



**Silesian University
of Technology**

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**

**ATTACHMENT TO THE DOCTORAL THESIS:
PUBLICATIONS**

Supplementary Information for all publications are available online

Supervisor: dr hab. Artur GÓRA, prof. PŚ

Gliwice, 2023



Review

Applications of water molecules for analysis of macromolecule properties



Karolina Mitusińska, Agata Raczyńska, Maria Bzówka, Weronika Bagrowska, Artur Góra*

Tunneling Group, Biotechnology Centre, Silesian University of Technology, Krzywoustego 8, Gliwice, Poland

ARTICLE INFO

Article history:

Received 15 October 2019
Received in revised form 26 January 2020
Accepted 1 February 2020
Available online 12 February 2020

Keywords:

Water molecules
Water placement
Macromolecule structure
Transportation
Water sites
Tunnel detection

ABSTRACT

Water molecules maintain proteins' structures, functions, stabilities and dynamics. They can occupy certain positions or pass quickly via a protein's interior. Regardless of their behaviour, water molecules can be used for the analysis of proteins' structural features and biochemical properties. Here, we present a list of several software programs that use the information provided by water molecules to: i) analyse protein structures and provide the optimal positions of water molecules for protein hydration, ii) identify high-occupancy water sites in order to analyse ligand binding modes, and iii) detect and describe tunnels and cavities. The analysis of water molecules' distribution and trajectories sheds a light on proteins' interactions with small molecules, on the dynamics of tunnels and cavities, on protein composition and also on the functionality, transportation network and location of functionally relevant residues. Finally, the correct placement of water molecules in protein crystal structures can significantly improve the reliability of molecular dynamics simulations.

© 2020 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Contents

1. Introduction	355
2. Software for protein hydration analysis	356
3. Software for water site detection and ligand binding analysis	358
4. Software for tunnel detection and transportation phenomena analysis	360
5. Summary and outlook	362
Funding	363
Conflict of interest	363
CRedit authorship contribution statement	363
References	363

1. Introduction

Life began to evolve in an aqueous milieu, and the unique properties of water determine the chemistry of all living organisms. Water is a ubiquitous and essential substance in cells, accounting for about 70% of their mass. It is not only the environment for biological processes, but also an integral part of them [1]. At a macromolecular level, water contributes to biomolecules' formation and their stability, dynamics and functions [2–4]. Water serves as a

reaction reagent or mediates ligand–protein and protein–protein interactions. Water molecules are small enough to penetrate a macromolecule's core, to stabilise its native structure and also to participate in processes occurring in the protein's core [5,6].

X-ray [7] and neutron diffraction [8] crystallography provide an insight into the spatial distribution of water molecules in the vicinity of biomolecular surfaces and confined regions such as active sites, pockets and cavities. Depending on the crystal quality, atomic resolutions can be achieved [9–11]. Protein structures deposited in the Protein Data Bank (PDB) [12] contain an abundance of information, i.e., alternative conformations of amino acid side chains and potential rearrangements of protein compartments. Information

* Corresponding author.

E-mail address: a.gora@tunnelinggroup.pl (A. Góra).

about water molecules' positions is usually incomplete or can be strongly influenced by experimental conditions. Therefore, it is unclear how closely the distribution of crystal water molecules resembles the native conditions of the biomolecule. Nuclear Magnetic Resonance (NMR) is useful for discovering the hydration properties of proteins, especially their dynamics. Unfortunately, this technique cannot provide information about the three-dimensional structure of the hydration sites, and its time scales are shorter by an order of magnitude than the residence times of water molecules [4,13].

The limitations of experimental methods can be overcome by computational techniques. *Ab initio* and DFT (Density Functional Theory) methods can be used for a precise description of a reaction mechanism, including the contribution of water [14,15]. Molecular dynamics (MD) and Monte Carlo (MC) simulations provide a detailed atomic description of a biomolecule and a solvent, along with their dynamics [16,17]. These simulations, however, do not tackle many equilibrium and long-time-scale kinetic properties [18].

The increasing awareness of the significant role of water molecules has given rise to a range of software focused on the analysis of water molecules' behaviour. Recent reviews focused on virtual screening strategies describe several docking software applications that are capable of utilising information related to water molecules [19–21]. This paper provides a review of the available computational methods that employ water molecules for the analysis of macromolecules' properties and structure dynamics. In the first part, we provide an overview of the techniques used for the prediction of water molecules' locations. The following chapter describes the water sites that may participate in ligand binding. Next, water molecules are analysed in terms of ligand transportation and the detection of tunnels and cavities. For all three chapters, a list of software along with their functionality and/or their characteristics and principles is provided. The last chapter comprises conclusions and general remarks, as well as perspectives on the further development of software.

2. Software for protein hydration analysis

Water molecules not only maintain the functions of proteins but also stabilise their native structure [13]. The presence of water molecules in proteins' internal cavities is conserved among homologous proteins families, as well as the key residues are [22]. It was shown, by reducing the amount of water during crystallisation [23] or by using mutants of particular proteins [24], that internal water molecules contribute to the structural folding and the stability of enzymes, ion channels, proton pumps and other macromolecules [25–27]. However, as we pointed out above, the experimental results are insufficient and can be inconsistent with each other [28]. The water molecules inside a protein's structure may also be distorted, and their position may depend on the orientation of a particular side chain. They may also be trapped inside a protein's cavity due to a process of large conformational changes.

The residence time of a water molecule buried in an internal cavity or trapped in a narrow cleft depends on its location and connectivity with the bulk solvent [29,30]. Fast minimisation and short equilibrium stages can provide insufficient or inaccurate solvation of the protein interior and can bias the results. Therefore, it is important to fill the internal cavities with water molecules precisely, prior to running MD simulations. Lengthening the minimisation and equilibration procedure can provide sufficient exchange of water molecules between the surroundings and the protein interior; however, it is strongly system-dependent. Application of software developed to place water molecules into a protein's cavities and its surroundings is proposed as an alternative strategy, especially for systems with large interior volumes, homology-modelled proteins or proteins with mutations introduced inside their cores (Table 1).

Three different methods (Fig. 1) have been implemented for the placement of water molecules in a protein's interior: i) based on the docking of water molecules, such as Dowser [31] and WaterDock [32], ii) based on the reference interaction site model (RISM), such as 3D-RISM [33], GAsol [34] and Placevent [35], and iii) based

Table 1
List of available software to predict water molecules' positions and orientation.

Software	Testing set	Accuracy*	Remarks
Docking-based			
Dowser [30]	14 crystal structures of OppA; D- and K-channels of cytochrome c oxidase;	63%	<i>Not available</i>
Dowser+ [31]	Photosystem II	74%	<i>Not available</i>
Dowser++ [39]		85%	Dowser++ standalone link
WaterDock [32]	14 crystal structures of OppA; HIV-1 protease; ribonuclease A; GluR2 ligand binding core; concanavalin A; glutathione S-transferase; carbonic anhydrase	88%	the code is available with the WaterDock 2.0 Pymol plugin: – link
WaterDock 2.0 [47]	14 crystal structures of OppA; HIV-1 protease; GluR2 ligand binding core; bovine pancreatic trypsin; glutathione S-transferase; HSP90; PIM1; series of 184 BRD4-BD1 complexes; androgen receptor; casein kinase II; thrombin; carbonic anhydrase	91%	WaterDock 2.0 standalone link WaterDock 2.0 Pymol plugin link
RISM-based			
3D-RISM [33]	Alanine dipeptide; HIV-1 protease	–	<i>Not available</i>
GAsol [34]	HIV-1 protease; neuraminidase; bovine pancreatic trypsin; series of 184 BRD4-BD1 complexes	94.3%	BSD 3-clause license link
Placevent [35]	HIV-1 protease; rotor ring of F-ATP synthase	water position error –0.5 Å	the code and tutorial link
Similarity-based			
ProBis H2O [37]	Src kinase with bound bosutinib; human programmed death 1 (hPD) with ligand (hPD-L1); DNA Gyrase B; human matrix metalloproteinase (hMMP-1)	–	GUI – PyMOL integrated link
PyWATER [36]	thrombin; trypsin; BPTI; bromodomain-containing protein 4; MHC class I proteins; class A β-lactamases	identified all crystallographic water molecules	GUI – PyMOL integrated link

*Accuracy was calculated based on the number and quality of identified crystallographic water molecules. The numbers were taken from original publications. Please note that there are some differences in the details of the accuracy measurements. Information about currently unavailable software is in *italics*.

on the assumption that internal water molecules are conserved among similar proteins' structures (PyWATER [36] and ProBiS H2O [37]).

The docking-based methods assume that the protein structure is the target and the water molecule is the ligand. Both Dowser and WaterDock utilise the commonly available docking software - AutoDock Vina [38]. These methods are fast and provide accurate positioning of the water molecules determined by crystallography. The water molecule docking algorithms have improved over new software versions, and for the Dowser software 'generations', the average accuracy of their predictions have increased from 63% in Dowser, to 74% in Dowser+ and up to 85% of the water molecules in Dowser++ [39], when compared to high-resolution crystallographic structures. WaterDock software presented a higher accuracy of crystallographic water molecules' prediction than Dowser ++: it was 88% for the original WaterDock and 91% for WaterDock 2.0; however, it should be kept in mind that there were some differences in the details of the accuracy measurements described in the original publications [39]. Along with their ability to predict water molecules' positions, both WaterDock software releases are also able to determine if water molecules are displaced or ordered. WaterDock 2.0 comes with an easy-to-use PyMOL plugin.

The RISM theory is used for calculating the distribution of solvent molecules around a solute and has its roots in statistical, mechanical integral equation theories (IET) of liquids [34]. Due to the fact that the distribution calculated by 3D-RISM theory is continuous, it is difficult to directly examine specific solvent interactions, especially when they are numerous. 3D-RISM has been successfully used to locate water molecules in proteins as compared to experiment [40] and simulation [41], to calculate hydration free energies [42] and to predict fragment and drug positions [43]. The Placevent algorithm gave an average error for water molecules' positions of about 0.5 Å [35]. The GASol software, in which the 3D-RISM theory was combined with a genetic algorithm and a desirability function, showed the highest accuracy,

with 93.4% of the predicted water molecules within 2 Å from their crystallographic positions [34]. Generally, RISM-based methods for water molecules' prediction are slower than docking-based ones, and the computational time is system-size-dependent. However, they can be more accurate, especially for complex systems (e.g., metalloproteins, proteins equipped with large cavities or in complexes with nucleic acids) [44]. Moreover, it was shown that the RISM theory may break down in larger systems and systematically underestimates the partial molar volume (PMV) of amino acids [13].

As an alternative to the methods based on the physicochemical properties comes a simple similarity-based approach, implemented in PyWATER [36] and ProBiS H2O [37], which both superimpose crystallographic structures similar to the target protein and cluster the positions of conserved water molecules inside the protein cavities. ProBiS H2O is the first software that utilises the ProBiS algorithm [45] to perform local superimposition of the detected conserved water molecules. It also reduces the bias introduced by comparing similar protein structures or structures in different conformations than the query protein. PyWATER searches for similar structures using the PDB database [46]. The accuracy of such an approach strongly depends on the number, similarity and quality of related structures. Generally, ProBiS H2O gives fewer clusters with more tightly packed water molecules in comparison to PyWATER due to the clustering algorithms used (PyWATER uses hierarchical clustering, while ProBiS H2O uses a Python implementation of 3D-DBSCAN (Three Dimensional Density-Based Spatial Clustering of Applications with Noise)). In addition, PyWATER stores information on the degree of conservation of each water molecule cluster with related atom numbers of water oxygen atoms from the superimposed protein structures.

All the tools mentioned above provide relatively fast, intuitive and accurate modelling of the water molecules in low-quality crystal structures and thus provide a more accurate starting point for an MD or MC study. Their usage can be also recommended for *in*

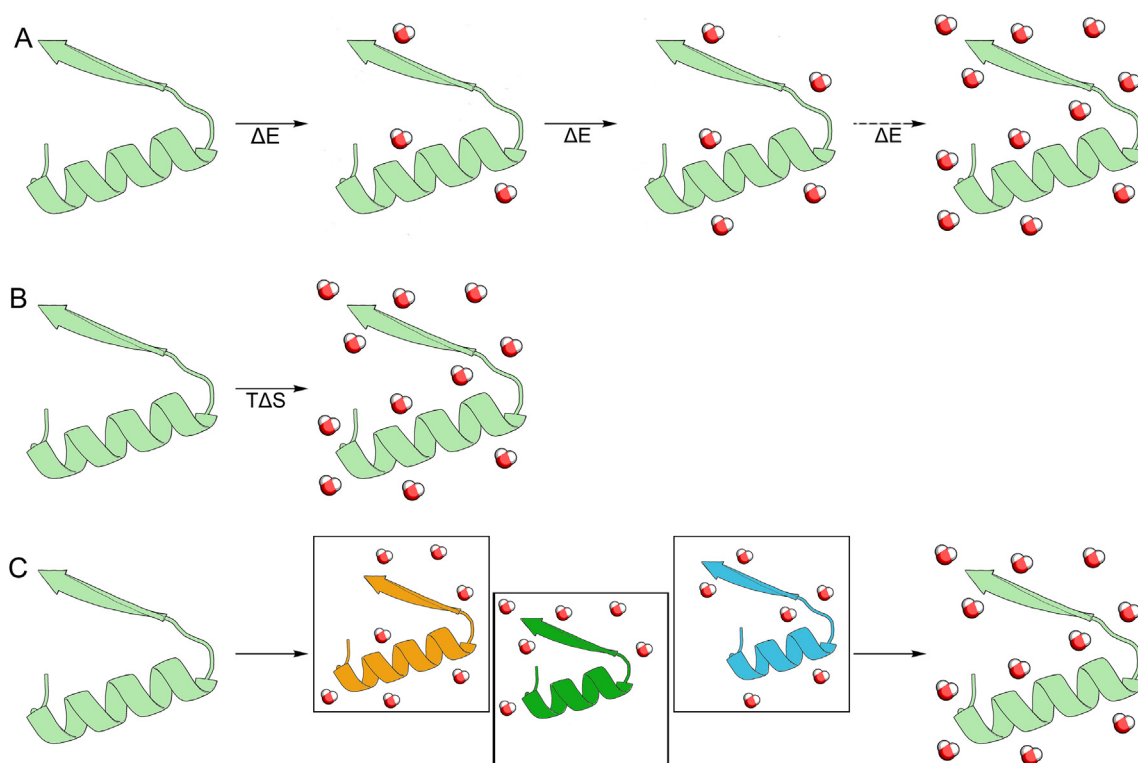


Fig. 1. Strategies for placement of water molecules in the protein's interior. (A) Docking-based, (B) RISM-based, and (C) similarity-based strategies.

silico prepared mutant structures, where substitution of residues enlarges or significantly reshapes internal cavities. However, the user should keep in mind that none of these methods assume flexibility or large conformational changes in the target structure. Also, in the case of preparing a very demanding protein structure, i.e., an unrefined homology model or a sole representative of a particular protein family, the user should be encouraged to avoid similarity-based or docking-based methods and focus on the RISM-based software to properly sample the positions of water molecules.

3. Software for water site detection and ligand binding analysis

Studies of protein–ligand interactions are crucial for a better understanding of the mechanisms of biological processes and their regulation [48]. Water-mediated interactions were found in 85% of 392 analysed protein–ligand complexes. Structural and thermodynamic data indicate that water molecules mediate interactions between proteins and ions, substrates, cofactors, inhibitors and other macromolecules [49,50]. Water molecules are placed methodically within the surroundings and inside the protein, and display a particular structure characterised by the presence of hydration or water sites – regions of high-water density. They act as locations that attract water and can be used to describe water behaviour around chemical molecules [7,51,52]. The hydration/dehydration balance is relevant for protein–ligand formation and binding affinity, involving both entropic and enthalpic contributions [53,54]. During the binding process, water molecules can either be displaced or conserved, bridging the protein–ligand interactions in the latter case [55]. The presence of water molecules in protein binding sites may imply different effects on the energy, entropy and enthalpy of the system. Depending on the situation, such effects may be favourable or unfavourable. For example, in a case where water molecules are trapped in a hydrophobic cavity filled by residues that cannot make appropriate hydrogen bonds, the enthalpic contribution is unfavourable. An opposite situation occurs when water molecules are engaged in forming hydrogen bonds to hydrophilic residues, and here the enthalpic effect may be favourable [56]. The displace-

ment of such water molecules can contribute to the binding free energy, impact affinity during ligands' association and govern enthalpy and entropy partitioning, according to the properties of the individual water molecules compared to those in the bulk phase [49]. Developing a ligand with a high binding affinity towards a specific target is one of the most important steps during the entire drug design process. Thus, a lot of effort is focused on the prediction of whether a water site should be displaced and whether this would cause an increase in a ligand's affinity.

Different approaches, both experimentally-based (i.e., location in crystal structures and B-factors) and knowledge-based (i.e., free energy, water's contribution to binding free energy, entropic contribution), have been reported to assess information about water sites [55]. One of the very first experimentally-based software programs, GRID [57], uses a regular array of 'grid points', established throughout and around the protein, to calculate the energetics of water probes inside a macromolecular binding site (Fig. 2a). GRID places a chemical probe and calculates an empirical interaction energy at all grid points [55]. An approach using crystallographic B-factors to determine which water molecules in a protein's structure are likely to be displaced has been implemented in Consolv [58] and WaterScore [59] software. Using geometric criteria can also indicate the positions of water molecules mediating protein–ligand interactions. Such a procedure has been included in the AcquaAlta program [60] for estimating the propensities of ligand hydration. In HINT software [61], the Gibbs free energy of non-covalent interactions is based on van der Waals interactions and partial atomic partition coefficients. A knowledge-based approach has been implemented, e.g., in AQUARIUS [62] or AQUARIUS2 [63] software. The probable positions for hydration sites are predicted based on solvent distributions surrounding particular amino acids derived from the analysis of a protein's structure. However, most of the software applications mentioned above are not currently used or are used very rarely. This is due to the fact that another class of methods, describing the thermodynamic properties of water by analysing data from MD and MC simulations, became very popular and easy to use.

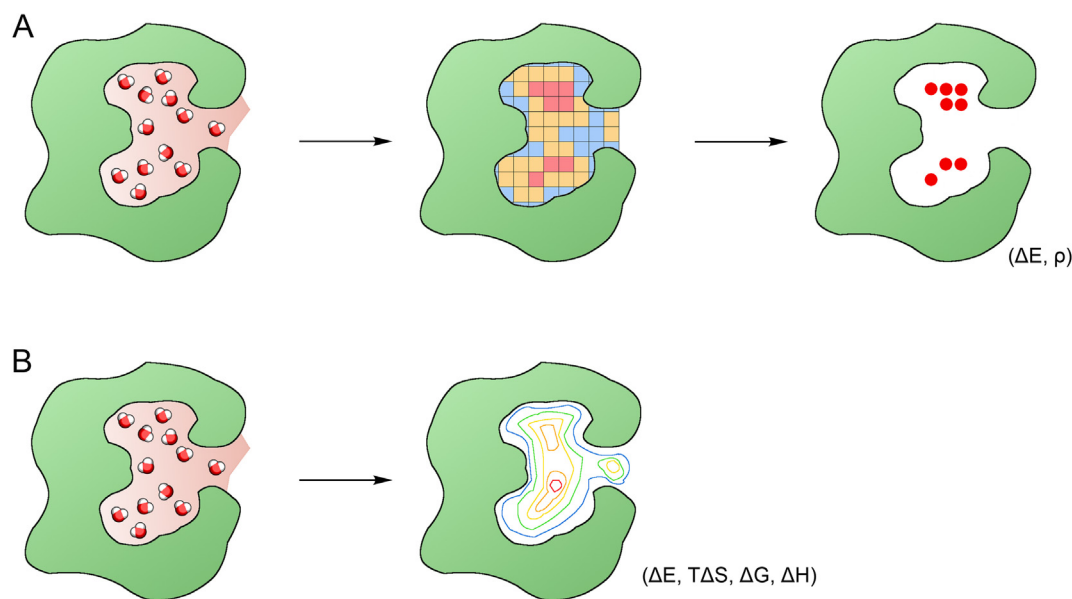


Fig. 2. Strategies of analysis of water sites and ligand binding modes. (A) Strategy using a grid to calculate energetics based on water local distribution, and (B) strategy using IFST (Inhomogeneous Fluid Solvation Theory) to assess the role of structural water molecules by calculating their contribution to the thermodynamics of protein solvation. Grid cells (squares at row A) are coloured according to increasing number of water molecules detected in cells (green – low, red – high). Cells with highest occupancy provide information about the energetically preferred position of the water molecules. Calculated isolines (row B) provide information about the same values of the thermodynamic factors calculated for water molecules in protein cavities. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 2

List of software for analysis of ligand binding and drug design with respect to the water molecules in the binding cavity.

Software	Testing set	Functionality	Remarks
Input – a single structure			
AcquaAlta [60]	trypsin; dihydrofolate reductase; thymidine kinase; VEGFR2; glycogen phosphorylase; human phosphodiesterase; beta trypsin; holo-glyceraldehyde 3P dehydrogenase; HSP90; AmpC beta-lactamase; 2CDK2; ACE; COMT; HIV-1 protease; non-nucleoside adenosine deaminase; ACK1; coagulation factor Xa; EGFR	generating of explicit water molecules at the ligand–protein interface; searching for water molecules interacting with generic functional groups of small organic molecules; generating water molecules bridging interactions between ligand and protein considering the hydration propensities of the involved functional groups and aromatic moieties	available on request: link
FLAP [73]	a set of 23 protein kinase structures	target-based pharmacophores; comparison of multiple protein targets; docking ligands into protein targets	commercial, standalone: link
JAWS [77]	neuraminidase; scytalone dehydratase; Major Urinary Protein 1; bovine β -lactoglobulin; cyclooxygenase-2	determining the optimal placement of water molecules in a binding site; binding free energy estimation	implemented in a modified version of MCPRO, v. 2.1 [87]
SZMAP [75]	HIV-1 protease; neuraminidase; trypsin; factor Xa; scytalone dehydratase; oligopeptide-binding protein (OppA);	computation of binding free energies and the corresponding thermodynamic components for water molecules in the binding site	commercial link
WaterFLAP [74]	adenosine A _{2A} StaR receptor in complex with triazine	generating and scoring water networks for both apo and ligand-complex structures; binding path prediction; lipophilic hot-spot calculation	commercial, standalone: link
WaterScore [59]	<i>cutinase; xylose isomerase; penicillopepsin; galactose/glucose binding protein; proteinase A; rhizopus pepsin; actinidin; DNase I; cholesterol oxidase; RNase A; thermitase; lipid binding protein; Fv fragment of mouse monoclonal antibody D1.3; dihydrofolate reductase</i>	<i>determine conserved water molecule positions; scoring of protein–ligand interactions and determination of ligand binding mode with respect to bound and displaced water molecules</i>	link (currently unavailable)
WATGEN [72]	<i>126 protein–peptide binding interface structures</i>	<i>identification of water sites; selection of the ‘best’ water sites for ligand docking; solvation thermodynamics; binding free energy estimation</i>	<i>no information available</i>
WATsite [71]	<i>three different structures of protein–ligand complexes of factor Xa</i>	<i>identification of water sites; free energy estimation</i>	link (currently unavailable)
WRAPPA [76]	vinculin binding-site; truncated SNARE complex; potassium channel fragment; human relaxin-3; RNA complexed with Rev peptide; Kv1.3 channel blocker Tc32	identification of water sites, referred to as dehydrons	web server: link
WScore [78]	<i>a set of 542 binding sites within 506 protein–ligand complexes, associated with 22 receptors</i>	<i>predicting the location of water sites; producing a detailed map of the water structure and displacement free energies; ligand docking and scoring</i>	<i>no information available</i>
Input – MD simulations			
AQUA-DUCT [79,80] AquaMMapS [85]	<i>Solanum tuberosum epoxide hydrolase [88] casein kinase 2; A_{2A} adenosine receptor</i>	<i>high-density water sites’ and/or co-solvent sites’ identification identification of spatial hot spots within the protein binding site</i>	<i>standalone: link no information available</i>
GIST [68]	Cucurbit[7]uril (CB7); factor Xa	high-density water sites’ identification; map of regions where the solvent interacts favourably with the surface or has unfavourable entropy	implemented in AmberTools
SPAM [84]	HIV-1 protease; hen egg-white lysozyme	qualitative estimation of the thermodynamic profile of water in hydration sites; binding free energy estimation	implemented in AmberTools
SSTMap [81] STOW [70]	Caspase 3 <i>HIV-1 protease–ligand complex; concanavalin A–carbohydrate complexes; cyclophilin A–ligand complexes</i>	identification of water sites <i>computation of contribution of discrete ordered water molecules to the solvation thermodynamics; determine and analyse water sites</i>	link <i>no information available</i>
WATCLUST [69]	AmpC beta-lactamase	determine and analyse water sites	VMD plugin: link the direct transfer of the information to Autodock
Water-swap [82]	neuraminidase in complex with oseltamivir	calculation of binding free energy by water-swap reaction coordinate	part of the Siremol’s Sire application: link
WaterMap [67,89]	streptavidin; Cox-2; antibody DB3; HIV-1 protease	identification of water sites; solvation thermodynamics; entropic and enthalpic contributions to the free energy	commercial, part of the Schrödinger package: link
WatMD [83]	<i>Green Fluorescent Protein; Mannitol 2-Dehydrogenase</i>	<i>identification of water sites</i>	<i>no information available</i>

*Information about currently unavailable software is in *italics*.

Most of the recently developed tools are based on Inhomogeneous Fluid Solvation Theory (IFST) derived in 1998 by Lazardis [64]. IFST is a statistical mechanical method that calculates free energy differences from short MD or MC simulations by quantifying the effect of a solute acting as a perturbation to bulk water. The solute may be different molecules, such as proteins, peptides or other molecules. One of the major advantages of IFST is that the

system is spatially decomposed to consider the contribution of specific regions to the total solvation free energies. The contributions of each individual water molecule to the enthalpy are calculated by computing the average interaction energies, whereas the contributions to the entropy are calculated from intermolecular correlations. The Gibbs free energy equation can be used to calculate the contribution to the free energy from the enthalpy and

entropy. The result is then compared with the contribution of one water molecule to the free energy of bulk water, again calculated using IFST. The results of IFST depend on the accuracy of the force field and the water model that was used: a detailed discussion can be found elsewhere [65,66]. IFST is implemented in WaterMap [67], GIST [68], WATCLUST [69], STOW [70] and WATsite [71] software. Methods based on IFST are limited to the analysis of high-occupancy hydration sites and therefore omit solvent molecules found in lower density regions [68] (Fig. 2b).

Some of the software listed in Table 2 perform calculations based on static input (WATsite, WATGEN [72], FLAP [73], WaterFLAP [74], SZMAP [75], WRAPPA [76], WaterScore [59], AcquaAlta [60], JAWS [77] and WScore [78]), while some of them rely on data obtained from MC or MD simulations (AQUA-DUCT [79,80], GIST [68], WATCLUST [69], STOW [70], SSTMap [81], WaterMap [67], Water-swap [82], WatMD [83], SPAM [84] and AquaMMapS [85]). In static structure-based software, hydration and water sites are determined by investigating potential binding sites through the placement of water probes. The only exception is WATsite, which conducts MD simulations based on the static input structure. In simulation-based software, the system is simulated with explicit water molecules, which are free to explore the system's space. These water molecules are then clustered in hydration sites, and their thermodynamic properties are calculated. Some of the software only identifies the water sites, without any further information, while some can estimate the binding free energy and the corresponding thermodynamic components for water molecules in the binding sites. In a very recent paper, the authors combined WATsite data with neural networks and deep learning to significantly improve the speed of water site identification and the calculation of the free energy contributions [86]. The authors claim that such an approach will allow the inclusion of solvation components, such as water-mediated interactions or enthalpically stable hydration networks in proximity to the protein–ligand complex, in structure-based ligand design.

4. Software for tunnel detection and transportation phenomena analysis

The intramolecular voids inside a protein structure, such as cavities, tunnels, channels and pores, are often important for protein functions [90]. While we have already shown the importance of cavities, this section focuses on the function of tunnels and channels. For proteins with a buried active site, tunnels facilitate substrate entry and enable product egress. Tunnels, as well as the whole protein structure, should not be seen as rigid bodies. In fact, a reasonable degree of flexibility is often required to maintain the catalytic reaction. The geometry and amino acid composition of a particular tunnel determine the shape and chemical properties of a potential ligand. Tunnels are also equipped with a much more sophisticated mechanism of small molecule discrimination – gates. Gates are capable of controlling substrate access to the active site, preventing solvent access to particular protein regions and synchronising processes occurring in distant parts of the protein [91]. Tunnels, pores, gates and cavities constitute a dynamic network inside a protein. Therefore, for proper tunnel detection, a single crystal structure of a protein may be insufficient. MD simulations provide a picture of a protein's movements in an aqueous solution. Reasonably long simulations give insights into the dynamics of the tunnel network. Well-defined tunnels allow fast water exchange over a time of about 10^{-9} s, while transient tunnels extend the required time up to 10^{-3} s. In comparison, the exchange time of water molecules at the protein surface with bulk ones is in the sub-nanosecond range [92]. Therefore, from the computational point of view, the lengths of the required molecular

dynamics simulations depend on the studied system. In the case of buried active sites linked with the solvent *via* a network of tunnels, hundreds of nanoseconds are enough to provide good sampling [88,93–96], whereas to observe the exchange of deeply buried water molecules with the bulk solvent up to as much as tens of milliseconds are necessary [30]. Shorter simulations can provide information about the potential pathways of such an exchange and can suggest mutations that can open an alternative tunnel [80]. The second parameter which might influence the required length of simulations is the frequency of gating phenomena. Gates defined by a single amino acid's rotation require shorter experiments than those defined by, e.g., loops or controlled by proteins' breathing motions [91].

The first tunnel detection software used a geometry-based approach to identify 'empty spaces' inside protein structures [90]. The most successful ones, such as Mole 2.0 [97], CAVER 3.0 [98], and CAVER Analyst 1.0 and 2.0 [99,100], are widely used by the scientific community, predominantly to describe tunnels identified in crystal structures. The most successful strategy employs the construction of a Voronoi diagram to detect and describe voids within the macromolecule [101,102]. Using a defined probe radius and internal cavity identification, the software is able to detect tunnels providing access from the selected area to the surroundings. Such a strategy assumes that the tunnels are a summation of connected cavities and is very often used for the analysis of single crystallographic structures. The structural information obtained on such a basis is mostly incomplete, due to tunnels' flexibility. Moreover, using spherical probes for tunnel exploration provides only an approximation of tunnels to tubes with symmetrical diameters and thus prevents analysis of tunnels' asymmetry. It is also difficult to analyse the regulation and direction of the solvent flow, as well as the contribution of tunnels to an enzyme's activity and selectivity. Some of these weaknesses were targeted in 2014 by a non-spherical approach by Benkaidali et al. [103]; however, due to its complex implementation the tool was rarely used. Results provided by geometry-based tunnel detection software were unable to answer questions about solvent flow direction and how tunnels contribute to this.

To analyse the solvent flow direction and tunnels' contribution to this parameter, we need to concentrate on solvent/ligand analysis. Several different methods have been implemented, based on very diverse approaches (Fig. 3). The first attempt was made in 2008 by Bidmon et al. [104], who introduced the Visual Abstractions of Solvent Pathlines method. The pathways of solvent molecules passing through the particular region of interest (so-called ROI) were pre-processed and visualised as Bézier curves. The next attempt at such an analysis was carried out in 2010 by Vasiliev et al. [105]. Their streamline tracing method was applied to photosystem II and was used to visualise water flux in particular regions of the protein. However, the results of the calculations are hard to interpret, and only a few applications of this method can be found in the literature [106]. In 2014, Benson and Pleiss proposed a solvent flux method to study water influx in the *Candida antarctica* lipase B protein cavity from the triglyceride–water interface [107]. They introduced a solvent concentration gradient and the reorientation and rescaling of the velocity vectors of selected water molecules in order to accelerate the influx and increase the probability of rare events in the study. Similarly to widely used strategies (aMD, REMD, SMD and RAMD), it was applied to investigate rare events in a reasonable computing time range (e.g., it overcame the significant energy barriers of slow biophysical events). In contrast to known methods, this technique allowed the flow of multiple molecules, including the selected solvent molecules, to be precisely investigated during a single simulation. Since artificial external forces are introduced to classic MD simulations, one could be concerned about misleading biases and the complicated proto-

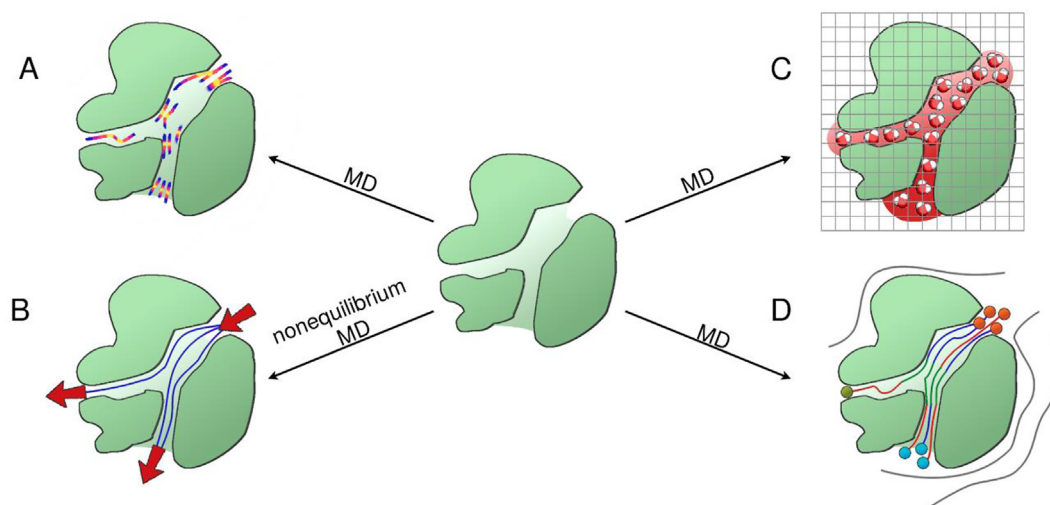


Fig. 3. Strategies for tunnel detection and description based on water molecule analysis. (A) Streamline tracing, (B) Solvent Flux, (C) *trj_cavity*, and (D) AQUA-DUCT water tracking approach.

Table 3

List of software and methods for tunnel detection and transportation phenomena observation.

Software or method	Applicability System	Functionality	Remarks
AQUA-DUCT [79,80]	<i>Mus musculus</i> epoxide hydrolase <i>D</i> -amino-acid oxidase [111] <i>Pyrococcus furiosus</i> phosphoglucose isomerase [93]	water molecule tracking tunnel and gating residue detection	can be used also for cavity and hot-spot detection Standalone: link
Solvent flux method [107]	Claudin-2 ion channel [112] <i>Candida antarctica</i> lipase B	water molecule tracking and occupancy analysis in the internal cavity ion transportation pathways identification <i>identification of water access pathway; hot-spot identification</i>	<i>based on an artificial gradient. Code not available.</i> visual analysis only; code available on request
Streamline tracing [105]	photosystem II squalene-hopene cyclase [106]	fibre tracing; tunnel detection; gating residue (access control points) identification changes in water flow after introducing amino acid substitution	visual analysis only; code available on request
trj_cavity [108]*	Der p 2 protein; TM pore; pullulanase polydicyclopentadiene [114]; herkinorin [115]; glycidoxypropyltrimethoxy silane [116]; mammalian translocator membrane protein [117]; human G-protein coupled receptors [118]; amyloid fibrils [119]; sperm whale myoglobin [120]; 07A metalloprotease [121]; human erythrocyte anion exchanger 1 (Band 3 protein) [122]; amorphous silica [123]; full-length TLR4 dimer [124]; profilin [125]; human serum albumin [126]; laccase [127]; dengue capsid protein (C protein) [128]; horse-radish peroxidase; lactoperoxidase [129]; OmpC–MlaA complex [130–132]; MATE transporter [133] cholesteryl ester transfer protein [134]; acyl carrier proteins [135] mouse myoglobin [109]	generating the trajectory of discovered cavities, quantification of time-dependent cavity volume, solvent presence inside a particular cavity; tunnel detection cavity analysis time-dependent cavity analysis	implemented in GROMACS: link
Visual Abstractions of Solvent Pathlines [104]	<i>TEM</i> β -lactamase [136]	analysis of the movement of ligands, movements within the cavities and tunnels of proteins <i>identification of the role of water in gating loop flexibility</i>	<i>visual analysis only; code not available</i>
Watergate [113]	haloalkane dehalogenase mutants	visualisation of water molecule trajectories	visual analysis only; code available on request

*For *trj_cavity* software only recent applications are presented (2017–2019). Information about currently unavailable software is in *italics*.

col that is dependent on the system. In 2014 another software program emerged as a GROMACS plugin, called *trj_cavity* [108]. *Trj_cavity* is capable of cavity and tunnel identification together with time-dependent calculations of their volume and solvent capacity. In the vast majority of research papers, *trj_cavity* is used only for

the identification of cavities and calculating their volumes and occupancy, while only one study was found where the authors used *trj_cavity* to actually trace ligands [109]. The existing gap between tools searching for tunnels and pathways, and advanced tools for accelerated water flux investigations was filled in 2017

by AQUA-DUCT [79], an easy-to-use tool facilitating analysis of the behaviour of water (and, if necessary, other solvent molecules) penetrating any selected region in a protein. AQUA-DUCT comprises a *Valve* module, which is capable of tracking water molecules, clustering their trajectories and enabling visualisation in PyMOL [110]. The *Valve* module was used to investigate relatively small proteins, such as D-amino-acid oxidase (DAAO) [111] and *Pyrococcus furiosus* phosphoglucose isomerase [93], as well as ion channels such as claudins [112]. In contrast to the geometry-based approach, water tracking analysis provides information about tunnels' functionality, allows their permeability to be compared and facilitates the detection of the gating residues controlling access to the binding cavity. At the same time, Watergate, a software application for statistical overview of the overall solvent flow, water trajectory clustering, and visualisation was developed [113]. The software programs using water molecules for tunnel detection are listed in Table 3.

Tools based on water molecules as a molecular probe for tunnel detection (listed in Table 3) can provide much more complex information about proteins than simple geometry-based methods. Since they are focused on the information provided by the solvent itself, they also take into account the physicochemical properties of the solute. Such information is useful for examining the effects of the introduced mutation on the solvent flow and thereby the enzyme's activity. By using the pathways of the solvent molecules, the user is able to identify the key residues important for the enzyme's activity and selectivity, and the amino acids that contribute to gating residues and control small ligands' entry/egress. These tools can additionally facilitate the description of cavity shape evolution during simulation time, which can be used for inhibitor design or hot-spot detection for substrate specificity modification, and also the identification of residues distant from the active site which contribute to the activity and selectivity, and thus can be considered as a safe alternative to smart mutant library design. However, it should be kept in mind that to properly sample events such as substrate entrance, product release or the exchange of the trapped solvent molecules with the bulk solvent, the analysed simulations must be of reasonable length and conducted in physiological-like conditions (please see the Summary and outlook section for more details).

5. Summary and outlook

The important role of water molecules in structural biology is reflected by a large number of different software programs dedicated to various types of water-molecule-based analysis. Most of the software presented here is focused only on particular aspects of water's presence in a macromolecular structure, such as its contribution to protein stability, ligand binding and drug design, or cavity and tunnel description.

Among all the described software, the role of software in optimising water placement inside proteins' cavities is probably the most underestimated, although placing water molecules is not a trivial task. Three different strategies, RISM theory, the docking of water molecules, and the analysis of conserved water molecules among similar proteins, are used and complement each other. The RISM-based software probably provides the most accurate model; however, it is time-consuming. Docking-based methods are the fastest; however, they may provide biased results for systems, such as metalloproteins, proteins with large cavities and protein-nucleic acids complexes, that are problematic for such software. Both methods can be considered for the prediction of differences in water rearrangement when mutant structures are designed. The third strategy requires a collection of similar structures and may

not be sensitive enough to provide correct predictions when a single mutation occurs. Therefore, depending on the investigated system, different strategies are optimal for water placement and can provide reliable starting points for molecular dynamics studies. As already stated, the water placement method should be considered as a standard approach for homology models or structures with introduced amino acid modifications.

Concerning the role of water in the description of ligand binding, the most commonly used methods are based on IFST. Given a water site, the software can predict how much free energy is gained (or lost) by displacing the water molecules that occupied the potential ligand-binding site. The solvent contribution cannot be neglected, as was shown in several excellent papers [137–139], and therefore continuous progress in both the accuracy and parallel analysis of alternative states is highly desired. Binding enthalpies and entropies of water molecules may also be calculated based on Grid Cell Theory (GCT) [140]. This is a recently developed method for investigating hydration thermodynamics from a molecular dynamics or Monte Carlo standpoint. In the GCT approach, the density, enthalpy, entropy and free energy of water are evaluated for an arbitrary region of space around a system of interest. These parameters refer to the water molecules which enter a particular hydration site of the protein(s) from the bulk concentration. However, this theory has not yet been implemented in known software. The recent AQUA-DUCT version provides an approach combining information on water and co-solvent high-occupancy sites. It can be used for pharmacophore design and suggests directions for future software development. All of the abovementioned methods can provide additional support to drug design and provide more accurate results in comparison to methods neglecting water molecules' contribution.

In the third group of software, water molecules are used as a molecular probe to sample the 'empty spaces' in proteins during molecular dynamics simulations to detect tunnels and cavities. This field is so far monopolised by a widely used geometry-based approach which on the one hand is very simple, but provides rather approximate results. It neglects the tunnels' asymmetry and the physicochemical properties of the tunnel-lining amino acids. Therefore, it is difficult to use such software for analysis of tunnels' functionality. The alternative approaches presented in our review are a most diverse group of software utilising water as a molecular probe. Depending on the implemented algorithms, they provide information about local water flow changes (such as the streamline tracing method), changes in cavity volumes (*trj_cavity*) or can provide a holistic picture of water flow *via* tunnel networks and an approximation of the energy profiles of particular pathways (AQUA-DUCT). The utilisation of water for cavity and tunnel description removes most of the limitations of the standard approach. It seems that using water molecules as a molecular probe enables more sophisticated analysis of the substrate transportation network provided by tunnels, handling tunnels and cavities together and describing the protein interior in a holistic way as a single entity. However, so far it is hard to provide an estimation of how accurate they are. This problem is caused not only by difficulties in experimental verification of their findings, nor the question of how accurately the hydrophobic cavities can be described, but also due to the lack of benchmarks for the performance inside protein structures of the different water models used in MD simulations.

Protein engineering is one of the most promising, but still largely unexplored, fields of application for software focused on the analysis of water as a molecular probe. So far, most of the examples of such studies are focused on understanding the changes introduced by mutant proteins' construction. However, there are papers showing the potential applicability of the water-based approach

for hot-spot detection [88] and mutant library design [111]. Successful verification of such a strategy can greatly facilitate protein engineering and provide an interesting and easy-to-apply technique. Also, using the information on the water molecules' (or small ligands, or other types of solvent molecules) tracking, the user gains knowledge on a protein's internal architecture, which might be used to develop a successful strategy for further modifications; for instance, to search for more potent inhibitors which will explore previously unused cavities, or to improve the protein's activity and/or selectivity by adjusting the pathways leading to and from the active site.

Since the analysis of water-mediated interactions has become of greater interest, we hope that the number and quality of software programs using water molecules to analyse macromolecules' properties will only increase. However, progress in this promising area cannot be achieved without the further joint efforts of theoreticians and experimentalists. One needs to consider that the majority of the tools described above depend on water models and force fields (e.g., tools based on Inhomogeneous Fluid Solvation Theory or benefiting from MD simulations). Both force fields and water models are being constantly upgraded to provide more accurate descriptions of studied systems. For example, the most recent papers of Huang, et al. provide force fields which can be used for both ordered and disordered proteins [141]. Recent four-point water models have improved the description of its thermodynamic properties; however, water molecules' non-bonded interactions still require validation [142]. Nevertheless, the majority of computational studies employ simple non-polarizable models of water (e.g., TIP3P, SPC/E, TIP4P) and assume that they will describe water molecules in macromolecular surroundings equally well as in the bulk water. Unfortunately, there is no study that can confirm such a presupposition, simply due to the limited access to experimental data providing insight into water's behaviour inside a protein's core. Moreover, even the benchmark analysis of a particular software data's dependency on the used parameters is very limited. As we mentioned above, the comparison of the IFST results obtained with different water models suggests that the quantitative application of IFST to biological systems is strictly model-dependent and has to be carefully analysed. Fortunately, several successful verifications of the findings guided by the software developed to analyse the behaviour and/or properties of water molecules have accelerated research in the field of protein research and each year bring to the scientific community new, optimised, versatile and reliable tools which greatly improve our understanding of nature.

Funding

This work was funded by the National Science Centre, Poland, grant no DEC-2013/10/E/NZ1/00649 and DEC-2015/18/M/NZ1/00427.

Conflict of interest

None declared.

CRediT authorship contribution statement

Karolina Mitusińska: Resources, Data curation, Writing - original draft, Writing - review & editing, Visualization. **Agata Raczyńska:** Resources, Data curation, Writing - original draft, Investigation, Visualization. **Maria Bzówka:** Resources, Data curation, Writing - original draft, Investigation, Writing - review & editing. **Weronika Bagrowska:** Resources, Data curation, Writing - original draft, Investigation. **Artur Góra:** Conceptualization, Super-

vision, Funding acquisition, Project administration, Writing - review & editing.

References

- [1] Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular biology of the cell. 4th ed. New York: Garland Science; 2002.
- [2] Vallone B, Brunori M. Roles for holes: are cavities in proteins mere packing defects? *Ital J Biochem* 2004;53(1):46–52.
- [3] J. A. Rupley, G. Careri, Protein hydration and function, 1991, 37–172.
- [4] Biedermannová L, Schneider B. Hydration of proteins and nucleic acids: advances in experiment and theory. A review. *Biochim Biophys Acta - Gen Subj* 2016;1860(9):1821–35.
- [5] Mikol V, Papageorgiou C, Borer X. The role of water molecules in the structure-based design of (5-hydroxynorvaline)-2-cyclosporin: synthesis, biological activity, and crystallographic analysis with cyclophilin A. *J Med Chem* 1995;38(17):3361–7.
- [6] Papoian GA, Ulander J, Eastwood MP, Luthey-Schulten Z, Wolynes PG. Water in protein structure prediction. *Proc Natl Acad Sci U S A* 2004;101(10):3352–7.
- [7] Nakasako M. Water-protein interactions from high-resolution protein crystallography. *Philos Trans R Soc London, Ser B Biol Sci*, 2004;359(1448):1191–206.
- [8] Niimura N, Chatake T, Kurihara K, Maeda M. Hydrogen and hydration in proteins. *Cell Biochem Biophys* 2004;40(3):351–70.
- [9] Savage H, Wlodawer A. Determination of water structure around biomolecules using X-ray and neutron diffraction methods. 1986, 162–183.
- [10] Levitt M, Park BH. Water: now you see it, now you don't. *Structure* 1993;1(4):223–6.
- [11] Carugo O, Bordo D. How many water molecules can be detected by protein crystallography? *Acta Crystallogr D Biol Crystallogr* 1999;55(2):479–83.
- [12] Berman HM. The Protein Data Bank. *Nucleic Acids Res* 2000;28(1):235–42.
- [13] Imai T, Hiraoka R, Kovalenko A, Hirata F. Locating missing water molecules in protein cavities by the three-dimensional reference interaction site model theory of molecular solvation. *Proteins Struct Funct Bioinf* 2006;66(4):804–13.
- [14] Hopmann KH, Himo F. Theoretical study of the full reaction mechanism of human soluble epoxide hydrolase. *Chemistry* 2006;12(26):6898–909.
- [15] Sheng X, Lind MES, Himo F. Theoretical study of the reaction mechanism of phenolic acid decarboxylase. *FEBS J* 2015;282(24):4703–13.
- [16] Abrioux C, Coasne B, Maurin G, Henn F, Jeffroy M, Boutin A. Cation behavior in Faujasite zeolites upon water adsorption: a combination of Monte Carlo and molecular dynamics simulations. *J Phys Chem C* 2009;113(24):10696–705.
- [17] Paquet E, Viktor HL. Molecular dynamics, Monte Carlo simulations, and Langevin dynamics: a computational review. *Biomed Res Int* 2015;2015:1–18.
- [18] Levy Y, Onuchic JN. Water and proteins: a love-hate relationship. *Proc Natl Acad Sci U S A* 2004;101(10):3325–6.
- [19] Lavecchia A, Giovanni C. Virtual screening strategies in drug discovery: a critical review. *Curr Med Chem* 2013;20(23):2839–60.
- [20] Sousa SF et al. Protein-ligand docking in the new millennium – A retrospective of 10 years in the field. *Curr Med Chem* 2013;20(18):2296–314.
- [21] Pagadala NS, Syed K, Tuszynski J. Software for molecular docking: a review. *Biophys Rev* 2017;9(2):91–102.
- [22] Finer-Moore JS, Kossiakoff AA, Hurley JH, Earnest T, Stroud RM. Solvent structure in crystals of trypsin determined by X-ray and neutron diffraction. *Proteins Struct Funct Genet* 1992;12(3):203–22.
- [23] Nagendra HG, Sukumar N, Vijayan M. Role of water in plasticity, stability, and action of proteins: the crystal structures of lysozyme at very low levels of hydration. *Proteins Struct Funct Genet* 1998;32(2):229–40.
- [24] Takano K, Yamagata Y, Yutani K. Buried water molecules contribute to the conformational stability of a protein. *Protein Eng Des Sel* 2003;16(1):5–9.
- [25] Meyer E. Internal water molecules and H-bonding in biological macromolecules: a review of structural features with functional implications. *Protein Sci* 1992;1(12):1543–62.
- [26] Zhou Y, Morais-Cabral JH, Kaufman A, MacKinnon R. Chemistry of ion coordination and hydration revealed by a K⁺ channel-Fab complex at 2.0 Å resolution. *Nature* 2001;414(6859):43–8.
- [27] Tanimoto T, Furutani Y, Kandori H. Structural changes of water in the Schiff base region of bacteriorhodopsin: proposal of a hydration switch model. *Biochemistry* 2003;42(8):2300–6.
- [28] Karplus PA, Faerman C. Ordered water in macromolecular structure. *Curr Opin Struct Biol* 1994;4(5):770–6.
- [29] Otting G, Liepinsh E, Wuthrich K. Protein hydration in aqueous solution. *Science* (80-) 1991;254(5034):974–80.
- [30] Persson F, Halle B. Transient access to the protein interior: simulation versus NMR. *J Am Chem Soc* 2013;135(23):8735–48.
- [31] Morozenko A, Leontyev IV, Stuchebrukhov AA. Dipole moment and binding energy of water in proteins from crystallographic analysis. *J Chem Theory Comput* 2014;10(10):4618–23.
- [32] Ross GA, Morris GM, Biggin PC. Rapid and accurate prediction and scoring of water molecules in protein binding sites. *PLoS ONE* 2012;7(3):e32036.
- [33] Sindhikara DJ, Hirata F. Analysis of biomolecular solvation sites by 3D-RISM theory. *J Phys Chem B* 2013;117(22):6718–23.

- [34] Fusani L, Wall I, Palmer D, Cortes A. Optimal water networks in protein cavities with GAsol and 3D-RISM. *Bioinformatics* 2018;34(11):1947–8.
- [35] Sindhikara DJ, Yoshida N, Hirata F. Placevent: an algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *J Comput Chem* 2012;33(18):1536–43.
- [36] Patel H, Gruning BA, Gunther S, Merfort I. PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering. *Bioinformatics* 2014;30(20):2978–80.
- [37] Jukić M, Konc J, Gobec S, Janežič D. Identification of conserved water sites in protein structures for drug design. *J Chem Inf Model* 2017;57(12):3094–103.
- [38] 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* 2009;1–18.
- [39] Morozenko A, Stuchebrukhov AA. Dowser++, a new method of hydrating protein structures. *Proteins Struct Funct Bioinf* 2016;84(10):1347–57.
- [40] Imai T, Hiraoka R, Kovalenko A, Hirata F. Water molecules in a protein cavity detected by a statistical–mechanical theory. *J Am Chem Soc* 2005;127(44):15334–5.
- [41] Stumpe MC, Blinov N, Wishart D, Kovalenko A, Pande VS. Calculation of local water densities in biological systems: a comparison of molecular dynamics simulations and the 3D-RISM-KH molecular theory of solvation. *J Phys Chem B* 2011;115(2):319–28.
- [42] Palmer DS, Frolov AI, Ratkova EL, Fedorov MV. Towards a universal method for calculating hydration free energies: a 3D reference interaction site model with partial molar volume correction. *J Phys: Condens Matter* 2010;22(49):492101.
- [43] Nikolić D, Blinov N, Wishart D, Kovalenko A. 3D-RISM-Dock: A new fragment-based drug design protocol. *J Chem Theory Comput* 2012;8(9):3356–72.
- [44] Truchon J-F, Pettitt BM, Labute P. A cavity corrected 3D-RISM functional for accurate solvation free energies. *J Chem Theory Comput* 2014;10(3):934–41.
- [45] Konc J, Janežič D. ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics* 2010;26(9):1160–8.
- [46] Bernstein FC et al. The protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 1977;112(3):535–42.
- [47] Sridhar A, Ross GA, Biggin PC. Waterdock 2.0: water placement prediction for Holo-structures with a pymol plugin. *PLoS ONE* 2017;12(2):e0172743.
- [48] Du X et al. Insights into protein–ligand interactions: mechanisms, models, and methods. *Int J Mol Sci* 2016;17(2):144.
- [49] Schiebel J et al. Intriguing role of water in protein–ligand binding studied by neutron crystallography on trypsin complexes. *Nat Commun* 2018;9(1):3559.
- [50] Lu Y, Wang R, Yang C-Y, Wang S. Analysis of ligand-bound water molecules in high-resolution crystal structures of protein–ligand complexes. *J Chem Inf Model* 2007;47(2):668–75.
- [51] Henchman RH, McCammon JA. Structural and dynamic properties of water around acetylcholinesterase. *Protein Sci* 2009;11(9):2080–90.
- [52] Irwin BWJ, Vukovic S, Payne MC, Huggins DJ. Large-scale study of hydration environments through hydration sites. *J Phys Chem B* 2019;123(19):4220–9.
- [53] Chervenak MC, Toone EJ. A direct measure of the contribution of solvent reorganization to the enthalpy of binding. *J Am Chem Soc* 1994;116(23):10533–9.
- [54] Klebe G, Böhm H-J. Energetic and entropic factors determining binding affinity in protein–ligand complexes. *J Recept Signal Transduct* 1997;17(1–3):459–73.
- [55] de Beer S, Vermeulen N, Oostenbrink C. The role of water molecules in computational drug design. *Curr Top Med Chem* 2010;10(1):55–66.
- [56] Uehara S, Tanaka S. AutoDock-GIST: incorporating thermodynamics of active-site water into scoring function for accurate protein–ligand docking. *Molecules* 2016;21(11):1604.
- [57] Goodford PJ. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J Med Chem* 1985;28(7):849–57.
- [58] Raymer ML, Sanschagrin PC, Punch WF, Venkataraman S, Goodman ED, Kuhn LA. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J Mol Biol* 1997;265(4):445–64.
- [59] García-Sosa AT, Mancera RL, Dean PM. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein–ligand complexes. *J Mol Model* 2003;9(3):172–82.
- [60] Rossato G, Ernst B, Vedani A, Smieško M. AcquaAlta: a directional approach to the solvation of ligand–protein complexes. *J Chem Inf Model* 2011;51(8):1867–81.
- [61] Eugene Kellogg G, Abraham DJ. Hydrophobicity: is $\text{LogP}(o/w)$ more than the sum of its parts? *Eur J Med Chem* 35(7–8), 651–661.
- [62] Pitt WR, Goodfellow JM. Modelling of solvent positions around polar groups in proteins. *Protein Eng Des Sel* 1991;4(5):531–7.
- [63] Pitt WR, Murray-Rust J, Goodfellow JM. AQUARIUS2: knowledge-based modeling of solvent sites around proteins. *J Comput Chem* 1993;14(9):1007–18.
- [64] Lazaridis T. Inhomogeneous fluid approach to solvation thermodynamics. 1. Theory. *J Phys Chem B* 1998;102(18):3531–41.
- [65] Huggins DJ, Payne MC. Assessing the accuracy of inhomogeneous fluid solvation theory in predicting hydration free energies of simple solutes. *J Phys Chem B* 2013;117(27):8232–44.
- [66] Hess B, van der Vegt NFA. Hydration thermodynamic properties of amino acid analogues: a systematic comparison of biomolecular force fields and water models. *J Phys Chem B* 2006;110(35):17616–26.
- [67] Cappel D, Sherman W, Beuming T. Calculating water thermodynamics in the binding site of proteins – applications of WaterMap to drug discovery. *Curr Top Med Chem* 2017;17(23).
- [68] Ramsey S, Nguyen C, Salomon-Ferrer R, Walker RC, Gilson MK, Kurtzman T. Solvation thermodynamic mapping of molecular surfaces in AmberTools: GIST. *J Comput Chem* 2016;37(21):2029–37.
- [69] López ED et al. WATCLUST: a tool for improving the design of drugs based on protein–water interactions: Fig. 1. *Bioinformatics* 2015;31(22):3697–9.
- [70] Li Z, Lazaridis T. Computing the thermodynamic contributions of interfacial water, 2012, 393–404.
- [71] Hu B, Lill MA. WATsite: hydration site prediction program with PyMOL interface. *J Comput Chem* 2014;35(16):1255–60.
- [72] Bui H-H, Schiewe AJ, Haworth IS. WATGEN: an algorithm for modeling water networks at protein–protein interfaces. *J Comput Chem* 2007;28(14):2241–51.
- [73] Baroni M, Cruciani G, Sciabola S, Perruccio F, Mason JS. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J Chem Inf Model* 2007;47(2):279–94.
- [74] Mason JS, Bortolato A, Weiss DR, Deflorian F, Tehan B, Marshall FH. High end GPCR design: crafted ligand design and druggability analysis using protein structure, lipophilic hotspots and explicit water networks. *Silico Pharmacol* 2013;1(1):23.
- [75] Bayden AS, Moustakas DT, Joseph-McCarthy D, Lamb ML. Evaluating free energies of binding and conservation of crystallographic waters using SZMAP. *J Chem Inf Model* 2015;55(8):1552–65.
- [76] Fraser CM, Fernandez A, Scott LR, WRAPPA: A Screening Tool for Candidate Dehydrogenation, 2011.
- [77] Michel J, Tirado-Rives J, Jorgensen WL. Prediction of the Water Content in Protein Binding Sites. *J Phys Chem B* 2009;113(40):13337–46.
- [78] Murphy RB et al. WScore: A Flexible and Accurate Treatment of Explicit Water Molecules in Ligand–Receptor Docking. *J Med Chem* May 2016;59(9):4364–84.
- [79] Magdziarz T et al. AQUA-DUCT: a ligands tracking tool. *Bioinformatics* 2017;33(13):2045–6.
- [80] Magdziarz T et al. AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective. *Bioinformatics* 2019. <https://doi.org/10.1093/bioinformatics/btz946>.
- [81] Haider K, Cruz A, Ramsey S, Gilson MK, Kurtzman T. Solvation structure and thermodynamic mapping (SSTMap): an open-source, flexible package for the analysis of water in molecular dynamics trajectories. *J Chem Theory Comput* 2018;14(1):418–25.
- [82] Woods CJ, Malaisree M, Hannongbua S, Mulholland AJ. A water-swap reaction coordinate for the calculation of absolute protein–ligand binding free energies. *J Chem Phys* 2011;134(5):054114.
- [83] Velez-Vega C, McKay DJJ, Aravamathan V, Pearlstein R, Duca JS. Time-averaged distributions of solute and solvent motions: exploring proton wires of GFP and PfM2DH. *J Chem Inf Model* 2014;54(12):3344–61.
- [84] Cui G, Swails JM, Manas ES. SPAM: a simple approach for profiling bound water molecules. *J Chem Theory Comput* 2013;9(12):5539–49.
- [85] Cuzzolin A, Deganutti G, Salmaso V, Sturlese M, Moro S. AquaMMAPS: an alternative tool to monitor the role of water molecules during protein–ligand association. *ChemMedChem* 2018;13(6):522–31.
- [86] Ghanbarpour A, Mahmoud A, Lill M. On-the-fly prediction of protein hydration densities and free energies using deep learning. 2020.
- [87] Jorgensen WL, Tirado-Rives J. Molecular modeling of organic and biomolecular systems using BOSS and MCPRO. *J Comput Chem* 2005;26(16):1689–700.
- [88] Mitusińska K, Magdziarz T, Bzówka M, Stańczak A, Gora A. Exploring Solanum tuberosum epoxide hydrolase internal architecture by water molecules tracking. *Biomolecules* 2018;8(4):143.
- [89] Yang Y, Lightstone FC, Wong SE. Approaches to efficiently estimate solvation and explicit water energetics in ligand binding: the use of WaterMap. *Expert Opin Drug Disc* 2013;8(3):277–87.
- [90] Brezovsky J, Chovancova E, Gora A, Pavelka A, Biedermannova L, Damborsky J. Software tools for identification, visualization and analysis of protein tunnels and channels. *Biotechnol Adv* 2013;31(1):38–49.
- [91] Gora A, Brezovsky J, Damborsky J. Gates of enzymes. *Chem Rev* 2013;113(8):5871–923.
- [92] Denisov VP, Halle B. Protein hydration dynamics in aqueous solution. *Faraday Discuss* 1996;103:227.
- [93] Subramanian K et al. Distant Non-obvious mutations influence the activity of a hyperthermophilic *Pyrococcus furiosus* phosphoglucose isomerase. *Biomolecules* 2019;9(6):212.
- [94] Brezovsky J et al. Engineering a de Novo Transport Tunnel. *ACS Catal* 2016;6(11):7597–610.
- [95] Fischer A, Don CG, Smieško M. Molecular dynamics simulations reveal structural differences among allelic variants of membrane-anchored cytochrome P450 2D6. *J Chem Inf Model* 2018;58(9):1962–75.
- [96] Watanabe G, Nakajima D, Hiroshima A, Suzuki H, Yoneda S. Analysis of water channels by molecular dynamics simulation of heterotetrameric sarcosine oxidase. *Biophys. Physicobiol* 2015;12:131–8.

- [97] Sehna D et al. MOLE 2.0: advanced approach for analysis of biomacromolecular channels. *J Cheminf* 2013;5(1):39.
- [98] Chovanova E et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput Biol* 2012;8(10):e1002708.
- [99] Kozlikova B et al. CAVER Analyst 1.0: graphic tool for interactive visualization and analysis of tunnels and channels in protein structures. *Bioinformatics* 2014;30(18):2684–5.
- [100] Jurcik A et al. CAVER Analyst 2.0: analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics* 2018;34(20):3586–8.
- [101] Kim D-S, Sugihara K. Tunnels and voids in molecules via voronoi diagram. In: 2012 Ninth International Symposium on Voronoi Diagrams in Science and Engineering. p. 138–43.
- [102] Pavelka A, Sebestova E, Kozlikova B, Brezovsky J, Sochor J, Damborsky J. CAVER: algorithms for analyzing dynamics of tunnels in macromolecules. *IEEE/ACM Trans Comput Biol Bioinf* 2016;13(3):505–17.
- [103] Benkaidali L et al. Computing cavities, channels, pores and pockets in proteins from non-spherical ligands models. *Bioinformatics* 2014;30(6):792–800.
- [104] Bidmon K, Grottel S, Bös F, Pleiss J, Ertl T. Visual abstractions of solvent pathlines near protein cavities. *Comput Graph Forum* 2008;27(3):935–42.
- [105] Vassiliev S, Comte P, Mahboob A, Bruce D. Tracking the flow of water through photosystem II using molecular dynamics and streamline tracing. *Biochemistry* 2010;49(9):1873–81.
- [106] Gustafsson C, Vassiliev S, Kürten C, Syrén P-O, Brinck T. MD simulations reveal complex water paths in squalene-hopene cyclase: tunnel-obstructing mutations increase the flow of water in the active site. *ACS Omega* 2017;2(11):8495–506.
- [107] Benson SP, Pleiss J. Solvent flux method (SFM): a case study of water access to *Candida antarctica* lipase B. *J Chem Theory Comput* 2014;10(11):5206–14.
- [108] Paramo T, East A, Garzón D, Ulmschneider MB, Bond PJ. Efficient characterization of protein cavities within molecular simulation trajectories: trj_cavity. *J Chem Theory Comput* 2014;10(5):2151–64.
- [109] Chintapalli SV, Anishkin A, Adams SH. Exploring the entry route of palmitic acid and palmitoylcarnitine into myoglobin. *Arch Biochem Biophys* 2018;655:56–66.
- [110] The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.
- [111] Subramanian K et al. Modulating D-amino acid oxidase (DAAO) substrate specificity through facilitated solvent access. *PLoS ONE* 2018;13(6):e0198990.
- [112] Irudayanathan FJ, Wang X, Wang N, Willsey SR, Seddon IA, Nangia S. Self-assembly simulations of classic claudins—insights into the pore structure, selectivity, and higher order complexes. *J Phys Chem B* 2018;122(30):7463–74.
- [113] Vad V et al. Watergate: visual exploration of water trajectories in protein dynamics. *Vcbm* 2017:33–42.
- [114] Elder RM, Long TR, Bain ED, Lenhart JL, Sirk TW. Mechanics and nanovoid nucleation dynamics: effects of polar functionality in glassy polymer networks. *Soft Matter* 2018;14(44):8895–911.
- [115] Marmolejo-Valencia AF, Martínez-Mayorga K. Allosteric modulation model of the mu opioid receptor by herkinorin, a potent not alkaloidal agonist. *J Comput Aided Mol Des* 2017;31(5):467–82.
- [116] Chowdhury SC, Elder RM, Sirk TW, van Duin ACT, Gillespie JW. Modeling of glycidoxypropyltrimethoxy silane compositions using molecular dynamics simulations. *Comput Mater Sci* 2017;140:82–8.
- [117] Zeng J et al. Structural prediction of the dimeric form of the mammalian translocator membrane protein TSPO: a key target for brain diagnostics. *Int J Mol Sci* 2018;19(9):2588.
- [118] Cao R, Giorgetti A, Bauer A, Neumaier B, Rossetti G, Carloni P. Role of extracellular loops and membrane lipids for ligand recognition in the neuronal adenosine receptor type 2A: an enhanced sampling simulation study. *Molecules* 2018;23(10):2616.
- [119] Choi H, Chang HJ, Lee M, Na S. Characterizing structural stability of amyloid motif fibrils mediated by water molecules. *ChemPhysChem* 2017;18(7):817–27.
- [120] Espinosa YR, Caffarena ER, Grigera JR. The role of hydrophobicity in the cold denaturation of proteins under high pressure: a study on apomyoglobin. *J Chem Phys* 2019;150(7):075102.
- [121] Polêto MD, Alves MP, Ligabue-Braun R, Eller MR, De carvalho AF. Role of structural ions on the dynamics of the *Pseudomonas fluorescens* 07A metalloprotease. *Food Chem., Jul.* 2019;286:309–15.
- [122] Kalli AC, Reithmeier RAF. Interaction of the human erythrocyte Band 3 anion exchanger 1 (AE1, SLC4A1) with lipids and glycophorin A: molecular organization of the Wright (Wr) blood group antigen. *PLoS Comput Biol* 2018;14(7):e1006284.
- [123] Chowdhury SC, Wise EA, Ganesh R, Gillespie JW. Effects of surface crack on the mechanical properties of Silica: a molecular dynamics simulation study. *Eng Fract Mech* 2019;207:99–108.
- [124] Patra MC, Kwon H-K, Batool M, Choi S. Computational insight into the structural organization of full-length toll-like receptor 4 dimer in a model phospholipid bilayer. *Front Immunol* 2018;9.
- [125] Kadirvel P, Anishetty S. Potential role of salt-bridges in the hinge-like movement of apicomplexa specific β -hairpin of Plasmodium and Toxoplasma profilins: A molecular dynamics simulation study. *J Cell Biochem* 2018;119(4):3683–96.
- [126] Tiwari G, Verma CS. Toward understanding the molecular recognition of albumin by p53-activating stapled peptide ATSP-7041. *J Phys Chem B* 2017;121(4):657–70.
- [127] Liu Y et al. Effect of surfactants on the interaction of phenol with laccase: MOLECULAR docking and molecular dynamics simulation studies. *J Hazard Mater* 2018;357:10–8.
- [128] Boon PLS et al. Partial intrinsic disorder governs the dengue capsid protein conformational ensemble. *ACS Chem Biol* 2018;13(6):1621–30.
- [129] Chen M, Zeng G, Xu P, Zhang Y, Jiang D, Zhou S. Understanding enzymatic degradation of single-walled carbon nanotubes triggered by functionalization using molecular dynamics simulation. *Environ Sci Nano* 2017;4(3):720–7.
- [130] Yeow J et al. The architecture of the OmpC–MlaA complex sheds light on the maintenance of outer membrane lipid asymmetry in *Escherichia coli*. *J Biol Chem* 2018;293(29):11325–40.
- [131] Hughes GW et al. Evidence for phospholipid export from the bacterial inner membrane by the Mla ABC transport system. *Nat Microbiol* 2019;4(10):1692–705.
- [132] Yeow J, Chong Z-S, Marzinek J, Bond P. Molecular basis for the maintenance of lipid asymmetry in the outer membrane of *Escherichia coli*. *BioRxiv* 2017. 192500.
- [133] Kraha A, Zachariae U. Insights into the ion-coupling mechanism in the MATE transporter NorM-VC. *Phys Biol* 2017;14(4):045009.
- [134] Revanasiddappa PD, Sankar R, Senapati S. Role of the bound phospholipids in the structural stability of cholesterol ester transfer protein. *J Phys Chem B* 2018;122(15):4239–48.
- [135] Farmer R, Thomas CM, Winn PJ. Structure, function and dynamics in acyl carrier proteins. *PLoS ONE* 2019;14(7):e0219435.
- [136] Bös F, Pleiss J. Multiple molecular dynamics simulations of TEM β -lactamase: dynamics and water binding of the Ω -loop. *Biophys J* 2009;97(9):2550–8.
- [137] Darby JF et al. Water networks can determine the affinity of ligand binding to proteins. *J Am Chem Soc* 2019;141(40):15818–26.
- [138] Syrén P-O, Hammer SC, Claasen B, Hauer B. Entropy is key to the formation of pentacyclic terpenoids by enzyme-catalyzed polycyclization. *Angew Chem Int Ed* 2014;53(19):4845–9.
- [139] Chen D et al. Effective lead optimization targeting the displacement of bridging receptor–ligand water molecules. *Phys Chem Chem Phys* 2018;20(37):24399–407.
- [140] Gerogiokas G, Calabro G, Henchman RH, Southey MWY, Law RJ, Michel J. Prediction of small molecule hydration thermodynamics with grid cell theory. *J Chem Theory Comput* 2014;10(1):35–48.
- [141] Huang J et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat Methods* 2017;14(1):71–3.
- [142] Beauchamp KA, Lin Y-S, Das R, Pande VS. Are protein force fields getting better? A systematic benchmark on 524 diverse NMR measurements. *J Chem Theory Comput* 2012;8(4):1409–14.

Structural bioinformatics

AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective

Tomasz Magdziarz, Karolina Mitusińska, Maria Bzówka, Agata Raczyńska, Agnieszka Stańczak, Michał Banas, Weronika Bagrowska and Artur Góra*

Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland

*To whom correspondence should be addressed.

Associate Editor: Arne Elofsson

Received on October 14, 2019; revised on December 11, 2019; editorial decision on December 16, 2019; accepted on December 17, 2019

Abstract

Motivation: Tunnels, pores, channels, pockets and cavities contribute to proteins architecture and performance. However, analysis and characteristics of transportation pathways and internal binding cavities are performed separately. We aimed to provide universal tool for analysis of proteins integral interior with access to detailed information on the ligands transportation phenomena and binding preferences.

Results: AQUA-DUCT version 1.0 is a comprehensive method for macromolecules analysis from the intramolecular voids perspective using small ligands as molecular probes. This version gives insight into several properties of macromolecules and facilitates protein engineering and drug design by the combination of the tracking and local mapping approach to small ligands.

Availability and implementation: <http://www.aqueduct.pl>.

Contact: info@aqueduct.pl

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

One of the most extensively used methods for the *in silico* study of macromolecules is molecular dynamics (MD) simulation. MD simulations have increased our knowledge of the conformational changes of proteins' regulatory elements such as gates (Góra *et al.*, 2013) or loops (Kreß *et al.*, 2018). They improved our understanding of the role of water in protein folding and stability, in shaping enzyme activity and selectivity, or in drug design (Mondal *et al.*, 2017; Spyraakis *et al.*, 2017). Finally MD simulations enabled analysis of intramolecular voids, described as cavities (Stank *et al.*, 2016) and tunnels (Kingsley and Lill, 2015; Marques *et al.*, 2016), contributing to the macromolecules' stability, functionality, activity and selectivity (Kokkonen *et al.*, 2019). More than 64% of enzymes are equipped with active sites buried inside the protein core (Pravda *et al.*, 2014), and investigation of the ligands' entry pathways is considered as essential for future improvements in *de novo* designed enzymes (Huang *et al.*, 2016). However, the description of protein interior dynamics is not a trivial problem, since the commonly used sphere approximation fails to give an accurate description of asymmetric volumes and neglects the physicochemical properties of the interior—factors essential for the transportation of reagents (Kaushik *et al.*, 2018).

2 Materials and methods

AQUA-DUCT 1.0 is an extension of the approach focused on molecules tracking (Magdziarz *et al.*, 2017). It goes beyond identification of the functionally relevant tunnels towards identification of structurally important residues and/or regions of macromolecules, approximation of free energy profiles of transportation pathways and an analysis of the evolution of the voids' and hot-spots dynamics (Fig. 1 and Supplementary Fig. S1). It reverses the standard approach of describing the evolution of macromolecules' dynamics through their atoms' movement analysis and enables investigation of macromolecules from the perspective of 'intramolecular voids'. To achieve this goal, we sample macromolecules' dynamics employing small entities in simulations (most frequent water molecules, but also other co-solvent, ions or other ligands). They are used as specific 'chemical probes', and their trajectories (Supplementary Figs S2 and S3) and occupancies (Supplementary Fig. S4) are analyzed to discriminate between functionally relevant compartments and to overcome the limitations of geometrically based approaches.

2.1 Small molecules tracking analysis

AQUA-DUCT 1.0 allows not only to detect, describe and compare tunnels' relevance and performance based on the number of

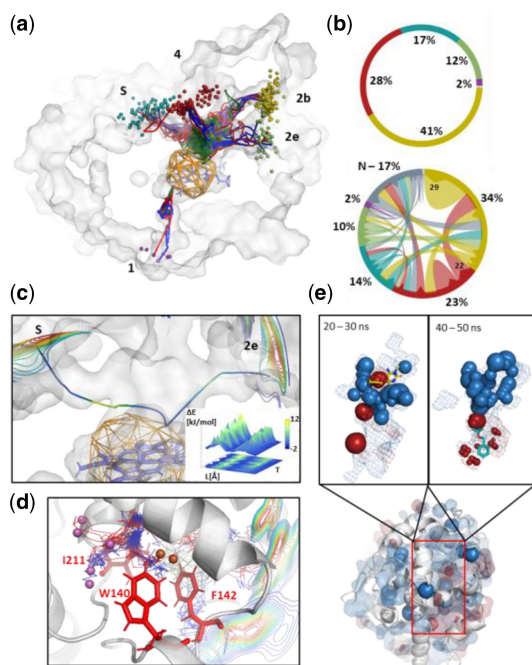


Fig. 1. An example of AQUA-DUCT analysis. (a) Paths (lines) and entry/exit locations (small balls) of water molecules passing via cytochrome P450 3A4-binding cavity during 50-ns MD simulation. (b) Statistical data of tunnels entry utilization (upper) and flow between tunnel entries (lower part). Colors reflect ones used in (a). N indicates trajectories which start or end in protein interior. (c) Energy profile of water transportation between '2e' and 's' tunnel entries (shown as isolines) in cytochrome P450 3A4. The smoothed path colored according to energy scale, the energy profile calculated in 10-ns time-window. (d) Rare events analysis—detected leakage of water molecules in LinB haloalkane dehalogenase via pathway used for *de novo* tunnel design. Main tunnels entries are shown as isolines, leaking molecules as small balls. Modified residues indicated by red sticks. (e) Hot-spots of water (blue spheres) and DMSO (red spheres) identified by distribution analysis in human epoxide hydrolase in 50-ns simulation (middle panel). Inner pockets shown in surface representations. Overlap of detected hot-spots and inhibitors during analysis of different time frames of simulations are shown on bottom (3-[4-(benzyloxy)phenyl]propan-1-ol) and top (6-amino-1-methyl-5-(piperidin-1-yl)pyrimidine-2,4(1H,3H)-dione) panel. (Color version of this figure is available at *Bioinformatics* online.)

molecules transported *via* a particular pathway (Fig. 1a and b), but also provides an approximation of transportation free energy profiles between pre-selected tunnels' entries (Fig. 1c). The analysis of solvent molecules' pathways allows for the identification of rare events which might correspond either to poorly sampled states, like *aqueduct* tunnel (W) in cytochrome P450 3A4 (Supplementary Fig. S5) or may suggest the localization of tunnels which can be designed *de novo* (Fig. 1d and Supplementary Fig. S6). Full statistical and quantitative analysis (Supplementary Fig. S7) is complemented by the visualization of raw and smoothed paths geometries (Supplementary Fig. S8), and the shape of ligands entry/exit areas (Supplementary Fig. S9).

2.2 Local-distribution analysis

The paths of molecules entering the protein interior can be structured and divided into distinct compartments corresponding to undisturbed passages and trapped molecules (Supplementary Fig. S2). The analysis of solvent trajectories can provide information about functionally relevant residues responsible for ligand trapping, which can vary depending on the tracked ligand (Supplementary Fig. S10). To simplify the identification of such residues, we calculate the local solvent distribution, which facilitates the detection of hot-spots, defined as compact volumes with high solvent occupancy (Supplementary Fig. S4). This approach can be used for the fast

identification of functionally important residues (e.g. gates) or molecules (e.g. catalytic water molecules), the description of hydrophilic/hydrophobic regions in the protein core (Supplementary Figs S11 and S12) and also for drug design (Fig. 1e).

2.3 Modes

The AQUA-DUCT 1.0 provides four distinct modes of analysis (Supplementary Fig. S13). The *standard mode* is used for the routine analysis of a single MD simulation. The *sandwich mode* enables the parallel analysis of multiple runs of individual simulations with different topologies (approximation of a macroscopic picture of the analyzed molecule). The *time-window mode* allows the analysis of long trajectories in pre-defined time windows and thus facilitates the identification of equivalent or alternative states (Supplementary Fig. S12). Different and rare conformations can be correctly described with the *consolidator mode* (Supplementary Fig. S14). Pre-selected frames of the simulation can be merged together to provide a pre-treated trajectory with enhanced sampling of a rare event [e.g. substrate entry (Supplementary Fig. S14) or the rare opening of an alternative pathway Fig. 1d] and efficiently analyzed. The obtained data can be used for the alternative design of enhanced catalysts or new inhibitors, as well as used as high-quality preliminary data comparable with Markov model results.

3 Conclusions

Our method is able to analyze dynamic changes in the spatial distribution of the physicochemical properties with user-defined time-scales and resolution, and also with easy and fast insight into the geometry of macromolecules' interiors and the approximation of transport energy barriers *via* particular pathways. The application of 'ligands-tracking' and 'local-distribution' approaches together with the introduction of a 'chemical probe' overcomes most of the limitations of currently available tools. The user receives direct access to information about the active site, potential hot-spots, functional residues, the network of internal transportation pathways and functional voids and cavities and benefits from modules that can facilitate the understanding of macromolecules, protein engineering and drug design.

Acknowledgement

This research was supported in part by PL-Grid Infrastructure.

Funding

This work was supported by the National Science Centre, Poland [DEC-2013/10/E/NZ1/00649 and DEC-2015/18/M/NZ1/00427].

Conflict of Interest: none declared.

References

- Gora, A. et al. (2013) Gates of enzymes. *Chem. Rev.*, **113**, 5871–5923.
- Huang, P.-S. et al. (2016) The coming of age of *de novo* protein design. *Nature*, **537**, 320–327.
- Kaushik, S. et al. (2018) Impact of the access tunnel engineering on catalysis is strictly ligand-specific. *FEBS J.*, **285**, 1456–1476.
- Kingsley, L.J. and Lill, M.A. (2015) Substrate tunnels in enzymes: structure-function relationships and computational methodology. *Proteins*, **83**, 599–611.
- Kokkonen, P. et al. (2019) Engineering enzyme access tunnels. *Biotechnol. Adv.*, **37**, 107386.
- Kreß, N. et al. (2018) Unlocked potential of dynamic elements in protein structures: channels and loops. *Curr. Opin. Chem. Biol.*, **47**, 109–116.
- Magdziarz, T. et al. (2017) AQUA-DUCT: a ligands tracking tool. *Bioinformatics*, **33**, 2045–2046.

- Marques, S. *et al.* (2016) Role of tunnels and gates in enzymatic catalysis. In: Svendsen, A. (ed.) *Understanding Enzymes: Function, Design, Engineering, and Analysis*. Pan Stanford Publishing, Singapore, pp. 421–463.
- Mondal, S. *et al.* (2017) Decomposition of total solvation energy into core, side-chains and water contributions: role of cross correlations and protein conformational fluctuations in dynamics of hydration layer. *Chem. Phys. Lett.*, **683**, 29–37.
- Pravda, L. *et al.* (2014) Anatomy of enzyme channels. *BMC Bioinformatics*, **15**, 379.
- Spyrakakis, F. *et al.* (2017) The roles of water in the protein matrix: a largely untapped resource for drug discovery. *J. Med. Chem.*, **60**, 6781–6828.
- Stank, A. *et al.* (2016) Protein binding pocket dynamics. *Acc. Chem. Res.*, **49**, 809–815.



Article

Structural and Evolutionary Analysis Indicate That the SARS-CoV-2 Mpro Is a Challenging Target for Small-Molecule Inhibitor Design

Maria Bzówka ^{1,†} , Karolina Mitusińska ^{1,†} , Agata Raczyńska ¹, Aleksandra Samol ¹,
Jack A. Tuszyński ^{2,3} and Artur Góra ^{1,*}

¹ Tunneling Group, Biotechnology Centre, ul. Krzywoustego 8, Silesian University of Technology, 44-100 Gliwice, Poland; m.bzowka@tunnelinggroup.pl (M.B.); k.mitusinska@tunnelinggroup.pl (K.M.); a.raczynska@tunnelinggroup.pl (A.R.); a.samol@tunnelinggroup.pl (A.S.)

² Department of Physics, University of Alberta, Edmonton, AB T6G 2E1, Canada; jact@ualberta.ca

³ DIMEAS, Politecnico di Torino, Corso Duca degli Abruzzi, 24, 10129 Turin, Italy

* Correspondence: a.gora@tunnelinggroup.pl; Tel.: +48-32-237-16-59

† These authors contributed equally to this work.

Received: 28 March 2020; Accepted: 26 April 2020; Published: 28 April 2020



Abstract: The novel coronavirus whose outbreak took place in December 2019 continues to spread at a rapid rate worldwide. In the absence of an effective vaccine, inhibitor repurposing or de novo drug design may offer a longer-term strategy to combat this and future infections due to similar viruses. Here, we report on detailed classical and mixed-solvent molecular dynamics simulations of the main protease (Mpro) enriched by evolutionary and stability analysis of the protein. The results were compared with those for a highly similar severe acute respiratory syndrome (SARS) Mpro protein. In spite of a high level of sequence similarity, the active sites in both proteins showed major differences in both shape and size, indicating that repurposing SARS drugs for COVID-19 may be futile. Furthermore, analysis of the binding site's conformational changes during the simulation time indicated its flexibility and plasticity, which dashes hopes for rapid and reliable drug design. Conversely, structural stability of the protein with respect to flexible loop mutations indicated that the virus' mutability will pose a further challenge to the rational design of small-molecule inhibitors. However, few residues contribute significantly to the protein stability and thus can be considered as key anchoring residues for Mpro inhibitor design.

Keywords: coronavirus; SARS-CoV; SARS-CoV-2; COVID-19; molecular dynamics simulations; ligand tracking approach; drug design; small-molecule inhibitors; evolutionary analysis

1. Introduction

In early December 2019, the first atypical pneumonia outbreak associated with the novel coronavirus of zoonotic origin (SARS-CoV-2) appeared in Wuhan City, Hubei Province, China [1,2]. In general, coronaviruses (CoVs) are classified into four major genera: Alphacoronavirus, Betacoronavirus (which primarily infect mammals), Gammacoronavirus, and Deltacoronavirus (which primarily infect birds) [3–5]. In humans, coronaviruses usually cause mild to moderate upper-respiratory tract illnesses, such as the common cold, however, the rarer forms of CoVs can be lethal. By the end of 2019, six kinds of human CoV have been identified: HCoV-NL63; HCoV-229E, belonging to Alphacoronavirus genera; HCoV-OC43; HCoV-HKU1; severe acute respiratory syndrome (SARS-CoV); and Middle East respiratory syndrome (MERS-CoV), belonging to Betacoronavirus genera [4]. Of the aforementioned CoVs, the latter two are the most dangerous and have been associated with the outbreak of two epidemics at the beginning of the 21st century [6]. In January

2020, SARS-CoV-2 was isolated and announced as a new, seventh type of human coronavirus. It was classified as a Betacoronavirus [2]. On the basis of the phylogenetic analysis of the genomic data of SARS-CoV-2, Zhang et al. indicated that SARS-CoV-2 is most closely related to two SARS-CoV sequences isolated from bats in 2015 and 2017. This suggests that the bat CoV and SARS-CoV-2 share a common ancestor, and the new virus can be considered as a SARS-like virus [7].

