length of the cap domain was correlated with that of the cap-loop. Overall, we identified both the cap domain and cap-loop as structural features which varied the most among the analysed enzymes. This is of a great importance, since both elements are considered to be responsible for controlling the access to the active site. On the other hand, the back-loop and the NC-loop, which connect the cap and main domains, showed similar lengths in most structures. Additionally, we assessed the similarities and differences among sEHs using multiple protein structure alignment, which revealed regions of high sequence and structural similarity in the main domain and the NC-loop. However, the cap domain region, especially the cap-loop and back-loop regions, exhibited structural differences among the analysed structures. Based on the sequence and structural similarities, we classified the analysed enzymes into three groups. Group I included mammalian sEHs and fungal TrEH, group IIa consisted of plant StEH1 and VrEH2, and group IIb included bacterial bmEH and thermophilic sEHs. Each group exhibited unique features, mostly in terms of cap-loop and back-loop lengths and conformations, and the orientation of helices in the α/β-hydrolase fold.

One of the main objectives of the study was to investigate the structural basis of the tunnel network in sEHs. Therefore, I used information provided by MD simulations trajectories to explore potential transport pathways of water molecules within analysed members of the sEH subfamily. The analysis of water molecule movements provided insight into the tunnel network in different regions of sEHs. As the active site in this subfamily is buried between the cap

and main domains, we expected the primarily utilised tunnels to be located within these two regions. Indeed, most of the identified pathways were associated with tunnels located in the main domain and at the border between the cap and main domains, which we named Tm1, and Tc/m, respectively. Nevertheless, we also identified other tunnels, such as Tm2, Tm3, Tm4, Tm5, Tg, and Tside in the main domain, Tcap1, Tcap2, and Tcap4 in the cap domain, and one more tunnel located between these two domains, namely Tc/m_side. Even though these additional tunnels were used by water molecules quite rarely, they should not be neglected. Based on the tunnel usage, we indicated that some members predominantly utilise both Tc/m and Tm1 tunnels (representatives of mammals and fungi), while others rely rather on one of these tunnels - Tc/m (representatives of bacterial and members from an unknown source) or Tm1 (representatives of plants), respectively (**Figure 7**). These distinct patterns were in agreement with the analysis of structural features, indicating a relationship between the structure and function in members of the soluble epoxide hydrolase subfamily. By analysing the dynamics of the structures, we observed that members of the sEH subfamily display various patterns regarding their flexibility, especially within the cap-loop and back-loop, but also in some helices surrounding the tunnels in the main domain. We correlated the variations in the flexibility with general patterns of tunnel utilisation. Overall, we found out that the structural and dynamic features of proteins translate into the shape, size and utility of individual tunnels and, consequently into the preference of recognised substrates and further catalytic efficiency.

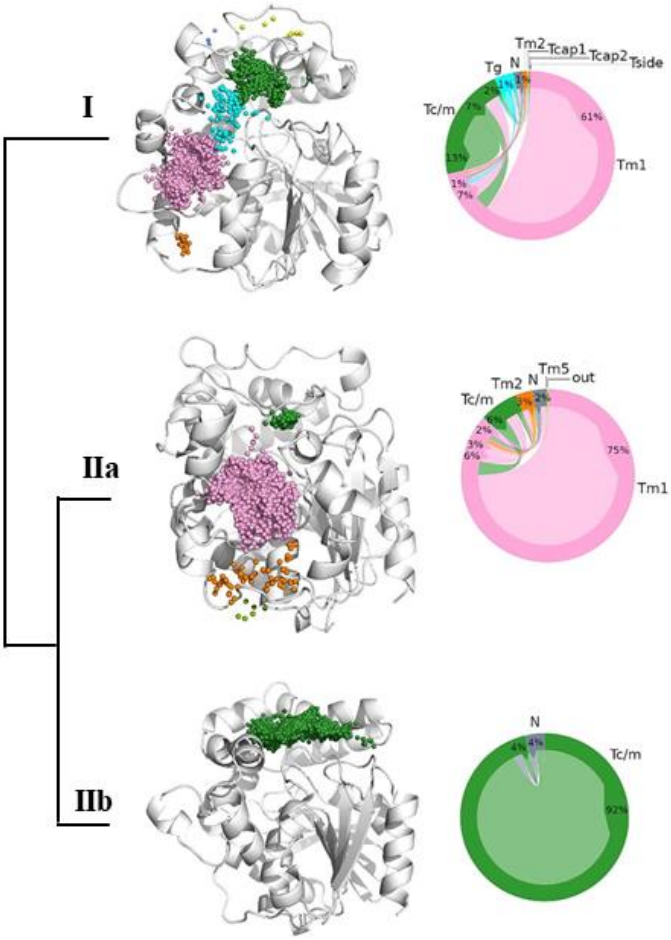

Figure 7 Results of the identified entries/exits to the tunnels selected members of soluble epoxide hydrolase (sEH) together with the intramolecular flow plot. The results are presented for three selected members – from Group I for *Homo sapiens* (hsEH), from Group IIa for *Solanum tuberosum* (StEH1), from Group IIb for *Bacillus megaterium* (bmEH), representing different patterns of tunnel utilisation. The intramolecular flow plot depicts the flow of water molecules through particular tunnels. Figure adapted from **Paper 6** [6] with some modifications.

While studying the soluble epoxide hydrolases, we came to the conclusion that we have a very good set of data that could be used to conduct a detailed comparison of different approaches aimed at identifying tunnels in proteins. Therefore, we made such a comparison and published the results in a subsequent **Paper 7** entitled "Geometry-based versus small-molecule tracking method for tunnel identification: benefits and pitfalls". We performed the analysis for the same set of data as in **Paper 6**. In our study, we employed two distinct tools to represent different approaches: already presented AQUA-DUCT 1.0 for the small-molecule tracking approach, and the commonly used CAVER 3.0 PyMOL plugin along with CAVER 3.02 [95] for the geometry-based approach.

In this article, I share the first authorship with Karolina Mitusińska. We conducted the analyses with the following division: I was involved in identifying tunnels in MD simulations using the small-molecule tracking approach, while Karolina Mitusisńska focused on the application of geometry-based approach. Together we also identified the tunnels in crystal structures using a geometry-based approach. Moreover, I was involved in refining the method for comparing tunnels found in crystal structures with those identified during MD simulations. I was involved in writing the manuscript, preparation of figures, reviewing and editing the final version, as well as in providing answers to the reviewers' comments.

Our goal was to mimic a typical workflow used for tunnel identification in proteins. Therefore, we started with the simplest geometry-based analysis of previously selected sEHs crystal structures. Then, we expanded the analysis, to include the information from the MD simulations and compare both geometry-based and small-molecule tracking methods. Such a comprehensive approach allowed for a comparison of results obtained with distinct methods, highlighting their respective advantages, limitations, and potential biases.
Using the simplest approach (CAVER 3.0 PyMOL plugin), we got information about the number of tunnels, their length, and bottleneck radius. Depending on the structure, we identified three to nine tunnels, with the bottleneck radius ranging from 0.9 Å in bmEH to 2.4 Å in the TrEH structure. We named the identified tunnels according to the nomenclature established while working on previous articles. Then, we used the MD simulation results to perform further geometry-based (CAVER 3.02) and small-molecule tracking (AQUA-DUCT 1.0) analyses. The number of tunnels identified by CAVER 3.02 in the MD simulations was generally higher than those identified in the crystal structures using CAVER 3.0 PyMOL plugin. This discrepancy is due to the conformational changes that proteins undergo during simulations. CAVER 3.02 also provided information on the occurrence of each tunnel, indicating the number of frames in which a tunnel was identified as open. Most structures had at least one tunnel open throughout the simulation, except for VrEH2. The Tm1 tunnel was consistently open in msEH, hsEH, TrEH, and StEH1, while the Tc/m tunnel was always open in bmEH and Sibe-EH, and the Tc/m_back tunnel was the most frequently open in CH65-EH. In VrEH2, the Tm1 tunnel was the most often open. In msEH, StEH1, CH65-EH, TrEH, and hsEH, the Tc/m tunnel was the second most frequently open tunnel. With AQ 1.0, I was able to track water molecules entering and exiting the active site cavities of sEHs and identify the functional tunnels, which were actually used. The number of the identified functional tunnels ranged from one in bmEH to nine in msEH.