The genome of coronaviruses typically contains a positive-sense, single-stranded RNA but it differs in size ranging between ≈ 26 and ≈ 32 kb. It also includes a variable number of open reading frames (ORFs), from 6 to 11. The first ORF is the largest, encoding nearly 70% of the entire genome and 16 non-structural proteins (nsps) [3,8]. Of the nsps, the main protease (Mpro, also known as a chymotrypsin-like cysteine protease, 3CLpro), encoded by nsp5, has been found to play a fundamental role in viral gene expression and replication, and thus it is an attractive target for anti-CoV drug design [9]. The remaining ORFs encode accessory and structural proteins, including spike surface glycoprotein (S), small envelope protein (E), matrix protein (M), and nucleocapsid protein (N).

On the basis of the three sequenced genomes of SARS-CoV-2 (Wuhan/IVDC-HB-01/2019, Wuhan/IVDC-HB-04/2019, and Wuhan/IVDC-HB-05/2019, provided by the National Institute for Viral Disease Control and Prevention, CDC, China), Wu et al. performed a detailed genome annotation. The results were further compared to related coronaviruses—1008 human SARS-CoV, 338 bat SARS-like CoV, and 3131 human MERS-CoV, indicating that the three strains of SARS-CoV-2 have almost identical genomes with 14 ORFs, encoding 27 proteins including 15 non-structural proteins (nsp1–10 and nsp12–16), 4 structural proteins (S, E, M, N), and 8 accessory proteins (3a, 3b, p6, 7a, 7b, 8b, 9b, and orf14). The only identified difference in the genome consisting of ≈ 29.8 kb nucleotides consisted of five nucleotides. The genome annotation revealed that SARS-CoV-2 is fairly similar to SARS-CoV at the amino acid level, however, there are some differences in the occurrence of accessory proteins, such as the fact that the 8a accessory protein, present in SARS-CoV, is absent in SARS-CoV-2, as well as the fact that the lengths of 8b and 3b proteins do not match. The phylogenetic analysis of SARS-CoV-2 showed it to be most closely related to SARS-like bat viruses, but no strain of SARS-like bat virus was found to cover all equivalent proteins of SARS-CoV-2 [10].

As previously mentioned, the main protease is one of the key enzymes in the viral life cycle. Together with other non-structural proteins (papain-like protease, helicase, RNA-dependent RNA polymerase) and the spike glycoprotein structural protein, it is essential for interactions between the virus and host cell receptor during viral entry [11]. Initial analyses of genomic sequences of the four nsps mentioned above indicate that those enzymes are highly conserved, sharing more than 90% sequence similarity with the corresponding SARS-CoV enzymes [12].

The first released crystal structure of the Mpro of SARS-CoV-2 (PDB ID: 6lu7) was obtained by Prof. Yang's group from ShanghaiTech by co-crystallisation with a peptide-like inhibitor N-[(5 methylisoxazol-3-yl)carbonyl]alanyl-L-valyl-N-1-((1R,2Z)-4-(benzyloxy)-4-oxo-1-[(3R)-2-oxopyrrolidin-3-yl]methyl]but-2-enyl)-L-leucinamide (N3 or PRD_002214) [13]. The same inhibitor was co-crystallised with other human coronaviruses, such as HCoV-NL63 (PDB ID: 5gwy), HCoV-KU1 (PDB ID: 3d23), or SARS-CoV (PDB ID: 2amq). This enzyme naturally forms a dimer, each of whose monomer consists of the N-terminal catalytic region and a C-terminal region [14]. Although 12 residues differ between both CoVs, only one, namely, S46 in SARS-CoV-2 (A46 in SARS-CoV), is located in the proximity of the entrance to the active site. However, such a small structural change would typically be not expected to substantially affect the binding of small molecules [12]. Such an assumption would routinely involve the generation of a library of derivatives and analogues on the basis of the scaffold of a drug that inhibits the corresponding protein in the SARS-CoV case. As shown in the present paper, regrettably, this strategy is not likely to succeed with SARS-CoV-2 for Mpro as a molecular target.

In this study, we investigated how only 12 different residues, located mostly on the protein's surface, may affect the behaviour of the active site pocket of the SARS-CoV-2 Mpro protein. To this end, we performed classical molecular dynamics simulations (cMD) of both SARS and SARS-CoV-2 Mpros, as well as mixed-solvents molecular dynamics simulations (MixMD) combined with small molecules'

tracking approach to analyse the conformational changes in the binding site. The experiment setup and methodology workflow is presented in Figure S1. Despite the structural differences in the active sites of both Mpro proteins, major issues involving plasticity and flexibility of the binding site could result in significant difficulties in inhibitor design for this molecular target. Indeed, an *in silico* attempt has already been made involving a massive virtual screening for Mpro inhibitors of SARS-CoV-2 using Deep Docking [15]. Other recent attempts focused on virtual screening for putative inhibitors of the same main protease of SARS-CoV-2 on the basis of the clinically approved drugs [16–21], and also on the basis of the compounds from different databases or libraries [22–24]. However, none of such attempts is likely to lead to clinical advances in the fight against SARS-CoV-2 for reasons we elaborate below.

2. Results

2.1. Crystal Structure Comparison, and Location of the Replaced Amino Acids Distal to the Active Site

The first SARS-CoV-2 main protease's crystallographic structure was made publicly available through the Protein Data Bank (PDB) [25] as a complex with an N3 inhibitor (PDB ID: 6lu7) [13]. Next, the structure without the inhibitor was also made available (PDB ID: 6y2e) [26]. We refer to these structures as SARS-CoV-2 Mpro^{N3} and SARS-CoV-2 Mpro, respectively. We also used two structures of the SARS-CoV main protease: one, referred to as SARS-CoV Mpro^{N3} (PDB ID: 2amq), co-crystallised with the same N3 inhibitor, and the other without an inhibitor (PDB ID: 1q2w), which we refer to as SARS-CoV Mpro. The SARS-CoV-2 Mpro and SARS-CoV Mpro structures differ by only 12 amino acids located mostly on the proteins' surface (Figure 1A, Table S1). Both enzymes share the same structural composition; they comprise three domains: domains I (residues 1–101) and II (residues 102–184) consist of an antiparallel β -barrel, and the α -helical domain III (residues 201–301) is required for the enzymatic activity [27]. Both enzymes resemble the structure of cysteine proteases, although their active site is lacking the third catalytic residue [28]; their active site comprises a catalytic dyad, namely, H41 and C145, and a particularly stable water molecule forms at least three hydrogen bond interactions with surrounding residues, including the catalytic histidine, which corresponds to the position of a third catalytic member (Figure 1B). It should be also noted that one of the differing amino acids in SARS-CoV-2 Mpro, namely, S46, is located on a C44-P52 loop, which is flanking the active site cavity.

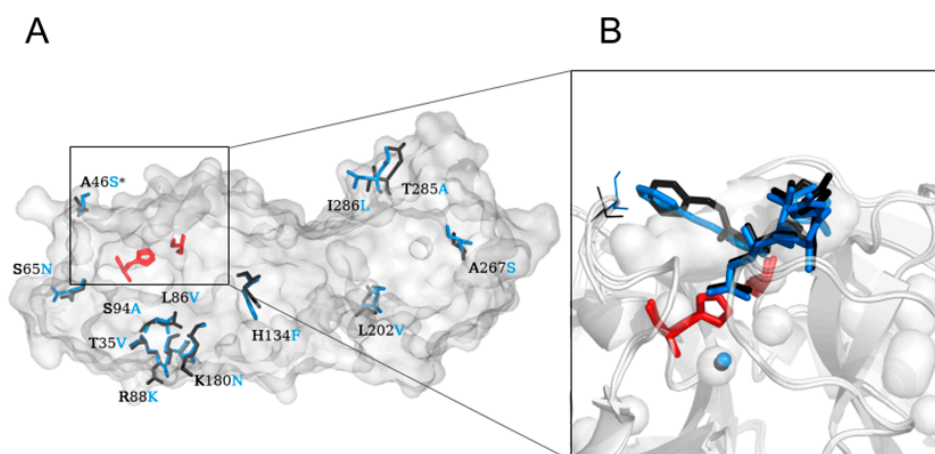


Figure 1. The differences between the severe acute respiratory syndrome coronavirus main protease (SARS-CoV Mpro) and SARS-CoV-2 Mpro structures. (A) The overall structure of both SARS-CoV and SARS-CoV-2 Mpros with differing amino acids are marked as black (SARS-CoV Mpro) and blue (SARS-CoV-2 Mpro). (B) Close-up of the active site cavity and bound N3 inhibitor into SARS-CoV (black sticks) and SARS-CoV-2 (blue sticks) Mpros. The catalytic water molecule that resembles the position of the third member of the catalytic triad adopted from the cysteine proteases is shown for both SARS-CoV (black sphere) and SARS-CoV-2 (blue sphere) Mpros. The active site residues are shown as

red sticks and the proteins' structures are shown in surface representation. The differing residues in position 46 located near the entrance to the active site are marked with an asterisk (*) on the (A) and as blue and black lines on the (B) panel.

2.2. Plasticity of the Binding Cavities

A total of 2 μ s classical molecular dynamics (cMD) simulations of both SARS-CoV-2 and SARS-CoV Mpros with different starting points were run to examine the plasticity of their binding cavities. As different starting points we used (i) SARS-CoV-2 Mpro apo structure; (ii) SARS-CoV-2 Mpro with an N3 ligand, which was removed before starting the simulation; (iii) SARS-CoV Mpro apo structure; and (iv) SARS-CoV Mpro with the same N3 ligand, which was also removed before starting the simulation. A total of 10 replicas of 50 ns classical molecular dynamics simulations were performed for each protein. To improve conformation sampling, the starting geometry for each system was kept but the initial vectors were randomly assigned. A combination of the cMD approach with water molecules used as molecular probes is assumed to provide a highly detailed picture of the protein's interior dynamics [29]. The small molecules tracking approach was used to determine the accessibility of the active site pocket in both SARS-CoV and SARS-CoV-2 Mpros, and a local distribution approach was used to provide information about an overall distribution of solvent in the proteins' interior. To properly examine the flexibility of both active site cavities, we used the time-window mode implemented in AQUA-DUCT software [30] to analyse the water molecules' flow through the cavity in a 10 ns time step and combined that with the outer pocket calculations to examine the plasticity and maximal accessible volume (MAV) of the binding cavity.

Surprisingly, despite their high similarity, the binding cavities of SARS-CoV and SARS-CoV-2 Mpros showed significantly different MAV (Wald test, $z = 2597$, $p < 0.05$). Both proteins reduced their MAV upon inhibitor binding by approximately 20%, but the maximal volume of SARS-CoV was over 50% larger than those of SARS-CoV-2 (Figure 2 and Figure S2).

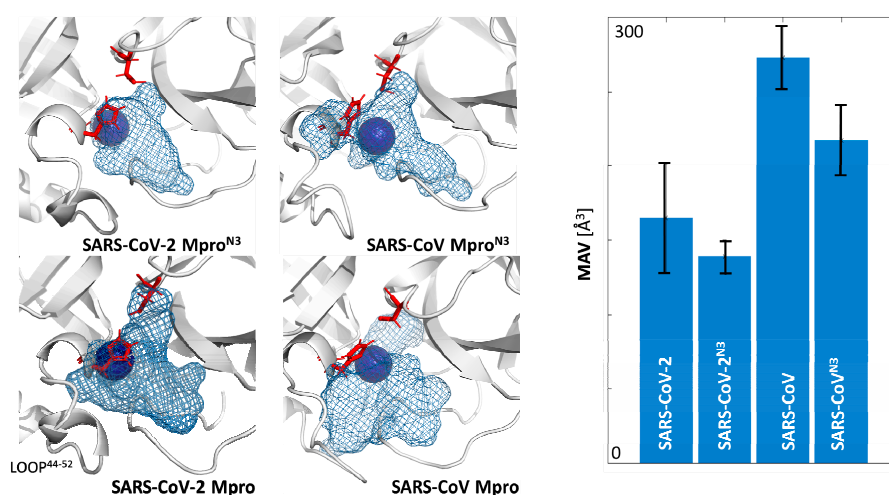


Figure 2. The differences between the maximal accessible volume of the binding cavities calculated during molecular dynamics (MD) simulations of both apo structures of Mpros (SARS-CoV and SARS-CoV-2) and structures with co-crystallised N3 inhibitor (SARS-CoV^{N3} and SARS-CoV-2^{N3}) used as different starting points for 10 replicas of 50 ns per structure. The position of the blue sphere (hot-spot with highest density) in each structure reflects the position of the catalytic water molecule.

2.3. Flexibility of the Active Site Entrance

To further examine the plasticity and flexibility of the main proteases binding cavities, we focused on the movements of loops surrounding their entrances and regulating the active sites' accessibility.

We found that one of the analysed loops of the SARS-CoV Mpro, namely, C44-P52 loop, was more flexible than the corresponding loops of SARS-CoV-2 Mpro structure, whereas the adjacent loops were mildly flexible (Figure 3). This could be indirectly assumed from the absence of the C44-P52 loop in the crystallographic structure of SARS-CoV Mpro structure. On the other hand, such flexibility could suggest that the presence of an inhibitor might stabilise the loops surrounding the active site. The analysis of B-factors of all deposited Mpro crystal structures fully confirmed these statements (Figure S3). It is worth adding that this loop was carrying the unique SARS-CoV-2 Mpro residue S46.

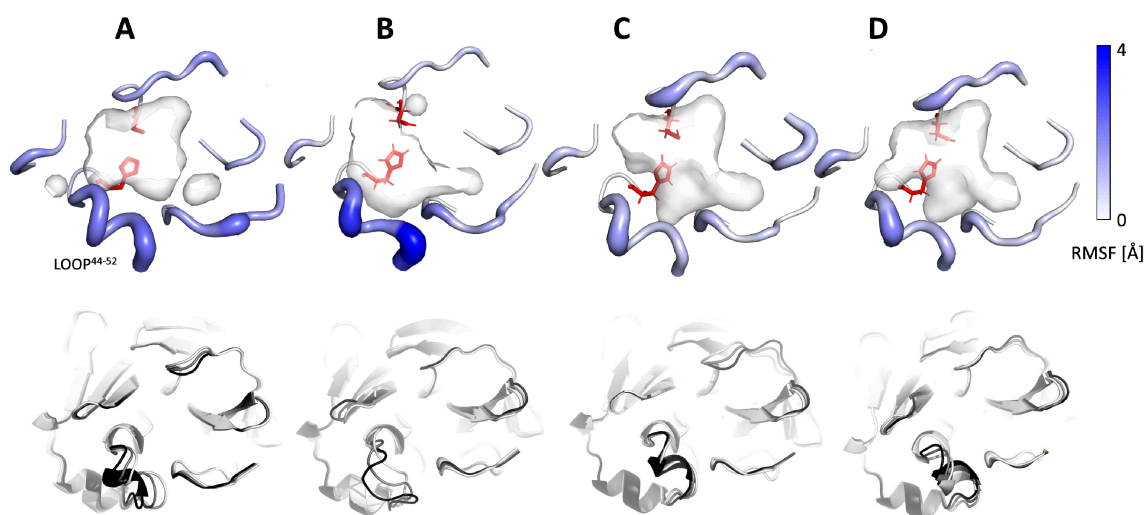


Figure 3. Flexibility of loops surrounding the entrance to the binding cavity of (A) SARS-CoV-2 Mpro, (B) SARS-CoV Mpro, (C) SARS-CoV Mpro^{N3}, and (D) SARS-CoV Mpro^{N3}. For the picture clarity, only residues creating loops were shown. Upper row: RMSF data. The active site residues are shown as red sticks, and the A46S replacement between SARS-CoV and SARS-CoV-2 main proteases is shown as light blue sticks. The width and colour of the shown residues reflect the level of loop flexibility. The wider and darker residues are more flexible. Lower row: the results of normal mode analysis as a superposition of active site surroundings; structures are coloured white—initial conformation, black—final conformation, gray—transient conformation.

2.4. Cosolvent Hot-Spots Analysis

The mixed-solvent MD simulations were run with six cosolvents: acetonitrile (ACN), benzene (BNZ), dimethylsulfoxide (DMSO), methanol (MEO), phenol (PHN), and urea (URE). Cosolvents were used as specific molecular probes, representing different chemical properties and functional groups that would complement the different regions of the binding site and the protein itself. Using small molecules tracking approach, we analysed the flow through the Mpros structures and identified the regions in which those molecules were being trapped and/or caged, located within the protein itself (global hot-spots; Figures S4 and S5) and inside the binding cavity (local hot-spots; Figure 4 and Figure S6). The size and location of both types of hot-spots differed and provided complementary information. The global hot-spots identified potential binding/interacting sites in the whole protein structure and additionally provided information about regions attracting particular types of molecules, whereas local hot-spots described the actual available binding space of a specific cavity.

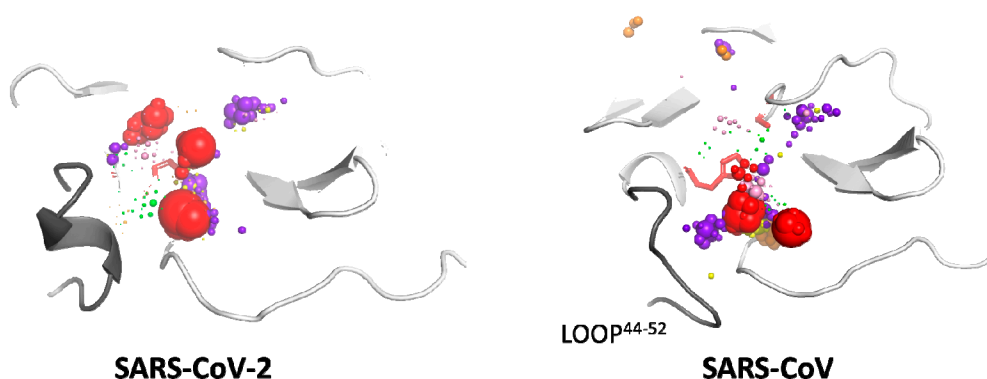


Figure 4. Localisation of the local hot-spots identified in the binding site cavities in SARS-CoV-2 and SARS-CoV main proteases. Hot-spots of individual cosolvents are represented by spheres, and their size reflects the hot-spot density. The colour coding is as follows: purple—urea, green—dimethylsulfoxide, yellow—methanol, orange—acetonitrile, pink—phenol, red—benzene. The active site residues are shown as red sticks, and the proteins' structures are shown in cartoon representation; loop 44–52 is grey. The proteins' structures come from the MD simulation snapshots (first frame of the production stage).

The general distribution of the global hot-spots from particular cosolvents was quite similar and verified specific interactions with the particular regions of the analysed proteins. A notable number of hot-spots were located around the amino acids that varied between the SARS-CoV-2 and SARS-CoV Mpros (Figures S4 and S5). The largest number and the densest hot-spots are located within the binding cavity and the region essential for Mpro dimerisation [31], between the II and III domains. The binding cavity is particularly occupied by urea, benzene, and phenol hot-spots, which is especially interesting because these solvents exhibit different chemical properties. In addition, the analysis performed by PARS server [32] detected three cavities located between domain II and III that could contribute significantly to the protein flexibility; however, none of them was found as conserved and therefore were not considered as regulatory sites.

A close inspection of the binding site cavity provided further details of cosolvent distribution. The benzene hot-spots for the SARS-CoV-2 Mpro structure are localised deep inside the active site cavity, whereas SARS-CoV Mpro features mostly benzene hot-spots at the cavity entrance (Figure 5). This is interesting because, in the absence of cosolvent molecules, the water accessible volume for SARS-CoV-2 Mpro was 50% smaller than in the case of SARS-CoV Mpro, underlining huge plasticity of the binding cavity and suggesting large conformational changes induced by interaction with a potential ligand. It is also interesting that both global and local hot-spots of the SARS-CoV Mpro structure are located in the proximity of the C44-P52 loop, which potentially regulates the access to the active site.

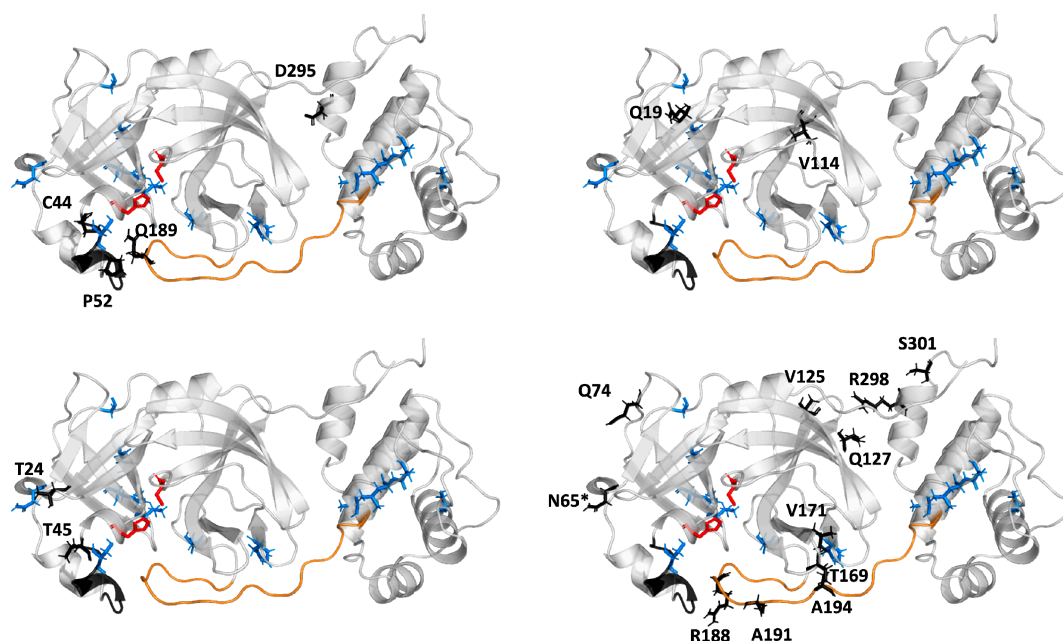


Figure 5. Localisation of the evolutionary-correlated residues of Mpro (black sticks). The correlated mutation analysis (CMA) analysis provided four groups of evolutionary-correlated residues. The SARS-CoV-2 Mpro structure is presented as a cartoon, the active site residues are shown as red sticks, the unique residues of the SARS-CoV-2 Mpro as blue sticks, and the asterisk (*) indicates the residue belonging to the evolutionary-correlated residues, unique for SARS-CoV Mpro. The loop C44-P52 is coloured black and the F185-T201 loop is orange. Please note that within one of the correlated groups (upper left), the residues from C44-P52 loop are correlated with Q189 from the linker loop and with residue from III domain.

2.5. Potential Mutability of SARS-CoV-2

In general, all the above-mentioned findings indicate potential difficulties in the identification of specific inhibitors toward Mpro proteins. First, the binding site itself is characterised by large plasticity (over 20% change of the MAV upon ligand binding) and probably even distant to active site mutations modify Mpro binding properties. Secondly, the C44-P52 loop regulates access to the active site and can contribute to the discrimination of potential inhibitors. Therefore, additional mutations in the above-mentioned regions, which could appear during further SARS-CoV-2 evolution, can significantly change the affinity between Mpro and its ligands. To verify potential threat of further mutability of the Mpro protein, we performed (i) correlated mutation analysis (CMA) on multiple sequence alignment, (ii) the analysis of the contribution of already identified differences between the SARS-CoV and SARS-CoV-2 Mpros to protein stability, and (iii) prediction of further possible mutations caused by the most probable mutations, the substitution of single nucleotides in the mRNA sequence of Mpro.

Indeed, the analysis performed with Comulator software [33] showed that within viral Mpros, evolutionary-correlated residues are dispersed throughout the structure. This indirectly supports our previous findings that distant amino acid mutation can contribute substantially to the binding site plasticity. It is worth adding that among evolutionary-correlated residues, we identified also those that differ between SARS-CoV-2 and SARS-CoV Mpros, located on the C44-P52 loop (Figure 5) and the F185-T201 linker loop. The CMA analysis indicated that particular residues in both loops are evolutionary-correlated. The Q189 from the linker loop correlates with residues from the C44-P52 loop, whereas R188, A191, and A194 correlate with selected residues from all domains, but not with the C44-P52 loop. As was shown, the mutation of amino acids distant from active site residues, which are evolutionary correlated, is most likely to modify the active site accessibility [34].

In the interest of examining the energetical effect of the 12 amino acid replacement in the SARS-CoV-2 Mpro structure, we calculated their energetic contributions to the protein's stability using FoldX [35]. As expected, the differences in total energies of the SARS-CoV Mpro and variants with introduced mutation from SARS-CoV-2 Mpro residue did not represent a significant energy change (Table S1). The biggest energy reduction was found for mutation H134F (-0.85 kcal/mol) and mutations R88K, S94A, T285A, and I286L only slightly reduced the total energy (Table S1).

To investigate further possible mutations of SARS-CoV-2 Mpro, single nucleotide substitutions were introduced to the SARS-CoV-2 main protease gene. If a substitution of a single nucleotide caused translation to a different amino acid compared to the corresponding residue in the wild-type structure, an appropriate mutation was proposed with FoldX calculations. The most energetically favourable potential mutations were chosen on the basis of a -1.5 kcal/mol threshold (Figure 6A, Table S2). Most of the energetically favourable potential mutations include amino acids that are solvent-exposed on the protein's surface, according to NetSurfP [36] results. These results show that in general, exposed amino acids are more likely to mutate.

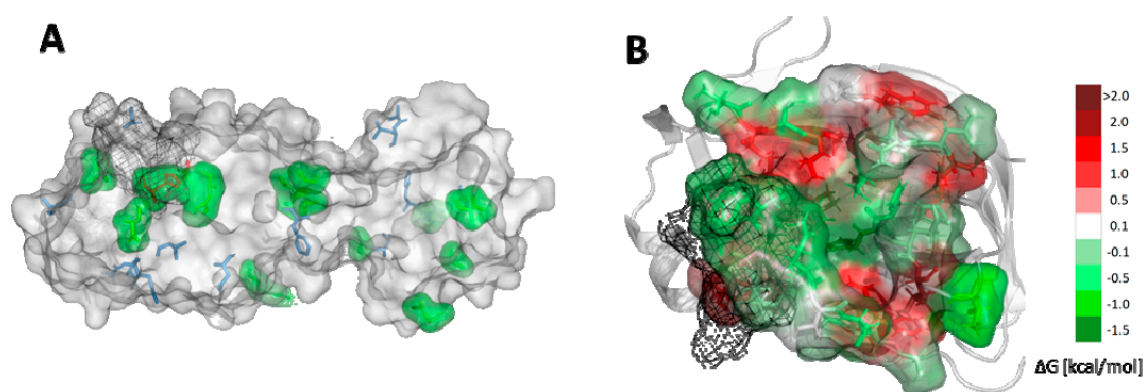


Figure 6. Potential mutability of SARS-CoV-2 Mpro. (A) Structure of SARS-CoV-2 Mpro with the most energetically favourable potential mutations of amino acids marked as green surface. Positions of amino acids that differ from the ones in SARS-CoV Mpro structure marked as blue sticks. Catalytic dyad marked as red. (B) The catalytic site of SARS-CoV-2 Mpro is shown as surface with the most energetically favourable potential mutations shown as green, neutral as white, and unfavourable as red. The C44-P52 loop is shown as black mesh.

Additionally, the potential mutability of the binding cavity was investigated. Residues belonging to the binding cavity were found within 7 \AA from the N3 inhibitor. Then, we calculated the differences in the Gibbs free energy of protein folding with respect to the wild-type protein (Table S3) and presented the results as a heat map. The most energetically favourable potential mutations are shown as green, neutral as white, and unfavourable as red (Figure 6B). Interestingly, residues forming the catalytic dyad, namely, H41 and C145, are also prone to mutate. However, probably the most important message comes from the analysis of the potential mutability of the C44-P52 loop. Mutation of four of them has a stabilising effect for the protein, and is near-neutral for the rest the effect. These results indicate that the future evolution of the Mpro protein can significantly reduce the potential use of this protein as a molecular target for coronavirus treatment due to a highly probable development of drug resistance of this virus through mutations.

3. Discussion

The analysis of water molecules' distribution and trajectories can be used for the analysis of proteins' structural features and biochemical properties. It also provides additional support to the drug design and investigation of protein interior [29,37,38]. As we have shown in previous research, tracking of water molecules in the binding cavity combined with the local distribution approach can identify catalytic water positions [39]. Indeed, despite differences in the size and dynamics of the

binding cavities of SARS-CoV and SARS-CoV-2 Mpros, the main identified water was always found in a position next to the H41 residue (Figure 2), and this location is assumed to indicate catalytic water of Mpro replacing the missing third catalytic site amino acid [28]. That was the first quality check of our methodology that approved our approach, and has initiated further investigations.

As reported in the previous research, the overall plasticity of Mpro is required for proper enzyme functioning [26,40]. In the case of SARS-CoV the truncation of the linker loop (F185-T201) gave rise to a significant reduction in protein activity and confirmed that the proper orientation of the linker allows the shift between dimeric and monomeric forms [41]. Dimerisation of the enzyme is necessary for its catalytic activity, and the proper conformation of the seven N-terminal residues (N-finger) is required [42]. In SARS-CoV-2 Mpro, the T285 is replaced by alanine, and the I286 by leucine. It has been shown that replacing S284, T285, and I286 by alanine residues in SARS-CoV Mpro leads to a 3.6-fold enhancement of the catalytic activity of the enzyme. This is accompanied by changes in the structural dynamics of the enzyme that transmit the effect of the mutation to the catalytic centre. Indeed, the T285A replacement observed in the SARS-CoV-2 Mpro allows the two domains III to approach each other a little closer [43].

The comparison of MD simulations of both main proteases initiated from different starting conformations (with and without N3 inhibitor) suggests that besides plasticity of the whole protein, there can be large differences between the accessibility to the binding cavity and/or the accommodation of the shape of the cavity in response to the inhibitor that can be bound. There are also differences in the outer pockets' maximal accessible volumes between the two structures of SARS-CoV main proteases; the apo SARS-CoV Mpro structure used as a starting point of MD simulations has shown the largest MAV of all the analysed systems. These results suggest that the SARS-CoV main proteases' binding cavity is highly flexible and changes both in volume and shape, significantly altering the ligand binding. This finding indicates a serious obstacle for a classical virtual screening approach and drug design in general. Numerous novel compounds that are considered as potential inhibitors of SARS-CoV have not reached the stage of clinical trials. The lack of success might be related to the above-mentioned plasticity of the binding cavity. Some of these compounds have been used for docking and virtual screening research aimed not only at SARS-CoV [44,45] but also at the novel SARS-CoV-2 [15–21]. Such an approach focuses mostly on the structural similarity between the binding pockets, but ignores the fact that the actual available binding space differs significantly. In general, a rational drug design can be a very successful tool in the identification of possible inhibitors in cases where the atomic resolution structure of the target protein is known. For a new target, when a highly homologous structure is available, a very logical strategy would be seeking chemically similar compounds or creating derivatives of this inhibitor, as well as finding those compounds that are predicted to have a higher affinity for the new target structure than the original one. This would be expected to work for SARS-CoV-2 proteins (such as Mpro) using SARS-CoV proteins as templates. However, our in-depth analysis indicates a very different situation taking place, with major shape and size differences emerging due to the binding site flexibility. Therefore, repurposing SARS drugs against COVID-19 may not be successful due to major shape and size differences, and despite docking methods, the enhanced sampling should be considered.

The continuous effort of Diamond Light Source group [46] performing massive XChem crystallographic fragment screen against Mpro has resulted in 22 non-covalent hits in the active site and 44 covalent hits in the active site (March 17th). Interestingly, two hits were identified on the dimer interface. The positions of the hits inside the active site overlap with the position of the maximal accessible volume calculated from MD simulations and supports our finding on large binding site flexibility (Figure S7).

The analysis of the water hot-spots shows the catalytic water hot-spot dominated water distribution inside the binding cavity. The remaining water hot-spots corresponded to a much lower water density level and were on the borders of the binding cavity, which suggests a rather hydrophobic or neutral interior of the binding cavity. The MixMD simulations performed with various cosolvents further

confirmed these observations. The largest number and the densest hot-spots were located within the binding cavity and the region essential for Mpro dimerisation [31], between the II and III domains. The deep insight into the local hot-spot distribution of the various cosolvents underlines the large differences in binding sites plasticity. The smaller binding cavity of the SARS-CoV-2 enlarged significantly in the presence of a highly hydrophobic cosolvent. The benzene hot-spots were detected deep inside the cavity, and also near the C44-P52 loop. In contrast, in the case of SARS-CoV, benzene hot-spots were located only in the vicinity of the C44-P52 loop. Such a conclusion may also imply that a sufficiently potent inhibitor of SARS-CoV and/or SARS-CoV-2 Mpros needs to be able to open its way to the active site before it can successfully bind to its cavity. These results support the regulatory role of the C44-P52 loop and again alert against unwarranted use of simplified approaches for drug repositioning or docking.

The difficulties in targeting the active site of the Mpros are also explained by evolutionary study and potential mutability analysis. As already pointed out, the C44-P52 loop is likely to regulate the access to the active site by enabling entrance of favourable small molecules and blocking the entry of unfavourable ones. The second important loop, F185-T201, which starts in the vicinity of the binding site and links I and II domains with the III domain contributes significantly to Mpro dimerisation [41].

The initial analysis of the effect of the 12 amino acid replacements in SARS-CoV Mpro on the SARS-CoV-2 Mpro structure stability was expected to provide neutral or stabilising contribution to protein folding. Indeed, all replacements were found to stabilise the protein's folding (e.g., H134F: -0.85 kcal/mol) or have an almost neutral character (e.g., R88K, S94A, T285A, I286L). The analysis of the potential risk of further Mpro structure evolution within the binding cavity suggests that mutations of residues that contribute to ligand binding or access to the active site are energetically favourable, and are likely to occur. Some of the residues that are prone to mutate would provide the inactive enzyme (e.g., the residues forming the catalytic dyad) and therefore could be considered as a blind alley in enzyme evolution, but others (e.g., amino acids from the C44-P52 loop, T45, S46, E47, L50) could significantly modify the inhibitors binding mode of Mpro. The locations of residues on the regulatory loop, which are prone to mutate puts in question the efforts to design inhibitors of the MPro active site as a viable long-term strategy. However, our results also indicate residues that are energetically unfavourable to mutate (e.g., P39, R40, P52, G143, G146, or L167), which could provide an anchor for successful drug design that can outlast coronavirus Mpro variability in future. Alternatively, we would suggest targeting the region between II and III domains, which contributes to the dimer formation.

4. Materials and Methods

4.1. Classical MD Simulations

The H++ server [47] was used to protonate the SARS-CoV-2 (PDB IDs: 6lu7, and 6y2e) and SARS-CoV main proteases' structures (PDB IDs: 2amq, and 1q2w) using standard parameters at pH 7.4. The missing 4-amino-acids-long loop of the 1q2w model was added using the corresponding loop of the 6lu7 model, and the quality of the loop refinement was confirmed by comparison with 2h2z structure of SARS-CoV (Figure S8). Additionally, 4 and 3 Na⁺ ions were added to the SARS-CoV-2 and the SARS-CoV, respectively. Water molecules were placed using the combination of 3D-RISM [48] and the Placevent algorithm [49]. The AMBER 18 LEaP [50] was used to immerse models in a truncated octahedral box with 12 Å radius of TIP3P water molecules and prepare the systems for simulation using the ff14SB force field. PMEMD CUDA package of AMBER 18 software [50] was used to run a total of 2 μs (10 replicas of 50 ns for each system) simulations of both SARS-CoV-2 and SARS-CoV Mpros systems using apo structures and structure with co-crystallised N3 inhibitor (removed before starting the simulations) to provide more starting points for simulations. To improve conformation sampling, the starting geometry for each system was kept but the initial vectors were randomly assigned. The minimisation procedure consisted of 2000 steps, involving 1000 steepest descent steps followed by 1000 steps of conjugate gradient energy minimisation, with decreasing constraints on the protein backbone

(500, 125, and 25 kcal \times mol⁻¹ \times Å⁻²) and a final minimisation with no constraints of conjugate gradient energy minimization. Next, gradual heating was performed from 0 K to 300 K over 20 ps using a Langevin thermostat with a collision frequency of 1.0 ps⁻¹ in periodic boundary conditions with constant volume. Equilibration stage was run using the periodic boundary conditions with constant pressure for 1 ns with 1 fs step using Langevin dynamics with a frequency collision of 1 ps⁻¹ to maintain temperature. Production stage was run for 50 ns with a 2 fs time step using Langevin dynamics with a collision frequency of 1 ps⁻¹ to maintain constant temperature. Long-range electrostatic interactions were modelled using the particle mesh Ewald method with a non-bonded cut-off of 10 Å and the SHAKE algorithm. The coordinates were saved at an interval of 1 ps. The number of added water molecules is shown in Table S4.

Because our analysis was focused on the binding site that is surrounded by short loops only, to keep a reasonable combination of the number and length of simulations, the single simulation length was set to 50 ns. As has been shown elsewhere [51,52], longer simulations do not provide additional information and could even have a tendency to move away from the native-like structures. This hypothesis was verified on 200 ns long simulations where all observed changes were combined with the movement of the III domain (Figure S9).

Normal mode analysis for each system was conducted using cptraj from AmberTools 18. Only heavy atoms of the protein were included for analysis.

4.2. Mixed-Solvent MD Simulations—Cosolvent Preparation

Six different cosolvents: acetonitrile (ACN), benzene (BNZ), dimethylsulfoxide (DMSO), methanol (MEO), phenol (PHN), and urea (URE) were selected to perform the mixed-solvent MD simulations. The chemical structures of cosolvent molecules were downloaded from the ChemSpider database [53], and a dedicated set of parameters was prepared. Parameters for ACN were adopted from the work by Nikitin and Lyubartsev [54], and parameters for URE were modified using the 8Mureabox force field to obtain parameters for a single molecule. For the rest of the co-solvent molecules, parameters were prepared using Antechamber [55] with Gasteiger charges [56]. The number of added water and cosolvents molecules is shown in Table S5, and the parameters for cosolvents are available in Tables S6–S11.

4.3. Mixed-Solvent MD Simulations—Initial Configuration

The Packmol software [57] was used to build the initial systems, consisting of protein (protonated according to the previously described procedure), water, and particular cosolvent molecules. We added 4 and 3 Na⁺ ions to the SARS-CoV-2 Mpro and the SARS-CoV Mpro, respectively. It was assumed that the percentage concentration of the cosolvent should not exceed 5% (in the case of ACN, DMSO, MEO, and URE), or should be about 1% in the case of BNZ and PHN phenol (see Table S5). The mixed-solvent MD simulation procedures (minimization, equilibration, and production) carried out using the AMBER 18 package were identical to the classical MD simulations. Only the heating stage differed—it was extended up to 40 ps. PMEMD CUDA package of AMBER 18 software [50] was used to run two replicas of 50 ns for each cosolvent of both SARS-CoV-2 and SARS-CoV Mpro systems using apo structures and structure with co-crystallised N3 inhibitor (removed before starting the simulations), thus providing a total of 2.4 μ s of MixMD simulations.

4.4. Water and Cosolvent Molecule Tracking

The AQUA-DUCT 1.0 software [30] was used to track water and cosolvent molecules. Molecules of interest, which entered the so-called *Object*, defined as a 5 Å sphere around the centre of geometry of active site residues, namely, H41, C145, H164, and D187, were traced within the *Scope* region, defined as the interior of a convex hull of both COVID-19 Mpro and SARS Mpro C α atoms. All visualisations were made in PyMol [58].

AQUA-DUCT was used to analyse maximal accessible volume (MAV), defined as the *outer* pocket [30]. Pockets were calculated by analysis of paths found during molecules tracking step in AQUA-DUCT. A regular grid was constructed, spanning all paths. The grid size was 1 Å. For each grid cell, the density of tracked molecules was calculated. Grid cells with nonzero density were used for pocket detection; the *outer* pocket represented the maximal possible space that could be explored by tracked molecules.

To analyse the significance of the changes of the maximal accessible volume between systems, we used generalized linear models with a Poisson distribution based on AIC comparisons and model fit. Wald tests were used to test the significance of the variables. All tests were performed in Statistica (StatSoft 2019).

4.5. Hot-Spot Identification and Selection

AQUA-DUCT [30] was used to detect regions occupied by molecules of interest, as well as to identify the densest sites using a local solvent distribution approach. Those so-called hot-spots could be calculated as local and/or global, on the basis of the distribution of tracked molecules that visited the *Object* (local) or just the *Scope* without visiting the *Object* (global); here, they were considered as potential binding sites. For clarity, the size of each sphere representing a particular hot-spot was changed to reflect its occupation level. The selection of the most significant hot-spots consisted of indicating points showing the highest density in particular regions. From the set of points in the space, small groups of hot-spots were determined. Groups were further defined by distance (radius) from each other. Any point found within a distance shorter than the determined radius (3 Å) from any other point being part of a given group was counted toward the group. For each so designated group of points, one showing the highest density was chosen as representing the place.

4.6. Obtaining SARS-CoV-2 Mpro Gene Sequence

SARS-CoV-2 Mpro was downloaded from the PDB as a complex with an N3 inhibitor (PDB ID: 6lu7). Tblastn [59] was run on the basis of the protein amino acid sequence. We obtained 100% identity with 10055–10972 region of SARS-CoV-2 Mpro complete genome (Sequence ID: MN985262.1). Blastx [60] calculations were run with the selected region, and orf1a polyprotein (NCBI reference sequence: YP_009725295.1) amino acid sequence, identical with the previously downloaded SARS-CoV-2 Mpro, was received.

4.7. Energetic Effect of Amino Acid Substitutions

FoldX software [35] was used to insert substitutions into the structures of SARS-CoV and SARS-CoV-2 Mpros. To analyse the changes in energetic contribution to the protein stability of the two structures, 12 single-point mutations were introduced to the SARS structure using the BuildModel module. The BuildModel module introduces substitution(s) of selected amino acid(s), optimizes the structure of a new variant, and calculates the difference in the Gibbs free energy of protein folding between the wild-type and mutant variant in kilocalories per mole. The lower the difference in energetical terms, the more stable the mutant variant should be. Each of the residues in SARS-CoV Mpro was mutated to the respective SARS-CoV-2 Mpro residue, and the difference in total energies between the wild-type SARS-CoV-2 Mpro and the mutant structures were calculated. Then, to investigate further possible mutations of SARS-CoV-2 Mpro, single nucleotide substitutions were introduced to the SARS-CoV-2 main protease gene. If a substitution of a single nucleotide caused translation to an amino acid different than the corresponding residue in the wild-type structure, an appropriate mutation was proposed with FoldX software.

4.8. Comulator Calculations of Correlation Between Amino Acids

SARS-CoV Mpro was downloaded from the PDB (PDB ID: 1q2w). Blast [61] was run on the basis of the amino acid sequence. As a result, 2643 sequences of viral main proteases similar to chain

A SARS-CoV Mpro were obtained. Clustal Omega [62] was used to prepare an alignment of those sequences. Comulator [33] was then employed to calculate the correlation between amino acids and, on the basis of the results, groups of positions in SARS-CoV Mpro sequence were selected whose amino acid occurrences strongly depended on each other.

5. Conclusions

In this paper, we reported on molecular dynamics simulations of the main protease (Mpro), whose crystal structures have been released. We compared the Mpro for SARS-CoV-2 with a highly similar SARS-CoV protein. In spite of a high level of sequence similarity between these two homologous proteins, their active sites showed major differences in both shape and size, indicating that repurposing SARS drugs for COVID-19 may be futile. Furthermore, a detailed analysis of the binding pocket's conformational changes during simulation time indicated its flexibility and plasticity, which dashes hope for rapid and reliable drug design. Moreover, our findings show the presence of a flexible C44-P52 loop regulating the access to the binding site pocket. A successful inhibitor may need to have an ability to relocate the loop from the entrance to bind to the catalytic pocket. However, mutations leading to changes in the amino acid sequence of the C44-P52 loop, although not affecting the folding of the protein, may result in the putative inhibitors' inability to access the binding pocket and provide a probable development of drug resistance. To avoid this situation in which the future evolution of the Mpros can undermine all our efforts, we should focus on key functional residues or those whose further mutation will destabilise the protein (e.g., P39, R40, P52, G143, G146, or L167). Alternatively, we would suggest targeting the region between II and III domains, which contributes to the dimer formation. Our results provide the basis for drug design efforts aimed at this important protein target as part of the multifaceted global effort to eradicate COVID-19. In view of the presented challenges to finding a potent drug targeting Mpro, in our opinion, the most successful strategy would be to screen a large database of compounds with diverse structures involving or designing inhibitors de novo using a fragment-based approach. Both of these strategies, unfortunately, take much longer than a currently preferred approach based on repurposing existing FDA-approved compounds and hence should be pursued as a long-term plan of preparedness for future outbreaks of COVID epidemics involving this and other strains of the virus.

Supplementary Materials: Supplementary materials can be found at <http://www.mdpi.com/1422-0067/21/9/3099/s1>.

Author Contributions: Conceptualization, A.G.; methodology, M.B., K.M., A.R., and A.S.; software, M.B., K.M., A.R., and A.S.; validation, M.B., K.M., and A.G.; formal analysis, M.B., K.M., A.R., and A.S.; investigation, M.B., K.M., A.R., and A.S.; resources, M.B. and K.M.; data curation, M.B. and K.M.; writing—original draft preparation, M.B. and K.M.; writing—review and editing, J.A.T. and A.G.; visualization, M.B., K.M., A.R., and A.G.; supervision, A.G.; project administration, A.G.; funding acquisition, J.A.T. and A.G. All authors have read and agreed to the published version of the manuscript.

Funding: K.M., M.B., A.R., A.S., and A.G.'s work was supported by the National Science Centre, Poland (grant no. DEC-2013/10/E/NZ1/00649 and DEC-2015/18/M/NZ1/00427). J.A.T. expresses gratitude for research support for this project received from IBM CAS and NSERC (Canada).

Acknowledgments: The authors would like to thank Tomasz Skalski for his helpful advice in statistical analysis of data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Mpro	Main protease
SARS	Severe acute respiratory syndrome
COVID-19	Coronavirus disease 2019
SARS-CoV-2	Severe acute respiratory syndrome coronavirus 2
CoVs	Coronaviruses
HCoV-NL63	Human coronavirus NL63
HCoV-229E	Human coronavirus 229E

HCoV-OC43	Human coronavirus OC43
HCoV-HKU1	Human coronavirus HKU1
SARS-CoV	Severe acute respiratory syndrome coronavirus
MERS-CoV	Middle East respiratory syndrome coronavirus
ORFs	Open reading frames
3CLpro	Chymotrypsin-like cysteine protease
S	Spike surface glycoprotein
E	Small envelope protein
M	Matrix protein
N	Nucleocapsid protein
N3(PRD_002214)	N-[(5 methylisoxazol-3-yl)carbonyl]alanyl-L-valyl-N~1-((1R,2Z)-4-(benzyloxy)-4-oxo-1-[(3R)-2-oxopyrrolidin-3-yl]methyl]but-2-enyl)-L-leucinamide
cMD	Classical molecular dynamics simulations
MixMD	Mixed-solvent molecular dynamics simulations
PDB	Protein Data Bank
MAV	Maximal accessible volume
ACN	Acetonitrile
BNZ	Benzene
DMSO	Dimethylsulfoxide
MEO	Methanol
PHN	Phenol
URE	Urea
CMA	Correlated mutation analysis
FDA	Food and Drug Administration

References

- Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X.; et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet* **2020**, *395*, 497–506. [[CrossRef](#)]
- Zhu, N.; Zhang, D.; Wang, W.; Li, X.; Yang, B.; Song, J.; Zhao, X.; Huang, B.; Shi, W.; Lu, R.; et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J. Med.* **2020**, *382*, 727–733. [[CrossRef](#)] [[PubMed](#)]
- Woo, P.C.Y.; Huang, Y.; Lau, S.K.P.; Yuen, K.-Y. Coronavirus Genomics and Bioinformatics Analysis. *Viruses* **2010**, *2*, 1804–1820. [[CrossRef](#)] [[PubMed](#)]
- Tang, Q.; Song, Y.; Shi, M.; Cheng, Y.; Zhang, W.; Xia, X.-Q. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. *Sci. Rep.* **2015**, *5*, 17155. [[CrossRef](#)] [[PubMed](#)]
- Cui, J.; Li, F.; Shi, Z.-L. Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol.* **2019**, *17*, 181–192. [[CrossRef](#)]
- Fehr, A.R.; Perlman, S. Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods Mol. Biol.* **2015**, *1282*, 1–23.
- Zhang, L.; Shen, F.; Chen, F.; Lin, Z. Origin and evolution of the 2019 novel coronavirus. *Clin. Infect. Dis.* **2020**. [[CrossRef](#)]
- Song, Z.; Xu, Y.; Bao, L.; Zhang, L.; Yu, P.; Qu, Y.; Zhu, H.; Zhao, W.; Han, Y.; Qin, C. From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses* **2019**, *11*, 59. [[CrossRef](#)]
- Xue, X.; Yu, H.; Yang, H.; Xue, F.; Wu, Z.; Shen, W.; Li, J.; Zhou, Z.; Ding, Y.; Zhao, Q.; et al. Structures of Two Coronavirus Main Proteases: Implications for Substrate Binding and Antiviral Drug Design. *J. Virol.* **2008**, *82*, 2515–2527. [[CrossRef](#)]
- Wu, A.; Peng, Y.; Huang, B.; Ding, X.; Wang, X.; Niu, P.; Meng, J.; Zhu, Z.; Zhang, Z.; Wang, J.; et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host Microbe* **2020**, *27*, 325–328. [[CrossRef](#)]
- Zumla, A.; Chan, J.F.W.; Azhar, E.I.; Hui, D.S.C.; Yuen, K.-Y. Coronaviruses—Drug discovery and therapeutic options. *Nat. Rev. Drug Discov.* **2016**, *15*, 327–347. [[CrossRef](#)] [[PubMed](#)]
- Liu, W.; Morse, J.S.; Lalonde, T.; Xu, S. Learning from the Past: Possible Urgent Prevention and Treatment Options for Severe Acute Respiratory Infections Caused by 2019-nCoV. *ChemBioChem* **2020**, *21*, 730–738.

13. Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of Mpro from COVID-19 virus and discovery of its inhibitors. *Nature* **2020**. [[CrossRef](#)] [[PubMed](#)]
14. Zhong, N.; Zhang, S.; Zou, P.; Chen, J.; Kang, X.; Li, Z.; Liang, C.; Jin, C.; Xia, B. Without Its N-Finger, the Main Protease of Severe Acute Respiratory Syndrome Coronavirus Can Form a Novel Dimer through Its C-Terminal Domain. *J. Virol.* **2008**, *82*, 4227–4234. [[CrossRef](#)] [[PubMed](#)]
15. Ton, A.-T.; Gentile, F.; Hsing, M.; Ban, F.; Cherkasov, A. Rapid Identification of Potential Inhibitors of SARS-CoV-2 Main Protease by Deep Docking of 1.3 Billion Compounds. *Mol. Inform.* **2020**. [[CrossRef](#)]
16. Xu, Z.; Peng, C.; Shi, Y.; Zhu, Z.; Mu, K.; Wang, X.; Zhu, W. Nelfinavir was predicted to be a potential inhibitor of 2019-nCoV main protease by an integrative approach combining homology modelling, molecular docking and binding free energy calculation. *bioRxiv* **2020**. [[CrossRef](#)]
17. Liu, X.; Wang, X.-J. Potential inhibitors for 2019-nCoV coronavirus M protease from clinically approved medicines. *J. Genet. Genomics* **2020**, *47*, 119–121. [[CrossRef](#)]
18. Li, Y.; Zhang, J.; Wang, N.; Li, H.; Shi, Y.; Guo, G.; Liu, K.; Hao, Z.; Zou, Q. Therapeutic Drugs Targeting 2019-nCoV Main Protease by High-Throughput Screening. *bioRxiv* **2020**. [[CrossRef](#)]
19. Nguyen, D.D.; Gao, K.; Chen, J.; Wang, R.; Wei, G.-W. Potentially highly potent drugs for 2019-nCoV. *bioRxiv* **2020**. [[CrossRef](#)]
20. Talluri, S. Virtual High Throughput Screening Based Prediction of Potential Drugs for COVID-19. *Preprints* **2020**. [[CrossRef](#)]
21. Chen, Y.W.; Yiu, C.-P.B.; Wong, K.-Y. Prediction of the SARS-CoV-2 (2019-nCoV) 3C-like protease (3CLpro) structure: virtual screening reveals velpatasvir, ledipasvir, and other drug repurposing candidates. *F1000Research* **2020**, *9*, 129. [[CrossRef](#)] [[PubMed](#)]
22. Fischer, A.; Sellner, M.; Neranjan, S.; Lill, M.A.; Smieško, M. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. *ChemRxiv* **2020**. [[CrossRef](#)]
23. Gentile, D.; Patamia, V.; Scala, A.; Sciortino, M.T.; Piperno, A.; Rescifina, A. Inhibitors of SARS-CoV-2 Main Protease from a Library of Marine Natural Products: A Virtual Screening and Molecular Modeling Study. *Mar. Drugs* **2020**, *18*, 225. [[CrossRef](#)]
24. Adem, S.; Eyupoglu, V.; Sarfraz, I.; Rasul, A.; Ali, M. Identification of potent COVID-19 main protease (Mpro) inhibitors from natural polyphenols: An in silico strategy unveils a hope against CORONA. *Preprints* **2020**. [[CrossRef](#)]
25. Berman, H.M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]
26. Zhang, L.; Lin, D.; Sun, X.; Curth, U.; Drosten, C.; Sauerhering, L.; Becker, S.; Rox, K.; Hilgenfeld, R. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α -ketoamide inhibitors. *Science* **2020**, *368*, 409–412. [[CrossRef](#)] [[PubMed](#)]
27. Bacha, U.; Barrila, J.; Velazquez-Campoy, A.; Leavitt, S.A.; Freire, E. Identification of Novel Inhibitors of the SARS Coronavirus Main Protease 3CL pro[†]. *Biochemistry* **2004**, *43*, 4906–4912. [[CrossRef](#)]
28. Anand, K. Coronavirus Main Proteinase (3CLpro) Structure: Basis for Design of Anti-SARS Drugs. *Science* **2003**, *300*, 1763–1767. [[CrossRef](#)]
29. 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. [[CrossRef](#)]
30. 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* **2019**. [[CrossRef](#)]
31. Li, C.; Qi, Y.; Teng, X.; Yang, Z.; Wei, P.; Zhang, C.; Tan, L.; Zhou, L.; Liu, Y.; Lai, L. Maturation Mechanism of Severe Acute Respiratory Syndrome (SARS) Coronavirus 3C-like Proteinase. *J. Biol. Chem.* **2010**, *285*, 28134–28140. [[CrossRef](#)]
32. Panjkovich, A.; Daura, X. PARS: A web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics* **2014**, *30*, 1314–1315. [[CrossRef](#)] [[PubMed](#)]
33. Kuipers, R.K.; Joosten, H.-J.; van Berkel, W.J.H.; Leferink, N.G.H.; Rooijen, E.; Ittmann, E.; van Zimmeren, F.; Jochens, H.; Bornscheuer, U.; Vriend, G.; et al. 3DM: Systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins Struct. Funct. Bioinform.* **2010**, *78*, 2101–2113. [[CrossRef](#)] [[PubMed](#)]

34. Subramanian, K.; Mitusińska, K.; Raedts, J.; Almourfi, F.; Joosten, H.-J.; Hendriks, S.; Sedelnikova, S.E.; Kengen, S.W.M.; Hagen, W.R.; Góra, A.; et al. Distant Non-Obvious Mutations Influence the Activity of a Hyperthermophilic *Pyrococcus furiosus* Phosphoglucose Isomerase. *Biomolecules* **2019**, *9*, 212. [[CrossRef](#)] [[PubMed](#)]
35. Schymkowitz, J.; Borg, J.; Stricher, F.; Nys, R.; Rousseau, F.; Serrano, L. The FoldX web server: An online force field. *Nucleic Acids Res.* **2005**, *33*, W382–W388. [[CrossRef](#)] [[PubMed](#)]
36. Klausen, M.S.; Jespersen, M.C.; Nielsen, H.; Jensen, K.K.; Jurtz, V.I.; Sønderby, C.K.; Sommer, M.O.A.; Winther, O.; Nielsen, M.; Petersen, B.; et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct. Funct. Bioinform.* **2019**, *87*, 520–527. [[CrossRef](#)]
37. Spyraakis, F.; Ahmed, M.H.; Bayder, A.S.; Cozzini, P.; Mozzarelli, A.; Kellogg, G.E. The Roles of Water in the Protein Matrix: A Largely Untapped Resource for Drug Discovery. *J. Med. Chem.* **2017**, *60*, 6781–6827. [[CrossRef](#)]
38. De Beer, S.B.A.; Vermeulen, N.P.E.; Oostenbrink, C. The Role of Water Molecules in Computational Drug Design. *Curr. Top. Med. Chem.* **2010**, *10*, 55–66. [[CrossRef](#)]
39. Mitusińska, K.; Magdziarz, T.; Bzówka, M.; Stańczak, A.; Góra, A. Exploring *Solanum tuberosum* Epoxide Hydrolase Internal Architecture by Water Molecules Tracking. *Biomolecules* **2018**, *8*, 143. [[CrossRef](#)]
40. Needle, D.; Lountos, G.T.; Waugh, D.S. Structures of the Middle East respiratory syndrome coronavirus 3C-like protease reveal insights into substrate specificity. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2015**, *71*, 1102–1111. [[CrossRef](#)]
41. Tsai, M.-Y.; Chang, W.-H.; Liang, J.-Y.; Lin, L.-L.; Chang, G.-G.; Chang, H.-P. Essential covalent linkage between the chymotrypsin-like domain and the extra domain of the SARS-CoV main protease. *J. Biochem.* **2010**, *148*, 349–358. [[CrossRef](#)]
42. Anand, K. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* **2002**, *21*, 3213–3224. [[CrossRef](#)] [[PubMed](#)]
43. Lim, L.; Shi, J.; Mu, Y.; Song, J. Dynamically-Driven Enhancement of the Catalytic Machinery of the SARS 3C-Like Protease by the S284-T285-I286/A Mutations on the Extra Domain. *PLoS ONE* **2014**, *9*, e101941. [[CrossRef](#)] [[PubMed](#)]
44. Chang, C.; Jeyachandran, S.; Hu, N.-J.; Liu, C.-L.; Lin, S.-Y.; Wang, Y.-S.; Chang, Y.-M.; Hou, M.-H. Structure-based virtual screening and experimental validation of the discovery of inhibitors targeted towards the human coronavirus nucleocapsid protein. *Mol. Biosyst.* **2016**, *12*, 59–66. [[CrossRef](#)] [[PubMed](#)]
45. Dayer, M.R.; Taleb-Gassabi, S.; Dayer, M.S. Lopinavir; A Potent Drug against Coronavirus Infection: Insight from Molecular Docking Study. *Arch. Clin. Infect. Dis.* **2017**, *12*. [[CrossRef](#)]
46. Resnick, E.; Bradley, A.; Gan, J.; Douangamath, A.; Krojer, T.; Sethi, R.; Geurink, P.P.; Aimon, A.; Amitai, G.; Bellini, D.; et al. Rapid Covalent-Probe Discovery by Electrophile-Fragment Screening. *J. Am. Chem. Soc.* **2019**, *141*, 8951–8968. [[CrossRef](#)] [[PubMed](#)]
47. Anandkrishnan, R.; Aguilar, B.; Onufriev, A.V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541. [[CrossRef](#)]
48. Luchko, T.; Gusarov, S.; Roe, D.R.; Simmerling, C.; Case, D.A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2010**, *6*, 607–624. [[CrossRef](#)]
49. Sindhikara, D.J.; Yoshida, N.; Hirata, F. Placevent: An algorithm for prediction of explicit solvent atom distribution-Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **2012**, *33*, 1536–1543. [[CrossRef](#)]
50. Case, D.A.; Ben-Shalom, I.Y.; Brozell, S.R.; Cerutti, D.S.; Cheatham III, T.E.; Cruzeiro, V.W.D.; Darden, T.A.; Duke, R.E.; Ghoreishi, D.; Gilson, M.K.; et al. *AMBER 2018*; University of California: San Francisco, CA, USA, 2018.
51. Heo, L.; Feig, M. What makes it difficult to refine protein models further via molecular dynamics simulations? *Proteins Struct. Funct. Bioinform.* **2018**, *86*, 177–188. [[CrossRef](#)]
52. Mitusińska, K.; Skalski, T.; Góra, A. Simple selection procedure to distinguish between static and flexible loops. *Int. J. Mol. Sci.* **2020**, *21*, 2293. [[CrossRef](#)]
53. Pence, H.E.; Williams, A. ChemSpider: An Online Chemical Information Resource. *J. Chem. Educ.* **2010**, *87*, 1123–1124. [[CrossRef](#)]

54. Nikitin, A.M.; Lyubartsev, A.P. New six-site acetonitrile model for simulations of liquid acetonitrile and its aqueous mixtures. *J. Comput. Chem.* **2007**, *28*, 2020–2026. [[CrossRef](#)] [[PubMed](#)]
55. Wang, J.; Wang, W.; Kollman, P.A.; Case, D.A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. Mol. Graph. Model.* **2006**, *25*, 247–260. [[CrossRef](#)]
56. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228. [[CrossRef](#)]
57. Martínez, L.; Andrade, R.; Birgin, E.G.; Martínez, J.M. PACKMOL: A package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **2009**, *30*, 2157–2164. [[CrossRef](#)] [[PubMed](#)]
58. *The PyMOL Molecular Graphics System, Version 2.0 Schrödinger*; Schrödinger LLC.: New York, NY, USA.
59. Gertz, E.M.; Yu, Y.-K.; Agarwala, R.; Schäffer, A.A.; Altschul, S.F. Composition-based statistics and translated nucleotide searches: Improving the TBLASTN module of BLAST. *BMC Biol.* **2006**, *4*, 41. [[CrossRef](#)]
60. Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T.L. BLAST+: Architecture and applications. *BMC Bioinform.* **2009**, *10*, 421. [[CrossRef](#)] [[PubMed](#)]
61. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [[CrossRef](#)]
62. Sievers, F.; Higgins, D.G. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Methods Mol. Biol.* **2014**, *1079*, 105–106. [[CrossRef](#)] [[PubMed](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).



Article

Computational Selectivity Assessment of Protease Inhibitors against SARS-CoV-2

André Fischer ^{1,†} , Manuel Sellner ^{1,†} , Karolina Mitusińska ^{2,†} , Maria Bzówka ^{2,†} , Markus A. Lill ^{1,*} , Artur Góra ^{2,*} and Martin Smieško ^{1,*}

¹ Computational Pharmacy, Department of Pharmaceutical Sciences, University of Basel, 4056 Basel, Switzerland; and.fischer@unibas.ch (A.F.); manuel.sellner@unibas.ch (M.S.)

² Tunneling Group, Biotechnology Centre, ul. Krzywoustego 8, Silesian University of Technology, 44-100 Gliwice, Poland; k.mitusinska@tunnelinggroup.pl (K.M.); m.bzowka@tunnelinggroup.pl (M.B.)

* Correspondence: markus.lill@unibas.ch (M.A.L.); a.gora@tunnelinggroup.pl (A.G.); martin.smiesko@unibas.ch (M.S.)

† These authors contributed equally to this work.

Abstract: The pandemic of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) poses a serious global health threat. Since no specific therapeutics are available, researchers around the world screened compounds to inhibit various molecular targets of SARS-CoV-2 including its main protease (M^{Pro}) essential for viral replication. Due to the high urgency of these discovery efforts, off-target binding, which is one of the major reasons for drug-induced toxicity and safety-related drug attrition, was neglected. Here, we used molecular docking, toxicity profiling, and multiple molecular dynamics (MD) protocols to assess the selectivity of 33 reported non-covalent inhibitors of SARS-CoV-2 M^{Pro} against eight proteases and 16 anti-targets. The panel of proteases included SARS-CoV M^{Pro}, cathepsin G, caspase-3, ubiquitin carboxy-terminal hydrolase L1 (UCHL1), thrombin, factor Xa, chymase, and prostaticin. Several of the assessed compounds presented considerable off-target binding towards the panel of proteases, as well as the selected anti-targets. Our results further suggest a high risk of off-target binding to chymase and cathepsin G. Thus, in future discovery projects, experimental selectivity assessment should be directed toward these proteases. A systematic selectivity assessment of SARS-CoV-2 M^{Pro} inhibitors, as we report it, was not previously conducted.

Keywords: coronavirus; SARS; protease; selectivity; structure-based design



Citation: 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. *Int. J. Mol. Sci.* **2021**, *22*, 2065. <https://doi.org/10.3390/ijms22042065>

Academic Editors: Cristina Belizna, Jan Willem Cohen Tervaert, Yehuda Shoenfeld and Alexander Makatsariya

Received: 4 January 2021

Accepted: 11 February 2021

Published: 19 February 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In late 2019, a novel coronavirus termed SARS-CoV-2 emerged and spread around the world causing coronavirus disease 2019 (COVID-19). Until today, over 62 million cases were reported accounting for over a 1.46 million of fatalities (as of 1 December 2020) [1]. While pharmaceutical interventions primarily remained symptomatic, multiple clinical trials are investigating novel treatments, mainly based on drug repurposing [2,3]. Thus, the treatment of this infection with specific drugs constitutes an urgent and unmet medical need. In the pharmaceutical treatment of viral infections such as human immunodeficiency virus and hepatitis C virus, the inhibition of viral proteases is a successfully applied strategy. Consequently, many computational and experimental efforts were directed toward targeting the main protease (M^{Pro}) of SARS-CoV-2 with small molecules leading to the discovery of multiple promising candidates [4–9]. Due to the high urgency and the competitive scientific field, off-target binding was rarely considered in the latest discovery projects. However, a large share of drug attrition in clinical trials, especially regarding compound safety, can be traced back to low target specificity and off-target binding [10–12]. To avoid the large cost associated with late stage drug failure, early off-target profiling, especially with comparably economical computational methods, offers an attractive strategy [12–14]. On the other hand, binding to multiple targets can be beneficial in specific

cases, such as pan inhibition of the viral proteases of SARS-CoV-2 and SARS-CoV [15]. Similarly, the concurrent inhibition of the coagulation protein factor Xa and SARS-CoV-2 M^{Pro} constitutes another example for a potentially synergistic effect of multi-target binding as COVID-19 infection is associated with life-threatening coagulopathies that can be treated with anticoagulants [16,17]. Computational methods such as molecular docking and specialized molecular dynamics (MD) protocols can be exploited to explore the selectivity of small-molecule compounds, as it was evidenced for various targets. For example, we previously applied docking combined with cosolvent MD simulations and determination of hydration hot-spots to investigate the selectivity of allosteric inhibitors against eight nuclear receptors to support targeted experimental profiling of novel compounds [18]. In general, it was discussed that cosolvent MD simulations can provide valuable insights for the development of potent and selective compounds [19,20]. Although multiple studies focused on the selectivity assessment among kinases using computational methods [21,22], there were only minor efforts to establish selectivity factors for small-molecules that bind to proteases [23,24].

Here, we examined the selectivity of 33 experimentally confirmed non-covalent SARS-CoV-2 M^{Pro} inhibitors against eight different proteases including SARS-CoV M^{Pro}, factor Xa, cathepsin G, caspase-3, prostatic, thrombin, ubiquitin carboxy-terminal hydrolase L1 (UCHL1), and chymase by molecular docking (Table 1). The proteases were selected based on a structural similarity search in the NCBI database, as well as considerations regarding their pharmacological relevance. First, we compared the active sites regarding pharmacophores and electrostatic potential. Furthermore, we performed classical as well as cosolvent MD simulations to identify water and small-molecule hot-spots of the respective active sites offering explanations for compound selectivity. Based on our results, experimental selectivity profiling in future discovery projects can be directed toward targets with an inherent high liability for off-target binding. Up to this date, such a comprehensive evaluation of off-target binding of SARS-CoV-2 M^{Pro} was not previously conducted and is of high importance to support antiviral drug development.

Table 1. Proteins considered in this study.

Protein	Function	Anti-Target ^a	Consequence of Inhibition
SARS-CoV-2 M ^{Pro}	Viral replication	-	antiviral activity
SARS-CoV M ^{Pro}	Viral replication	no	antiviral activity
Caspase-3	Apoptosis	yes	interference with development
Factor Xa	Coagulation	no	prevention of coagulopathies
Cathepsin G	Immune system	yes	interference with immune response
UCHL1	Protein degradation	yes	interference with development and homeostasis
Prostatic	Sodium balance	yes	alters homeostasis
Thrombin	Coagulation	no	prevention of coagulopathies
Chymase	Vasoconstriction	yes	interference with blood pressure

^a Description if the protein is regarded as anti-target.

2. Results and Discussion

2.1. Sequence and Active Site Comparison

We selected eight proteases to assess off-target binding and to structurally compare them to SARS-CoV-2 M^{Pro}. The proteases were selected based on their catalytic residues, sequence and structural similarity, availability of structural information, as well as their pharmacological and physiological relevance. Except for SARS-CoV M^{Pro}, they exhibited different overall folds and showed low global sequence similarity to SARS-CoV-2 M^{Pro} (Table 2). Furthermore, the volumes of the active site cavity (Table S1) of the SARS-CoV-2 M^{Pro} was the smallest among all analyzed proteases, and regarding absolute values, most similar to chymase as opposed to SARS-CoV M^{Pro}. However, when we compared the active sites of all analyzed proteases, they presented a remarkable degree of similarity. The root

mean-square deviation (RMSD) of their catalytic residues did not exceed 2 Å, except for UCHL1 with 2.8 Å. Furthermore, when using FuzCav [25] to determine the similarity of the active site pockets of the panel of proteases, we observed that not only their catalytic residues, but also the complete active sites are similar relative to SARS-CoV-2 M^{Pro} (Supplementary Figure S1). Active sites exceeding a similarity value of 0.16 can be regarded as similar [25]. We also compared the electrostatic potentials of the binding cavities of all analyzed proteases using PIPSA [26] software. The electrostatic potentials were compared quantitatively by calculating the Hodgkin similarity index (Table 2, Figure S2). Using the Hodgkin similarity index it is possible to determine the correlation between the potentials of the analysed proteases (+1 indicates that potentials are identical, 0 indicates that potentials are fully uncorrelated, and -1 indicates that potentials are anti-correlated). Three out of eight proteases (SARS-CoV M^{Pro}, prostaticin, and thrombin) presented a positive correlation, whereas the remaining ones indicated an anti-correlation in relation to the SARS-CoV-2 M^{Pro} binding cavity. In the case of UCHL1, the high anti-correlation was caused by a specific binding site spot with reversed distribution of electrostatic potentials. Prostaticin, thrombin and SARS-CoV M^{Pro} also showed high similarity relative to the active site of SARS-CoV-2 M^{Pro} using FuzCav. Thus, even though the substrate specificity of the panel of proteases is diverse [27], their exceptionally similar active sites compared to SARS-CoV-2 may suggest a potential for off-target binding of small-molecules.

Table 2. Similarity of proteins and active sites.

Protein	Catalytic Residues	Global Identity ^a	AS RMSD (Å) ^b	Site Similarity ^c	Fold	Similarity Index ^d
SARS-CoV M ^{Pro}	H41, C145	96.1%	0.2	0.91	α/β	0.90
Caspase-3	H121, C163	11.6%	1.9	0.67	α/β	-0.74
Factor Xa	H57, D102, S195	11.6%	1.8	0.71	all- β	-0.67
Cathepsin G	H59, D103, S196	14.5%	1.8	0.61	all- β	-0.71
UCHL1	C90, H161, D176	15.7%	2.8	0.74	α/β	-0.98
Prostaticin	H85, D134, S154	13.1%	1.6	0.69	all- β	0.26
Thrombin	H57, D102, S195	12.5%	1.8	0.74	all- β	0.31
Chymase	H45, D89, S182	19.0%	1.9	0.64	all- β	-0.49

^a Sequence identity of the catalytic unit to SARS-CoV-2 M^{Pro}; ^b RMSD between histidine, cysteine/serine, and aspartic acid residues relative to SARS-CoV-2 M^{Pro}; ^c Similarity of the binding sites to SARS-CoV-2 M^{Pro} determined by FuzCav; ^d Hodgkin similarity index of the binding sites comparison relative to the SARS-CoV-2 M^{Pro} determined by PIPSA.

2.2. Hydration and Small-Molecule Hot-Spots

In the next step, we further characterized and compared the selected proteases according to their hydration and small-molecule hot-spots by using different molecular probes including water, acetonitrile, isopropanol, and pyridine. The use of specific functional groups represented by the different organic probes associating with the active sites can be used to fine-tune the selectivity profile of protease inhibitors [18,28]. Similarly, selectively targeting hydration sites occurring in one protein, but not in an anti-target, offers potential to be exploited in structure-based design. It should however be mentioned that whether the displacement of water molecules is favorable or not depends on the thermodynamic profile of the respective hydration site [29], which was not assessed in this work. The comparison of the hot-spots in the vicinity of the active site residues revealed distinct similarities (Figure 1).

For SARS-CoV-2 M^{Pro}, we identified two hydration sites located in the vicinity of H41, as well as a small-molecule hot-spot for acetonitrile molecules at the same location (Figure 2). While we could not detect a hydration site in the vicinity of C141, association of pyridine and isopropanol was detected. In the case of SARS-CoV M^{Pro}, however, four hydration sites could be identified. Potentially, the increased flexibility of SARS-CoV M^{Pro} allowed for increased solvent accessibility in comparison to SARS-CoV-2 M^{Pro} (Table S1) [30]. Furthermore, the increased magnitude of cosolvent densities in SARS-CoV M^{Pro} confirmed this observation, although they mainly occupied the vicinity of H41. Even though one hydration site between the M^{Pro}s overlapped, there were signifi-

cant differences between the small-molecule hot-spots of the SARS-CoV proteases, which have to be accounted for in the design of pan inhibitors against the two coronaviruses (Figure 2). For caspase-3, two unique hydration sites were identified in the active site cavity distant from the catalytic residues, which were not observed in the SARS-CoV-2 M^{Pro}. One of the hydration sites overlapped with the occupancy of multiple organic probes, which sampled a unique region not observed in any other protease besides UCHL1. Three hydration sites were identified in the vicinity of the active site residues in UCHL1. One of the hydration sites was located between three catalytic residues as in caspase-3, but none of the other cysteine proteases. Thus, these proteases are similar, while presenting distinct differences to the M^{Pro}s despite their shared catalytic mechanism using cysteine for the nucleophilic attack of the substrate (Table 2). In the active site of factor Xa, we detected two hydration sites matching the position in prostatic and chymase indicating that they are conserved. As one of the sites (denoted as site B in Figure 1) could not be observed in the SARS-CoV-2 M^{Pro}, it might contribute to ligand specificity. Remarkably, the cosolvent densities among factor Xa, cathepsin G, thrombin, and chymase presented a high degree of similarity, with an additional density for acetonitrile compared to SARS-CoV-2 M^{Pro}. The same region was occupied by pyridine probes. A common density of pyridine in the center of the sites, however, suggested a common preference for hydrophobic or aromatic moieties among the aforementioned enzymes. Comparing the organic probe density of factor Xa to SARS-CoV-2 M^{Pro}, a common preference for acetonitrile on the distal side of the catalytic histidine could be observed. This could guide the placement of an amphipathic moiety in this region to inhibit both proteases with future antivirals. While cathepsin G and thrombin shared one of the most commonly observed hydration sites, they lack the common site observed in the viral M^{Pro}s which indicates that the displacement of this water molecule (denoted as site A in Figure 1) could contribute to selective binding. Both thrombin and chymase presented a high number of hydration sites within their active sites sharing the above-mentioned hydration site B not observed in the viral proteases, as well as the previously discussed acetonitrile density near the backbone of the catalytic histidine residue. Thus, the inherent potential for off-target binding of novel antivirals to these targets is small, similar to caspase-3 and UCHL1. Compared to the volume of its active site cavity, the site of chymase seemed highly accessible to the surrounding solvent. Overall, the highest similarity among the proteases regarding hydration sites and small-molecule binding hot-spots, could be observed for factor Xa and chymase. Differences in hydration site locations identified for individual simulations are most probably related to conformational changes of the proteins, as the volumes of the active sites underlined (Table S1).

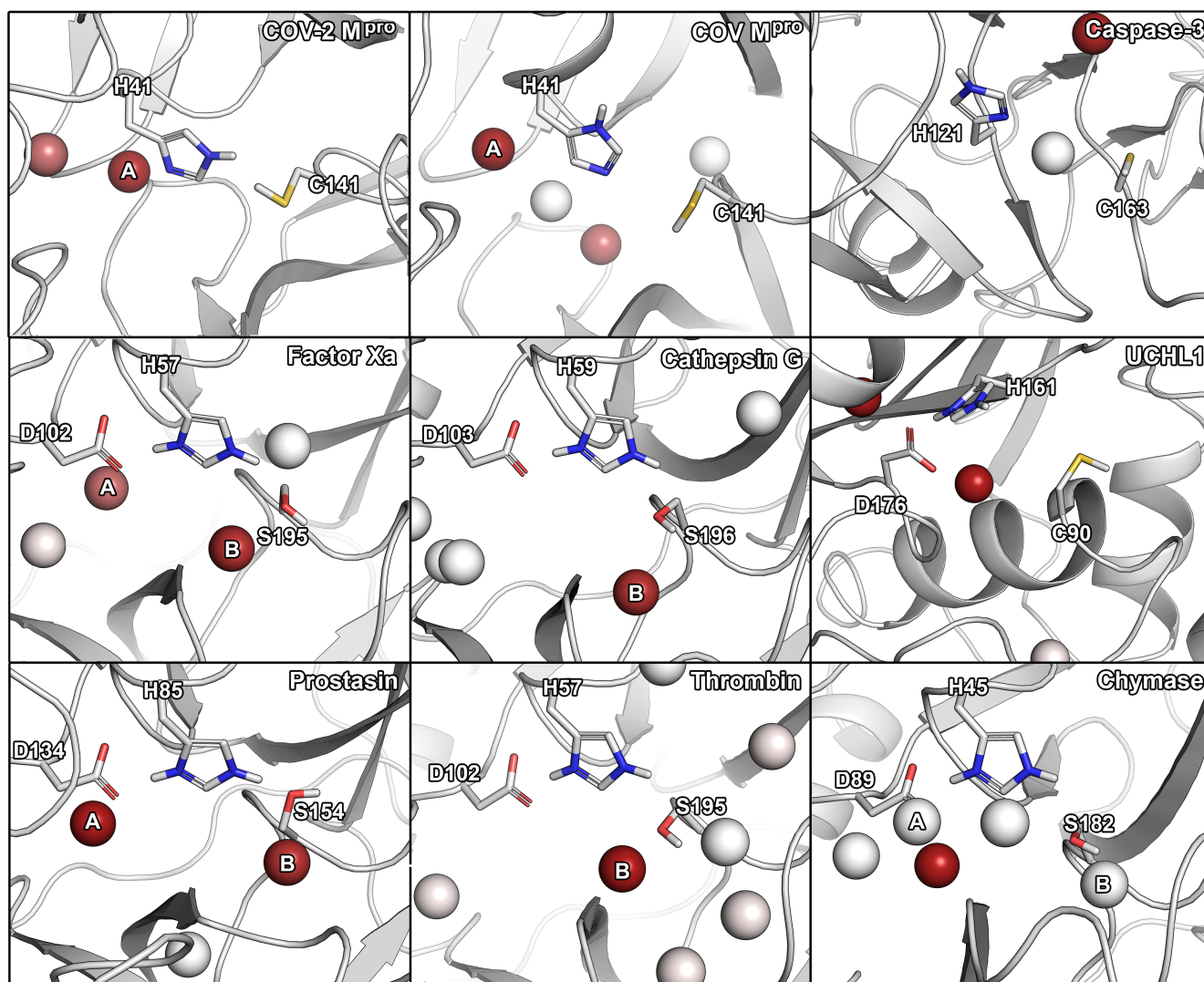


Figure 1. Hydration hot-spots of the selected panel of proteases in relation to the catalytic residues. To allow a direct comparison, the protein structures were aligned according to their catalytic residues. Two most consistently occurring hydration sites are indicated by A and B. The hot-spots are color-coded according to the occupancy of a particular region by identified hydration hot-spots. Hydration hot-spots with highest occupancy are colored in dark red, those with low occupancy in white.

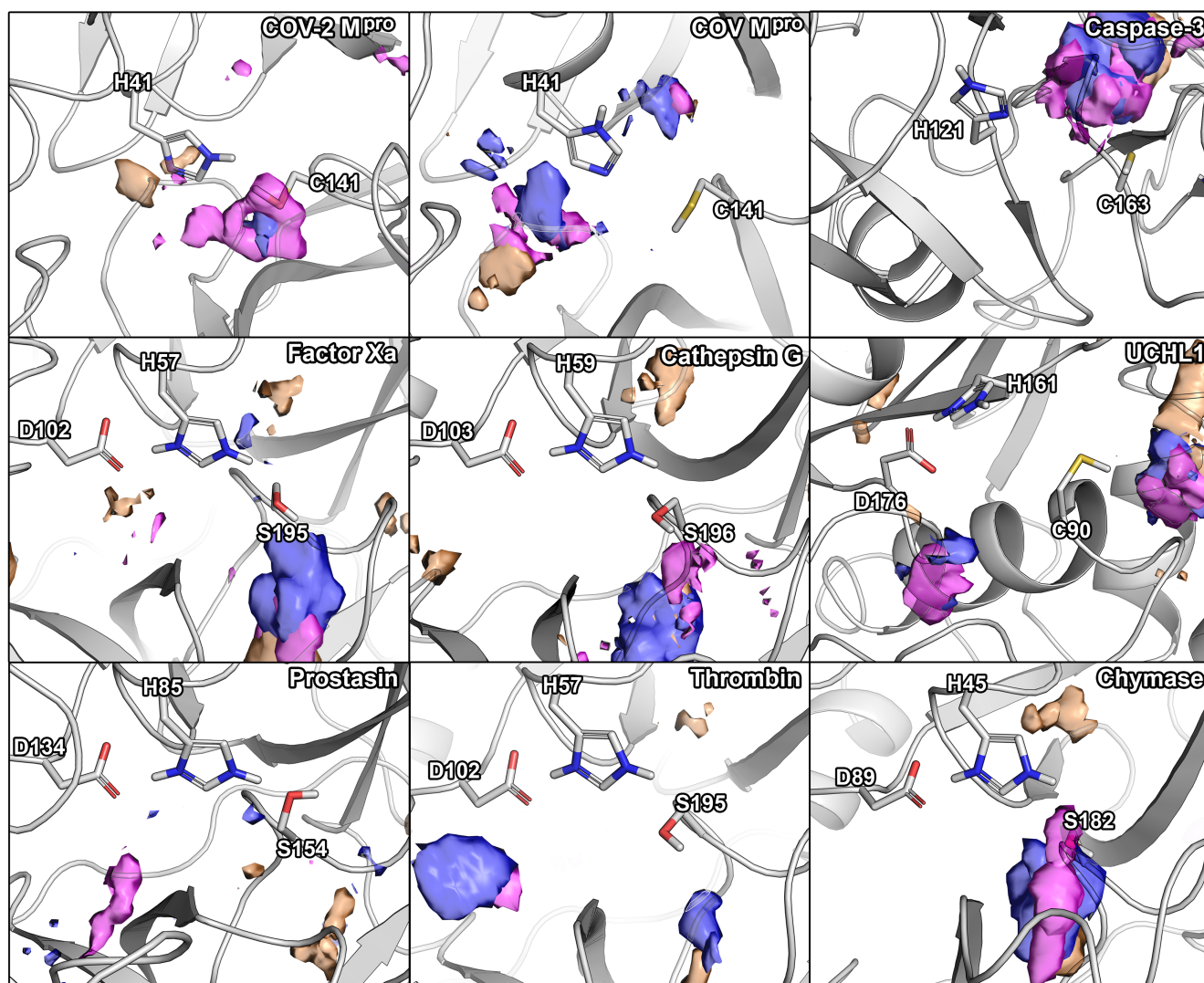


Figure 2. Small-molecule hot-spots of the selected panel of proteases in relation to their catalytic residues. Blue densities correspond to isopropanol, pink densities to pyridine, and orange densities to acetonitrile.

2.3. Protease Selectivity Assessed by Molecular Docking

Binding modes obtained from molecular docking have been widely used to establish selectivity factors toward different targets [13,18,31]. Here, we compiled compound sets comprising experimentally verified ligands of nine proteases to assess the selectivity of recently reported SARS-CoV-2 M^{Pro} inhibitors (Figure 3) by cross-docking them into the respective protein active sites. These known binders were either retrieved from a set of cocrystallized ligands, the PubChem BioAssay database [32], or the literature (Tables S2–S10, Figures S3–S18). First, to ensure accurate pose-prediction of our computational models, we retrieved numerous crystal structures from the Protein Data Bank for each target and cross-docked their cocrystallized ligands, which is a common procedure in virtual screening projects [33]. Based on the obtained RMSD values between predicted and native binding poses, ensembles of protein structures that yielded best-possible pose prediction quality for non-covalent ligands were identified. In the case of unsatisfactory performance of these structures, short MD simulations were performed to enrich the proteins' structural diversity. These procedures were performed for the Glide standard precision (SP) and smina docking protocol, to address known differences among docking programs. This resulted in excellent docking accuracy for most protein systems which ranged from 75% to 100% of cocrystallized ligands being predicted below an RMSD threshold of 2.5 Å.

The only exception was prostatic, for which we only found an accurate pose for one of the two available ligands (Table S11). Unfortunately, no non-covalent small molecule was cocrystallized with UCHL1 which prevented us from computing these metrics in this case. In a next step, we evaluated the performance of the selected ensembles to distinguish between known actives and randomly selected decoy molecules based on the Area Under the Curve (AUC) of the Receiver Operator Characteristic (ROC) curves. Considering the best score of each compound against the ensemble, acceptable ROC AUC values between 0.631 and 0.953 were obtained demonstrating the accuracy of our models and procedures in both detecting the actives and predicting bioactive conformations (Table S12).