Results from AQUA-DUCT not only allowed to identify the real pathways of water molecule transportation but also provided information on the number of inlets in each ingress/egress area as well as their distribution. Additionally, the intramolecular flow plot showed the exchange and flow direction of water molecules.

A comparison of tunnels identified in crystal structures and MD simulations with a geometry-based approach revealed that they generally have their counterparts. However, differences in tunnel shape and size can occur, possibly due to packing inaccuracies or poor resolution in the crystal structure. Even though the overall shape and size of tunnels found in crystal structures were preserved in MD simulations, the regions closer to the protein surface showed more variability. Thus, as we concluded, a simple analysis of crystal structure may lead to an incomplete picture of the tunnel network. Comparison between geometry-based and small-molecule tracking approaches using MD simulations revealed more differences in tunnel identification. AQUA-DUCT could detect functional tunnels that were missed by CAVER, while some tunnels identified by the geometry-based methods were not functional thus were not found by the small-molecule tracking approach. The latter approach provided additional insights into tunnel functionality and the transport of specific molecules (**Figure 8**).

It proved useful for identifying rarely occurring tunnels and assessing their potential usability, e.g. for protein re-engineering and opening novel pathways to reach the active site. Overall, we highlighted that MD simulations may offer a more comprehensive understanding of protein tunnel networks, and that the small-molecule tracking approach complements the geometry-based methods. However, this has its computational cost, which also needs to be considered.

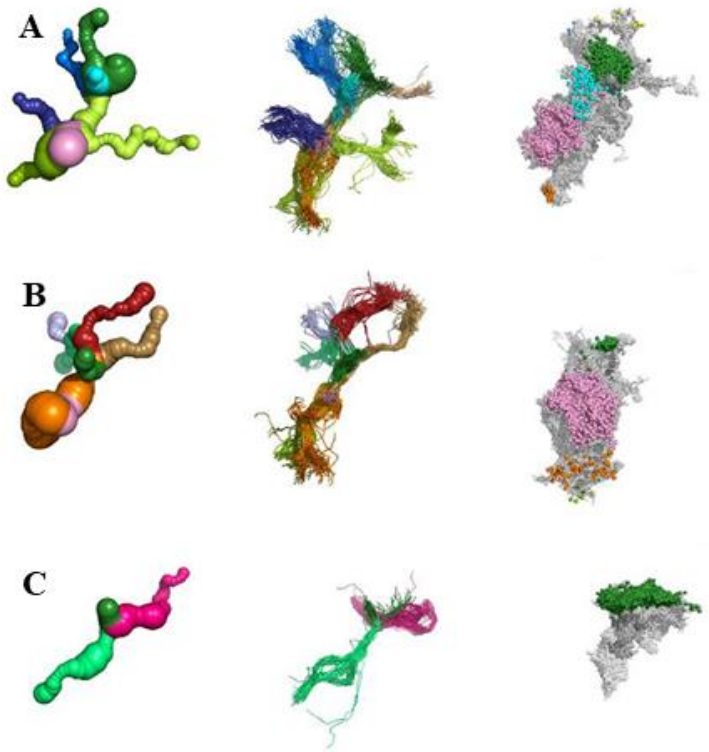

Figure 8 Comparison of tunnels identified in soluble epoxide hydrolase subfamily (sEH) using the geometry-based and small-molecule tracking approaches. The results are presented for three selected members – **(A)** for *Homo sapiens* (hsEH), **(B)** for *Solanum tuberosum* (StEH1), and **(C)** for *Bacillus megaterium* (bmEH). Each panel comprises tunnels obtained by the CAVER 3.0 PyMOL plugin for the crystal structure, tunnels centerlines obtained by CAVER 3.02 software for molecular dynamics simulations, and clusters of inlets together with water molecule pathways obtained by AQUA-DUCT 1.0 from the same MD simulations. Corresponding tunnels, centerlines, and clusters of inlets are marked with the same colour. The colour scheme is consistent with the one proposed in **Figure 7**. Figure adapted from **Paper 7** [7] with some modifications.

The culmination of our research on tunnels in sEHs was the analysis of their evolution. The results were published as **Paper 8** entitled "Evolution of tunnels in α/β-hydrolase fold proteins – What can we learn from studying epoxide hydrolases". This publication is of special importance since the evolution of tunnels has not been addressed in such a comprehensive way in any previous paper. In **Paper 8**, I again share the first authorship with Karolina Mitusińska. In this article, we presented the entire pipeline of evolutionary analysis of tunnels (and also other structurally important compartments) in the sEHs subfamily. We conducted the research on the same set of data as in **Paper 6** and **Paper 7**, excluding the VrEH2 structure. Specifically, I performed the evolutionary analysis of the residues making up particular compartments (residues forming the active site, buried and surface residues, main and cap domains, NC-loop, cap-loop, α-helices, loops, and β-strands used as reference) and all tunnels in each of the selected member of sEHs. For that, I used the entropy analysis implemented

in the BALCONY package [96] (previously, I was involved in the work related to the development of this tool). I evaluated the overall variability for both the referential compartments and tunnels. Besides, I participated in providing characteristics about the general evolutionary analysis of tunnels and a detailed analysis of the selected cases which culminated in proposing the perforation mechanism of the tunnel formation that could be applied as a strategy for *de novo* tunnel design. I was involved in writing the manuscript, preparation of figures, reviewing and editing the final version, as well as in providing answers to the Reviewers' comments.

I performed the variability analyses for all the selected compartments and tunnels. Lists of amino acids building a given compartment were prepared based on the previously conducted structural studies. Lists of residues lining particular tunnels were obtained from the MD simulations and CAVER 3.02 results. Given that tunnels often branch near the surface, which could lead to an overrepresentation of surface residues, we performed separate analyses for all tunnel-lining residues, surface tunnel-lining residues, and tunnel-lining residues after exclusion of those classified as surface residues. The evolutionary analysis for reference compartments revealed that the active site amino acids were conserved, while surface residues were the most variable. This confirmed the well-known observation that amino acids comprising the macromolecule surface evolve faster, while the active site residues remain well-preserved. Generally, buried residues showed lower variability. The cap-loop and NC-loop compartments were classified as a variable in all sEHs, except for the NC-loop in CH65-EH. α-helices and loops were also variable, except for hsEH and StEH1 where the variability of loops was not statistically significant. B-strands were conserved in all analysed proteins, except for msEH. Both main and cap domains were classified as variables.
For the tunnel evolution study, we hypothesised that tunnels should be rather conserved structure features, but equipped with some variable parts which could be responsible, e.g. for different substrate specificity profiles. We based this hypothesis on two assumptions. The first one was that surface residues are more variable in comparison to the buried residues and the second was that access to the active site cavity should be preserved to sustain the activity of the enzyme. Indeed, our results confirmed that. We found out that almost all analysed tunnels could be considered conserved. Plots of entropy distribution showed that tunnel-lining residues exhibit different levels of variability, with surface residues generally displaying higher entropy values. The exclusion of the surface residues from tunnel-lining residues further enhanced the conserved nature of the analysed tunnels.

To get better insight into the location of the variable/conserved residues along tunnels, we performed some additional analyses. For that, we selected three different tunnels, identified in various sEHs. We chose the Tc/m tunnel of hsEH and the Tm1 tunnel of StEH1 which were identified before as the most common tunnels in sEHs, and the Tc/m_back tunnel of bmEH as the case of a tunnel which was previously engineered. Also, we chose these tunnels because they exhibited different distributions of the entropy values.

In the case of the Tc/m tunnel of hsEH whose mouth is located between the cap and main domains, we concluded that it can be characterised as an ancestral tunnel, formed during insertion of the cap domain and preserved in almost all epoxide hydrolases. It is lined with residues with both low and high values of entropy. Residues with high entropy values are located close to the surface or at the interface between both domains. Our detailed analysis indicated that the residue with the highest entropy value – F497 is situated in the middle of the tunnel, between two less-variable residues (D496 and V498) and that it might act as a molecular gate. We confirmed this hypothesis of F497 being a molecular gate by analysis of various crystal structures of hsEH in which the conformations of the F497 residue differ substantially (**Figure 9A**). As for the Tm1 tunnel of StEH1 whose mouth is located in the main domain, near the NC-loop and hinge region, the vast majority of the tunnel-lining residues showed relatively high entropy values, with only several residues with lower entropy located mostly near the active site. In our previous article [97], we already identified three residues, P188, L266, and I270, as potentially useful during protein engineering processes. Here, we also confirmed these residues to be variable, which could mean that their mutation and substitution might not affect protein stability (**Figure 9B**). In the case of the Tc/m_back of bmEH whose mouth is positioned on the other side of the enzyme, we showed that residues with lower entropy values were located within the binding cavity, while the residues with higher entropy were located in the region in the neighbourhood of pocket on the protein surface (**Figure 9C**). By the evaluation of the entropy values for each tunnel-lining residue, we confirmed that two residues that were previously modified [98] – F128 and M145 provided a successful way to open a new tunnel providing access to the bmEH active site. The introduced mutations turned an inaccessible tunnel into an accessible one, enhancing the enzyme's functionality. As we showed, a detailed analysis of residues' entropy can be helpful in the identification of both variable and conserved residues, which can be of great importance when considering potential targets for single-point mutations.

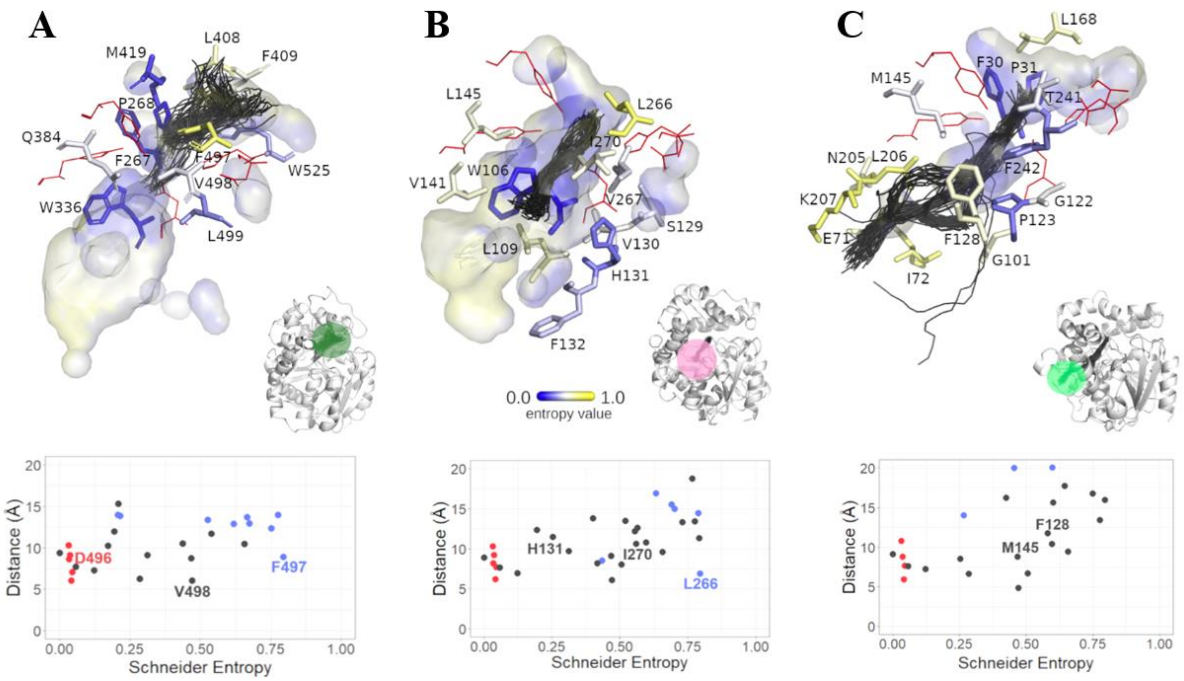

Figure 9 Variability analysis of the representative tunnels from the selected members of the soluble epoxide hydrolase (sEH) subfamily. Results are presented for **(A)** Tc/m tunnel of the *Homo sapiens* (hsEH), **(B)** Tm1 tunnel of the *Solanum tuberosum* (StEH1), and **(C)** Tc/m_back tunnel of the *Bacillus megaterium* (bmEH). The upper panel shows close-up of tunnel-lining residues. For the clarity, only the most frequently detected amino acids are shown (sticks coloured according to their entropy values). Active site residues are shown as red lines. The active site cavity is shown as the interior surface. The lower panel shows the distribution of the entropy values of the tunnel-lining residues in relation to the distance from the geometric centre of the α carbons of the enzyme's active site residues. Active site residues are marked red, buried residues are marked grey, and surface residues are marked blue. Figure adapted from **Paper 8** [8] with some modifications.

All our findings led us to propose a mechanism for tunnel formation. We hypothesise that new tunnels can appear through mutations occurring not only on the protein surface but also at the border of large cavities, affecting surface cavities. In detail, mutations in variable residues can spontaneously drive the evolution of active site accessibility through surface perforation or the joining of internal cavities (**Figure 10**). Such a mechanism can be adapted for enzyme modification and can significantly improve enzyme performance by e.g. separating substrate/product transport pathways from water delivery pathways. Thus, identifying residues prone to causing such an effect can be valuable in protein re-engineering processes.

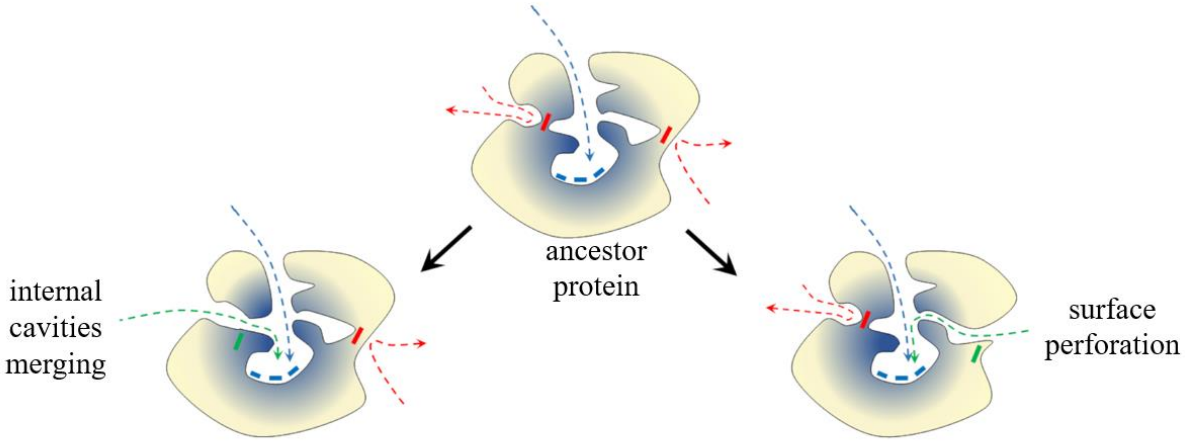

Figure 10 Schematic representation of the proposed theory for tunnel evolution – 'perforation model'. The first possibility shows the appearance of a new tunnel as a result of the merging internal cavities, while the second possibility shows the evolution of a new tunnel as a result of surface perforation. Figure adapted from **Paper 8** [8] with some modifications.

I believe that all of the presented results, starting from the identification of tunnels, through the assessment of their functionality and the study of aspects related to their evolution, can significantly improve the approach to the rational design of proteins and their regulation in catalytic processes.