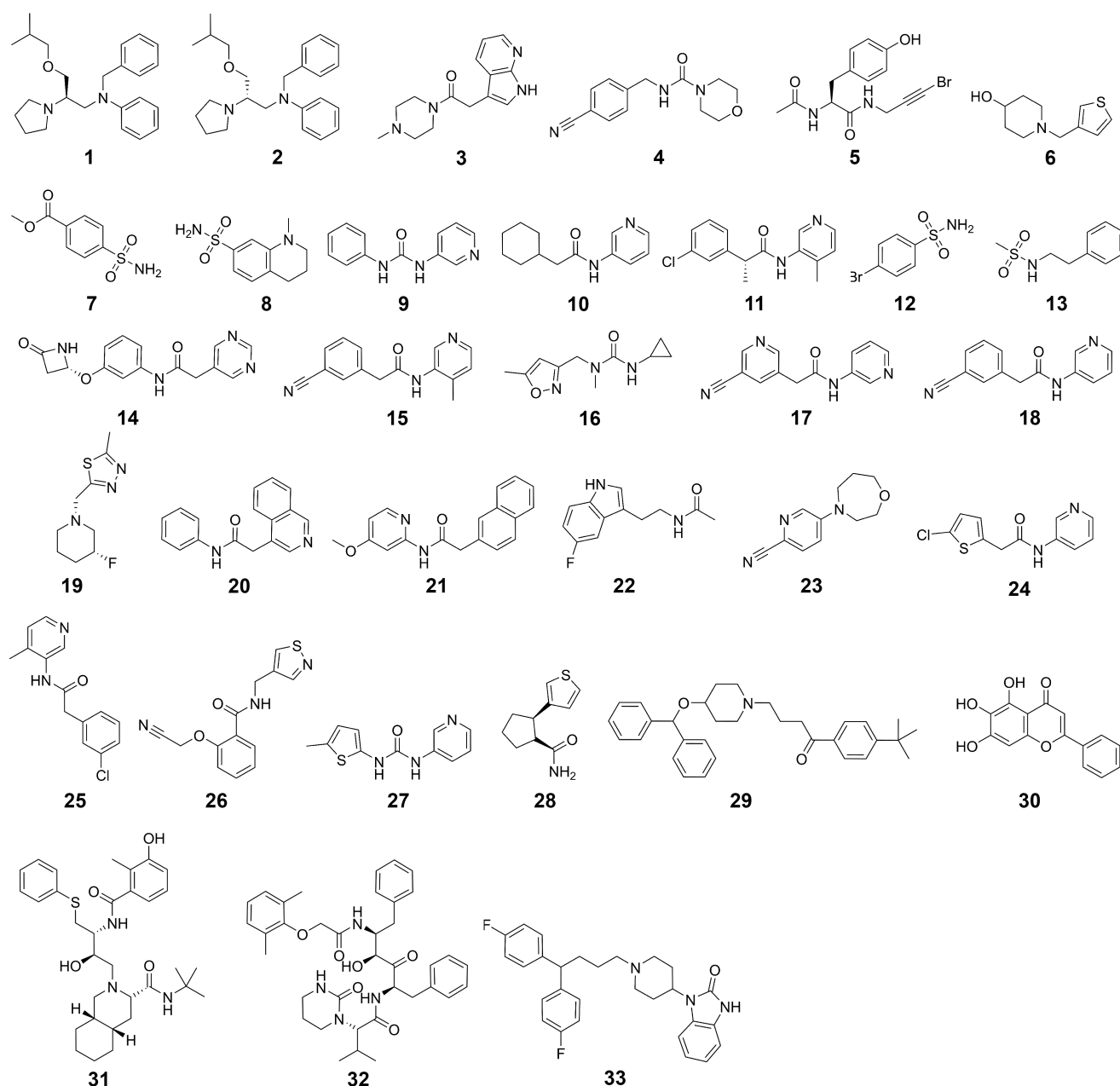


Figure 3. SARS-CoV-2 M^{pro} inhibitors considered in this study.

After the validation of the docking protocols, we performed the selectivity analysis based on docking scores. In detail, the SARS-CoV-2 M^{pro} inhibitors were docked to every selected protease and their docking scores were compared with those of the native

ligands of the respective enzyme. Again, we determined the ROC AUC metrics with the SARS-CoV-2 compounds regarded as decoys. Except for cathepsin G and chymase, the docking calculations with SARS-CoV-2 M^{Pro} inhibitors as decoys displayed higher ROC AUC values compared to the docking calculations with randomly selected decoy molecules. This indicates an overall low potential for off-target binding based on this metric. However, in addition to the ROC AUC metric, we depicted the scores in histograms for every target (Figure 4A). The docking scores of the SARS-CoV-2 compounds were predicted to be comparably high in magnitude with the majority of compounds only scored slightly better than -6.0 kcal/mol, even when docked to SARS-CoV-2 M^{Pro} itself. This is not surprising as the experimentally measured affinity for those compounds only reached micromolar IC₅₀ values. As already established by the ROC AUC values, cathepsin G presented the highest score overlap between the compound sets suggesting a risk for off-target inhibition of this protease involved in antigen processing. Other proteins with a comparatively high overlap were SARS-CoV M^{Pro}, UCHL1, and chymase. Further, some inhibitors of caspase-3 presented an overlap with the best-scoring SARS-CoV-2 M^{Pro} inhibitors, while the majority of compounds were highly separated. In SARS-CoV M^{Pro}, thrombin, and chymase several SARS-CoV-2 M^{Pro} inhibitors yielded a similar docking score as some of the actives for that target, even though the peaks of the score distribution were well separated. Especially, in the case of thrombin, concurrent binding could benefit COVID-19 patients suffering from coagulopathies such as venous thromboembolism or sepsis-induced coagulopathy [16,17]. In addition to caspase-3, the distribution in factor Xa and prostatic presented a clear separation of the peaks for each compound category indicating a low potential for concurrent binding. In order to obtain more confidence in the results from molecular docking, we used the complexes and subjected them to MD simulations followed by molecular mechanics-generalized Born surface area (MM/GBSA) post-processing. This methodology is considered to be more precise as opposed to docking scores for a multitude of biomolecular systems [34]. Even though the spread of the values was higher using this protocol, the general trends remained highly similar, especially for chymase, cathepsin G, and caspase-3 (Figure S19). There was a slightly higher overlap of the scores for factor Xa and SARS-CoV-M^{Pro}, which would indicate a higher potential for a compound to hit both targets.

Interestingly, when the actives of each target were docked to SARS-CoV-2 M^{Pro}, prostatic inhibitors yielded better docking scores compared to the native ligands (Figure S20). We noticed similar, but less pronounced trends for inhibitors of thrombin, factor Xa, chymase, and caspase-3, while cathepsin G, SARS-CoV M^{Pro}, and UCHL1 compounds presented nearly identical maxima of their docking scores compared to SARS-CoV-2 M^{Pro} inhibitors. In conclusion, the distribution of the scores indicate promiscuity toward chymase, UCHL1, and especially cathepsin G, even though the majority of SARS-CoV-2 M^{Pro} inhibitors presented inferior binding scores towards all assessed proteins. Notably, empirical scoring functions, as they were used in this project, have a known degree of inaccuracy, and thus, the absolute numbers should be regarded only as trends.

To acquire structural insights into the selectivity factors of each protease, we visualized binding modes of ligands presenting either low or high binding affinities for each target. According to these complexes, compounds intended to inhibit SARS-CoV-2 M^{Pro} should present π - π stacking with the catalytic residue H41 as well as high complementarity with the available subpockets of the active site (Figure 4B). To achieve potent and selective interaction with cathepsin G, the binding modes suggest a salt bridge to K192 or H57 as well as a deeply buried hydrophobic moiety to be optimal (Figure 4C). Similarly, ionic interactions, especially if they were buried, seemed to play a role for selective binding toward thrombin (Figure 4D), caspase-3 (Figure 5A), and chymase (Figure 5B). Interestingly, compounds hitting the presumably desired off-target factor Xa also strongly relied on shape complementarity as for SARS-CoV-2 M^{Pro} (Figure 5C), which might explain the concurrent binding to these targets, as we have previously detected in a virtual screening project [4].

Two-dimensional (2D) depictions of all discussed binding modes are presented in the Supplementary Information (Figures S21 and S22).

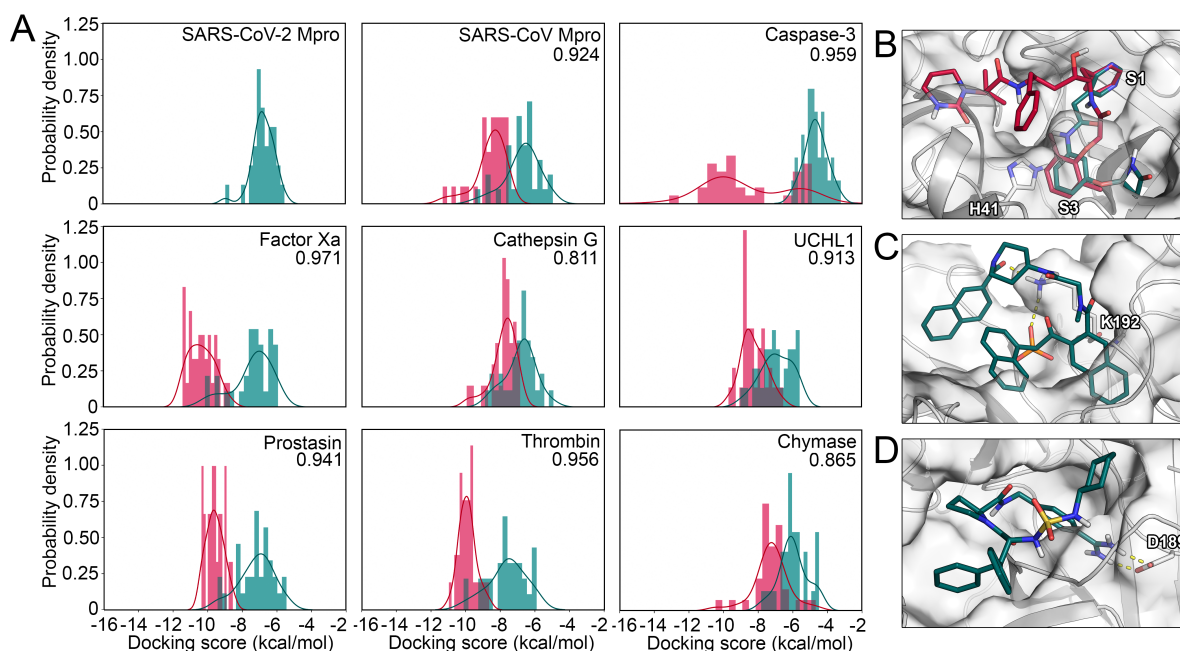


Figure 4. (A) Score distribution of SARS-CoV-2 M^{pro} inhibitors docked to the selected panel of proteases. The compounds designed against SARS-CoV-2 M^{pro} are shown in pine green, while the known actives for the remaining targets are shown in red. ROC AUC values with the SARS-CoV-2 M^{pro} inhibitors regarded as decoys for every target are shown. (B) Binding mode of compound 32 (red) and compound 14 (pine green) toward SARS-CoV-2 M^{pro}. (C) Binding mode of compound 199 toward cathepsin G. (D) Binding mode of compound 177 toward thrombin.

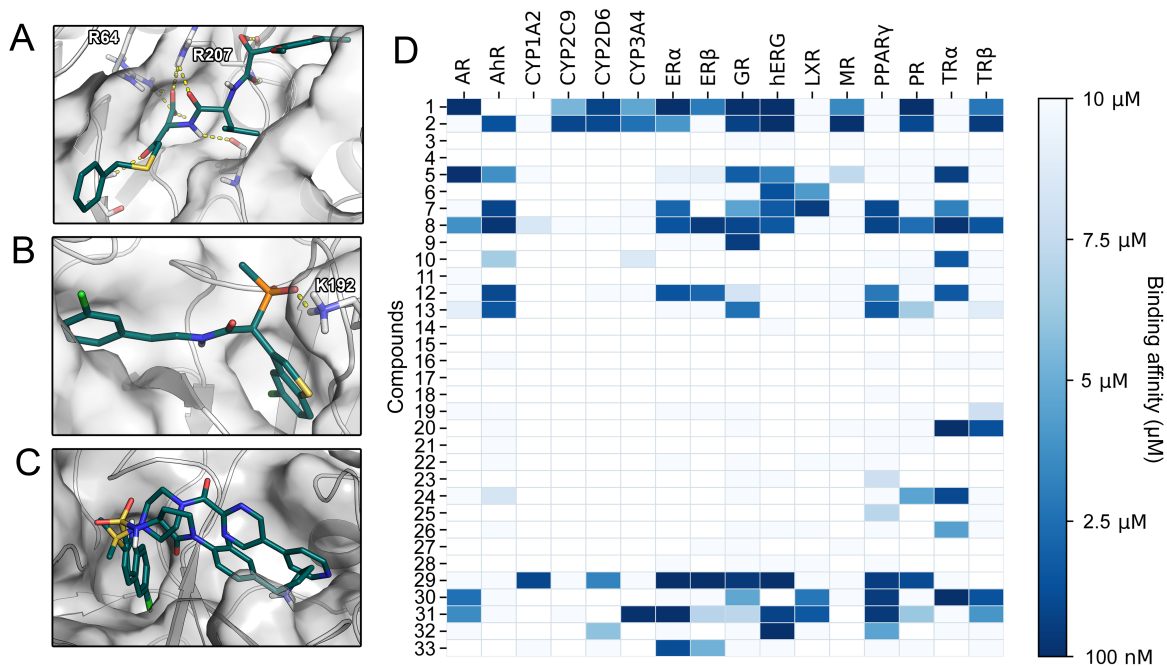


Figure 5. (A) Binding mode of compound 86 toward caspase-3. (B) Binding mode of compound 293 toward chymase. (C) Binding mode of compounds 130 and 137 toward factor Xa. (D) SARS-CoV-2 M^{pro} inhibitors examined with VTL. The predicted binding affinities of the assessed compounds 1-33 for 16 anti-targets are given.

2.4. Toxicity Profiling

As mentioned in the introduction, affinity toward anti-targets is frequently responsible for drug attrition [12]. To establish toxicologically relevant binding profiles of drug candidates our lab has developed the VirtualToxLab (VTL) evaluating their interaction with 16 anti-targets relevant for endocrine disruption, cardiac adverse effects, and extensive or undesired metabolism [13]. Besides estimates for binding affinities against the anti-targets, the VTL provides a parameter referred to as toxic potential serving as a consensus readout for potential undesired effects of the respective compound. SARS-CoV-2 M^{Pro} inhibitors with toxic potential significantly higher than 0.5 (Table S13) included compounds **1** ((*R*)-bepiridil), **2** ((*S*)-bepiridil), **29**, **31** (nelfinavir), and **32** (lopinavir). While compounds **1** and **2** were predicted to interact with multiple nuclear receptors, the hERG channel, and various cytochromes, compounds **29**, **31**, and **32** presented affinity toward a more narrow spectrum of anti-targets (Figure 5D). Compound **29**, for example, almost exclusively bound to nuclear receptors, resulting in a high estimated risk for endocrine disruption [13]. A common feature of compounds **29**, **31**, and **32** was their prediction as hERG binders. The hERG potassium channel is one of the most frequently tested anti-targets in drug development due to its involvement in fatal arrhythmias [12,13,35]. The results regarding the HIV protease inhibitors nelfinavir and lopinavir included in our study confirmed the predictive power of the VTL protocol, as in vitro experiments evidenced their hERG inhibition [35]. Thus, bepiridil (compounds **1** and **2**) displaying a strong interaction for hERG might be at risk to cause cardiac arrhythmia. A large share of the reported SARS-CoV-2 M^{Pro} inhibitors have comparably low molecular weights with 26 of 33 reported compounds below 300 g/mol (Figure S23). Since such fragment-like compounds frequently display low specificity [36], the expansion of these scaffolds might generally decrease their potential for off-target binding and at the same time improve their moderate potency.

2.5. Selectivity from Different Perspectives

We analyzed the selectivity of nine proteases from different perspectives including sequence and active site similarity, the location of hydration hot-spots and preference for certain chemical probes, as well as molecular docking and toxicological profiling. At the first sight, the low sequence similarity among the proteases (Table 1), the different cleavage sites, as well as the different volumes of the active sites may suggest a low risk of off-target binding. However, all investigated off-target proteins showed a considerable active site similarity based on 3D fingerprints and the positioning of catalytic residues (Table 2). Based on these parameters, prostatic and factor Xa were the most similar proteases compared to SARS-CoV-2 M^{Pro}. In six analyzed proteases, a hydration site indicated by water hot-spot was identified near the histidine (denoted as site A). Only for cathepsin G, caspase-3, and thrombin, we could not identify this hydration site. Thus, the displacement of this water molecule would not add any ligand selectivity in this regard. Further, for five of the nine proteases, another hydration site was identified, near the serine/cysteine residues in factor Xa, cathepsin G, prostatic, thrombin, and chymase. Regarding the observed cosolvent densities, we detected a density of acetonitrile in factor Xa, cathepsin G, UCHL1, thrombin, and chymase, but not SARS-CoV-2 M^{Pro}, where this region was explored by pyridine. Distinct placement of pharmacophores matching these differences in density could be exploited to improve inhibitor selectivity. The similarity of the overall densities in factor Xa, cathepsin G, and chymase coupled to the dissimilarity to SARS-CoV-2 M^{Pro}, indicated low potential for off-target binding toward these proteases. A common density of pyridine in the center of multiple sites including the one of SARS-CoV-2 M^{Pro}, however, showed the preference of an aromatic or hydrophobic moiety in this region. Caspase-3 and UCHL1 presented the most unique densities indicating a low potential for off-target binding of inhibitors targeting the remaining proteases.

The docking results of 33 non-covalent SARS-CoV-2 M^{Pro} suggested an overall high potential for binding to factor Xa, thrombin, and cathepsin G. On the other hand, the distribution of the docking scores indicated a low potential of the 33 compounds for binding

toward UCHL1 and caspase-3 (Figure 4A). As the individual enzymes offered different structural factors relevant for potent ligand-protein interaction, their consideration might improve the design of selective inhibitors. Regarding individual compounds, we identified four compounds which had the highest binding affinity toward more than a half of the analyzed proteases: nelfinavir (31) [4,37,38], lopinavir (32) [39–41], pimoziide (33) [42], and baicalein (30) [43] (Figure S24). Similarly, the aforementioned compounds, as well as both stereoisomers of beiperidil (1-2) were predicted to interact with a large panel of known anti-targets. Especially, interactions with the hERG potassium channel, as it was also observed in laboratory experiments, raised safety concerns for several compounds (Figure 5D). In this regard, nelfinavir (32) was not only predicted to interact with other proteases such as factor Xa, but also towards the hERG channel indicating low selectivity of this compound. Interestingly, when we focused on the compound with the highest predicted binding affinity toward a particular enzyme, we could distinguish three groups: one in which nelfinavir binds best (including SARS-CoV M^{Pro}, UCHL1, thrombin, and chymase), second in which pimoziide binds best (including prostaticin, factor Xa, and caspase-3), and third in which baicalein binds best (including SARS-CoV-2 M^{Pro} and cathepsin). A closer look into the first group of enzymes revealed that none of them share the same cleavage site, moreover both SARS-CoV and SARS-CoV-2 M^{Pro}s bind best to different compounds (nelfinavir and baicalein, respectively). As we highlighted selectivity from different perspectives, the different metrics are inherently not always consistent for a single target. To conclude, our predictions indicate the highest potential for off-target binding of SARS-CoV-2 M^{Pro} inhibitors for factor Xa, SARS-CoV M^{Pro}, and cathepsin G. Low potential was determined for prostaticin, thrombin, and to the largest degree, for caspase-3 (Table 3).

Table 3. Conclusions for selectivity of SARS-CoV-2 M^{Pro} inhibitors.

Protein	Docking	Cosolvents	Hydration	Site Similarity ^a
SARS-CoV M ^{Pro}	**	*	**	***
Caspase-3	none	none	none	*
Factor Xa	**	**	*	*
Cathepsin G	***	*	none	*
UCHL1	***	*	*	none
Prostaticin	**	none	*	**
Thrombin	**	none	*	**
Chymase	**	*	*	*

The potential for off-target binding of SARS-CoV-2 M^{Pro} inhibitors against the selected panel of proteases based on molecular docking, cosolvent MD simulations, and hydration site analysis. The potential was defined according to asterisks from none to three. ^a Consensus of binding site similarity determined with FuzCav and PIPSA.

3. Materials and Methods

3.1. Selection of Proteases

The panel of proteases for this work were selected using the VAST+ tool [44] by using the SARS-CoV-2 M^{Pro} structure (PDB ID: 6Y2E) as input. The VAST+ protocol determines similar macromolecules to a query structure by computing the superposition of three dimensional protein structures relying purely on geometric measures. The results were filtered to only match only human proteases. The next criterion was the reaction mechanism of the selected protease, to ensure a representation of both cysteine and serine proteases in our study. Finally, we examined the availability of crystal structures with cocrystallized ligands leading to the selection of proteases listed in Table 1. The preprocessing of the structures is given in the Supplementary Materials.

3.2. Similarity of Proteins and Active Sites

The sequence identity was determined using FASTA sequences derived from the UniProt database [45] (Table S14). In the case of both SARS-CoV-2 and SARS-CoV M^{Pro}, as well as factor Xa and thrombin, we truncated the sequences to cover the entry in the

respective crystal structures limiting the analysis to the catalytic unit. The sequences were aligned with the ClustalW algorithm [46] in the UGENE suite (v34.0) [47]. The sequence identity was computed based on matches of the respective protein to SARS-CoV-2 M^{PRO} in respect to the length of the sequence.

The active sites of the proteases were aligned in PyMOL using the pair_fit command. Each protease was aligned to the reference structure: the SARS-CoV-2 M^{PRO} (PDB ID: 6Y2E), fitting the protease catalytic histidine residue with the H41 of the SARS-CoV-2 M^{PRO}, protease catalytic serine or cysteine residue with the C145 of the SARS-CoV-2 M^{PRO}, and the protease aspartic acid residue with the catalytic water molecule (ID 582) of the SARS-CoV-2 M^{PRO}. The RMSD values were computed according to the superimposition of the catalytic residues in PyMOL.

To determine the similarity of the active sites of the considered off-targets to SARS-CoV-2 M^{PRO}, we used the FuzCav [25] routine. This routine computes the similarity based on fingerprints for each binding site incorporating pharmacophoric properties from the coordinates of surrounding α -carbon atoms. As input, we selected residues in 5 Å around the cocrystallized ligands in the structures.

To compare the electrostatic interaction properties of the binding cavities, we used the PIPSA [26] software. First, we preprocessed the structures using PDB2PQR tool [48], and we calculated the Adaptive Poisson-Boltzmann Solver (APBS) electrostatics potentials [49] setting the grid spacing to 0.6. Then, we calculated the similarity matrix (Hodgkin index) of the binding cavities from APBS grids. The binding pocket was set as a sphere with a radius of 12.5 Å around the geometric centre of the catalytic amino acids after superposition.

3.3. Cosolvent MD Simulations

The cosolvent MD simulations were conducted with the Mixed Solvent MD workflow of the Desmond (v2019-1) simulation engine [50] with acetonitrile, isopropanol, and pyridine as probe molecules as they are water-miscible and feature a low potential for aggregation. The concentration of the probe molecules was selected at 5% (by volume) and, if required for system setup, the water buffer parameter was increased from 12.0 to 15.0, as described in the documentation of the workflow. From the above-mentioned protein structures, monomers were retained to reduce the computational cost of the simulations. Furthermore, the ligands were removed from the structures to sample the respective binding sites. The simulations were performed with the default specifications at a temperature of 300 K and the OPLS_2005 force field in an NPT ensemble. After an equilibration of 15 ns, production runs of each probe were individually executed for 5 ns with 10 replica simulations resulting in a cumulative simulation time of 600 ns per protein.

3.4. Classical MD Simulations

The H++ server [51] was used to protonate all proteins structures listed in Table S15 using standard parameters at pH 7.4. The missing 4-amino-acids-long loop of the 1Q2W SARS-CoV M^{PRO} model was added using the corresponding loop of the 6LU7 model (from SARS-CoV-2 M^{PRO}), and its quality was confirmed by comparison with another crystal structure of SARS-CoV M^{PRO} (PDB ID: 2H2Z). Counter ions were added to neutralize the systems as shown in Table S15. Water molecules were placed using the combination of 3D-RISM [52] and the Placevent algorithm [53]. AMBER 18 LEaP [54] was used to immerse models in a truncated octahedral box with 12 Å radius of TIP3P water molecules and prepare the systems for simulation using the ff14SB force field [55]. The number of added water molecules is shown in Table S15. The PMEMD CUDA package of AMBER 18 software [54] was used to run 10 replicas of 50 ns for each system. The starting geometry for each system was kept, but the initial vectors were randomly assigned to enrich conformational sampling. The minimization procedure consisted of 2000 steps, involving 1000 steepest descent steps followed by 1000 steps of conjugate gradient energy minimization, with decreasing constraints on the protein backbone (500, 125 and 25 kcal·mol⁻¹·Å⁻²) and a conjugate gradient minimization with no constraints. Next,

the systems were gradually heated from 0 to 300 K over 20 ps using a Langevin thermostat with a collision frequency of 1.0 ps^{-1} in periodic boundary conditions with constant volume. Equilibration stage was run using the periodic boundary conditions with constant pressure for 1 ns (10 ns in the case of caspase-3 and factor Xa structures to ensure proper equilibration) with 1 fs step using Langevin dynamics with a frequency collision of 1 ps^{-1} to maintain temperature. Production stage was run for 50 ns with a 2 fs time step using Langevin dynamics with a collision frequency of 1 ps^{-1} to maintain constant temperature. Long-range electrostatic interactions were treated using the particle mesh Ewald method with a non-bonded cut-off of 10 \AA and the SHAKE algorithm. The coordinates were saved at an interval of 1 ps. The computation of the maximum available volume (MAV) is given in the Supplementary Information.

3.5. Water Molecules Tracking, Hot-Spots Identification

AQUA-DUCT 1.0 software [56] was used to track water molecules for all proteases in each simulation replica. Tracking of water molecules was conducted in two specific regions: the *Object*, which represents the cavity of a particular interest, and the *Scope* representing the whole macromolecule. The *Object* was defined as a 4 \AA sphere around the centroid of the active site residues of each protein (catalytic residues listed in Table 2), and the *Scope* was defined as the interior of a convex hull of α -carbon atoms in all structures. AQUA-DUCT was also used for identification of hot-spots, defined as the regions of the highest density of traced molecules within the protein interior. AQUA-DUCT is able to calculate hot-spots using two types of data: (i) using only the pathways of those molecules that entered the *Object* to calculate local hot-spots, and (ii) using the pathways of all molecules that entered the *Scope* region to calculate the global hot-spots. Both types of hot-spots were calculated for each of the simulation replica, and then simplified using the `hs_gsimplifier.py` script. The `hs_simplifier.py` script was used to analyze the positions of all identified hot-spots and grouped those hot-spots which were located within a radius of 3 \AA in the case of global hot-spots, and 2 \AA in the case of local hot-spots. Then it provided the information about the particular simulation replica in which the hot-spot was identified. The information was kept and the simplified hot-spots are colour-coded according to their occupation. Those which were the most common were colored dark red, and those which were rare are white.

3.6. Molecular Docking and Validation

To ensure a high accuracy in pose prediction, we determined a fitting structural ensemble for each target that was able reproduce to binding modes of a maximal portion of non-covalent cocrystallized ligands. We evaluated the Glide standard-precision (SP) [57] as well as the `smina` [58] docking protocol throughout this study. While default setting were retained for Glide including the grid generation, an exhaustiveness of 16, a cubic search space with a side length of 21 \AA , as well as a random seed of 42 was configured for `smina`. The centroids for the respective search spaces were determined by computing the mass center of the cocrystallized ligand. The RMSD between the docked pose and the respective cocrystallized ligand was computed using the `rmsd.py` script that comes with Maestro after protein structure alignment. In the case of unsatisfying pose prediction, we created structural ensembles (Tables S16–S18) by clustering representative structures of MD simulations as detailed in our previous work [4]. For each protein, the docking protocol combined with the structural ensemble correctly reproducing the highest number of cocrystallized ligands was selected to assess the potential to discriminate random decoy compounds from known binders. Actives for each target were collected from various sources including crystal structures, the literature, as well as the PubChem database [59]. The respective decoy compounds were generated using the novel DUDE-Z web server [60] with SMILES strings as input. Since LigPrep frequently generated multiple plausible protonation states and stereoisomers of the decoys, we chose the best docking score against the ensemble of each isomer. Together with the results from the known binders, the docking scores were submitted to the Screening Explorer web server [61] to determine enrichment metrics including

the maximal reachable enrichment as well as the ROC AUC. The following selectivity assessment was conducted by using SARS-CoV-2 M^{PRO} inhibitors as decoy compounds combined with known binders as actives. The ROC AUC metric was again obtained from the Screening Explorer web server and the scores were compared in a histogram computed using the Matplotlib [62] python library. To compare the absolute values of the docking scores among the different proteins, smina was used to rescore the poses obtained from Glide SP docking in the case of SARS-CoV-2 M^{PRO}, caspase-3, and chymase. Lastly, we aimed on deducing the structural factors for selective binding by visual inspection of the binding modes.

3.7. MD and MM/GBSA Post-Processing

To obtain more confidence in the results from docking, we post-processed the docking poses with the MM/GBSA protocol. Using Desmond (v2019-1), we conducted 2 ns simulations of all 576 ligand-protein complexes with different targets. We used the OPLS_2005 force field in an NPT ensemble with a temperature of 310 K maintained by the Nose-Hoover thermostat and atmospheric pressure maintained by the Martyna-Tobias-Klein barostat. The orthorhombic periodic boundary system was solvated with TIP3P water molecules. Long-range interactions were treated with the u-series algorithm [63] and short-range interactions were cut off at 9 Å, while bonds to hydrogen atoms were constrained with the M-SHAKE algorithm. The default relaxation protocol in Desmond was applied before the production phase. Atomic coordinates were deposited in an interval of 20 ps and the thermal_mmgbsa.py script that comes with Maestro (v2019-4) was applied to obtain binding free energies of the last 10 frames of the simulations, which were averaged thereafter.

4. Conclusions

Due to the current COVID-19 pandemic and the lack of specific therapeutics, many small molecules that inhibit SARS-CoV-2 M^{PRO} have been proposed. This work aims to support further development of these compounds in order to avoid safety issues due to off-target binding, which is one of the major reasons for late stage drug attrition. We addressed the concern of selectivity and off-target binding of 33 published, experimentally confirmed SARS-CoV-2 M^{PRO} inhibitors by predicting their affinity toward eight different proteases and profiling their active sites regarding hydration site and small-molecule hot-spots. Even though the selected off-target proteins presented a low global sequence identity to SARS-CoV-2 M^{PRO}, their binding sites were considerably similar. This similarity could explain the predicted affinities of the SARS-CoV-2 M^{PRO} inhibitors, which presented a considerable overlap with actives against chymase, UCHL1, and cathepsin G. Interestingly, inhibitors of prostatic displayed higher predicted binding affinities to SARS-CoV-2 M^{PRO} than its native inhibitors. Refining the SARS-CoV-2 M^{PRO} inhibitors may be necessary to achieve higher affinities toward their designated target, as well as to improve selectivity and thereby decrease off-target binding. Around one third of the investigated compounds presented medium to high potential for endocrine disruption, altered drug metabolism, or cardiac adverse events based on the prediction of binding affinities towards 16 well established anti-targets. Our work showed that, while there are many proposed SARS-CoV-2 M^{PRO} inhibitors, they generally exhibit poor selectivity and may cause pharmacological undesired effects by off-target binding. Even though the panel of proteases share a comparably low sequence identity and different substrate specificity, enzymes such as cathepsin G, factor Xa, as well as UCHL1 could be relevant off-targets for novel antivirals. If experimental testing and compound optimization efforts will be guided to achieve selectivity over the suggested anti-targets, novel antivirals could have an improved safety profile.

Supplementary Materials: The following are available online at <https://www.mdpi.com/1422-0067/22/4/2065/s1>, Figures S1 and S2: Active site similarity, Table S1: Maximal available volumes, Figures S3–S18: Structures of actives for each protease, Tables S2–S10: Origin of active compounds, Figure S19: Binding free energies obtained by MM/GBSA calculations, Figure S20: Score distribution of SARS-CoV-2 actives against every target, Figures S21 and S22: 2D depiction of discussed binding

modes, Tables S11 and S12: metrics of docking protocol validation, Table S13: Toxic potential of SARS-CoV-2 actives, Figure S23: Molecular weight of SARS-CoV-2 actives, Figure S24: Comparison of docking scores, Supporting text: Model preparation, Supporting text: MAV calculation, Table S14: UniProt identifiers for all proteases, Table S15: Ions and water molecules for conventional MD, Tables S16–S18: Crystal structures considered in docking. References [64–66] are cited in the Supplementary Materials.

Author Contributions: Conceptualization, A.F.; methodology, A.F.; formal analysis and investigation, A.F., M.S. (Manuel Sellner), M.B., K.M.; writing—original draft preparation, A.F., K.M., M.B.; writing—review and editing, A.F., M.S. (Manuel Sellner), K.M., M.B., M.A.L., A.G., M.S. (Martin Smieško); visualization, A.F., K.M., M.B.; supervision, M.A.L., A.G., M.S. (Martin Smieško); project administration, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Raw data obtained from molecular docking, VTL, MM/GBSA, co-solvent MD, and conventional MD are available in a public repository on GitHub under https://github.com/mmodbasel/SARS-CoV2_selectivity, accessed on 4 January 2021.

Acknowledgments: We gratefully acknowledge the support of NVIDIA Corporation with the donation of a Titan Xp GPU used for this research. This research was supported in part by PL-Grid Infrastructure.

Conflicts of Interest: The authors declare no conflict of interest.

References

- World Health Organisation. *Novel Coronavirus (2019-nCoV) Situation Reports*; WHO: Geneva, Switzerland, 2020.
- Kalil, A.C. Treating COVID-19—Off-Label Drug Use, Compassionate Use, and Randomized Clinical Trials During Pandemics. *JAMA* **2020**, *323*, 1897–1898. [[CrossRef](#)]
- Chen, N.; Zhou, M.; Dong, X.; Qu, J.; Gong, F.; Han, Y.; Qiu, Y.; Wang, J.; Liu, Y.; Wei, Y.; et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *Lancet* **2020**, *395*, 507–513. [[CrossRef](#)]
- Fischer, A.; Sellner, M.; Neranjan, S.; Smieško, M.; Lill, M.A. Potential inhibitors for novel coronavirus protease identified by virtual screening of 606 million compounds. *Int. J. Mol. Sci.* **2020**, *21*, 3626. [[CrossRef](#)]
- Dai, W.; Zhang, B.; Jiang, X.M.; Su, H.; Li, J.; Zhao, Y.; Xie, X.; Jin, Z.; Peng, J.; Liu, F.; et al. Structure-based design of antiviral drug candidates targeting the SARS-CoV-2 main protease. *Science* **2020**, *368*, 1331–1335. [[CrossRef](#)] [[PubMed](#)]
- Jin, Z.; Du, X.; Xu, Y.; Deng, Y.; Liu, M.; Zhao, Y.; Zhang, B.; Li, X.; Zhang, L.; Peng, C.; et al. Structure of Mpro from SARS-CoV-2 and discovery of its inhibitors. *Nature* **2020**, *582*, 289–293. [[CrossRef](#)]
- He, J.; Hu, L.; Huang, X.; Wang, C.; Zhang, Z.; Wang, Y.; Zhang, D.; Ye, W. Potential of coronavirus 3C-like protease inhibitors for the development of new anti-SARS-CoV-2 drugs: Insights from structures of protease and inhibitors. *Int. J. Antimicrob. Agents* **2020**, 106055. [[CrossRef](#)]
- Ma, C.; Sacco, M.D.; Hurst, B.; Townsend, J.A.; Hu, Y.; Szeto, T.; Zhang, X.; Tarbet, B.; Marty, M.T.; Chen, Y.; et al. Boceprevir, GC-376, and calpain inhibitors II, XII inhibit SARS-CoV-2 viral replication by targeting the viral main protease. *Cell Res.* **2020**, *30*, 678–692. [[CrossRef](#)]
- Kandeel, M.; Al-Nazawi, M. Virtual screening and repurposing of FDA approved drugs against COVID-19 main protease. *Life Sci.* **2020**, *251*, 117627. [[CrossRef](#)] [[PubMed](#)]
- Van Vleet, T.R.; Liguori, M.J.; Lynch, J.J.; Rao, M.; Warder, S. Screening Strategies and Methods for Better Off-Target Liability Prediction and Identification of Small-Molecule Pharmaceuticals. *SLAS Discov.* **2019**, *24*, 1–24. [[CrossRef](#)]
- Bowes, J.; Brown, A.J.; Hamon, J.; Jarolimek, W.; Sridhar, A.; Waldron, G.; Whitebread, S. Reducing safety-related drug attrition: The use of in vitro pharmacological profiling. *Nat. Rev. Drug Discov.* **2012**, *11*, 909–922. [[CrossRef](#)] [[PubMed](#)]
- Waring, M.J.; Arrowsmith, J.; Leach, A.R.; Leeson, P.D.; Mandrell, S.; Owen, R.M.; Pairaudeau, G.; Pennie, W.D.; Pickett, S.D.; Wang, J.; et al. An analysis of the attrition of drug candidates from four major pharmaceutical companies. *Nat. Rev. Drug Discov.* **2015**, *14*, 475–486. [[CrossRef](#)] [[PubMed](#)]
- Vedani, A.; Dobler, M.; Hu, Z.; Smieško, M. OpenVirtualToxLab-A platform for generating and exchanging in silico toxicity data. *Toxicol. Lett.* **2015**, *232*, 519–532. [[CrossRef](#)] [[PubMed](#)]
- Rao, M.S.; Gupta, R.; Liguori, M.J.; Hu, M.; Huang, X.; Mantena, S.R.; Mittelstadt, S.W.; Blomme, E.A.G.; Van Vleet, T.R. Novel Computational Approach to Predict Off-Target Interactions for Small Molecules. *Front. Big Data* **2019**, *2*, 1–17. [[CrossRef](#)]
- Xia, S.; Liu, M.; Wang, C.; Xu, W.; Lan, Q.; Feng, S.; Qi, F.; Bao, L.; Du, L.; Liu, S.; et al. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res.* **2020**, *30*, 343–355. [[CrossRef](#)]

16. Tang, N.; Bai, H.; Chen, X.; Gong, J.; Li, D.; Sun, Z. Anticoagulant treatment is associated with decreased mortality in severe coronavirus disease 2019 patients with coagulopathy. *J. Thromb. Haemost.* **2020**, 1094–1099. [[CrossRef](#)] [[PubMed](#)]
17. Connors, J.M.; Levy, J.H. COVID-19 and its implications for thrombosis and anticoagulation. *Blood* **2020**, *135*, 2033–2040. [[CrossRef](#)] [[PubMed](#)]
18. Fischer, A.; Smieško, M. Allosteric binding sites on nuclear receptors: Focus on drug efficacy and selectivity. *Int. J. Mol. Sci.* **2020**, *21*, 534. [[CrossRef](#)] [[PubMed](#)]
19. Ghanakota, P.; Carlson, H.A. Driving Structure-Based Drug Discovery through Cosolvent Molecular Dynamics. *J. Med. Chem.* **2016**, *59*, 10383–10399. [[CrossRef](#)]
20. Kimura, S.R.; Hu, H.P.; Ruvinsky, A.M.; Sherman, W.; Favia, A.D. Deciphering Cryptic Binding Sites on Proteins by Mixed-Solvent Molecular Dynamics. *J. Chem. Inf. Model.* **2017**, *57*, 1388–1401. [[CrossRef](#)] [[PubMed](#)]
21. Keretsu, S.; Bhujbal, S.P.; Joo Cho, S. Computational study of paroxetine-like inhibitors reveals new molecular insight to inhibit GRK2 with selectivity over ROCK1. *Sci. Rep.* **2019**, *9*, 1–14. [[CrossRef](#)]
22. Ferguson, F.M.; Gray, N.S. Kinase inhibitors: The road ahead. *Nat. Rev. Drug Discov.* **2018**, *17*, 353–376. [[CrossRef](#)]
23. Chaudhury, S.; Gray, J.J. Identification of Structural Mechanisms of HIV-1 Protease Specificity Using Computational Peptide Docking: Implications for Drug Resistance. *Structure* **2009**, *17*, 1636–1648. [[CrossRef](#)] [[PubMed](#)]
24. Boyd, S.E.; Garcia de la Banda, M.; Pike, R.N.; Whisstock, J.C.; Rudy, G.B. PoPS: A computational tool for modeling and predicting protease specificity. *Proc. IEEE Comp. Syst. Bioinform. Conf.* **2004**, 372–381. [[CrossRef](#)]
25. Weill, N.; Rognan, D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J. Chem. Inf. Model.* **2010**, *50*, 123–135. [[CrossRef](#)] [[PubMed](#)]
26. Blomberg, N.; Gabdoulline, R.R.; Nilges, M.; Wade, R.C. Classification of protein sequences by homology modeling and quantitative analysis of electrostatic similarity. *Protein. Struct. Funct. Genet.* **1999**, *37*, 379–387. [[CrossRef](#)]
27. Perona, J.J.; Craik, C.S. Structural basis of substrate specificity in the serine proteases. *Protein. Sci.* **1995**, *4*, 337–360. [[CrossRef](#)]
28. Tan, Y.S.; Spring, D.R.; Abell, C.; Verma, C.S. The Application of Ligand-Mapping Molecular Dynamics Simulations to the Rational Design of Peptidic Modulators of Protein-Protein Interactions. *J. Chem. Ther. Comput.* **2015**, *11*, 3199–3210. [[CrossRef](#)]
29. Yang, Y.; Hu, B.; Lill, M.A. *WATsite2.0 with PyMOL Plugin: Hydration Site Prediction and Visualization BT—Protein Function Prediction: Methods and Protocols*; Springer: New York, NY, USA, 2017; pp. 123–134. [10](#). [[CrossRef](#)]
30. 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. *Int. J. Mol. Sci.* **2020**, *21*, 3099. [[CrossRef](#)] [[PubMed](#)]
31. Ortiz, A.R.; Gomez-Puertas, P.; Leo-Macias, A.; Lopez-Romero, P.; Lopez-Viñas, E.; Morreale, A.; Murcia, M.; Wang, K. Computational approaches to model ligand selectivity in drug design. *Curr. Top. Med. Chem.* **2006**, *6*, 41–55. [[CrossRef](#)]
32. Wang, Y.; Bryant, S.H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B.A.; Thiessen, P.A.; He, S.; Zhang, J. PubChem BioAssay: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D955–D963. [[CrossRef](#)]
33. Kumar, A.; Zhang, K.Y.J. A cross docking pipeline for improving pose prediction and virtual screening performance. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 163–173. [[CrossRef](#)]
34. Wang, E.; Sun, H.; Wang, J.; Wang, Z.; Liu, H.; Zhang, J.Z.; Hou, T. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chem. Rev.* **2019**, *119*, 9478–9508. [[CrossRef](#)] [[PubMed](#)]
35. Anson, B.D.; Weaver, J.G.R.; Ackerman, M.J.; Akinsete, O.; Henry, K.; January, C.T.; Badley, A.D. Blockade of HERG channels by HIV protease inhibitors. *Lancet (Lond. UK)* **2005**, *365*, 682–686. [[CrossRef](#)]
36. Chen, Y.; Shoichet, B.K. Molecular docking and ligand specificity in fragment-based inhibitor discovery. *Nat. Chem. Biol.* **2009**, *5*, 358–364. [[CrossRef](#)] [[PubMed](#)]
37. Huynh, T.; Wang, H.; Luan, B. In Silico Exploration of the Molecular Mechanism of Clinically Oriented Drugs for Possibly Inhibiting SARS-CoV-2's Main Protease. *J. Phys. Chem. Lett.* **2020**, *11*, 4413–4420. [[CrossRef](#)]
38. Yamamoto, N.; Matsuyama, S.; Hoshino, T.; Yamamoto, N. Nelfinavir inhibits replication of severe acute respiratory syndrome coronavirus 2 in vitro. *bioRxiv* **2020**. [[CrossRef](#)]
39. Liu, X.; Wang, X.J. Potential inhibitors against 2019-nCoV coronavirus M protease from clinically approved medicines. *J. Genet. Genom.* **2020**, *47*, 119–121. [[CrossRef](#)]
40. Kumar, Y.; Singh, H.; Patel, C.N. In silico prediction of potential inhibitors for the main protease of SARS-CoV-2 using molecular docking and dynamics simulation based drug-repurposing. *J. Infect. Public Health* **2020**, *13*, 1210–1223. [[CrossRef](#)] [[PubMed](#)]
41. Narkhede, R.; Cheke, R.; Ambhore, J.; Shinde, S. The Molecular Docking Study of Potential Drug Candidates Showing Anti-COVID-19 Activity by Exploring of Therapeutic Targets of SARS-CoV-2. *Eurasian J. Med. Oncol.* **2020**, *4*, 185–195. [[CrossRef](#)]
42. Vatansever, E.; Yang, K.; Kratch, K.; Drelich, A.; Cho, C.; Mellott, D.; Xu, S.; Tseng, C.; Liu, W. Bepridil is potent against SARS-CoV-2 In Vitro. *bioRxiv* **2020**. [[CrossRef](#)]
43. Eleftheriou, P.; Amanatidou, D.; Petrou, A.; Geronikaki, A. In Silico Evaluation of the Effectivity of Approved Protease Inhibitors against the Main Protease of the Novel SARS-CoV-2 Virus. *Molecules* **2020**, *25*, 2529. [[CrossRef](#)] [[PubMed](#)]
44. Madej, T.; Lanczycki, C.J.; Zhang, D.; Thiessen, P.A.; Geer, R.C.; Marchler-Bauer, A.; Bryant, S.H. MMDB and VAST+: Tracking structural similarities between macromolecular complexes. *Nucleic Acids Res.* **2014**. [[CrossRef](#)]
45. Consortium, T.U. UniProt: A worldwide hub of protein knowledge. *Nucleic Acids Res.* **2019**, *47*, D506–D515. [[CrossRef](#)]

46. Thompson, J.D.; Higgins, D.G.; Gibson, T.J. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680. [[CrossRef](#)]
47. Okonechnikov, K.; Golosova, O.; Fursov, M.; Varlamov, A.; Vaskin, Y.; Efremov, I.; German Grehov, O.G.; Kandrov, D.; Rasputin, K.; Syabro, M.; et al. Unipro UGENE: A unified bioinformatics toolkit. *Bioinformatics* **2012**, *28*, 1166–1167. [[CrossRef](#)] [[PubMed](#)]
48. Dolinsky, T.J.; Nielsen, J.E.; McCammon, J.A.; Baker, N.A. PDB2PQR: An automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* **2004**, *32*, W665–W667. [[CrossRef](#)]
49. Baker, N.A.; Sept, D.; Joseph, S.; Holst, M.J.; McCammon, J.A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 10037–10041. [[CrossRef](#)] [[PubMed](#)]
50. Bowers, K.; Chow, E.; Xu, H.; Dror, R.; Eastwood, M.; Gregersen, B.; Klepeis, J.; Kolossvary, I.; Moraes, M.; Sacerdoti, F.; et al. Scalable Algorithms for Molecular Dynamics Simulations on Commodity Clusters. In Proceedings of the ACM/IEEE SC 2006 Conference (SC'06), Tampa, FL, USA, 11–17 November 2006; p. 43. [[CrossRef](#)]
51. Anandkrishnan, R.; Aguilar, B.; Onufriev, A.V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541. [[CrossRef](#)] [[PubMed](#)]
52. Luchko, T.; Gusarov, S.; Roe, D.R.; Simmerling, C.; Case, D.A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Ther. Comput.* **2010**, *6*, 607–624. [[CrossRef](#)]
53. Sindhikara, D.J.; Yoshida, N.; Hirata, F. Placevent: An algorithm for prediction of explicit solvent atom distribution—Application to HIV-1 protease and F-ATP synthase. *J. Comput. Chem.* **2012**, *33*, 1536–1543. [[CrossRef](#)] [[PubMed](#)]
54. Case, D.A.; Walker, R.C.; Cheatham, T.E.; Simmerling, C.; Roitberg, A.; Merz, K.M.; Luo, R.; Darden, T. *Amber 18*; University of California: San Francisco, CA, USA, 2018.
55. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Comput.* **2015**, *11*. [[CrossRef](#)] [[PubMed](#)]
56. 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*, 2599–2601. [[CrossRef](#)]
57. Halgren, T.A.; Murphy, R.B.; Friesner, R.A.; Beard, H.S.; Frye, L.L.; Pollard, W.T.; Banks, J.L. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **2004**, *47*, 1750–1759. [[CrossRef](#)] [[PubMed](#)]
58. Koes, D.R.; Baumgartner, M.P.; Camacho, C.J. Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise. *J. Chem. Inf. Model.* **2013**, *53*, 1893–1904. [[CrossRef](#)]
59. Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B.A.; Thiessen, P.A.; Yu, B.; et al. PubChem 2019 update: Improved access to chemical data. *Nucleic Acids Res.* **2019**, *47*, D1102–D1109. [[CrossRef](#)] [[PubMed](#)]
60. Mysinger, M.M.; Carchia, M.; Irwin, J.J.; Shoichet, B.K. Directory of Useful Decoys, Enhanced (DUD-E): Better Ligands and Decoys for Better Benchmarking. *J. Med. Chem.* **2012**, *55*, 6582–6594. [[CrossRef](#)] [[PubMed](#)]
61. Empereur-Mot, C.; Zagury, J.F.; Montes, M. Screening Explorer—An Interactive Tool for the Analysis of Screening Results. *J. Chem. Inf. Model.* **2016**, *56*, 2281–2286. [[CrossRef](#)]
62. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [[CrossRef](#)]
63. Shaw, D.E.; Grossman, J.P.; Bank, J.A.; Batson, B.; Butts, J.A.; Chao, J.C.; Deneroff, M.M.; Dror, R.O.; Even, A.; Fenton, C.H.; et al. Anton 2: Raising the Bar for Performance and Programmability in a Special-Purpose Molecular Dynamics Supercomputer. In Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, SC, Austin TX, USA, 15 November 2015; pp. 41–53. [[CrossRef](#)]
64. Madhavi Sastry, G.; Adzhigirey, M.; Day, T.; Annabhimoju, R.; Sherman, W. Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. *J. Comput.-Aided Mol. Des.* **2013**, *27*, 221–234. [[CrossRef](#)]
65. Schrödinger, LCC. *Maestro Small-Molecule Drug Discovery Suite 2019-3*; Schrödinger, LCC: New York, NY, USA, 2019.
66. Berman, H.M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T.N.; Weissig, H.; Shindyalov, I.N.; Bourne, P.E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242. [[CrossRef](#)] [[PubMed](#)]



Computational insights into the known inhibitors of human soluble epoxide hydrolase

Maria Bzówka^{a,b}, Karolina Mitusińska^a, Katarzyna Hopko^c, Artur Góra^{a,*}

^aTunneling Group, Biotechnology Centre, ul. Krzywoustego 8, Silesian University of Technology, Gliwice 44-100, Poland

^bDepartment of Organic Chemistry, Bioorganic Chemistry and Biotechnology, ul. Krzywoustego 4, Faculty of Chemistry, Silesian University of Technology, Gliwice 44-100, Poland

^cBiotechnology Centre, ul. Krzywoustego 8, Silesian University of Technology, Gliwice 44-100, Poland

Human soluble epoxide hydrolase (hsEH) is involved in the hydrolysis of epoxyeicosatrienoic acids (EETs), which have potent anti-inflammatory properties. Given that EET conversion generates nonbioactive molecules, inhibition of this enzyme would be beneficial. Past decades of work on hsEH inhibitors resulted in numerous potential compounds, of which a hundred hsEH–ligand complexes were crystallized and deposited in the Protein Data Bank (PDB). We analyzed all deposited hsEH–ligand complexes to gain insight into the binding of inhibitors and to provide feedback on the future drug design processes. We also reviewed computationally driven strategies that were used to propose novel hsEH inhibitors.

Keywords: Human soluble epoxide hydrolase (hsEH); hsEH inhibitors; Protein–ligand interactions; Pharmacophore; Drug design

Introduction

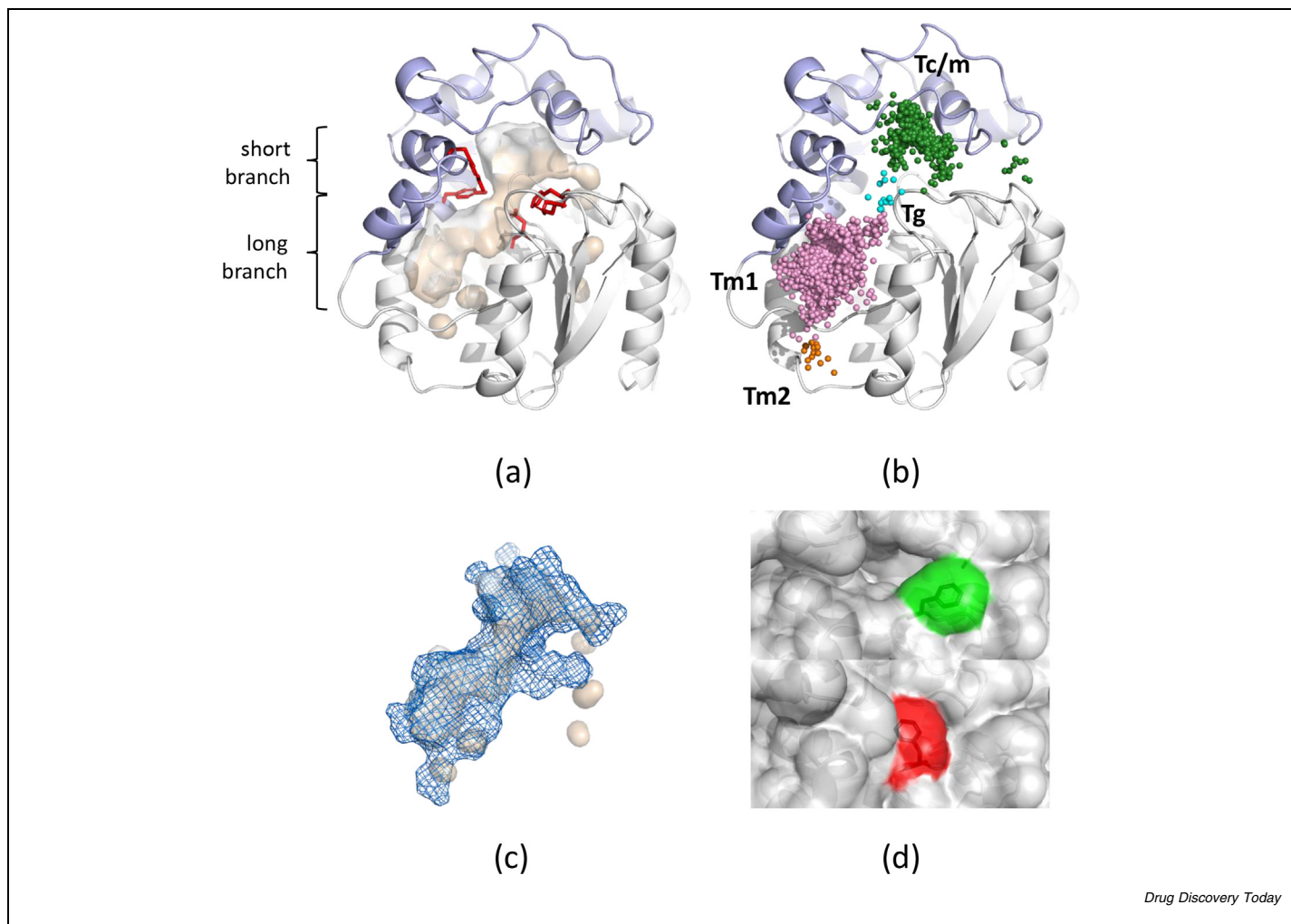
hsEH is a bifunctional homodimeric enzyme that is a member of the epoxide hydrolase family. It comprises two independently folded domains: an N-terminal domain (NTD) with phosphatase activity and a C-terminal domain (CTD) with hydrolase activity. hsEH is encoded by *EPHX2* and occurs in cytosol and peroxisomes [1]. The enzyme is distributed in most tissues, including liver, kidney, intestine, adipocytes, neurons, and blood vessels [2,3]. Its CTD hydrolyzes arachidonic acid epoxides and other natural epoxy-fatty acids, especially the transformation of regioisomers of 5,6-,8,9-,11,12-, and 14,15-EETs to the corresponding dihydroxyeicosatrienoic acids (DHETs) [4]. The conversion of EETs by hsEH generates nonbioactive molecules; thus, enzyme inhibition would be expected to enhance EET bioavailability and their beneficial properties. hsEH was proposed as a molecular target in many diseases and disorders, including cardiovascular, metabolic, renal, ocular, neurodegenerative, and psychiatric disorders, as reviewed elsewhere [5–11]. Significant efforts have

been made to develop inhibition strategies against hsEH. Although numerous potential inhibitors were proposed, none was successfully applied. In this review, we emphasize the contribution of computational approaches to hsEH studies and provide feedback from the analysis of crystal structures of hsEH with bound inhibitors underlying the potential impact of intramolecular voids, tunnels, or regulatory elements in drug development [12–14].

Architecture of the hsEH hydrolase domain

The hsEH CTD comprises two parts, the core (main) domain (M235-P369 and M469-M555) and the cap (lid) domain (S370-R468 with a flexible cap-loop F409-T443) (Fig. 1a). The so-called NC-loop (Y348-S370) connects both domains. The active site of hsEH is buried inside the protein core in an 'L'-shaped pocket. The two branches that comprise the internal pocket are 15 Å (long branch) and 10 Å (short branch) long. The entire pocket is hydrophobic. The branches are connected by a small

* Corresponding author. Góra, A. (a.gora@tunnelinggroup.pl)

**FIGURE 1**

(a) Crystal structure of the hydrolase domain of human soluble epoxide hydrolase (hsEH). The main domain is in grey and the cap domain is in blue. Internal pockets are shown and are in beige; the residues building the active site are presented as red sticks. (b) Localization of areas of water molecule entry/egress (clusters) to the active site of hsEH. Small balls represent single inlets of water molecule entry/egress, colors correspond to identified tunnels (pink, Tm1; green, Tc/m; cyan, Tg; orange, Tm2). (c) Comparison of the water-accessible volumes of hsEH identified during molecular dynamics (MD) simulation (blue mesh) with the protein internal pockets identified in hsEH crystal structure (Protein Data Bank (PDB) ID: 1s8o; beige surface). (d) The 'open' (upper panel) and 'closed' (lower panel) conformation of the F497 residue. The side-chain rotation of the F497 residue regulates access to the Tc/m tunnel.

bottleneck in which the catalytic triad and two stabilizing residues are located (D335, D496, H524, Y383, and Y466) [15–17]. The active site catalyzes the epoxide hydrolysis reaction, as summarized by Hopmann and Himo [18].

The active site is connected with the environment through tunnels. The presence of two branches suggests that the protein only has two entrances to its interior. However, tracking of water molecules during MD simulations implemented in the AQUADUCT software [19] provided information about four entrance/exit locations for molecules penetrating the hsEH (Fig. 1b). We have marked them according to previously published nomenclature [20]: Tm1, used by most (~78%) of the identified water molecules, permanently open, located in the long branch between helices $\alpha 4$ and $\alpha 10$; Tc/m, used by ~20% of water molecules, located at the border of domains, in the short branch (between loops E520-H524 and A411-S415); Tg, transient (<1%) gorge linking Tc/m with Tm1 (regulated by E494-L499 loop); and Tm2,

rarely used (<1%), separated by two loops (the N359-M369 part of the NC-loop and the S479-P488 loop), adjacent to the Tm1 tunnel. Secondary structure motifs are presented in Fig. S1 and Table S1 in the supplemental information online. The performed analysis also depicts that the internal cavity of hsEH is substantially larger than that observed in the crystal structure (Fig. 1c).

Overview of hsEH inhibitors

Origin of the co-crystallized hsEH inhibitors

Some of the earliest hsEH inhibitors were *trans*-3-phenylglycidols and chalcone oxides, although most inhibitors contained the following scaffolds: urea, amide, carbamate, thiourea, thioamide, thioester, carbonate, ester, amidine, and guanidine, as well as heterocycles, aminoheterocycles, and/or aminoheteroaryls [21]. Among countless synthesized inhibitors, only individual compounds were qualified into clinical trials. However, none has received approval for use in the clinic.

In this review, we focus on inhibitors co-crystallized with hSEH. The crystal structures of enzyme–inhibitor complexes are the final experimental verification of the predicted or observed binding affinity of the proposed compounds to hSEH. They were proposed with the use of methods such as structure–activity relationships (SARs), virtual screening (VS) docking, or fragment-based crystallography (Table S2 in the supplemental information online).

The first crystal structures deposited in the PDB [22] were proposed by Gomez et al. [15,23] based on previous quantitative SAR (QSAR) analyses [24,25]. SAR along with the lead optimization disclosed a series of potent arylamides [26], urea derivatives [27–32], pyrazoles [33], and piperidine-derived non-ureas [34]. Hiesinger et al. used a combination of selective optimization of the side activities approach (SOSA) with computer-aided target deorphanization to identify talinolol as a potent inhibitor of hSEH [35]. VS studies were also used to identify novel inhibitors [36], which were generally amide or urea derivatives. By contrast, Morisseau et al. obtained fulvestrant as potent hSEH inhibitor [37]. Pilger et al. used docking together with spin diffusion-based nuclear magnetic resonance (NMR) methods to propose urea- and indazole-based inhibitors [38]. A combination of a VS, docking, and SAR approach was used by Xing et al. to propose benzoxazole and derivatives [39], whereas Thalji et al. proposed 1-(1,3,5-triazin-yl)piperidine-4-carboxamide [40] as potential hSEH inhibitors.

The highest number of hSEH–inhibitor complexes deposited in PDB were obtained by fragment-based crystallography. Amano et al. identified numerous fragments that inhibit sEH and revealed their binding modes. They discovered novel inhibitor scaffolds, such as aminothiazole, benzimidazole derivatives, and *N*-ethylmethylamine [17,41]. Complexes obtained by Öster et al. showed the possibility of dual conformations or multiple binding sites for some compounds. The results also revealed that the ligand properties, such as potency, efficiency and, to some degree, clogP, influence the successful generation of crystal structures with bound ligands [42]. Xue et al. used a combination of fragment-based crystallography and high-throughput VS (HTVS) to map the hSEH active site pocket and identified two scaffolds, oxindoline, and 2-phenylbenzimidazole-4-sulfonamide [43]. Lately, a novelty among inhibitors was the crystallization of the first endogenous ligand, 15-deoxy- Δ 12,14-prostaglandin-J2 (15d-PGJ₂), with hSEH [44]. Details of the studies leading to the crystallographic structures of hSEH–inhibitor complexes can be found in Table S2 in the supplemental information online.

Computational efforts to propose novel hSEH pharmacophore models and inhibitors

Hammock's group reported a series of 1,3-disubstituted ureas, related amides, and carbamates as first pharmacophores. They then expanded the idea toward dicyclohexyl urea and 12-(3-adamantan-1-yl-ureido)dodecanoic acid AUDA compounds to propose a model that comprises primary, secondary, and tertiary pharmacophores (P1, 1,3-disubstituted urea, and polar groups P2, P3) connected by the hydrophobic linkers L1 and L2 [21,45]. Scientists have also adopted multiple approaches, often computationally driven, such as HTVS, SAR, and docking, to

improve the components of the above-mentioned model, or to propose new ones (Fig. S2 and Table S3 in the supplemental information online) [46–49].

Moser et al. developed a pharmacophore model based on co-crystallized inhibitors of sEH [46]. The pharmacophore comprises two donor/acceptor features (F3 and F4), and three hydrophobic regions (F1, F2, and F5). Two acceptor and two donor features are optional (F6–F9). The model was later used to find dual 5-lipoxygenase (5-LO)/sEH inhibitors [50]. This dual target was also studied by Nandha et al., who optimized the benzimidazole derivative hits by incorporating SAR studies [51]. Waltenberger et al. created a set of multiple structure- and ligand-based pharmacophore models to cover different putative binding modes of the ligands [47]. Tripathi et al. applied a 3D-QSAR approach to design a pharmacophore model containing four features: one hydrogen bond acceptor, two hydrophobic rings, and one aromatic ring [48]. Bhagwati et al. presented three ligand-based pharmacophore: ADH, ADHH, and AADHH (A, hydrogen bond acceptor; D, hydrogen bond donor; and H, hydrophobic feature) [49]. All models are presented in Fig. S2 and study details are summarized in Table S3 in the supplemental information online. All of the above-mentioned studies targeted the 'L'-shaped internal cavity, with the main focus on the active site surrounding, underlying the importance of D335, T383, and Y466 while binding. Studies also confirmed that urea- and amide-based inhibitors are the most commonly used.

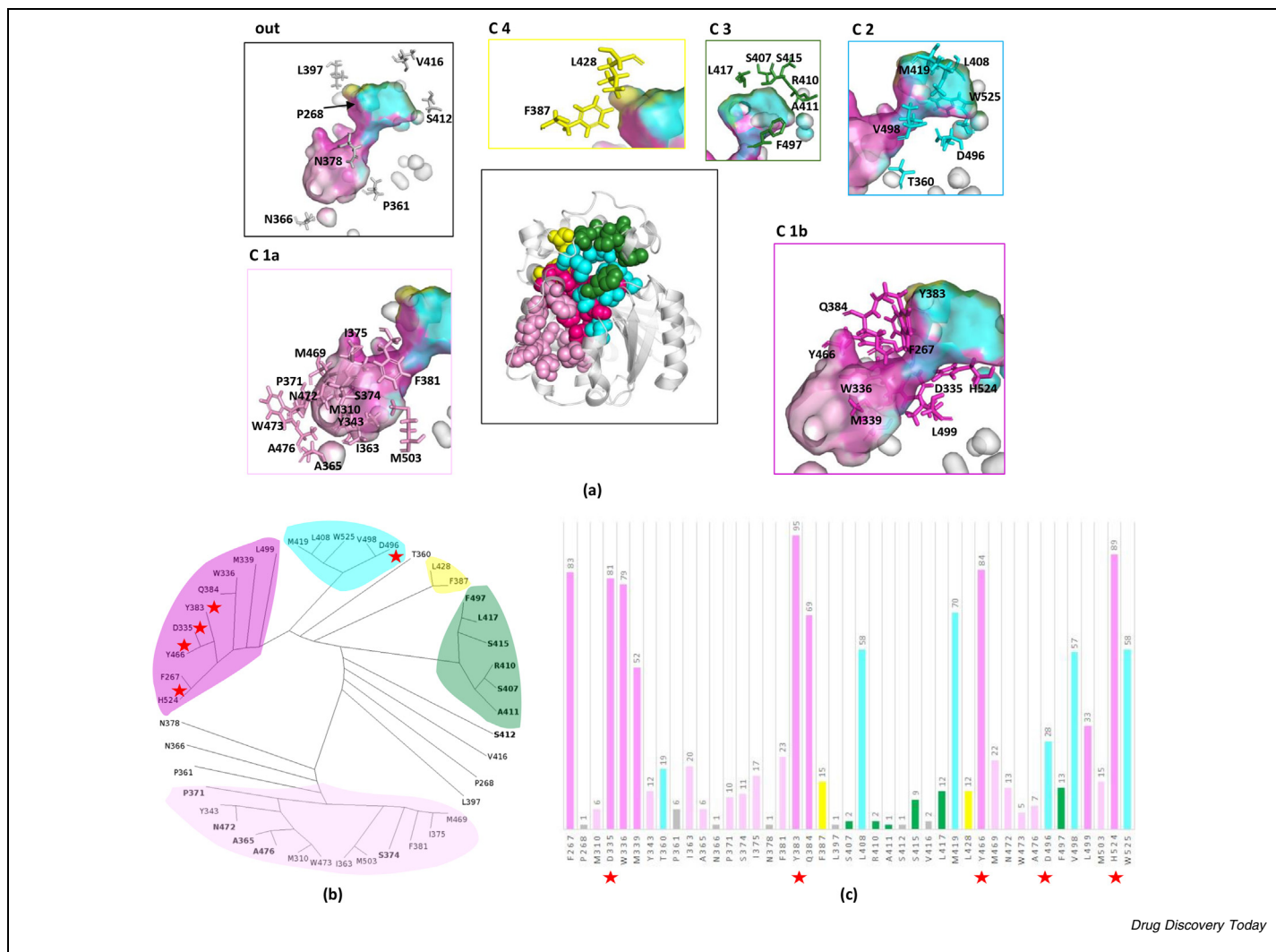
Recently, Scholz et al. created and screened a small carbonylcarboxamide compound library and identified *meta*-carbonyl derivatives as suitable 3D-pharmacophores that could extend the chemical space in drug discovery, targeting sEH [52]. Abis et al. revealed novel, dual inhibitory mechanisms, in which the hSEH can be inhibited by reversible binding of 15d-PGJ₂ in the catalytic pocket, and by covalent locking onto C423 and C522, distanced from the active site [44]. Furthermore, searching for potential inhibitors among natural products [53,54] has also impacted hSEH inhibitor development [55–59].

Analysis of known inhibitors co-crystallized with hSEH CTD

Efforts to understand hSEH function and selectivity resulted in 105 crystal structures deposited in the PDB, of which 101 are complexes containing ligands relevant to CTD inhibition. We revisited these structures using methods described in Box 1 to map the interactions between the protein and inhibitors, and to summarize targeted parts of the CTD (Table S4 in the supplemental information online).

Binding residues clustering

In total, 43 amino acids were identified as interacting with ligands. Amino acids were clustered based on the similarity of the interaction pattern between residues and inhibitors. This approach revealed five principal clusters: C-1a, C-1b, C-2, C-3, C-4, and a set of outliers (Fig. 2). Clusters C-1a (light pink) and C-1b (deep pink) cover the long branch of the 'L'-shaped pocket and most of the bottleneck, where the active site is located. They surround the main entrance to the interior of the protein, the Tm1 tunnel. Clusters C-2 (cyan), C-3 (green), and C-4 (yellow)

**FIGURE 2**

Localization of the amino acids identified for particular clusters. Amino acids are shown in two different representations: as spheres in the figure presenting the whole protein and as sticks in figures showing specific cluster locations. The internal cavities of the protein crystal structure are colored according to the colors of the amino acid clusters that occupy them. (b) Amino acids clustering results. Clusters are presented in the form of a radial cladogram and stars indicate active site residues. Surface amino acids are in bold. (c) Number of inhibitors identified by LigPlot as interacting with a particular residue. The same color scheme is provided for all figures: C-1a (light pink), C-1b (deep pink), C-2 (cyan), C-3 (green), C-4 (yellow), and a set of outliers (grey).

surround the short branch and cover the remaining part of the bottleneck. Cluster C-1a is formed by 13 amino acids, and cluster C-1b by nine residues. D335 and H524 from the catalytic triad, and both stabilizing tyrosines were identified within cluster C-1b. Residues from this cluster interact with the majority of analyzed inhibitors. Cluster C-2, formed by six amino acids, is located between the cap and main domains, close to the active site, and surrounds the Tg tunnel. Within this cluster, the remaining catalytic residue, D496, was identified. It interacts with known inhibitors less often compared with the remaining active site residues. Cluster C-3, formed by six amino acids from the cap domain, mostly from the cap-loop (R410, A411, S415, and L417), is located above the entrance to the Tc/m tunnel. F497 residue potentially works as a gate and regulates access to the active site for bulky inhibitors *via* the Tc/m tunnel (Fig. 1d). This residue was also highlighted by Amano et al. as

being involved in specific interactions [17]. Cluster C-4, formed only by two amino acids, is also located in the cap domain, closer to the hinge region that links the cap and main domains, and contributes to the structure flexibility.

All remaining amino acids are specific for interactions with individual inhibitors and, therefore, were clustered as outliers. These residues are located on: (i) the NC-loop, near the entrance to the rarely used Tm2 tunnel (N366) and facing one of the walls from the long branch (P361); (ii) the loop between domains, near the active site (P268); (iii) close to the hinge region (N378); and (iv) in the cap domain (S412, V416, and L397), of which the first two amino acids are part of the cap loop.

Inhibitor clustering

Inhibitor clustering was performed based on the interaction pattern between residues and inhibitors, enabling some structural

similarities of the inhibitors to be observed. Clustering of inhibitors indicated four major binding locations in the hSEH interior, named as C-I (red), C-II (orange), C-III (blue), and C-IV (light green) (Fig. 3, and Table S4 in the supplemental information online). Eight of the co-crystallized compounds were assigned as outliers (Fig. S3 in the supplemental information online). Interestingly, most of the outliers are compounds that were crystallized in multiple positions. Residues interacting with inhibitors from particular clusters are listed in Table S5 in the supplemental information online and inhibitors clustering results together with their chemical structures are shown in Fig. S4 in the supplemental information online.

Nine inhibitors gathered in cluster C-I (red on Fig. 3) bind to the lower part of the long branch. They are positioned close to the Tm1 tunnel entry and do not reach the active site. All compounds from this group contain at least one aromatic ring in their structure. In most of the structures, the aromatic ring is surrounded by M310, Y343, A365, N472, W473, and A476 residues.

Five inhibitors (one crystallized in two positions) grouped in cluster C-II (orange on Fig. 3) bind to the upper part of the long branch, close to the active site. All compounds contain at least two aromatic rings in their structures, and none of the structures in this group has a urea or amide moiety. All compounds interact with D335 from the catalytic triad, with W336, and (except one inhibitor) with at least one of the following S374 and I375.

Most of the inhibitors (66) are gathered in cluster C-III (blue on Fig. 3). This group is the most structurally diverse; it comprises both small inhibitors (crystallized in the bottleneck), and those with extensive structures, consisting of several rings and aliphatic linkers (occupying the whole 'L'-shaped pocket). Most of the derivatives contain disubstituted urea or a carboxamide motif in the central part of their structure, which targets at least three active site residues (D335, Y383, and Y466). Given their specific interactions, they can be divided into ten subclusters (C-IIIa–j).

The subcluster C-IIIa groups together eight rather short non-urea derivatives. At a first glance, they share a common motif of the aromatic ring with the attached halogen atom(s). However, closer inspection shows that this motif can be oriented toward different residues. Besides interacting with amino acids from the active site, compounds can interact with F267, W336, M339, Q384, and V498. Five compounds grouped in subcluster C-IIIb share a disubstituted urea motif connected (in most cases) with one aromatic ring with an attached halogen atom(s). Those compounds interact with the same residues as compounds from subcluster C-IIIa and additionally with H524 and W525 residues. The subcluster C-IIIc gathers nine inhibitors, which are generally carboxamide derivatives. Although these inhibitors interact with the same residues as subcluster C-IIIb, they form additional interactions with L408, M419, and L428 because their structures are more complex. The subcluster C-IIId also contains nine inhibitors. Most inhibitors are disubstituted urea or carboxamide derivatives (with one exception). Their interaction pattern is the most similar to subcluster C-IIIb, but they rarely interact with W525. Instead, they form an additional interaction with L499. The subcluster C-IIIE comprises 13 disubstituted or urea derivatives (with three exceptions). In most cases, the urea motif is con-

nected with an aromatic ring. In some structures, the trifluoromethyl group is attached to the aromatic ring through the other heteroatom. Although the interaction pattern is similar to the pattern of the subcluster C-IIIc, the interactions with L428 and W525 are rare. Additionally, interactions with I375 and F391 are observed. The subcluster C-IIIf contains four inhibitors, which are disubstituted urea and carboxamide derivatives with at least two aromatic rings. Although they also share a similar pattern to subcluster C-IIIc, they do not interact with L428; instead, they interact often with I363, F387, L499, and M503. The subcluster C-IIIg groups 11 rather compact compounds, most of which are amine or heteroaromatic derivatives. Only two urea and one carboxamide derivatives are found in this group. Their interaction pattern corresponds to that observed in the subcluster C-IIIc. Interactions with active site amino acids are mostly formed by heteroatoms from aromatic rings. Given the size of the compounds, interactions with M339 and L428 are not observed and the compounds rarely interact with Q384 and V498 residues. The subcluster C-IIIf contains eight inhibitors, five of which share a disubstituted urea motif, the others being amide, amine, and pyrazole derivatives. The interaction pattern in this subcluster is similar to that in subcluster C-IIIb. Compared with members of other subclusters, inhibitors from subclusters C-IIIf and C-IIIf often interact with T360. The subcluster C-IIIf comprises five compounds, most of which are pyrazole or benzimidazole derivatives. None of them has urea or carboxamide motives. Despite the residues from the active site, they interact with F267 and W525. The unique inhibitor from the subcluster C-IIIf is fulvestrant. Given its size, it has many unique interactions, but also shares a common interaction pattern with other compounds from cluster C-III.

Twenty-four inhibitors (two crystallized in more than one position) are gathered in cluster C-IV (light green on Fig. 3). They are crystallized between the main and cap domains in the hSEH structure. Almost all compounds from this cluster commonly interact with F267, Y383, F387, L417, M419, F497, V498, and H524. Given their specific interactions, the compounds can be further divided into four subclusters (C-IVa–d).

The subcluster C-IVa groups 12 compounds. Most of them comprise at least two aromatic rings. Often they contain cyclic carboxamide derivatives in their structures. One of the compounds is unique because of its two long aliphatic chains. Most of the remaining inhibitors from subclusters C-IVb–d contain two condensed or isolated aromatic rings. Often, they have heteroatoms in their structures. Compounds from subcluster C-IVa specifically interact with L408, and W525. In most cases, the heteroatoms form hydrogen bonds with D496 and F497. Compounds from subcluster C-IVb interact with L408 and Y466; however, they do not interact with Y383, F497, and W525. Compounds from subcluster C-IVc interact with L408 but do not interact with Y383, F387, L417, and D496. Inhibitors from subcluster C-IVd form bonds with Y383 and F387, but do not interact with F267, L408, and F497.

Inhibitors characterized as outliers (six structures, one crystallized in multiple positions) bind to the long branch or at the border of cap and main domains (Fig. S3 in the supplemental information online).

Box 1 Methods

Interactions analysis

In total, 101 crystal structures of hsEH–inhibitor complexes (as of December 2020) were downloaded from the PDB database. The CTD was selected for further analysis. Some inhibitors were crystallized at various positions, which resulted in an analysis of 124 protein–ligand complexes. Amino acids involved in interactions between hsEH and inhibitors were identified by LIGPLOT [64]. The first step of the analysis involved reading the 3D coordinates of the protein structures and identification of atoms belonging to inhibitors. A list of both hydrogen and nonbonded interactions between hsEH and ligands was generated. In the case of identifying the hydrogen bonds, the program computes all possible positions for hydrogen atoms (H) attached to donor atoms (D) that satisfy geometrical criteria with acceptor atoms (A) in the vicinity. The criteria were as follows: H–A distance was <2.7 Å and the D–A distance was <3.35 Å. The program also lists all possible nonbonded contacts between atoms that are less than a specified distance apart. The cutoff was set to 3.9 Å. All possible interactions between hsEH and inhibitors are summarized in Table S4 in the supplemental information online.

Clustering

The hierarchical clustering was used to group amino acids and inhibitors, based on the interaction similarity expressed in binary form (1 indicating an interaction, 0 indicating no interaction). A cluster package from the SciPy library [65] was used with the Jaccard metric to compute the distance between the points and the average method to perform the linkage.

Molecular dynamics simulation

The crystal structure of hsEH (PDB ID: 1s8o) was downloaded from the PDB database. The NTD and the co-crystallized hexaethylene glycol were manually removed from the structure. MD simulations were carried out according to the protocol described by Mitusińska et al. [20].

Tunnels identification

The AQUA-DUCT 1.0 software was used to track water molecules and to identify areas of molecule entry/egress to the active site. Molecules that entered the so-called ‘Object’, defined as a 5-Å sphere around the center of geometry of active site residues, namely D333, Y383, Y466, D496, and H524, were traced within the ‘Scope’ region, defined as the interior of a convex hull of hsEH CTD C α atoms. Points at which the molecules of interest enter or leave the Scope (so-called ‘inlets’) were clustered to provide information about the tunnel network of the protein.

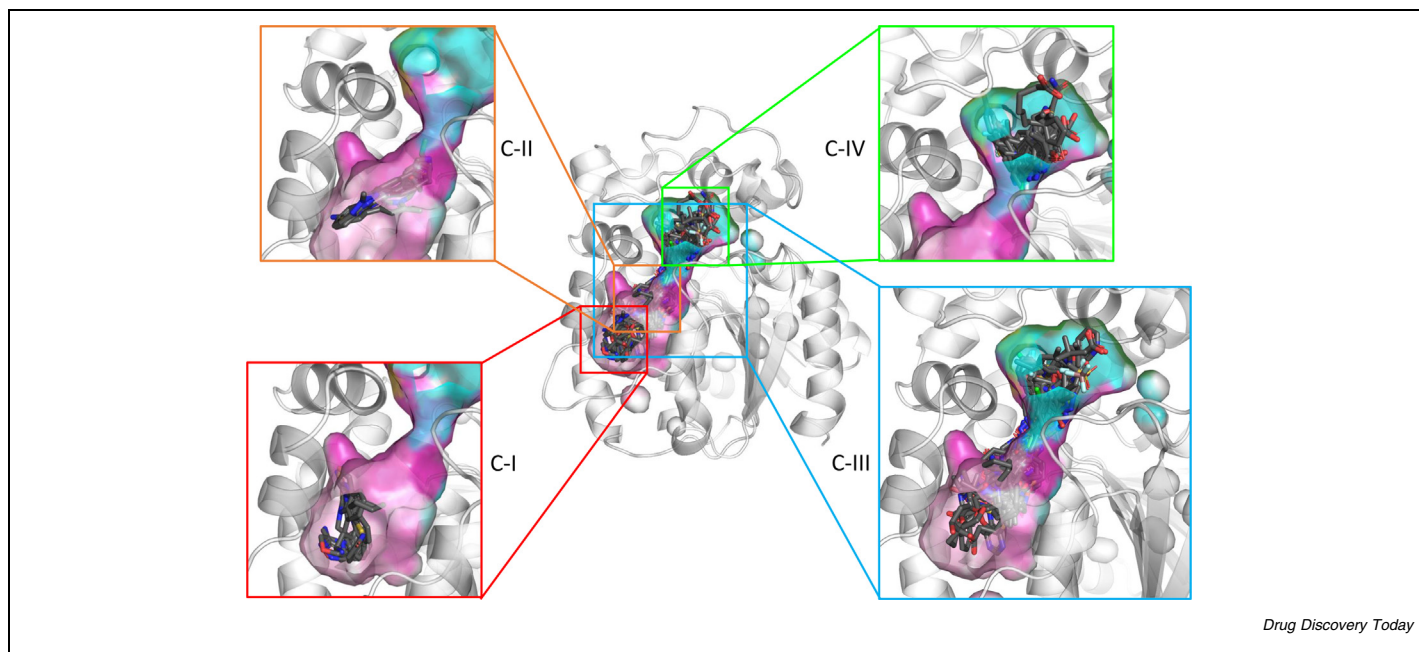
Concluding remarks and perspectives

Most of the research conducted so far highlights the importance of the side chain residues comprising the active site, D335 from the catalytic triad, and stabilizing Y383 and Y466 residues. The hydrogen-bonding network between the known inhibitors and these residues is a highly conserved feature across the reported hsEH inhibitors, as demonstrated by numerous research studies [17,36,41–43]. Despite the active site, the hydrophobic interior of the cavities, defined by F267, Y383, L408, M419, V498, and

W525 (the so-called F267 pocket), as well as the hydrophobic surface defined by W336, M339, and L499 (the so-called W336 niche) were often targeted to provide high binding affinity and selectivity of proposed inhibitors [15,16]. The hydrophobic interior means that a large part of the inhibitor structure needs to be hydrophobic, which, as a consequence, reduces their solubility. All these findings are supported by the interaction map provided in our review of the known inhibitors (Table S2 in the supplemental information online). However, close inspection of the inhibitors from the most abundant cluster C-III revealed that the aromatic moieties that interact with H524 and/or W525 were found in >75% of structures. This underlines the importance of the presence of aromatic groups in the F267 pocket, which was neglected in most of the proposed pharmacophore models (Fig. S2 in the supplemental information online).

The active site of hsEH is buried inside the core of the protein. Thus, in all analyzed crystal structures, the potential inhibitors are located inside the interior of the protein's. Although such a location of potential inhibitors adds new constraints for inhibitor design, it might also provide new opportunities. As described by Marques et al., the inhibitors can target not only the active site itself, but also the tunnels providing access to it [14]. Our analysis indicated that, in the case of hsEH, the interaction with active site residues and their surrounding is not vital for successful inhibitor design. An excellent example are members of the C-I inhibitor cluster, which bind close to the Tm1 entry and do not reach the active site residues. In addition, some of the inhibitors from the C-II and C-IV clusters are located exactly within the entry/egress areas, mostly near the Tm1 and Tg tunnels. In contrast to the inhibitors from cluster C-II, which occupy the large hydrophobic moiety, small inhibitors positioned on the border of the buried and surface-exposed residues 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 the low solubility of inhibitors. Additionally, analysis of all deposited hsEH–inhibitors complexes indicated that the inhibitors do not fully occupy the available internal pocket and that there is still some unused space at the ends of both the short and long branches of the ‘L’-shaped pocket that could host specific inhibitors with increased solubility and still guarantee high selectivity.