3.4.Roles of water molecules in the enzymatic reaction

During my doctoral studies, besides relatively small enzymes, I also focused on studying the behaviour of more complex biological macromolecules, such as transmembrane proteins. Specifically, I was investigating Toll-like receptors (TLRs) - biomolecules that play an important role in the functioning of our immunity. These receptors are able to recognise various molecular patterns in the host organism. Recognition of those ligands activates downstream signalling cascades that lead to the induction of the innate immune system. In general, studying TLRs presents numerous challenges, due to their complexity, diversity, dynamic nature, and quite complicated interplay with various ligands, adaptor proteins, and downstream effectors. Nevertheless, I wanted to try to apply the techniques and strategies I was familiar with to understand at least a bit better how these receptors function.

Since I wanted to use computational methods to study the regulatory mechanisms of TLRs, it was necessary to summarise research performed so far and find out what has been achieved and what remains a challenge. With that, **Paper 9** entitled "Recent Advances in Studying Toll-like Receptors with the Use of Computational Methods" was published. In this article, we focused on reviewing TLRs regarding both their function and mechanism of action.
I am the first author of this paper. I performed the vast majority of the literature revision and organised all the data. I was writing the manuscript, reviewing and editing the final version, as well as providing answers to the reviewers' comments.

While reviewing the literature, we noticed that *in silico* studies on TLRs were mostly focused on designing and evaluating novel modulators. Since Toll-like receptors are a potential therapeutic target in various diseases and conditions, it is not surprising that conducting this type of research is a priority for most research groups. As we pointed out, the complexity of TLRs results in a limited understanding of the structural basis of their modes of action, leading to a need for substantial effort in studying TLR dynamics at various levels, from particular domains to the entire receptor structure. Some of the works indeed addressed topics related to dynamic changes, however, such analyses still constitute a relatively small part of all studies that are performed on these receptors. In our review, we identified several areas in TLR research that require further development. Here, I would like to focus on one of them - the need to investigate the proteolytic cleavage reaction in some members of the TLR family. In general, when a ligand binds to a TLR, it either prompts the formation of a receptor dimer or alters the conformation of a preexisting dimer, which further enables adaptor proteins to bind and trigger the response. In some members of the TLR family (TLR7-9), an additional step is needed to allow ligand recognition. This involves the above-mentioned proteolytic cleavage of the long-loop region (so-called Z-loop). Since this process is assumed to serve as one of the regulatory mechanisms for TLRs, understanding its molecular basis would be very beneficial.

Considering that the presence of water is crucial in the proteolytic cleavage reaction, I proposed to analyse the role of water in the TLR-protease system, within the entire reaction cycle. Specifically, I wanted to focus on investigating the behaviour of water molecules within the reaction site. I planned to carry out such an analysis for each reaction species formed during the enzymatic reaction - from the reactant, through subsequent intermediate states, to the product. Performing such research was made feasible through the integration of a variety

of computational methodologies including AI-based structure prediction, MD simulations, QM-only and QM/MM calculations, and small-molecule tracking and local-distribution analysis.

The results of these analyses were deposited as a **Preprint 1** "The proteolytic cleavage of TLR8 Z-loop by furin protease - molecular recognition, reaction mechanism and role of water molecules" which currently is under consideration for publication in a peer-reviewed journal. In this article, I am the first author and I performed the majority of analyses related to the prediction of the TLR8-furin complex (AI-based structure prediction), analysis of the dynamics of the complex and interaction network (MD simulations for each reaction species), and analysis of the reorganisation of water molecules (AQUA-DUCT calculations). I was involved in writing the manuscript, preparation of figures, as well as the overall data organisation.

At this point, I would like to emphasise that carrying out the above-mentioned calculations would not be possible without obtaining a reaction profile from which the coordinates of individual reaction species could be extracted and optimised. The QM and QM/MM calculations were mostly performed by Agnieszka Stańczak (Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences) and Katarzyna Szleper (Tunneling Group), with methodological support provided by Professor Tomasz Borowski (Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences).

We focused on investigating the proteolytic cleavage of TLR8 since it was confirmed that the cleaved form of the receptor is predominant in immune cells and that the uncleaved Z-loop is unable to form a dimer, which is essential for proper functioning. We relied on information suggesting the involvement of furin protease in this process. Indeed, analysis of the sequence at the TLR8 cleavage site, R452-K453-R454-R455↓S456 (RKRR↓S), indicated that furin could be involved in the proteolytic cleavage reaction, as the R-X-K/R-R↓ motif is preferentially recognised by this enzyme.

We started the investigation of the proteolytic cleavage reaction with the prediction of the TLR8-furin complex. Using AlphaFold-Multimer, it was possible to obtain a prediction that indicated not only a good orientation of both macromolecules toward each other, specifically within the RKRR↓S fragment from TLR8 and furin's catalytic site but also strong electrostatic compatibility. Analysis of MD simulations for such a complex indicated that the system can adapt such a spatial geometry within the reaction site that would be necessary

to initiate the enzymatic reaction. The obtained results were the basis for QM and QM/MM calculations which were used to propose the mechanism and reaction profile for the proteolytic cleavage in TLR8. The proposed putative reaction pathway for the proteolytic cleavage of the TLR8 Z-loop by furin comprises acylation, deacylation, and product release steps, being in line with the general mechanism of serine proteases. In short, our studies indicated that the rate limiting step of the analysed reaction is formation of the first tetrahedral intermediate (INT1), which is in agreement with studies performed for other serine proteases. We observed that the formation of this reaction species involves a simultaneous transfer of a proton from furin's catalytic S368 to H194 and a nucleophilic attack on the peptide bond in TLR8. For the subsequent acyl-enzyme species (INT2), we observed its energy being lower than for INT1, but not lower than the energy of the reactant (RE). As for the deacylation process, we observed the formation of tetrahedral intermediate (INT3) only slightly higher in energy than the previous INT2. Finally, the reaction led to the formation of the energetically favourable product (PROD), which indicates that overall, the entire enzymatic reaction is exothermic.

As mentioned, for each reaction species (RE, INT1-INT3, and PROD), I performed MD simulations (in multiple repetitions) to analyse the complex dynamics, interactions network, and the reorganisation of water molecules. Generally, at the reaction site, we observed quite subtle reorientations of side chains. Even though these changes are minor, they seem crucial to achieving the optimal positioning, primarily to initiate the enzymatic reaction, and then to enable subsequent steps to occur. Analysis of the interaction network among residues from TLR8, furin protease and solvent molecules indicated that the general distribution of these interactions may differ between individual reaction steps (**Figure 11**). For instance, we noticed significant differences in water molecule distribution at various stages of the reaction. At the very beginning, there were almost no water molecules present in the vicinity of the reaction site. As the reaction progressed, solvent molecules first occupied the reaction site and furin's interior, and then moved towards the TLR8-furin interface.

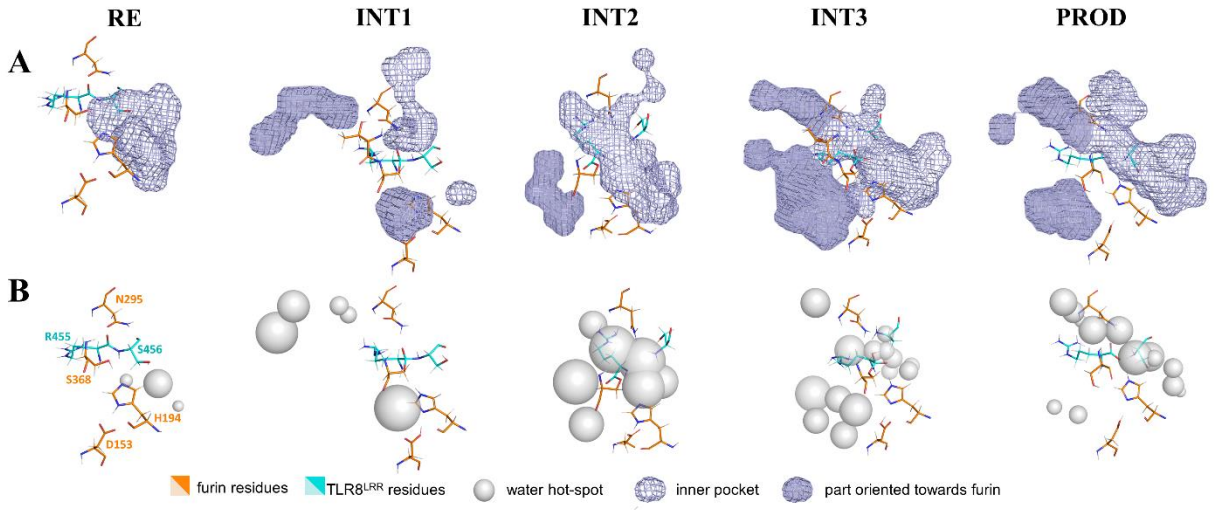

Figure 11 Analysis of water molecules reorganisation within the entire cycle of the proteolytic cleavage of the TLR8 Z-loop by furin protease. **(A)** Visualisation of the internal pockets penetrated by water molecules within the reaction site. Pockets are shown as purple mesh. **(B)** Localisation of the identified high-density water hot-spots. Hot-spots are shown as grey spheres, and their size reflects the hot-spots density. Figure adapted form **Preprint 1** [10] with some modifications.