Surprisingly, we also identified the limited use of advanced computational approaches that could explore the potential conformational changes of the interior of the protein. Including protein dynamics during drug development would increase the conformational sampling of the protein interior and enable the design of compounds that could bind to the unexplored regions of hsEH. One of the best examples of such strategy is the design of selective inhibitors for nitric oxide synthase, where detection of the side chain rotation of a single amino acid providing access to the additional pocket resulted in a successful design of selective inhibitors [60]. Together with protein interior dynamics, it is also important to take into consideration the role of water molecules during drug design [61]. Several tools incorporating the information provided by water molecules have already been proposed [62]. Similarly, mixed-solvent MD studies could propose new pharmacophore models based on the most probable



Drug Discovery Today

FIGURE 3

Localization of inhibitors identified for the main clusters of human soluble epoxide hydrolase (hsEH): C-I (red), C-II (orange), C-III (cyan), C-IV (light green), and C-V (blue). Inhibitors are shown as sticks. The internal cavities of the protein crystal structure are colored according to the colors of the amino acid clusters that occupy them.

location of a particular solvent during MD simulations. Such a concept is widely used in drug discovery, as reviewed by Ghana-kota and Carlson [63].

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The authors' work was supported by the National Science Centre, Poland (grant no. DEC-2013/10/E/NZ1/00649).

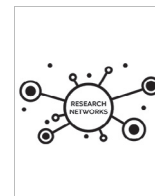
Appendix A. Supplementary material

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.drudis.2021.05.017>.

References

- [1] J.W. Newman et al., Epoxide hydrolases: their roles and interactions with lipid metabolism, *Prog. Lipid Res.* 44 (2005) 1–51.
- [2] K.M. Wagner et al., Soluble epoxide hydrolase as a therapeutic target for pain, inflammatory and neurodegenerative diseases, *Pharmacol. Ther.* 180 (2017) 62–76.
- [3] S. Norwood et al., Epoxyeicosatrienoic acids and soluble epoxide hydrolase: potential therapeutic target for inflammation and its induced carcinogenesis, *Am. J. Translat. Res.* 2 (2019) 447–457.
- [4] K. Wagner et al., The role of long chain fatty acids and their epoxide metabolites in nociceptive signaling, *Prostaglandins Other Lipid Mediat.* 113–115 (2014) 2–12.
- [5] J. He et al., Soluble epoxide hydrolase: a potential target for metabolic diseases, *J. Diabetes* 8 (2016) 305–313.
- [6] K. Hashimoto, Role of soluble epoxide hydrolase in metabolism of PUFAs in psychiatric and neurological disorders, *Front. Pharmacol.* 10 (2019) 36.
- [7] J.-Y. Liu, Inhibition of soluble epoxide hydrolase for renal health, *Front. Pharmacol.* 9 (2019) 1551.
- [8] B. Park, T.W. Corson, Soluble epoxide hydrolase inhibition for ocular diseases: vision for the future, *Front. Pharmacol.* 10 (2019) 95.
- [9] Q. Ren, Soluble epoxide hydrolase inhibitor: a novel potential therapeutic or prophylactic drug for psychiatric disorders, *Front. Pharmacol.* 10 (2019) 420.
- [10] R.D. Jones et al., Epoxy-oxylipins and soluble epoxide hydrolase metabolic pathway as targets for NSAID-induced gastroenteropathy and inflammation-associated carcinogenesis, *Front. Pharmacol.* 10 (2019) 731.
- [11] M.F. Domingues et al., Soluble epoxide hydrolase and brain cholesterol metabolism, *Front. Mol. Neurosci.* 12 (2020) 325.
- [12] A. Gora et al., Gates of enzymes, *Chem. Rev.* 113 (2013) 5871–5923.
- [13] S. Marques et al., Role of tunnels and gates in enzymatic catalysis, in: *Understanding Enzymes: Function, Design, Engineering, and Analysis*, Pan Stanford Publishing, 2016, pp. 421–463.
- [14] S.M. Marques et al., Enzyme tunnels and gates as relevant targets in drug design, *Med. Res. Rev.* 37 (2017) 1095–1139.
- [15] G.A. Gomez et al., Structure of human epoxide hydrolase reveals mechanistic inferences on bifunctional catalysis in epoxide and phosphate ester hydrolysis, *Biochemistry* 43 (2004) 4716–4723.
- [16] C. Morisseau, B.D. Hammock, Epoxide hydrolases: mechanisms, inhibitor designs, and biological roles, *Annu. Rev. Pharmacol. Toxicol.* 45 (2005) 311–333.
- [17] Y. Amano et al., Structural insights into binding of inhibitors to soluble epoxide hydrolase gained by fragment screening and X-ray crystallography, *Bioorg. Med. Chem.* 22 (2014) 2427–2434.
- [18] K.H. Hopmann, F. Himo, Theoretical study of the full reaction mechanism of human soluble epoxide hydrolase, *Chem. – Eur. J.* 12 (26) (2006) 6898–6909.
- [19] T. Magdziarz et al., AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective, *Bioinformatics* 36 (2020) 2599–2601.
- [20] K. Mitusińska et al., Exploring *Solanum tuberosum* epoxide hydrolase internal architecture by water molecules tracking, *Biomolecules* 8 (2018) 143.
- [21] H.C. Shen, B.D. Hammock, Discovery of inhibitors of soluble epoxide hydrolase: a target with multiple potential therapeutic indications, *J. Med. Chem.* 55 (2012) 1789–1808.
- [22] H.M. Berman, The Protein Data Bank, *Nucleic Acids Res.* 28 (2000) 235–242.

- [23] G.A. Gomez et al., Human soluble epoxide hydrolase: structural basis of inhibition by 4-(3-cyclohexylureido)-carboxylic acids, *Protein Sci.* 15 (2006) 58–64.
- [24] C. Morisseau et al., Structural refinement of inhibitors of urea-based soluble epoxide hydrolases, *Biochem. Pharmacol.* 63 (2002) 1599–1608.
- [25] N.R. McElroy et al., QSAR and classification of murine and human soluble epoxide hydrolase inhibition by urea-like compounds, *J. Med. Chem.* 46 (2003) 1066–1080.
- [26] A.B. Eldrup et al., Structure-based optimization of arylamides as inhibitors of soluble epoxide hydrolase, *J. Med. Chem.* 52 (2009) 5880–5895.
- [27] A.B. Eldrup et al., Optimization of piperidyl-ureas as inhibitors of soluble epoxide hydrolase, *Bioorg. Med. Chem. Lett.* 20 (2010) 571–575.
- [28] K.S.S. Lee et al., Optimized inhibitors of soluble epoxide hydrolase improve *in vitro* target residence time and *in vivo* efficacy, *J. Med. Chem.* 57 (2014) 7016–7030.
- [29] K. Takai et al., Three-dimensional rational approach to the discovery of potent substituted cyclopropyl urea soluble epoxide hydrolase inhibitors, *Bioorg. Med. Chem. Lett.* 25 (2015) 1705–1708.
- [30] S.D. Kodani et al., Identification and optimization of soluble epoxide hydrolase inhibitors with dual potency towards fatty acid amide hydrolase, *Bioorg. Med. Chem. Lett.* 28 (2018) 762–768.
- [31] A. Lukin et al., Discovery of polar spirocyclic orally bioavailable urea inhibitors of soluble epoxide hydrolase, *Bioorg. Chem.* 80 (2018) 655–667.
- [32] K. Hiesinger et al., Design, synthesis, and structure–activity relationship studies of dual inhibitors of soluble epoxide hydrolase and 5-lipoxygenase, *J. Med. Chem.* 63 (2020) 11498–11521.
- [33] H.Y. Lo et al., Substituted pyrazoles as novel sEH antagonist: investigation of key binding interactions within the catalytic domain, *Bioorg. Med. Chem. Lett.* 20 (2010) 6379–6383.
- [34] S. Pecic et al., Synthesis and structure-activity relationship of piperidine-derived non-urea soluble epoxide hydrolase inhibitors, *Bioorg. Med. Chem. Lett.* 23 (2013) 417–421.
- [35] K. Hiesinger et al., Computer-aided selective optimization of side activities of talinanol, *ACS Med. Chem. Lett.* 10 (2019) 899–903.
- [36] D. Tanaka et al., A practical use of ligand efficiency indices out of the fragment-based approach: Ligand efficiency-guided lead identification of soluble epoxide hydrolase inhibitors, *J. Med. Chem.* 54 (2011) 851–857.
- [37] C. Morisseau et al., Inhibition of soluble epoxide hydrolase by fulvestrant and sulfoxides, *Bioorg. Med. Chem. Lett.* 23 (2013) 3818–3821.
- [38] J. Pilger et al., A combination of spin diffusion methods for the determination of protein-ligand complex structural ensembles, *Angew. Chem. Int. Ed.* 54 (2015) 6511–6515.
- [39] L. Xing et al., Discovery of potent inhibitors of soluble epoxide hydrolase by combinatorial library design and structure-based virtual screening, *J. Med. Chem.* 54 (2011) 1211–1222.
- [40] R.K. Thalji et al., Discovery of 1-(1,3,5-triazin-2-yl)piperidine-4-carboxamides as inhibitors of soluble epoxide hydrolase, *Bioorg. Med. Chem. Lett.* 23 (2013) 3584–3588.
- [41] Y. Amano et al., Identification of N-ethylmethylamine as a novel scaffold for inhibitors of soluble epoxide hydrolase by crystallographic fragment screening, *Bioorg. Med. Chem.* 23 (2015) 2310–2317.
- [42] L. Öster et al., Successful generation of structural information for fragment-based drug discovery, *Drug Discovery Today* 20 (2015) 1104–1111.
- [43] Y. Xue et al., Fragment screening of soluble epoxide hydrolase for lead generation-structure-based hit evaluation and chemistry exploration, *ChemMedChem* 11 (2016) 497–508.
- [44] G. Abis et al., 15-deoxy- Δ 12,14-Prostaglandin J2 inhibits human soluble epoxide hydrolase by a dual orthosteric and allosteric mechanism, *Commun. Biol.* 2 (2019) 188.
- [45] S.-X. Huang et al., Incorporation of piperazino functionality into 1,3-disubstituted urea as the tertiary pharmacophore affording potent inhibitors of soluble epoxide hydrolase with improved pharmacokinetic properties, *J. Med. Chem.* 53 (2010) 8376–8386.
- [46] D. Moser et al., Evaluation of structure-derived pharmacophore of soluble epoxide hydrolase inhibitors by virtual screening, *Bioorg. Med. Chem. Lett.* 22 (2012) 6762–6765.
- [47] B. Waltenberger et al., Discovery of potent soluble epoxide hydrolase (seh) inhibitors by pharmacophore-based virtual screening, *J. Chem. Inf. Model.* 56 (2016) 747–762.
- [48] N. Tripathi et al., Discovery of novel soluble epoxide hydrolase inhibitors as potent vasodilators, *Sci. Rep.* 8 (2018) 14604.
- [49] S. Bhagwati, M.I. Siddiqi, Identification of potential soluble epoxide hydrolase (sEH) inhibitors by ligand-based pharmacophore model and biological evaluation, *J. Biomol. Struct. Dyn.* 38 (2019) 4956–4966.
- [50] D. Moser et al., Dual-target virtual screening by pharmacophore elucidation and molecular shape filtering, *ACS Med. Chem. Lett.* 3 (2012) 155–158.
- [51] B. Nandha et al., Synthesis of substituted fluorobenzimidazoles as inhibitors of 5-lipoxygenase and soluble epoxide hydrolase for anti-inflammatory activity, *Arch. Pharm.* 351 (2018) 1800030.
- [52] M.S. Scholz et al., Soluble epoxide hydrolase inhibitors with carboranes as non-natural 3-D pharmacophores, *Eur. J. Med. Chem.* 185 (2020) 111766.
- [53] A.L. Harvey et al., The re-emergence of natural products for drug discovery in the genomics era, *Nat. Rev. Drug Discovery* 14 (2015) 111–129.
- [54] L. Zhang et al., The strategies and techniques of drug discovery from natural products, *Pharmacol. Ther.* 216 (2020) 107686.
- [55] N.P. Thao et al., In silico investigation of cycloartane triterpene derivatives from *Cimicifuga dahurica* (Turcz.) Maxim. roots for the development of potent soluble epoxide hydrolase inhibitors, *Int. J. Biol. Macromol.* 98 (2017) 526–534.
- [56] L.B. Vinh et al., Soluble epoxide hydrolase inhibitory activity of phenolic glycosides from *Polygala tenuifolia* and in silico approach, *Med. Chem. Res.* 27 (2018) 726–734.
- [57] J.H. Kim et al., *In vitro* and in silico investigation of anthocyanin derivatives as soluble epoxide hydrolase inhibitors, *Int. J. Biol. Macromol.* 112 (2018) 961–967.
- [58] I.S. Cho et al., Inhibitory activity of quercetin 3-O-arabinofuranoside and 2-oxopomolic acid derived from *Malus domestica* on soluble epoxide hydrolase, *Molecules* 25 (2020) 4352.
- [59] A. Das Mahapatra et al., Small molecule soluble epoxide hydrolase inhibitors in multitarget and combination therapies for inflammation and cancer, *Molecules* 25 (2020) 5488.
- [60] E.D. Garcin et al., Anchored plasticity opens doors for selective inhibitor design in nitric oxide synthase, *XXXX* 4 (2008) 700–707.
- [61] S.B.A. de Beer et al., The role of water molecules in computational drug design, *Curr. Top. Med. Chem.* 10 (2010) 55–66.
- [62] K. Mitusińska et al., Applications of water molecules for analysis of macromolecule properties, *Comput. Struct. Biotechnol. J.* 18 (2020) 355–365.
- [63] P. Ghanakota, H.A. Carlson, Driving structure-based drug discovery through cosolvent molecular dynamics, *J. Med. Chem.* 59 (2016) 10383–10399.
- [64] A.C. Wallace et al., LIGPLOT : a program to generate schematic diagrams of protein-ligand interactions, *Protein Eng.* 8 (1995) 127–134.
- [65] P. Virtanen et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nat. Methods* 17 (2020) 261–272.



Structure-function relationship between soluble epoxide hydrolases structure and their tunnel network

Karolina Mitusińska, Piotr Wojsa, Maria Bzówka, Agata Raczyńska, Weronika Bagrowska, Aleksandra Samol, Patryk Kapica, Artur Góra*

Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland



ARTICLE INFO

Article history:

Received 21 August 2021
Received in revised form 21 October 2021
Accepted 23 October 2021
Available online 13 December 2021

Keywords:

Soluble epoxide hydrolases
Structure–function relationship
Tunnel network
Protein engineering

ABSTRACT

Enzymes with buried active sites maintain their catalytic function *via* a single tunnel or tunnel network. In this study we analyzed the functionality of soluble epoxide hydrolases (sEHs) tunnel network, by comparing the overall enzyme structure with the tunnel's shape and size. sEHs were divided into three groups based on their structure and the tunnel usage. The obtained results were compared with known substrate preferences of the studied enzymes, as well as reported in our other work evolutionary analyses data. The tunnel network architecture corresponded well with the evolutionary lineage of the source organism and large differences between enzymes were observed from long fragments insertions. This strategy can be used during protein re-engineering process for large changes introduction, whereas tunnel modification can be applied for fine-tuning of enzyme.

© 2021 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

1. Introduction

Enzymes are proteins that facilitate catalytic reactions in their active site. In most known enzymes, the active site is buried inside their structure [1], and connected with the environment by tunnels. These tunnels enable and regulate not only substrate entrance and product release, but also ensure specific conditions within the active site for the reaction to occur. For example, cytochrome CYP3A4 structure is equipped with a tunnel called an aqueduct, which is used only for water molecules transport [2]. The aqueduct is regulated by R375 side chain, which is able to rotate and switch between different conformations, thus allowing a fine control of the presence of water molecules in the active site cavity [3]. Therefore, tunnels may have a regulatory mechanism, known as molecular gates that can form *via* one or more amino acids, often bulky

or charged residues, able to rotate their side chain, hence, controlling access to the active site [4]. Despite controlling transport of substrates, products, and additional solvent molecules or even ions, gates in tunnels are also capable of controlling and synchronizing the reaction, or protecting the active site from poisoning [5,6].

Since tunnels are involved in the functioning of enzymes with buried active sites, it may seem that they are evolutionary conserved structural features, which has been supported by reports [7–9]. It is known that the most conserved structural feature of an enzyme is its active site [10]. Additionally, the amino acids forming the protein core are also more conserved [11,12] and tend to evolve slower than the surface residues, excluding those involved in protein–protein and/or protein–ligand interactions [13,14]. However, in our other study [15] we elucidate that in the case of the soluble epoxide hydrolases (sEHs) most of their tunnels should be considered as variable structural features with only one exception - the tunnel located at the border between the main and cap domains. These counterintuitive findings has inspired the investigation of the structure–function relationship of sEHs in more detail.

Epoxide hydrolases (EHs) have been subjects of several structural and genome analyses. Heikinheimo *et al.* [16] provided four requirements, to distinguish epoxide hydrolases from other α/β -hydrolase fold members. A structure is an α/β -hydrolase fold mem-

Abbreviations: sEHs, soluble epoxide hydrolases; msEH, *Mus musculus* soluble epoxide hydrolase; hsEH, *Homo sapiens* soluble epoxide hydrolase; StEH1, *Solanum tuberosum* soluble epoxide hydrolase; VrEH2, *Vigna radiata* soluble epoxide hydrolase; TrEH, *Trichoderma reesei* soluble epoxide hydrolase; bmEH, *Bacillus megaterium* soluble epoxide hydrolase; CH65-EH, soluble epoxide hydrolase from an unknown source, sampled in hot springs in China; Sibe-EH, soluble epoxide hydrolase from an unknown source, sampled in hot springs in Russia.

* Corresponding author at: Biotechnology Centre, Krzywoustego 8, 44-100 Gliwice, Poland.

E-mail address: a.gora@tunnelinggroup.pl (A. Góra).

<https://doi.org/10.1016/j.csbj.2021.10.042>

2001-0370/© 2021 The Authors. Published by Elsevier B.V. on behalf of Research Network of Computational and Structural Biotechnology. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

ber, when it fulfils at least two of them and maintains the sequence order of the catalytic triad. The four epoxide hydrolases features are as follows: i) the sequence order of the catalytic triad is nucleophile-acid-histidine, with the nucleophile on the canonical strand $\beta 5$; ii) the “catalytic elbow” on top of the $\beta 5$ strand with a sequence pattern that is often Gly-X-Nuc-X-Gly; iii) the structure starts from strand $\beta 3$ and is at least five strands long, including the cross-over connection at the nucleophile (strands 43567); iv) a long loop at the end of $\beta 7$ strand allows the side chains of the triad to form a hydrogen bond. Barth *et al.* [17] systematically compared known EHs based on their sequences, structures and biochemical properties. They identified three conserved and three variable regions mixed together within the protein’s sequence: i) the highly variable *N*-terminal region, which is absent in plant and most bacterial EHs, while in mammalian and insect microsomal EHs this region act as a membrane anchor; ii) the conserved first half of the α/β -hydrolase core domain; iii) the variable NC-loop, which starts directly after the $\beta 6$ strand and ends before the first cap domain helix, linking the *N*-terminal part of the core domain with the cap domain; iv) the conserved mostly helical cap domain; v) a variable cap-loop inserted between helix $\alpha 3$ and $\alpha 4$ of the cap domain, and vi) the conserved C-terminal half of the core domain consisting of two β -strands and two α -helices. A comprehensive genome analysis by van Loo *et al.* [18] supports the work by Barth *et al.* [17]. They screened various genomic databases for EHs of the α/β -hydrolase family and divided them into 8 groups from a phylogenetic tree. Thus, identifying the following: i) sequences with proteobacterial origin and proteins with *N*-terminal signal peptides related to association with membranes (group 1); ii) sequences of bacterial, archaeal and eukaryotic origins, and even from multicellular organisms that have *N*-terminal extensions of unknown function (group 2); iii) sequence of mostly putative EHs from actinobacteria, β -proteobacteria and fungi (group 3); iv) sequence of both EHs and haloalkane dehalogenases (group 4); v) sequences of mammalian, bacterial and fungal microsomal EHs and the insect juvenile hormone EHs (group 5); vi) sequence of both fluoroacetate dehalogenases and EHs with the charge-relay aspartate located at the loop after $\beta 6$ strand position (group 6); vii) sequences of EHs similar to those from group 6 with the first conserved ring-opening tyrosine and charge-relay aspartate located at different positions (group 7), and viii) sequence of a large number of known plant and mammalian EHs, including the mammalian sEHs (group 8). Unfortunately, structures of certain group members remain unknown.

Herein, we analyzed the available crystal structures of sEHs and performed a detailed analysis of their tunnel network. We focused on functional tunnels, i.e., tunnels in which we identified pathways of water molecules leading to/from the active site. We used water molecules as a molecular probe that enabled the investigation of the protein intramolecular voids and provided insights into the protein internal architecture. Thus, we were able to describe the structural basis of the tunnel network of sEHs. The available information on the sEHs substrate preferences were analyzed and combined with the data of the shape and size of their tunnel network. This study provides insight into the relationship between the structure of enzyme, usage of tunnels, and substrate preferences.

2. Materials and methods

2.1. Structure selection and preparation for analysis

Eight crystal structures of sEHs were downloaded from the Protein Data Bank (PDB) database [19]. The selected structure represent different clades of animals (*Mus musculus* (msEH, PDB ID: 1cqz [20]) and *Homo sapiens* (hsEH, PDB ID: 1s8o [21])), plants

(*Solanum tuberosum* (StEH1, PDB ID: 2cjp [22]) and *Vigna radiata* (VrEH2, PDB ID: 5xm6 [23])), fungi (*Trichoderma reesei* (TrEH, PDB ID: 5uro [24])), and bacteria (*Bacillus megaterium* (bmEH, PDB ID: 4nzz [25])), as well as two thermophilic enzymes collected from hot springs in Russia (Sibe-EH, PDB ID: 5ng7 [26]) and China (CH65-EH, PDB ID: 5nfq [26]) from an unknown source. Incomplete structures with missing structural information regarding the position of amino acids were discarded (structures were collected in December 2019). Such structures may introduce bias into the results of the water molecules flow analysis. Additional ligands and ions were manually removed, as well as the *N*-terminal phosphatase domain of msEH and hsEH.

2.2. Multiple structure alignment

The selected sEHs structures, comprising only of the selected EH domains, were submitted to the mTM-align webserver [27]. The structural alignment was carried out using default parameters. The obtained alignment was viewed and processed by SeaView [28].

2.3. Molecular dynamics simulations

The H++ server [29] was used to protonate the analyzed structures using standard parameters at reported optimal pH for the enzyme activity (Supplementary Table 1). Additionally, counterions were added in the structures. Water molecules were placed using the combination of 3D-RISM theory [30] and Placevent algorithm [31]. Water molecules were added to fill the internal cavities and pockets of the proteins’ structures [32]. The Amber14 tLEaP package [33] was used to immerse the models in a truncated octahedral box with 10 Å radius of TIP3P water molecules and the ff14SB force field [34] was used for the parametrization of each system. PMEMD CUDA package of AMBER 14 software was used to run a set of 50 ns MD simulations of selected EHs. To improve conformation sampling, the starting geometry for each system was kept but the initial vectors were randomly assigned. The minimization procedure consisted of 2000 steps, involving 1000 steepest descent steps followed by 1000 steps of conjugate gradient energy minimization, with decreasing constraints on the protein backbone (500, 125, and 25 kcal \times mol⁻¹ \times Å⁻²) and a final minimization with no constraints of conjugate gradient energy minimization. Next, gradual heating was performed from 0 K to 300 K over 20 ps using a Langevin thermostat with a collision frequency of 1.0 ps⁻¹ in periodic boundary conditions with constant volume. Equilibration stage was conducted using the periodic boundary conditions with constant pressure for the time stated in Supplementary Table 1 with 1 fs time step using Langevin dynamics with a frequency collision of 1 ps⁻¹ to maintain temperature. Production stage was conducted for 50 ns with a 2 fs time step using Langevin dynamics with a collision frequency of 1 ps⁻¹ to maintain constant temperature. Long-range electrostatic interactions were modelled using the particle mesh Ewald method with a non-bonded cut-off of 10 Å and SHAKE algorithm. The coordinates were saved at 1 ps intervals. The number of added water molecules and ions is shown in Supplementary Table 1.

2.4. Water path analysis

AQUA-DUCT software [35] version 1.0 was used to trace paths of all water molecules that were found within a defined distance from the center of masses of atoms in the catalytic center (as listed in Supplementary Table 2). The *Scope* was defined as an interior of the convex hull of C-alpha atoms of the protein. Each water molecule path was cut to fit to the protein surface (auto_barber set to protein). All inlets were then clustered using the barber method

with cutting sphere correction to the van der Waals radius of the closest atom (auto_barber_tovdw set to True).

3. Results

For the purpose of water molecules flow analysis we selected only crystal structures, which were unique and complete i.e. no information was missing about the position of a particular residue. In the case of repeated structures in PDB database, the apo structure and/or the one with the best resolution was selected. It was considered that incomplete or low resolution structures may introduce bias into the results. Eight sEHs structures were chosen that represent the clades of animals (*Mus musculus* (msEH) and *Homo sapiens* (hsEH)), plants (*Solanum tuberosum* (StEH1) and *Vigna radiata* (VrEH2)), fungi (*Trichoderma reesei* (TrEH)), and bacteria (*Bacillus megaterium* (bmEH)), as well as two thermophilic enzymes collected from hot springs in Russia (Sibe-EH) and China (CH65-EH) from an unknown source (structures were collected in December 2019). The obtained sEHs structures were compared using mTM-align web server [27] for multiple protein structure alignment (MSTA) analysis (Fig. 1). Then, the functional tunnels were identified using AQUA-DUCT [35] to compare the transport pathways. The functional tunnels were defined as those that display in which pathways of water molecules leading to/from the active site were identified. In this study we examined if using only crystal structures of the available sEHs and the information about the usage of tunnels, could the obtained results of such analyses be convergent with those of the evolutionary studies based on multiple sequences of EHs. Finally, we combined all data to investigate and evaluate the structural basis of the tunnel network of sEHs.

3.1. Structure comparison

To perform the structural comparison, we analyzed the structural features of sEHs using quantitative and qualitative descriptors (length, location/position). Nomenclature from the work of Barth *et al.* [17] was employed. The lengths of structural compartments were determined, such as the active site, cap and main domains, cap-loop and NC-loop, as well as some additional compartments (Supplementary Table 3, Fig. 1, and Fig. 2). All analyzed structures consisted of the main and cap domains, which were characteristic of the α/β -hydrolase fold. The additional N-terminal phosphatase domain - which is a known mammalian sEHs feature - was excluded from the structural analysis. The length of the analyzed structures varied from 284 (bmEH) to 333 amino acids (TrEH). The number of main domain amino acids varied from 197 (Sibe-EH) to 233 (TrEH), while the number of cap domain amino acids was correlated with the length of the cap-loop. The shortest cap domain and cap-loop were found in bmEH, with 70 and 4 amino acids, respectively, and the longest in msEH structure, with 100 and 35 amino acids, respectively. The length of NC-loop connecting the cap and main domains was similar in almost all analyzed structures (22 amino acids), except for the thermophilic enzymes (Sibe-EH with 13, and CH65-EH with 16 amino acids) and bmEH (15 amino acids). To provide a precise description of the structural differences between sEHs another loop connecting the cap domain with the main domain, which is referred to as the back-loop in this study, was distinguished. The back-loop was defined as a loop between the α D helix of the cap domain and β 7 strand of the main domain together with β 7 strand (dark blue on Figs. 1 and 2). Length of the back-loop varied the most between two thermophilic sEHs: the back-loop of Sibe-EH consists of 15, while in the case of CH65-EH it consisted of 33 amino acids.

MSTA of selected sEHs structures provided by the mTM-align web server were analyzed (Fig. 2 and Supplementary Fig. 1). Thus, regions of higher structural similarity were identified as those that display greater differences. The regions of high sequence similarity corresponded mostly to the main domain. The part of the main domain that connects with the cap domain by the NC-loop showed high structural similarity. The NC-loop region also showed high similarity, however, it is clearly the NC-loop of the group IIb enzymes differed from the other analyzed sEHs structures. Hence, the cap domain region was less similar with two exceptions (alignment positions 190–206, and 288–320, Fig. 2) that corresponded to the α -helical regions between the NC-loop and cap-loop, and between the cap-loop and back-loop. These α -helices formed two layers of the cap domain (Fig. 2). The cap-loop and back-loop regions displayed high structural differences. Additionally, the main domain region connected with the cap domain by the back-loop displayed higher structural similarity.

The β -barrel shape of the EHs main domain was observed in all analyzed structures, as well as the location of the cap domain. The active site was located in the buried cavity between the cap and main domains. In order to determine the structural factors that characterize for specific sEHs, the sequences and structures were aligned using mTM-align webserver. This enabled separation of structures into three groups (Fig. 3). Surprisingly, both mammalian sEHs were grouped with fungal TrEH (group I), while the second group consisted of plant StEH1 and VrEH2 (group IIa), and the third was bacterial bmEH and thermophilic EHs (group IIb). Structures were grouped together that displayed some common unique features. Enzymes from group I had relatively long cap-loops. The back-loop of TrEH was longer than that of msEH and hsEH, and slightly shifted away from the main domain. The α -helix located after the back-loop, α E helix, was parallel to the adjacent β -strands, β 7, and β 8 (secondary structure was derived from the work of Barth *et al.* [17], Supplementary Table 3 and Supplementary Table 4). Enzymes from group IIa had relatively long cap-loops which were positioned closer to the α D and α E helices compared to group I enzymes. Additionally, in contrast to group I enzymes, their α E helix rotated towards the α D helix region adjacent the NC-loop. Enzymes from group IIb had relatively short cap-loops, and the longest back-loops. Interestingly, the part of the back-loop closest to the cap domain was unfolded, whereas enzymes from groups I and IIa formed an α -helix. The α D helix was close to the α E helix. The α E helix had similar orientation as in group I enzymes' structure. The main domain structure was similar in shape in all analyzed structures.

3.2. Water molecules transport analysis

The main aim of our study was to describe the structural basis of sEHs tunnel network with focus on access to the active site and the implication of structural differences on the dynamics of the analyzed proteins. In order to conduct such analysis, five repetitions of molecular dynamics (MD) simulations were conducted (total of 250 ns per system) to mimic the conformational changes that occur in physiological conditions to a protein in solution. Then, the potential transport pathways were investigated, using water molecules as a molecular probe. During explicit solvent MD simulations, the protein was immersed in solvent (such as water) molecules that were able to penetrate the protein's interior. Given the number of identified water molecules which entered the protein's active site, provides the rate of exchange between the enzyme's interior and environment. Analysis of these movements provided detailed information on the tunnel network as well as their usage, while maintaining the required simulation time at a relatively low level. In our previous report, we showed that 50 ns of a *Solanum tuberosum* sEHs was enough to sample rare events of water mole-

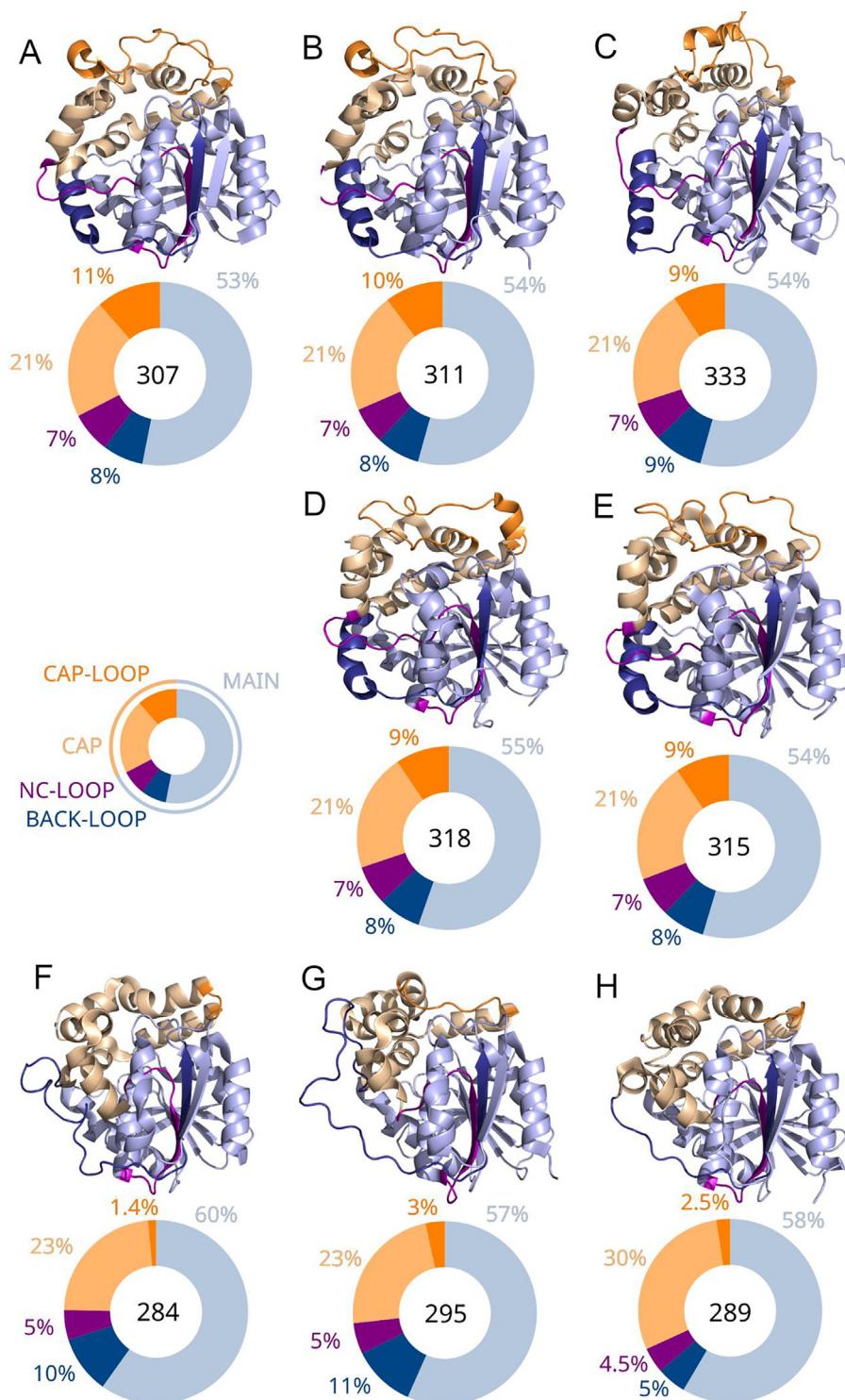


Fig. 1. Crystal structures of selected soluble epoxide hydrolases (sEHs) and a pie chart representing the size of particular compartments. A) *Mus musculus* sEH (msEH), B) *Homo sapiens* sEH (hsEH), C) *Trichoderma reesei* sEH (TrEH), D) *Solanum tuberosum* sEH (StEH1), E) *Vigna radiata* sEH (VrEH2), F) *Bacillus megaterium* sEH (bmEH), and thermophilic G) CH65-EH, and H) Sibe-EH from an unknown organism. The data herein represents the data shown in Supplementary Table 3. The proteins are shown as cartoons. Pie charts under each protein structure show the share of particular compartments in the overall structure, while the number inside the pie chart stands for the number of amino acids comprising of the whole soluble epoxide hydrolase structure.

cules entering a particular tunnel [37]. The water molecules transport analysis was facilitated by the AQUA-DUCT software and was used exclusively to trace pathways of water molecules that entered the active site of analyzed EHs. The utilization of such workflow allowed the observation of changes to the protein's structure, which, for example, opens or closes a particular pathway leading

to the active site. The obtained results gave insight into such pathways in three regions of sEHs structure - the main domain, cap domain, and border between these domains. In order to further clarify these results, the observed pathways were associated with tunnels. Therefore, the identified tunnels leading to/from the active site were marked, according to their localization; Tm - main

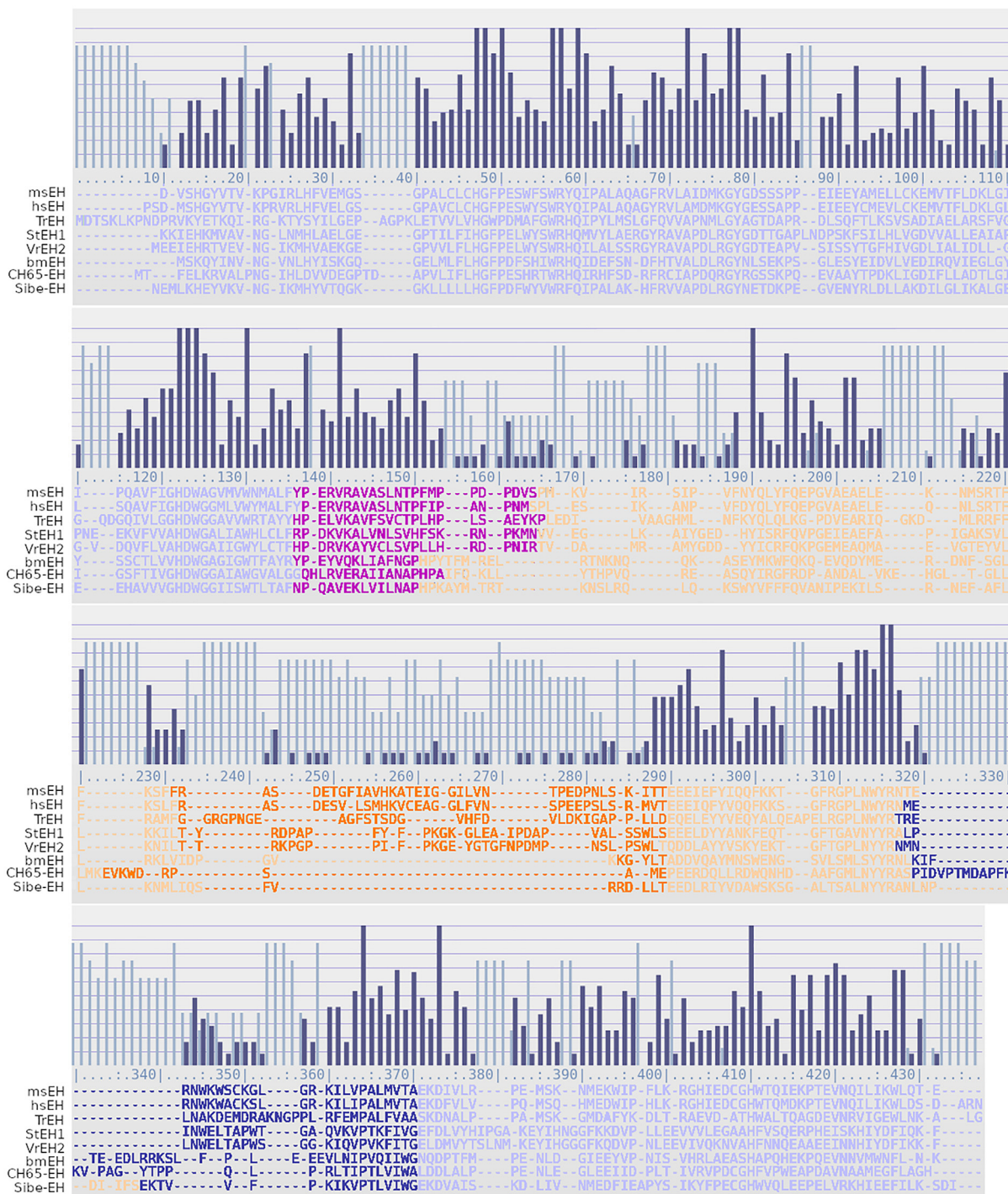


Fig. 2. Multiple Protein Structures Alignment (MSTA) of selected soluble epoxide hydrolases (sEHs). The proteins' sequences are color-coded (cap-loop – dark orange, cap – orange, NC-loop – violet, back-loop – dark blue, main domain – lilac). The dark blue bars indicate regions of higher structural similarity, while the light blue bars indicate regions of lower structural similarity. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

domain, Tcap – cap domain, and Tc/m – border between those domains (Fig. 4). The regions in which water molecules entered and/or left the protein interior are shown as small balls (so-called inlets) in Fig. 4. The inlets were then clustered to represent tunnels entries reported elsewhere [15]. It should be noted that not all tunnels were represented in other structures.

Comparison of all analyzed structures revealed only two tunnels that were predominantly utilized by water molecules (above 30% of all identified inlets, Supplementary Table 5): Tc/m tunnel

located at the border between the main and cap domain, and Tm1 tunnel located in the main domain. Moreover, we identified other tunnels, such as Tm2, Tm3, Tm4, Tm5, Tg, and Tside in the main domain, and another tunnel located between these two domains, namely Tc/m_side, and Tcap1, Tcap2, and Tcap4 in the cap domain; those tunnels were, however, rarely used by water molecules. Information regarding the predominant tunnel allowed the determination of three different patterns of tunnel usage of sEHs: i) both Tc/m and Tm1 tunnels are predominantly used

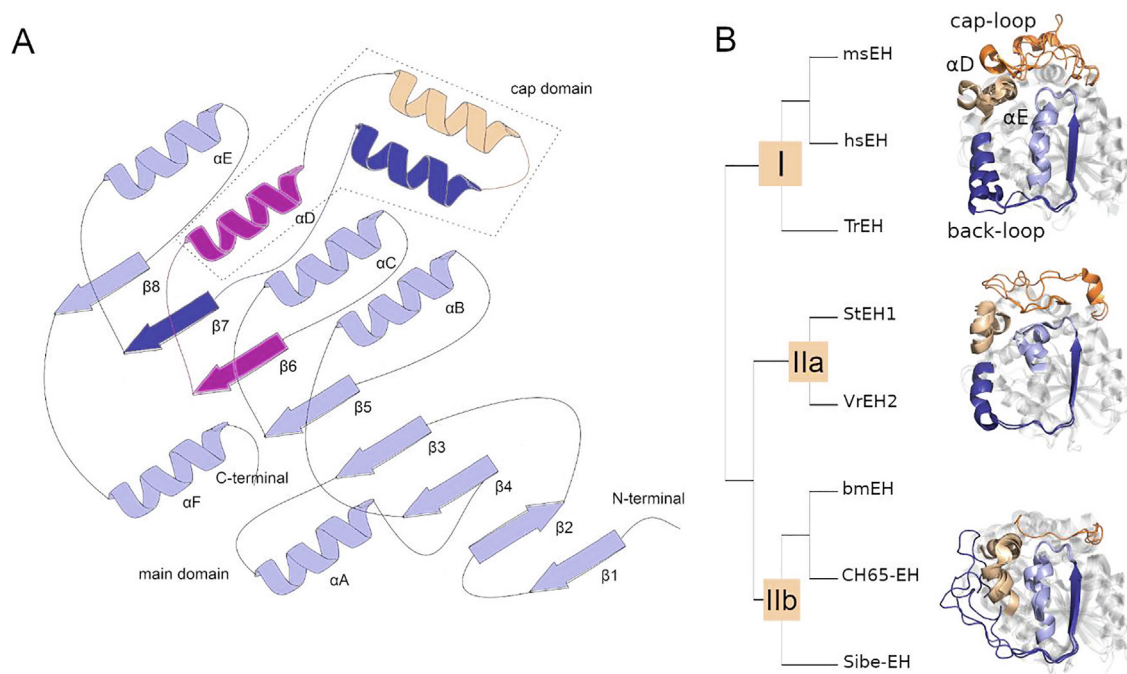


Fig. 3. Structural similarity analysis of soluble epoxide hydrolases (sEHs). A) Schematic representation of sEHs structure. The nomenclature used was in accordance with [36]. B) Cladogram of analyzed sEHs structures. The proteins' structures are shown as cartoons and for clarity only the most unique regions are shown.

(msEH, hsEH, and TrEH); ii) only Tm1 tunnel was predominantly used (StEH1, and VrEH2), and iii) only Tc/m tunnel was predominantly used (bmEH, CH65-EH, and Sibe-EH) (Supplementary Table 5, Fig. 4). Interestingly, the observed patterns corresponded to structural analysis described above. In mammalian and fungal sEHs two tunnels were predominantly used by water molecules – Tm1 and Tc/m. In contrast to TrEH, mammalian sEHs additionally utilized Tg and Tm3 tunnels, as well as Tcap1. In the case of plant sEHs, Tm1 tunnel was utilized by 92% of water molecules entering the active sites. Additionally, plant EHs employed several other tunnels, such as Tc/m, Tm2, and Tm5. Finally, in both thermophilic enzymes, CH65-EH, and Sibe-EH, Tc/m tunnel was used by the vast majority (78% and 98%, respectively) of water molecules entering the active site cavity. Both enzymes also used additional tunnels located in the main domain and the border between the main and cap domains. Similarly, bmEH utilizes mostly Tc/m tunnel; however, in a single MD simulation several water molecules employed other tunnels. VrEH2 enzyme was the only sEH unable to utilize Tc/m tunnel, instead the Tc/m_{side} tunnel was used. The number of inlets per simulation nanosecond was examined to determine the water molecules' flux. bmEH, which utilizes only one tunnel, was found as the most 'open' structure (96 inlets/ns), while VrEH2 was the most 'closed' structure (only 7 inlets/ns). The inlets/ns values were similar for hsEH, msEH, StEH1, and TrEH (40–45 inlets/ns) (Supplementary Table 5). Therefore, the number of functional tunnels does not reflect the water molecules flow through the enzyme's active site.

To complement the small-molecules transport analysis, we investigated the most flexible regions of the selected sEHs (according to RMSF data from MD simulations). The obtained results showed that both mammalian sEHs most flexible regions were α D and α E helices, and the cap-loop region, whereas TrEH only the α D helix and part of the cap-loop were identified (Fig. 5). Moreover, the accumulated movements of these regions were weaker than in case of mammalian sEHs. Similarly, the cap-loop and α D helix were the most flexible regions in plant sEHs, with little movement observed in the α E helix. Finally, in the case of bacterial and

thermophilic sEHs, the most flexible regions were the back-loop, cap-loop, and part of the α E helix. Mammalian and thermophilic sEHs showed the greatest overall flexibility.

4. Discussion

To date, little is known about EHs, however, several studies have been conducted to investigate their structural features [15–18]. sEHs belong to the α/β -hydrolases family that display a modular structure with a central catalytic domain, the main domain, formed by eight superhelically twisted β -strands [17,36,38]. This superfamily can tolerate large insertions to the scaffold without losing their catalytic activity [39]. The most important modification of the fold is the insertion after the β 6 strand, which forms the cap domain. This domain has great impact on substrate recognition and catalysis [40–42]. The cap-loop covers the active site, and thus, limits the pathways of substrate and products transport to a specific tunnel. Tunnel locations can, therefore, be constructed as a natural consequence of the active site positioning between both domains. Tunnels have been identified passing through the main and cap domains as well as the interdomain space. The cap domain is connected with the main domain by two flexible loops acting as hinges, namely NC-loop and back-loop. The NC-loop is considered to participate in substrate binding by defining the binding pocket and regulating the access to the active site [43]. Therefore, the tunnel network of sEHs may be regulated through a set of structural features: i) intramolecular voids in the main domain, ii) intramolecular voids in the cap domain, iii) hinge loops connecting both domains. Since the NC- and back-loops act as hinges they can regulate the tunnel network either by positioning the cap domain on the main domain, while affecting the entrances/exits of Barth *et al.* and Bauer *et al.* regarding the modular structure of EHs [17,38]. Analysis of the MSTA suggested that sEHs consisted of several modules (compartments), including the main and cap domains, as well as the NC-loop, cap-loop, and back-loop.

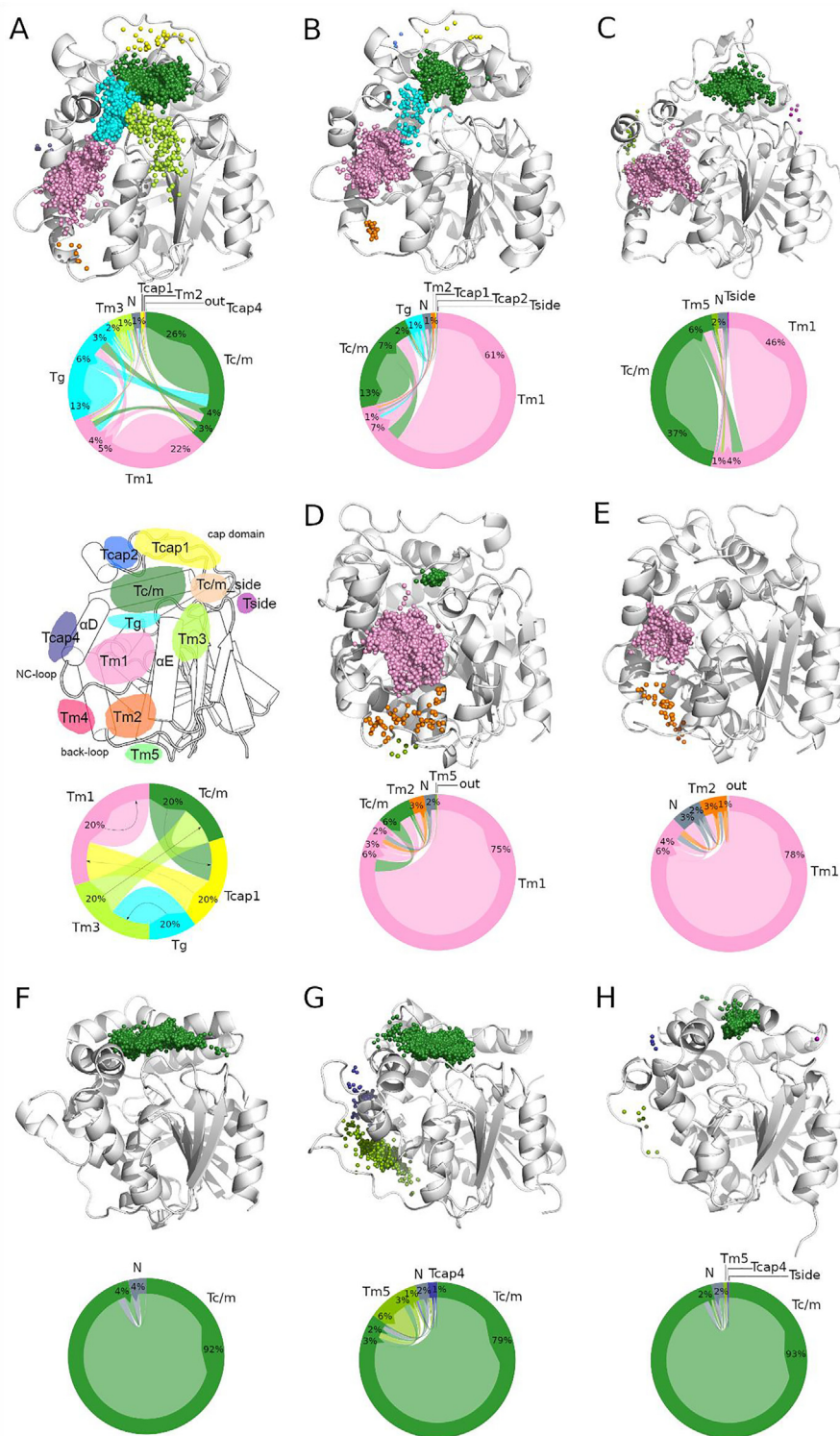


Fig. 4. Identified entries/exits of selected epoxide hydrolases (EHs) and the intramolecular flow plot. The intramolecular flow plot (also known as the migration flow plot) depicts the flow of water molecules through particular tunnels. The outer ring represents the size of the tunnel, while the size of the inner part of the plot (called here flow) represents a particular transport pathway by direction (shown in the legend by small arrows). In the sample plot five flows are shown: Tc/m to Tcap1, Tcap1 to Tm1, Tg to Tm3, Tm3 to Tc/m, Tm3 to Tc/m, and Tm1 to Tm1. The 'out' flow stands for the pathways that do not belong to specific clusters, while 'N' flow stands for the pathways that started and/or ended within the protein structure. The figure represents data presented in [Supplementary Table 6](#). A) *Mus musculus* EH (msEH), B) *Homo sapiens* EH (hsEH), C) *Trichoderma reesei* EH (TrEH), D) *Solanum tuberosum* EH (StEH1), E) *Vigna radiata* EH (VrEH2), F) *Bacillus megaterium* EH (bmEH), and thermophilic G) CH65-EH, and H) Sibe-EH from an unknown organism. The proteins are shown as cartoons, and the entries/exits are marked as small balls (so-called inlets). For picture clarity only the epoxide hydrolase domain of msEH and hsEH structures are shown.

Moreover, it was found that the main domain regions and mostly helical region of the cap domain displayed a high level of structural similarity, whereas the NC-loop, cap-loop, and the back-

loop regions display dissimilarity (Fig. 2). These findings also suggest that the most dissimilar regions are more prone to modifications.

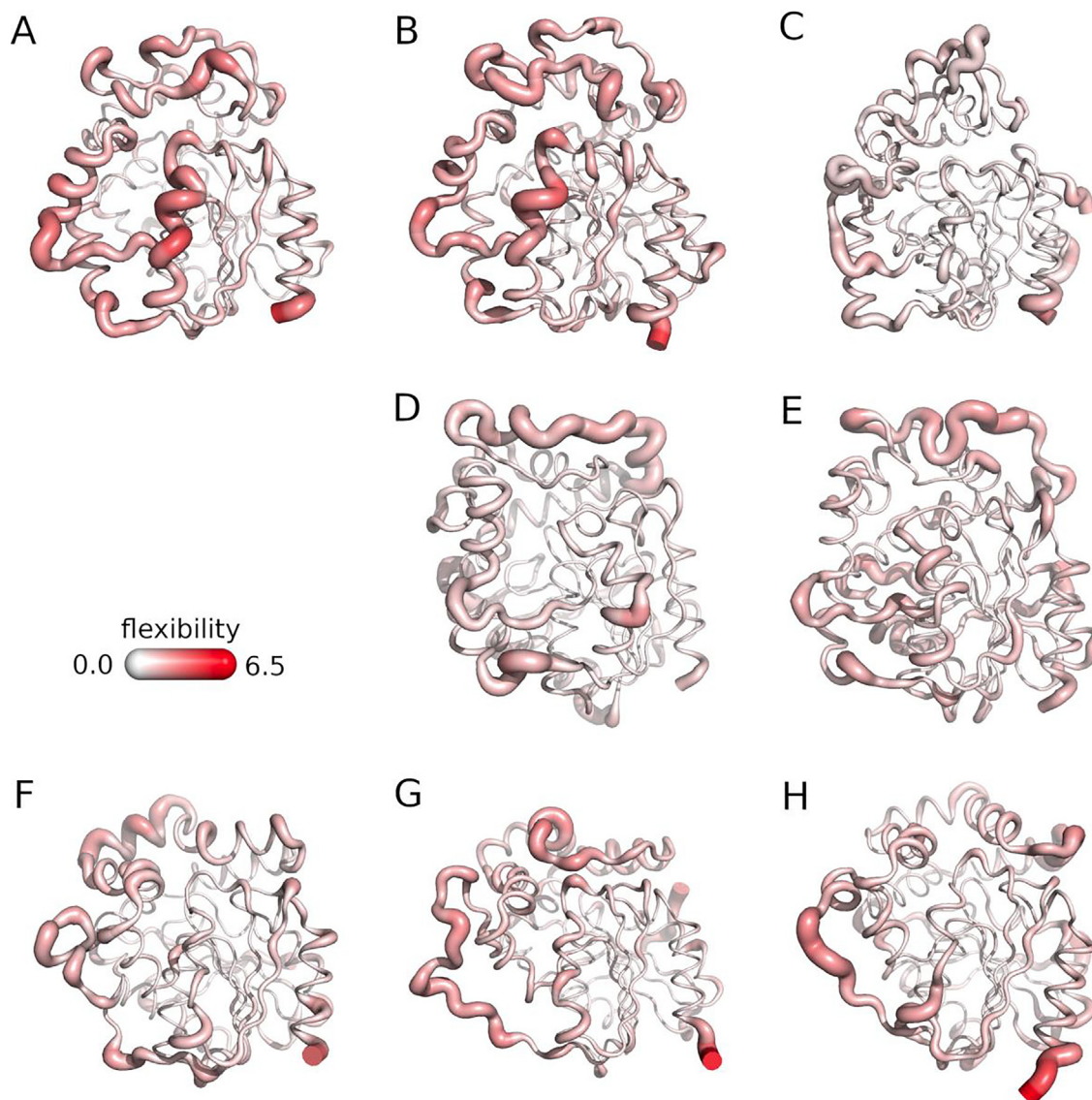


Fig. 5. The overall flexibility of the selected soluble epoxide hydrolases. A) *Mus musculus* EH (msEH), B) *Homo sapiens* EH (hsEH), C) *Trichoderma reesei* EH (TrEH), D) *Solanum tuberosum* EH (StEH1), E) *Vigna radiata* EH (VrEH2), F) *Bacillus megaterium* EH (bmEH), and thermophilic G) CH65-EH, and H) Sibe-EH from an unknown source organism. The proteins are shown as springs, with the thin white lines indicating lower flexibility, and thicker reddish lines - higher flexibility.

In our other study [15], sEHs were employed as a sample system in order to investigate the evolution of tunnels. It was determined that most tunnels should be considered as variable structural features of proteins. Tc/m tunnel was found to be the only exception, located between the cap and main domains. We proposed that insertion of the cap domain defined the buried active site cavity and the tunnel linking it with the environment. Such structural arrangement was preserved in most of the EHs which supports the hypothesis regarding the origin of the positioning of the active site between both domains. However, according to other reports [15,21,37,44,45], this was not the only pathway leading to the active site. Other tunnels were located in the cap and main domains, as well as between those domains. It should be noted that when predominant tunnels were used for substrate and/or product transport, the rarely used tunnels should not be neglected because they could be used, for example, for water molecules transport during the hydrolysis step.

In this study, we focused on the functionality of the sEHs tunnel network. Functional tunnels were defined as those which were used by water molecules to reach the active site cavity. A relation-

ship was found between the protein structure and the shape and size of its tunnel network. Hence, despite overall structural similarity, the sEHs structures were divided into three groups, based on their structure and tunnel usage (Figs. 3 and 4). The results of the structural compartments and tunnels usage analyses suggested a close evolutionary relationship of the proteins, which were considered unprecedented. Notably, the obtained results were based on a relatively low number of structures that represented different clades. Nonetheless, the sampling of the EHs family is only fragmentary. However, our results were in good agreement with that of Barth *et al.* [17] based not only on tunnel usage, but on the analysis of multiple sequences of EHs. Moreover, the presented data were in-line with the theory that animals and fungi were more closely related than animals and plants [46].

Mammalian (hsEH and msEH) and fungal (TrEH) structures were assigned to group I. Members of this group shared common features such as relatively long back-loop and cap-loop. Enzymes in this group primarily utilize two main tunnels - Tc/m, and Tm1. In all sEHs from the group I, Tc/m tunnel was found conserved [15]. This was also the case for Tm1 tunnel, but only in the case

of msEH [15]. The results of the structure flexibility analysis (Fig. 5) showed significant differences between sEHs that represent mammalian and fungal families. Mammalian sEHs were more flexible, and the regions with high RMSF values surround Tc/m, Tg, and Tm1 tunnels' entries/exits regions. Furthermore, during MD simulations we observed that those regions merged and created a long gorge. In contrast, these regions were quite rigid for TrEH structure, and consequently both tunnels clearly separated. Such differences had substantial implications on the substrate preferences, which will be discussed below.

Plant sEHs (StEH1 and VrEH2) were assigned to group IIa. The interaction between the cap and main domains was much tighter, thus Tc/m tunnel was narrower relative to other analyzed sEHs. A subtle rearrangement of the α D helix region adjacent to the NC-loop caused narrowing of the Tc/m tunnel's mouth and dramatically limited the tunnel usage (StEH1) or closed it permanently (VrEH2). Moreover, access to the active site through the cap domain was also nearly completely blocked. Similar to mammalian and fungal sEHs, plant sEHs structures had relatively long cap-loop and back-loop, however, the enzymes predominantly utilize the Tm1 tunnel, which was identified as a variable feature in StEH1 structure [15]. The flexibility analysis results of plant sEHs showed that the most flexible regions were distant to the tunnel entries and, therefore, the conformational changes were only limited to slight effect on for catalytic efficiency (if any).

Finally, bacterial (bmEH), and thermophilic enzymes from an unknown organism (CH65-EH and Sibe-EH) were assigned to group IIb. Members of this group had relatively short cap-loops and longest back-loops. They mainly utilized the Tc/m tunnel. Their α D helix was close to the α E helix, which caused narrowing of the tunnel mouths located in the main domain, namely Tm1, Tg, and Tm2. As a result the location of other tunnels on the other side of the back-loop, namely Tcap4 and Tm5, they could be opened. However, since the vast majority of water molecules were transported via the Tc/m tunnel, the role of the other tunnels for substrates/products transportation can be neglected. This observation supported the hypothetical origin of sEHs via insertion resulting in active site positioning between cap and main domains. Surprisingly, in the case of IIb group enzymes the Tc/m tunnel was found to be a variable feature [15]. This could be due a small number of residues lining the walls of the tunnel, which was significantly shorter in comparison to Tc/m tunnels in other sEHs. It was shown that the cap-loop, back-loop, and part of the α E helix were the most flexible regions. The results suggested that the movement of the back-loop may cause opening of the neighboring tunnels, such as Tcap4 and Tm5, which were commonly used by the enzymes.

Barth et al. [17] found a correlation between the length of the NC-loop and cap-loop and the type of catalyzed substrates and evolutionary lineage of the source organism. Their results indicated that sEHs of eukaryotes had long cap-loops and medium-sized NC-loops, as well as being more active towards aliphatic epoxides, while participating in fatty acid metabolism. It must be noted that the function of an enzyme and often its name reflects substrate preference based on a very limited data set. Therefore, the functional names given to enzymes early in their investigation can bias a whole field. Additionally, the overall knowledge on the functions and applications of EHs in humans and other organisms is also limited. Known compounds synthesized by and/or tested on all analyzed sEHs are shown at Supplementary Figs. 2–8. Mammalian EHs are involved in the xenobiotic metabolism and the degradation of endogenously derived epoxy fatty acids [47,48], as well as hydrolysis of *trans*-epoxy alcohol 1 (skin-related allylic epoxide) to RSR triol-3, which is the most abundant triol isomer in human and porcine epidermis [49]. Plant EHs are involved in the biosynthesis of essential aliphatic cuticular compounds [50],

detoxification of epoxy fatty acids in seeds [51], and conversion of the epoxides that accumulate during stress into less reactive compounds [52]. EHs in plants are also involved in the defence system, where their activity can be enhanced by water deprivation, wounding or during virus infection [53–55]. For example, *NtEH-1* gene encoding an EHs product of the *Nicotiana tabacum* L. is induced in the presence of the tobacco mosaic virus (TMV) [54,56]. Also, not all plants metabolize the epoxy fatty acids in their seeds. Large amounts of fatty acids in the form of triacylglycerols are used as sources of energy and biosynthetic intermediates [57]. Epoxy fatty acids are common storage lipids in seeds of certain species, such as *Astraceae* which store about 70% of their lipids in this form [57,58]. This may be the origin of the observed structural rearrangement discussed previously allowing mammalian EHs to transform long-chained epoxy fatty acids. Summerer et al. [59] highlighted a substantial difference in the catalysis of 9,10-epoxystearic acid between mammalian (from rat liver) and plant (from soybean) sEHs. They found that although the reaction catalyzed by the plant sEH was highly enantioselective towards (*R*)-configured carbon, the mammalian sEH catalysis involved non-enantioselective hydrolysis. Hence, the different binding mode may be related to the structural features which enable positioning of the epoxide ring. This conclusion was in agreement with that of Pineau et al. [60] in which different inhibition patterns were observed between the plant (*Arabidopsis thaliana*) and mammalian EHs. The presented data showed the observed differences may stem from different substrate preferences between these enzymes, and combined with the work of Mowbray et al. [22], showed that *Solanum tuberosum* EH may be very efficient in metabolizing substrates with aliphatic substituents of the epoxide ring. The presented tunnel network analysis sheds light on its potential mechanism. The sEHs flexibility analysis suggested a plausible mechanism of substrate/product transport, instead of the cap domain movement, which could facilitate large substrate access, or highly improbable passage of long-chained substrate entering through one tunnel and leaving by another. Additionally, the secondary structure elements surrounding the two main tunnels could move away, merging Tc/m, Tm1 and Tg tunnels into one long gorge (Fig. 6), which could encompass even long-chained substrates. In the case of F497 residue in hsEH, which is located between Tc/m and Tm1 tunnels, two different orientations in the crystal structure were detected [44]. Due to phenylalanine side chain bulky character, it operates as a molecular gate controlling access through the gorge and promoting proper positioning of the epoxide ring, or closing the tunnel to create a hydrophobic environment for the reaction to occur. This gate may provide constraints affecting a particular substrate preferences. Indeed, the mammalian sEHs have not been used in industry due to their limitations of accepting other epoxide substrates and difficulties in engineering their regioselectivity [61].

Furthermore, plant sEHs predominantly utilize only one funnel-shape tunnel. In our case, the size of the tunnel's mouth and its funnel shape facilitated substrate access to the active site. Thus, in contrast to the long expandable tunnel of mammalian sEHs (which were created by merging of Tc/m, Tm1, and Tg tunnels), Tm1 remained open for a wide range of substrates without causing steric hindrances. Additionally, Tm1 tunnel was capable of transporting the substrates and product, whereas the side tunnels transported water molecules. Therefore, the active site cavity and its surroundings were easily modified and such enzymes were often used in industry. Several reports on StEH1 highlight the high potential of plant EHs as regioselective catalysts [62–65]. Notably, plant EHs are usually highly regio- and/or stereoselective [22,66–68] towards specific substrates. This issue was carefully analyzed both experimentally and theoretically in StEH1, VrEH2 and *Phaseolus vulgaris* PvEH3 [23,69–73]. Reports have shown that modifica-

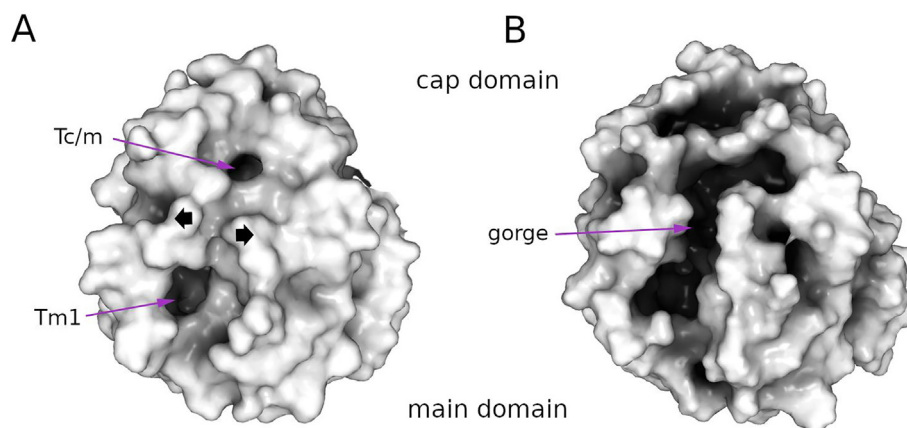


Fig. 6. Opening of the gorge in mouse soluble epoxide hydrolase (msEH): A) the entrance to the active site in the crystal structure (PDB ID: 1cqz), and B) during molecular dynamic simulation. The protein structure is shown as white surface.

tion of the stereo- and regioselectivity of the enzymes is related to the direction of the attack of water molecules and stabilization of particular transition states [72]. Moreover, modifications of the NC-loop, located near the entrance to the Tm1 tunnel, may also enhance the enantioselectivity of PvEH3 [73]. The aforementioned properties can be easily modified in plant EHs with a funnel-shaped entrance to the active site pocket than in mammalian EHs with occluded active site.

Bacterial sEHs have the shortest cap- and NC-loops among all analyzed EHs, which may be related to several bacterial EHs that accept small substrates such as styrene oxide, and mono- and disubstituted epoxides [17]. The main tunnel identified in bacterial structures is short and well-defined, which prevents conversion of long-chained substrates. The short and well-defined tunnel makes such enzymes an easy system for future modification and applicability in industrial processes.

As previously mentioned, a relationship was observed between the enzyme's structure and its overall flexibility and substrate preferences. Due to the limited number of analyzed structures it was hypothesized that in the case of sEHs, divergent evolution occurred. All analyzed enzymes belonged to the same α/β -hydrolase superfamily and shared the same fold, however, they present different substrate preferences profiles, which implied the presence of a common ancestor. Indeed, in the case of other enzymes related to the same superfamily the ancestral protein was identified [74], and displayed enhanced thermal stability and higher specific activity than the extant enzymes. Therefore, bacteria, the most primitive group of living organisms, would present the least complicated mechanism. Due to EHs being required to transport both the substrate and water molecule to the active site cavity, a multi-purpose tunnel or a tunnel network may be needed, in which the substrate could be transported by one tunnel, and the water molecule by another. Our results suggested that bmEH utilized only one tunnel to maintain transport of these two reagents, as well as the reaction product. The Tc/m tunnel was identified in almost all analyzed sEHs (excluding VrEH2), and was considered to be evolutionarily preserved due to its location within the intramolecular voids between the main and cap domains. Other analyzed sEHs utilized predominantly at least one tunnel (Tc/m, Tm1, or both), whereas additional tunnels were rarely used. This may suggest that other sEHs - mammalian, plant and fungal - were more specialized relative to the bacterial bmEH, as they functioned through separate transport water molecules from the transport of substrate and/or products. In our study sEHs of multicellular organisms were divided into two groups - mammalian and fungal (group I, Fig. 3), and plant (group IIa, Fig. 3). Structures of mam-

malian and fungal sEHs although seemed similar, they display a different pattern of overall flexibility. Two mammalian sEHs displayed high flexibility of the cap domain, as well as the α D and α E helices, and in TrEH only the cap-loop and α D helix were slightly flexible. Therefore, the substrates and/or products in the case of mammalian sEHs could possibly be transported *via* the widest and always open tunnel, such as the large gorge formed by merging of Tc/m, Tm1, and Tg tunnels. In the case of TrEH, two separated tunnel entries were observed (Tc/m and Tm1), therefore, it could suggest that the substrate entry and product release occurred *via* different tunnels. Water molecules may enter the active site cavity by a side tunnel. In the case of another α/β -hydrolase superfamily member, dehalogenases, a strategy of separating the substrate entry and product release pathways resulted in the most active dehalogenase identified to date [75]. Moreover, mammalian sEHs are bifunctional enzymes, a product of gene fusion event [76,77] with an N-terminal domain exhibiting phosphatase activity, and the C-terminal domain being an actual EH [78]. Additionally, sEHs in plants and other multicellular organisms evolve independently. We speculated that insertion resulting in cap domain formation was the starting point of the specialization of the EHs. The cap domain covered the active site pocket, and thus enabled precise control of the conditions of the enzymatic reaction. In mammalian and fungal enzymes, the ancestral Tc/m tunnel was preserved and insertions resulted in higher flexibility of the enzyme structure and more complex tunnel network, whereas in plants, the inserts had closed (partially or fully) origin Tc/m tunnel and a new predominant tunnel located in main domain overcharge the substrate/products transportation. Furthermore, the sequences of the plant EHs were divided into two clades - EH1 and EH2 [56]. Moreover, it was shown that the main differences between those clades were located in the cap domain [79]. Among all examined sEHs, mammalian sEHs were subjected to the most precise and rigorous control. The most complicated tunnel network was observed, as well as extensive water exchange between all potential pathways and long range conformational changes capable of merging or separating the particular tunnels.

Our study also explored expansive strategies employed for protein re-engineering. Several approaches have been previously proposed to fine-tune enzyme's activity and/or selectivity through the introduction of additional tunnel or modification of an existing one. Reetz and Kotik's groups focused on *Aspergillus niger* EH existing tunnel leading to the active site and targeted the tunnel-lining and adjacent residues for mutagenesis [80–85]. Thus, they obtained highly enantioselective variants which were useful for producing enantiopure terminal epoxides containing various side

chains. A more advanced approach was shown by Kong *et al.* who introduced an additional tunnel in bmEH using targeted mutagenesis to unblock the steric hindrance in the active pocket [25]. This resulted in the formation of an EH with unusual (*R*)-enantioselectivity and much higher activity toward α -naphthyl glycidyl ether. Brezovsky *et al.* furthered this work by engineering a *de novo* tunnel in a haloalkane dehalogenase LinB, an enzyme closely related to EHs and shared very similar structural features [75]. They opened a novel tunnel by modifications of three residues W140A/F143L/I211L resulting in surface perforation and tunnel opening. The successful tunnel engineering strategies also supported the surface perforation model proposed in our other study on sEHs [15], which described the evolution of tunnels. The presented model suggested that tunnels appeared through even a single-point mutation promoting the formation of two adjacent cavities or permanently opening of an existing cavity. Our study suggested that the tunnel network was also vulnerable to more dramatic modifications such as large fragment indels (insertions/deletions), as depicted the cap domain formation. The MST data suggested that the longer region of the cap-loop identified in sEHs from group I and IIa stem from insertion, as well as the unfolded region of the back-loop of sEHs from group IIb. A strategy based on longer fragment insertion may lead to more complex modifications of the enzyme's activity and/or selectivity, however, such results will be difficult to predict, while the previously described approach based on cavity perforation and tunnel modification could be used for enzyme fine-tuning.

5. Conclusions

This paper is an extension of our other work, in which the evolution of tunnels is studied. We found that tunnels are mostly variable structural features of proteins and a surface perforation model was proposed to describe the mechanism of tunnel appearance. Additionally, interconnection of the protein structure, shape and size of its tunnel network and the substrate preferences are explored. Moreover, our results suggest that tunnels may appear not only due to a single-point mutation, but also by more dramatic structural modifications such as large fragments indels. Finally, sEHs were divided into three groups based on their structure, usage of tunnels, and substrate preferences, indicating that these features are mutually connected.

Funding

This work was funded by the National Science Centre, Poland, grant no DEC-2013/10/E/NZ1/00649. Publication supported by the Own Scholarship Fund of the Silesian University of Technology in the year 2019/2020 (Grant No 919/RN2/RR4/2019).

CRediT authorship contribution statement

Karolina Mitusińska: Conceptualization, Data curation, Funding acquisition, Investigation, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing. **Piotr Wojsa:** Data curation, Funding acquisition, Investigation, Methodology, Software, Validation, Visualization. **Maria Bzówka:** Data curation, Funding acquisition. **Agata Raczyńska:** Data curation, Funding acquisition, Visualization. **Aleksandra Samol:** Data curation, Funding acquisition. **Patryk Kapica:** Data curation. **Artur Góra:** Conceptualization, Funding acquisition, Methodology, Project administration, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.csbj.2021.10.042>.