Water's role in proteolytic cleavage reaction is primarily to act as a substrate in the hydrolysis of the acyl-enzyme, a function we have confirmed. Nevertheless, we also proposed that water can play supporting roles in other steps of the reaction. For instance, water molecules could potentially stabilise the high-energy tetrahedral species INT1 and INT3. Typically, amino acids forming the oxyanion hole ensure such stabilisation. We confirmed that in the complex we studied, the negative charge developing on the R455 oxygen atom from the TLR8 cleavage site is stabilised by the oxyanion hole residues of furin, namely N295, S368, and T367. However, while analysing the MD simulations for these species, we could observe that in some repetitions, water molecules were close enough to either assist or even take over the stabilising function from these residues. Moreover, the identified water hot-spots in the RE, INT3, and PROD species suggest that water molecules might act as a proton shuttle between furin's catalytic histidine H194 and other residues. We hypothesise that this water-mediated proton transfer could open up alternative reaction pathways and affect the energy profile. Finally, our observation of water movement towards the TLR8-furin interface might indicate the involvement of solvent molecules in the dissociation process of these macromolecules. An increased presence of water molecules between these macromolecules could potentially weaken the strong electrostatic interactions holding the complex together, thereby aiding its separation. I believe that the hypotheses we have presented regarding the supplementary roles of water molecules in catalytic reactions can open up the way for new and exciting research opportunities to be explored.

4. Conclusions and Future Perspectives

In my doctoral thesis, I illustrated how, by incorporating the analysis of water molecules in biological systems such as protein, scientists can contribute to a much better understanding of macromolecule structure, dynamics, functioning and overall regulation. I showed that by analysing the behaviour of water molecules in proteins, we can contribute to such fields as drug design and protein engineering. I also confirmed that when studying water in enzymatic processes, we can reveal its multiple roles.

Conducting such comprehensive research would not be possible without the obtained knowledge about computational tools (**Paper 1**) that apply water molecules to the studies on macromolecules' properties. However, most importantly, it would be very hard to carry out the research without one software - AQUA-DUCT 1.0, in the development of which I was involved (**Paper 2**).

In the field of drug design - I presented how, by using a combination of small-molecule tracking (for water and co-solvent molecules) and local-distribution approaches, it is possible to describe the variations in the dynamics of the internal pockets within the macromolecules and identify novel potential sites for ligand binding. By conducting such an analysis for the main protease of SARS-CoV-2, it was possible to indicate that targeting the active site binding pocket might not be the best strategy when designing inhibitors and that targeting other regions could be an alternative option (**Paper 3**). Additionally, the above-mentioned approach was also found useful in the assessment of the potential risk of off-target binding while studying it for SARS-CoV-2 Mpro and a panel of various proteases (**Paper 4**). Also, by using this approach, I was able to propose new potential binding sites for human soluble epoxide hydrolase (**Paper 5** and [92]). By conducting further *in silico* studies where these new sites were targeted, I was able to observe that there was a tendency for new potential inhibitors to bind strongly to these regions. During experimental validation, some of these compounds showed quite a strong inhibition effect, which gives hope that the predictions obtained using computational methods are correct.

Regarding the protein engineering and the general protein regulation - I showed that, by tracking of water molecules during MD simulations, it is possible to describe in detail whole tunnel networks and the transportation phenomena in proteins. By using this approach, it was possible to determine the relationship between the structure and function in proteins from the soluble epoxide hydrolase subfamily (**Paper 6**). Additionally, it was finally possible to conduct a comprehensive comparison of geometry-based and small-molecule tracking methods for detecting and analysing tunnels in proteins (**Paper 7**). By conducting an evolutionary analysis of the tunnels, a new theory was proposed on how tunnels can be formed in proteins (**Paper 8**). This might be very useful, especially in research aimed at carrying out the rational engineering of proteins, e.g. by *de novo* tunnel opening.

Finally, as for the enzymatic reaction - I showed that water molecules can play various roles during the entire reaction cycle. Taking the example of the analysed proteolytic cleavage reaction in TLR8, even though the primary role of water (at a certain step of the reaction) is its catalytic role; it does not mean that water would not also have other functions. Based on the results, several additional roles of water molecules were hypothesised, e.g. provide the stabilisation for certain intermolecular interactions, act as a potential mediator in shuttling the proton between the specific amino acids or participate in the dissociation process of the protein-protein complex (**Paper 9, Preprint 1**).

I believe that the results of projects in which I was involved opened up many opportunities for further, deeper analyses of macromolecules, especially proteins. In the future, I would like to continue work related to the application of water (and co-solvent) molecules in drug design studies, but I would also like to focus more on the aspect of the role of water molecules in protein-protein interactions.

Certainly, as a scientific community, we should also face various challenges that can affect, even bias the results when studying the roles of water molecules in biological systems with the use of computational tools. Some of these challenges are as follows: accurate modelling of solvent effects and parametrisation of water models, the accuracy of the force fields, adequate sampling of the conformational space, and algorithmic limitations. We should also make an effort and try to do our best to integrate both computational and experimental approaches.

List of Figures

Information about the funding

Some of the results presented in this thesis were a continuation of projects awarded by the National Science Centre, Poland, grant no. DEC-2013/10/E/NZ1/00649 "The effect of gating amino acids and anchor amino acids on the regulation of enzyme activity and selectivity" and DEC-2015/18/M/NZ1/00427 "Reversible modification of enzyme activity and substrate specificity by engineering of the ligand exchange pathways".
The work was also supported by the Ministry of Science and Higher Education, Poland from the budget for science for the years 2019-2023, as a research project under the "Diamond Grant" programme [Project number: DIA2018 014148, Agreement Number: 0141/DIA/2019/48] "Molecular aspects of Toll-like receptors regulation considering water molecules as a potential mediator in protein-ligand and protein-protein interactions".

I gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Centers: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2019/013235, PLG/2019/013237, PLG/2020/013874, PLG/2021/014751, and PLG/2022/015577. Part of the computations was also performed using the Poznan Supercomputing and Networking Center infrastructure.

During my PhD studies, I had the opportunity to participate in research internships under the PROM programme - International scholarship exchange of doctoral students and academic staff and Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under the H2020 Programme; in particular, I gratefully acknowledge the support of Horacio Pérez-Sánchez of Universidad Católica de Murcia (UCAM) – Computer Engineering Department and the computer resources and technical support provided by Barcelona Supercomputing Center.