References

- [1] Pravda L, Berka K, Svobodová Vařeková R, Sehnal D, Banáš P, Laskowski RA, et al. Anatomy of enzyme channels. *BMC Bioinf* 2014;15(1). <https://doi.org/10.1186/s12859-014-0379-x>.
- [2] Mancini G, Zazza C, Soares CM. F429 regulation of tunnels in cytochrome P450 2B4: a top down study of multiple molecular dynamics simulations e0137075. *PLoS ONE* 2015;10(9). <https://doi.org/10.1371/journal.pone.0137075>.
- [3] Fishelovitch D, Shaik S, Wolfson HJ, Nussinov R. How does the reductase help to regulate the catalytic cycle of cytochrome P450 3A4 using the conserved water channel? *J Phys Chem B* 2010;114(17):5964–70. <https://doi.org/10.1021/jp101894k>.
- [4] Gora A, Brezovsky J, Damborsky J. Gates of enzymes. *Chem Rev* 2013;113(8):5871–923. <https://doi.org/10.1021/cr300384w>.
- [5] Marques SM, Daniel L, Buryška T, Prokop Z, Brezovsky J, Damborsky J. Enzyme tunnels and gates as relevant targets in drug design. *Med Res Rev* 2017;37(5):1095–139. <https://doi.org/10.1002/med.21430>.
- [6] Kingsley LJ, Lill MA. Substrate tunnels in enzymes: Structure-function relationships and computational methodology. *Proteins Struct Funct Bioinforma* 2015;83(4):599–611. <https://doi.org/10.1002/prot.24772>.
- [7] Kim J, Raushel FM. Perforation of the tunnel wall in carbamoyl phosphate synthetase derails the passage of ammonia between sequential active sites. *Biochemistry* 2004;43:5334–40. <https://doi.org/10.1021/bi049945+>.
- [8] Nakamura A, Yao M, Chimnarong S, Sakai N, Tanaka I. Ammonia channel couples glutaminase with transamidase reactions in GatCAB. *Science* (80-) 2006;312(5782):1954–8.
- [9] Thangapandian S, John S, Lee Y, Arulalapperumal V, Lee KW, Gaetano C. Molecular modeling study on tunnel behavior in different histone deacetylase isoforms e49327. *PLoS ONE* 2012;7(11). <https://doi.org/10.1371/journal.pone.0049327>.
- [10] Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol* 2009;26(10):2387–95. <https://doi.org/10.1093/molbev/msp146>.
- [11] Tseng YY, Liang J. Estimation of amino acid residue substitution rates at local spatial regions and application in protein function inference: a Bayesian Monte Carlo approach. *Mol Biol Evol* 2006;23:421–36. <https://doi.org/10.1093/molbev/msj048>.
- [12] Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* 2011;188:479–88. <https://doi.org/10.1534/genetics.111.128025>.
- [13] Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. *Proc Natl Acad Sci* 2012;109(14):E831–40. <https://doi.org/10.1073/pnas.1117408109>.
- [14] Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among protein sites. *Nat Rev Genet* 2016;17(2):109–21. <https://doi.org/10.1038/nrg.2015.18>.
- [15] Bzówka M, Mitusińska K, Raczyńska A, Skalski T, Samol A, Bagrowska W, et al. Evolution of tunnels in α/β -hydrolases fold proteins – what can we learn from studying epoxide hydrolases? *bioRxiv* 2021. <https://doi.org/10.1101/2021.12.08.471815>.
- [16] Heikinheimo P, Goldman A, Jeffries Cy, Ollis DL. Of barn owls and bankers: a lush variety of α/β hydrolases. *Structure* 1999;7(6):R141–6. [https://doi.org/10.1016/S0969-2126\(99\)80079-3](https://doi.org/10.1016/S0969-2126(99)80079-3).
- [17] Barth S, Fischer M, Schmid RD, Pleiss J. Sequence and structure of epoxide hydrolases: a systematic analysis. *Proteins Struct Funct Bioinforma* 2004;55(4):846–55. <https://doi.org/10.1002/prot.20013>.
- [18] van Loo B, Kingma J, Arand M, Wubolts MG, Janssen DB. Diversity and biocatalytic potential of epoxide hydrolases identified by genome analysis. *Appl Environ Microbiol* 2006;72(4):2905–17. <https://doi.org/10.1128/AEM.72.4.2905-2917.2006>.
- [19] Berman HM, Battistuz T, Bhat TN, Bluhm WF, Bourne PE, Burkhardt K, et al. The protein data bank. *Acta Crystallogr D Biol Crystallogr* 2002;58(6):899–907. <https://doi.org/10.1107/S0907444902003451>.
- [20] Argiriadi MA, Morisseau C, Hammock BD, Christianson DW. Detoxification of environmental mutagens and carcinogens: Structure, mechanism, and evolution of liver epoxide hydrolase. *Proc Natl Acad Sci* 1999;96(19):10637–42. <https://doi.org/10.1073/pnas.96.19.10637>.
- [21] Gomez GA, Morisseau C, Hammock BD, Christianson DW. Structure of human epoxide hydrolase reveals mechanistic inferences on bifunctional catalysis in epoxide and phosphate ester hydrolysis. *Biochemistry* 2004;43(16):4716–23. <https://doi.org/10.1021/bi036189j>.
- [22] Mowbray SL, Elfström LT, Ahlgren KM, Andersson CE, Widersten M. X-ray structure of potato epoxide hydrolase sheds light on substrate specificity in

- plant enzymes. *Protein Sci* 2006;15(7):1628–37. <https://doi.org/10.1110/ps.051792106>.
- [23] Li F-L, Kong X-D, Chen Qi, Zheng Y-C, Xu Q, Chen F-F, et al. Regioselectivity engineering of epoxide hydrolase: near-perfect enantioconvergence through a single site mutation. *ACS Catal* 2018;8(9):8314–7. <https://doi.org/10.1021/acscatal.8b02622>.
- [24] Wilson C, De Oliveira GS, Adriani PP, Chambergo FS, Dias MVB. Structure of a soluble epoxide hydrolase identified in *Trichoderma reesei*. *Biochim Biophys Acta - Proteins Proteomics* 2017;1865(8):1039–45. <https://doi.org/10.1016/j.bbapap.2017.05.004>.
- [25] Kong X-D, Yuan S, Li L, Chen S, Xu J-H, Zhou J. Engineering of an epoxide hydrolase for efficient bioremediation of bulky pharmaco substrates. *Proc Natl Acad Sci U S A* 2014;111(44):15717–22. <https://doi.org/10.1073/pnas.1404915111>.
- [26] Ferrandi EE, Sayer C, De Rose SA, Guazzelli E, Marchesi C, Saneei V, et al. New thermophilic α/β Class epoxide hydrolases found in metagenomes from hot environments. *Front Bioeng Biotechnol* 2018;6. <https://doi.org/10.3389/fbioe.2018.00144>.
- [27] Dong R, Peng Z, Zhang Y, Yang J. mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics* 2018;34:1719–25. <https://doi.org/10.1093/bioinformatics/btx828>.
- [28] Gouy M, Guindon S, Gascuel O. Sea view version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* 2010;27(2):221–4. <https://doi.org/10.1093/molbev/msp259>.
- [29] Anandakrishnan R, Aguilar B, Onufriev AV. 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res* 2012;40(W1):W537–41. <https://doi.org/10.1093/nar/gks375>.
- [30] Luchko T, Gusarov S, Roe DR, Simmerling C, Case DA, Tuszynski J, et al. Three-dimensional molecular theory of solvation coupled with molecular dynamics in amber. *J Chem Theory Comput* 2010;6(3):607–24. <https://doi.org/10.1021/ct900460m>.
- [31] Sindhikara DJ, Yoshida N, Hirata F. Placevent: an algorithm for prediction of explicit solvent atom distribution-application to HIV-1 protease and F-ATP synthase. *J Comput Chem* 2012;33(18):1536–43. <https://doi.org/10.1002/jcc.22984>.
- [32] Mitusińska K, Raczynska A, Bzówka M, Bagrowska W, Góra A. Applications of water molecules for analysis of macromolecule properties. *Comput Struct Biotechnol J* 2020;18:355–65. <https://doi.org/10.1016/j.csbj.2020.02.001>.
- [33] Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, et al. AMBER14. *Univ Calif*; 2014.
- [34] 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(8):3696–713. <https://doi.org/10.1021/acs.jctc.5b00255>.
- [35] Magdziarz T, Mitusińska K, Bzówka M, Raczynska A, Stańczak A, Banas M, et al. AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective. *Bioinformatics* 2020;36:2599–601. <https://doi.org/10.1093/bioinformatics/btz946>.
- [36] Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, et al. The α/β hydrolase fold. *Protein Eng Des Sel* 1992;5(3):197–211. <https://doi.org/10.1093/protein/5.3.197>.
- [37] Mitusińska K, Magdziarz T, Bzówka M, Stańczak A, Góra A. Exploring solanum tuberosum epoxide hydrolase internal architecture by water molecules tracking. *Biomolecules* 2018;8:143. <https://doi.org/10.3390/biom8040143>.
- [38] Bauer TL, Buchholz PCF, Pleiss J. The modular structure of α/β -hydrolases. *FEBS J* 2020;287(5):1035–53. <https://doi.org/10.1111/febs.15071>.
- [39] Jochens H, Hessler M, Stiba K, Padhi SK, Kazlauskas RJ, Bornscheuer UT. Protein Engineering of α/β -Hydrolase Fold Enzymes. *ChemBioChem* 2011;12(10):1508–17. <https://doi.org/10.1002/cbic.201000771>.
- [40] Miled N, Bussetta C, De caro A, Rivière M, Berti L, Cnaan S. Importance of the lid and cap domains for the catalytic activity of gastric lipases. *Comp Biochem Physiol Part B Biochem Mol Biol* 2003;136(1):131–8. [https://doi.org/10.1016/S1096-4959\(03\)00183-0](https://doi.org/10.1016/S1096-4959(03)00183-0).
- [41] Dugi KA, Dichek HL, Talley GD, Brewer HB, Santamarina-Fojo S. Human lipoprotein lipase: the loop covering the catalytic site is essential for interaction with lipid substrates. *J Biol Chem* 1992;267(35):25086–91.
- [42] Li C, Hu B-C, Wen Z, Hu D, Liu Y-Y, Chu Q, et al. Greatly enhancing the enantioselectivity of PVEH2, a Phaseolus vulgaris epoxide hydrolase, towards racemic 1,2-epoxyhexane via replacing its partial cap-loop. *Int J Biol Macromol* 2020;156:225–32. <https://doi.org/10.1016/j.ijbiomac.2020.04.071>.
- [43] Li B, Yang G, Wu L, Feng Y, Oberer M. Role of the NC-loop in catalytic activity and stability in lipase from *Fervidobacterium changbaicum* e46881. *PLoS ONE* 2012;7(10). <https://doi.org/10.1371/journal.pone.0046881>.
- [44] Bzówka M, Mitusińska K, Hopko K, Góra A. Computational insights into the known inhibitors of human soluble epoxide hydrolase. *Drug Discov Today* 2021;26(8):1914–21. <https://doi.org/10.1016/j.drudis.2021.05.017>.
- [45] Gomez GA. Human soluble epoxide hydrolase: structural basis of inhibition by 4-(3-cyclohexylureido)-carboxylic acids. *Protein Sci* 2006;15(1):58–64. <https://doi.org/10.1110/ps.051720206>.
- [46] Simpson AGB, Roger AJ. The real “kingdoms” of eukaryotes. *Curr Biol* 2004;14(17):R693–6. <https://doi.org/10.1016/j.cub.2004.08.038>.
- [47] Zeldin DC, Wei SZ, Falck JR, Hammock BD, Snapper JR, Capdevila JH. Metabolism of epoxyeicosatrienoic acids by cytosolic epoxide hydrolase: substrate structural determinants of asymmetric catalysis. *Arch Biochem Biophys* 1995;316(1):443–51. <https://doi.org/10.1006/abbi.1995.1059>.
- [48] Fretland AJ, Omiecinski CJ. Epoxide hydrolases: biochemistry and molecular biology. *Chem Biol Interact* 2000;129(1-2):41–59. [https://doi.org/10.1016/S0009-2797\(00\)00197-6](https://doi.org/10.1016/S0009-2797(00)00197-6).
- [49] Yamanashi H, Boeglin WE, Morisseau C, Davis RW, Sulikowski GA, Hammock BD, et al. Catalytic activities of mammalian epoxide hydrolases with cis and trans fatty acid epoxides relevant to skin barrier function. *J Lipid Res* 2018;59(4):684–95. <https://doi.org/10.1194/jlr.M082701>.
- [50] Blee E, Schuber F. Biosynthesis of cutin monomers: involvement of a lipoxygenase/peroxygenase pathway. *Plant J* 1993;4(1):113–23. <https://doi.org/10.1046/j.1365-3113.1993.04010113.x>.
- [51] Arahira M, Nong VH, Uda K, Fukazawa C. Purification, molecular cloning and ethylene-inducible expression of a soluble-type epoxide hydrolase from soybean (*Glycine max*[L.] Merr.). *Eur J Biochem* 2000;267:2649–57. <https://doi.org/10.1046/j.1432-1327.2000.01276.x>.
- [52] Murray GI, Paterson PJ, Weaver RJ, Even SWB, Melvin WT, Burke MD. The expression of cytochrome P-450, epoxide hydrolase, and glutathione S-transferase in hepatocellular carcinoma. *Cancer* 1993;71:36–43. [https://doi.org/10.1002/1097-0142\(19930101\)71:1<36::AID-CNCR2820710107>3.0.CO;2-J](https://doi.org/10.1002/1097-0142(19930101)71:1<36::AID-CNCR2820710107>3.0.CO;2-J).
- [53] Kiyosue T, Beetham JK, Pinot F, Hammock BD, Yamaguchi-Shinozaki K, Shinozaki K. Characterization of an Arabidopsis cDNA for a soluble epoxide hydrolase gene that is inducible by auxin and water stress. *Plant J* 1994;6(2):259–69. <https://doi.org/10.1046/j.1365-3113.1994.6020259.x>.
- [54] Guo A, Durner J, Klessig DF. Characterization of a tobacco epoxide hydrolase gene induced during the resistance response to TMV. *Plant J* 1998;15. <https://doi.org/10.1046/j.1365-3113.1998.00241.x>.
- [55] Gomi K, Yamamoto H, Akimitsu K. Epoxide hydrolase: a mRNA induced by the fungal pathogen *Alternaria alternata* on rough lemon (*Citrus jambhiri* Lush). *Plant Mol Biol* 2003;53(1/2):189–99. <https://doi.org/10.1023/B:PLAN.000009287.95682.24>.
- [56] Huang F-C, Schwab W. Molecular characterization of NbeH1 and NbeH2, two epoxide hydrolases from *Nicotiana benthamiana*. *Phytochemistry* 2013;90:6–15. <https://doi.org/10.1016/j.phytochem.2013.02.020>.
- [57] Stark A, Lundholm A-K, Meijer J. Comparison of fatty acid epoxide hydrolase activity in seeds from different plant species. *Phytochemistry* 1995;38(1):31–3. [https://doi.org/10.1016/0031-9422\(94\)00646-B](https://doi.org/10.1016/0031-9422(94)00646-B).
- [58] Badami R, Patil K. Structure and occurrence of unusual fatty acids in minor seed oils. *Prog Lipid Res* 1980;19(3-4):119–53. [https://doi.org/10.1016/0163-7827\(80\)90002-8](https://doi.org/10.1016/0163-7827(80)90002-8).
- [59] Summerer S, Hanano A, Utsumi S, Arand M, Schuber F, Blée E. Stereochemical features of the hydrolysis of 9,10-epoxystearic acid catalysed by plant and mammalian epoxide hydrolases. *Biochem J* 2002;366(2):471–80.
- [60] Pineau E, Xu L, Renault H, Trolet A, Navrot N, Ullmann P, et al. Arabidopsis thaliana EPOXIDE HYDROLASE1 (ATEH1) is a cytosolic epoxide hydrolase involved in the synthesis of poly-hydroxylated cutin monomers. *New Phytol* 2017;215(1):173–86. <https://doi.org/10.1111/nph.14590>.
- [61] Decker M, Arand M, Cronin A. Mammalian epoxide hydrolases in xenobiotic metabolism and signalling. *Arch Toxicol* 2009;83(4):297–318. <https://doi.org/10.1007/s00204-009-0416-0>.
- [62] Thomaus A, Naworyta A, Mowbray SL, Widersten M. Removal of distal protein-water hydrogen bonds in a plant epoxide hydrolase increases catalytic turnover but decreases thermostability. *Protein Sci* 2008;17(7):1275–84. <https://doi.org/10.1110/ps.034173.107>.
- [63] Lindberg D, Ahmad S, Widersten M. Mutations in salt-bridging residues at the interface of the core and lid domains of epoxide hydrolase STEH1 affect regioselectivity, protein stability and hysteresis. *Arch Biochem Biophys* 2010;495(2):165–73. <https://doi.org/10.1016/j.abb.2010.01.007>.
- [64] Carlsson AJ, Bauer P, Ma H, Widersten M. Obtaining optical purity for product diols in enzyme-catalyzed epoxide hydrolysis: contributions from changes in both enantio- and regioselectivity. *Biochemistry* 2012;51(38):7627–37. <https://doi.org/10.1021/bi3007725>.
- [65] Bauer P, Carlsson AJ, Amrein BA, Dobritzsch D, Widersten M, Kamerlin SCL. Conformational diversity and enantioconvergence in potato epoxide hydrolase 1. *Org Biomol Chem* 2016;14(24):5639–51. <https://doi.org/10.1039/C6OB00060F>.
- [66] Zhang C, Li C, Zhu X, Liu Y, Zhao J, Wu M. Highly regio- and enantio-selective hydrolysis of two racemic epoxides by GmEH3, a novel epoxide hydrolase from *Glycine max*. *Int J Biol Macromol* 2020;164:2795–803. <https://doi.org/10.1016/j.ijbiomac.2020.08.011>.
- [67] Hu B-C, Hu D, Li C, Xu X-F, Wen Z, Wu M-C. Near-perfect kinetic resolution of racemic p-chlorostyrene oxide by SIEH1, a novel epoxide hydrolase from *Solanum lycopersicum* with extremely high enantioselectivity. *Int J Biol Macromol* 2020;147:1213–20. <https://doi.org/10.1016/j.ijbiomac.2019.10.091>.
- [68] Blée E, Schuber F. Regio- and enantioselectivity of soybean fatty acid epoxide hydrolase. *J Biol Chem* 1992;267(17):11881–7. [https://doi.org/10.1016/S0021-9258\(19\)49780-9](https://doi.org/10.1016/S0021-9258(19)49780-9).
- [69] Elfström LT, Widersten M. Catalysis of potato epoxide hydrolase, STEH1. *Biochem J* 2005;390:633–40. <https://doi.org/10.1042/BJ20050526>.
- [70] Carlsson JA, Bauer P, Dobritzsch D, Kamerlin SCL, Widersten M. Epoxide hydrolysis as a model system for understanding flux through a branched reaction scheme. *IUCrj* 2018;5(3):269–82. <https://doi.org/10.1107/S2052252518003573>.
- [71] Carlsson JA, Bauer P, Dobritzsch D, Nilsson M, Kamerlin SCL, Widersten M. Laboratory-evolved enzymes provide snapshots of the development of

- enantioconvergence in enzyme-catalyzed epoxide hydrolysis. *ChemBioChem* 2016;17(18):1693–7. <https://doi.org/10.1002/cbic.201600330>.
- [72] Lind MES, Himo F. Quantum chemical modeling of enantioconvergence in soluble epoxide hydrolase. *ACS Catal* 2016;6(12):8145–55. <https://doi.org/10.1021/acscatal.6b01562>.
- [73] Zhang C, Liu Y, Li C, Xu Y, Su Y, Li J, et al. Significant improvement in catalytic activity and enantioselectivity of a *Phaseolus vulgaris* epoxide hydrolase, PvEH3, towards ortho-cresyl glycidyl ether based on the semi-rational design. *Sci Rep* 2020;10(1). <https://doi.org/10.1038/s41598-020-58693-1>.
- [74] Babkova P, Sebestova E, Brezovsky J, Chaloupkova R, Damborsky J. Ancestral haloalkane dehalogenases show robustness and unique substrate specificity. *ChemBioChem* 2017;18(14):1448–56. <https://doi.org/10.1002/cbic.201700197>.
- [75] Brezovsky J, Babkova P, Degtjarik O, Fortova A, Gora A, Iermak I, et al. Engineering a de Novo Transport Tunnel. *ACS Catal* 2016;6(11):7597–610. <https://doi.org/10.1021/acscatal.6b02081>.
- [76] Harris TR, Aronov PA, Hammock BD. Soluble epoxide hydrolase homologs in *Strongylocentrotus purpuratus* suggest a gene duplication event and subsequent divergence. *DNA Cell Biol* 2008;27(9):467–77. <https://doi.org/10.1089/dna.2008.0751>.
- [77] Harris TR, Hammock BD. Soluble epoxide hydrolase: gene structure, expression and deletion. *Gene* 2013;526(2):61–74. <https://doi.org/10.1016/j.gene.2013.05.008>.
- [78] Cronin A, Mowbray S, Durk H, Homburg S, Fleming I, Fisslthaler B, et al. The N-terminal domain of mammalian soluble epoxide hydrolase is a phosphatase. *Proc Natl Acad Sci* 2003;100(4):1552–7. <https://doi.org/10.1073/pnas.0437829100>.
- [79] Wijekoon CP, Goodwin PH, Valliani M, Hsiang T. The role of a putative peroxisomal-targeted epoxide hydrolase of *Nicotiana benthamiana* in interactions with *Colletotrichum destructivum*, *C. orbiculare* or *Pseudomonas syringae* pv. *tabaci*. *Plant Sci* 2011;181(2):177–87. <https://doi.org/10.1016/j.plantsci.2011.05.004>.
- [80] Reetz MT, Torre C, Eipper A, Lohmer R, Hermes M, Brunner B, et al. Enhancing the enantioselectivity of an epoxide hydrolase by directed evolution. *Org Lett* 2004;6(2):177–80. <https://doi.org/10.1021/ol035898m10.1021/ol035898m.s001>.
- [81] Reetz MT, Bocola M, Wang L-W, Sanchis J, Cronin A, Arand M, et al. Directed evolution of an enantioselective epoxide hydrolase: uncovering the source of enantioselectivity at each evolutionary stage. *J Am Chem Soc* 2009;131(21):7334–43. <https://doi.org/10.1021/ja809673d>.
- [82] Reetz MT, Zheng H. Manipulating the expression rate and enantioselectivity of an epoxide hydrolase by using directed evolution. *ChemBioChem* 2011;12(10):1529–35. <https://doi.org/10.1002/cbic.201100078>.
- [83] Reetz MT, Wang L-W, Bocola M. Directed evolution of enantioselective enzymes: Iterative cycles of CASTing for probing protein-sequence space. *Angew Chemie - Int Ed* 2006;45(8):1236–41. <https://doi.org/10.1002/anie.200502746>.
- [84] Kotik M, Štěpánek V, Kyslík P, Marešová H. Cloning of an epoxide hydrolase-encoding gene from *Aspergillus niger* M200, overexpression in *E. coli*, and modification of activity and enantioselectivity of the enzyme by protein engineering. *J Biotechnol* 2007;132(1):8–15. <https://doi.org/10.1016/j.jbiotec.2007.08.014>.
- [85] Kotik M, Archelas A, Faměrová V, Oubrechtová P, Křen V. Laboratory evolution of an epoxide hydrolase – towards an enantioconvergent biocatalyst. *J Biotechnol* 2011;156(1):1–10. <https://doi.org/10.1016/j.jbiotec.2011.08.003>.



Corrigendum

Corrigendum to “Structure–function relationship between soluble epoxide hydrolases structure and their tunnel network” [Comput. Struct. Biotechnol. J. 20 (2022) 193–205]



Karolina Mitusińska, Piotr Wojsa, Maria Bzówka, Agata Raczyńska, Weronika Bagrowska, Aleksandra Samol, Patryk Kapica, Artur Góra*

Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland

The authors regret that during the revision process of the follow-up paper (reference 15 Bzówka et al. ‘Evolution of tunnels in α/β -hydrolases fold proteins – what can we learn from studying epoxide hydrolases?’ – previously published as a preprint), the multiple sequence alignment (MSA) used for calculating the tunnels’ variability was not adequately post-processed. After verifying all the results, the variability of the tunnels changed and almost all identified tunnels can be described as conserved features. These findings do not call into question the results presented in the article ‘Structure–function relationship between soluble epoxide hydrolases structure and their tunnel network’ published in the Computational and Structural Biotechnology Journal but they influenced the results presented in the reference 15. After introducing the changes, the preprint has been updated and has also been approved for publication in the Plos Computational Biology Journal. Taking responsibility for our work and the quality of the article published in the Computational and Structural Biotechnology Journal, we would like to provide a corrigendum note. We marked the changed text bold, so the changes can be easier to follow.

We would like to introduce the following changes:

Introduction section:

Second paragraph:

Original text: “However, in our other study [15] we elucidate that in the case of the soluble epoxide hydrolases (sEHs) most of their tunnels should be considered **as variable structural features with only one exception – the tunnel located at the border between the main and cap domains. These counterintuitive findings** has inspired the investigation of the structure–function relationship of sEHs in more detail.”

Changed text: “In our other study [15] we elucidate that in the case of the soluble epoxide hydrolases (sEHs) most of their tunnels should be considered **as conserved structural features. These**

findings have inspired the investigation of the structure–function relationship of sEHs in more detail.”

Discussion section:

Second paragraph:

Original text: “In our other study [15], sEHs were employed as a sample system in order to investigate the evolution of tunnels. It was determined that most tunnels should be considered as **variable** structural features of proteins. **Tc/m tunnel was found to be the only exception, located between the cap and main domains.** We proposed that insertion of the cap domain defined the buried active site cavity and the tunnel linking it with the environment. Such structural arrangement was preserved in most of the EHs which supports the hypothesis regarding the origin of the positioning of the active site between both domains.

Changed text: “In our other study [15], sEHs were employed as a sample system in order to investigate the evolution of tunnels. It was determined that most tunnels should be considered as **conserved** structural features of proteins **with Tc/m tunnel identified in all analyzed structures, between the cap and main domains.** We proposed that insertion of the cap domain defined the buried active site cavity and the tunnel linking it with the environment. Such structural arrangement was preserved in most of the EHs which supports the hypothesis regarding the origin of the positioning of the active site between both domains”.

Fourth paragraph:

Original text: “Mammalian (hsEH and msEH) and fungal (TrEH) structures were assigned to group I. Members of this group shared common features such as relatively long back-loop and cap-loop. Enzymes in this group primarily utilize two main tunnels – Tc/m, and Tm1. In all sEHs from the group I, T/cm tunnel was found conserved [15]. This was also the case for Tm1 tunnel, **but only in the case of msEH** [15]. The results of the structure flexibility analysis (Fig. 5) showed significant differences between sEHs that represent mammalian and fungal families.”

Changed text: “Mammalian (hsEH and msEH) and fungal (TrEH) structures were assigned to group I. Members of this group shared common features such as relatively long back-loop and cap-loop. Enzymes in this group primarily utilize two main tunnels – Tc/m, and Tm1. In all sEHs from the group I, T/cm tunnel was found con-

DOI of original article: <https://doi.org/10.1016/j.csbj.2021.10.042>

* Corresponding author at: Biotechnology Centre, Krzywoustego 8, 44-100 Gliwice, Poland.

E-mail address: a.gora@tunnelinggroup.pl (A. Góra).

served. This was also the case for the Tm1 tunnel [15]. The results of the structure flexibility analysis (Fig. 5) showed significant differences between sEHs that represent mammalian and fungal families.”

Fifth paragraph:

Original text: “Similar to mammalian and fungal sEHs, plant sEHs structures had relatively long cap-loop and back-loop, however, the enzymes predominantly utilize the Tm1 tunnel, which was identified as a **variable feature in StEH1 structure** [15]. The flexibility analysis results of plant sEHs showed that the most flexible regions were distant to the tunnel entries and, therefore, the conformational changes were only limited to slight effect on for catalytic efficiency (if any).”

Changed text: “Similar to mammalian and fungal sEHs, plant sEHs structures had relatively long cap-loop and back-loop, however, the enzymes predominantly utilize the Tm1 tunnel, which was identified as a **conserved feature in StEH1 structure** [15]. The flexibility analysis results of plant sEHs showed that the most

flexible regions were distant to the tunnel entries and, therefore, the conformational changes were only limited to slight effect on for catalytic efficiency (if any).”

Sixth paragraph:

Original text: “This observation supported the hypothetical origin of sEHs via insertion resulting in active site positioning between cap and main domains. **Surprisingly, in the case of Ilb group enzymes the Tc/m tunnel was found to be a variable feature [15]. This could be due a small number of residues lining the walls of the tunnel, which was significantly shorter in comparison to Tc/m tunnels in other sEHs.**”

Changed text: “This observation supported the hypothetical origin of sEHs via insertion resulting in active site positioning between cap and main domains. **In the case of Ilb group enzymes the Tc/m tunnel was found to be a conserved structural feature [15].**”

The authors would like to apologise for any inconvenience caused.

Geometry-Based versus Small-Molecule Tracking Method for Tunnel Identification: Benefits and Pitfalls

Karolina Mitusińska,[‡] Maria Bzówka,[‡] Tomasz Magdziarz, and Artur Góra*

 Cite This: *J. Chem. Inf. Model.* 2022, 62, 6803–6811

 Read Online

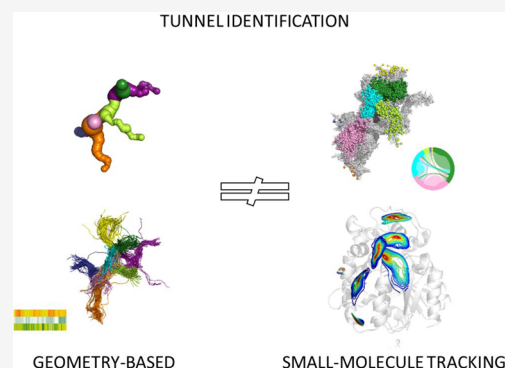
ACCESS |

 Metrics & More

 Article Recommendations

 Supporting Information

ABSTRACT: Different methods for tunnel identification, geometry-based and small-molecule tracking approaches, were compared to provide their benefits and pitfalls. Results obtained for both crystal structures and molecular dynamics (MD) simulations were analyzed to investigate if a more computationally demanding method would be beneficial. Careful examination of the results is essential for the low-diameter tunnel description, and assessment of the tunnel functionality based only on their geometrical parameters is challenging. We showed that the small-molecule tracking approach can provide a detailed description of the system; however, it can also be the most computationally demanding.



INTRODUCTION

Most bioinformatics workflows start with an application of a simple approach providing a general description of the problem followed by the application of more complex and time-consuming solutions that guarantee a deeper understanding of the described phenomena. The same pipeline is observed in structural biology studies, such as tunnel identification in protein structure.¹ In our study, a tunnel is defined as a pathway connecting the protein surface with an internal cavity or a pathway connecting more than one cavity (definition taken from Prokop et al.²). Tunnels gain significant importance due to the set of functions they maintain in enzymes, i.e., control of the activity and selectivity and reaction synchronization.^{3–5} More than half of currently known protein structures are equipped in tunnels; therefore, tunnel identification is carried out as a standard procedure, especially in enzymes with buried active sites.⁶

First and still the most commonly used approach for tunnel identification is the geometry-based approach reviewed in ref 1. This approach employs the construction of a Voronoi diagram to detect and describe voids within a macromolecule structure. Then, tunnels are identified using a predefined probe radius and internal “empty spaces”. However, this approach is usually used to analyze crystal structures or single molecular dynamics (MD) simulation snapshots. This approach was implemented in different software, such as CAVER 3.0, MOLE 2.0, MolAxis, or ChExVis (which do not differ substantially as shown by Brezovsky et al.¹). Of those geometry-based methods, only CAVER 3.02 analyzes the whole MD simulation, which provides a general overview of the potential tunnels connecting the active site with the enzyme’s environment. CAVER 3.02 is

applied to a series of frames derived from MD simulations where the dynamic of tunnel-lining residues is taken into account. In each analyzed snapshot (the user can select which frames they want to analyze further), tunnels are identified based on the diameter of the defined probe. The clustering algorithm implemented in CAVER 3.02 is applied to compare and group identified tunnels, and thus, the tunnel opening and closing events can be observed.⁷ Still, this approach considers only the geometry of the tunnels, while the physical and chemical properties of potential molecules transported via tunnels are neglected. This simplification may not be considered an obstacle if there is only one tunnel leading to the active center. However, this is not always the case.^{4,8} The choice of the transport pathway for a given substrate/product is no longer trivial in the case of multiple tunnels connecting the active center to the environment. Tunnels in proteins maintain different functions, such as transport of ligand, product, solvent, and/or ions to and from the active site. Description of a tunnel’s functionality, even *in silico*, is a complex process that requires lots of computational efforts, such as MD simulations combined with tracking of small molecules through the tunnels.^{4,8,9}

A different approach to tunnel identification has been proposed by the developers of the AQUA-DUCT software.^{10,11}

Special Issue: Advancing Women in Chemistry

Received: August 2, 2022

Published: November 14, 2022



The software uses a small-molecule tracking approach to provide information on the flow direction and tunnel contribution during MD simulations. AQUA-DUCT traces water molecules (or other selected small molecules present in the simulated system) penetrating the protein's interior. Thus, in contrast to the geometry-based methods, it includes physicochemical properties of the tunnel-lining residues and identifies only those tunnels which are capable of transporting water or other small molecules of interest. However, this approach requires analysis of bigger files (MD simulation trajectory files consisting of protein and solvent molecules) and relatively large sampling (in terms of the number of frames) of MD simulations to draw and analyze the pathways of the analyzed small molecules (for benchmark of the AQUA-DUCT resource usage and effect of the trajectory time-step on the obtained results, see ref 11). Therefore, the small-molecule tracking approach may be more demanding compared with the geometry-based approach in terms of preparing the MD simulation trajectory files and their storage.

So far, no comparison of the above-mentioned approaches has been made. An extensive comparison of the geometry-based methods was made by Brezovsky et al.,¹ in which they also stressed the existing limitations of those types of analysis, such as lack of information on electrostatics, hydrophobicity, or dynamics of identified pathways. However, it remains unknown whether it is beneficial to use outcomes of the MD simulations or if the analysis of crystallographic structures is sufficient. In this study, we collated the profits of using more advanced tools with the oversights or misinterpretations of using the simplest techniques. As a model system, we chose representative members of the soluble epoxide hydrolases (sEH), a group of enzymes which belong to the α/β -hydrolases fold family,^{12–14} due to their diverse tunnel network.¹⁵ We hope that our results shed light on tunnel identification in protein structure and the interpretation of the results and will help researchers with an adequate selection of the method corresponding with their requirements and expectations.

MATERIALS AND METHODS

Obtaining Protein Structures for Analysis. Eight unique and complete crystal structures were downloaded from the Protein Data Bank (PDB)¹⁶ representing the same set of structures as used elsewhere:^{15,17} *Homo sapiens* (hsEH, PDB ID: 1S8O), *Mus musculus* (msEH, PDB ID: 1CQZ), *Solanum tuberosum* (StEH1, PDB ID: 2CJP), *Vigna radiata* (VrEH2, PDB ID: 5XM6), (*Trichoderma reesei* (TrEH, PDB ID: SURO), *Bacillus megaterium* (bmEH, PDB ID: 4NZZ), and two structures from an unknown source organism collected from hot springs in Russia and China (Sibe-EH, PDB ID: 5NG7; CH65-EH, PDB ID: 5NFQ).

MD Simulations. The H++ server¹⁸ was used to protonate the analyzed structures using standard parameters at the reported optimal pH for the enzyme activity (Table S1). Counterions were added to the structures to neutralize the systems. Water molecules were placed using the combination of 3D-RISM theory¹⁹ and Placevent algorithm.²⁰ The Amber14 tLEaP²¹ package was used to immerse the models in a truncated octahedral box with a 10 Å radius of TIP3P water molecules, and the ff14SB force field²² was used for parametrization of each system. A PMEMD CUDA package of Amber14 software was used to run a single repetition of a 50 ns MD simulation of selected sEHs. The minimization

procedure consisted of 2000 steps, involving 1000 steepest descent steps followed by 1000 steps of conjugate gradient energy minimization with decreasing constraints on the protein backbone (500, 125, and 25 kcal \times mol⁻¹ \times Å⁻²) and a final minimization with no constraints of conjugate gradient energy minimization. Next, gradual heating was performed from 0 to 300 K over 20 ps using a Langevin thermostat with a collision frequency of 1.0 ps⁻¹ in periodic boundary conditions with constant volume. The equilibration stage was conducted using the periodic boundary conditions with constant pressure for the time stated in Table S1 with a 1 fs time step using Langevin dynamics with a collision frequency of 1.0 ps⁻¹ to maintain a constant temperature. The production stage was conducted for 50 ns with a 2 fs time step using Langevin dynamics with a collision frequency of 1.0 ps⁻¹ to maintain a constant temperature. Long-range electrostatic interactions were modeled using the particle mesh Ewald method with a nonbonded cutoff of 10 Å and SHAKE algorithm. The coordinates were saved at 1 ps intervals. The number of added water molecules and ions is shown in Table S1.

Tunnel Identification: CAVER Analysis. Tunnel identification and analysis in each system was carried out using CAVER software²³ in two steps: (i) the crystal structure of the enzyme was analyzed by the CAVER plugin for PyMOL;²³ (ii) tunnels were identified and analyzed in 50,000 snapshots of multiple MD simulations by CAVER 3.02 software.²³ Parameters used for both steps are shown in Table S2. The tunnels found during MD simulations and in crystal structures were ranked and numbered on the basis of their throughput value.²³

Tunnel Identification: AQUA-DUCT Analysis. AQUA-DUCT analysis was carried out according to the protocol described elsewhere.^{15,24} A small-molecule tracking approach implemented in AQUA-DUCT^{10,11} was used for tunnel identification and assessment of their functionality. Tunnel's functionality was defined as the ability of the tunnel to transport small molecules (such as water molecules, ions, ligands, or cosolvents, such as methanol, phenol, etc.).

Tunnels Comparison. Tunnels were identified in both crystal structures and during MD simulations and then compared with each other to find their corresponding counterparts. First, the tunnels identified during MD simulations and in crystal structures were maintained using the same approach as described elsewhere.¹⁷ In the case of tunnels identified in MD simulations by CAVER 3.02 but for which no corresponding counterpart was found in the crystal structures by CAVER plugin for PyMOL, their tunnel-lining residues were selected based on the cutoff threshold of 0.65. This value was chosen on the basis of quantile computations for the tunnels identified in MD simulation, which had their counterparts in the crystallographic structures.

Tunnel functionality was then assessed based on a small-molecule tracking approach implemented in AQUA-DUCT by superposing the paths of water molecules, and their entry/exit areas with tunnels were identified by CAVER in both the crystal structures and during MD simulations. A visual comparison allowed matching of the water molecule pathways and tunnels identified by CAVER.

RESULTS

Here, we chose the same set of eight sEHs as presented in ref 15. We mimicked a typical approach used in various studies regarding tunnel identification in protein structure (Figure 1).

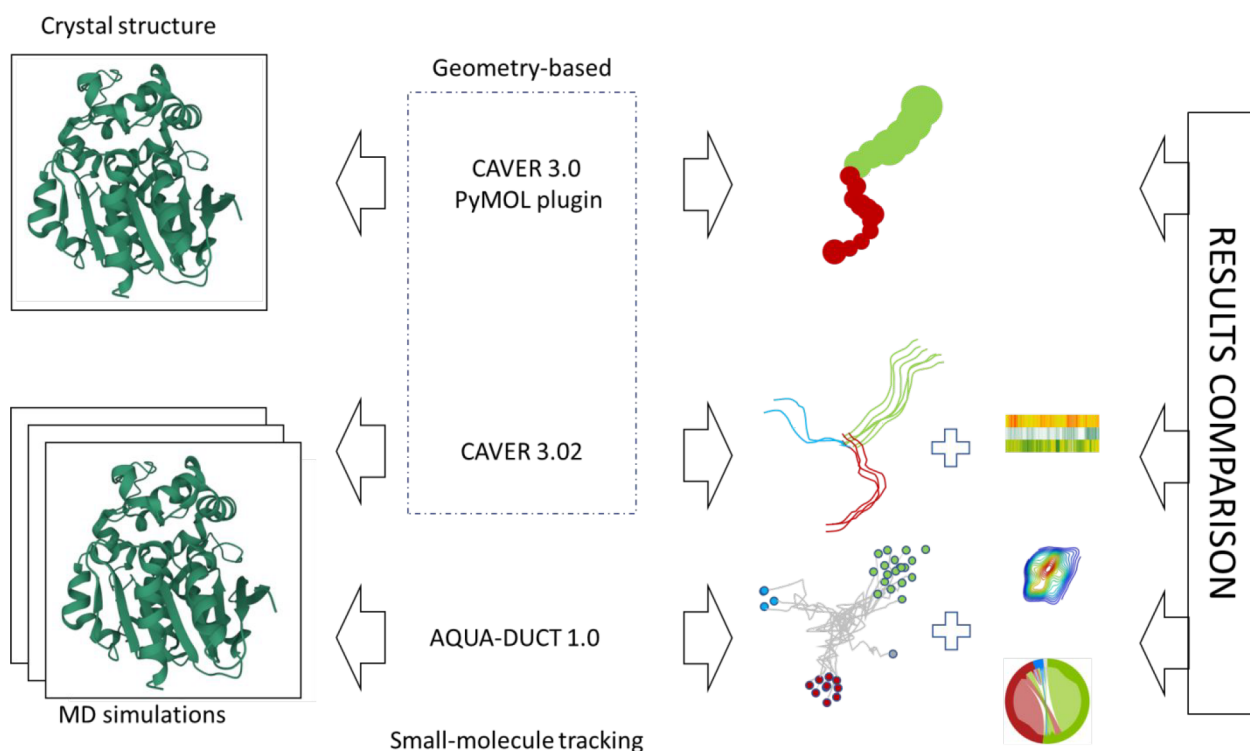


Figure 1. Tunnel identification and schematic comparison of the obtained results: (upper row) CAVER 3.0 PyMOL plugin, tunnel visualization in crystal structures; (middle row) CAVER 3.02, centerlines of the identified tunnels and tunnel occurrence heatmap; (bottom row) AQUA-DUCT, inlet clusters with water molecule pathways, entry/exit areas, and intramolecular flowchart.

The terms tunnel and channel are often used interchangeably in the scientific literature; therefore, based on Prokop et al.,² we used unifying terminology: that a tunnel is a pathway connecting the protein surface with an internal cavity or a pathway connecting two or more cavities. A channel is then a pathway leading throughout the protein structure without any interruption by an internal cavity with both sides open to the surrounding solvent. We started with a simple analysis of crystal structures downloaded from Protein Data Bank using a geometry-based tool, CAVER 3.0 PyMOL plugin, which is one of the most widely used tools for tunnel identification. Then, we expanded our analysis to tunnel identification during MD simulations. We used CAVER 3.02 software to analyze a single repetition of MD per protein, as is often the case in other studies. CAVER 3.02 is based on the same principles as its plugin counterpart with the advantage of taking into account the information from MD simulations. Lastly, we used AQUA-DUCT 1.0, which uses the small-molecule tracking approach during MD simulations. As an input, we used the same MD simulations as were used during CAVER 3.02 analysis. Thus, for each structure, we obtained results from three different approaches: (i) geometry-based approach applied on a crystal structure, (ii) geometry-based approach applied on an MD simulation, and (iii) small-molecule tracking approach applied on an MD simulation. A comparison of these results (Tables 1 and S2–S9) provided insights on when it is best to use a particular approach as well as their benefits and pitfalls and how they can bias the bigger picture.

Tunnels Identification in Crystal Structures by CAVER 3.0 PyMOL Plugin. The simplest approach aims to identify tunnels in crystal structures. The CAVER 3.0 PyMOL plugin provides information about the number of tunnels, their length, and bottleneck radius. In our study, it identified three

(in CH65-EH) to nine (in hsEH) tunnels in the analyzed protein structures with the maximal bottleneck ranging from 0.9 Å in bmEH to 2.4 Å in the TrEH structure (Figure 2, Table 1). We used the same naming for the identified tunnels as in our previous studies,^{15,17} based on the region in which the tunnel was identified (Tcap, for tunnels found in the cap domain; Tm, for tunnels identified in the main domain; Tc/m, for tunnel identified at the border between both domains). The detailed list of tunnels identified in the crystal structures is in Table 1.

Tunnels Identification in MD Simulations by CAVER 3.02. We ran a single repetition of MD simulations for each sEH and analyzed them using CAVER 3.02 software, which processed a set of snapshots from an MD simulation and identified tunnels in each of them. Then, CAVER 3.02 performed clustering on tunnels which it considers similar; i.e., tunnels whose portions lead through the same part of the structure. Clustering provided a clear picture of tunnels in the same conformation. The number of tunnels identified by the CAVER 3.02 software in an MD simulation was often higher than the number of tunnels identified by the CAVER 3.0 PyMOL plugin in the crystal structure (Figure 2, Table 1). This is due to the formational changes which proteins undergo. Comparison of tunnels and selection of the best corresponding counterparts (Tables S3–S10) were done based on the similarity of tunnel-lining residues (see Materials and Methods section for a detailed description of the comparison procedure).

Additionally, CAVER 3.02 provided information on the tunnel's occurrence, which is measured as the number of frames in which the tunnel was identified. In most structures, except VrEH2, at least one tunnel was identified as open during the whole simulation time. In four structures, msEH,

Table 1. Comparison of Results Obtained by the Geometry-Based and Small-Molecule Tracking Approaches in the Crystal Structures and during MD Simulations^a

enzyme	tunnel	CAVER 3.0 PyMOL plugin		CAVER 3.02			AQUA-DUCT	
		crystal structure		MD simulations			inlets	cluster area [Å ²]
		rank	max_bottleneck [Å]	rank	max_bottleneck [Å]	occurrence [%]		
hsEH	Tc/m1	1	1.58	2	2.70	59	554 (22%)	67.0
	Tm1	2	1.78	1	2.87	100	1830 (79%)	93.2
	Tg	3	1.10	8	1.86	11	94 (4%)	82.5
	Tc/m2	4	1.34	15	1.96	1		
	Tm5	5	1.18	5	1.82	54	1 (<1%)	
	Tcap4	6	1.15	16	1.34	1		
	Tcap2b	7	0.95	7	1.52	19	3 (<1%)	1.4
	Tm3	8	0.92	14	1.11	2		
	Tm2			3	2.57	95	24 (1%)	15.7
	Tc/m_side			14	2.00	48	1 (<1%)	
msEH	Tcap1						7 (<1%)	46.5
	Tc/m1	1	2.09	1	3.18	99	1627 (38%)	70.7
	Tm1	2	2.01	2	3.14	100	1400 (32%)	77.7
	Tcap4	3	1.12	6	1.93	51	6 (<1%)	11.7
	Tm2	4	1.03	5	2.23	57	7 (<1%)	21.7
	Tm3	5	1.02	3	2.37	88	215 (5%)	108.9
	Tside	6	0.96	9	1.21	11	5 (<1%)	
	Tg			4	2.96	71	1031 (24%)	46.5
TrEH	Tcap1			7	2.61	15	33 (1%)	86.4
	Tc/m1	1	2.40	1	2.93	96	986 (41%)	234.7
	Tm1	2	1.45	2	2.65	100	1200 (54%)	212.3
	Tcap4	3	1.28	3	1.95	53	23 (1%)	41.9
	Tside	4	0.96	4	2.33	53	10 (<1%)	26.5
	Tm5	5	0.91	6	1.81	12		
	Tback	7	0.91	13	1.1	2		
StEH1	Tc/m_back						1 (<1%)	
	Tm1	1	1.79	1	3.07	100	1774 (89%)	149.4
	Tc/m1	2	1.40	3	2.11	98	149 (7%)	7.8
	Tm2	3	1.79	4	2.52	65	65 (3%)	112.0
	Tcap3	4	1.14	12	1.42	13	1 (<1%)	
	Tcap6	5	0.97	7	1.36	28		
	Tc/m_back	6	1.10	16	1.28	8		
VrEH2	Tcap5	7	0.93	9	1.44	19		
	Tm5			6	2.04	21	6 (<1%)	11.8
	Tm1	1	1.41	1	2.84	89	563 (93%)	68.5
	Tcap1	2	1.30	4	1.61	53	1 (<1%)	
	Tside	3	1.31	5	1.51	53		
	Tm5	4	1.14	2	1.98	80	10 (2%)	24.8
	Tm2			3	2.00	77	27 (4%)	30.9
bmEH	Tcap2b			7	1.51	30	1 (<1%)	
	Tcap7			13	1.49	7	1 (<1%)	
	Tc/m1	2	1.92	1	2.74	100	5256 (100%)	88.4
CH65-EH	Tc/m_back	3	1.06	3	1.88	18		
	Tcap7	4	0.90	4	1.17	1		
	Tc/m1	1	1.45	2	2.29	97	3375 (88%)	67.29
Sibe-EH	Tc/m_back	2	1.56	1	2.48	100		
	Tc/m_side	3	1.19	5	1.54	28		
	Tm4			4	2.06	46	359 (9%)	126.1
	Tcap4			6	1.72	25	84 (2%)	104.9
Sibe-EH	Tc/m1	1	1.89	1	2.51	100	1011 (98%)	30.1
	Tc/m3	2	1.11	3	1.88	23		
	Tc/m_back	3	1.16	9	1.20	4	20 (2%)	86.5
	Tcap4	4	1.05	6	1.70	14	1 (<1%)	
	Tc/m2	5	1.91	12	1.44	2		
Tside	6	0.91	13	1.24	2	1 (<1%)		

^aPlease note that the table comprises the best matches between tunnels identified in crystal structures and during MD simulations. The detailed tunnel comparison results are provided in Tables S3–S10.

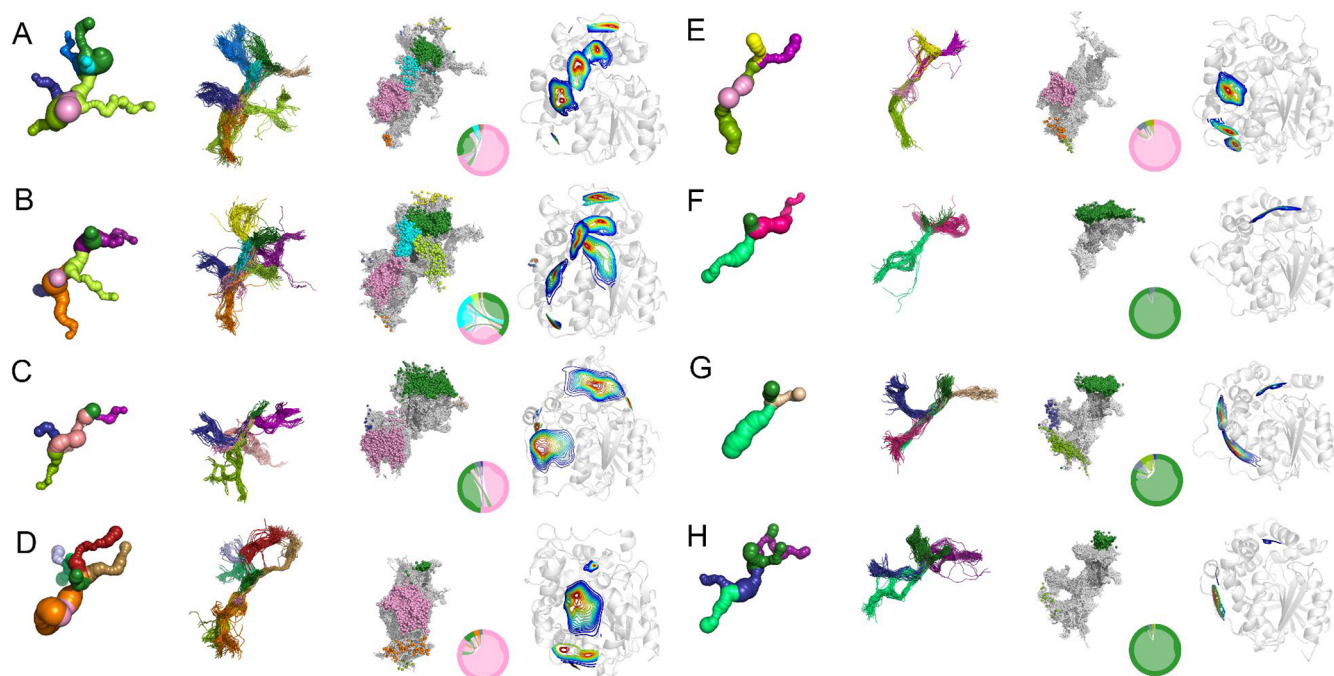


Figure 2. Comparison of results obtained from the geometry-based and small-molecule tracking approaches for the following epoxide hydrolases: (A) *Homo sapiens* (hsEH), (B) *Mus musculus* (msEH), (C) *Trichoderma reesei* (TrEH), (D) *Solanum tuberosum* (StEH1), (E) *Vigna radiata* (VrEH2), (F) *Bacillus megaterium* (bmEH), and (G) Sibe-EH and (H) CH65-EH identified in hot springs. Each panel comprises tunnels obtained by the CAVER 3.0 PyMOL plugin for the crystal structure, tunnel centerlines obtained by CAVER 3.02 software for molecular dynamics (MD) simulations, and inlet clusters with water molecule pathways, entry/exit areas, and an intramolecular flowchart obtained by AQUA-DUCT from MD simulations. Corresponding tunnels, centerlines, and inlet clusters are marked with the same color; entry/exit areas are colored according to their density: blue represents the overall shape of the entry/exit area and red, the region in which the highest number of inlets was identified.

hsEH, TrEH, and StEH1, the Tm1 tunnel was identified as the always open tunnel, while for bmEH and Sibe-EH, it was the Tc/m tunnel and for CH65-EH, the Tc/m_back tunnel. In the case of VrEH2, the most often open tunnel was the Tm1 tunnel; it was open for 89% of the simulation time. However, it is worth noting that in the case of msEH, StEH1, CH65-EH, TrEH, and hsEH the second most often tunnel is the Tc/m tunnel, which is open for 99%, 98%, 97%, 96%, and 59% of the simulation time, respectively (Table 1).

Tunnel Entrance/Exit Area Identification in MD Simulations by AQUA-DUCT. The same MD simulations were examined using AQUA-DUCT to identify the tunnel entrance/exit areas using the small-molecule tracking approach. AQUA-DUCT traced water molecules which entered and/or left the active site cavities of sEHs. Thus, it identified tunnels which were actually used by water molecules, which we will be referring to as the functional tunnels.

AQUA-DUCT identified one (in bmEH) to nine (in msEH) functional tunnels in the analyzed sEHs (Figure 2, Table 1). Tunnels were named using the previously established scheme regarding their exit location in the sEH structure. It should be noted that some tunnels found by CAVER do not have their functional counterparts identified by AQUA-DUCT. The opposite situation, when tunnels identified by the small-molecule tracking approach do not have their counterparts identified by CAVER, also occurred.

While CAVER 3.02 software provided the tunnel's occurrence, AQUA-DUCT provided the information on the number of inlets identified in each entrance/exit area. An inlet is a representation of the point in which a traced molecule entered or left the active site cavity. It can be assumed that the main tunnel will transport the highest number of water

molecules, i.e., will have the highest number of inlets. The distribution of inlets approximated the shape and size of the tunnel entry/exit area (Figure 2). Moreover, the intramolecular flow plot provided information regarding the water molecules' exchange and flow direction (Figure 2). According to AQUA-DUCT results, the sEHs can be divided into three groups: (i) in which the Tc/m1 tunnel was the main tunnel (bmEH, CH65-EH, and Sibe-EH), (ii) in which the Tm1 was the main tunnel (StEH1 and VrEH2), and (iii) in which the Tc/m1 and Tm1 were the main tunnels (msEH, hsEH, and TrEH).¹⁵

DISCUSSION

Computational identification of tunnels in proteins, based on their crystal structures, has been performed since the beginning of the 21st century.¹ With the introduction of MD simulations, this capability was soon extended.^{23,25} However, the vast majority of previously investigated tunnels have been described on the basis of solely crystal structures. While a crystal structure provides information on a single potential protein conformation, MD simulations provide a more detailed picture of protein motion and conformational changes. However, it should be kept in mind that the obtained picture depends on the force field and/or water models used during the MD simulation, which is out of the scope of our research. Importantly, the assessment of tunnel functionality still remains a nontrivial task due to several reasons, e.g., (i) variety of functions maintained by different tunnels in a protein (substrate entry, product egress, solvent accessibility control) and (ii) lack of a direct experimental method for small molecule transport assessment. Only indirect analyses can be provided, which require mutant design and kinetic studies

supported by advanced *in silico* methods as shown by Biedermannová et al.²⁶ The mentioned study suggests that results derived by advanced computational methods such as combination of Random Acceleration Molecular Dynamics (RAMD) and Adaptive Biasing Force (ABF) can be approximated by the geometry-based methods. So far, the performance of the recently developed small-molecule tracking method was not compared with commonly used approaches. We would like to point out that our research presents the first ever reported comparison of the geometry-based and small-molecule tracking approaches for tunnel identification in proteins. We hope that our results will help researchers to adequately select the method corresponding to their requirements and expectations. A comparison of the results obtained using both approaches on sEHs provided a systematic overview of the benefits and pitfalls of those methods.

Comparison of Results Obtained with Geometry-Based Approach in Crystal Structures and MD Simulations. Our results suggest that most tunnels identified during MD simulations have their counterparts in tunnels identified in crystal structures. However, closer inspection of the systems chosen for our study shows that reported tunnel shape and size can differ substantially in some cases (Figure S1). Differences may be attributable to packing inaccuracies or poor resolutions of crystal structures.^{27–30} Here, the structure with the poorest resolution (msEH, 2.80 Å resolution) also had the highest average difference measured between bottlenecks in corresponding tunnels identified in the crystal structure and during the MD simulation. Crystal structures with poor resolutions are prone to be inaccurate, especially within the most flexible regions, such as loops^{29,31} or gating residues within tunnels.^{3,5} Therefore, in some cases, a simple analysis of crystal structure leads to an incomplete picture of the enzyme's tunnel network.

CAVER software ranks the tunnels identified in crystal structures according to their priority, which is computed by averaging tunnel throughput, which is a measure of the cost of each pathway and can range from 0 (worst) to 1 (best). For tunnels identified during MD simulation, priority was averaged over all MD simulation frames in which a tunnel was identified. Analysis of the tunnel ranking showed no correlation between crystal structures and MD simulations. The differences in tunnel ranking between crystal structures and MD simulations may be associated with several factors, e.g., dense packing of the structures in crystals or multiple conformations that are accessible during MD simulation. However, comparison of the maximal bottleneck radii showed good correlation between the corresponding tunnels in crystal structures and MD simulations ($r = 0.82$) (Figure S1). On average, the difference between the measured bottleneck radii of the crystal structure was smaller than the bottleneck measured during the MD simulation by about 0.7 Å.

A high correlation between measured bottlenecks in corresponding tunnels from the crystal structure and MD simulations may suggest that the shape and size of the tunnels present in a crystal structure are preserved despite the potential conformational changes which may affect overall protein structure. However, closer analysis of the tunnels identified in crystal structures and during MD simulations by CAVER showed that the tunnels identified in crystal structures are well-defined; however, their parts located closer to the protein surface are, in some cases, coiled. For most tunnels identified during MD simulations, the interior parts of tunnels were well-

defined, whereas the tunnels' mouths were widely distributed on the protein surface. Such an observation might suggest that those regions are tightly packed and/or lined by bulky residues, which can change their conformation to open/close a particular tunnel. Therefore, we recommend tunnel identification using MD simulations instead of a single crystal structure. However, the geometry-based approach has issues related to the asymmetrical shape of the tunnel: multiple tunnels identified by CAVER during MD simulations may in fact be the same tunnel, as it was shown in the case of the Tc/m tunnel (Figure S2). Part of the tunnels can be seen as short-lasting cavities, which rarely connect with other internal voids,¹⁷ and as such, they are difficult to identify using the geometry-based approach.

Comparison of the Results Obtained with Geometry-Based and Small-Molecule Tracking Approaches from MD Simulations. Using the same MD simulations as an input for two different approaches provided an opportunity to compare the results. While CAVER was developed to find all possible entrances to the enzymes' active sites, defined as a space accessible for the probe with a defined size in particular frames, AQUA-DUCT is focused on the tunnel's functionality, defined as the ability of a tunnel (or cavity) to transport small molecules of interest. However, we observed that both tools were able to identify the main tunnels (the most often open/the most used by water molecules) in the analyzed sEHs. The difference between the approaches is more visible when comparing the side tunnels (rarely open/used by less water molecules). Here, we would like to point out that the aim of the study was not only to compare both approaches but also to equip the user with a set of guidelines on how to carefully interpret the information on the tunnel network provided by each tool. We noticed that in several cases AQUA-DUCT was unable to detect tunnels identified by the geometry-based approach in both crystal structure and during MD simulation. This may be caused by the physicochemical properties of the tunnel-lining residues, which could block the transport of particular molecules via the selected pathway. According to our analysis, such nonfunctional tunnels were rather common, not rare, cases. They were found by CAVER 3.0 PyMOL plugin and CAVER 3.02 in seven out of eight analyzed sEHs (all except msEH). Such tunnels may not be used for the transport of small molecules; however, their modification can lead to improved (thermo)stability of the protein.³² We also noticed tunnels which were identified by AQUA-DUCT but not by the geometry-based approach. At first glance, such a finding for MD simulation analysis is unexpected because of the effective radius of the water molecule, which was bigger than the probe used in our investigation. However, this can be observed due to two factors: (i) "rare events" or "water leakage" and (ii) clustering algorithm. Rare events were previously discussed in the case of StEH1.³³ Rare events can be identified by AQUA-DUCT even during relatively short MD simulations (50 ns), but their identification by CAVER may be challenging. When a water molecule is transported from one internal cavity to another during the course of an MD simulation, it can leak through the protein region, which is equipped with a set of connected cavities and not a permanent tunnel (which is a must for a geometry-based approach). Protein motions can promote molecule passage, and therefore, longer or advanced sampling MD simulations need to be performed to detect such a void continuum by a geometry-based approach. Another option is to use a smaller probe, which will make computations

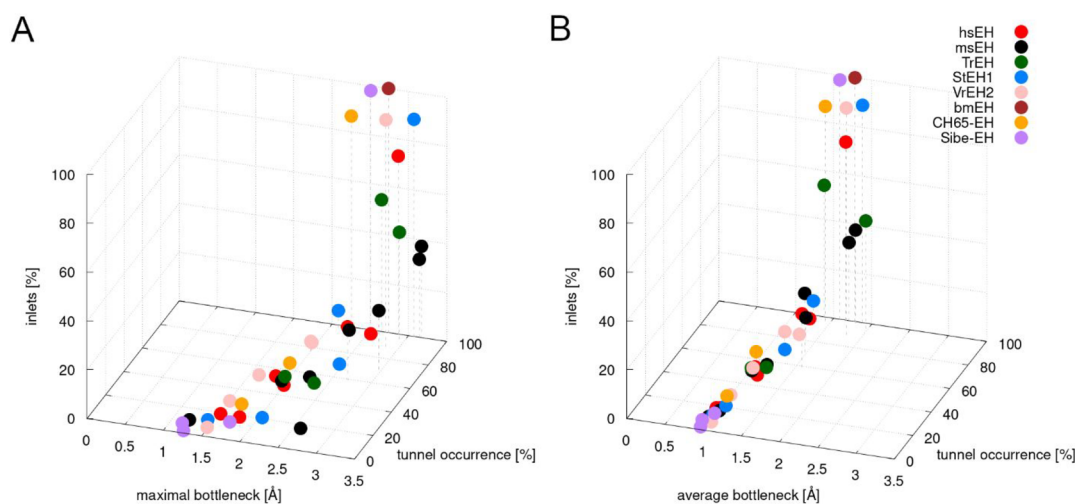


Figure 3. Relationship between number of inlets, tunnel occurrence, and average and maximal bottleneck obtained for all analyzed proteins. Please note that even those tunnels which were always open and had wide bottlenecks were not always identified as functional tunnels.

longer and analysis more challenging. The clustering approach used by CAVER searched for the similarities between detected pathways, and thus, it omitted the rarely occurring tunnels. Therefore, we recommend a careful analysis of the clustering results. It is worth noting that tunnels which were identified by AQUA-DUCT and not by CAVER 3.02 during MD simulations may be used for modifying an identified “rare event” tunnel whose opening may lead to improved protein activity^{34–36} or selectivity.^{37–40} Importantly, AQUA-DUCT is designed to track all types of small molecules,^{11,24} such as water molecules, ions, ligands, and additional cosolvents, such as methanol, urea, dimethyl sulfoxide, acetonitrile, or phenol (the use of AQUA-DUCT for the analysis of cosolvents was shown by Bzówka et al.⁴¹). For this study, we used water molecules; because of the analyzed system, sEHs require a catalytic water molecule to convert substrates to product(s), and therefore, they should be equipped in tunnels maintaining water molecule transport. However, in the case of proteins whose tunnels transport hydrophobic molecules, such as AlkL, which was proposed to facilitate a passive transport function increasing the rate of alkane diffusion,⁴² using a hydrophobic probe would be recommended.

A comparison of the geometry-based and the small-molecule tracking approaches also showed other differences. Tunnels identified by the small-molecule tracking approach can be considered functional for particular types of traced molecules. The number of molecules passing through a particular entry/exit reflects the tunnels’ usability, whereas in the case of tunnels identified by the geometry-based approach, we were unable to determine the tunnels’ functionality. Taking into account all analyzed sEHs, no correlation between the number of inlets and average bottleneck radius, maximal bottleneck radius, or tunnel occurrence was found (Figure 3). However, for particular enzymes, such correlation can be observed. In msEH, both Tc/m and Tm1 tunnels have similar bottleneck radii and occurrence according to the CAVER 3.02 results (Table 1), and AQUA-DUCT showed that they are used to transport 38% and 32% of the identified water molecules, respectively. Interestingly, the Tg tunnel, which was not present in the crystal structure, also has a similar bottleneck radius but occurred only in 71% of the simulation time. The Tg tunnel was used to transport around 24% of the identified

water molecules. All above-mentioned tunnels can be considered similar. In contrast, in StEH1, the Tm1 and Tm2 tunnels which have similar maximal bottlenecks (>2.5 Å) differ substantially in terms of functionality: Tm1 was used by 89% of water molecules and Tm2, by 3%. On the other hand, the Tm1 and Tc/m1 tunnels were almost always open, and they also differed in terms of their functionality (Tc/m1 was used by 7% of water molecules).

In this study, we compared the software for tunnel identification, namely, the geometry-based approach of CAVER 3.02 and small-molecule tracking approach of AQUA-DUCT. We used the crystal structures and a set of MD simulation trajectories to compare the results. Moreover, we wanted to raise awareness among the users of tunnel identification software that the geometry-based approach has its flaws, which may be overcome or supplemented by using the small-molecule tracking approach. Even though the main tunnels in the analyzed proteins seem to be described in a comparable way by both approaches, the results differ when it comes to the side tunnels, which can be of great importance for catalysis. Those differences are related to the way in which both approaches search for tunnels. However, it must be kept in mind that we were analyzing tunnels with relatively narrow bottlenecks (1.0–2.0 Å radii) in which a subtle conformational change may cause opening or closing of a particular pathway. Therefore, the described differences between the approaches may not be observable in wider tunnels and channels. We also showed that MD simulations provide much more information on the tunnels and protein dynamics. The small-molecule tracking approach was shown to solve some limitations of the geometry-based approach; however, in some aspects, both approaches are complementary and may be useful for further protein engineering. Because tunnel detection in the crystal structures using the geometry-based approach is easier compared with other approaches using MD simulation data, it may be the most commonly used. We hope that our work will increase awareness among researchers using a geometry-based approach about its limitations and will provide a guide for the selection of methods according to their needs.

DATA AND SOFTWARE AVAILABILITY

Tunnel identification in the crystal structures was carried out using a CAVER 3.0 plugin for PyMOL.²³ Parameters used for the analysis are specified in Table S1. The classical MD simulations of each protein were carried out using the CUDA version of the pmemd program available in Amber14.²¹ Tunnel identification during MD simulations was performed by CAVER 3.02 software²³ using the same set of parameters as for the analysis in the crystal structures (as in Table S1) and using AQUA-DUCT 1.0 version.^{11,24} For AQUA-DUCT analysis, the water molecules which entered and/or left the active site cavity (called the Object) were traced within the protein's interior (called the Scope).

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.2c00985>.

Number of added ions and water molecules along with the protonation pH and the duration of the equilibration step for each of the analyzed systems; list of parameters set; comparison of tunnels identified with the geometry-based approach in both crystal structure and during MD simulation for all analyzed proteins; correlation between maximal bottleneck radii measured in corresponding tunnels identified in both the crystal structure and in the MD simulation for each protein structure; comparison between Tc/m tunnels identified by CAVER 3.02 software during MD simulations and the cluster of inlets identified by AQUA-DUCT (PDF)

AUTHOR INFORMATION

Corresponding Author

Artur Góra – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland;

orcid.org/0000-0003-2530-6957; Email: a.gora@tunnelinggroup.pl

Authors

Karolina Mitusińska – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland

Maria Bzówka – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland;

orcid.org/0000-0001-6802-8753

Tomasz Magdziarz – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jcim.2c00985>

Author Contributions

[‡]K.M. and M.B. contributed equally.

Funding

This work was supported by the National Science Centre, Poland (grant number DEC-2013/10/E/NZ1/00649).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors would like to thank Agata Raczyńska, Weronika Bagrowska, Aleksandra Samol, and Piotr Wojsa for their help

in preparing the files and running MD simulations for four proteins.

ABBREVIATIONS

sEH, soluble epoxide hydrolase; hsEH, human soluble epoxide hydrolase; msEH, mouse soluble epoxide hydrolase; TrEH, *Trichoderma reesei* soluble epoxide hydrolase; StEH1, *Solanum tuberosum* soluble epoxide hydrolase; VrEH2, *Vigna radiata* soluble epoxide hydrolase; bmEH, *Bacillus megaterium* soluble epoxide hydrolase; Sibe-EH, soluble epoxide hydrolase from hot springs in Russia; CH65-EH, soluble epoxide hydrolase from hot springs in China; MD, molecular dynamics; PDB, Protein Data Bank

REFERENCES

- (1) Brezovsky, J.; Chovancova, E.; Gora, A.; Pavelka, A.; Biedermannova, L.; Damborsky, J. Software Tools for Identification, Visualization and Analysis of Protein Tunnels and Channels. *Biotechnol. Adv.* **2013**, *31*, 38–49.
- (2) Prokop, Z.; Gora, A.; Brezovsky, J.; Chaloupkova, R.; Stepankova, V.; Damborsky, J. Engineering of Protein Tunnels: Keyhole-Lock-Key Model for Catalysis by Enzymes with Buried Active Sites. In *Protein Engineering Handbook*; Lutz, S., Bornscheuer, U. T., Eds.; Wiley-VCH: Weinheim, 2012; Vol. 3, pp 421–464.
- (3) Gora, A.; Brezovsky, J.; Damborsky, J. Gates of Enzymes. *Chem. Rev.* **2013**, *113*, 5871–5923.
- (4) Kingsley, L. J.; Lill, M. A. Substrate Tunnels in Enzymes: Structure-Function Relationships and Computational Methodology. *Proteins Struct. Funct. Bioinforma.* **2015**, *83*, 599–611.
- (5) Marques, S. M.; 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.
- (6) Pravda, L.; Berka, K.; Svobodová Vařeková, R.; Sehnal, D.; Banáš, P.; Laskowski, R. A.; Koča, J.; Otyepka, M. Anatomy of Enzyme Channels. *BMC Bioinformatics* **2014**, *15*, 379.
- (7) Pavelka, A.; Sebestova, E.; Kozlikova, B.; Brezovsky, J.; Sochor, J.; Damborsky, J. CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **2016**, *13*, 505–517.
- (8) Urban, P.; Lautier, T.; Pompon, D.; Truan, G. Ligand Access Channels in Cytochrome P450 Enzymes: A Review. *Int. J. Mol. Sci.* **2018**, *19*, 1617.
- (9) Rydzewski, J.; Nowak, W. Ligand Diffusion in Proteins via Enhanced Sampling in Molecular Dynamics. *Phys. Life Rev.* **2017**, *22*–23, 58–74.
- (10) Magdziarz, T.; Mitusińska, K.; Goldowska, S.; Phuciennik, A.; Stolarczyk, M.; Ługowska, M.; Góra, A. AQUA-DUCT: A Ligands Tracking Tool. *Bioinformatics* **2017**, *33*, 2045–2046.
- (11) 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*, 2599–2601.
- (12) Nardini, M.; Dijkstra, B. W. α/β Hydrolase Fold Enzymes: The Family Keeps Growing. *Curr. Opin. Struct. Biol.* **1999**, *9*, 732–737.
- (13) Marchot, P.; Chatonnet, A. Enzymatic Activity and Protein Interactions in Alpha/Beta Hydrolase Fold Proteins: Moonlighting Versus Promiscuity. *Protein Pept. Lett.* **2012**, *19*, 132–143.
- (14) Bauer, T. L.; Buchholz, P. C. F.; Pleiss, J. The Modular Structure of α/β -hydrolases. *FEBS J.* **2020**, *287*, 1035–1053.
- (15) 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. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 193–205.
- (16) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.

- (17) Bzówka, M.; Mitusińska, K.; Raczyńska, A.; Skalski, T.; Samol, A.; Bagrowska, W.; Magdziarz, T.; Góra, A. Evolution of Tunnels in α/β -Hydrolase Fold Proteins—What Can We Learn from Studying Epoxide Hydrolases? *PLoS Comput. Biol.* **2022**, *18*, No. e1010119.
- (18) Anandakrishnan, R.; Aguilar, B.; Onufriev, A. V. H++ 3.0: Automating PK Prediction and the Preparation of Biomolecular Structures for Atomistic Molecular Modeling and Simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541.
- (19) Luchko, T.; Gusarov, S.; Roe, D. R.; Simmerling, C.; Case, D. A.; Tuszynski, J.; Kovalenko, A. Three-Dimensional Molecular Theory of Solvation Coupled with Molecular Dynamics in Amber. *J. Chem. Theory Comput.* **2010**, *6*, 607–624.
- (20) Sindhikara, D. J.; Yoshida, N.; Hirata, F. Placevent: An Algorithm for Prediction of Explicit Solvent Atom Distribution-Application to HIV-1 Protease and F-ATP Synthase. *J. Comput. Chem.* **2012**, *33*, 1536–1543.
- (21) Case, D. A.; Babin, V.; Berryman, J. T.; Betz, R. M.; Cai, Q.; Cerutti, D. S.; Cheatham, T. E., III; Darden, T. A.; Duke, R. E.; Gohlke, H.; Goetz, A. W.; Gusarov, S.; Homeyer, N.; Janowski, P.; Kaus, J.; Kolossváry, I.; Kovalenko, A.; Lee, T. S.; LeGrand, S.; Luchko, T.; Luo, R.; Madej, B.; Merz, K. M.; Paesani, F.; Roe, D. R.; Roitberg, A.; Sagui, C.; Salomon-Ferrer, R.; Seabra, G.; Simmerling, C. L.; Smith, W.; Swails, J.; Walker, R. C.; Wang, J.; Wolf, R. M.; Wu, X.; Kollman, P. A. *AMBER14*; University of California: San Francisco, 2014.
- (22) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. Ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.
- (23) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **2012**, *8*, No. e1002708.
- (24) 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*, 1.
- (25) Sehnal, D.; Svobodová Vařeková, R.; Berka, K.; Pravda, L.; Navrátilová, V.; Banáš, P.; Ionescu, C.-M.; Otyepka, M.; Koča, J. MOLE 2.0: Advanced Approach for Analysis of Biomacromolecular Channels. *J. Cheminform.* **2013**, *5*, 39.
- (26) Biedermannová, L.; Prokop, Z.; Gora, A.; Chovancová, E.; Kovács, M.; Damborský, J.; Wade, R. C. A Single Mutation in a Tunnel to the Active Site Changes the Mechanism and Kinetics of Product Release in Haloalkane Dehalogenase LinB. *J. Biol. Chem.* **2012**, *287*, 29062–29074.
- (27) DePristo, M. A.; de Bakker, P. I.; Blundell, T. L. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. *Structure* **2004**, *12*, 831–838.
- (28) Furnham, N.; Blundell, T. L.; DePristo, M. A.; Terwilliger, T. C. Is One Solution Good Enough? *Nat. Struct. Mol. Biol.* **2006**, *13*, 184–185.
- (29) Burra, P. V.; Zhang, Y.; Godzik, A.; Stec, B. Global Distribution of Conformational States Derived from Redundant Models in the PDB Points to Non-Uniqueness of the Protein Structure. *Proc. Natl. Acad. Sci. U. S. A.* **2009**, *106*, 10505–10510.
- (30) Lamb, A. L.; Kappock, T. J.; Silvaggi, N. R. You Are Lost without a Map: Navigating the Sea of Protein Structures. *Biochim. Biophys. Acta - Proteins Proteomics* **2015**, *1854*, 258–268.
- (31) Djinovic-Carugo, K.; Carugo, O. Missing Strings of Residues in Protein Crystal Structures. *Intrinsically Disord. Proteins* **2015**, *3*, No. e1095697.
- (32) Damborsky, J.; Prokop, Z.; Koudelakova, T.; Stepankova, V.; Chaloupkova, R.; Chovancova, E.; Gora, A. W.; Brezovsky, J. Method of Thermostabilization of a Protein and/or Stabilization towards Organic Solvents. US 20130102763 A1, April 25, 2013.
- (33) 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.
- (34) Kong, X.-D.; Yuan, S.; Li, L.; Chen, S.; Xu, J.-H.; Zhou, J. Engineering of an Epoxide Hydrolase for Efficient Bioresolution of Bulky Pharmacological Substrates. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 15717–15722.
- (35) Hamre, A. G.; Frøberg, E. E.; Eijsink, V. G. H.; Sørli, M. Thermodynamics of Tunnel Formation upon Substrate Binding in a Processive Glycoside Hydrolase. *Arch. Biochem. Biophys.* **2017**, *620*, 35–42.
- (36) Chaplin, V. D.; Valliere, M. A.; Hangasky, J. A.; Knapp, M. J. Investigations on the Role of a Solvent Tunnel in the α -Ketoglutarate Dependent Oxygenase Factor Inhibiting HIF (FIH). *J. Inorg. Biochem.* **2018**, *178*, 63–69.
- (37) Subramanian, K.; Mitusińska, K.; Raedts, J.; Almourfi, F.; Joosten, H.-J.; Hendriks, S.; Sedelnikova, S. E.; Kengen, S. W. M.; Hagen, W. R.; Góra, A.; Martins dos Santos, V. A. P.; Baker, P. J.; van der Oost, J.; Schaap, P. J. Distant Non-Obvious Mutations Influence the Activity of a Hyperthermophilic Pyrococcus Furiosus Phosphoglucose Isomerase. *Biomolecules* **2019**, *9*, 212.
- (38) Kokkonen, P.; Bednar, D.; Pinto, G.; Prokop, Z.; Damborsky, J. Engineering Enzyme Access Tunnels. *Biotechnol. Adv.* **2019**, *37*, 107386.
- (39) Li, G.; Yao, P.; Gong, R.; Li, J.; Liu, P.; Lonsdale, R.; Wu, Q.; Lin, J.; Zhu, D.; Reetz, M. T. Simultaneous Engineering of an Enzyme's Entrance Tunnel and Active Site: The Case of Monoamine Oxidase MAO-N. *Chem. Sci.* **2017**, *8*, 4093–4099.
- (40) Gorskova, I. N.; Mei, X.; Atkinson, D. Arginine 123 of Apolipoprotein A-I Is Essential for Lecithin:Cholesterol Acyltransferase Activity. *J. Lipid Res.* **2018**, *59*, 348–356.
- (41) Bzówka, M.; Mitusińska, K.; Raczyńska, A.; Samol, A.; Tuszynski, J. A.; 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*, 3099.
- (42) Schubeis, T.; Le Marchand, T.; Daday, C.; Kopec, W.; Tekwani Movellan, K.; Stanek, J.; Schwarzer, T. S.; Castiglione, K.; de Groot, B. L.; Pintacuda, G.; Andreas, L. B. A β -Barrel for Oil Transport through Lipid Membranes: Dynamic NMR Structures of AlkL. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117*, 21014–21021.


RESEARCH ARTICLE

Evolution of tunnels in α/β -hydrolase fold proteins—What can we learn from studying epoxide hydrolases?

Maria Bzówka¹ , Karolina Mitusińska¹ , Agata Raczyńska¹ , Tomasz Skalski² , Aleksandra Samol¹ , Weronika Bagrowska¹ , Tomasz Magdziarz¹ , Artur Góra^{1*} 

1 Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice, Poland,

2 Biotechnology Centre, Silesian University of Technology, Gliwice, Poland

 These authors contributed equally to this work.

* a.gora@tunnelinggroup.pl


 OPEN ACCESS

Citation: Bzówka M, Mitusińska K, Raczyńska A, Skalski T, Samol A, Bagrowska W, et al. (2022) Evolution of tunnels in α/β -hydrolase fold proteins—What can we learn from studying epoxide hydrolases? PLoS Comput Biol 18(5): e1010119. <https://doi.org/10.1371/journal.pcbi.1010119>

Editor: Marco Punta, San Raffaele Hospital: IRCCS Ospedale San Raffaele, ITALY

Received: September 25, 2021

Accepted: April 19, 2022

Published: May 17, 2022

Copyright: © 2022 Bzówka et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting Information](#) files.

Funding: This work was supported by the National Science Centre, Poland (grant number DEC-2013/10/E/NZ1/00649) (AG, MB, TM, KM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Abstract

The evolutionary variability of a protein's residues is highly dependent on protein region and function. Solvent-exposed residues, excluding those at interaction interfaces, are more variable than buried residues whereas active site residues are considered to be conserved. The abovementioned rules apply also to α/β -hydrolase fold proteins—one of the oldest and the biggest superfamily of enzymes with buried active sites equipped with tunnels linking the reaction site with the exterior. We selected soluble epoxide hydrolases as representative of this family to conduct the first systematic study on the evolution of tunnels. We hypothesised that tunnels are lined by mostly conserved residues, and are equipped with a number of specific variable residues that are able to respond to evolutionary pressure. The hypothesis was confirmed, and we suggested a general and detailed way of the tunnels' evolution analysis based on entropy values calculated for tunnels' residues. We also found three different cases of entropy distribution among tunnel-lining residues. These observations can be applied for protein reengineering mimicking the natural evolution process. We propose a 'perforation' mechanism for new tunnels design *via* the merging of internal cavities or protein surface perforation. Based on the literature data, such a strategy of new tunnel design could significantly improve the enzyme's performance and can be applied widely for enzymes with buried active sites.

Author summary

So far very little is known about proteins tunnels evolution. The goal of this study is to evaluate the evolution of tunnels in the family of soluble epoxide hydrolases—representatives of numerous α/β -hydrolase fold enzymes. As a result two types of tunnels evolution analysis were proposed (a general and a detailed approach), as well as a 'perforation' mechanism which can mimic native evolution in proteins and can be used as an additional strategy for enzymes redesign.

Introduction

Protein evolution mechanisms, and the factors determining protein evolution rate, have drawn attention in the past decades. Comprehensive studies regarding protein evolution resulted in a set of principles linking protein evolution with their structural and functional features. The most crucial assumption is that functionally important residues evolve at slower rates compared with the less important residues [1]. Moreover, residues buried in the protein core and those on the protein surface were shown to have different substitution patterns [2], which may be related to different packing densities in the macromolecule [3]. These findings provided the groundwork for various experimental techniques [4] and bioinformatic tools used intensively to carry out protein engineering [5,6] to search for particular ancestral proteins [7–9], and to explore the evolution of enzyme functions within superfamilies [10]. The distinction between residues that evolve more slowly or more quickly (i.e. conserved and variable residues, respectively) can be used to inform preselection of target regions for function or stability improvement, and in the design of smart libraries, while also providing explanations for unsuccessful attempts which resulted in dysfunctional or unstable mutants [11–15].

The amino acids comprising an enzyme's catalytic site (regardless of its location) represent one of the most evident examples of conserved residues. In contrast, solvent-exposed residues which do not contribute to protein-protein or protein-ligand recognition are more variable, since they are not essential for either the enzyme's function nor its structural stability [3,16,17]. The evolutionary rate of secondary structure elements has also been investigated by several research groups. In a study by Sitbon and Pietrokovski [18], the authors suggest that, due to their regular repetitive structure, helices and strands might be more conserved than loops. On the other hand, Liu *et al.* showed that loops might tend to be evolutionary conserved since functional sites are overrepresented by loop-rich regions [19]. However, other results suggest that β -sheet regions evolve more slowly in comparison to helical regions, and that random coil regions evolve the fastest [3,18,20,21].

Meanwhile, the results of site-directed mutagenesis experiments demonstrated that even mutations positioned relatively far from catalytic residues can attenuate an enzyme's catalytic activity [22,23]. However, frequently distal mutations are fine-tuning the conformational ensembles of enzymes by evolutionary conformational selection [24,25] but that approach can also modify the allosteric mechanism of an enzyme [26,27], or its tunnel utilised to maintain ligands transport [28,29]. Growing evidence of a large number of tunnels in protein structures [28,30] and their importance for an enzyme's catalytic performance has led to the assumption that, while respecting evolutionary pressures, tunnels are generally preserved during protein evolution. So far, only a few individual studies have addressed this question. Evolutionarily preserved tunnels, or their parts, were reported in glutamine amidotransferases [31], carbamoyl phosphate synthetase [32], and histone deacetylases [33]. In contrast, a faster rate of evolution was proposed for residues constituting gates in cytochromes [34].

Limited information about the variability of the tunnel-lining residues encouraged us to perform the first systematic study on the determination of a tunnel's evolution in the soluble epoxide hydrolases (sEHs) family. We chose representative members of the sEHs due to three facts: i) that they belong to one of the oldest and the biggest enzymes superfamily—the α/β -hydrolases fold family [35–37], ii) that the crystal structures of different clade members (mammals, plants, fungi, and bacteria) were available, and iii) that sEHs catalyse the conversion of a broad spectrum of substrates and exhibit a diverse tunnel network in their structures. Such a tunnel network connects the conserved active site buried between the main and more structurally variable cap domain with the environment. We hypothesised that tunnels are conserved structural features equipped with variable parts, e.g. gates responsible for different substrate

specificity in closely related family members. Additionally, we raised the following question: are there any mechanisms or schemes that can be adopted during protein engineering to mimic new tunnels' appearance? Our results indicate that most tunnels in soluble epoxide hydrolases can be considered as conserved features, and we have proposed a “perforation” model that can be applied as a strategy for *de novo* tunnel design. Due to high structural similarity between members of α/β -hydrolases superfamily, our results could be expanded and applied into other superfamily members including acetylcholinesterase, dienelactone hydrolase, lipase, thioesterase, serine carboxypeptidase, proline iminopeptidase, proline oligopeptidase, haloalkane dehalogenase, haloperoxidase, epoxide hydrolase, hydroxynitrile lyase and others [38]. We need to emphasise that since we analysed tunnels identified in relatively small protein structures with narrow tunnels (usually 1.0–2.0 Å), some processes leading to tunnel formation or modification cannot be covered. This includes long insertion or deletion, dimerization, or quaternary protein structure organisation.

Results

For this study, we chose only the unique and complete structures of sEHs deposited in the Protein Data Bank (PDB) [39]. Any structures with information missing about the positions of any of their amino acid residues could have provided bias, and therefore were excluded. The resulting selection of seven epoxide hydrolase structures represent the clades of animals (*Mus musculus*, msEH, PDB ID: 1CQZ, and *Homo sapiens*, hsEH, PDB ID: 1S8O), plants (*Solanum tuberosum*, StEH1, PDB ID: 2CJP), fungi (*Trichoderma reesei*, TrEH, PDB ID: 5URO), bacteria (*Bacillus megaterium*, bmEH, PDB ID: 4NZZ) and thermophilic enzymes collected in hot springs in Russia and China from an unknown source organism (Sibe-EH, and CH65-EH, PDB IDs: 5NG7, and 5NFQ, respectively).

Model description and referential compartment evolutionary analysis

sEHs consist of two domains: the main domain, featuring eight β -strands surrounded by six α -helices; and the mostly helical cap domain, which sits atop the main domain. The cap domain is inserted between the strands of the main domain and is connected by an element called the NC-loop. The cap-loop is inserted between two helices of the cap domain [40]. The active site of the sEHs is buried inside the main domain, and therefore the transportation of substrates and products is facilitated by tunnel (either single or in a network) [29].

We performed an entropy analysis of the residues making up particular protein compartments with the use of the Schneider entropy metric implemented in the BALCONY package [41]. As an input BALCONY requires multiple sequence alignment (MSA) and a list of residues building up particular compartments. We analysed the compartments listed in **S1 Table** (i.e. residues forming the active site; buried and surface residues; main and cap domains; NC-loop; cap-loop; and α -helices, loops, and β -strands). In order to determine the positions' variability, we used Schneider entropy metric [42] calculated for each position in the MSA. To avoid bias and position-specific conservation scores we trimmed the MSA removing positions that did not correspond to the analysed proteins' sequences. To evaluate the overall compartments' variability we calculated the difference between the median distances of positions of the proteins' compartments and the remaining positions of the trimmed MSA (**Fig 1 and S2 Table**, see also **Methods** section for the description of the MSA trimming). Negative values of the difference between median distances of the selected proteins' compartments and the trimmed MSA (**S2 Table**) indicate compartments with lower variability, and positive values indicate compartments with higher variability in comparison to the remaining positions in the trimmed MSA. For quantitative statistical analysis, we compared the calculated Schneider

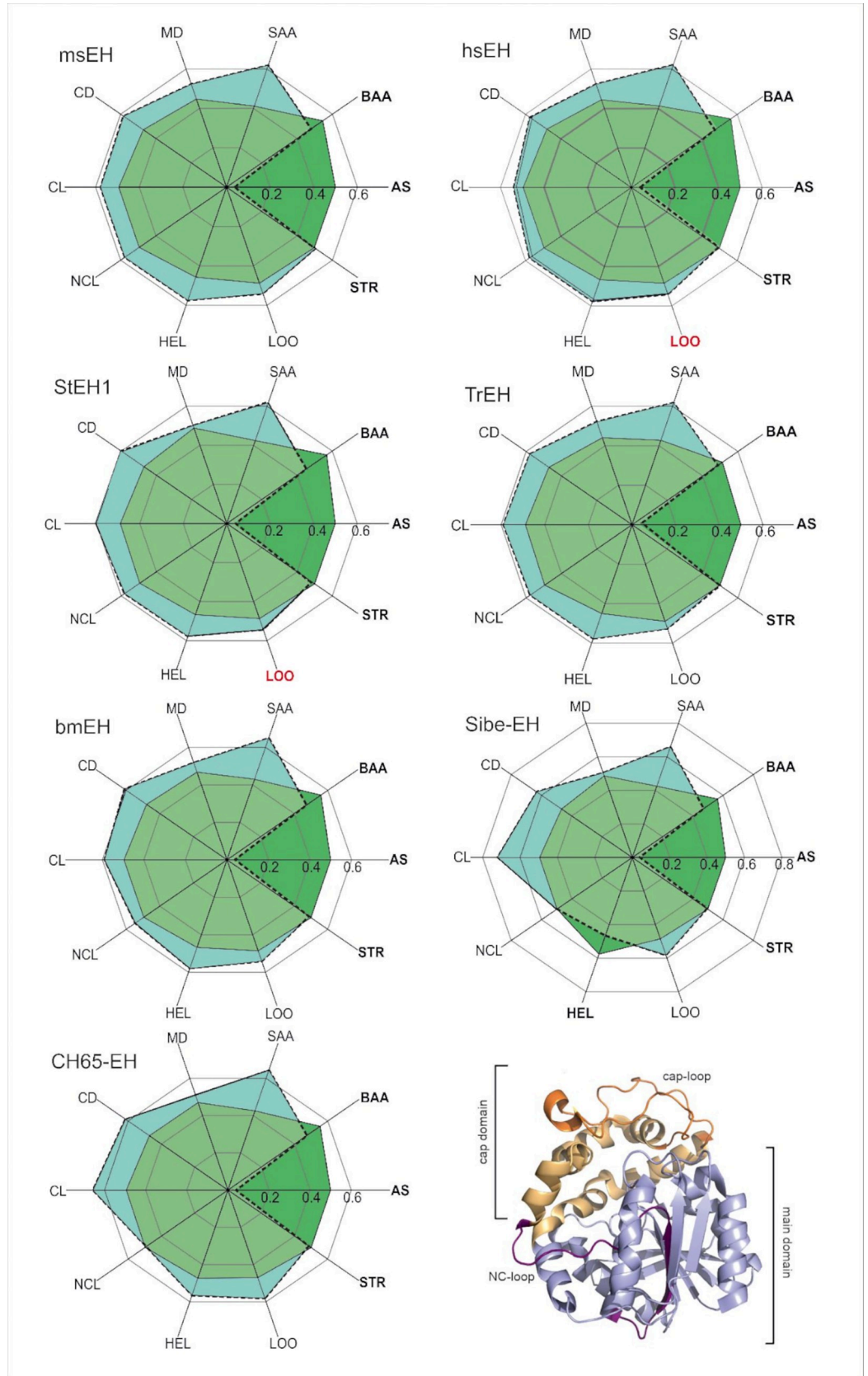


Fig 1. Radial plot of the median entropy values of referential compartments (green) and the remaining positions of the trimmed MSA (turquoise). When the median entropy values of the components cover the median entropy values of the trimmed MSA, it means that the particular compartment is more conserved than the remaining positions of the MSA (dark green). The compartments considered as conserved are written in bold. The MSA contained 1455 sequences and 419 positions. Figure represents data shown in [S2 Table](#). All pairwise differences (except for loops (LOO) in hsEH and StEH1, marked red) are statistically significant (Epps-Singleton test). In the bottom right corner, a schematic representation of the analysed structure-specific compartments is provided. Abbreviations: AS—active site; BAA—buried amino acids; SAA—surface amino acids; MD—main domain; CD—cap domain; CL—cap-loop; NCL—NC-loop; HEL—helices; LOO—loops; STR—strands.

<https://doi.org/10.1371/journal.pcbi.1010119.g001>

entropy values of these compartments with the remaining positions of the trimmed MSA using the Epps-Singleton test [43].

Based on the obtained differences in median distances and the results of the Epps-Singleton test, the active site residues were classified as conserved, i.e. with lower entropy scores in comparison to the remaining positions in the MSA. The surface residues (classified as solvent-exposed residues according to the NetSurfP server [44]) were classed as the most variable. Entropy analysis showed that the variability of the buried residues was significantly lower than the variability of the surface residues (**Fig 1 and S2 Table**). These results are in agreement with the general findings mentioned previously [3,16,45]. With regards to the structural elements specific to sEHs, all compartments (main domain, cap domain, cap-loop, and NC-loop (except for the NC-loop in CH65-EH)) were classified as variable among all the selected sEHs. In all analysed proteins, α -helices and loops were also classified as variable (however, in the case of hsEH and StEH1 the information about the variability of loops was not statistically significant). In all analysed proteins, except for msEH, β -strands were found to be conserved which stays in agreement with the work of Sitbon and Pietrokovski [18] (**Fig 1 and S2 Table**).

Tunnel identification and comparison

We identified tunnels providing access to the active site using a geometry-based approach implemented in CAVER software [46] for both crystal structures and in molecular dynamics (MD) simulations, and then compared their geometries (for details see the [Methods](#) section). CAVER software identified between three and nine tunnels in the analysed crystal structures. Those tunnels were then compared with the tunnels identified during MD simulations to find their corresponding counterparts ([S3 Table](#)), based on the similarity of their tunnel-lining residues (for more details see the [Methods](#) section). We marked all identified tunnels according to their localisation within the epoxide hydrolase's domains as was shown in our previous work [47]. We identified tunnels passing through three regions of the sEH structure: i) the main domain (marked as Tg, Tm, Tback, and Tside), ii) the cap domain (marked as Tcap), as well as iii) the border between the cap and main domains (marked as Tc/m).

We identified seven tunnels in the main domain, six in the cap domain, and three at the border between those domains (**Fig 2**). It should be pointed out that the Tc/m tunnel was identified as multiple tunnels by CAVER (Tc/m1, Tc/m2, and Tc/m3). This issue is related to the asymmetric shape of the Tc/m tunnel, which makes it difficult to classify in a geometry-based approach ([S1 Fig](#)).

Closer analysis of the tunnels identified in crystal structures and during MD simulations by CAVER showed that the tunnels identified in crystal structures are well-defined; however, their parts located closer to the protein surface are, in some cases, coiled. For most tunnels identified during MD simulations, the interior parts of tunnels were well-defined, whereas the tunnels' mouths were widely distributed on the protein surface. Such an observation might suggest that those regions are tightly packed and/or lined by bulky residues which can change their conformation to open/close a particular tunnel.