Information about conferences, courses and internships attended after obtaining a Master's degree

2020-05
Conference VII Śląskie Spotkania Naukowe [online]
Oral contribution: Poszukiwanie nowej strategii terapeutycznej skierowanej przeciwko głównej proteazie wirusa SARS-COV-2

2020-09
Webinar Molecular dynamics simulations analysis with molecular probes [online]
Co-organizer

2021-04
Workshop Computer Simulation and Theory of Macromolecules organised by Max Planck Institute for Biophysical Chemistry [online]
Oral contribution: Analysis of molecular dynamics simulations from the "intramolecular voids" perspective with the use of small molecular probes

2021- 07
Conference The 45th FEBS Virtual Congress [online]
Poster: SARS-CoV-2 Mpro as a challenging molecular target for small-molecule inhibitor design

2021-09
Internship PROM Programme - International scholarship exchange of PhD candidates and academic staff
Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), Catholic University of Murcia, Spain

2021-10
Workshop Advances and Challenges in Biomolecular Simulations organised by European Molecular Biology Organization [online]
Poster: Preparation of Toll-like receptor structures for molecular dynamics simulations

2022-02 - 2022-04
Internship HPC-Europa3 Transnational Access Programme
Structural Bioinformatics and High Performance Computing Research Group (BIO-HPC), Catholic University of Murcia, Spain and Barcelona Supercomputing Center

2022-03
Conference II Symposium on Chemical and Physical Sciences for Young Researchers, Murcia, Spain
Oral contribution: Searching for novel potential binding sites and small-molecule inhibitors towards human soluble epoxide hydrolase

2022-04
Conference Computational Oncology and Personalized Medicine COPM2022
THE CHALLENGES OF THE FUTURE
Oral contribution: Application of computational methods in searching novel binding sites and inhibitors towards human soluble epoxide hydrolase

2022-07
Conference 2022 IUBMB–FEBS–PABMB Young Scientists' Forum (YSF 2022) and 2022 IUBMB–FEBS–PABMB The Biochemistry Global Summit; Vimeiro and Lisbon, Portugal
Poster: Readdressing the molecular basis of proteolytic cleavage of TLR8 Z-loop (Best Poster Award during IUBMB–FEBS–PABMB Young Scientists' Forum)

2022-09
Conference 21st European Conference on Computational Biology; Sitges, Barcelona, Spain
Poster: The importance of protein-protein interactions in Toll-like receptor 8 functioning

2023-05
Conference 6th edition of the SBDD - Structured Based Drug Design Conference; Advances in Drug Discovery; Sestri Levante, Italy
Poster: Application of computational methods in searching novel binding sites and inhibitors towards human soluble epoxide hydrolase

2023-07
Webinar 3rd COZYME Webinar [online]
Oral contribution: Water molecules as a key for enzymes interior understanding and reshaping

2023-08
Conference The 6th Quantum Bio-Inorganic Chemistry Conference; Warsaw, Poland
Poster: Studying the molecular basics of TLR8 Z-loop proteolytic cleavage by furin protease with an emphasis on the role of water molecules in the system

2023-09
Course EMBO Lecture Course Structural biophysics of biomolecular complexes; Istanbul, Turkiye
Oral contribution: Insight into the TLR8 functioning with the use of computational modelling and simulations

References

1. Mitusińska K, Raczyńska A, Bzówka M, Bagrowska W, Góra A. Applications of water molecules for analysis of macromolecule properties. Comput Struct Biotechnol J. 2020;18: 355–365. doi:10.1016/j.csbj.2020.02.001
2. Magdziarz T, Mitusińska K, Bzówka M, Raczyńska A, Stańczak A, Banas M, et al. AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective. Bioinformatics. 2020;36: 2599–2601. doi:10.1093/bioinformatics/btz946
3. Bzówka M, Mitusińska K, Raczyńska A, Samol A, Tuszyński JA, Góra A. Structural and evolutionary analysis indicate that the sars-COV-2 mpro is a challenging target for small-molecule inhibitor design. Int J Mol Sci. 2020;21. doi:10.3390/ijms21093099
4. Fischer A, Sellner M, Mitusińska K, Bzówka M, Lill MA, Góra A, et al. Computational Selectivity Assessment of Protease Inhibitors against SARS-CoV-2. Int J Mol Sci. 2021;22: 2065. doi:10.3390/ijms22042065
5. Bzówka M, Mitusińska K, Hopko K, Góra A. Computational insights into the known inhibitors of human soluble epoxide hydrolase. Drug Discov Today. 2021;26: 1914–1921. doi:10.1016/j.drudis.2021.05.017
6. Mitusińska K, Wojsa P, Bzówka M, Raczyńska A, Bagrowska W, Samol A, et al. Structure-function relationship between soluble epoxide hydrolases structure and their tunnel network. Comput Struct Biotechnol J. 2022;20: 193–205. doi:10.1016/j.csbj.2021.10.042
7. Mitusińska K, Bzówka M, Magdziarz T, Góra A. Geometry-Based versus Small-Molecule Tracking Method for Tunnel Identification: Benefits and Pitfalls. J Chem Inf Model. 2022;62: 6803–6811. doi:10.1021/acs.jcim.2c00985
8. Bzówka M, Mitusińska K, Raczyńska A, Skalski T, Samol A, Bagrowska W, et al. Evolution of tunnels in α/β-hydrolase fold proteins—What can we learn from studying epoxide hydrolases? PLOS Comput Biol. 2022;18: e1010119. doi:10.1371/journal.pcbi.1010119
9. Bzówka M, Bagrowska W, Góra A. Recent Advances in Studying Toll-like Receptors with the Use of Computational Methods. J Chem Inf Model. 2023;63: 3669–3687. doi:10.1021/acs.jcim.3c00419
10. Bzówka M, Szleper K, Stańczak A, Borowski T, Góra A. The proteolytic cleavage of TLR8 Z-loop by furin protease - molecular recognition, reaction mechanism and role of water molecules. Res Sq. 2023. doi:10.21203/rs.3.rs-3590328/v1
11. Zwier KR. Methodology in Aristotle's Theory of Spontaneous Generation. J Hist Biol. 2018;51: 355–386. doi:10.1007/s10739-017-9494-7
12. Cavaillon J-M, Legout S. Louis Pasteur: Between Myth and Reality. Biomolecules. 2022;12: 596. doi:10.3390/biom12040596
13. Kawaguchi Y. Panspermia Hypothesis: History of a Hypothesis and a Review of the Past, Present, and Future Planned Missions to Test This Hypothesis. Astrobiology. Singapore: Springer Singapore; 2019. pp. 419–428. doi:10.1007/978-981-13-3639-3_27
14. Oparin A. Origin and evolution of metabolism. Comp Biochem Physiol. 1962;4: 371–377. doi:10.1016/0010-406X(62)90018-X
15. Miller SL. A Production of Amino Acids Under Possible Primitive Earth Conditions. Science (80-). 1953;117: 528–529. doi:10.1126/science.117.3046.528
16. Ferris JP, Hagan WJ. HCN and chemical evolution: The possible role of cyano compounds in prebiotic synthesis. Tetrahedron. 1984;40: 1093–1120. doi:10.1016/S0040-4020(01)99315-9
17. Maden B. No soup for starters? Autotrophy and the origins of metabolism. Trends Biochem Sci. 1995;20: 337–341. doi:10.1016/S0968-0004(00)89069-6
18. Cleaves HJ. Prebiotic Chemistry: What We Know, What We Don't. Evol Educ Outreach. 2012;5: 342–360. doi:10.1007/s12052-012-0443-9
19. Bada JL. New insights into prebiotic chemistry from Stanley Miller's spark discharge experiments. Chem Soc Rev. 2013;42: 2186. doi:10.1039/c3cs35433d
20. Cairns-Smith AG. The origin of life and the nature of the primitive gene. J Theor Biol.