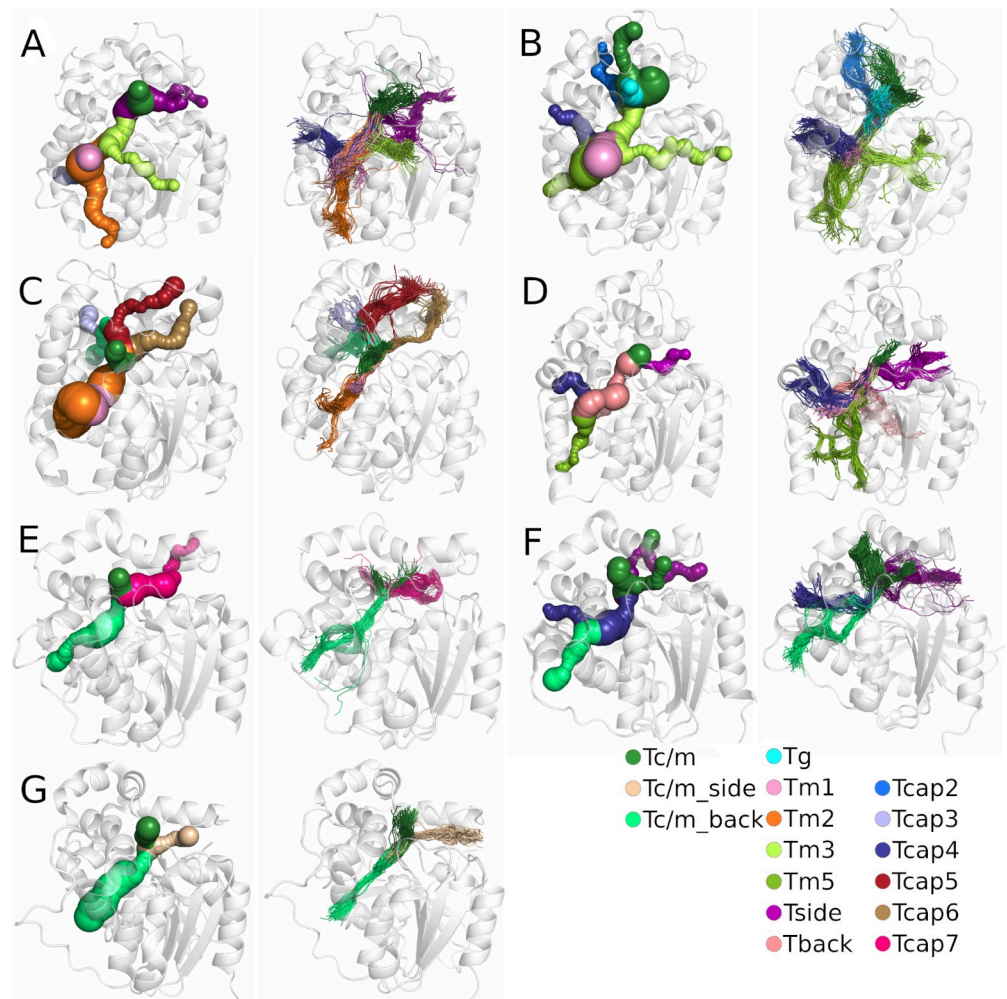


Fig 2. Comparison of tunnels identified in the crystal structure (left) and the results after molecular dynamics (MD) simulation (right) for each system: A) *M. musculus* soluble epoxide hydrolase (msEH), B) *H. sapiens* soluble epoxide hydrolase (hsEH), C) *S. tuberosum* soluble epoxide hydrolase (StEH1), D) *T. reesei* soluble epoxide hydrolase (TrEH), E) *B. megaterium* soluble epoxide hydrolase (bmEH), and thermophilic soluble epoxide hydrolases from an unknown source organism F) Sibe-EH, and G) CH65-EH. Protein structures are shown as white transparent cartoons. Matching tunnels are marked with the same colour as spheres (in crystal structures) and lines (in MD simulations).

<https://doi.org/10.1371/journal.pcbi.1010119.g002>

Tunnel evolutionary analysis

In the case of sEHs, tunnels can perform several distinct functions: i) transport and positioning of substrates and products, ii) control of the solvent access to the catalytic cavity, and iii) transport of catalytic water. Only those tunnels which maintain at least one of those functions can undergo evolutionary pressure. As we confirmed during the referential compartments' evolutionary analysis, surface residues are more variable than buried residues. Indeed, Fig 3 shows protein structures coloured according to Schneider entropy values, where thin blue lines represent regions with lower entropy, and yellow thick lines represent regions with higher entropy values. We also coloured the identified tunnels according to their frequency of detection (i.e. based on the number of frames in which they were identified) in MD simulations (darker = more frequent). The overall position of the tunnels was similar among all the protein structures; however, there were large differences concerning their frequency during the MD

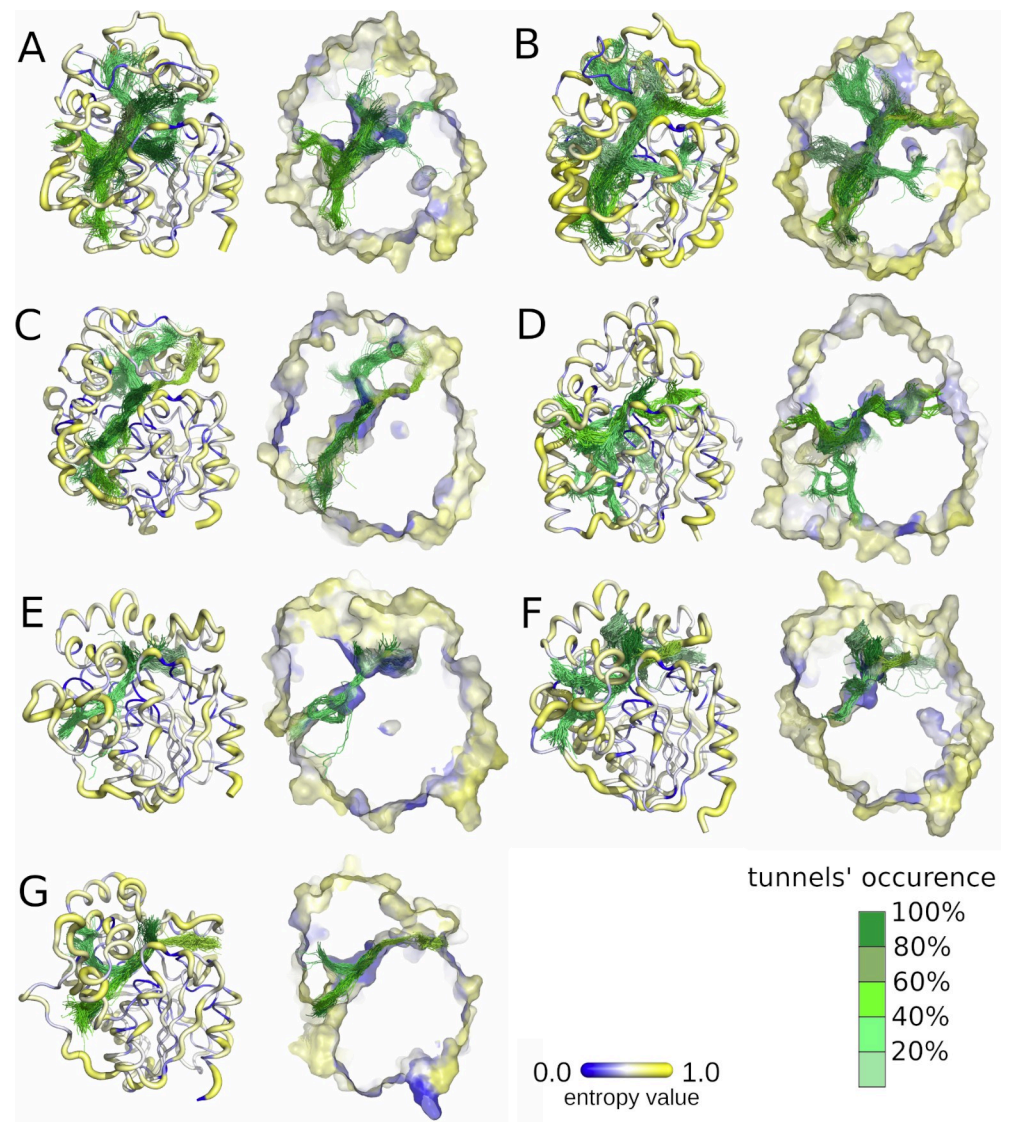


Fig 3. Visualisation of the entropy score of each protein residue (right), and frequency of tunnels identified with CAVER during molecular dynamics (MD) simulations (left) for each system: A) *M. musculus* soluble epoxide hydrolase (msEH), B) *H. sapiens* soluble epoxide hydrolase (hsEH), C) *S. tuberosum* soluble epoxide hydrolase (StEH1), D) *T. reesei* soluble epoxide hydrolase (TrEH), E) *B. megaterium* soluble epoxide hydrolase (bmEH), and thermophilic soluble epoxide hydrolases from an unknown source organism F) Sibe-EH, and G) CH65-EH. Protein residues are shown according to their entropy score: low values of entropy are marked as thin blue lines and higher values as thick yellow lines. Tunnel centerlines are coloured according to the frequency of their occurrence during MD simulations (the tunnels occurrence was calculated based on the numbers of the MD simulation frames in which the tunnel was identified; 100% means that the tunnel remained open in all 50,000 MD simulation frames): dark green indicates the most frequently identified tunnels, and light green those very rarely identified. The right side of each pair shows cross-sections of protein surfaces coloured according to the entropy score of each amino acid residue.

<https://doi.org/10.1371/journal.pcbi.1010119.g003>

simulations. Cross-sections of these structures suggest that the protein core is composed of residues with lower variability (lower entropy values), whereas the tunnel mouths, located at the protein surfaces, are surrounded by residues of both higher and lower variability (higher and lower entropy values, respectively).

We identified the residues lining these particular tunnels during the MD simulations. During MD simulations, the protein is not a rigid body and the residues gain some level of

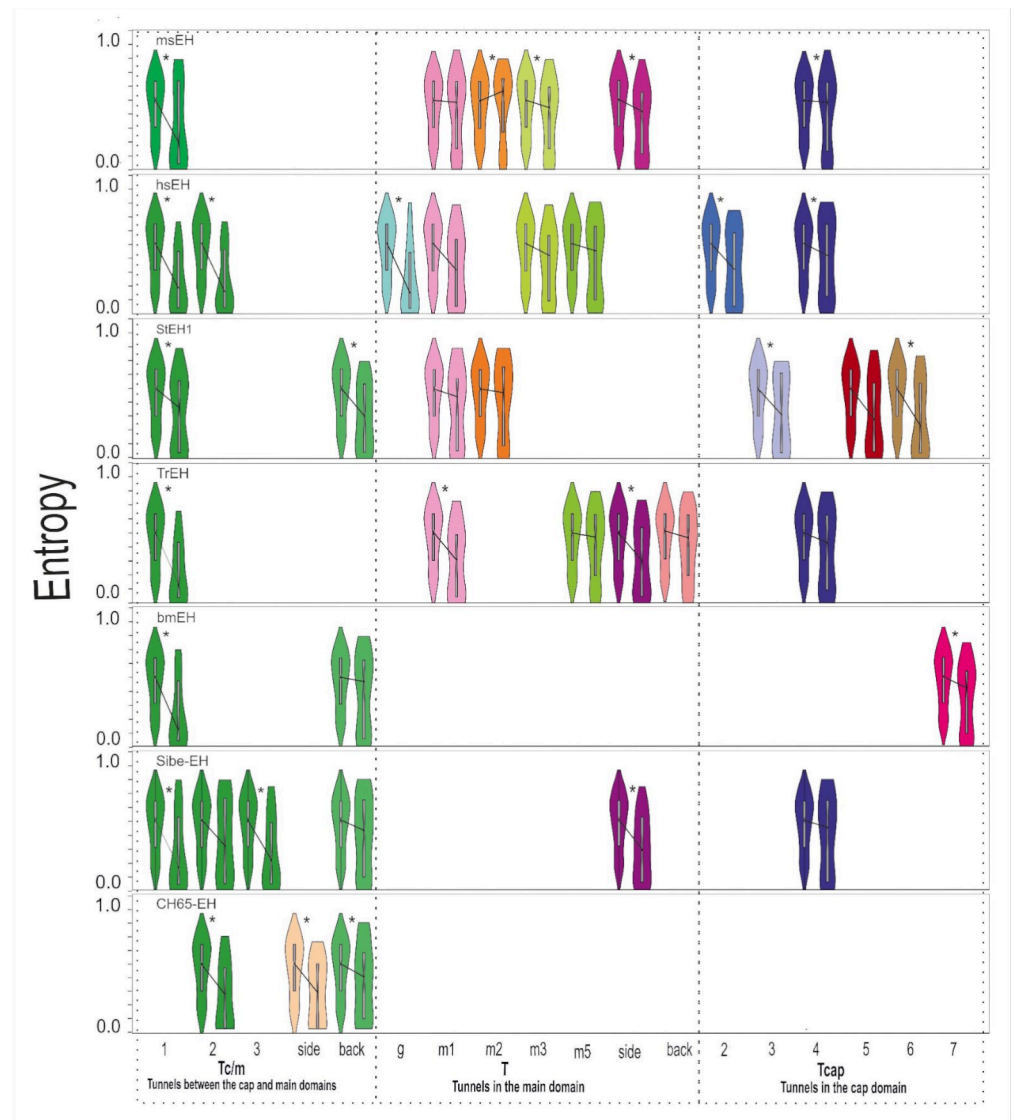


Fig 4. Distribution of the entropy values and median entropy values of tunnel-lining residues without the surface residues, and the remaining positions of the trimmed MSA (violin plots), for all analysed soluble epoxide hydrolase structures. Figure represents the data shown in [S12 Table](#). Statistically significant pairwise differences in median distances are marked by a star (*).

<https://doi.org/10.1371/journal.pcbi.1010119.g004>

flexibility, which may cause the opening and closing of identified tunnels. Moreover, due to the residues' movements, the identified tunnels may branch (either near the active site, in the middle of the tunnel, or near the surface). Since we observed many cases of tunnels branching near the surface, the list of identified tunnel-lining residues may be overrepresented by the surface residues. Therefore, we decided to perform an entropy analysis of: i) all tunnel-lining residues; ii) surface tunnels-lining residues; and iii) tunnel-lining residues without the surface residues. An evolutionary analysis of the tunnel-lining residues without the surface residues is presented in [Fig 4](#). Analysis was performed using the same procedure as in the case of the referential compartments. Complete lists of tunnel-lining residues are shown in [S5–S11 Tables](#). A detailed analysis of the sEHs tunnels is shown in [S3–S9 Figs](#). Median distances of all analysed proteins are listed in [S12 Table](#).

Based on the median distances between all tunnel-lining residues and the remaining residues' positions in the MSA, we concluded that almost all analysed tunnels should be considered as conserved. Following exclusion of the surface residues from the tunnel-lining residues, differences in median values decreased indicating that the conserved character of tunnels comes from the buried residues (**Fig 4 and S12 Table**). It is clear that the surface tunnel-lining residues generally reach higher entropy values than the other analysed tunnel-lining residues (**S3–S9 Figs**).

Presented violin plots (**Fig 4**) provide insight into tunnel's residues entropy distribution. To perform that, the right violin shape from each pair has to be analysed. For example, in bmEH the distribution of the entropy values among residues creating Tc/m1 tunnel shows a triangle-like shape with a wide base of residues with low entropy values, which corresponds to the prevailing contribution of conserved residues. In contrast, the distribution of the entropy values among residues lining the Tc/m_back tunnel resembles a rectangle or even hourglass-like shape which means that both variable and conserved residues build that tunnel. Thus, analysis of the shape of the violin plots provides descriptive information about the variability of the residues creating each tunnel. The differences between violin plots for all tunnel-lining residues, and those with excluded surface residues, clearly confirm the variable character of tunnels' entries.

Detailed analysis of selected tunnels

The violin plots provide information about the general variability of the tunnel-lining residues, but do not give insight into the location of the variable/conserved residues along the tunnel. To analyse that we performed a more advanced analysis. We selected three different tunnels which were identified in three different sEHs. The Tc/m tunnel of hsEH and the Tm1 tunnel of StEH1 represent the most commonly identified tunnels, and the Tc/m_back tunnel of bmEH represents an interesting case of a tunnel which already was engineered. The entropy values of the tunnel-lining residues are presented in **S13 Table**.

As we pointed out elsewhere [47] the Tc/m tunnel whose mouth is located between the main and cap domains can be seen as an ancestral tunnel created during cap domain insertion and preserved in nearly all epoxide hydrolases. In hsEH this tunnel (**Fig 5A**) has an average length of ~ 13.3 Å. It was open during 59% of the simulation time, with an average bottleneck radius of 1.6 Å, reaching a maximum of 2.7 Å. It is lined by residues with both low and high values of entropy, which makes the overall entropy distribution nearly flat (when surface residues are included) or exponential (when surface residues are excluded) which corresponds to the hourglass-like and triangle-like shape of the violin plot, respectively. The majority of variable residues is located close to the surface or at the interface between the cap and main domains. Close inspection of the tunnel revealed also a highly variable residue (i.e. with higher entropy value)—F497 (Schneider entropy value 0.7946)—located approximately in the middle of the tunnel and situated between two less-variable residues (i.e. with lower entropy values)—D496 (Schneider entropy value 0.0336), from the active site, and V498 (Schneider entropy value 0.4713). The F497 residue might act as a molecular gate [48] since its position in several other crystal structures differs substantially, and was identified as a surface residue (**S10 Fig**).

The Tm1 tunnel of StEH1 is the shortest identified in this structure (**Fig 5B**). Similar tunnels were identified in three other analysed sEHs: msEH, hsEH, and TrEH. The tunnel mouth is located in the main domain, near the NC-loop and hinge region. A close inspection of this tunnel revealed that it was ~ 13 Å long on average, and was always open during MD simulation. It had an average bottleneck radius of 1.9 Å, with a potential to increase up to 3.1 Å. The analysis of the violin plots suggests overrepresentation of the variable residues (reversed triangle-

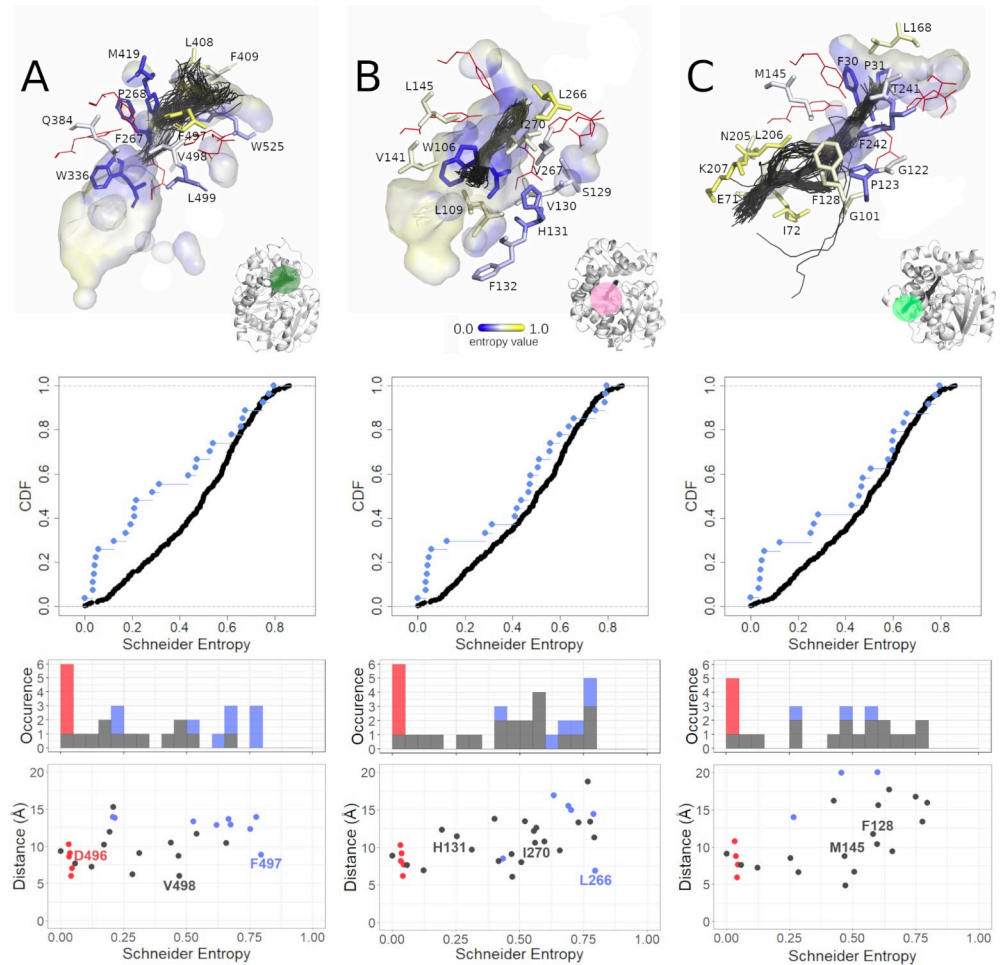


Fig 5. Analysis of the selected tunnels of soluble epoxide hydrolases (sEHs). A) The Tc/m tunnel of the *H. sapiens* soluble epoxide hydrolase (hsEH) structure, B) the Tm1 tunnel of the *S. tuberosum* soluble epoxide hydrolase (StEH1) structure, and C) the Tc/m_back tunnel of the *B. megaterium* soluble epoxide hydrolase (bmEH). Each panel consists of three parts: top section—close-up of tunnel residues. Residues are coloured according to entropy score. For the sake of clarity, less-frequently detected amino acid residues were omitted, and those creating the active site are shown as red lines. The active site cavity is shown as the interior surface, and the representative tunnel detected during molecular dynamics (MD) simulations as centerlines; middle section—cumulative distribution function (CDF) of entropy score for the tunnel-lining residues without the surface residues (cyan dots) and corresponding counterpart (black dots); and bottom section—scatterplot of the tunnel residues' entropy values relative to distance from the geometric centre of the α carbons of the enzyme, along with a marginal histogram of entropy value counts in respective intervals. Scatterplot points as well as histogram counts grouped into classes based on residue classification (active site—red; surface residues—blue; buried—grey).

<https://doi.org/10.1371/journal.pcbi.1010119.g005>

like shape of the violin plot, when surface residues are included), and nearly flat distribution of entropy values (hourglass-like violin shape, when surface residues are excluded). The majority of the tunnel-lining residues showed relatively high entropy values, while the residues with lower entropy values were located in proximity to the active site. In our previous analysis of StEH1 [49], we identified three residues, namely P188 (Schneider entropy value 0.5117) (not shown in Fig 5), L266 (Schneider entropy value 0.7946), and I270 (Schneider entropy value 0.5594), as potentially useful during protein engineering. Here we present that those residues are also variable, which may suggest that their substitution might not affect protein stability. Interestingly, approximately in the middle of the tunnel length, a less-variable H131 (Schneider entropy value 0.2524) residue was also identified.

The last example is the Tc/m_back tunnel of bmEH (Fig 5C) which was already engineered by Kong *et al.* [50]. This tunnel was identified as a third tunnel during MD simulation and had an average length of 26.7 Å. It was open only for 18% of the simulation time, with an average bottleneck radius of 1.0 Å, and the potential to increase up to 1.8 Å. The mouth of this tunnel was located in the main domain. Both violin plots (with surface residues included and excluded) show a similar hourglass-like shape. Close inspection of the tunnel revealed that residues with lower entropy values contributed to the binding cavity inside the main domain, while residues with higher entropy were located in the area surrounding a deep pocket on the protein surface. We also found two residues, namely F128 (Schneider entropy value 0.5798) and M145 (Schneider entropy value 0.4678), which had lower entropy values than their neighbours. Those residues were successfully modified to create a novel tunnel leading to the bmEH active site, allowing conversion of bulky substrates [50].

Discussion

In our study, we focused on sEHs, which are enzymes belonging to the α/β -hydrolase superfamily. Members of this superfamily share a barrel-like scaffold of eight anti-parallel β -strands surrounded by α -helices with a mostly helical cap domain sitting on top of the entrance to the active site [45], which seems to be also the oldest [51] and most stable [52] fold used by one of the largest groups of enzymes [53]. Structural and evolutionary analyses of EHs have been reported systematically [40,47,54–56], providing valuable insights into their structural and functional features. In our work, we first assessed the system-specific compartments described previously by Barth *et al.*, such as the main and cap domains, the NC-loop, and the cap-loop, along with secondary structure elements such as strands, helices, and loops. Based on an alignment of 95 EH sequences, three available crystal structures, and several homology models, they showed that the main and cap domains are conserved, while the NC-loop and cap-loop are variable [40].

Here, we analysed an alignment of 1455 EH sequences and additionally performed an in-depth analysis of the seven complete crystal structures representing different clades (animals, plants, fungi, and bacteria). By calculating the difference between median distances of Schneider entropy values of a selected compartment and the remaining positions of the trimmed MSA—we were able to determine the variability of each compartment. The calculated median distances for all analysed compartments confirmed well-known observations: active sites comprised highly conserved residues, with greater variability exhibited by surface residues than by buried residues [3,16,45]. Our results were also consistent with the work of Barth *et al.* [40], showing that the cap-loop and NC-loop should be considered as variable features. However, in contrast to their work, for such a large set of sequences, the whole main and cap domains were considered variable. In all analysed proteins, α -helices, and loops were found to be variable (S2 Table), while β -strands were found to be conserved in all analysed proteins, except for msEH. Such a tendency was shown previously for other systems elsewhere [18]. Further, since we were able to identify structural compartments of the seven sEHs analysed, observations regarding the modularity of EHs are still applicable [37].

The main aim of our analysis was to perform what was, to our knowledge, the first systematic analysis of the evolution of tunnels in a large family of sEHs. Therefore, our results can be applied mostly to the EHs, and—with some minor adjustments—to other members of the α/β -hydrolases superfamily. We identified multiple tunnels of different sizes and shapes, located in three different regions: the cap and main domains, as well as at the border between those domains. We hypothesised that tunnels are conserved structural features equipped with variable parts, such as gates responsible for different substrate specificity profiles in closely related

family members. This hypothesis was based on two assumptions: i) that the surface residues are more variable in comparison to the buried residues, and ii) that access to the active site cavity should be preserved to sustain the catalytic activity of the enzyme. Our results confirmed both assumptions. Moreover, we identified the Tc/m tunnel which was present in all analysed sEHs, and is located in the border between the cap and main domains. The cap domain is thought to be a result of a large insertion into the α/β -hydrolase main domain [45,47]. Both domains interact, creating a hydrogen bond network [57]; they co-evolved to preserve access to the buried active site while also ensuring the flexibility required for transport of the substrate and the products [58]. Most of the residues with lower entropy values in the cap domain face the main domain. This finding confirms previously presented information about the main and cap domains' relative flexibility [40,59–62].

We also proposed two ways of the analysis of the tunnel residues variability. The violin plots allow analysis of the contribution of variable and conserved residues, which provides a general overview of each tunnel. They also allow assessment of the variability of a particular compartment relative to the remaining positions of the MSA (as shown in Fig 1). The scatter-plots (similar to those in Fig 5A–5C) provide detailed insight and can be used to draw further conclusions regarding the distribution of entropy values of tunnel-lining residues along an analysed tunnel. They can also be used to identify the most variable and conserved tunnel-lining residues. In general, after excluding the active site and surface residues, the analysed examples (Fig 5) show three cases of entropy distribution among tunnel-lining residues: i) the flat distribution of the entropy values (Fig 5A); ii) the overrepresentation of residues with higher entropy values (Fig 5B); and iii) the quasi-sigmoidal distribution (Fig 5C; most of the residues have values of the entropy in the range of 0.25–0.7).

Our results confirmed the conserved character of the tunnels. Moreover, we found that even conserved tunnels can be lined with more variable residues, located not only at the surface (tunnels' entry). Close inspection of the Tc/m tunnel of hsEH allowed us to detect variable S412 and F497 residues (Schneider scores 0.618 and 0.795, respectively), among which phenylalanine was observed to be the most flexible amino acid, and which was even observed in a different conformation in crystal structures (S10 Fig). This indicates a potential role for F497 as a gate, controlling access through this tunnel [48]. On the other hand, the Tc/m tunnel is also lined by more conserved residues, such as the highly conserved substrate-stabilising tyrosine located in the cap domain (Y466 in hsEH, Schneider score 0.0323) [63,64].

Analysis of the variability of particular amino acid positions could be used in the search for feasible key amino acids (hot-spots) [65]. More variable positions might be considered as favourable locations for the introduction of mutations. Such residues can be detected even for the shortest tunnels, and have already been shown to enable fine-tuning of enzyme properties [66]. For example, the Tm1 tunnel of StEH1 is lined with several variable residues which may have a role to play in the fine-tuning of the enantioselectivity of that enzyme [67]. Such a strategy is acknowledged as one of the most likely to succeed, since it does not significantly disturb protein activity and stability, and the different locations of hot-spots along the transport pathway may enable modification of geometric/electrochemical constraints, thus contributing to the enzyme selectivity.

In our other study, we showed a relationship between a tunnel's shape and location, and the enzyme's function [47]. Thus, the evolution of the tunnel network can be considered as an additional mechanism that allows the enzyme to adapt and catalyse the conversion of different substrates. Mimicking such a process could provide a straightforward strategy for enzyme re-engineering. As we pointed out above, the insertion of the cap domain has created the buried active site cavity and the Tc/m tunnel ensuring access to that cavity. This tunnel can be considered as an ancestral tunnel and it seems to be well-preserved among nearly all sEHs family

members. However, the insertion of large fragments into existing structures appears to be a high-risk strategy. Based on our results, we can suggest a much easier approach that can be used for tunnel network redesign.

Perforation mechanism of the tunnel formation

The observed entropy values of tunnel-lining residues usually range from 0.25 to 0.7 (S13 Table). As we showed, the scatterplots can be used to identify the most variable and conserved residues. Variable residues are considered potentially safe hot-spots for single-point mutations [65]. We can imagine that new tunnels providing access to the protein interior can appear as a result of a “perforation” *via* a mutation occurring: i) in the surface layer of protein or ii) at the border of large cavities affecting surface integrity (Fig 6). Such a process can be easily mimicked and adopted for enzyme modification. We showed [47] that, in some cases, tunnels behave more like a series of small cavities which are rarely open. In the case of such tunnels, a mutation resulting in a permanently open cavity might be a driving force for future tunnel widening and modulation of selectivity or activity of enzymes, or otherwise provide additional regulation of activity.

The appearance of a new tunnel, resulting from a single-point mutation, *via* the proposed perforation mechanism provided significant freedom and flexibility for α/β -hydrolases to modify their activity and selectivity. Since the mechanism for hydrolysis performed by the sEHs involves deprotonation of the nucleophile in the hydrolysis step (proton shuttling) and water attack, it requires precise transport of water molecules. New tunnels could significantly improve the enzyme’s performance by separating the substrate/products transport pathways from water delivery tracks.

A perfect example of the mimicking of the proposed surface perforation model is the transformation of the Tc/m_back tunnels of the bmEH shown by Kong *et al.* [50] in which they turned a substrate inaccessible tunnel into an accessible one in order to improve the enzyme’s functionality. As we showed here these two residues whose substitution to alanine led to the opening of a side tunnel, improving the activity of bmEH upon α -naphthyl glycidyl ether, had

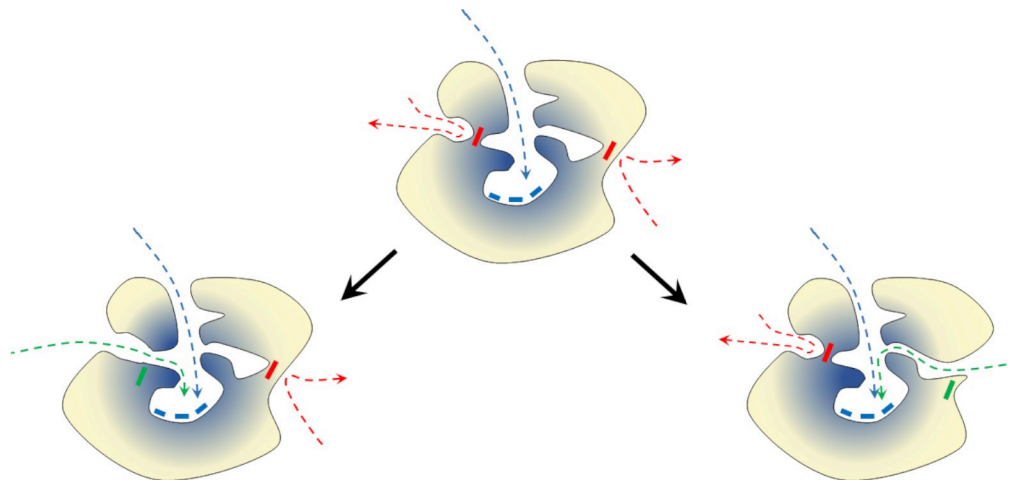


Fig 6. Schematic representation of the ‘perforation’ model of protein tunnel evolution. The ancestor protein (middle) and two modification pathways leading to a new enzyme by merging internal cavities (left) or by surface perforation (right). Yellow—variable residues; blue—conserved residues. Boxes represent residues: blue—conserved active site residues, red—potentially mutable (variable) residue, and green—mutated residue. Arrows represent pathways leading to the active site: blue—actual pathway, red—potential novel pathway, and green—novel open pathway.

<https://doi.org/10.1371/journal.pcbi.1010119.g006>

higher entropy values than their neighbours. This work also led us to a hypothesis that mutations of such variable residues could also appear spontaneously and may drive the evolution of the active site accessibility *via* surface perforation and/or joining of internal cavities. Identification of such residues which are prone to cause such an effect might easily be adopted as part of protein reengineering processes. These conclusions are supported by the observations of Aharoni *et al.* [68], who noticed that most mutations affecting protein functionality (mostly activity and selectivity) were located either on the protein surface or within the active site cavity. Indeed, the investigation of long and narrow tunnels, not obviously relevant at first glance during protein engineering, should be regarded as a strategy for new pathway creation, as illustrated by Brezovsky *et al.* in their *de novo* tunnel design study which resulted in the most active dehalogenases known so far [69]. Dehalogenases are closely related to sEHs and belong also to the α/β -hydrolases superfamily, thus further supporting the rationality of our approach.

The tunnels described in our findings which we consider conserved provided substantial information about their origin, and about the evolution of enzymes' families more broadly. On the other hand, our results suggest that after the ancestral occlusion of the active site, the further evolution of α/β -hydrolases may be driven by perforation of either the surface or of the internal cavities, which mostly comprised variable residues. Tunnels themselves can be equipped with both conserved residues, which are potentially indispensable for their performance, as well as highly variable ones, which can be easily used for fine-tuning an enzyme's properties. Such hotspots can be easily identified using the approach presented here.

Methods

Workflow

Evolutionary analysis was divided into two parts: system-specific compartment analysis, and tunnel analysis. Prior to those analyses, the positions of the residues that contribute to compartments and tunnels needed to be mapped in an MSA comprising sequences of epoxide hydrolases. The identified residues were then used as input for an evolutionary analysis using the BALCONY software [41]. Tunnels were identified by CAVER software [46] in both crystal structures and during MD simulations and then compared with each other to find their corresponding counterparts. Finally, the tunnel-lining residues, the surface tunnel-lining residues, and the tunnel-lining residues without surface residues were used for the evolutionary analysis using BALCONY software (Fig 7).

Obtaining protein structures for analysis

Seven unique and complete crystal structures were downloaded from the PDB database [39]. The selected structures all belong to the α/β hydrolase superfamily, share the same core fold scheme [45], and consist of a main and a cap domains [40]. Five structures represent different clades. They belong to clades of animals (*M. musculus* (msEH, PDB ID: 1CQZ)), *H. sapiens* (hsEH, PDB ID: 1S8O)), plants (*S. tuberosum* (StEH1, PDB ID: 2CJP)), fungi (*T. reesei* (TrEH, PDB ID: 5URO)), and bacteria (*B. megaterium* (bmEH, PDB ID: 4NZZ)). Two structures were collected from an unknown source organism in hot springs in Russia and China (Sibe-EH, PDB ID: 5NG7, CH65-EH, and PDB ID: 5NFQ).

Structure preparation

Ligands were manually removed from each structure, and only one chain was used for the analysis. For the msEH and hsEH structures, only the C-terminal domain, with the hydrolytic activity, was used. Several referential structural compartments were selected for further

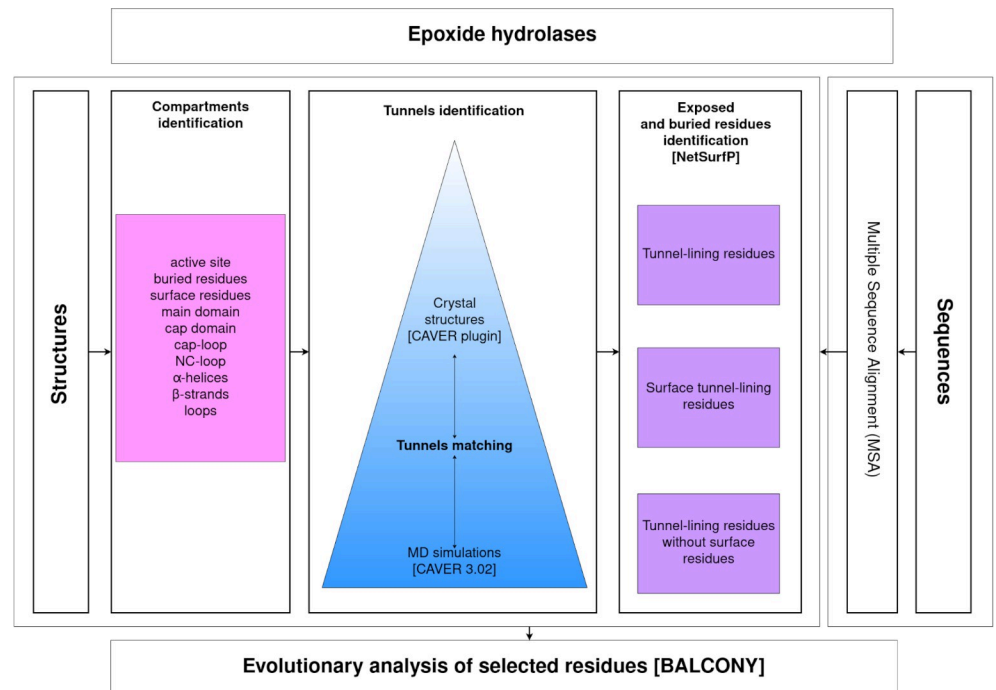


Fig 7. Research workflow.

<https://doi.org/10.1371/journal.pcbi.1010119.g007>

analysis (see [S2 Table](#)): the active site; buried and surface residues; main and cap domains; cap-loop; NC-loop; and α -helices, β -strands, and loops. The definitions of the cap-loop and NC-loop were taken from the works of Barth *et al.* [40] and of Smit and Labuschagne [70]. The NetSurfP service [44] was used to identify both buried and surface residues. Tunnels identified by CAVER software were also selected for further analysis.

MD simulations

MD simulations for msEH (PDB ID: 1CQZ), hsEH (PDB ID: 1S8O), StEH1 (PDB ID: 2CJP), TrEH (PDB ID: 5URO), bmEH (PDB ID: 4NZZ), Sibe-EH (PDB ID: 5NG7), and CH65-EH (PDB ID: 5NFQ) were carried out according to the protocol described by Mitusińska *et al.* [47].

CAVER analysis

Tunnel identification and analysis in each system were carried out using CAVER 3.02 software [46] in two steps: i) the crystal structure of the enzyme was analysed by the CAVER plugin for PyMOL [71]; ii) tunnels were identified and analysed in 50,000 snapshots of multiple MD simulations by the standalone CAVER 3.02 software. The parameters used for both steps are shown in [S14 Table](#). The tunnels found in MD simulations and in crystal structures were ranked and numbered based on their throughput value [46].

Tunnels comparison

The tunnels identified during MD simulations and in crystal structures were compared based on the occurrence of tunnel-lining residues. For crystal structures, the occurrence was defined as the number of atoms of a particular amino acid that were identified as tunnel-forming

atoms. For the sake of simplicity, no weighting scheme was used: C α , backbone atoms, and side-chain atoms were considered to be of the same importance. For MD simulations' results, tunnel occurrence was defined differently: as the number of MD snapshots in which particular amino acid was detected for a particular tunnel cluster. Therefore, this number could vary between 1 and the number of MD snapshots (50,000 in the performed analyses).

Despite the different definitions of occurrence used for crystal structures and for MD results, interpretation can be conducted in exactly the same way for each. Therefore, the above-defined occurrences can be directly used for fine-tuning the list of residues that form tunnels, i.e. by applying certain threshold values. In this study, the threshold was a number in the open range (0,1) and amino acids were retained only if the condition $o > \max(o) \times \tau$ was satisfied, where o is the occurrence and τ is the threshold value.

For sets of tunnels detected in both the crystal structures and in the MD data, a distance matrix was calculated using the Jaccard distance formula [72]:

$$d_{T_A T_B} = \frac{|T_A \cup T_B| - |T_A \cap T_B|}{|T_A \cup T_B|}$$

where T_A and T_B are A and B tunnels, respectively, and d is the Jaccard distance.

Elements of the matrix with lower distance values correspond to crystal structures/MD data pairs of similar tunnels. Further improvements in distance calculation accuracy were achieved by fine-tuning the tunnels' amino acids with thresholds. For each of the compared pairs, two independent thresholds were used, and τ values for both lists of tunnel-forming residues in the crystal structure and MD simulations were scanned in the range of [0.05, 0.95] with a step of 0.05 (361 combinations in total). The combination of τ values which yielded the minimal distance was selected as the optimal one.

Obtaining protein sequences, and MSA

Each of the amino acid sequences of the selected sEHs (PDB IDs: 1S8O, 1CQZ, 2CJP, 4NZZ, 5URO, 5NFQ, and 5NG7) was used as a separate query for a BLAST [73] search of similar protein sequences. The obtained results were merged and duplicates were removed, providing 1484 unique sequences (including those primarily selected). The 12 outlying sequences were detected and individually checked in the Uniprot database [74]. Nine sequences were trimmed according to the hydrolase domain, and three were removed since there was no information or similarity with other sequences. Next, in order to eliminate proteins other than EHs from the set of sequences, the conserved motifs described by van Loo *et al.*[55] were used, and only sequences with motifs H-G-X-P and G-X-Sm-X-S/T were preserved (where X is usually an aromatic residue, and Sm is a small residue). As a result, 29 sequences were discarded during the analysis. In the last step of MSA preparation, additional domains (e.g. phosphatase domain) were removed. To detect sequences with an additional domain, a histogram of sequence lengths was prepared, and long sequences (> 420 residues) were trimmed all at once in a temporary MSA. In the end, a final MSA of 1455 epoxide hydrolase sequences was prepared with Clustal Omega [75] using default parameters (S11 Fig).

BALCONY analysis

BALCONY (Better ALignment CONsensus analYsis) [41], an R package, was used to analyse the MSA and map selected structural compartments/tunnels onto the correct positions in aligned reference UniProt sequences. The Schneider metric [42] was calculated for each alignment position. Selected structures of *M. musculus*, *H. sapiens*, *S. tuberosum*, *T. reesei*, and *B. megaterium* sEHs, as well as the two thermophilic enzymes collected in hot springs (respective

PDB IDs: 1CQZ, 1S8O, 2CJP, 5URO, 4NZZ, 5NG7, and 5NFQ), were divided into compartments/tunnels as shown in **S1 and S5–S11 Tables**. The compartment/tunnel residues were then appropriately mapped with MSA, and Schneider entropy values were collected for each position in the MSA.

Variability analysis

To assess the variability of a particular tunnel/compartment, their positions were compared with selected positions of the MSA. The MSA was trimmed only to positions where at least one residue was present of the seven structures (PDB IDs: 1CQZ, 1S8O, 2CJP, 5URO, 4NZZ, 5NG7, and 5NFQ) (**S12 Fig**). The MSA containing 1455 sequences was trimmed from 722 to 419 positions. This way, for each comparison, Schneider entropy values of a compartment/tunnel positions were compared to the Schneider entropy values of selected positions of the MSA in which were present: i) neither one of residues of the currently analysed compartment/tunnel, and ii) at least one residue of the seven analysed structures. In order to determine whether a compartment was to be classed as variable, a median distance was calculated, which was defined as a difference between medians of Schneider entropy values of a selected compartment/tunnel and the selected positions in the MSA. If the median distance was > 0 , then the analysed compartment was considered variable. To compare the distributions of entropy scores of analysed compartments/tunnels with the distribution of the selected positions of the MSA, the Epps–Singleton two-sample test [43] was used. The advantage of this test is the comparison of the empirical characteristic functions (the Fourier transform of the observed distribution function) instead of the observed distributions. The comparison analysis was performed using the `es.test()` function from GitHub repository [76]. In an attempt to visualise the variability of selected tunnels (**Fig 5**), the collected entropy values of selected tunnel-lining residues without the surface residues and the selected MSA positions were sorted separately, and cumulative distribution functions (CDF) were calculated. For each position in the selected tunnel, a paired one from the selected position in the MSA was found, based on the minimal CDF. Plots of CDF as a function of entropy score were prepared.

Supporting information

S1 Table. List of amino acids forming particular compartments of analysed protein structures.

(XLSX)

S2 Table. Differences in median Schneider entropy values between the median entropy values of the selected proteins' compartments and remaining positions of the trimmed Multiple Sequence Alignment (MSA). Negative values indicate compartments with lower variability and positive values indicate compartments with higher variability in comparison to the remaining positions in the trimmed MSA. Differences in median distances values that are marked bold passed the statistical significance of the Epps–Singleton two-sample test (p -value < 0.05).

(XLSX)

S3 Table. List of corresponding tunnels identified in both the crystal structure and in the MD simulation for each analysed protein structure.

(XLSX)

S4 Table. Comparison of maximal bottleneck radii measured in corresponding tunnels identified in both the crystal structure (CR) and in the MD simulation (MD) for each

protein structure.

(XLSX)

S5 Table. List of amino acids forming analysed tunnels in msEH structure.

(XLSX)

S6 Table. List of amino acids forming analysed tunnels in hsEH structure.

(XLSX)

S7 Table. List of amino acids forming analysed tunnels in StEH1 structure.

(XLSX)

S8 Table. List of amino acids forming analysed tunnels in TrEH structure.

(XLSX)

S9 Table. List of amino acids forming analysed tunnels in bmEH structure.

(XLSX)

S10 Table. List of amino acids forming analysed tunnels in Sibe-EH structure.

(XLSX)

S11 Table. List of amino acids forming analysed tunnels in CH65-EH structure.

(XLSX)

S12 Table. Differences in Schneider entropy values between the median distance of selected tunnel-lining residues and the median distances of the remaining positions of the trimmed Multiple Sequence Alignment (MSA). Negative values indicate compartments with lower variability and positive values indicate compartments with higher variability than the remaining positions of the trimmed MSA. Table consists of: the MD simulation tunnel ranking (MDRank), tunnel name, its average length, number of tunnel-lining residues, number of tunnel-lining residues without the surface residues, number of surface tunnel-lining residues, and the last three columns comprise of the differences between the median distances of particular tunnel-lining residues and the remaining positions of the trimmed MSA: median distances of tunnel-lining residues (7th column), median distances of the tunnel-lining residues without the surface residues (8th column), and median distances of the surface tunnel-lining residues (9th column). Differences between median distances values that are marked bold passed the statistical significance of the Epps-Singleton two-sample test (p -value < 0.05). Hyphen (-) indicates that there were no surface residues. NA means that the number of surface residues was insufficient to obtain the p -value median distance of the Epps-Singleton two-sample test.

(XLSX)

S13 Table. Entropy values for selected tunnels, Tm1 from StEH1, Tc/m1 from hsEH, and Tc/m_back from bmEH. Surface amino acids are in italics. Active site residues are marked by an asterisks (*).

(XLSX)

S14 Table. The list of parameters set for both CAVER plugin and CAVER 3.0 tunnels identification for each of the analysed systems.

(XLSX)

S1 Fig. Issue related to identification of asymmetrical tunnels based on the example of the Tc/m tunnels identified by CAVER software during MD simulations.

(TIFF)

S2 Fig. Correlation between maximal bottleneck radii measured in corresponding tunnels identified in both the crystal structure and in the MD simulation for each protein structure.

(TIFF)

S3 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the *M. musculus* soluble epoxide hydrolase (msEH). Statistically significant pairwise differences in the median distance values are marked by a star (*).

(TIFF)

S4 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the *H. sapiens* soluble epoxide hydrolase (hsEH). Statistically significant pairwise differences in the median distance values are marked by a star (*).

(TIFF)

S5 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the *S. tuberosum* soluble epoxide hydrolase (StEH1). Statistically significant pairwise differences in the median distance values are marked by a star (*).

(TIFF)

S6 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the *T. resei* soluble epoxide hydrolase (TrEH). Statistically significant pairwise differences in the median distance values are marked by a star (*). NA by the violin plots means that the number of surface residues was insufficient to obtain the p-value of the Epps–Singleton two-sample test. In the case of the Tside tunnel, no surface residues were identified.

(TIFF)

S7 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the *B. megaterium* soluble epoxide hydrolase (bmEH). Statistically significant pairwise differences in the median distance values are marked by a star (*). NA by the violin plots means that the number of surface residues was insufficient to obtain the p-value median distance of the Epps–Singleton two-sample test.

(TIFF)

S8 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the thermophilic enzyme collected in hot springs in Russia (Sibe-EH). Statistically significant pairwise differences in the median distance values are marked by a star (*).

(TIFF)

S9 Fig. The distribution of the entropy values and the median entropy values of all tunnel-lining residues, tunnel-lining residues without the surface residues and the remaining positions of the trimmed MSA, and surface tunnel-lining residues (the violin and box plot) for the thermophilic enzyme collected in hot springs in China (CH65-EH). Statistically significant pairwise differences in the median distance values are marked by a star (*).

(TIFF)

S10 Fig. The open and closed position of the F497 residue of hSEH. The protein is shown as cartoon and surface, and the F497 residue is shown as sticks.

(TIFF)

S11 Fig. Representation of the created Multiple Sequence Alignment (MSA) of the epoxide hydrolases sequences. The red brace marks the sequences of the soluble epoxide hydrolases with known crystal structures. Gaps are marked in blue. MSA was pictured using the DECIPHER library for R (<https://www.rdocumentation.org/packages/DECIPHER/versions/2.0.2>).

(TIFF)

S12 Fig. Representation of the trimmed Multiple Sequence Alignment (MSA) of the epoxide hydrolases sequences. The red brace marks the sequences of the soluble epoxide hydrolases with known crystal structures. Gaps are marked in blue. MSA was pictured using the DECIPHER library for R (<https://www.rdocumentation.org/packages/DECIPHER/versions/2.0.2>).

(TIFF)

Author Contributions

Conceptualization: Artur Góra.

Data curation: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska.

Formal analysis: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska, Tomasz Skalski, Aleksandra Samol, Weronika Bagrowska, Tomasz Magdziarz.

Funding acquisition: Artur Góra.

Investigation: Maria Bzówka, Agata Raczyńska, Tomasz Skalski, Aleksandra Samol, Weronika Bagrowska, Tomasz Magdziarz.

Methodology: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska, Tomasz Skalski, Tomasz Magdziarz, Artur Góra.

Project administration: Artur Góra.

Resources: Aleksandra Samol, Artur Góra.

Software: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska, Tomasz Skalski, Tomasz Magdziarz.

Supervision: Artur Góra.

Validation: Maria Bzówka, Karolina Mitusińska.

Visualization: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska, Tomasz Skalski, Artur Góra.

Writing – original draft: Maria Bzówka, Karolina Mitusińska, Agata Raczyńska.

Writing – review & editing: Maria Bzówka, Karolina Mitusińska, Artur Góra.

References

1. Kimura M, Ohta T. On Some Principles Governing Molecular Evolution. *Proc Natl Acad Sci U S A*. 1974; 71(7):2848–52. <https://doi.org/10.1073/pnas.71.7.2848> PMID: 4527913
2. Tseng YY, Liang J. Estimation of Amino Acid Residue Substitution Rates at Local Spatial Regions and Application in Protein Function Inference: A Bayesian Monte Carlo Approach. *Mol Biol Evol*. 2005; 23(2):421–36. <https://doi.org/10.1093/molbev/msj048> PMID: 16251508
3. Franzosa EA, Xia Y. Structural determinants of protein evolution are context-sensitive at the residue level. *Mol Biol Evol*. 2009; 26(10):2387–95. <https://doi.org/10.1093/molbev/msp146> PMID: 19597162
4. Jäckel C, Kast P, Hilvert D. Protein Design by Directed Evolution. *Annu Rev Biophys*. 2008; 37:153–73. <https://doi.org/10.1146/annurev.biophys.37.032807.125832> PMID: 18573077
5. Damborsky J, Brezovsky J. Computational tools for designing and engineering enzymes. *Curr Opin Chem Biol*. 2014; 19:8–16. <https://doi.org/10.1016/j.cbpa.2013.12.003> PMID: 24780274
6. Garcia-Guevara F, Avelar M, Ayala M, Segovia L. Computational Tools Applied to Enzyme Design – a review. *Biocatalysis*. 2015; 1:109–17.
7. Hochberg GKA, Thornton JW. Reconstructing Ancient Proteins to Understand the Causes of Structure and Function. *Annu Rev Biophys*. 2017; 46(1):247–69. <https://doi.org/10.1146/annurev-biophys-070816-033631> PMID: 28301769
8. Siddiq MA, Hochberg GK, Thornton JW. Evolution of protein specificity: insights from ancestral protein reconstruction. *Curr Opin Struct Biol*. 2017; 47:113–22. <https://doi.org/10.1016/j.sbi.2017.07.003> PMID: 28841430
9. Arenas M, Bastolla U. ProtASR2: Ancestral reconstruction of protein sequences accounting for folding stability. *Methods Ecol Evol*. 2019; 11(2):248–57.
10. Cuesta SM, Rahman SA, Furnham N, Thornton JM. The Classification and Evolution of Enzyme Function. *Biophys J*. 2015; 109(6):1082–6. <https://doi.org/10.1016/j.bpj.2015.04.020> PMID: 25986631
11. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSpot: database of computational hot spots in protein interfaces. *Nucleic Acids Res*. 2008; 36(Database):D662–6. <https://doi.org/10.1093/nar/gkm813> PMID: 17959648
12. Pavelka A, Chovancova E, Damborsky J. HotSpot Wizard: a web server for identification of hot spots in protein engineering. *Nucleic Acids Res*. 2009; 37(Web Server):W376–83. <https://doi.org/10.1093/nar/gkp410> PMID: 19465397
13. Verma R, Schwaneberg U, Roccatano D. MAP2.03D: A Sequence/Structure Based Server for Protein Engineering. *ACS Synth Biol*. 2012; 1(4):139–50. <https://doi.org/10.1021/sb200019x> PMID: 23651115
14. Martinez R, Schwaneberg U. A roadmap to directed enzyme evolution and screening systems for biotechnological applications. *Biol Research*. 2013; 46(4). <https://doi.org/10.4067/S0716-97602013000400011> PMID: 24510142
15. Amaury A, Bartlett IGJ, Hegedüs Z, Dutt S, Hobot F, Horner KA, et al. Predicting and Experimentally Validating Hot-Spot Residues at Protein–Protein Interfaces. *ACS Chem Biol*. 2019; 14(10):2252–63. <https://doi.org/10.1021/acscchembio.9b00560> PMID: 31525028
16. Ramsey DC, Scherrer MP, Zhou T, Wilke CO. The Relationship Between Relative Solvent Accessibility and Evolutionary Rate in Protein Evolution. *Genetics [Internet]*. 2011 Jun; 188(2):479–88. Available from: <http://www.genetics.org/lookup/doi/10.1534/genetics.111.128025> PMID: 21467571
17. Shahmoradi A, Sydykova DK, Spielman SJ, Jackson EL, Dawson ET, Meyer AG, et al. Predicting Evolutionary Site Variability from Structure in Viral Proteins: Buriedness, Packing, Flexibility, and Design. *J Mol Evol [Internet]*. 2014 Oct 13; 79(3–4):130–42. Available from: <http://link.springer.com/10.1007/s00239-014-9644-x> PMID: 25217382
18. Sitbon E, Pietrovski S. Occurrence of protein structure elements in conserved sequence regions. *BMC Struct Biol*. 2007; 7(3):1–15. <https://doi.org/10.1186/1472-6807-7-3> PMID: 17210087
19. Liu J, Tan H, Rost B. Loopy Proteins Appear Conserved in Evolution. *J Mol Biol [Internet]*. 2002 Sep; 322(1):53–64. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022283602007362> [https://doi.org/10.1016/s0022-2836\(02\)00736-2](https://doi.org/10.1016/s0022-2836(02)00736-2) PMID: 12215414
20. Goldman N, Thorne JL, Jones DT. Assessing the Impact of Secondary Structure and Solvent Accessibility on Protein Evolution. *Genetics*. 1998; 149:445–58. <https://doi.org/10.1093/genetics/149.1.445> PMID: 9584116
21. Siltberg-Liberles J, Grahnen JA, Liberles DA, Siltberg-Liberles J, Grahnen JA, Liberles DA. The Evolution of Protein Structures and Structural Ensembles Under Functional Constraint. *Genes (Basel) [Internet]*. 2011 Oct 28 [cited 2018 Oct 24]; 2(4):748–62. Available from: <http://www.mdpi.com/2073-4425/2/4/748> <https://doi.org/10.3390/genes2040748> PMID: 24710290

22. Jack BR, Meyer AG, Echave J, Wilke CO. Functional Sites Induce Long-Range Evolutionary Constraints in Enzymes. *PLoS Biol* [Internet]. 2016 [cited 2018 Oct 24]; 14(5):1002452. Available from: <https://journals.plos.org/plosbiology/article/file?id=10.1371/journal.pbio.1002452&type=printable> PMID: 27138088
23. Subramanian K, Mitusińska K, Raedts J, Almourfi F, Joosten HJ, Hendriks S, et al. Distant Non-Obvious Mutations Influence the Activity of a Hyperthermophilic *Pyrococcus furiosus* Phosphoglucose Isomerase. *Biomolecules* [Internet]. 2019 May 31; 9(6):212. Available from: <https://www.mdpi.com/2218-273X/9/6/212> <https://doi.org/10.3390/biom9060212> PMID: 31159273
24. Otten R, Liu L, Kenner LR, Clarkson MW, Mavor D, Tawfik DS, et al. Rescue of conformational dynamics in enzyme catalysis by directed evolution. *Nat Commun*. 2018 Dec; 9(1):1314. <https://doi.org/10.1038/s41467-018-03562-9> PMID: 29615624
25. Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. Conformational dynamics and enzyme evolution. *J R Soc Interface*. 2018 Jul; 15(144):20180330. <https://doi.org/10.1098/rsif.2018.0330> PMID: 30021929
26. Carlson GM, Fenton AW. What Mutagenesis Can and Cannot Reveal About Allostery. *Biophys J*. 2016 May; 110(9):1912–23. <https://doi.org/10.1016/j.bpj.2016.03.021> PMID: 27166800
27. Weinkam P, Chen YC, Pons J, Sali A. Impact of Mutations on the Allosteric Conformational Equilibrium. *J Mol Biol*. 2013 Feb; 425(3):647–61. <https://doi.org/10.1016/j.jmb.2012.11.041> PMID: 23228330
28. 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* [Internet]. 2017 Sep; 37(5):1095–139. Available from: <http://doi.wiley.com/10.1002/med.21430> PMID: 27957758
29. Kokkonen P, Bednar D, Pinto G, Prokop Z, Damborsky J. Engineering enzyme access tunnels. *Biotechnol Adv* [Internet]. 2019 Nov; 37(6):107386. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0734975019300679> <https://doi.org/10.1016/j.biotechadv.2019.04.008> PMID: 31026496
30. Kingsley LJ, Lill MA. Substrate tunnels in enzymes: Structure-function relationships and computational methodology. *Proteins Struct Funct Bioinforma* [Internet]. 2015 Apr; 83(4):599–611. Available from: <http://doi.wiley.com/10.1002/prot.24772> PMID: 25663659
31. Nakamura A, Yao M, Chimnarong S, Sakai N, Tanaka I. Ammonia Channel Couples Glutaminase with Transamidase Reactions in GatCAB. *Science* (80-) [Internet]. 2006 Jun 30; 312(5782):1954–8. Available from: <https://www.sciencemag.org/lookup/doi/10.1126/science.1127156> PMID: 16809541
32. Kim J, Raushel FM. Perforation of the Tunnel Wall in Carbamoyl Phosphate Synthetase Derails the Passage of Ammonia between Sequential Active Sites. *Biochemistry*. 2004; 43:5334–40. <https://doi.org/10.1021/bi049945+> PMID: 15122899
33. Thangapandian S, John S, Lee Y, Arulalapperumal V, Lee KW. Molecular Modeling Study on Tunnel Behavior in Different Histone Deacetylase Isoforms. Gaetano C, editor. *PLoS One* [Internet]. 2012 Nov 29; 7(11):e49327. Available from: <https://dx.plos.org/10.1371/journal.pone.0049327> PMID: 23209570
34. Zawaira A, Coulson L, Gallotta M, Karimanzira O, Blackburn J. On the deduction and analysis of singlet and two-state gating-models from the static structures of mammalian CYP450. *J Struct Biol* [Internet]. 2011 Feb; 173(2):282–93. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S1047847710002959> <https://doi.org/10.1016/j.jsb.2010.09.026> PMID: 20932908
35. Nardini M, Dijkstra BW. α/β Hydrolase fold enzymes: the family keeps growing. *Curr Opin Struct Biol* [Internet]. 1999 Dec; 9(6):732–7. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0959440X99000378> [https://doi.org/10.1016/s0959-440x\(99\)00037-8](https://doi.org/10.1016/s0959-440x(99)00037-8) PMID: 10607665
36. Marchot P, Chatonnet A. Enzymatic Activity and Protein Interactions in Alpha/Beta Hydrolase Fold Proteins: Moonlighting Versus Promiscuity. *Protein Pept Lett* [Internet]. 2012 Feb 1; 19(2):132–43. Available from: <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=0929-8665&volume=19&issue=2&spage=132> <https://doi.org/10.2174/092986612799080284> PMID: 21933125
37. Bauer TL, Buchholz PCF, Pleiss J. The modular structure of α/β -hydrolases. *FEBS J* [Internet]. 2020 Mar 10; 287(5):1035–53. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1111/febs.15071> PMID: 31545554
38. Holmquist M. Alpha Beta-Hydrolase Fold Enzymes Structures, Functions and Mechanisms. *Curr Protein Pept Sci* [Internet]. 2000 Sep 1; 1(2):209–35. Available from: <http://www.ingentaselect.com/rpsv/cgi-bin/cgi?ini=xref&body=linker&reqdoi=10.2174/1389203003381405> PMID: 12369917
39. Berman HM. The Protein Data Bank. *Nucleic Acids Res* [Internet]. 2000 Jan 1; 28(1):235–42. Available from: <https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/28.1.235> PMID: 10592235
40. Barth S, Fischer M, Schmid RD, Pleiss J. Sequence and structure of epoxide hydrolases: A systematic analysis. *Proteins Struct Funct Bioinforma* [Internet]. 2004 Apr 2; 55(4):846–55. Available from: <http://doi.wiley.com/10.1002/prot.20013> PMID: 15146483

41. Pluciennik A, Stolarczyk M, Bzówka M, Raczynska A, Magdziarz T, Góra A. BALCONY: an R package for MSA and functional compartments of protein variability analysis. *BMC Bioinformatics* [Internet]. 2018 Dec 14; 19(1):300. Available from: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2294-z> PMID: 30107777
42. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins Struct Funct Genet* [Internet]. 1991 Jan; 9(1):56–68. Available from: <http://doi.wiley.com/10.1002/prot.340090107> PMID: 2017436
43. Epps TW, Singleton KJ. An omnibus test for the two-sample problem using the empirical characteristic function. *J Stat Comput Simul* [Internet]. 1986 Dec; 26(3–4):177–203. Available from: <http://www.tandfonline.com/doi/abs/10.1080/00949658608810963>
44. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, et al. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins Struct Funct Bioinforma* [Internet]. 2019 Jun 9; 87(6):520–7. Available from: <https://onlinelibrary.wiley.com/doi/abs/10.1002/prot.25674> PMID: 30785653
45. Ollis DL, Cheah E, Cygler M, Dijkstra B, Frolow F, Franken SM, et al. The α/β hydrolase fold. "Protein Eng Des Sel [Internet]. 1992; 5(3):197–211. Available from: <https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/5.3.197>
46. 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. Pric A, editor. *PLoS Comput Biol* [Internet]. 2012 Oct 18; 8(10):e1002708. Available from: <https://dx.plos.org/10.1371/journal.pcbi.1002708> PMID: 23093919
47. Mitusińska K, Wojsa P, Bzówka M, Raczynska A, Bagrowska W, Samol A, et al. Structure-function relationship between soluble epoxide hydrolases structure and their tunnel network. *Comput Struct Biotechnol J* [Internet]. 2022; 20:193–205. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S2001037021005225> <https://doi.org/10.1016/j.csbj.2021.10.042> PMID: 35024092
48. Bzówka M, Mitusińska K, Hopko K, Góra A. Computational insights into the known inhibitors of human soluble epoxide hydrolase. *Drug Discov Today* [Internet]. 2021 Aug; 26(8):1914–21. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S135964462100252X> <https://doi.org/10.1016/j.drudis.2021.05.017> PMID: 34082135
49. Mitusińska K, Magdziarz T, Bzówka M, Stańczak A, Góra A. Exploring Solanum tuberosum Epoxide Hydrolase Internal Architecture by Water Molecules Tracking. *Biomolecules* [Internet]. 2018 Nov 12; 8(4):143. Available from: <http://www.mdpi.com/2218-273X/8/4/143> <https://doi.org/10.3390/biom8040143> PMID: 30424576
50. Kong XD, Yuan S, Li L, Chen S, Xu JH, Zhou J. Engineering of an epoxide hydrolase for efficient biore-solution of bulky pharmaco substrates. *Proc Natl Acad Sci* [Internet]. 2014 Nov 4; 111(44):15717–22. Available from: <http://www.pnas.org/cgi/doi/10.1073/pnas.1404915111> PMID: 25331869
51. Caetano-Anollés G, Wang M, Caetano-Anollés D, Mittenthal JE. The origin, evolution and structure of the protein world. *Biochem J* [Internet]. 2009 Feb 1; 417(3):621–37. Available from: <https://portlandpress.com/biochemj/article/417/3/621/44492/The-origin-evolution-and-structure-of-the-protein> <https://doi.org/10.1042/BJ20082063> PMID: 19133840
52. Minary P, Levitt M. Probing Protein Fold Space with a Simplified Model. *J Mol Biol* [Internet]. 2008 Jan; 375(4):920–33. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022283607014581> <https://doi.org/10.1016/j.jmb.2007.10.087> PMID: 18054792
53. Suplatov DA, Besenmatter W, Svedas VK, Svendsen A. Bioinformatic analysis of alpha/beta-hydrolase fold enzymes reveals subfamily-specific positions responsible for discrimination of amidase and lipase activities. *Protein Eng Des Sel* [Internet]. 2012 Nov 1; 25(11):689–97. Available from: <https://academic.oup.com/peds/article-lookup/doi/10.1093/protein/gzs068> PMID: 23043134
54. Heikinheimo P, Goldman A, Jeffries C, Ollis DL. Of barn owls and bankers: a lush variety of α/β hydro-lases. *Structure* [Internet]. 1999 Jun; 7(6):R141–6. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0969212699800793> [https://doi.org/10.1016/s0969-2126\(99\)80079-3](https://doi.org/10.1016/s0969-2126(99)80079-3) PMID: 10404588
55. van Loo B, Kingma J, Arand M, Wubbolts MG, Janssen DB. Diversity and Biocatalytic Potential of Epox-ide Hydrolases Identified by Genome Analysis. *Appl Environ Microbiol* [Internet]. 2006 Apr 1; 72(4):2905–17. Available from: <http://aem.asm.org/cgi/doi/10.1128/AEM.72.4.2905-2917.2006> PMID: 16597997
56. Dimitriou PS, Denesyuk A, Takahashi S, Yamashita S, Johnson MS, Nakayama T, et al. Alpha/beta-hydrolases: A unique structural motif coordinates catalytic acid residue in 40 protein fold families. *Proteins Struct Funct Bioinforma* [Internet]. 2017 Oct; 85(10):1845–55. Available from: <http://doi.wiley.com/10.1002/prot.25338> PMID: 28643343
57. Lindberg D, Ahmad S, Widersten M. Mutations in salt-bridging residues at the interface of the core and lid domains of epoxide hydrolase StEH1 affect regioselectivity, protein stability and hysteresis. *Arch*

- Biochem Biophys [Internet]. 2010 Mar; 495(2):165–73. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0003986110000202> <https://doi.org/10.1016/j.abb.2010.01.007> PMID: 20079707
58. Jeon J, Nam HJ, Choi YS, Yang JS, Hwang J, Kim S. Molecular Evolution of Protein Conformational Changes Revealed by a Network of Evolutionarily Coupled Residues. *Mol Biol Evol* [Internet]. 2011 Sep; 28(9):2675–85. Available from: <https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/msr094> PMID: 21470969
 59. Schiøtt B, Bruice TC. Reaction Mechanism of Soluble Epoxide Hydrolase: Insights from Molecular Dynamics Simulations. *J Am Chem Soc* [Internet]. 2002 Dec; 124(49):14558–70. Available from: <https://pubs.acs.org/doi/10.1021/ja021021r> PMID: 12465965
 60. Bahl CD, Morisseau C, Bomberger JM, Stanton BA, Hammock BD, O'Toole GA, et al. Crystal Structure of the Cystic Fibrosis Transmembrane Conductance Regulator Inhibitory Factor Cif Reveals Novel Active-Site Features of an Epoxide Hydrolase Virulence Factor. *J Bacteriol* [Internet]. 2010 Apr 1; 192(7):1785–95. Available from: <https://jb.asm.org/content/192/7/1785> <https://doi.org/10.1128/JB.01348-09> PMID: 20118260
 61. Lindberg D, de la Fuente Revenga M, Widersten M. Temperature and pH Dependence of Enzyme-Catalyzed Hydrolysis of trans-Methylstyrene Oxide. A Unifying Kinetic Model for Observed Hysteresis, Cooperativity, and Regioselectivity. *Biochemistry* [Internet]. 2010 Mar 16; 49(10):2297–304. Available from: <https://pubs.acs.org/doi/10.1021/bi902157b> PMID: 20146441
 62. Hvorecny KL, Bahl CD, Kitamura S, Lee KSS, Hammock BD, Morisseau C, et al. Active-Site Flexibility and Substrate Specificity in a Bacterial Virulence Factor: Crystallographic Snapshots of an Epoxide Hydrolase. *Structure* [Internet]. 2017 May; 25(5):697–707.e4. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S096921261730062X> <https://doi.org/10.1016/j.str.2017.03.002> PMID: 28392259
 63. Mowbray SL, Elfström LT, Ahlgren KM, Andersson CE, Widersten M. X-ray structure of potato epoxide hydrolase sheds light on substrate specificity in plant enzymes. *Protein Sci* [Internet]. 2006 Jul; 15(7):1628–37. Available from: <http://doi.wiley.com/10.1110/ps.051792106> PMID: 16751602
 64. Hasan K, Gora A, Brezovsky J, Chaloupkova R, Moskalikova H, Fortova A, et al. The effect of a unique halide-stabilizing residue on the catalytic properties of haloalkane dehalogenase Data from *Agrobacterium tumefaciens* C58. *FEBS J* [Internet]. 2013 Jul; 280(13):3149–59. Available from: <http://doi.wiley.com/10.1111/febs.12238> PMID: 23490078
 65. Swint-Kruse L. Using Evolution to Guide Protein Engineering: The Devil IS in the Details. *Biophys J* [Internet]. 2016 Jul; 111(1):10–8. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0006349516303551> <https://doi.org/10.1016/j.bpj.2016.05.030> PMID: 27410729
 66. Chaloupková R, Sýkorová J, Prokop Z, Jesenská A, Monincová M, Pavlová M, et al. Modification of Activity and Specificity of Haloalkane Dehalogenase from *Sphingomonas paucimobilis* UT26 by Engineering of Its Entrance Tunnel. *J Biol Chem* [Internet]. 2003 Dec; 278(52):52622–8. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0021925820752749> <https://doi.org/10.1074/jbc.M306762200> PMID: 14525993
 67. Janfalk Carlsson Å, Bauer P, Dobritzsch D, Nilsson M, Kamerlin SCL, Widersten M. Laboratory-Evolved Enzymes Provide Snapshots of the Development of Enantioconvergence in Enzyme-Catalyzed Epoxide Hydrolysis. *ChemBioChem* [Internet]. 2016 Sep 15; 17(18):1693–7. Available from: <http://doi.wiley.com/10.1002/cbic.201600330> PMID: 27383542
 68. Aharoni A, Gaidukov L, Khersonsky O, Gould SM, Roodveldt C, Tawfik DS. The “evolvability” of promiscuous protein functions. *Nat Genet* [Internet]. 2005 Jan 28; 37(1):73–6. Available from: <http://www.nature.com/articles/ng1482> <https://doi.org/10.1038/ng1482> PMID: 15568024
 69. Brezovsky J, Babkova P, Degtjarik O, Fortova A, Gora A, Iermak I, et al. Engineering a de Novo Transport Tunnel. *ACS Catal* [Internet]. 2016 Nov 4; 6(11):7597–610. Available from: <https://pubs.acs.org/doi/10.1021/acscatal.6b02081>
 70. Smit M, Labuschagne M. Diversity of Epoxide Hydrolase Biocatalysts. *Curr Org Chem* [Internet]. 2006 Jul 1; 10(10):1145–61. Available from: <http://www.eurekaselect.com/openurl/content.php?genre=article&issn=1385-2728&volume=10&issue=10&page=1145>
 71. Schrödinger. The PyMOL Molecular Graphic Systems. Schrödinger, LLC;
 72. Gilbert G. Distance between Sets. *Nature* [Internet]. 1972 Sep; 239(5368):174–174. Available from: <http://www.nature.com/articles/239174c0>
 73. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* [Internet]. 1990 Oct; 215(3):403–10. Available from: <https://linkinghub.elsevier.com/retrieve/pii/S0022283605803602> [https://doi.org/10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2) PMID: 2231712
 74. UniProt Consortium T. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* [Internet]. 2018 Mar 16; 46(5):2699–2699. Available from: <https://academic.oup.com/nar/article/46/5/2699/4841658> <https://doi.org/10.1093/nar/gky092> PMID: 29425356

75. Sievers F, Higgins DG. Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. In 2014. p. 105–16. Available from: http://link.springer.com/10.1007/978-1-62703-646-7_6
76. Ileppane. <https://github.com/ileppane/statistics/blob/master/ESTest.R>. 2014.

Recent Advances in Studying Toll-like Receptors with the Use of Computational Methods

Maria Bzówka,* Weronika Bagrowska, and Artur Góra



Cite This: *J. Chem. Inf. Model.* 2023, 63, 3669–3687



Read Online

ACCESS |



Metrics & More

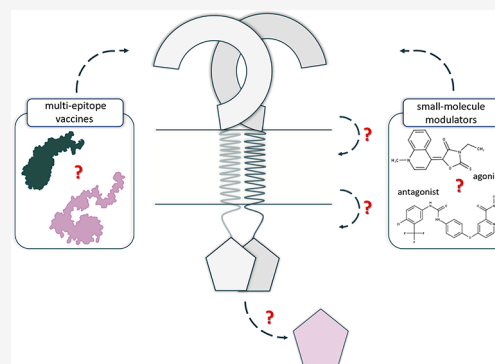


Article Recommendations



Supporting Information

ABSTRACT: Toll-like receptors (TLRs) are transmembrane proteins that recognize various molecular patterns and activate signaling that triggers the immune response. In this review, our goal is to summarize how, in recent years, various computational solutions have contributed to a better understanding of TLRs, regarding both their function and mechanism of action. We update the recent information about small-molecule modulators and expanded the topic toward next-generation vaccine design, as well as studies of the dynamic nature of TLRs. Also, we underline problems that remain unsolved.



KEYWORDS: immune response, pattern recognition receptors, small-molecule modulators, Toll-like receptors, vaccine design, protein–ligand interactions, protein–protein interactions signaling

INTRODUCTION

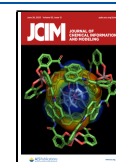
Toll-like receptors (TLRs) represent one of the families of pattern recognition receptors (PRRs) and are an important part of the innate immune system.^{1,2} They are able to recognize various molecular patterns (MPs) in the host organism: damage/danger-, microbial/microbe-, pathogen- or xenobiotic-associated (DAMPs, MAMPs, PAMPs, or XAMPs, respectively).^{3–5} Recognition of those MPs activates downstream signaling cascades that lead to the induction of the innate immune system.^{6–8} In humans, TLRs comprise ten functional members (TLR1–10) that share similar domain organization: an N-terminal domain containing the leucine-rich repeats (LRRs), a single transmembrane helix (TM), and a C-terminal cytoplasmic Toll-interleukin-1 receptor (TIR) domain (Figure 1A). TLR7–9 possess an additional long-inserted loop region (so-called Z-loop) in their LRR domain (Figure 1B) that needs to be cleaved proteolytically. The LRR domain is responsible for ligand recognition, while the TIR domain interacts with adaptor proteins and is responsible for initiating signal transduction. A characteristic feature of the TIR domain in all TLRs is the conserved and functionally important BB-loop (Figure 1C). TLRs are expressed either on the cell surface (TLR1, 2, 4, 5, 6, 10; occasionally TLR7) or in the various intracellular compartments (TLR3, 7, 8, 9; occasionally TLR4). The location of TLRs determines the spectrum of ligands they are able to recognize. For instance, TLRs expressed on the cell surface primarily recognize microbial membranes and/or components of the cell wall, while intracellular TLRs principally recognize nucleic

acids.^{9–11} The full list of the recognized ligands is much larger and has been discussed in several papers.^{11–14} The binding of ligands to a TLR either induces the formation of a receptor dimer or changes the conformation of a preexisting dimer (Figure 1D), which subsequently allows adaptor proteins to bind and trigger an immune response.¹⁵ TLRs can recruit various adaptor proteins; however, myeloid differentiation primary-response protein 88 (MyD88) and TIR domain-containing adaptor protein inducing interferon- β (TRIF) are the most important ones. Two distinct signaling pathways used by TLRs start from them—MyD88-dependent and TRIF-dependent pathways. In general, the MyD88-dependent pathway is utilized by all TLRs, except TLR3, and leads to the production of various proinflammatory cytokines. The TRIF-dependent pathway is utilized by TLR3 and 4 and is associated with the stimulation of type-I interferon^{16–19} (Figure 1E).

Toll-like receptors are a potential therapeutic target in various diseases and conditions. Thus, searching for and designing compounds that can act as agonists or antagonists is the objective of many studies. The distinction between

Received: March 16, 2023

Published: June 7, 2023



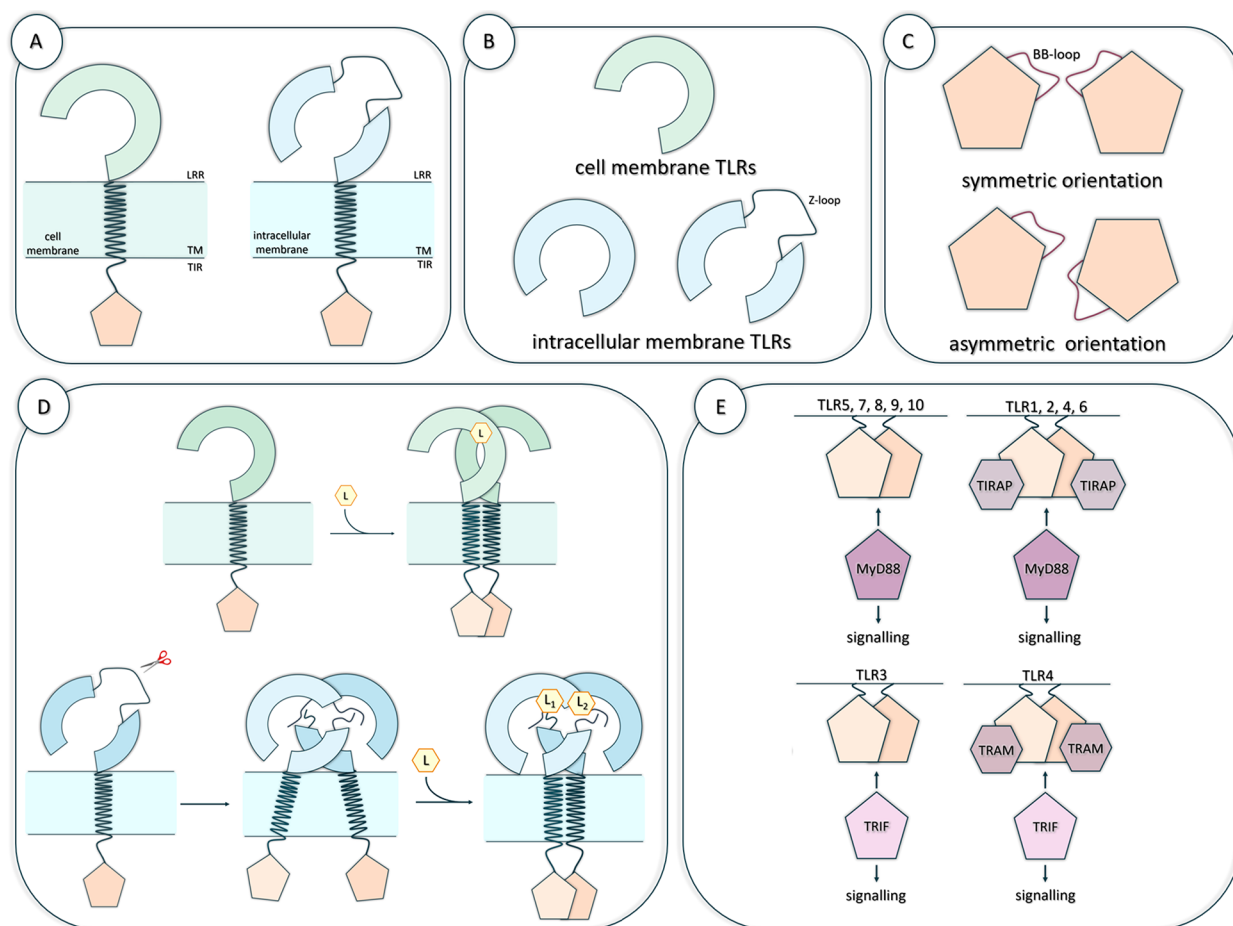


Figure 1. Structural organization and potential Toll-like receptors (TLRs) mechanism of action. (A) The general structure of the TLRs' monomers. (B) Differences in the TLRs' LRR domains between the cell membrane and intracellular membrane TLRs. (C) Various orientations (symmetric and asymmetric) of the TIR domain subunits in the TLRs' TIR dimer. (D) Potential mechanisms of the TLRs activation. The upper panel shows the mechanisms of the cell membrane TLRs activation, while the lower panel presents the mechanisms of the intracellular membrane TLRs containing a Z-loop. (L) indicates the ligand, while the scissors symbol indicates the proteolytic cleavage of the Z-loop. (E) Binding of the adaptor proteins, MyD88 and TRIF, to the respective TLRs' TIR dimer.

agonists and antagonists for TLRs is crucial since they are used to treat different conditions. For instance, TLR agonists have been developed to treat allergies, asthma, different types of cancer, and chronic infections by upregulating the innate immune system. Moreover, since TLRs induce the response of the body's defenses, they are also promising targets for designing vaccines. On the other hand, TLR antagonists have been used to treat many inflammatory conditions such as acute/chronic inflammation, sepsis, chronic obstructive pulmonary diseases, cardiovascular diseases, neuropathic and chronic pain, and various autoimmune diseases.^{20–23}

In recent years, multiple studies have been published, in which TLRs were the main object of research. Particular studies were focused on the following aspects regarding Toll-like receptors: their structure, ligand recognition, signal transduction, and modulator design. Some of these works were done with the use of *in silico* methods. Due to the increase in the use of computational techniques, it was our goal to summarize how various *in silico* solutions have contributed to a better understanding of TLRs. More than five years have passed since the last published reviews on this topic,^{24–26} and we decided to gather the latest relevant results in this paper. We summarized the research conducted so far, while also emphasizing in which areas we still lack knowledge or

solutions. In this work, we focused exclusively on research on human Toll-like receptors (hTLRs).

AVAILABLE STRUCTURES OF TLRs

The first solved structures of hTLRs—TIR domains of TLR1 and TLR2—have been available since 2000,²⁷ while the LRR domain of TLR3 has been available since 2005.^{28,29} In the case of the TM helix, the first structures were elucidated in 2014 as the result of an NMR experiment.³⁰ The vast majority of available structures have been deposited in the Protein Data Bank (PDB)³¹ in the past decade (Supplementary Table S1). However, almost all are single domains of TLRs. Obtaining full-length structures of TLRs remains a challenge. So far, only the LRR and TM domains of TLR3 and TLR7 have been determined together as a result of the Cryo-EM experiment.³² Furthermore, there is a large disproportion in the number of structures between the individual members of the TLRs family. The biggest number of structures has been deposited for the LRR domain of TLR8. In contrast, other TLRs have very few (or none) representative structures of their particular domains. Investigation of the available structures revealed that a part of them miss a number of residues, which worsens their overall quality. Moreover, some deposited LRR domains of TLR1, TLR2, and TLR4 are hybrids of human TLR with hagfish

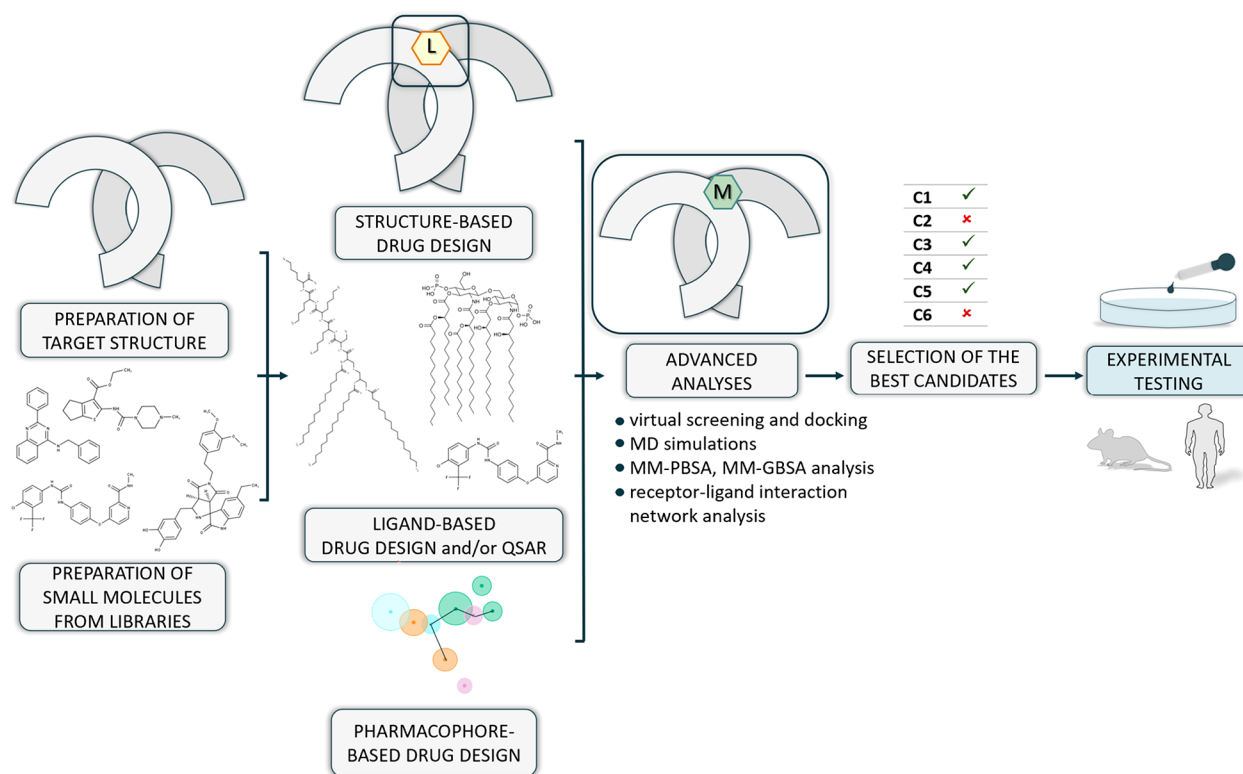


Figure 2. A general protocol for small-molecule modulators design targeting the LRR domain of TLRs. The subunits of the LRR domain are colored gray, indicating both TLRs located in the cell membrane and TLRs located in the intracellular compartments. (L) indicates the location of the ligand binding site, while (M) points out the designed modulator and (C) the selected candidate(s).

variable lymphocyte receptor B. Those factors make not only the structural analysis but also studies on ligand binding, receptor activation, signal transduction, and modulator design not trivial. An interesting combination of computational and experimental approaches was applied for the identification and understanding of the Zn binding to the TIR domain.³³ Lushpa et al. proposed a hypothesis in which Zn²⁺ ions can bind to the TLR1 TIR domain BB-loop and stabilize the conformation of the domain, which interact with TLR2 TIR domain or adaptor proteins. With the use of the NMR experiment, the authors confirmed that the computationally obtained two modes correspond to distinct conformations of the BB-loop and that Zn binding may affect the dynamics and conformational landscape of the BB-loop in the TIR domain. Another example of the use of solution NMR combined with computational simulations has been recently published.³⁴ Kornilov et al. contributed to resolving one of the major “blank spots” in the structure of TLRs, which was the conformation of their transmembrane domains and cytoplasmic juxtamembrane (JM) regions. The authors identified a new structural element, the cytoplasmic hydrophobic JM α -helix, which plays an important role in TLR activation and connects the transmembrane and cytoplasmic parts of the receptor. As they pointed out, the role of the JM region is more complicated than that of a TM-TIR linker and should not be underestimated in further studies.

Recently, we have entered an era where we have gained relatively straightforward access to the prediction of structures. Models of full-length TLRs structures in their monomeric form can be found in the repository of the AlphaFold Protein Structure Database.^{35,36} Still, one needs to remember that in

the case of the predicted structures, they need to be carefully assessed in terms of their quality and usability.

■ COMPUTATIONAL STUDIES ON TLRs

Review articles on computational methods applied in the Toll-like receptors research published before 2017 cover mostly the topics related to designing small-molecule modulators of TLRs.^{24–26} For instance, Murgueitio et al.²⁴ described three main application areas of computational methods to the discovery of TLR modulators: (i) exploration of the structure and function of the receptor, (ii) analysis of receptor–ligand interactions, and (iii) rational design of novel TLR agonists and antagonists by virtual screening (VS). In another work, Pérez-Regidor et al.²⁵ focused almost exclusively on the search for novel chemical modulators for TLRs employing VS techniques. Not only did the authors provide information about the available results for five members of the TLRs family—TLR2, 3, 4, 7, and 8—but also they described the available information about the databases, protocols, and techniques used in virtual screening. In their review, Billod et al.²⁶ focused on TLR4 exclusively and summarized the following aspects: a perspective of the TLR4/MD2/ligand recognition and dimerization, mutant studies, binding mode modulators analysis, and VS strategies for various types of modulators. In 2020 Wang et al. published an article aimed at the progress in developing TLR signaling pathway modulators.³⁷ They mainly focused on the results provided by Yin and Wang laboratories and discussed the identification and characterization of new chemical entities, their modes of action, and further applications. For works that used computational methods, they provided such information in the paper. Based on the results summarized in those reviews, it

is clear that almost all the studies focused on finding small-molecule modulators for the LRR domain of the TLRs. As rightly noted by Wang et al.,³⁷ TM domains are usually considered “undruggable” and TIR domains among TLRs are highly conserved, which is why most modulators are designed to target the LRR domain of TLRs.

Below, we summarized substantial studies that have been published in recent years in which computational methods have been employed. First, we gathered the recent works that focused primarily on designing modulators for TLRs. In particular, we focused on two types of modulators: small-molecule and vaccine components. While small-molecule compounds have been extensively studied, vaccine components have not been reviewed in detail. Second, we reviewed studies principally focused on the investigation of the dynamic nature of TLRs, which is crucial for understanding their function and mechanism of action.

■ MODULATORS OF TLRs

The search for new chemical entities as potential TLRs modulators is an ongoing process, especially because relatively few compounds with therapeutic potential have been tested in clinical trials. Additionally, the use of a strategy involving the TLRs as a driving force for the design of next-generation vaccines has become increasingly popular recently. Since different types of modulators (small-molecule or part of the vaccines, e.g., epitopes) require various methods and techniques for their identification, we reviewed both classes separately.

Novel Potential Small-Molecule Agents. The general protocol used for the search for novel small-molecule TLRs modulators has remained the same in most of the studies conducted so far. It consists of the following steps: (i) preparation of the target structure, (ii) preparation of small molecules from available libraries, (iii) structure-, ligand, and/or pharmacophore-based virtual screening combined with molecular docking, (iv) selection of best candidates, (v) experimental testing, and (vi) identification of potential drug candidates. Before the selection of the best candidates, more advanced computational methods are sometimes used, e.g., molecular dynamics (MD) simulations, MM-PBSA, MM-GBSA binding free energy analysis, combined with receptor–ligand interaction network analysis (Figure 2). By applying those advanced methods it is possible to gain better insight into the molecular basis of ligand recognition. Usually, all-atom MD simulations of the receptor–ligand complex are performed.

For VS, scientists have various commercial, public, or in-house databases at their disposal. Many groups have concentrated on modifying the previously identified small-molecule compounds or mimicking the native ligands within known binding sites. Nevertheless, there are also examples revealing novel chemical classes of potential modulators. Studies conducted so far are still mainly focused on targeting the LRR domain of TLRs. There has been no noticeable progress in the design of modulators for the TIR domain.

Many recent studies have been carried out on TLR2. For instance, Murgueitio et al.³⁸ performed a shape- and feature-based similarity VS (with the use of ROCS software) to screen some commercially available databases (LifeChemicals, Maybridge, Chembridge, Enamine HTS Collection, Asinex, and Specs). For the similarity search, they used the previously discovered TLR modulators from Guan et al.³⁹ and Liang et

al.⁴⁰ The authors tested selected hits, and four (AG1–AG4) were found to synergistically increase the nuclear factor kappa-light-chain-enhancer of activated B cells (NF- κ B) activation induced by the known lipopeptide ligand Pam₃CSK₄. Further studies indicated that the tested compounds could act as allosteric modulators of TLR2. To investigate the binding modes of the identified compounds, the authors run docking calculations (GOLD Suite), using the crystal structure of the human TLR2/1 heterodimer in complex with Pam₃CSK₄ (PDB ID: 2Z7X). They inspected the docking poses in LigandScout and identified a putative binding site in the vicinity of the Pam₃CSK₄ binding site which is formed during the heterodimerization process. Information about the characterized interacting residues can be found in [Supplementary Table S2](#). For other compounds described in the following parts of this section, details on the identified interacting amino acids (if available in the cited publications) are also provided in [Supplementary Table S2](#).

Durai et al.⁴¹ used receptor–ligand- and ligand-based VS to prepare the pharmacophore models and to screen in-house libraries comprising nearly seven million compounds. They focused on the nonpeptide TLR2 antagonists, distinct from several known inhibitors with fatty acid chains. For the receptor–ligand-based model, the authors prepared the protein–lipopeptide complex (PDB ID: 2Z7X),⁴² while for the ligand-based model, they selected compounds from Guan et al.³⁹ They used the Receptor–ligand Pharmacophore Generation and Common Feature Pharmacophoric Generation protocols (Discovery Studio Visualizer Software, 4.0), respectively. The next steps involved screening the compounds that mapped to the pharmacophore features and filtering them using Lipinski and Veber rules, as well as ADMET properties. The authors evaluated the best hits for their ability to bind directly to the lipopeptide binding site of the human recombinant TLR2. For that, they performed a two-step molecular docking (CDOCKER and AutoDock Vina) and tested the selected protein–ligand docking complexes by MD simulations combined with MM-PBSA binding free energy calculations (GROMACS with CHARMM27 force fields and SPC216 water model). The authors selected promising TLR2/1 antagonists using surface plasmon resonance experiments and tested their ability to inhibit the synthesis and secretion of IL-8 in human embryonic kidney cells overexpressing TLR2. Two molecules—C11 and C13—displayed both direct binding to TLR2 extracellular domain and reduced Pam₃CSK₄-induced IL-8 production. Those antagonists showed no toxic effect in cell viability assays and seemed to have good pharmacological properties. The results supported the possibility that C11 and C13 can disrupt TLR2/1 heterodimerization.

Chen et al.⁴³ performed a structure-based VS (Glide) of the ZINC database. Based on the scoring results, including shape, chemical-feature, and drug-like properties, they identified potential agonists targeting the TLR2 heterodimer and modulating the TLR2/1 response. For the most promising candidates, which shared a motif of an amine conjugated with an acid substituent, they tested their activity *in vitro*. The results revealed that two compounds showed a high TLR2 activation effect and that one compound—ZINC6662436 (SMU127)—stimulated the NF- κ B and promoted tumor necrosis factor- α in human macrophage and mononuclear cells. Also, the *in vivo* results showed signs of inhibition of breast cancer tumor growth in BABL/c mice. In a later study,

Chen et al.⁴⁴ improved the potency of the SMU127 by modifying the ring system, while keeping all other structural features. One of the modified compounds—SMU-C13 possessed the highest TLR2 activity. This compound was docked into the 2Z7X structure (Glide) and evaluated regarding its putative binding. The *in silico* simulation indicated a tight fit into the known binding site of Pam₃CSK₄ and TLR2/1. Based on the structure–activity relationship (SAR) results, the authors concluded that the introduced piperidine ring contributed to the increased activity against TLR2.

Grabowski et al.⁴⁵ performed both ligand- and structure-based VS using commercial databases of nearly six million compounds (Asinex, LifeChemicals, Mybridge, Chembridge, Enamine, Otava, Specs, Vitas-M, KeyOrganics, and ChemDiv). The authors selected two well-characterized chemotypes of small-molecule modulators to build their models—(i) m1 proposed in previous work by Murgueitio et al.⁴⁶ and (ii) CU-CPT22 and the other benzotropolones discovered by Yin et al.⁴⁷ For that, they applied a standard protocol for pharmacophore-based screening (LigandScout). For all modeling studies, the authors used a TLR2 monomer from the TLR1-TLR2 heterodimer (PDB ID: 2Z7X). Screening of compounds was followed by their filtering using shape- and feature-based properties. Then, the authors carried out docking (GOLD), rescoring, and visual inspection analyses and selected the best hits for biological testing to confirm their ability to inhibit TLR2-mediated responses. Selected compounds were tested in HEK293-hTLR2 cells, THP-1 macrophages, and peripheral blood mononuclear cells. The most active compound, a pyrogallol derivative named MMG-11 inhibited both TLR2/1 and TLR2/6 signaling. It also showed a higher potency than the previously discovered CU-CPT22. Additionally, in a subsequent paper,⁴⁸ Grabowski et al. confirmed that another potent compound (named compound 8) showed a TLR2 inhibition and additionally reduced TLR7/8 responses. Encouraged by these results, they applied a computationally guided synthesis approach to get an analogue of that compound which showed dual inhibition of TLR2 and 8. For docking studies (GOLD), the authors used the crystal structures with cocrystallized ligands; SWYZ for TLR8 and 2ZJX for TLR2/1. The authors selected the putative binding modes based on pharmacophore fit rescoring using previously reported TLR2 antagonist MMG-11 and CUCPT9b for TLR8. The results showed that the selected compound 24 is able to simultaneously and selectively target both surface- and endosomal-located TLRs. This compound showed also high efficacy with low cytotoxicity and a noncompetitive antagonist behavior. Also, in another work, Bermudez et al.⁴⁹ explored the chemical space around the pyrogallol-containing antagonists to improve synthetic accessibility and chemical stability.

Boger's lab proposed a new and potent class of TLRs agonists—diprovocims.⁵⁰ They obtained results from a compound library designed to promote cell surface receptor dimerization. The discovered class of compounds had no structural similarity to any known natural and synthetic TLR agonist, and selected members were confirmed to be active in both human and murine systems. Comprehensive SAR studies improved the potency 800-fold over the screening leads, providing diprovocim-1 and diprovocim-2. The compound 3 of the diprovocim-1 scaffold, later referred as Diprovocim, showed full agonist activity at very low concentrations in human THP-1 cells, being more potent than any other known small-molecule TLR agonist. Later, the basis of TLR2/TLR1

activation by Diprovocim was studied by Su et al.⁵¹ They combined analysis of the structures of Diprovocim-bound TLR2 homodimer and TLR2/TLR1 in a complex with Pam₃CSK₄ with docking, MD simulations (AMBER with ff14SB and GAFF force fields and TIP3P water model), MM-PBSA, MM-GBSA binding free energy and mutagenesis analyses. *In silico* results indicated that binding two Diprovocim molecules to the TLR2/TLR1 heterodimer was slightly less energetically efficient than binding a single molecule. Further analyses revealed that the new modulator interacts with TLR2/TLR1 at the same binding pocket as Pam₃CSK₄. However, the observed conformations around the ligand binding sites were different. The Diprovocim-bound TLR2 homodimer showed a larger distance between the C-termini of the TLR2 LRR domain than the Pam₃CSK₄-bound TLR2/TLR1 heterodimer, suggesting that the TLR2 homodimer may not be able to activate downstream signaling. The authors noticed the widespread hydrophobic interactions and a hydrogen-bonding network between the receptor and Diprovocim molecules within the ligand binding pocket, while in the Pam₃CSK₄-bound receptor complex, such a network was absent. These differences could explain the greater potency of Diprovocim in activating TLR2/TLR1-mediated signaling. The mutagenesis analysis was focused on the identification of which amino acid on TLR1 and TLR2 are important for the binding of Diprovocim and Pam₃CSK₄, and all details can be found in the paper of Su et al.⁵¹

For the TLR4 receptor associated with myeloid differentiation factor 2 (MD2), Mishra and Pathak⁵² aimed at the identification of small-molecule protein–protein inhibitors based on a pharmacophore mapping-based approach. For that, they used information about the generated hot-spot residues (DrugScore^{PP1}, KFC2, HotPoint, HotRegion) and their corresponding pharmacophoric features (PocketQuery and ZINCpharmer) on the protein–protein interaction interfaces in the TLR4/MD2 homodimer complex (PDB ID: 3FXI). The authors ran VS with molecular docking (FlexX) and performed extensive post-VS filtration based on ADMET properties, oral bioavailability, and possible side effects—off-targeting and environmental hazard. From selected hits, two (C11 and C15) with the predicted best inhibitory concentration were confirmed to form a stable complex with the target protein during MD simulation analysis (NAMD with CHARMM force field and TIP3P water model). In other studies, Facchini et al.⁵³ and Cochet et al.⁵⁴ focused on designing the monosaccharide mimetics of lipid A, which is a known agonist. The authors successfully designed mimetics through docking with MD2 (AutoDock Vina and AutoDock) and confirmed the stability of the modulators by performing MD simulations (AMBER). The compounds were predicted to bind inside the MD2 hydrophobic pocket with favorable predicted binding scores. Subsequently, compounds were synthesized and tested to confirm their ability to bind to MD2 and inhibit LPS-stimulated TLR4 activation.

In general, many known TLR4 modulators are LPS mimetics; however, alternative strategies for finding non-LPS-like modulators have also been applied. A lot of studies focused on the use of opioids and their derivatives. For instance, morphine, cocaine, and methamphetamine (METH) were found to interact with TLR4 to initiate neuroimmune signaling.^{55–57} In their work, Wang et al. performed docking (AutoDock Vina) of METH to the TLR4 receptor (PDB ID: 3VQ2) to investigate how the compound interacts with TLR4/

MD2. METH was docked into the dimerization interface of the TLR4/MD-2 complex, and further MD simulation (NAMD, AMBER force field) suggested that the binding of the compound stabilizes the TLR4/MD-2 tetrameric form, which could shift the equilibrium and potentially activate TLR4 signaling as a nonclassic agonist. In another work, Wang et al.⁵⁸ revealed the molecular mechanism of (+)-naltrexone and (+)-naloxone underlying the effects of opioid isomers on TLR4 signaling as the first biased inhibitors of TLR4, which inhibit only the TRIF-dependent signaling with no effects on the MyD88 signaling. These results became the basis for the design of more promising TLR4 antagonists based on known opioids. For instance, Selfridge et al.⁵⁹ designed and synthesized compounds based on (+)-naltrexone and (+)-noroxymorphone. In another study, Zhang et al.⁶⁰ used the previously established protocols to investigate in detail the molecular interactions between (+)-naltrexone, its derivatives, and MD2 of TLR4. Results showed that hydrophobic residues in the MD2 cavity interacted directly with the (+)-naltrexone-based TLR4 antagonists and were essential for ligand binding. Increasing hydrophobicity of the substituted group improved TLR4 antagonistic activity, while charged groups disfavored binding with MD2. MD simulations (NAMD with AMBER03 and GAFF force fields and TIP3P water model) demonstrated that (+)-naltrexone or its derivatives bound to MD2, stabilized its conformation, and blocked TLR4 signaling. The idea of improving naltrexone-based compounds was also developed in later works. An example is the work of Zhang et al.,⁶¹ who designed bivalent ligands by connecting two naltrexone units through a rigid pyrrole spacer. In a very recent study, Pérez-Regidor et al.⁶² focused on finding non-LPS-like modulators among the approved drugs and drug-like molecules from commercial, public, and in-house libraries of compounds. Based on the structure-, ligand-based VS and docking (FLAP, GLIDE, AutoDock and AutoDock Vina) combined with biological results, the authors presented a common scaffold consisting of two hydrophobic moieties separated by a polar linker. They showed that one large hydrophobic moiety occupies the hydrophobic MD2 cavity, while the second moiety is associated with the same hydrophobic region as one of the lipid A alkyl chains, and the polar linker occupies the entrance to the pocket. Another approach was proposed by Gao et al.⁶³ They focused on the computational design of macrocyclic peptides (Rosetta Peptidrive), based on the fragment of MD2 mediating the association of the TLR4/MD-2 complex. The authors synthesized proposed constructs and experimentally evaluated their ability to activate the TLR4 signaling. Application of such approach could potentially overcome the existing problem of targeting protein–protein interaction interfaces which are usually flat and may not be suitable for binding of organic compounds.

An interesting study was performed by Borges et al.⁶⁴ The authors investigated the effect of the natural limonoid gedunin on different TLRs (2, 3, and 4) activation. They performed *in vitro*, *in vivo*, and *in silico* studies. The experimental results confirmed that gedunin is able to impair inflammasome activation, and cytokine production and induce anti-inflammatory factors in macrophages. The docking studies (AutoDock Vina) revealed that the investigated compound can efficiently bind to the TLR2, TLR3, MD2 protein of TLR4, and also to the caspase-1, making gedunin considered a multitarget compound. The authors used the following PDB structures: human caspase-1 (PDB ID: 1RWX), TLR2 (PDB

ID: 1O77), and TLR3 (PDB ID: 1ZIW). For both TLR2 and TLR4, gedunin bound within the known ligand binding site, while for TLR3 two distinct binding sites were predicted. The authors pointed out that one of the predicted regions for TLR3 is involved in the dimerization of TLR3 and is considered the dsRNA binding site, thus it might be the most prominent. Still, as pointed out by the authors, further biochemical assays are required to confirm gedunin binding.

For endosomal TLRs—TLR3 and TLR7–9—Talukdar et al.⁶⁵ recently published a perspective paper regarding the structural evolution of their small-molecule agonists and antagonists. They concluded in detail information about structural features around binding sites of both types of modulators, and their evolution and provided information about the development of various chemotypes, e.g., guanosine-, oxoadenine-, 3-deazapurine-, imidazoquinoline-, quinoline-, benzimidazole-, imidazole-, pyridopyrimidine-, pyrrolopyrimidine-, pyrimidine-, quinazoline-, chromene-, benzoxazole-, indole-, triazole-, indazole-, and benzanilide-based.

Here, we wanted to highlight a few studies not included in the above-mentioned publication. One example is the work performed by Gupta et al.⁶⁶ They used the known ligand-based pharmacophore modeling approach to find novel human TLR7 modulators based on the set of TLR7 agonists with confirmed experimental activity. The data set was divided into training and test sets based on criteria such as structural diversity and activity range. The authors created a pharmacophore model (HypoGen algorithm available in 3D-QSAR pharmacophore generation protocol of Discovery Studio) and screened the natural hit compounds from the InterBioScreen Natural product database. They filtered the screened compounds and based on molecular docking (LibDock) and further interaction analyses, they selected the most interesting compound - STOCK1N-65837 (an indoline derivative natural alkaloid). The compound was further validated with MD simulation (GROMACS with GROMOS96 43a1 force field and SPC216 water model). The authors found that STOCK1N-65837 formed hydrogen bond interactions with residues from LRR15 and LRR16 of hTLR7, which was in the good agreement with previous findings that amino acids within that region crucial for ligand binding. The authors underlined that further experimental validation is necessary to confirm the activity of the compound; however, their results already provided a basis for further designing of natural modulators targeting TLRs.

Sribar et al.⁶⁷ used the previously established approach consisting of structure- and pharmacophore-based computational studies (ROCS, GOLD, LigandScout), combined with MD simulations (Desmond), followed up by experimental validation to find novel inhibitors of TLR8. They performed two rounds of VS. The authors used the best hit from the first round of VS and performed its optimization by shape- and chemistry-based screening. Later, they prioritised them according to their diversity and physicochemical properties. Based on that approach, they found a novel pyrimidine scaffold for TLR modulators. Experimental validation of the most promising compounds from the second round of VS revealed their low cytotoxicity, suggesting that they are relevant for further lead optimization.

Recently, Wang et al.⁶⁸ focused on revealing the mechanism of action of known agonists for TLR7 and TLR8 - imidazoquinoline derivatives (Resiquimod (R), Hybrid-2 (H), Gardiquimod (G)). They carried out MD simulations (GROMACS with AMBER ff99SB and GAFF force fields and

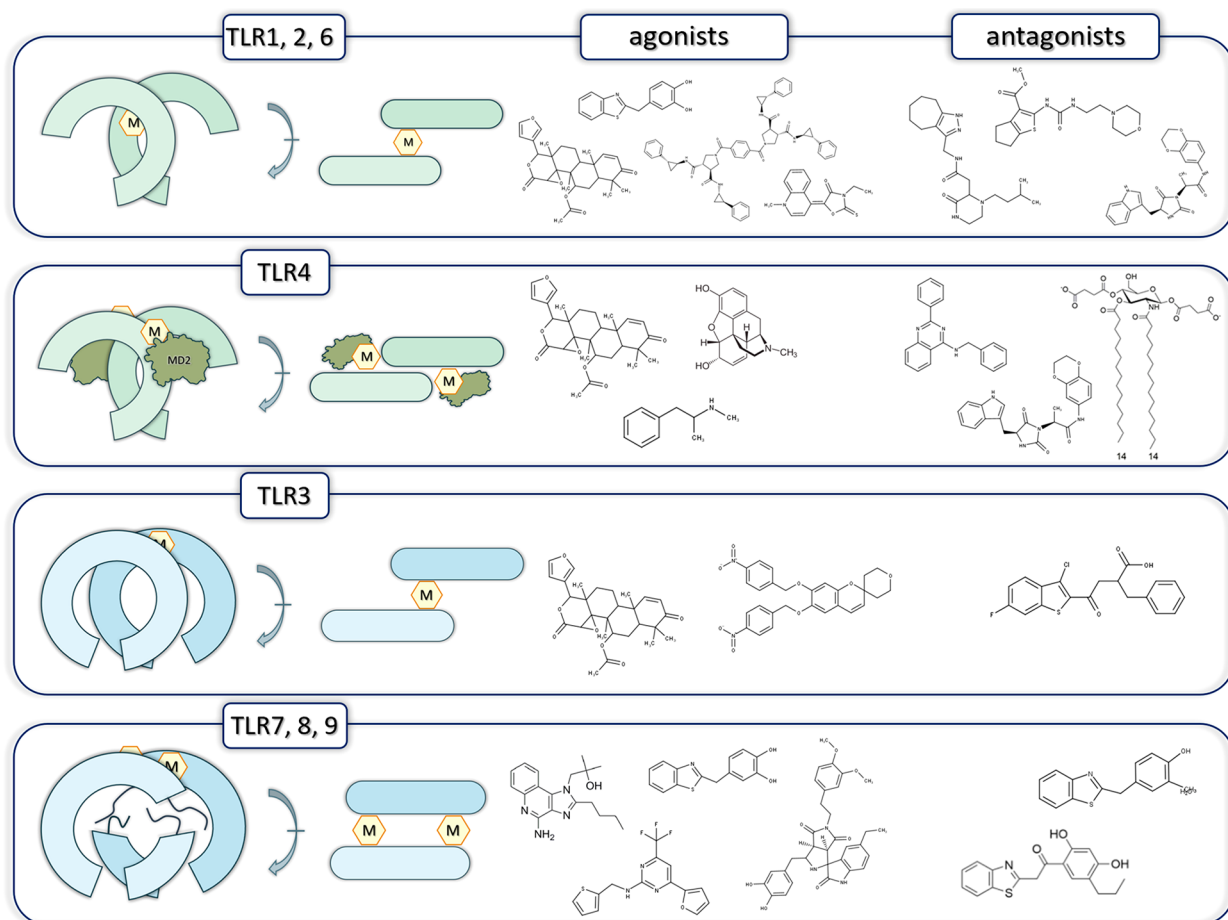


Figure 3. Examples of scaffolds of small-molecule modulators targeting the LRR domain of TLRs. The left panel shows the approximate location of small-molecule modulators (M) with respect to the LRR subunits of the TLR dimers described in this review. Agonists are presented on the middle panel, while antagonists are on the right panel. TLR4 was shown with the associated myeloid differentiation factor 2 (MD2). Please note that one of the agonists' scaffolds is shown for more than one TLRs. This indicates the possibility of targeting both surface- and endosomal-located TLRs by a given modulator.

SPC water model) for both TLR7 and TLR8 apo structures and TLR7 and TLR8 with bound antagonists, followed by the MM-GBSA calculations. Their analysis showed that TLR7-R and TLR7-G complexes formed open conformations during the simulation, while the others were kept in closed conformations. They found that the binding pocket of TLR7 was less flexible than in TLR8, thus the binding of the antagonist was tighter. Moreover, these *in silico* predictions were in agreement with the experimental data.

In Figure 3 we presented examples of scaffolds of both agonists and antagonists targeting the LRR domain of TLRs proposed in reviewed publications. Also, we showed the localization of the designed small-molecule modulators in relation to the subunits of the TLRs. In Supplementary Table S2 we gathered the structures of all the best hits from the reviewed research papers, as well as the information about the interacting amino acids (if available).

As can be seen from the above-mentioned studies, many groups used the information from the previously designed modulators either for introducing some modifications aimed at increasing their activity or for obtaining models for VS and further studies. In the reviewed papers we encountered both the strategy to design modulators structurally similar to known ligands and compounds with a completely different structure. Interestingly, the targeting sites remain the same, which

highlights the challenges in the reconstruction of TLRs structure and difficulties with the identification of other potential binding sites which could affect TLRs function. We could also notice that some of the proposed modulators were able to influence the signaling pathways in various TLRs. Nevertheless, the molecular bases of their selectivity have not been thoroughly examined. Therefore, one needs to keep in mind that we still need in-depth studies revealing the differences in the mechanism of action in relation to different receptors. We believe that in the coming years, more groups will include analyses related to potential off-targeting effects, as well as that there will be an increase in interest in the screening of natural compounds databases for proposing novel small-molecule modulators. Regarding methods, we are expecting an increased contribution of AI-supported screening, especially in ligand-based screening.

Next-Generation Vaccines. Subunit vaccines are considered one of the next-generation vaccines. They consist of pieces of a pathogen, instead of the whole organism. Evidently, this also means they do not contain any live pathogen and thus show significantly lower immunogenicity. The immunogenicity of the subunit vaccines can be improved by several factors, e.g., addition of adjuvants, choice of different delivery systems, usage of multiple antigens or epitopes, and optimization of vaccine dosage. TLRs are excellent targets for such multi-

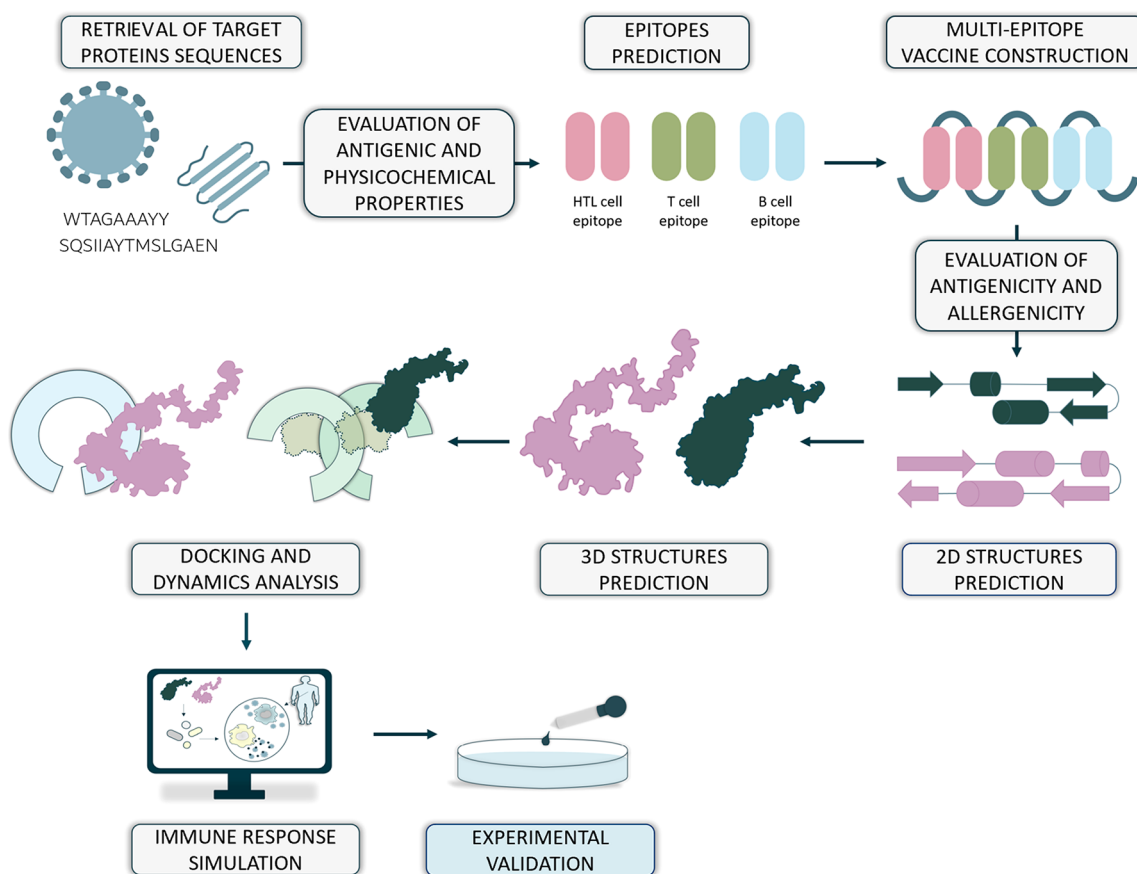


Figure 4. General protocol for next-generation multi-epitope vaccine design. The ability of binding different epitopes (shown as dark green and pink shapes, respectively) to LRR subunits of the TLRs located both in the cell membrane (light green) and in the intracellular membrane (light blue) has been shown.

epitope vaccines to provide a signal to induce an effective immune response that in turn leads to long-lasting protection.^{23,69,70} The protocol used for the search for multi-epitope modulators is substantially different from the one used for small-molecule modulators. The general protocol consists of multiple steps: (i) retrieval of target proteins sequences, (ii) evaluation of antigenic and physicochemical properties of the target proteins, (iii) epitopes prediction, (iv) multi-epitope vaccine construction, (v) evaluation of antigenicity and allergenicity of the vaccine combined with the exploration of the physicochemical parameters, (vi) prediction of secondary and tertiary structure, (vii) molecular docking to the immune receptors, and (viii) dynamics' analysis of the complexes. Some studies also include further computational immune simulation to assess the vaccine's ability to stimulate the immune response (Figure 4).

Each step of this protocol is quite elaborate and usually requires the usage of several tools/servers. As information about vaccine construction has not previously been addressed in computational reviews about TLRs, a brief summary is given here. Target sequences might be obtained from databases like PDB or UniProt.⁷¹ Then, they are submitted, e.g., to the VaxiJen⁷² to check the antigenicity and to ExPASyProtParam⁷³ to investigate the physicochemical properties. Multiple servers can be used to predict the epitopes, depending on the type. Among them, there are NetCTL,⁷⁴ NetMHCIIpan,⁷⁵ Immune Epitope Database,⁷⁶ BepiPred,⁷⁷ and BCPREDS.⁷⁸ Antigenicity, promiscuity, and allergenicity of epitopes can be evaluated

with the use of AllerTop,⁷⁹ AlgPred,^{80,81} VaxiJen, and ToxinPred^{82,83} servers. Structural evaluation of the vaccine begins with the prediction of secondary structure, which is usually done by the SOPMA server.⁸⁴ Later, the tertiary structure can be predicted, often by the I-TASSER.⁸⁵ However, the obtained models still need further refinement. For that, ModRefiner⁸⁶ and GalaxyRefine⁸⁷ are common choices. At this stage, it is evident that the way to obtain a structure of this type of modulator is quite demanding. Molecular docking of the epitope involves predicting the proper orientation and conformation of the epitope when it interacts with the immune receptor's binding site. The ClusPro server⁸⁸ is able to perform such computations. Further investigation of the dynamical properties is usually performed using Normal Mode Analysis (NMA) rather than all-atom MD simulations. However, the latter one (if used) can provide better and more detailed insight. A simulation of a possible immune response, which usually concludes the *in silico* part, is often performed using the C-ImmSim tool.⁸⁹

In studying TLRs, molecular docking combined with the investigation of the dynamical stabilities and prediction of the vaccine's ability to stimulate the immune response are the most crucial. The above-mentioned protocol and its variations have been used multiple times for vaccine design. Undoubtedly, vaccines against severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) have received the most attention in recent years.^{90–93} However, studies on other vaccine designs have also been carried out, both before and after the outbreak of

COVID-19. The following examples are studies focused on designing vaccines against Middle East respiratory syndrome (MERS),⁹⁴ Hepatitis C virus (HCV),⁹⁵ human immunodeficiency virus (HIV),⁹⁶ Neo-Coronavirus (NeoCoV),⁹⁷ Human cytomegalovirus (HCMV),⁹⁸ Kaposi Sarcoma,⁹⁹ as well as infections like dengue,¹⁰⁰ chikungunya,¹⁰¹ or those caused by *Taenia solium*,¹⁰² *Klebsiella oxytoca*,¹⁰³ *Klebsiella pneumoniae*,¹⁰⁴ or *Mycobacterium tuberculosis*.^{105,106} What is also worth mentioning in the context of next-generation vaccine design is the potential use of TLR agonists as vaccine adjuvants. Since TLR agonists are capable of stimulating innate immune responses, which also trigger adaptive immune responses, they can likewise be used to improve vaccine efficacy.^{69,107,108} For instance, monophosphoryl lipid A (MPL) and CpG-1018 have been used as adjuvants in licensed vaccines, and other TLR agonists are under the investigation.

Below, we want to elaborate more on vaccines against SARS-CoV-2, although the ultimate goal remains similar in all the studies—to get a stable protein-vaccine complex that triggers the immune response.

Different groups focused on studies of multiepitope vaccines against various TLRs. For instance, Oladipo et al.⁹⁰ studied the TLR2, TLR3, TLR4, and TLR9, while Rafi et al.⁹¹ focused on TLR2 and TLR4, and Ysrafil et al.⁹² investigated TLR3, TLR4, and TLR8, as well as angiotensin-converting enzyme 2 (ACE2) as the entry receptors of SARS-CoV-2. Drawing upon the structure of the SARS-CoV-2 spike (S) glycoprotein (and nucleocapsid (N) protein and open reading frame 1a (ORF1a) protein in the case of Ysrafil et al.⁹²), the authors tried to develop a potent multiepitope subunit vaccine. The groups received different predictions of the epitopes, depending on the particular settings used while executing the general protocol which was described earlier. Therefore, the final models of the multiepitope vaccine constructs were different, dependent on the sequences that build the individual epitopes. Here, we wanted to provide more details about the interesting study proposed by Pitaloka et al.⁹³ The authors focused on designing a vaccine for protection against *Mycobacterium tuberculosis* (MTB) and SARS-CoV-2 coinfections. They used web servers—Bepipred-2.0 for B-cells epitopes, NetCTL.1.2 for Cytotoxic T Lymphocytes (CTL) epitopes, and Net MHC II pan 3.2 for Helper T Lymphocyte (HTL) epitopes—to screen potential epitopes from outer membrane protein A Rv0899 (OmpATb) of MTB and S protein of SARS-CoV-2. Epitope domains were selected from identified immunodominant areas and filtered out (by BLASTp) based on shared homology with humans. Then, at the vaccine's N-terminus, the authors introduced the 50S ribosomal protein L7/L12 adjuvant using a commonly used EAAAK linker, while AAY and GPGPG linkers were used to connect the particular epitopes. In general, all the results showed that the proposed multiepitope vaccine candidates were nontoxic, capable of initiating the immunogenic response and not inducing an allergic reaction. Also, the molecular docking results revealed rather strong and stable interactions between the constructed vaccines and particular receptors within their LRR domains. During the computational simulations of the potential immune response using the C-ImmSim tool, the authors noticed a rise in the production of immune defenses, i.e. rise in the HTL cell population with memory T and B cells development, an increase in IgM, IgG1 + IgG2, and IgG + IgM antibody levels. The stability of the complexes of various vaccines was confirmed by studying their dynamic properties. For instance,

Oladipo et al.⁹⁰ and Pitaloka et al.⁹³ used NMA to study the stability and mobility of selected receptor–vaccine complexes. In the first study, as a result, the vaccine protein and its receptor were predicted to spin toward each other. In the second study, based on the detected correlations in the covariance matrix between pairs of residues, the authors confirmed the stability of the vaccine candidate model. Rafi et al.⁹¹ performed classical MD simulations to check the stability of the constructed vaccine with the extracellular subunit of TLR2 and TLR4/MD2. The results indicated that the TLR–vaccine complexes were both stable and compact during the simulations. Especially for the TLR4–vaccine complex, a strong hydrogen bond network was pointed out, suggesting reduced flexibility of the vaccine when bound to the receptor, improved binding strength, and increased vaccine–receptor stability. Furthermore, the authors expanded their analysis by using the full-length heterodimer TLR4/MD2–vaccine complex, which was placed in a membrane to imitate the dynamic behavior during the MD simulation of the vaccine in biological systems. This study is one of the first where the full-length models of TLR receptors from the AlphaFold Protein Structure Database were used. For both TLR2 and TLR4 complexes, significant structural transitions toward membrane bilayer were observed, but the crucial interactions between the vaccine and the extracellular domain of receptors remained stable. Based on the observations made in the above-mentioned papers, one can speculate that during the binding, potentially well-designed vaccines may have a stabilizing effect on the TLRs in the system.

Although at first glance epitopes may be treated similarly to small-molecule modulators, the specificity of their search is quite different. It takes into account not only the process of binding to the TLR but also the stability and specificity of the epitope. Research on epitopes has the potential to reveal the mechanism of action of TLRs and their specificity to a greater extent. In the near future, this type of research can contribute to a much better understanding of the functioning of our immune system and the recognition of threats. We also anticipate that the contribution of AI-based methods will allow for a better understanding of the signaling pathways and their interrelations.

■ DYNAMIC NATURE OF TLRs

The complexity of TLRs has consequences in the relatively weak understanding of the structural basis of their modes of action. Therefore, significant effort is required to comprehend TLR dynamics at the level of particular domains, the full-length receptor, and the dimerization process. Here, in the first part, we gave an outline of the studies that examined the effect of certain mutations on the receptor's dynamics. In the second part, we summarized the works that focused on the characterization of the dynamical properties and conformational changes of full-length TLRs.

Mutations' Effects on the TLRs Dynamics. It is known that even a single mutation can induce substantial changes in terms of the macromolecule's structure and function. For TLRs, one can hypothesize that depending on the mutation location, the ligand recognition or the adaptor protein binding could be disturbed. Below, we summarized studies focused on examining the effect of various mutations on TLRs. Those studies have usually focused on the analysis of individual domains of TLRs—the LRR or TIR domains.

Regarding the LRR domain, Anwar and Choi¹⁰⁹ examined the structure–activity relationship in TLR4 mutants by the application of MD simulations (GROMACS with AMBER99SB-ILDN force field and TIP3P water model) together with principal component (PCA) and residue interaction network (RIN) analyses (RINalyzer, Cytoscape). To evaluate the influence of single nucleotide polymorphisms (SNPs), they examined four different models: (i) wild-type TLR4 (TLR4WT; PDB ID: 3FXI); (ii) a double mutant—aspartic acid-to-glycine at position 299 and threonine-to-isoleucine at position 399 (TLR4GI; PDB ID: 4G8A); (iii) the aspartic acid-to-glycine mutant (TLR4G299); and (iv) the threonine-to-isoleucine mutant (TLR4I399). Those mutations were classified as eliminating signaling activity; however, they did not disturb the ligand recognition nor did they establish contact with the associated MD2 protein. The single mutant structures were generated with the use of Chimera software. Computational studies revealed differences in the dynamic properties of the analyzed variants. The authors pointed out that the mutated complexes were less cohesive and displayed both local and global variation in the secondary structure, which could affect the proper exploration of conformational phase space. In particular, results from PCA confirmed that the mutated variants displayed unique low-frequency motions, which could be linked to the differential behaviors in these TLR4 variants. The authors also showed that decay in the rotational correlation function together with the observed density distributions and alteration of the number of hydrogen bonds between the protein and ligand could result in the loss of function.

Gosu et al.¹¹⁰ performed MD simulations (GROMACS with AMBER-ff99SB-ILDN force field and TIP3P water model) of human wild-type and mutant TLR3 to get insights into the dynamic nature of the dsRNA-bound TLR3 complex. They investigated several complexes: dsRNA-unbound TLR3 wild-type dimer (apo_dTLR3WT), dsRNA-bound TLR3 wild-type dimer (dTLR3WT-dsRNA), dsRNA-bound TLR3 dimer with a leucine-to-phenylalanine mutation at position 412 (dTLR3L412F-dsRNA), and dsRNA-bound TLR3 dimer with a proline-to-leucine mutation at position 680 (dTLR3P680L-dsRNA). In TLR3, L412F polymorphism was associated with several human diseases, while the P680L mutation was found as one that reduces the binding affinity of dsRNA to TLR3 and affects subsequent signaling. A human TLR3 dimer model was built by homology modeling using the mouse TLR3 dimer crystal structure (PDB ID: 3CIY) as a template to obtain an accurate structure conformation. The mutations were introduced using Discovery Studio Visualizer. The authors performed MD simulations (GROMACS with AMBER-ff99SB-ILDN force field and TIP3P water model) together with PCA, RIN, hydrogen bond, and protein–nucleic acid interaction analyses to investigate the global motions and the distribution of crucial residues for signal transduction. They claimed that the apo wild-type preformed dimer is unlikely to be stable in physiological conditions. Thus, they proposed that TLR3 might exist as a monomer in a solution. Further, the interaction energies and hydrogen bonds analyses indicated that the mutations induced certain conformational changes that could disturb the TLR3 signaling. The interaction sites between TLR3 and dsRNA were observed at both the N-terminal and C-terminal ends of TLR3 LRR, while the dimerization interface was confirmed at the C-terminal site but only for dTLR3WT-dsRNA and dTLR3L412F-dsRNA. It

might suggest that P680 is crucial for maintaining the dimer interface for ligand binding. This hypothesis seems to be confirmed by the MD simulations in which the mutation dTLR3P608L disrupted the dimer interface in two out of three runs.

In the case of TLR3, we also want to underline one of the possible post-translational processes that the protein may undergo—glycosylation. TLR3 is a receptor with multiple glycosylation sites. Although most of these sites are not associated with dsRNA recognition, the *N*-glycan located at N413 has been observed to be in direct contact with viral dsRNA. In their work, Sun et al.¹¹¹ reported that mutations of two independent glycosylation sites (N247and/or N413) in TLR3 resulted in the abolishing activity of ligand-induced TLR3 downstream signaling, which indicates that *N*-glycosylation at N413 is important in ligand recognition. Very recently, Wang et al.¹¹² published a paper in which they analyzed the role of *N*-glycan in TLR3, specifically at the N413 position via both classical and umbrella sampling MD simulation (NAMD with CHARMM36m force field) combined with NMA. They prepared six systems to assess the stability of TLR3s: TLR3 (N413 unglycosylated) with/without dsRNA, TLR3 with the paucimannosidic glycan (N413-Man₃GlcNAc₂) with/without dsRNA, and TLR3 with the oligomannosidic glycan (N413-Man₆GlcNAc₂) with/without dsRNA. The authors used the glycosylated TLR3 LRR complexed with dsRNA from the PDB (PDB ID: 3CIY). For N413, glycosylation states were built using the Glycan Reader and Modeler module. The authors found that the loop region of LRR12 in TLR3 is important for interacting with dsRNA via the formation of hydrogen bonds. The glycan at N413 stabilized dsRNA in the TLR3 binding site and altered the dynamics of the binding process, with its size, length, and branch affecting the thermodynamics and dynamics of TLR3 recognition with dsRNA. These findings provide a new perspective for modulating TLR3 function and extend our understanding of the biological role of glycans in innate immune recognition.

Regarding the TIR domain, Mahita and Sowdhamini investigated the effect of key mutations on the conformational dynamics, based on TLR2 and TLR3.¹¹³ For that, they used a combination of MD simulations (GROMOS96 54a7 force field), protein–protein interaction (PPCheck), and protein structure network analyses. They carried out the analyses for eight different complexes, including not only wild-type and mutant dimers but also wild-type and mutant trimers (TIR dimers with different adaptor proteins). To build the complexes of the receptors with the adaptor proteins, the authors performed a protein–protein docking (HADDOCK). The following computational studies highlighted the significant differences between the dimer interfaces of the wild type and mutant forms and also provided a possible explanation of how the introduced mutations may affect adaptor binding to the receptor. For the proline-to-histidine (P681H) mutation in the TIR domain of TLR2, they observed an increase in the stability of the TLR1–TLR2 heterodimer. This mutation also affected the surface of the putative adaptor-binding platform causing it to become slightly more curved. For the alanine-to-proline (A795P) mutation in the TIR domain of TLR3, they pointed out that individual subunits in a mutant tilt slightly more toward each other in comparison to the wild type. Such a subtle change may influence the orientation of the BB-loops (important for mediating interactions between dimer subunits)

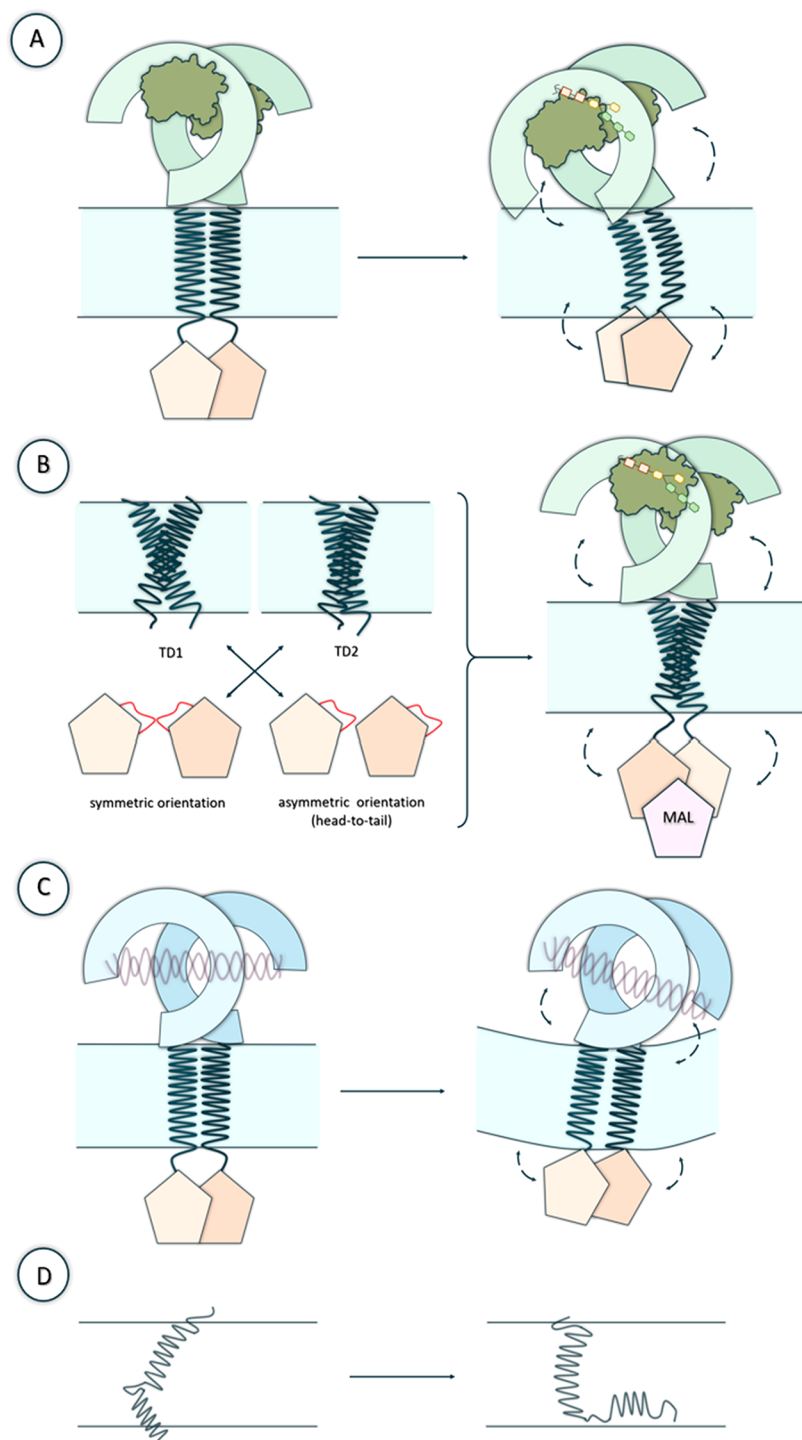


Figure 5. Examples of potential dynamical changes of TLRs observed in cited studies. (A and B) Structural transitions that the particular domains of TLR4 may exhibit (based on Patra et al.¹¹⁶ and Matamoros-Recio et al.¹¹⁷ works). TLR4 was shown with the associated myeloid differentiation factor 2 (MD2) and with the bound lipopolysaccharide LPS (C) Structural rearrangements of TLR3 domains and membrane (based on Patra et al.¹¹⁸ study). TLR3 was shown bound with dsRNA. (D) Differences in the structural organization of the transmembrane helix (TM) and cytoplasmic juxtamembrane (JM) regions that may occur in TLRs (based on Kornilov et al.³⁴ work).

on the homodimer, and thus also the binding of the adaptor proteins—MyD88 and TRIF. The authors pointed out that the obtained results were based on the assumption that TLR2 and the TLR3 TIR dimer adopt a similar conformation as that of the TLR10 TIR dimer crystal structure. As they admitted, this does not rule out the possibility of the dimers adopting a

different TIR dimer conformation during signal transduction, e.g., an asymmetrical arrangement.

Ghosh et al.¹¹⁴ showed that by applying the random alanine scanning mutation (with Robetta, using Computational Interface Alanine Scanning Server), it was possible to validate how much the residues from the BB- and DD-loops of the TIR domain contribute to TLR2 heterodimer complex formation.

For that, the binding free energy ($\Delta\Delta G_{\text{binding}}$) of the interface residues was computed. The residues with positive cutoff values >0.5 kcal/mol were accepted as the residues of importance in the dimer stability for human TLR1–2 and TLR2–6. The authors concluded that for the hTLR1–TLR2 complex, three residues—Q97, N99, Y136 of TLR1— and two residues—E55, K62 of TLR2—impact the binding energy of the complex. For the hTLR2–TLR6 complex, the following residues were predicted to have a significant role: Y44, W45 of TLR2 and E159, K160 of TLR6. While combining the results of alanine scanning mutation studies with sequence alignment, structure prediction and superimposition, molecular docking (ZDOCK), and MD simulations (GROMACS with GROMOS96 54a7 force field and SPC water model), the authors presented two key conclusions. The first was that the subtle conformational variations in the TLR structures might play a crucial role during special circumstances. The second was that the role of TLR2 BB-loop residues and TLR1/TLR6 near-DD-loop residues is important for the process of heterodimerization and for initiating differential downstream signaling.

In the summarized studies,^{109,110,113,114} the authors showed that the analysis of mutations' effect can be helpful not only in studying the TLRs' structural dynamics but also in uncovering their mechanism of action, especially in the context of ligand or adaptor protein binding. However, we still have limited knowledge regarding the particular TLRs. Given the fact that many more mutations in TLRs are reported (e.g., in the UniProt or ClinVar¹¹⁵ databases), more research should be carried out to clarify the effect of those substitutions.

Full-Length TLRs. Due to the complexity of the TLR structure and the presence of the lipid bilayer, the study of the dynamics of the full-length receptor is difficult. However, some studies have been published in recent years and they provided important insights, especially regarding the possible structure rearrangement and mechanism of action of TLRs. In Figure 5 we present the dynamical changes that particular TLRs can undergo which have been revealed and described in the recent years.

One of the first extensive studies of a full-length TLR in a membrane-aqueous environment was the work by Patra et al.¹¹⁶ The authors focused on TLR4 (TLR4/MD2/LPS homoheterodimer; TLR4 associated with MD2 protein and lipopolysaccharide LPS) and provided key insights into the orientation and interaction of LRR (named ECD in the paper), TM, and TIR domains with respect to the dipalmitoylphosphatidylcholine (DPPC) bilayer. To reach these results, they successfully applied homology modeling methods, followed by protein–protein docking and MD simulations. Additionally, they used molecular docking and binding free energy calculations to get insight into the binding of the TAK-242 ligand with the TLR4–TIR dimer. For each of the domains, the protocol had to be adapted accordingly to obtain the best possible models that could be included in a full-length structure. For instance, the dimeric LPS-bound LRR structure was obtained from the PDB (PDB ID: 3FXI), and missing residues were added (SWISS-MODEL), while the TM domain was modeled as a single α -helix and protein–protein docking (ZDOCK) was carried out to obtain a dimeric structure. The TIR domain was obtained via homology modeling using the crystal structure of TLR10 (PDB ID: 2J67) and consecutive superimposition of monomeric TLR4–TIR over the two subunits of dimeric TLR10–TIR resulted in a dimeric TLR4–TIR domain. Then, all three individual domains were aligned

on a straight axis and peptide bonds were patched between the extreme C- and N-terminal residues to adjacent domains (Discovery Studio). The constructed model could be finally inserted into the pre-equilibrated bilayer and used for further MD simulation (GROMACS with Gromos96 54a7 and Barger-lipid hybrid force field and SPC water model) and molecular docking. The authors showed that each domain of TLR4 exhibits several structural transitions (Figure 5A). The results revealed that LRR and TIR domains may be partially immersed in the membrane bilayer and that the TM domain tilts and bends to overcome the hydrophobic mismatch with the bilayer core. The authors claimed that the dynamic properties of TLR4–LRR had little effect on the interactions between LPS and MD2. For the TLR4–TM, the authors pointed out the possibility of an alternate dimerization or a potential oligomerization interface, as previously found for TLR3–TM.³⁰ Patra et al. also observed that the gradual absorption of the TLR4–TIR domain to the membrane leaflet could be a consequence of the electrostatic interactions and the bending/twisting actions of the LRR and TM domains. Their analyses indicated that even though TLR4–TIR surfaces are potentially membrane-absorbed, they also include the solvent-exposed part dedicated to interactions with other proteins. Thus, such a partial immersion is unlikely to prevent these segments from contacting the adaptor or other binding components. In the case of TLR4, the MyD88 adaptor protein is guided to TLR4–TIR by the membrane-anchored adaptor, TIR domain-containing adaptor protein (TIRAP). Hence, it is probable that the activated receptor complex TLR4/TIRAP/MyD88 is close to the membrane. For TAK-242, Patra et al. constructed two possible homodimerization interfaces—first, where helix α C and the BB loop of both TIR subunits form the dimer interface, and second, where helix α C is exposed toward the solvent and places helix α E and the BB loop in between the dimer interface. Results obtained from estimated binding free energy revealed that the first model—the α C– α C dimer—had a greater binding affinity and that the affinity of TAK-242 for the α C– α C dimer was stronger than for the α E–BB dimer. This could be an indication that the α C– α C/BB–BB model might represent the physiological dimeric interface of TLR4. However, the TAK-242 binding inside the TIR dimer cavity remains speculative, since in the case of separate simulation of full-length TLR4 as well as simulation of full-length TLR4 with TAK-242, the binding cavity of the ligand was partially blocked due to the rotation and upward movement of the TIR dimer.

In the following years, Matamoros-Recio et al.¹¹⁷ also studied the full-length model of the agonist LPS-bound TLR4. The complete model was obtained from combining the individual domains that were previously optimized by different protocols. For the dimeric LPS-bound LRR structure, the crystal structure (PDB ID: 3FXI) was selected and optimized (Maestro), while the structure of the TM domain was predicted by submitting their sequences to TMDOCK and PREDDIMER web servers. Finally, the homology modeling (implemented in YASARA) was used to predict the TIR domain dimer, using the templates of the TIR domains of human TLR1, TLR2, TLR6, and TLR10. The authors combined *ab initio* calculations with molecular docking, all-atom MD simulations, and thermodynamics calculations to provide the complete 3D models of the active TLR4 complex embedded into a membrane system. In total, they analyzed four full-length models, different dimerization interfaces for TM domain and orientations of TIR domain were observed

(Figure 5B). They showed that the interactions on different interfaces—TLR4/TLR4*, TLR4/MD-2*, and TLR4*/MD2—were kept within the simulations and that both subunits in the dimeric complex show a mutual stabilizing role. Also, they confirmed that the transmembrane domain and the following hydrophobic region (HR) indicate plasticity, depending on the membrane composition. Such plasticity may determine the dimerization of the intracellular domain. These observations are supported by a recent study by Kornilov et al.³⁴ in which the results of MD simulations (GROMACS with AMBER ff14SB and slipids force fields and TIP3P water model) indicated that juxtamembrane (JM) regions of various TLRs interact with lipids and are immersed into the bilayer membrane. The simulations showed that both TM and JM generally retain their secondary structure but adapt to the nonpolar environment by changing their tilt to the membrane and by rotating to find the optimal location of charged and nonpolar residues at the lipid–water interface (Figure 5D). In their study, Matamoros-Recio et al.¹¹⁷ proposed two models of TM-TM* (named TD-TD* in the paper) and pointed out that TM-HR can adopt different conformations, thus changing the mode of dimerization depending on the environment, regulated by TLR4 localization. The authors described also two models for the TIR-TIR* dimer (named ID-ID* in the paper)—symmetrical and asymmetrical. In the first model, the α C helix and the BB-loop in TIR domains were facing the dimerization interface, while in the second model, the dimerization interface was preserved in a head-to-tail way. The authors pointed out that both models were capable of binding the adaptor proteins. It could mean that the dimerization mechanism, and thus the receptor's activation depends on (among others) the membrane composition (localization of TLR4) and structural rearrangement. They also showed that both symmetric and asymmetric TIR-TIR* models are suitable for MyD88-adaptor-like (MAL) binding, supporting the hypothesis that both models could coexist, and have a direct implication in the activation of distinct TLR4 pathways.

In their other work, Patra et al. studied the structure and dynamics of a full-length dimer of TLR3 immersed in a bilayer of 1-palmitoyl-2-oleoyl-*sn*-glycero-3-phosphocholine (POPC).¹¹⁸ They used a similar set of molecular modeling methods as in the case of TLR4.¹¹⁶ They studied three membrane-solvated complexes of the TLR3 homodimer bound with the dsRNA. Their analyses indicated that the TLR3-TIR homodimer built from the TLR6-TIR structure led to obtaining a full-length receptor structure with the stability necessary to maintain key intermolecular interactions with the ligand and with the membrane. Furthermore, they showed that flexible juxtamembrane loops of TLR3 allow for the simultaneous bending of the LRR and TIR domains on both surfaces of the membrane. They also observed that the complex immersed in the bilayer progressively tilted on the bilayer surface due to the electrostatic attraction between the charged parts of both the protein and phospholipids from the bilayer (Figure 5C). In that case, the LRR-NT was only partially absorbed by the lipid headgroups. That was in contrast to the LRR-NT from their previously reported TLR4 model that was completely buried in the bilayer surface. They assumed that it is possible that the negatively charged dsRNA restricted the insertion of LRR-NT into the membrane surface. During the simulations, the dsRNA kept its structural integrity while bound to TLR3. The observed distortions in the TLR3-

TM domain were distinct from the previously reported TLR4-TM. Thus, the authors concluded that the orientation and conformational changes of each TLR type may vary, depending on their location in the cell or the lipid composition in the membrane. Based on the MD simulations analysis, Patra et al. indicated the probable interface involving residues from the α C and α D helix and the CD and DE loops of both TIR monomers. The BB-loop of one subunit was completely solvent-exposed, while the other was partially involved in dimer packing. The solvent-exposed part confirmed the importance of this segment in TRIF recruitment by the activated receptor.

The reviewed papers revealed important insight into TLRs dynamics. In summarized studies, the authors presented relevant information on possible changes in position and conformation that receptors embedded in the cell membrane or intracellular compartments may undergo. Also, an important message regarding the potential mechanism of TIR domain dimerization and binding of the adaptor protein came from the analyzed models of both symmetrical and asymmetrical domains. This may be helpful for designing new types of TLR modulators, especially those targeting the TIR domain. One should remember that the presented studies on full-length receptors refer only to TLR3 and TLR4, which means that for now, the conclusions cannot be unified for all other receptors. As in the case of studying the effect of mutations, it seems that research regarding the dynamics of TLRs is just beginning. Considering the differences in TLRs structure, substrate recognition, dimerization requirements, and association with adaptor proteins, along with the importance of understanding the TLRs signal transduction pathway, we can expect a significant increase in interest in this field in the coming years.

CONCLUSIONS

Toll-like receptors are one of the most crucial components of the immune system. Given their importance, it was not a surprise that the 2011 Nobel Prize in Physiology or Medicine was awarded to Dr. Jules A. Hoffmann and Dr. Bruce A. Beutler for their discoveries of the role of TLRs in innate immunity. It happened relatively quickly after the discovery of TLRs, only within 15 years. Since that time, tens of thousands of papers have been published in which TLRs have been the main subject of research. TLRs are complicated in terms of their structure, dynamics, and functioning, and this complexity is a challenge despite the enormous progress in the development of both experimental and computational methods. In our review, we aimed to highlight the progress made in recent years with the use of *in silico* methods for TLRs studies. Also, we wanted to point out the areas that still await their discoverers. One of the main limitations in understanding the function of TLRs is difficulty in the proper characterization of receptor structure at various stages of signal transduction. Even the latest breakthrough in AI-based structure prediction is not yet widely used in research aimed at revealing the mechanism of action of TLRs.

Based on the results presented in the reviewed papers, we can conclude that still, the most attention is paid to the use of computational solutions for the design of small-molecule modulators. The use of *in silico* methods to design other types of modulators, such as multiepitope vaccines, is gaining more popularity, but yet, it is not as common as in the case of small-molecule compounds. Both small-molecule and multiepitope modulators are designed in such a way as to target the LRR of

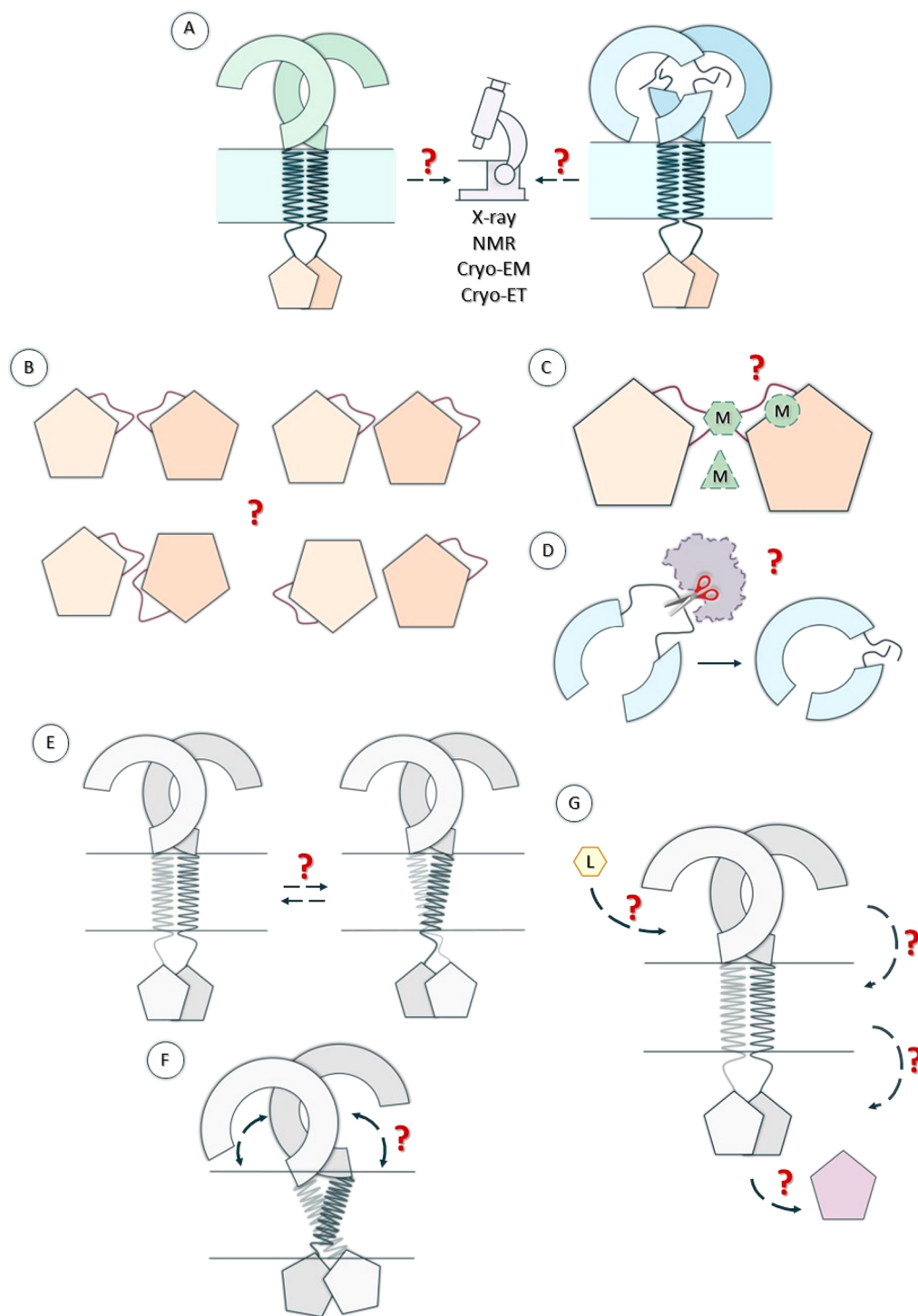


Figure 6. Areas in TLR research that still require further development. (A) Experimental verification of the predicted structures. (B) Studying the orientation of the subunits of the TIR domain dimers of TLRs. (C) Designing small-molecule modulators (M) targeting the TIR domain of TLRs. (D) Studying the proteolytic cleavage of the Z-loop in TLR7–9. (E) Analyzing potential changes in the subunits dynamics in TLRs. (F) Analyzing the conformational changes and structural rearrangements in both TLR receptors and bilayer membrane. (G) Studying the whole process of ligand recognition through the signaling cascade to the immune response.

TLRs. There was no breakthrough in the design of small-molecule modulators targeting the TIR domain. Among things that scientists will want to keep improving is obtaining the best binding affinity and stability of the modulators. Regarding the dynamics of TLRs, scientists have shown that studying the mutations' effect can contribute to a better understanding of the potential mechanism of action of the receptors. That is of

special interest for both ligand and adaptor protein binding. More demanding, both in terms of system preparation and computing power, is the analysis of the dynamics of the full-length TLR complex. So far, only TLR3 and 4 have been built as full-length models embedded in the lipid bilayer. Those studies presented relevant information on possible conformational changes that may occur in the receptor's structure. Thus,

it would be very important to perform similar studies for members of the TLR family. Since now we have easier access to the predictions of large macromolecule structures, we expect that in the coming years, we will witness progress in research on the TLRs' dynamics and mechanism of action.

In Figure 6 we presented the main areas in TLR research that still require further studies. Figure 6A illustrates the necessity of the experimental verification of the predicted structures. Despite the great progress in AI-based methods to predict the tertiary structures of macromolecules, experimental validation is a must to confirm the compliance of the obtained predictions. Access to experimentally solved structures of transmembrane proteins is also important in order to confirm the orientation of individual domains or subunits of the structure toward each other. Obtaining information about the orientation of the subunits of the TIR domain dimers of TLRs is of special interest (Figure 6B). So far, we have information about possible symmetrical or asymmetric orientations. However, we lack a systematic review of what orientations are preferred by specific receptors and how the orientation of the subunits can determine the binding of the adaptor proteins and the initiation of the signal cascade. This issue is also related to the design of small-molecule modulators targeting the TIR domain (Figure 6C). Without details about the orientation of the subunits, it is difficult to properly select the best binding site for modulators.

As we mentioned in the Introduction of this review, some TLRs (7–9) require the proteolytic cleavage of the Z-loop in their LRR domain (Figure 6D). This is needed to allow ligands to bind and to further activate the receptor. Very little is known about the molecular basis of this process. Basically, only the information about the examples of proteases potentially involved in cleavage is available. To our best knowledge, there are no *in silico* studies attempting to explain this process. We are aware that one of the obstacles may be the size of the system and that no accurate structure predictions of the TLR-protease complex have been available so far. However, we hope that with the increase of the computational resources and the possibility to predict the structure of complexes using, e.g., AlphaFold Multimer, this issue will be soon addressed.

In Figure 6E,F, we wanted to highlight the importance of conducting further research on the dynamics and conformational changes of TLRs. As we mentioned, studies presented to date have mainly focused on TLR3 and TLR4. Very little is known about other receptors, e.g., how the conformational changes occur in individual subunits or how full-length receptors behave in relation to the membrane in which they are immersed. In particular, we would like to know whether the location of the receptor (cell membrane or intracellular compartments) determines the TLRs' dynamics and the subsequent ability to bind the adaptor proteins. Figure 6G illustrates the ultimate goal of studying the Toll-like receptors with the use of computational methods, which is to get deep insight into each stage of the receptor functioning. Thus, the challenge is to combine all the information, starting from the recognition of the ligand by the receptor, through the triggering of the signaling cascade, to the immune response.

■ ASSOCIATED CONTENT

Data Availability Statement

Information about human Toll-like receptors domains deposited in the Protein Data Bank and information about chemical structures of the best hits together (small-molecule

agonists and antagonists) with the identified chemical interactions from the reviewed research papers are provided in the Supporting Information.

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.3c00419>.

Supplementary Table S1. Overview of human Toll-like receptors domains deposited in the Protein Data Bank. Supplementary Table S2. Chemical structures of the best hits (small-molecule agonists and antagonists) from the reviewed research papers. (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Maria Bzówka – Tunneling Group, Biotechnology Centre and Department of Organic Chemistry, Bioorganic Chemistry and Biotechnology, Faculty of Chemistry, Silesian University of Technology, 44-100 Gliwice, Poland; orcid.org/0000-0001-6802-8753; Email: maria.bzowka@polsl.pl, m.bzowka@tunnelinggroup.pl

Authors

Weronika Bagrowska – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland

Artur Góra – Tunneling Group, Biotechnology Centre, Silesian University of Technology, 44-100 Gliwice, Poland; orcid.org/0000-0003-2530-6957

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jcim.3c00419>

Funding

The work was 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” program [Project Number: DI2018 014148; Agreement Number: 0141/DIA/2019/48].

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The work was 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: DI2018 014148; Agreement Number: 0141/DIA/2019/48].

■ ABBREVIATIONS

TLRs, Toll-like receptors; PRRs, pattern recognition receptors; MPs, molecular patterns; DAMPs, damage/danger-associated molecular patterns; MAMPs, microbial/microbe-associated molecular patterns; PAMPs, pathogen-associated molecular patterns; XAMPs, xenobiotic-associated molecular patterns; LRR, leucine-rich repeats domain; TM, transmembrane domain; TIR, Toll-interleukin-1 receptor domain; MyD88, myeloid differentiation primary-response protein 88; TRIF, TIR domain-containing adaptor protein inducing interferon- β ; hTLRs, human Toll-like receptors; PDB, Protein Data Bank; JM, juxtamembrane; VS, virtual screening; MD, molecular dynamics; MM-PBSA, Molecular Mechanics Poisson–Boltzmann Surface Area; MM-GBSA, Molecular Mechanics with Generalized Born and Surface Area; NF- κ B, nuclear factor

kappa-light-chain-enhancer of activated B cells; SAR, structure–activity relationship; MD2, myeloid differentiation factor 2; ADMET, absorption, distribution, metabolism, excretion, toxicity properties; IL, interleukine; METH, methamphetamine; NMA, Normal Mode Analysis; SARS-CoV-2, severe acute respiratory syndrome coronavirus 2; MERS, Middle East respiratory syndrome; HCV, Hepatitis C virus; HIV, human immunodeficiency virus; NeoCoV, Neo-Coronavirus; MPL, monophosphoryl lipid A; S protein, SARS-CoV-2 spike glycoprotein; N protein, nucleocapsid protein; ORF1a, open reading frame 1a protein; MTB, *Mycobacterium tuberculosis*; OmpATb, outer membrane protein A Rv0899; HTL, human thymus lymphoid; IgM, immunoglobulin M; IgG, immunoglobulin G; PCA, principal component analysis; RIN, residue interaction network; SNPs, single nucleotide polymorphisms; WT, wild type; LPS, lipopolysaccharide; DPPC, dipalmitoyl-phosphatidylcholine bilayer; POPC, 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine; TIRAP, TIR domain-containing adaptor protein

REFERENCES

- (1) Kawai, T.; Akira, S. Toll-like Receptors and Their Crosstalk with Other Innate Receptors in Infection and Immunity. *Immunity* **2011**, *34*, 637–650.
- (2) Vijay, K. Toll-like Receptors in Immunity and Inflammatory Diseases: Past, Present, and Future. *Int. Immunopharmacol.* **2018**, *59*, 391–412.
- (3) Thompson, M. R.; Kaminski, J. J.; Kurt-Jones, E. A.; Fitzgerald, K. A. Pattern Recognition Receptors and the Innate Immune Response to Viral Infection. *Viruses* **2011**, *3*, 920–940.
- (4) Amarante-Mendes, G. P.; Adjemian, S.; Branco, L. M.; Zanetti, L. C.; Weinlich, R.; Bortoluci, K. R. Pattern Recognition Receptors and the Host Cell Death Molecular Machinery. *Front. Immunol.* **2018**, *9*. DOI: 10.3389/fimmu.2018.02379.
- (5) Li, D.; Wu, M. Pattern Recognition Receptors in Health and Diseases. *Signal Transduct. Target. Ther.* **2021**, *6*, 291.
- (6) Takeuchi, O.; Akira, S. Pattern Recognition Receptors and Inflammation. *Cell* **2010**, *140*, 805–820.
- (7) Behzadi, P.; García-Perdomo, H. A.; Karpinski, T. M. Toll-Like Receptors: General Molecular and Structural Biology. *J. Immunol. Res.* **2021**, *2021*, 1–21.
- (8) Kawai, T.; Akira, S. The Role of Pattern-Recognition Receptors in Innate Immunity: Update on Toll-like Receptors. *Nat. Immunol.* **2010**, *11*, 373–384.
- (9) Bell, J. K.; Mullen, G. E. D.; Leifer, C. A.; Mazzoni, A.; Davies, D. R.; Segal, D. M. Leucine-Rich Repeats and Pathogen Recognition in Toll-like Receptors. *Trends Immunol.* **2003**, *24*, 528–533.
- (10) Matsushima, N.; Tanaka, T.; Enkhbayar, P.; Mikami, T.; Taga, M.; Yamada, K.; Kuroki, Y. Comparative Sequence Analysis of Leucine-Rich Repeats (LRRs) within Vertebrate Toll-like Receptors. *BMC Genomics* **2007**, *8*, 124.
- (11) Asami, J.; Shimizu, T. Structural and Functional Understanding of the Toll-like Receptors. *Protein Sci.* **2021**, *30*, 761–772.
- (12) West, A. P.; Koblansky, A. A.; Ghosh, S. Recognition and Signaling by Toll-Like Receptors. *Annu. Rev. Cell Dev. Biol.* **2006**, *22*, 409–437.
- (13) Yu, L.; Wang, L.; Chen, S. Endogenous Toll-like Receptor Ligands and Their Biological Significance. *J. Cell. Mol. Med.* **2010**, *14*, 2592–2603.
- (14) Gao, D.; Li, W. Structures and Recognition Modes of Toll-like Receptors. *Proteins Struct. Funct. Bioinforma.* **2017**, *85*, 3–9.
- (15) Manavalan, B.; Basith, S.; Choi, S. Similar Structures but Different Roles—an Updated Perspective on TLR Structures. *Front. Physiol.* **2011**, *2*, 1–13.
- (16) Kawasaki, T.; Kawai, T. Toll-Like Receptor Signaling Pathways. *Front. Immunol.* **2014**, *5*. DOI: 10.3389/fimmu.2014.00461.
- (17) O'Neill, L. A. J.; Bowie, A. G. The Family of Five: TIR-Domain-Containing Adaptors in Toll-like Receptor Signalling. *Nat. Rev. Immunol.* **2007**, *7*, 353–364.
- (18) Troutman, T. D.; Bazan, J. F.; Pasare, C. Toll-like Receptors, Signaling Adaptors and Regulation of the pro-Inflammatory Response by PI3K. *Cell Cycle* **2012**, *11*, 3559–3567.
- (19) El-Zayat, S. R.; Sibaii, H.; Mannaa, F. A. Toll-like Receptors Activation, Signaling, and Targeting: An Overview. *Bull. Natl. Res. Cent.* **2019**, *43*, 187.
- (20) Hennessy, E. J.; Parker, A. E.; O'Neill, L. A. J. Targeting Toll-like Receptors: Emerging Therapeutics? *Nat. Rev. Drug Discovery* **2010**, *9*, 293–307.
- (21) Anwar, M. A.; Shah, M.; Kim, J.; Choi, S. Recent Clinical Trends in Toll-like Receptor Targeting Therapeutics. *Med. Res. Rev.* **2019**, *39*, 1053–1090.
- (22) Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G. R.; Wang, J.; Cong, Q.; Kinch, L. N.; Schaeffer, R. D.; Millán, C.; Park, H.; Adams, C.; Glassman, C. R.; DeGiovanni, A.; Pereira, J. H.; Rodrigues, A. V.; van Dijk, A. A.; Ebrecht, A. C.; Opperman, D. J.; Sagmeister, T.; Buhlheller, C.; Pavkov-Keller, T.; Rathinaswamy, M. K.; Dalwadi, U.; Yip, C. K.; Burke, J. E.; Garcia, K. C.; Grishin, N. V.; Adams, P. D.; Read, R. J.; Baker, D. Accurate Prediction of Protein Structures and Interactions Using a Three-Track Neural Network. *Science (80-.)* **2021**, *373*, 871–876.
- (23) Sartorius, R.; Trovato, M.; Manco, R.; D'Apice, L.; De Berardinis, P. Exploiting Viral Sensing Mediated by Toll-like Receptors to Design Innovative Vaccines. *npj Vaccines* **2021**, *6*, 127.
- (24) Murgueitio, M. S.; Rakers, C.; Frank, A.; Wolber, G. Balancing Inflammation: Computational Design of Small-Molecule Toll-like Receptor Modulators. *Trends Pharmacol. Sci.* **2017**, *38*, 155–168.
- (25) Pérez-Regidor, L.; Zarioh, M.; Ortega, L.; Martín-Santamaría, S. Virtual Screening Approaches towards the Discovery of Toll-Like Receptor Modulators. *IJMS* **2016**, *17*, 1508.
- (26) Billod, J. M.; Lacetera, A.; Guzmán-Caldentey, J.; Martín-Santamaría, S. Computational Approaches to Toll-Like Receptor 4 Modulation. *Molecules* **2016**, *21*, 994.
- (27) Xu, Y.; Tao, X.; Shen, B.; Horng, T.; Medzhitov, R.; Manley, J. L.; Tong, L. Structural Basis for Signal Transduction by the Toll/Interleukin-1 Receptor Domains. *Nature* **2000**, *408*, 111–115.
- (28) Choe, J.; Kelker, M. S.; Wilson, I. A. Crystal Structure of Human Toll-Like Receptor 3 (TLR3) Ectodomain. *Science (80-.)* **2005**, *309*, 581–585.
- (29) Bell, J. K.; Botos, I.; Hall, P. R.; Askins, J.; Shiloach, J.; Segal, D. M.; Davies, D. R. The Molecular Structure of the Toll-like Receptor 3 Ligand-Binding Domain. *Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 10976–10980.
- (30) Mineev, K. S.; Goncharuk, S. A.; Arseniev, A. S. Toll-like Receptor 3 Transmembrane Domain Is Able to Perform Various Homotypic Interactions: An NMR Structural Study. *FEBS Lett.* **2014**, *588*, 3802–3807.
- (31) Berman, H. M. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (32) Ishida, H.; Asami, J.; Zhang, Z.; Nishizawa, T.; Shigematsu, H.; Ohto, U.; Shimizu, T. Cryo-EM Structures of Toll-like Receptors in Complex with UNC93B1. *Nat. Struct. Mol. Biol.* **2021**, *28*, 173–180.
- (33) Lushpa, V. A.; Goncharuk, M. V.; Lin, C.; Zalevsky, A. O.; Talyzina, I. A.; Luginina, A. P.; Vakhrameev, D. D.; Shevtsov, M. B.; Goncharuk, S. A.; Arseniev, A. S.; Borshchevskiy, V. I.; Wang, X.; Mineev, K. S. Modulation of Toll-like Receptor 1 Intracellular Domain Structure and Activity by Zn²⁺ Ions. *Commun. Biol.* **2021**, *4*, 1003.
- (34) Kornilov, F. D.; Shabalkina, A. V.; Lin, C.; Volynsky, P. E.; Kot, E. F.; Kayushin, A. L.; Lushpa, V. A.; Goncharuk, M. V.; Arseniev, A. S.; Goncharuk, S. A.; Wang, X.; Mineev, K. S. The Architecture of Transmembrane and Cytoplasmic Juxtamembrane Regions of Toll-like Receptors. *Nat. Commun.* **2023**, *14*, 1503.
- (35) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Židek, A.;

- Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.; Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
- (36) Varadi, M.; Anyango, S.; Deshpande, M.; Nair, S.; Natassia, C.; Yordanova, G.; Yuan, D.; Stroe, O.; Wood, G.; Laydon, A.; Židek, A.; Green, T.; Tunyasuvunakool, K.; Petersen, S.; Jumper, J.; Clancy, E.; Green, R.; Vora, A.; Lutfi, M.; Figurnov, M.; Cowie, A.; Hobbs, N.; Kohli, P.; Kleywegt, G.; Birney, E.; Hassabis, D.; Velankar, S. AlphaFold Protein Structure Database: Massively Expanding the Structural Coverage of Protein-Sequence Space with High-Accuracy Models. *Nucleic Acids Res.* **2022**, *50*, D439–D444.
- (37) Wang, Y.; Zhang, S.; Li, H.; Wang, H.; Zhang, T.; Hutchinson, M. R.; Yin, H.; Wang, X. Small-Molecule Modulators of Toll-like Receptors. *Acc. Chem. Res.* **2020**, *53*, 1046–1055.
- (38) Murgueitio, M. S.; Ebner, S.; Hörtnagl, P.; Rakers, C.; Bruckner, R.; Henneke, P.; Wolber, G.; Santos-Sierra, S. Enhanced Immunostimulatory Activity of in Silico Discovered Agonists of Toll-like Receptor 2 (TLR2). *Biochim. Biophys. Acta - Gen. Subj.* **2017**, *1861*, 2680–2689.
- (39) Guan, Y.; Omueti-Ayoade, K.; Mutha, S. K.; Hergenrother, P. J.; Tapping, R. I. Identification of Novel Synthetic Toll-like Receptor 2 Agonists by High Throughput Screening. *J. Biol. Chem.* **2010**, *285*, 23755–23762.
- (40) Liang, Q.; Wu, Q.; Jiang, J.; Duan, J.; Wang, C.; Smith, M. D.; Lu, H.; Wang, Q.; Nagarkatti, P.; Fan, D. Characterization of Sparstolonin B, a Chinese Herb-Derived Compound, as a Selective Toll-like Receptor Antagonist with Potent Anti-Inflammatory Properties. *J. Biol. Chem.* **2011**, *286*, 26470–26479.
- (41) Durai, P.; Shin, H.-J.; Achek, A.; Kwon, H.-K.; Govindaraj, R. G.; Panneerselvam, S.; Yesudhas, D.; Choi, J.; No, K. T.; Choi, S. Toll-like Receptor 2 Antagonists Identified through Virtual Screening and Experimental Validation. *FEBS J.* **2017**, *284*, 2264–2283.
- (42) Jin, M. S.; Kim, S. E.; Heo, J. Y.; Lee, M. E.; Kim, H. M.; Paik, S.-G.; Lee, H.; Lee, J.-O. Crystal Structure of the TLR1-TLR2 Heterodimer Induced by Binding of a Tri-Acylated Lipopeptide. *Cell* **2007**, *130*, 1071–1082.
- (43) Chen, Z.; Cen, X.; Yang, J.; Tang, X.; Cui, K.; Cheng, K. Structure-Based Discovery of a Specific TLR1–TLR2 Small Molecule Agonist from the ZINC Drug Library Database. *Chem. Commun.* **2018**, *54*, 11411–11414.
- (44) Chen, Z.; Cen, X.; Yang, J.; Lin, Z.; Liu, M.; Cheng, K. Synthesis of Urea Analogues Bearing N-Alkyl-N'-(Thiophen-2-Yl) Scaffold and Evaluation of Their Innate Immune Response to Toll-like Receptors. *Eur. J. Med. Chem.* **2019**, *169*, 42–52.
- (45) Grabowski, M.; Murgueitio, M. S.; Bermudez, M.; Rademann, J.; Wolber, G.; Weindl, G. Identification of a Pyrogallol Derivative as a Potent and Selective Human TLR2 Antagonist by Structure-Based Virtual Screening. *Biochem. Pharmacol.* **2018**, *154*, 148–160.
- (46) Murgueitio, M. S.; Henneke, P.; Glossmann, H.; Santos-Sierra, S.; Wolber, G. Prospective Virtual Screening in a Sparse Data Scenario: Design of Small-Molecule TLR2 Antagonists. *ChemMedChem.* **2014**, *9*, 813–822.
- (47) Cheng, K.; Wang, X.; Zhang, S.; Yin, H. Discovery of Small-Molecule Inhibitors of the TLR1/TLR2 Complex. *Angew. Chemie Int. Ed.* **2012**, *51*, 12246–12249.
- (48) Grabowski, M.; Bermudez, M.; Rudolf, T.; Šribar, D.; Varga, P.; Murgueitio, M. S.; Wolber, G.; Rademann, J.; Weindl, G. Identification and Validation of a Novel Dual Small-Molecule TLR2/8 Antagonist. *Biochem. Pharmacol.* **2020**, *177*, 113957.
- (49) Bermudez, M.; Grabowski, M.; Murgueitio, M. S.; Tiemann, M.; Varga, P.; Rudolf, T.; Wolber, G.; Weindl, G.; Rademann, J. Biological Characterization, Mechanistic Investigation and Structure-Activity Relationships of Chemically Stable TLR2 Antagonists. *ChemMedChem.* **2020**, *15*, 1364–1371.
- (50) Morin, M. D.; Wang, Y.; Jones, B. T.; Mifune, Y.; Su, L.; Shi, H.; Moresco, E. M. Y.; Zhang, H.; Beutler, B.; Boger, D. L. Diprovocims: A New and Exceptionally Potent Class of Toll-like Receptor Agonists. *J. Am. Chem. Soc.* **2018**, *140*, 14440–14454.
- (51) Su, L.; Wang, Y.; Wang, J.; Mifune, Y.; Morin, M. D.; Jones, B. T.; Moresco, E. M. Y.; Boger, D. L.; Beutler, B.; Zhang, H. Structural Basis of TLR2/TLR1 Activation by the Synthetic Agonist Diprovocim. *J. Med. Chem.* **2019**, *62*, 2938–2949.
- (52) Mishra, V.; Pathak, C. Structural Insights into Pharmacophore-Assisted in Silico Identification of Protein–Protein Interaction Inhibitors for Inhibition of Human Toll-like Receptor 4 – Myeloid Differentiation Factor-2 (HTLR4–MD-2) Complex. *J. Biomol. Struct. Dyn.* **2019**, *37*, 1968–1991.
- (53) Facchini, F. A.; Zaffaroni, L.; Minotti, A.; Rapisarda, S.; Calabrese, V.; Forcella, M.; Fusi, P.; Airoidi, C.; Ciaramelli, C.; Billod, J.-M.; Schromm, A. B.; Braun, H.; Palmer, C.; Beyaert, R.; Lapenta, F.; Jerala, R.; Pirianov, G.; Martin-Santamaria, S.; Peri, F. Structure–Activity Relationship in Monosaccharide-Based Toll-Like Receptor 4 (TLR4) Antagonists. *J. Med. Chem.* **2018**, *61*, 2895–2909.
- (54) Cochet, F.; Facchini, F. A.; Zaffaroni, L.; Billod, J.-M.; Coelho, H.; Holgado, A.; Braun, H.; Beyaert, R.; Jerala, R.; Jimenez-Barbero, J.; Martin-Santamaria, S.; Peri, F. Novel Carboxylate-Based Glycolipids: TLR4 Antagonism, MD-2 Binding and Self-Assembly Properties. *Sci. Rep.* **2019**, *9*, 919.
- (55) Wang, X.; Northcutt, A. L.; Cochran, T. A.; Zhang, X.; Fabisiak, T. J.; Haas, M. E.; Amat, J.; Li, H.; Rice, K. C.; Maier, S. F.; Bachtell, R. K.; Hutchinson, M. R.; Watkins, L. R. Methamphetamine Activates Toll-Like Receptor 4 to Induce Central Immune Signaling within the Ventral Tegmental Area and Contributes to Extracellular Dopamine Increase in the Nucleus Accumbens Shell. *ACS Chem. Neurosci.* **2019**, *10*, 3622–3634.
- (56) Northcutt, A. L.; Hutchinson, M. R.; Wang, X.; Baratta, M. V.; Hiranita, T.; Cochran, T. A.; Pomrenze, M. B.; Galer, E. L.; Kopajtic, T. A.; Li, C. M.; Amat, J.; Larson, G.; Cooper, D. C.; Huang, Y.; O'Neill, C. E.; Yin, H.; Zahniser, N. R.; Katz, J. L.; Rice, K. C.; Maier, S. F.; Bachtell, R. K.; Watkins, L. R. DAT Isn't All That: Cocaine Reward and Reinforcement Require Toll-like Receptor 4 Signaling. *Mol. Psychiatry* **2015**, *20*, 1525–1537.
- (57) Hutchinson, M. R.; Zhang, Y.; Shridhar, M.; Evans, J. H.; Buchanan, M. M.; Zhao, T. X.; Slivka, P. F.; Coats, B. D.; Rezvani, N.; Wieseler, J.; Hughes, T. S.; Landgraf, K. E.; Chan, S.; Fong, S.; Phipps, S.; Falke, J. J.; Leinwand, L. A.; Maier, S. F.; Yin, H.; Rice, K. C.; Watkins, L. R. Evidence That Opioids May Have Toll-like Receptor 4 and MD-2 Effects. *Brain. Behav. Immun.* **2010**, *24*, 83–95.
- (58) Wang, X.; Zhang, Y.; Peng, Y.; Hutchinson, M. R.; Rice, K. C.; Yin, H.; Watkins, L. R. Pharmacological Characterization of the Opioid Inactive Isomers (+)-Naltrexone and (+)-Naloxone as Antagonists of Toll-like Receptor 4. *Br. J. Pharmacol.* **2016**, *173*, 856–869.
- (59) Selfridge, B. R.; Wang, X.; Zhang, Y.; Yin, H.; Grace, P. M.; Watkins, L. R.; Jacobson, A. E.; Rice, K. C. Structure–Activity Relationships of (+)-Naltrexone-Inspired Toll-like Receptor 4 (TLR4) Antagonists. *J. Med. Chem.* **2015**, *58*, 5038–5052.
- (60) Zhang, X.; Cui, F.; Chen, H.; Zhang, T.; Yang, K.; Wang, Y.; Jiang, Z.; Rice, K. C.; Watkins, L. R.; Hutchinson, M. R.; Li, Y.; Peng, Y.; Wang, X. Dissecting the Innate Immune Recognition of Opioid Inactive Isomer (+)-Naltrexone Derived Toll-like Receptor 4 (TLR4) Antagonists. *J. Chem. Inf. Model.* **2018**, *58*, 816–825.
- (61) Zhang, X.; Peng, Y.; Grace, P. M.; Metcalf, M. D.; Kwilas, A. J.; Wang, Y.; Zhang, T.; Wu, S.; Selfridge, B. R.; Portoghesi, P. S.; Rice, K. C.; Watkins, L. R.; Hutchinson, M. R.; Wang, X. Stereochemistry and Innate Immune Recognition: (+)-norbinaltorphimine Targets Myeloid Differentiation Protein 2 and Inhibits Toll-like Receptor 4 Signaling. *FASEB J.* **2019**, *33*, 9577–9587.
- (62) Pérez-Regidor, L.; Guzmán-Caldentey, J.; Oberhauser, N.; Punzón, C.; Balogh, B.; Pedro, J. R.; Falomir, E.; Nurisso, A.; Mátyus, P.; Menéndez, J. C.; de Andrés, B.; Fresno, M.; Martín-Santamaría, S. Small Molecules as Toll-like Receptor 4 Modulators Drug and In-House Computational Repurposing. *Biomedicines* **2022**, *10*, 2326.