1966;10: 53–88. doi:10.1016/0022-5193(66)90178-0
21. Cairns-smith AG. The chemistry of materials for artificial Darwinian systems. Int Rev Phys Chem. 1988;7: 209–250. doi:10.1080/01442358809353213
22. Corliss JB, Barossa JA, Hiffmann SE. An Hypothesis Concerning the Relationships Between Submarine Hot Springs and the Origin of Life on Earth. Oceanol Acta. 1981;4: 59–69.
23. Miller SL, Bada JL. Submarine hot springs and the origin of life. Nature. 1988;334: 609–611. doi:10.1038/334609a0
24. Gilbert W. Origin of life: The RNA world. Nature. 1986;319: 618–618. doi:10.1038/319618a0
25. Coveney P V., Swadling JB, Wattis JAD, Greenwell HC. Theory, modelling and simulation in origins of life studies. Chem Soc Rev. 2012;41: 5430. doi:10.1039/c2cs35018a
26. Das T, Ghule S, Vanka K. Insights Into the Origin of Life: Did It Begin from HCN and H2O? ACS Cent Sci. 2019;5: 1532–1540. doi:10.1021/acscentsci.9b00520
27. Meisner J, Zhu X, Martínez TJ. Computational Discovery of the Origins of Life. ACS Cent Sci. 2019;5: 1493–1495. doi:10.1021/acscentsci.9b00832
28. Enchev V, Angelov I, Dincheva I, Stoyanova N, Slavova S, Rangelov M, et al. Chemical evolution: from formamide to nucleobases and amino acids without the presence of catalyst. J Biomol Struct Dyn. 2021;39: 5563–5578. doi:10.1080/07391102.2020.1792986
29. Alberts B, Heald R, Johnson A, Morgan D, Raff M, Roberts K, et al. Molecular Biology of the Cell (Seventh Edition). W. W. Norton & Company; 2022.
30. Ramazi S, Zahiri J. Post-translational modifications in proteins: resources, tools and prediction methods. Database. 2021;2021. doi:10.1093/database/baab012
31. Dodd MS, Papineau D, Grenne T, Slack JF, Rittner M, Pirajno F, et al. Evidence for early life in Earth's oldest hydrothermal vent precipitates. Nature. 2017;543: 60–64. doi:10.1038/nature21377
32. Bonner JT. The origins of multicellularity. Integr Biol Issues, News, Rev. 1998;1: 27–36. doi:10.1146/annurev.genet.42.110807.091513
33. Fisher RM, Shik JZ, Boomsma JJ. The evolution of multicellular complexity: the role of relatedness and environmental constraints. Proc R Soc B Biol Sci. 2020;287. doi:10.1098/rspb.2019.2963
34. de Vries J, Archibald JM. Plant evolution: landmarks on the path to terrestrial life. New Phytol. 2018;217: 1428–1434. doi:10.1111/nph.14975
35. Irisarri I, Meyer A. The Identification of the Closest Living Relative(s) of Tetrapods: Phylogenomic Lessons for Resolving Short Ancient Internodes. Syst Biol. 2016;65: 1057–1075. doi:10.1093/sysbio/syw057
36. Meyer A, Meyer A, Dolven SI. Molecules , Fossils , and the Origin of Tetrapods Molecules , Fossils , and the Origin of Tetrapods. 2017; 102–113.
37. Szent-Györgyi A. Biology and pathology of water. Perspect Biol Med. 1971;14: 239–249. doi:10.1353/pbm.1971.0014
38. Milo R, Phillips R. Cell Biology by the Numbers. Garland Science; 2015.
39. Lynden-Bell RM, Morris SC, Barrow JD, Finney JL, Harper C. Water and Life: The Unique Properties of H2O. CRC Press; 2010.
40. Brini E, Fennell CJ, Fernandez-Serra M, Hribar-Lee B, Lukšič M, Dill KA. How Water's Properties Are Encoded in Its Molecular Structure and Energies. Chem Rev. 2017;117: 12385–12414. doi:10.1021/acs.chemrev.7b00259
41. Westall F, Brack A. The Importance of Water for Life. Space Sci Rev. 2018;214: 1–23. doi:10.1007/s11214-018-0476-7
42. Rhee YM, Sorin EJ, Jayachandran G, Lindahl E, Pande VS. Simulations of the role of water in the protein-folding mechanism. Proc Natl Acad Sci. 2004;101: 6456–6461. doi:10.1073/pnas.0307898101
43. Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Water in protein structure prediction. Proc Natl Acad Sci. 2004;101: 3352–3357. doi:10.1073/pnas.0307851100
44. Levy Y, Onuchic JN. Water Mediation in Protein Folding and Molecular Recognition. Annu Rev Biophys Biomol Struct. 2006;35: 389–415. doi:10.1146/annurev.biophys.35.040405.102134
45. Bellissent-Funel M-C, Hassanali A, Havenith M, Henchman R, Pohl P, Sterpone F, et al. Water

Determines the Structure and Dynamics of Proteins. Chem Rev. 2016;116: 7673–7697. doi:10.1021/acs.chemrev.5b00664
46. Baldwin RL. Dynamic hydration shell restores Kauzmann's 1959 explanation of how the hydrophobic factor drives protein folding. Proc Natl Acad Sci. 2014;111: 13052–13056. doi:10.1073/pnas.1414556111
47. Camilloni C, Bonetti D, Morrone A, Giri R, Dobson CM, Brunori M, et al. Towards a structural biology of the hydrophobic effect in protein folding. Sci Rep. 2016;6: 28285. doi:10.1038/srep28285
48. Nick Pace C, Scholtz JM, Grimsley GR. Forces stabilizing proteins. FEBS Lett. 2014;588: 2177–2184. doi:10.1016/j.febslet.2014.05.006
49. Levy Y, Onuchic JN. Water and proteins: A love-hate relationship. Proc Natl Acad Sci. 2004;101: 3325–3326. doi:10.1073/pnas.0400157101
50. Fogarty AC, Laage D. Water Dynamics in Protein Hydration Shells: The Molecular Origins of the Dynamical Perturbation. J Phys Chem B. 2014;118: 7715–7729. doi:10.1021/jp409805p
51. Laage D, Elsaesser T, Hynes JT. Water Dynamics in the Hydration Shells of Biomolecules. Chem Rev. 2017;117: 10694–10725. doi:10.1021/acs.chemrev.6b00765
52. England JL, Haran G. Role of Solvation Effects in Protein Denaturation: From Thermodynamics to Single Molecules and Back. Annu Rev Phys Chem. 2011;62: 257–277. doi:10.1146/annurev-physchem-032210-103531
53. Sun Q, Fu Y, Wang W. Temperature effects on hydrophobic interactions: Implications for protein unfolding. Chem Phys. 2022;559: 111550. doi:10.1016/j.chemphys.2022.111550
54. Seelig J, Seelig A. Protein Unfolding—Thermodynamic Perspectives and Unfolding Models. Int J Mol Sci. 2023;24: 5457. doi:10.3390/ijms24065457
55. Damjanović A, Schlessman JL, Fitch CA, García AE, García-Moreno E. B. Role of Flexibility and Polarity as Determinants of the Hydration of Internal Cavities and Pockets in Proteins. Biophys J. 2007;93: 2791–2804. doi:10.1529/biophysj.107.104182
56. Jeszenői N, Bálint M, Horváth I, van der Spoel D, Hetényi C. Exploration of Interfacial Hydration Networks of Target–Ligand Complexes. J Chem Inf Model. 2016;56: 148–158. doi:10.1021/acs.jcim.5b00638
57. Zsidó BZ, Hetényi C. The role of water in ligand binding. Curr Opin Struct Biol. 2021;67: 1–8. doi:10.1016/j.sbi.2020.08.002
58. Schiebel J, Gaspari R, Wulsdorf T, Ngo K, Sohn C, Schrader TE, et al. Intriguing role of water in protein-ligand binding studied by neutron crystallography on trypsin complexes. Nat Commun. 2018;9: 3559. doi:10.1038/s41467-018-05769-2
59. Zsidó BZ, Bayarsaikhan B, Börzsei R, Szél V, Mohos V, Hetényi C. The Advances and Limitations of the Determination and Applications of Water Structure in Molecular Engineering. Int J Mol Sci. 2023;24: 11784. doi:10.3390/ijms241411784
60. Leitner DM, Hyeon C, Reid KM. Water-mediated biomolecular dynamics and allostery. J Chem Phys. 2020;152. doi:10.1063/5.0011392
61. Grimaldo M, Roosen-Runge F, Zhang F, Schreiber F, Seydel T. Dynamics of proteins in solution. Q Rev Biophys. 2019;52. doi:10.1017/s0033583519000027
62. Maurer M, Oostenbrink C. Water in protein hydration and ligand recognition. J Mol Recognit. 2019;32. doi:10.1002/jmr.2810
63. Spyrakis F, Ahmed MH, Bayden AS, Cozzini P, Mozzarelli A, Kellogg GE. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. J Med Chem. 2017;60: 6781–6827. doi:10.1021/acs.jmedchem.7b00057
64. Matsuoka D, Nakasako M. Probability distributions of hydration water molecules around polar protein atoms obtained by a database analysis. J Phys Chem B. 2009;113: 11274–11292. doi:10.1021/jp902459n
65. Gnesi M, Carugo O. How many water molecules are detected in X-ray protein crystal structures? J Appl Crystallogr. 2017;50: 96–101. doi:10.1107/S1600576716018719
66. Sliz P, Harrison SC, Rosenbaum G. How does radiation damage in protein crystals depend on X-ray dose? Structure. 2003;11: 13–19. doi:10.1016/S0969-2126(02)00910-3
67. McPherson A, Gavira JA. Introduction to protein crystallization. Acta Crystallogr Sect F Struct Biol Commun. 2014;70: 2–20. doi:10.1107/S2053230X13033141

68. Jorge C, Marques BS, Valentine KG, Wand AJ. Characterizing Protein Hydration Dynamics Using Solution NMR Spectroscopy. 1st ed. Methods in Enzymology. Elsevier Inc.; 2019. doi:10.1016/bs.mie.2018.09.040
69. Zanotti JM, Bellissent-Funel MC, Parello J. Hydration-coupled dynamics in proteins studied by neutron scattering and NMR: The case of the typical EF-hand calcium-binding parvalbumin. Biophys J. 1999;76: 2390–2411. doi:10.1016/S0006-3495(99)77395-9
70. Trainor K, Palumbo JA, MacKenzie DWS, Meiering EM. Temperature dependence of NMR chemical shifts: Tracking and statistical analysis. Protein Sci. 2020;29: 306–314. doi:10.1002/pro.3785
71. Li C, Pielak GJ. Using NMR to distinguish viscosity effects from nonspecific protein binding under crowded conditions. J Am Chem Soc. 2009;131: 1368–1369. doi:10.1021/ja808428d
72. Doerr A. Structural analysis of macromolecular assemblies. Nat Methods. 2008;5: 23. doi:10.1038/nmeth1160
73. Thompson RF, Walker M, Siebert CA, Muench SP, Ranson NA. An introduction to sample preparation and imaging by cryo-electron microscopy for structural biology. Methods. 2016;100: 3–15. doi:10.1016/j.ymeth.2016.02.017
74. Alavi S. Molecular Simulations: Fundamentals and Practice. Wiley-VCH; 2020.
75. Onufriev A V., Izadi S. Water models for biomolecular simulations. Wiley Interdiscip Rev Comput Mol Sci. 2018;8. doi:10.1002/wcms.1347
76. Kadaoluwa Pathirannahalage SP, Meftahi N, Elbourne A, Weiss ACG, McConville CF, Padua A, et al. Systematic Comparison of the Structural and Dynamic Properties of Commonly Used Water Models for Molecular Dynamics Simulations. J Chem Inf Model. 2021;61: 4521–4536. doi:10.1021/acs.jcim.1c00794
77. Cramer CJ, Truhlar DG. Implicit Solvation Models: Equilibria, Structure, Spectra, and Dynamics. Chem Rev. 1999;99: 2161–2200. doi:10.1021/cr960149m
78. Izadi S, Anandakrishnan R, Onufriev A V. Building Water Models: A Different Approach. J Phys Chem Lett. 2014;5: 3863–3871. doi:10.1021/jz501780a
79. Fichthorn KA, Weinberg WH. Theoretical foundations of dynamical Monte Carlo simulations. J Chem Phys. 1991;95: 1090–1096. doi:10.1063/1.461138
80. Bergazin TD, Ben-Shalom IY, Lim NM, Gill SC, Gilson MK, Mobley DL. Enhancing water sampling of buried binding sites using nonequilibrium candidate Monte Carlo. J Comput Aided Mol Des. 2021;35: 167–177. doi:10.1007/s10822-020-00344-8
81. Van Mourik T, Bühl M, Gaigeot MP. Density functional theory across chemistry, physics and biology. Philos Trans R Soc A Math Phys Eng Sci. 2014;372. doi:10.1098/rsta.2012.0488
82. Groenhof G. Introduction to QM/MM Simulations. Biomolecular Simulations. 2013. pp. 43–66. doi:10.1007/978-1-62703-017-5_3
83. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, et al. AMBER 2014. San Francisco: University of California; 2014.
84. Case DA, Ben-Shalom IY, Brozell SR, Cerutti DS, Cheatham TE, Cruzeiro VWD, et al. AMBER 2018. San Francisco: University of California; 2018.
85. Case DA, Aktulga HM, K. B, Ben-Shalom IY, Berryman JT, Brozell SR, et al. Amber 2022. San Francisco: University of California; 2022.
86. Jorgensen WL, Chandrasekhar J, Madura JD, Impey RW, Klein ML. Comparison of simple potential functions for simulating liquid water. J Chem Phys. 1983;79: 926–935. doi:10.1063/1.445869
87. Maier JA, Martinez C, Kasavajhala K, Wickstrom L, Hauser KE, Simmerling C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. J Chem Theory Comput. 2015;11: 3696–3713. doi:10.1021/acs.jctc.5b00255
88. Mitusińska K, Raczyńska A, Wojsa P, Bzówka M, Góra A. AQUA-DUCT: Analysis of Molecular Dynamics Simulations of Macromolecules with the use of Molecular Probes [Article v1.0]. Living J Comput Mol Sci. 2020;2. doi:10.33011/livecoms.2.1.21383
89. Ghanakota P, Carlson HA. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. J Med Chem. 2016;59: 10383–10399. doi:10.1021/acs.jmedchem.6b00399
90. Defelipe L, Arcon J, Modenutti C, Marti M, Turjanski A, Barril X. Solvents to Fragments to

Drugs: MD Applications in Drug Design. Molecules. 2018;23: 3269. doi:10.3390/molecules23123269
91. Marques SM, Daniel L, Buryska T, Prokop Z, Brezovsky J, Damborsky J. Enzyme Tunnels and Gates As Relevant Targets in Drug Design. Med Res Rev. 2017;37: 1095–1139. doi:10.1002/med.21430
92. Bzówka M, Mitusińska K, Góra A. Novel Potential Binding Sites for Selective Inhibitor Design of Human Soluble Epoxide Hydrolase. Res Sq. 2020. doi:10.21203/rs.3.rs-29814/v1
93. Trott O, Olson AJ. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010;31: 455–461. doi:10.1002/jcc.21334
94. Stroganov O V., Novikov FN, Stroylov VS, Kulkov V, Chilov GG. Lead Finder: An Approach To Improve Accuracy of Protein−Ligand Docking, Binding Energy Estimation, and Virtual Screening. J Chem Inf Model. 2008;48: 2371–2385. doi:10.1021/ci800166p
95. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, et al. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. PLoS Comput Biol. 2012;8. doi:10.1371/journal.pcbi.1002708
96. Płuciennik A, Stolarczyk M, Bzówka M, Raczyńska A, Magdziarz T, Góra A. BALCONY: an R package for MSA and functional compartments of protein variability analysis. BMC Bioinformatics. 2018;19: 300. doi:10.1186/s12859-018-2294-z
97. Mitusińska K, Magdziarz T, Bzówka M, Stańczak A, Gora A. Exploring Solanum tuberosum Epoxide Hydrolase Internal Architecture by Water Molecules Tracking. Biomolecules. 2018;8: 143. doi:10.3390/biom8040143
98. Kong X-D, Yuan S, Li L, Chen S, Xu J-H, Zhou J. Engineering of an epoxide hydrolase for efficient bioresolution of bulky pharmaco substrates. Proc Natl Acad Sci. 2014;111: 15717–15722. doi:10.1073/pnas.1404915111