- (63) Gao, M.; London, N.; Cheng, K.; Tamura, R.; Jin, J.; Schueler-Furman, O.; Yin, H. Rationally Designed Macrocyclic Peptides as Synergistic Agonists of LPS-Induced Inflammatory Response. *Tetrahedron* **2014**, *70*, 7664–7668.
- (64) Borges, P. V.; Moret, K. H.; Raghavendra, N. M.; Maramaldo Costa, T. E.; Monteiro, A. P.; Carneiro, A. B.; Pacheco, P.; Temerozo, J. R.; Bou-Habib, D. C.; das Graças Henriques, M.; Penido, C. Protective Effect of Gedunin on TLR-Mediated Inflammation by Modulation of Inflammasome Activation and Cytokine Production: Evidence of a Multitarget Compound. *Pharmacol. Res.* **2017**, *115*, 65–77.
- (65) Talukdar, A.; Ganguly, D.; Roy, S.; Das, N.; Sarkar, D. Structural Evolution and Translational Potential for Agonists and Antagonists of Endosomal Toll-like Receptors. *J. Med. Chem.* **2021**, *64*, 8010–8041.
- (66) Gupta, C. L.; Babu Khan, M.; Ampasala, D. R.; Akhtar, S.; Dwivedi, U. N.; Bajpai, P. Pharmacophore-Based Virtual Screening Approach for Identification of Potent Natural Modulatory Compounds of Human Toll-like Receptor 7. *J. Biomol. Struct. Dyn.* **2019**, *37*, 4721–4736.
- (67) Šribar, D.; Grabowski, M.; Murgueitio, M. S.; Bermudez, M.; Weindl, G.; Wolber, G. Identification and Characterization of a Novel Chemotype for Human TLR8 Inhibitors. *Eur. J. Med. Chem.* **2019**, *179*, 744–752.
- (68) Wang, X.; Chen, Y.; Zhang, S.; Deng, J. N. Molecular Dynamics Simulations Reveal the Selectivity Mechanism of Structurally Similar Agonists to TLR7 and TLR8. *PLoS One* **2022**, *17*, e0260565.
- (69) Luchner, M.; Reinke, S.; Milicic, A. TLR Agonists as Vaccine Adjuvants Targeting Cancer and Infectious Diseases. *Pharmaceutics* **2021**, *13*, 142.
- (70) Pulendran, B.; Arunachalam, P. S.; O'Hagan, D. T. Emerging Concepts in the Science of Vaccine Adjuvants. *Nat. Rev. Drug Discovery* **2021**, *20*, 454–475.
- (71) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Agivretova, R.; Ahmad, S.; Alpi, E.; Bowler-Barnett, E. H.; Britto, R.; Bursteinas, B.; Bye-A-Jee, H.; Coetzee, R.; Cukura, A.; Da Silva, A.; Denny, P.; Dogan, T.; Ebenezer, T.; Fan, J.; Castro, L. G.; Garmiri, P.; Georghiou, G.; Gonzales, L.; Hatton-Ellis, E.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Jokinen, P.; Joshi, V.; Jyothi, D.; Lock, A.; Lopez, R.; Luciani, A.; Luo, J.; Lussi, Y.; MacDougall, A.; Madeira, F.; Mahmoudy, M.; Menchi, M.; Mishra, A.; Moulang, K.; Nightingale, A.; Oliveira, C. S.; Pundir, S.; Qi, G.; Raj, S.; Rice, D.; Lopez, M. R.; Saidi, R.; Sampson, J.; Sawford, T.; Speretta, E.; Turner, E.; Tyagi, N.; Vasudev, P.; Volynkin, V.; Warner, K.; Watkins, X.; Zaru, R.; Zellner, H.; Bridge, A.; Poux, S.; Redaschi, N.; Aimo, L.; Argoud-Puy, G.; Auchincloss, A.; Axelsen, K.; Bansal, P.; Baratin, D.; Blatter, M.-C.; Bolleman, J.; Boutet, E.; Breuza, L.; Casals-Casas, C.; de Castro, E.; Echioukh, K. C.; Coudert, E.; Cuhe, B.; Doche, M.; Dornevil, D.; Estreicher, A.; Famiglietti, M. L.; Feuermann, M.; Gasteiger, E.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz-Gumowski, N.; Hinz, U.; Hulo, C.; Hyka-Nouspikel, N.; Jungo, F.; Keller, G.; Kerhornou, A.; Lara, V.; Le Mercier, P.; Lieberherr, D.; Lombardot, T.; Martin, X.; Masson, P.; Morgat, A.; Neto, T. B.; Paesano, S.; Pedruzzi, I.; Pilbout, S.; Pourcel, L.; Pozzato, M.; Pruess, M.; Rivoire, C.; Sigrist, C.; Sonesson, K.; Stutz, A.; Sundaram, S.; Tognolli, M.; Verbregue, L.; Wu, C. H.; Arighi, C. N.; Arminski, L.; Chen, C.; Chen, Y.; Garavelli, J. S.; Huang, H.; Laiho, K.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Q.; Wang, Y.; Yeh, L.-S.; Zhang, J.; Ruch, P.; Teodoro, D. UniProt: The Universal Protein Knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.
- (72) Doytchinova, I. A.; Flower, D. R. VaxiJen: A Server for Prediction of Protective Antigens, Tumour Antigens and Subunit Vaccines. *BMC Bioinformatics* **2007**, *8*, 4.
- (73) Gasteiger, E.; Hoogland, C.; Gattiker, A.; Duvaud, S.; Wilkins, M. R.; Appel, R. D.; Bairoch, A. Protein Identification and Analysis Tools on the ExPASy Server. *The Proteomics Protocols Handbook*; Humana Press: Totowa, NJ, 2005; pp 571–607.
- (74) Larsen, M. V.; Lundegaard, C.; Lamberth, K.; Buus, S.; Lund, O.; Nielsen, M. Large-Scale Validation of Methods for Cytotoxic T-Lymphocyte Epitope Prediction. *BMC Bioinformatics* **2007**, *8*, 424.
- (75) Jensen, K. K.; Andreatta, M.; Marcatili, P.; Buus, S.; Greenbaum, J. A.; Yan, Z.; Sette, A.; Peters, B.; Nielsen, M. Improved Methods for Predicting Peptide Binding Affinity to MHC Class II Molecules. *Immunology* **2018**, *154*, 394–406.
- (76) Vita, R.; Mahajan, S.; Overton, J. A.; Dhanda, S. K.; Martini, S.; Cantrell, J. R.; Wheeler, D. K.; Sette, A.; Peters, B. The Immune Epitope Database (IEDB): 2018 Update. *Nucleic Acids Res.* **2019**, *47*, D339–D343.
- (77) Jespersen, M. C.; Peters, B.; Nielsen, M.; Marcatili, P. BepiPred-2.0: Improving Sequence-Based B-Cell Epitope Prediction Using Conformational Epitopes. *Nucleic Acids Res.* **2017**, *45*, W24–W29.
- (78) EL-Manzalawy, Y.; Dobbs, D.; Honavar, V. Predicting Linear B-cell Epitopes Using String Kernels. *J. Mol. Recognit.* **2008**, *21*, 243–255.
- (79) Dimitrov, I.; Flower, D. R.; Doytchinova, I. AllerTOP - a Server for in Silico Prediction of Allergens. *BMC Bioinformatics* **2013**, *14*, S4.
- (80) Saha, S.; Raghava, G. P. S. AlgPred: Prediction of Allergenic Proteins and Mapping of IgE Epitopes. *Nucleic Acids Res.* **2006**, *34*, W202–W209.
- (81) Sharma, N.; Patiyal, S.; Dhall, A.; Pande, A.; Arora, C.; Raghava, G. P. S. AlgPred 2.0: An Improved Method for Predicting Allergenic Proteins and Mapping of IgE Epitopes. *Brief. Bioinform.* **2021**, *22*. DOI: 10.1093/bib/bbaa294.
- (82) Gupta, S.; Kapoor, P.; Chaudhary, K.; Gautam, A.; Kumar, R.; Raghava, G. P. S. In Silico Approach for Predicting Toxicity of Peptides and Proteins. *PLoS One* **2013**, *8*, e73957.
- (83) Sharma, N.; Naorem, L. D.; Jain, S.; Raghava, G. P. S. ToxinPred2: An Improved Method for Predicting Toxicity of Proteins. *Brief. Bioinform.* **2022**, *23*. DOI: 10.1093/bib/bbac174.
- (84) Geourjon, C.; Deléage, G. SOPMA: Significant Improvements in Protein Secondary Structure Prediction by Consensus Prediction from Multiple Alignments. *Bioinformatics* **1995**, *11*, 681–684.
- (85) Yang, J.; Yan, R.; Roy, A.; Xu, D.; Poisson, J.; Zhang, Y. The I-TASSER Suite: Protein Structure and Function Prediction. *Nat. Methods* **2015**, *12*, 7–8.
- (86) Xu, D.; Zhang, Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. *Biophys. J.* **2011**, *101*, 2525–2534.
- (87) Heo, L.; Park, H.; Seok, C. GalaxyRefine: Protein Structure Refinement Driven by Side-Chain Repacking. *Nucleic Acids Res.* **2013**, *41*, W384–W388.
- (88) Kozakov, D.; Hall, D. R.; Xia, B.; Porter, K. A.; Padhorny, D.; Yueh, C.; Beglov, D.; Vajda, S. The ClusPro Web Server for Protein–Protein Docking. *Nat. Protoc.* **2017**, *12*, 255–278.
- (89) Rapin, N.; Lund, O.; Bernaschi, M.; Castiglione, F. Computational Immunology Meets Bioinformatics: The Use of Prediction Tools for Molecular Binding in the Simulation of the Immune System. *PLoS One* **2010**, *5*, e9862.
- (90) Oladipo, E. K.; Ajayi, A. F.; Ariyo, O. E.; Onile, S. O.; Jimah, E. M.; Ezediuno, L. O.; Adebayo, O. I.; Adebayo, E. T.; Odeyemi, A. N.; Oyeleke, M. O.; Oyewole, M. P.; Oguntomi, A. S.; Akindiya, O. E.; Olamoyegun, B. O.; Aremu, V. O.; Arowosaye, A. O.; Aboderin, D. O.; Bello, H. B.; Senbadejo, T. Y.; Awoyelu, E. H.; Oladipo, A. A.; Oladipo, B. B.; Ajayi, L. O.; Majolagbe, O. N.; Oyawoye, O. M.; Oloke, J. K. Exploration of Surface Glycoprotein to Design Multi-Epitope Vaccine for the Prevention of Covid-19. *Informatics Med. Unlocked* **2020**, *21*, 100438.
- (91) Rafi, M. O.; Al-Khafaji, K.; Sarker, M. T.; Taskin-Tok, T.; Rana, A. S.; Rahman, M. S. Design of a Multi-Epitope Vaccine against SARS-CoV-2: Immunoinformatic and Computational Methods. *RSC Adv.* **2022**, *12*, 4288–4310.
- (92) Ysrafil, Y.; Sapiun, Z.; Astuti, I.; Anasiru, M. A.; Slamet, N. S.; Hartati, H.; Husain, F.; Damiti, S. A. Designing Multi-Epitope Based Peptide Vaccine Candidates against SARS-CoV-2 Using Immunoinformatics Approach. *BioImpacts* **2022**, DOI: 10.34172/bi.2022.23769.

- (93) Pitaloka, D. A. E.; Izzati, A.; Amirah, S.; Syakuran, L. A. Multi Epitope-Based Vaccine Design for Protection Against Mycobacterium Tuberculosis and SARS-CoV-2 Coinfection. *Adv. Appl. Bioinforma. Chem.* **2022**, *15*, 43–57.
- (94) Srivastava, S.; Kamthania, M.; Singh, S.; Saxena, A.; Sharma, N. Structural Basis of Development of Multi-Epitope Vaccine against Middle East Respiratory Syndrome Using in Silico Approach. *Infect. Drug Resist.* **2018**, *11*, 2377–2391.
- (95) Ikram, A.; Zaheer, T.; Awan, F. M.; Obaid, A.; Naz, A.; Hanif, R.; Paracha, R. Z.; Ali, A.; Naveed, A. K.; Janjua, H. A. Exploring NS3/4A, NSSA and NSSB Proteins to Design Conserved Subunit Multi-Epitope Vaccine against HCV Utilizing Immunoinformatics Approaches. *Sci. Rep.* **2018**, *8*, 16107.
- (96) Pandey, R. K.; Ojha, R.; Aathmanathan, V. S.; Krishnan, M.; Prajapati, V. K. Immunoinformatics Approaches to Design a Novel Multi-Epitope Subunit Vaccine against HIV Infection. *Vaccine* **2018**, *36*, 2262–2272.
- (97) Aziz, S.; Waqas, M.; Halim, S. A.; Ali, A.; Iqbal, A.; Iqbal, M.; Khan, A.; Al-Harrasi, A. Exploring Whole Proteome to Conceive Multi-Epitope-Based Vaccine for NeoCoV: An Immunoinformatics and in-Silico Approach. *Front. Immunol.* **2022**, *13*. DOI: 10.3389/fimmu.2022.956776.
- (98) Chauhan, V.; Singh, M. P. Immuno-Informatics Approach to Design a Multi-Epitope Vaccine to Combat Cytomegalovirus Infection. *Eur. J. Pharm. Sci.* **2020**, *147*, 105279.
- (99) Chauhan, V.; Rungta, T.; Goyal, K.; Singh, M. P. Designing a Multi-Epitope Based Vaccine to Combat Kaposi Sarcoma Utilizing Immunoinformatics Approach. *Sci. Rep.* **2019**, *9*, 2517.
- (100) Ali, M.; Pandey, R. K.; Khatoon, N.; Narula, A.; Mishra, A.; Prajapati, V. K. Exploring Dengue Genome to Construct a Multi-Epitope Based Subunit Vaccine by Utilizing Immunoinformatics Approach to Battle against Dengue Infection. *Sci. Rep.* **2017**, *7*, 9232.
- (101) Narula, A.; Pandey, R. K.; Khatoon, N.; Mishra, A.; Prajapati, V. K. Excavating Chikungunya Genome to Design B and T Cell Multi-Epitope Subunit Vaccine Using Comprehensive Immunoinformatics Approach to Control Chikungunya Infection. *Infect. Genet. Evol.* **2018**, *61*, 4–15.
- (102) Kaur, R.; Arora, N.; Jamakhani, M. A.; Malik, S.; Kumar, P.; Anjum, F.; Tripathi, S.; Mishra, A.; Prasad, A. Development of Multi-Epitope Chimeric Vaccine against Taenia Solium by Exploring Its Proteome: An in Silico Approach. *Expert Rev. Vaccines* **2020**, *19*, 105–114.
- (103) Yousafi, Q.; Amin, H.; Bibi, S.; Rafi, R.; Khan, M. S.; Ali, H.; Masroor, A. Subtractive Proteomics and Immuno-Informatics Approaches for Multi-Peptide Vaccine Prediction Against Klebsiella Oxytoca and Validation Through In Silico Expression. *Int. J. Pept. Res. Ther.* **2021**, *27*, 2685–2701.
- (104) Mahapatra, S. R.; Dey, J.; Kaur, T.; Sarangi, R.; Bajoria, A. A.; Kushwaha, G. S.; Misra, N.; Suar, M. Immunoinformatics and Molecular Docking Studies Reveal a Novel Multi-Epitope Peptide Vaccine against Pneumonia Infection. *Vaccine* **2021**, *39*, 6221–6237.
- (105) Bhatt, P.; Sharma, M.; Prakash Sharma, P.; Rathi, B.; Sharma, S. Mycobacterium Tuberculosis Dormancy Regulon Proteins Rv2627c and Rv2628 as Toll like Receptor Agonist and as Potential Adjuvant. *Int. Immunopharmacol.* **2022**, *112*, 109238.
- (106) Cheng, P.; Wang, L.; Gong, W. In Silico Analysis of Peptide-Based Biomarkers for the Diagnosis and Prevention of Latent Tuberculosis Infection. *Front. Microbiol.* **2022**, *13*. DOI: 10.3389/fmicb.2022.947852.
- (107) Kayesh, M. E. H.; Kohara, M.; Tsukiyama-Kohara, K. An Overview of Recent Insights into the Response of TLR to SARS-CoV-2 Infection and the Potential of TLR Agonists as SARS-CoV-2 Vaccine Adjuvants. *Viruses* **2021**, *13*, 2302.
- (108) Yang, J.-X.; Tseng, J.-C.; Yu, G.-Y.; Luo, Y.; Huang, C.-Y. F.; Hong, Y.-R.; Chuang, T.-H. Recent Advances in the Development of Toll-like Receptor Agonist-Based Vaccine Adjuvants for Infectious Diseases. *Pharmaceutics* **2022**, *14*, 423.
- (109) Anwar, M. A.; Choi, S. Structure-Activity Relationship in TLR4 Mutations: Atomistic Molecular Dynamics Simulations and Residue Interaction Network Analysis. *Sci. Rep.* **2017**, *7*, 43807.
- (110) Gosu, V.; Son, S.; Shin, D.; Song, K.-D. Insights into the Dynamic Nature of the DsRNA-Bound TLR3 Complex. *Sci. Rep.* **2019**, *9*, 3652.
- (111) Sun, J.; Duffy, K. E.; Ranjith-Kumar, C. T.; Xiong, J.; Lamb, R. J.; Santos, J.; Masarapu, H.; Cunningham, M.; Holzenburg, A.; Sarisky, R. T.; Mbow, M. L.; Kao, C. Structural and Functional Analyses of the Human Toll-like Receptor 3. *J. Biol. Chem.* **2006**, *281*, 11144–11151.
- (112) Wang, Y.; Wu, S.; Zhang, C.; Jin, Y.; Wang, X. Dissecting the Role of N-Glycan at N413 in Toll-like Receptor 3 via Molecular Dynamics Simulations. *J. Chem. Inf. Model.* **2022**, *62*, S258–S266.
- (113) Mahita, J.; Sowdhagini, R. Investigating the Effect of Key Mutations on the Conformational Dynamics of Toll-like Receptor Dimers through Molecular Dynamics Simulations and Protein Structure Networks. *Proteins Struct. Funct. Bioinforma.* **2018**, *86*, 475–490.
- (114) Ghosh, S. K.; Saha, B.; Banerjee, R. Insight into the Sequence-Structure Relationship of TLR Cytoplasm's Toll/Interleukin-1 Receptor Domain towards Understanding the Conserved Functionality of TLR 2 Heterodimer in Mammals. *J. Biomol. Struct. Dyn.* **2021**, *39*, 5348–5357.
- (115) Landrum, M. J.; Lee, J. M.; Benson, M.; Brown, G. R.; Chao, C.; Chitipiralla, S.; Gu, B.; Hart, J.; Hoffman, D.; Jang, W.; Karapetyan, K.; Katz, K.; Liu, C.; Maddipatla, Z.; Malheiro, A.; McDaniel, K.; Ovetsky, M.; Riley, G.; Zhou, G.; Holmes, J. B.; Kattman, B. L.; Maglott, D. R. ClinVar: Improving Access to Variant Interpretations and Supporting Evidence. *Nucleic Acids Res.* **2018**, *46*, D1062–D1067.
- (116) Patra, M. C.; Kwon, H.-K.; Batool, M.; Choi, S. Computational Insight Into the Structural Organization of Full-Length Toll-Like Receptor 4 Dimer in a Model Phospholipid Bilayer. *Front. Immunol.* **2018**, *9*. DOI: 10.3389/fimmu.2018.00489.
- (117) Matamoros-Recio, A.; Franco-Gonzalez, J. F.; Perez-Regidor, L.; Billod, J. M.; Guzman-Caldentey, J.; Martin-Santamaria, S. Full-Atom Model of the Agonist LPS-Bound Toll-like Receptor 4 Dimer in a Membrane Environment. *Chem. - A Eur. J.* **2021**, *27*, 15406–15425.
- (118) Patra, M. C.; Batool, M.; Haseeb, M.; Choi, S. A Computational Probe into the Structure and Dynamics of the Full-Length Toll-like Receptor 3 in a Phospholipid Bilayer. *Int. J. Mol. Sci.* **2020**, *21*, 2857.

The proteolytic cleavage of TLR8 Z-loop by furin protease - molecular recognition, reaction mechanism and role of water molecules

Maria Bzówka (✉ maria.bzowka@polsl.pl)

Silesian University of Technology <https://orcid.org/0000-0001-6802-8753>

Katarzyna Szleper

Silesian University of Technology

Agnieszka Stańczak

Czech Academy of Sciences

Tomasz Borowski

Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences

Artur Góra

Silesian University of Technology <https://orcid.org/0000-0003-2530-6957>

Article

Keywords:

Posted Date: November 14th, 2023

DOI: <https://doi.org/10.21203/rs.3.rs-3590328/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

1 The proteolytic cleavage of TLR8 Z-loop by furin protease - molecular recognition, reaction
2 mechanism and role of water molecules

3
4 Maria Bzówka^{1,2*}, Katarzyna Szleper¹, Agnieszka Stańczak^{3,4}, Tomasz Borowski⁵, Artur
5 Góra¹

6
7 ¹Tunneling Group, Biotechnology Centre, Silesian University of Technology, Gliwice,
8 Poland

9 ²Department of Organic Chemistry, Bioorganic Chemistry and Biotechnology, Faculty of
10 Chemistry, Silesian University of Technology, Gliwice, Poland

11 ³Institute of Organic Chemistry and Biochemistry, Czech Academy of Sciences, Prague,
12 Czech Republic

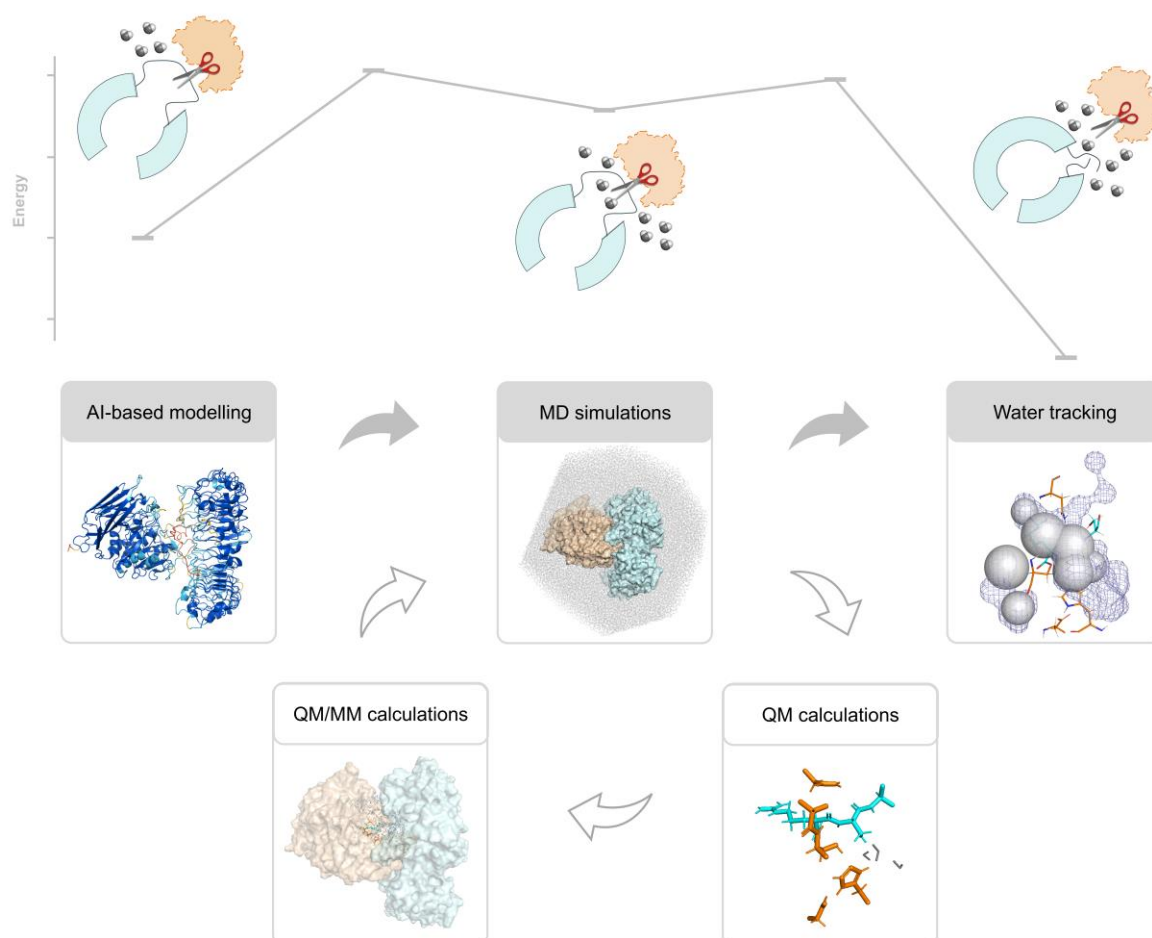
13 ⁴Faculty of Science, Charles University, Prague, Czech Republic

14 ⁵Jerzy Haber Institute of Catalysis and Surface Chemistry, Polish Academy of Sciences,
15 Kraków, Poland

16

17 * Corresponding Author: maria.bzowka@polsl.pl or m.bzowka@tunnelinggroup.pl

18



19

20

21 **Abstract:**

22

23 Understanding the mechanisms underlying the immune response is crucial for advancing our
24 knowledge of an organism's defence. One such mechanism is the proteolytic cleavage of the
25 Z-loop in some members of the Toll-like receptor (TLR) family. This process is essential for
26 several reasons: it allows proper receptor dimerisation, facilitates its activation, and
27 introduces a control mechanism by preventing inappropriate or excessive immune reactions.
28 In our study, we focused on investigating the proteolytic cleavage of TLR8 by furin protease,
29 for which the mechanism of this process has not been widely investigated. We employed
30 various computational methods not only to propose the reaction pathway but also to explore
31 the role of water molecules within the reaction site. Those included AI-based structure
32 prediction, molecular dynamics simulations, quantum mechanics and quantum
33 mechanics/molecular mechanics calculations, as well as small-molecule tracking combined
34 with local-distribution methods.

35

36

37

38 Toll-like receptors (TLRs) are transmembrane proteins that play an important role in
39 recognising molecular patterns (MPs) associated with pathogens or damage. When a ligand
40 binds to a TLR, it either prompts the formation of a receptor dimer or alters the conformation
41 of a preexisting dimer, which enables adaptor proteins to bind and trigger an innate immune
42 response¹⁻⁴. For some members of the TLR family, an additional step is needed to allow MP
43 recognition and further activation of the receptor. This involves the proteolytic cleavage of
44 the long-loop region, (Z-loop), inserted in the N-terminal domain containing leucine-rich
45 repeats motifs (LRR domain) of TLR7-9. TLR7-9 can be found in the intracellular
46 compartments of the cell, mostly in the endosomal membrane, and primarily recognise
47 nucleic acids from viruses and bacteria⁵⁻¹¹. It is assumed that proteolytic cleavage not only
48 serves as a regulatory mechanism but also ensures that the immune system does not
49 inappropriately target self-nucleic acids. Dysregulation of this process can lead to excessive
50 immune responses, potentially contributing to autoimmune disorders.

51
52 In our study, we focused on investigating the proteolytic cleavage of TLR8. Ishii *et al.*¹²
53 confirmed that the cleaved form of this receptor is predominant in immune cells. Moreover,
54 Tanji *et al.*¹¹ showed that TLR8 with the uncleaved Z-loop is unable to form a dimer, which
55 is essential for proper functioning. Yet, the molecular bases of the Z-loop cleavage have not
56 been explored in depth⁴. The literature indicates that furin-like proprotein convertase and
57 cathepsins might contribute to TLR8 cleavage. Analysis of the tetrabasic amino acid
58 sequence before the proteolytic cleavage site, R452-K453-R454-R455↓S456 (RKRR↓S),
59 suggests that furin might be primarily involved in the enzymatic reaction. This is because the
60 R-X-K/R-R↓ motif is preferentially recognised by this enzyme¹³⁻¹⁵. Therefore, we chose furin
61 to investigate the cleavage mechanism of TLR8 with the use of *in silico* methods. While
62 studying this process we relied on the general mechanism of serine proteases¹⁶, since the
63 catalytic site of furin is composed of serine, histidine and aspartate. Also, we used the
64 information about the first step of the acylation process in furin complexed with the H5N1
65 avian influenza virus¹⁷. Besides studying the reaction *per se*, we explored the dynamics of the
66 system, which consists of two large macromolecules. In particular, we focused on the role of
67 water molecules throughout the enzymatic reaction. Considering the challenges associated
68 with analysing protein-protein complexes, our study also aimed to offer a methodological
69 guide, illustrating how various computational methods can complement each other in the
70 description of such biological systems.

71

72 **Results**

73

74 *Prediction of the TLR8^{LRR}-furin complex*

75

76 Despite the availability of crystal structures of the TLR8 LRR domain, none has the Z-loop
77 region fully resolved, where the proteolytic cleavage site is located. Moreover, there are no
78 complexes of TLR8 with any protease potentially involved in the Z-loop cleavage. To
79 investigate this mechanism and determine whether furin may be involved, it was essential to
80 obtain an accurate prediction of the TLR8^{LRR}-furin complex.

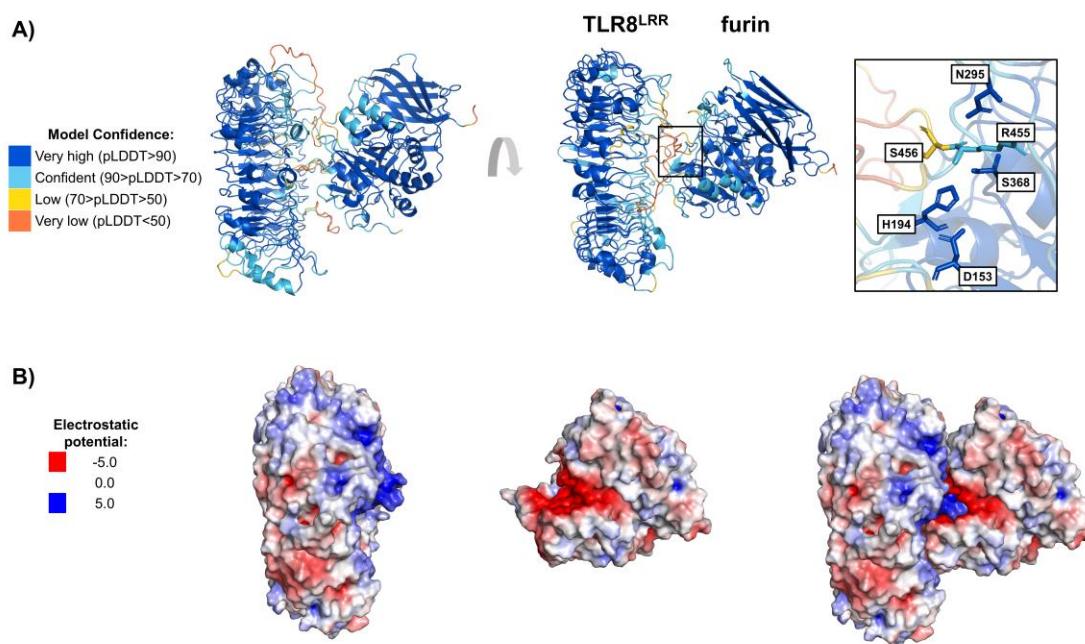
81 Using AlphaFold-Multimer¹⁸, we generated 25 of such predictions. They differed regarding
82 the estimated accuracy (interface predicted TM-score+predicted TM-score; ipTM+pTM
83 ranging from 0.28 to 0.71; scores gathered in **Supplementary Table S1**) and the orientations
84 of the macromolecules towards each other. The highest-ranked predictions, “0” to “6”
85 showed a consistent positioning of the TLR8 and furin, with a Z-loop in a ‘solvent-exposed’
86 conformation, maintaining the proximity between the RKRR↓S fragment and furin’s catalytic
87 site. The estimated accuracy for these predictions was quite high, beginning at 0.71 and
88 decreasing to 0.66. For the predictions “7” to “13”, with scores declining from 0.56 to 0.40,
89 there was a progressive loosening in the maintenance of the interface between the proteolytic
90 cleavage and catalytic sites. The remaining predictions, with scores below 0.40, displayed
91 separation between the subunits. These lower-ranked predictions were excluded from further
92 analyses. We also tried to achieve such an exposed conformation of the Z-loop using standard
93 modelling methods. However, this region tended to curl around the surface of the receptor.
94 Conducting molecular dynamics (MD) simulations also failed to yield the expected
95 arrangement. Therefore, classical protein-protein docking methods were inadequate for
96 accurately predicting the TLR8^{LRR}-furin complex (data not shown).

97 In the Alpha Fold-Multimer predictions “0” to “6”, most of the residues (excluding these
98 from the disordered Z-loop region) displayed high or very high values of the predicted local
99 difference distance test (pLDDT; values in a range 0-100), corresponding to local structural
100 accuracy. To select the best prediction, we focused on evaluating the pLDDT values obtained
101 for Z-loop residues. Notably, we observed that amino acids from the RKRR↓S motif had
102 higher pLDDT values than the remaining residues of the Z-loop. Specifically, for the top-
103 ranked prediction (ranked_0) (**Figure 1A**), the analysed motif displayed the best values:
104 55.83, 67.86, 71.34, 75.45, and 62.43, respectively. Additionally, this prediction showed
105 strong electrostatic compatibility between the subunits. Our analysis confirmed a positive
106 charge of the Z-loop’s proteolytic cleavage region and a contrasting negative charge around
107 the furin’s binding pocket (**Figure 1B**). We also observed favourable MolProbity Score
108 (1.12), Clash Score (0.81), and Ramachandran Favoured percentages (94.72%) which support
109 the geometrical and stereochemical quality of the top prediction. Detailed results of the
110 structural assessment are shown in **Supplementary Figure S1 and Table S2**.

111

112

113



114 **Figure 1 A)** Results of the AlphaFold-Multimer top-ranked prediction of the TLR8^{LRR}-furin
 115 complex (ranked_0) with a close-up of the proteolytic cleavage region and furin catalytic
 116 amino acids and oxyanion hole. The complex is coloured according to the values of the
 117 predicted local difference distance test (pLDDT). R455 and S456 from the TLR8 proteolytic
 118 cleavage site, D153, H194, N295, and S368 from the furin active site and oxyanion hole, are
 119 shown in stick representation. **B)** Results of the electrostatic potential analysis for the top-
 120 ranked prediction. The TLR8 is shown on the left side, furin is shown in the middle and
 121 TLR8^{LRR}-furin complex is shown on the right side. Surface regions with negative
 122 electrostatic potentials are coloured in red, those with positive electrostatic potentials are
 123 coloured in blue.

124
 125

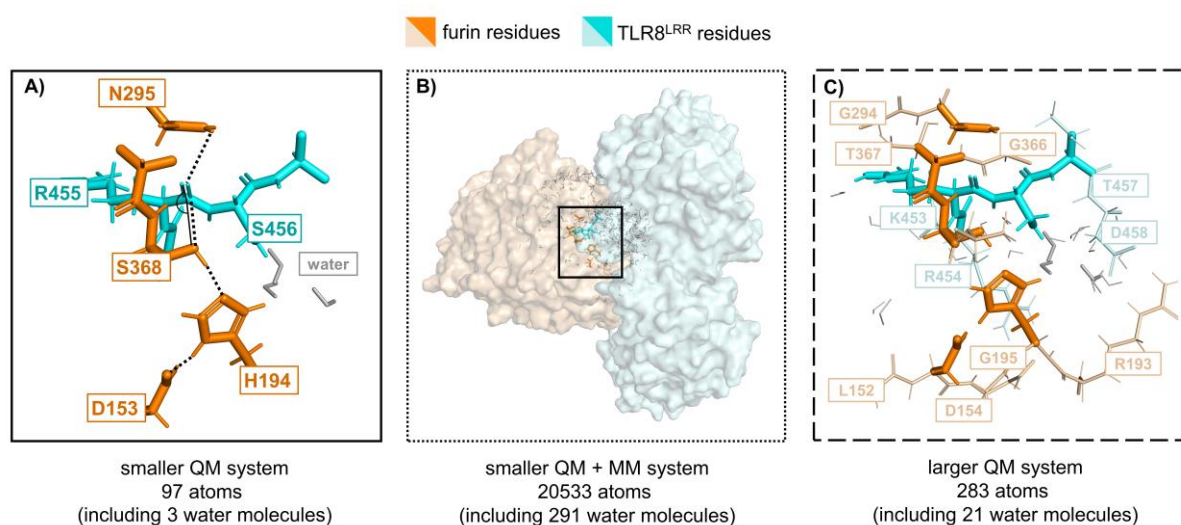
126 *Proposed mechanism and reaction profile of the proteolytic cleavage in TLR8*

127

128 Due to the lack of experimental data for this system, we performed QM calculations to
 129 investigate the proteolytic cleavage mechanism of the TLR8^{LRR}-furin complex. The input
 130 structure for calculations was selected based on the analysis of MD simulations for the
 131 TLR8^{LRR}-furin reactant and the fulfilment of the near attack conformation criteria (NAC)
 132 outlined in the **Methods** section. We started the investigation using QM calculations for the
 133 simplest QM-cluster model comprising amino acids involved in the analysed reaction
 134 (**Figure 2A**). While we successfully obtained converged minima, the corresponding energy
 135 values were not satisfactory. In parallel, we attempted to employ the QM/MM ONIOM
 136 approach to optimise this pre-defined QM-cluster region in the whole TLR8^{LRR}-furin
 137 complex (**Figure 2B**). Due to the prolonged optimisation time for such a large system, it was
 138 not possible to rely solely on the QM/MM method to determine the reaction coordinates.
 139 Therefore, we incorporated the previously optimised geometries of intermediate states and
 140 the product from QM calculations as QM region in QM/MM calculations. Nevertheless, the

141 energy values of the converged minima remained excessively high. We performed additional
 142 QM calculations for the expanded QM-cluster region (**Figure 2C**). This was intended to more
 143 accurately represent the vicinity of the reaction site in the complex. Indeed, this change in the
 144 system size influenced the energy values we obtained (**Supplementary Table S3**). We
 145 further explored the impact of various functionals. The obtained energy values highlighted
 146 that the energy profile with the lowest energy barriers was achieved using the BP86-
 147 D3(BJ)/def2-TZVPD level of theory combined with the Conductor-like screening model for
 148 the realistic solvation (COSMO-RS)¹⁹. COSMO-RS has been known to perform effectively in
 149 characterising enzymatic reactions^{20,21}.

150



151 **Figure 2** Representation of models used for QM and QM/MM calculations. Furin residues
 152 are represented as orange sticks with surface coloured in pale orange. TLR8 residues are
 153 represented as cyan sticks with surface coloured in pale cyan. Water molecules are shown as
 154 grey sticks. **A)** Smaller QM-cluster model. The internal coordinates used to define NAC
 155 criteria are depicted; black dashed lines indicate crucial interactions and black solid line
 156 shows the angle for nucleophilic attack angle. **B)** System partitioning into the QM region
 157 (defined by smaller QM-cluster model atoms) and MM region (remaining residues of
 158 TLR8^{LRR}-furin complex and water molecules within 20 Å of the reaction site). **C)** Larger
 159 QM-cluster model.

160

161 The presented putative reaction pathway consists of four steps (**Figure 3**). The energy values
 162 presented below were calculated using the BP86-D3(BJ)/def2-TZVPD level of theory with
 163 COSMO-RS solvation model including ΔG_{freq} correction.

164 (i) In the first step, after TLR8^{LRR}-furin complex formation, there was a straightforward path
 165 from the reactant (RE) *via* TS1 to the first tetrahedral intermediate (INT1), which involved C-
 166 O bond formation between R455 and S368. In the RE complex, the key C-O distance was $R_{\text{C-O}} = 3.09$ Å. H194 was a plausible H-acceptor and formed a hydrogen bond with S368 ($R_{\text{H-S}} = 2.82$ Å). Along the reaction coordinates C-O bond was formed (distances of 1.87 Å and 1.51

169 Å, in TS1 and INT1, respectively) concomitant with proton transfer from S368 to H194,
170 which resulted in the tetrahedral intermediate. The ΔG values (set at 0.0 kcal/mol for RE)
171 were 21.8 and 20.7 kcal/mol for TS1 and INT1, respectively. This step might or might not be
172 also mediated by a water molecule present between H194 and S368. This water molecule
173 could serve first as a proton shuttle and then stabilise the tetrahedral intermediate. The RE
174 complex with a water molecule present between H194 and S368 differed in energy from the
175 initial structure only by 1.3 kcal/mol. However, TS1 in a water-mediated scenario had ΔG of
176 24.8 kcal/mol, hence we did not proceed with this pathway.

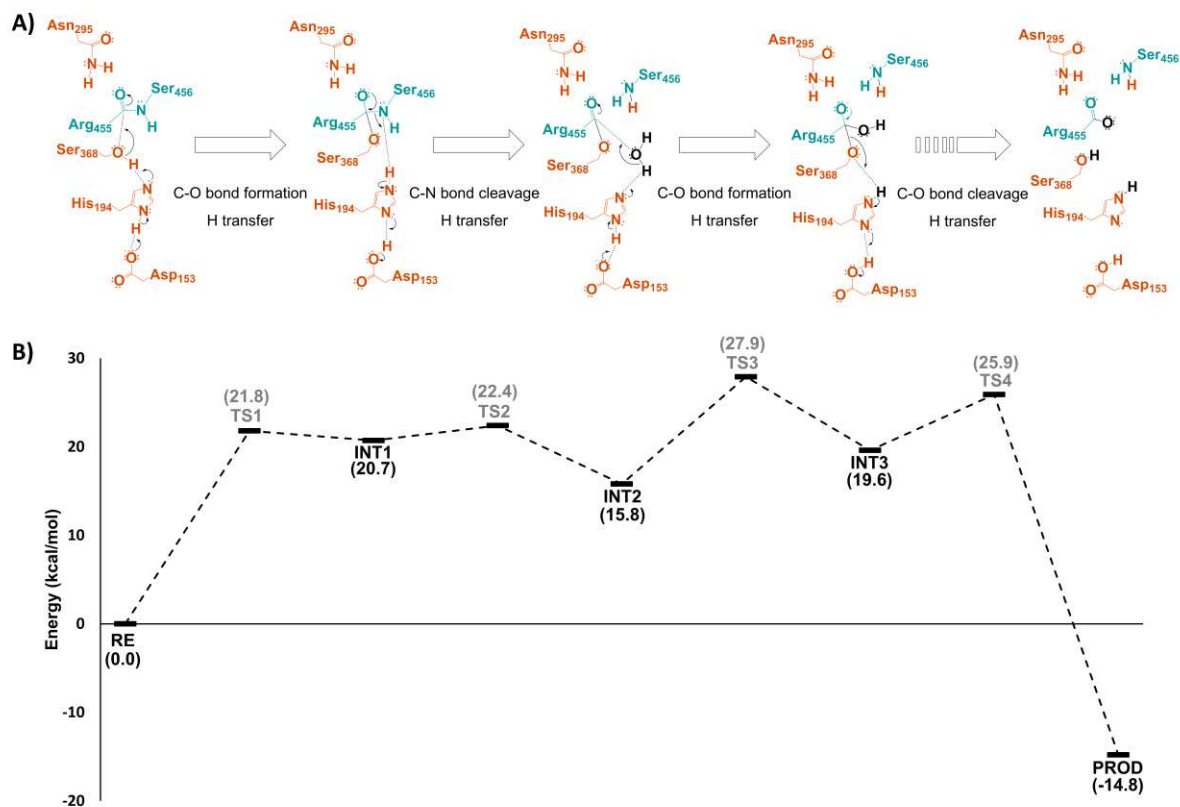
177 (ii) The second step was the C-N bond cleavage of R455 in a newly formed and rather
178 unstable tetrahedral intermediate ($R_{C-N} = 1.43$ Å in INT1). Elongation of this bond between
179 R455 and S456 resulted in a second intermediate – INT2 (acyl-enzyme). The key distances of
180 the C-N bond and ΔG values were $R_{C-N} = 2.30$ Å and 22.4 kcal/mol for TS2, and $R_{C-N} = 3.35$
181 Å and 15.8 kcal/mol for INT2. This bond cleavage was accompanied by proton transfer from
182 H194 to the N-terminus of S456.

183 (iii) Further, to hydrolyse the ester bond between R455 and S456, the catalytic water
184 molecule came into the active site. After activation by H194 (as a proton acceptor), this water
185 molecule formed the C-O bond, which resulted in a following tetrahedral intermediate (INT3)
186 *via* the TS3 – energy barrier of this step was 12.1 kcal/mol. The ΔG values of TS3 and INT3
187 were 27.9 and 19.6 kcal/mol, respectively. The C-O bond distance decreased from 3.8 Å of
188 INT2 *via* 1.90 Å in TS3 towards a fully created C-O bond of 1.47 Å in INT3.

189 (iv) The last step of the reaction led to the regeneration of the active site and the release of the
190 final product. The C-O bond of R455 and S368 was elongated and fully cleaved ($R_{C-O} = 1.47$
191 Å, 1.87 Å, and 3.26 Å in INT3, TS4, and PROD, respectively). The barrier of this last step
192 was 6.3 kcal/mol and the ΔG values of TS4 and PROD were: 25.9, and -14.8 kcal/mol,
193 respectively. During the investigation, the first considered product contained a carboxyl
194 group in R455. Since this group was in close proximity to H194, it resulted in a spontaneous
195 deprotonation of R455 and formation of the most stable product. Alternatively, we observed a
196 possibility of transferring the proton from the R455 carboxyl group to the S456, however, it
197 did not have the lowest energy. Comparison of product structures and their energies is
198 provided in **Supplementary Figure S2**.

199 Cartesian coordinates of all optimised structures at various levels of theory are available in
200 the provided repository.

201



202

203 **Figure 3 A)** Proposed reaction mechanism of the proteolytic cleavage of the TLR8 Z-loop by
 204 furin protease (residue numbering as in crystal structures). Furin residues are coloured in
 205 orange, TLR8 residues are coloured in cyan. **B)** The calculated energy profile of the reaction
 206 for key intermediates and transition states. Energies calculated at the BP86-D3(BJ)/def2-
 207 TZVPD level of theory with COSMO-RS solvation model including ΔG_{freq} correction.

208

209

210 *Analysis of complex dynamics and interactions network*

211

212 Assuming that the rate of the analysed reaction allows the rearrangement of amino acids' side
 213 chains, we used structures optimised by the QM/MM ONIOM to run MD simulations and
 214 analyse the dynamics for each reaction species. For all data, we performed the following
 215 analyses: root mean square deviation (RMSD), root mean square fluctuation (RMSF), key
 216 distances and hydrogen bond network (**Supplementary Figure S3-S12, Supplementary**
 217 **Table S4-S9**). Based on the results, we described and illustrated selected 100-ns fragments of
 218 simulations that ensured the TLR8^{LRR}-furin complex stability and best illustrated the
 219 proposed reaction mechanism (**Figure 4**).

220

221 For the RE complex, RMSD values oscillated in the range of 1-1.8Å, reflecting minor
 222 variations in the residues' side chains within the defined QM region. We attributed it to the
 223 complex's tendency to achieve and maintain the optimal geometry for initiating the
 224 enzymatic reaction. Importantly, throughout all the repetitions, the TLR8^{LRR}-furin complex
 225 remained stable. We described the part of the simulation (the second 100-ns fragment from

226 the first repetition) that was used to select the frame for the QM and QM/MM calculations.
227 We observed that the ϵ -nitrogen of H194 and the hydroxyl hydrogen of S368 were
228 consistently positioned within 2 to 3.3 Å and could form a hydrogen bond for 58% of the
229 time. Similarly, the distance between the hydroxyl oxygen of S368 and the carbonyl carbon
230 of R455 maintained a value of \sim 3.3 Å. The values of the angle measured between the
231 hydroxyl oxygen of S368 and the carbonyl group of R455 fluctuated around 80°, which
232 corresponded to the nucleophilic attack criterium requiring a value close to 90°. The distance
233 between the carbonyl oxygen of R455 and amino hydrogen of N295 displayed small
234 fluctuations within a range from 2 to 4 Å. Notably, we identified an interaction at \sim 3.3 Å of
235 R455 carbonyl oxygen with main chain amide hydrogen of T367. We observed the possibility
236 of forming the hydrogen bond between the carbonyl oxygen of R455 and main chain amide
237 hydrogen of S368 (28%). Furthermore, the distance of about 2 Å between the carboxylate
238 oxygen of D153 and the hydrogen atom of the δ -nitrogen of H194 remained consistent. For
239 these atoms, we confirmed the possibility of forming a hydrogen bond (96%). Additionally,
240 we identified the possibility of forming hydrogen bonds with solvent molecules by D153
241 carboxylate oxygen (18%), H194 ϵ -nitrogen (17%) and S368 hydroxyl group (11%).
242

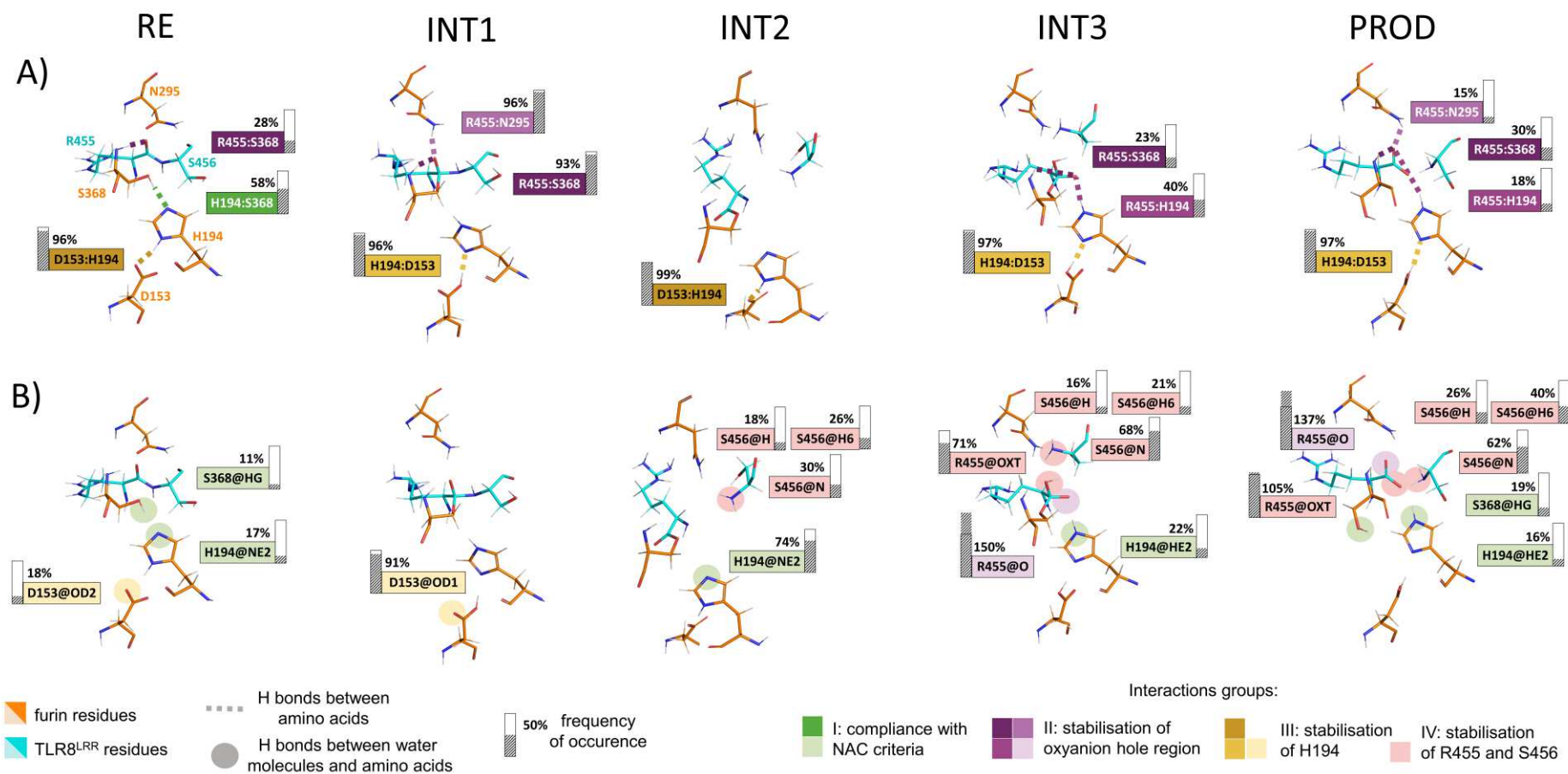
243 For the INT1 species, the RMSD values were in the range of 0.5-1.2 Å. Amino acids
244 involved in the reaction retained a stable arrangement during the analysed MD simulations.
245 To provide details, we selected the second 100ns-fragment from the fourth repetition. We
246 focused on the scenario resulting from the reaction profile, where the stabilisation of the
247 oxyanion takes place directly by an amino acid from the oxyanion hole. We observed that the
248 negatively charged oxygen of R455 was stabilised by both the amino group of N295 (96% of
249 the time) and main chain amide hydrogen of S368 (93%), positioned at the distance close to 2
250 Å. Similarly to the RE, the distance between the R455 carbonyl oxygen and N-terminus
251 amino hydrogen of T367 was below 3.3 Å. The interaction between the δ -nitrogen of H194
252 and the carboxyl hydrogen of D153 was stable (96%) and maintained at \sim 2 Å. Additionally,
253 we identified the hydrogen bond between one of the carboxyl oxygens of D153 and the
254 solvent (91%). We also noticed that the hydrogen atom placed on the ϵ -nitrogen of H194 was
255 oriented towards the N-terminus nitrogen of S456, keeping the distance within 3-4 Å.
256

257 RMSD values for the INT2 species ranged from 0.8 to 1.5 Å. We linked the observed
258 fluctuations with the increased flexibility of the S456, which is a direct consequence of the
259 cleavage of the peptide bond between R455 and S456, as well as with the flexibility of the
260 remaining amino acids from the TLR8^{LRR}-furin interface, surrounding the defined QM
261 region. To characterise details, we focused on the second 100-ns fragment from the fifth
262 repetition. We indicated the interactions of H194 ϵ -nitrogen with solvent molecules
263 throughout 74% of the time. Also, the carboxylate oxygen of D153 and the hydrogen atom
264 placed on the δ -nitrogen of H194 were consistently positioned within close proximity of 2 Å,
265 with a possibility of forming hydrogen bond (99%). Notably, we observed the possibility of
266 forming multiple hydrogen bonds with solvent molecules by the newly released S456 N-
267 terminus amino group (nitrogen - 30%, hydrogens - 26% and 18%).
268

269 For the INT3 species, RMSD values ranged from 0.6 to 1.5 Å, with some peaks up to 2 Å

270 observed in a single repetition. Again, we attributed the fluctuations to the flexibility of
271 residues outside the defined QM region and the movement of the S456. We chose the first
272 100-ns fragment from the first repetition of the MD simulation to describe details. Depending
273 on its orientation, the negatively charged oxygen of R455 could be partially stabilised by
274 hydrogen bonds with either ϵ -nitrogen of H194 (40% of the time) or the N-terminus amino
275 group of S368 (23%). However, much stronger stabilisation was achieved by interaction with
276 solvent molecules (150%). We observed that the δ -nitrogen of H194 and the carboxyl
277 hydrogen of D153 were positioned at ~ 2 Å from each other and formed a hydrogen bond
278 (97%). Similarly to INT2, we indicated the possibility of forming multiple hydrogen bonds
279 with solvent molecules by the S456 N-terminus amino group (nitrogen - 68%, hydrogens -
280 21%, and 16%) and additionally by the R455 tetrahedral intermediate hydroxyl group (71%).
281

282 For the PROD species, we observed that the RMSD values were consistently within 0.5-1.2
283 Å, with a single exception reaching 1.8 Å. These fluctuations were independent of residues
284 within the defined QM region. To provide details, we selected the first 100-ns fragment from
285 the fifth repetition of the MD simulation. The carboxylate group of R455 could be stabilised
286 by hydrogen bonds formed with the hydrogen positioned on ϵ -nitrogen of H194 (18% of the
287 time), the amino group of N295 (15%) and the main chain amide group of S368 (30%).
288 Additionally, we observed strong interactions between the oxygens of the R455 carboxylate
289 group and water molecules (137%, 105%). In the analysed fragment, the orientation of R455
290 carboxylate group varied, often positioning it within a distance of less than 3.3 Å from the ϵ -
291 nitrogen of H194, main chain amide hydrogen of T367, and main chain amide group of S368.
292 D153 residue stabilised catalytic H194 by hydrogen bond (97%), keeping the distance
293 between the δ -nitrogen and the carboxyl hydrogen ~ 2 Å. Additionally, we found interactions
294 between solvent molecules and a hydrogen atom placed on the ϵ -nitrogen (16%) and S368
295 hydroxyl group (19%).



296 **Figure 4** Analysis of interaction network for the consecutive steps of the proteolytic cleavage reaction for the TLR8^{LRR}-furin complex (RE,
 297 INT1-INT3, PROD). Results are shown for the selected 100-ns fragments of MD simulations. **A)** Hydrogen bond formation frequencies between
 298 amino acids in TLR8^{LRR}-furin complex. **B)** Hydrogen bond formation frequencies between water molecules and amino acids in the TLR8^{LRR}-
 299 furin complex. Only those hydrogen bonds are shown, for which the percentage of occurrence was above 10% in the analysed fragment of
 300 simulation. The frequency exceeding 100% means that more solvent molecules fulfilled the criteria to form hydrogen bonds.

301 *Analysis of water molecules reorganisation*

302

303 To get a better understanding of the behaviour of solvent during each step of the reaction,
304 first, we performed radial distribution function (RDF) analysis. We evaluated the probability
305 of finding water molecules within the reaction site, defined as the region within the centre of
306 geometry of carbonyl carbon of R455 from TLR8 and ϵ -nitrogen of H194 from furin. This
307 analysis showed differences in the function's variability between the reaction steps. In
308 general, we observed a notable increase in the probability of finding water molecules within
309 the reaction site, as the reaction progressed (**Supplementary Figure S13**). However, the
310 RDF analysis did not specify the exact location of water molecules. Therefore, we tracked
311 water molecules within the reaction site, computed the volumes of regions (inner pockets)
312 penetrated by water molecules and described the high-density water sites (hot-spots)^{22,23}.
313 Information about the volumes of inner pockets is presented in **Supplementary Table S10**.
314 Below, we describe RDF, inner pockets and the identified hot-spots for previously selected
315 100-ns fragments of MD simulations and show the results in **Figure 5**.

316

317 For the RE complex, we did not detect any prominent peaks near the pre-defined reaction
318 site. Within a distance 2.25 to 6.75 Å from the reference point, normalised water density
319 values gradually increased to reach a value of 0.23. Despite a minor decline, values
320 consistently ranged between 0.20-0.30. Based on the distribution of the inner pocket (volume
321 of 113 Å³), we observed that only the vicinity of H194 and S456 could be penetrated by
322 water molecules. We observed relatively small hot-spots, with only one located in the
323 proximity to the reaction site, near the ϵ -nitrogen of H194.

324

325 For the INT1 species, we observed a gradual incline of RDF within 2.75 to 8.25 Å, where the
326 maximum value of 0.28 was reached. Further, values slightly dropped but remained stable
327 within the 0.20-0.26 range. The region for possible penetration by water molecules was larger
328 than for the RE complex (156 Å³), and was divided into several pockets. Similarly to the RE
329 complex, none of the observed pockets covered the cleavage site. We identified a few hot-
330 spots, however, none of them was located in a direct proximity to the reaction site. The
331 biggest hot-spot was trapped between D153, H194, and S368.

332

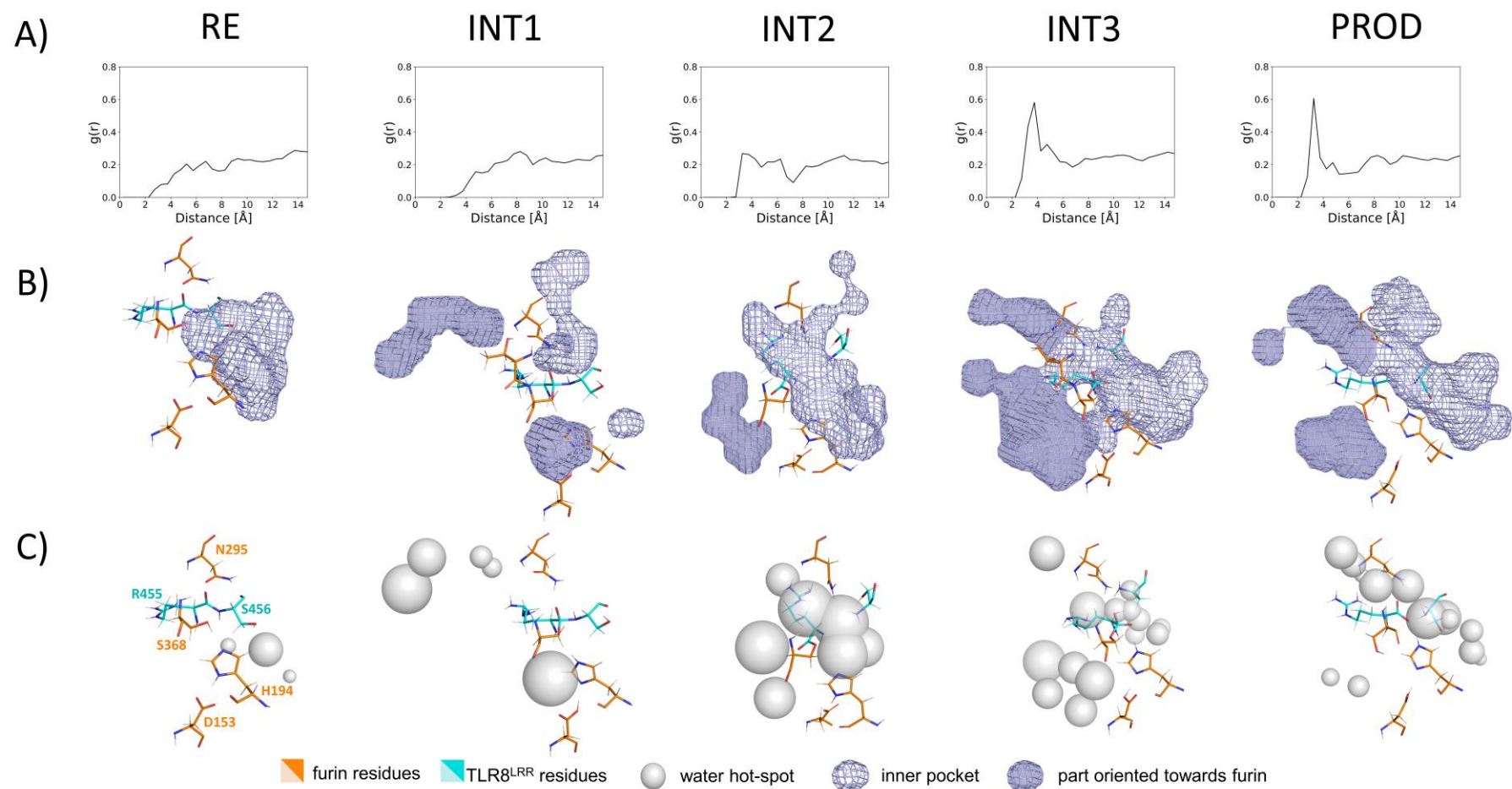
333 For the INT2 species, we observed a different pattern of RDF. There was a broad peak from
334 2.75 to 6.25 Å reaching the maximum of 0.27 at the distance of 3.25 Å. In comparison to
335 previous complexes, increased width of the peak suggests a higher possibility of finding
336 water molecules within the area of interest. The distribution of the inner pocket (140 Å³)
337 reflected the available space resulting from the cleavage of the peptide bond between R455
338 and S456. Additionally, we observed the appearance of a separate, relatively small pocket,
339 yet tightly filled by water, located more deeply in furin. We observed several massive hot-
340 spots crowded in the reaction site. We could associate these findings with the fact that during
341 this step of the reaction, a water molecule is needed to serve as a substrate in bond hydrolysis.

342

343 For the subsequent INT3 species, unlike for INT2, the first peak was sharp and reached a
344 maximum of 0.58 at 3.75 Å. Further, we observed a gradual decline to 0.19 at 6.75 Å,

345 followed by a stabilisation ~ 0.25 . The distribution of the inner pocket indicated a large
346 internal space available for water penetration (373 \AA^3). Notably, a significant part of the inner
347 pocket was oriented towards furin's core, indicating that water molecules may be crowded in
348 the interior of the furin. In comparison to other species, we observed the highest number of
349 middle-size hot-spots. They were placed not only in the reaction site as in the INT2 species,
350 but also deeper in the furin's core.

351
352 For the PROD species, the pattern of the function remained similar as for the INT3 species.
353 However, the observed peak was sharper and narrower, with the maximum of 0.60 at 3.25 \AA
354 distance. Then, RDF values declined gradually to reach 0.14 at 5.25 \AA . Following a small
355 increase, values stabilised ~ 0.25 . The inner pocket (277 \AA^3) was extended throughout a
356 relatively narrow and elongated region between the furin and TLR8 interface. It was
357 accompanied by an additional smaller void, reaching the furin's core. We identified hot-spots
358 mostly placed within the above mentioned interface. Particularly, the biggest hotspot was
359 located between H194 and the C-terminal carboxylate group of R455. Two other big hot-
360 spots were observed close to the carboxylate group of R455 and one in the vicinity of the N-
361 terminal amino group.



362 **Figure 5** Analysis of water molecules reorganisation for the consecutive steps of the proteolytic cleavage reaction for the TLR8^{LRR}-furin
 363 complex (RE, INT1-INT3, PROD). Results are shown for the selected 100-ns fragments of MD simulations. **A)** Radial distribution function
 364 (RDF) plots illustrating the probability of finding water molecules around the reference point set as the midpoint between the carbonyl carbon of
 365 R455 in TLR8 and ϵ -nitrogen of H194 in furin. **B)** Distribution of the solvent densities - inner pockets (shown as violet mesh). **C)** Identified
 366 high-density water hot-spots (shown as grey spheres). The density of hot-spots is reflected by the size of the presented spheres.

367 Discussion

368

369 Understanding the mechanisms underlying the immune response is crucial for advancing our
370 knowledge of an organism's defence. One such mechanism is the proteolytic cleavage of the
371 Z-loop in some members of the TLR family. This process is essential for several reasons: it
372 allows proper receptor dimerisation, facilitates its activation, and introduces a control
373 mechanism by preventing inappropriate or excessive immune reactions^{5,6,9,10}. In our study,
374 we focused on investigating the molecular bases of the proteolytic cleavage of TLR8. We
375 relied on information suggesting the involvement of furin protease in this process¹¹. We
376 employed *in silico* methods not only to propose the reaction pathway but also to shed light on
377 the potential changes in this system, emphasising the role of water molecules.

378

379 We proposed a putative reaction pathway for the proteolytic cleavage of the TLR8 Z-loop by
380 furin, comprising acylation, deacylation, and product release steps, in line with the general
381 mechanism of serine proteases¹⁶. To the best of our knowledge, no computational studies
382 have provided theoretical details of the entire cleavage mechanism by this enzyme.
383 Rungrotmongkol *et al.*¹⁷ studied only the first step of the acylation process in furin
384 complexed with the H5N1 avian influenza virus. They reported that the formation of
385 tetrahedral intermediate (INT1) involves a simultaneous transfer of a proton from S368 to
386 H194 and a nucleophilic attack on the peptide bond, observations we also made. Such
387 concerted acylation distinguishes furin from other proteases, where this part of reaction is
388 typically step-wise²⁴. It is generally accepted that the rate-limiting step in serine proteases is
389 determined by the formation of the INT1^{25,26}. Our results align with this, as the initial step in
390 the proposed reaction pathway has the highest energy barrier, at 21.8 kcal/mol. In their study,
391 Rungrotmongkol *et al.* reported an energy barrier of 16.2 kcal/mol which was in agreement
392 with experimental activation energies (14-21 kcal/mol), converted from the observed rate of
393 cleavage reaction by furin at various substrates and experimental conditions²⁷⁻²⁹. A
394 subsequent study³⁰ carried out for various synthetic peptides also provided kinetic parameters
395 which, when recalculated according to the Eyring equation³¹, gave similar activation energy
396 values. Notably, among the peptides studied, one featured the same RKRR↓S motif as present
397 in the analysed TLR8 Z-loop. For this peptide, the calculated energy barrier was 16.8
398 kcal/mol. We obtained activation energy higher only by 5 kcal/mol which is in a good
399 agreement with experimental values, considering the expected accuracy of DFT. It should be
400 noted that in the experimental study, only small synthetic peptides were examined, whereas
401 our computational work was intended to investigate the cleavage within a large
402 macromolecule. The interaction of these two macromolecules requires continuous adjustment
403 between them, which may also contribute to a rather slow reaction rate. As we demonstrated,
404 the energy of INT1 is higher compared to the reactant (RE) complex (20.7 kcal/mol). This
405 indicates that the first step of acylation is endothermic, which is in general agreement with
406 previous study¹⁶. However, even though we observed that the subsequent species, the acyl-
407 enzyme (INT2), has energy lower than the INT1 (15.8 kcal/mol), it is not lower than the RE.
408 Thus, we could not confirm that the acylation process catalysed by furin is exothermic, as
409 shown in the studies of other serine proteases³²⁻³⁴. As we mentioned, the deacylation process
410 in the furin protease has not been previously studied with theoretical methods. In our study,

411 in the first step of this process, we observed the formation of tetrahedral intermediate (INT3),
412 slightly higher in energy (19.6 kcal/mol) than the previous INT2. Later, the reaction led to the
413 formation of the energetically favourable product (-14.8 kcal/mol), which indicates that the
414 entire enzymatic reaction is exothermic. Our findings correspond to the proposed general
415 energy profile of serine proteases²⁴.

416 In the final step of the proteolytic cleavage, one might expect the presence of a carboxyl
417 group in R455, since such an arrangement of atoms was present in the previous step (INT3).
418 Considering that the reaction takes place in an aqueous solution (pH ~6.5), we could expect
419 ionisation of the terminal groups. The pKa of arginine carboxyl group and serine amino
420 group (1.8 and 9.2, respectively), indicate potential proton transfer. Such a product was
421 obtained and it exhibited lower energy. Another plausible scenario involved transfer of the
422 proton from the carboxyl group of R455 to ϵ -nitrogen of H194, which is also supported by
423 pKa of histidine side chain (6.04). Indeed, those atoms were positioned in close proximity,
424 facilitating the spontaneous deprotonation of R455 and formation of the most energetically
425 stable product.

426
427 Given the assumption of a relatively slow reaction rate, we analysed potential rearrangements
428 in residues' side chains and water molecules within the reaction site. Overall, we observed
429 that reorientation of side chains are quite subtle, however, they still occur. They primarily
430 aim to achieve the optimal positioning for initiating the enzymatic reaction, and subsequently,
431 they adjust to facilitate further reaction steps. Notably, any significant deviations in the side
432 chain dynamics could disturb the reaction progress. Thus, our observations of only subtle
433 changes, in addition to the results obtained from the QM calculations, may indicate that the
434 analysed reaction is feasible. The changes in the positioning of water molecules during the
435 catalytic cycle are more substantial and crucial. First of all, we noticed significant differences
436 in water molecules distribution at various stages of the reaction. At the very beginning, there
437 were almost no water molecules present in the vicinity of the reaction site. As the reaction
438 progressed, solvent molecules first occupied the reaction site and furin's interior, and then
439 moved towards the TLR8^{LRR}-furin interface. Not only positioning of water molecules may
440 vary, but also roles which these molecules exhibit.

441 In the proteolytic cleavage, the fundamental role of water is to act as a substrate in the acyl-
442 enzyme hydrolysis³⁵, which we confirmed. However, we also postulate that it plays a
443 supporting role in the remaining reaction steps. For the high-energy tetrahedral INT1 and
444 INT3 species, it was crucial to ensure a proper stabilisation of the negative charge that
445 develops on the R455 oxygen atom as the reaction progresses. We confirmed that the
446 oxyanion hole region formed by N295 together with S368 and T367 provides such
447 stabilisation. However, in some repetitions of MD simulations, we noticed water molecules
448 being close enough to assist or even take over the stabilising function from these residues.
449 Similar role of water molecules in the stabilisation of oxyanion was also observed in other
450 serine proteases³⁶⁻³⁸. Moreover, the detected water hot-spots in RE, INT3 and PROD species
451 indicate that solvent molecules could function as a proton shuttle between H194 and other
452 residues. We hypothesise that water-mediated proton transfer has the potential to lead to
453 alternative reaction pathways and influence the energy profile. Lastly, our observation of
454 water movement towards the TLR8^{LRR}-furin interface might indicate involvement of solvent

455 molecules in the dissociation process of these macromolecules. An increased presence of
456 water molecules between these macromolecules could potentially weaken the strong
457 electrostatic interactions holding the complex together, facilitating its separation. As Meyer
458 reported, water molecules may occupy the space previously filled by a ligand, thus aiding the
459 release of products in serine proteases³⁹. The presented hypotheses, however, require further
460 investigation that will consider not only the molecular interactions over extended timescales
461 but also the accurate modelling of the environment, for instance, through the selection of
462 suitable water models.

463

464 In this study, we showed how various *in silico* methods can be combined to characterise not
465 only the putative enzymatic reaction mechanism but also the dynamic changes occurring in
466 the system, especially for water molecules. This includes AI-based structure prediction, MD
467 simulations, QM-only and QM/MM calculations, as well as small-molecule tracking
468 combined with local-distribution methods. Such a methodological guide can offer an
469 alternative to the *ab initio* molecular dynamics approach, whose applicability is limited when
470 dealing with large biological systems, like the analysed complex.

471

472

473

474

475

476

477 **Methods**

478

479 A complete description of employed methods and computational setups is available in
480 **Supplementary Information**. This includes:

- 481 i) Prediction of the TLR8^{LRR}-furin complex performed with AlphaFold-Multimer¹⁸,
482 ii) MD simulations of TLR8^{LRR}-furin reactant (RE) complex performed with AMBER18⁴⁰,
483 iii) QM calculations performed with Turbomole 7.6 software⁴¹,
484 iv) QM/MM ONIOM calculations performed with Gaussian 16, Revision C.01⁴²,
485 v) MD simulations of TLR8^{LRR}-furin intermediate species and product (INT1-INT3, PROD)
486 performed with AMBER18 and AMBER22⁴³,
487 vi) Analysis of the interaction network among residues from TLR8, furin protease and
488 solvent molecules performed with AmberTools *cpptraj* program⁴⁴,
489 vii) Tracking of water molecules and identification of hot-spots performed with AQUA-
490 DUCT 1.0 software^{22,23}.

491

492 Generated data is available in the provided repository:

493 <https://doi.org/10.5281/zenodo.10082614>

494

495 **Acknowledgement:**

496

497 The authors would like to thank Jeremy Esque (Toulouse Biotechnology Institute, Université
498 de Toulouse, CNRS, INRAE, INSA) and Weronika Bagrowska (Tunneling Group,
499 Biotechnology Centre) for their help with setting up the AlphaFold calculations.

500

501

502 **Funding:**

503

504 The work of MB was supported by the Ministry of Science and Higher Education, Poland
505 from the budget for science for the years 2019–2023, as a research project under the
506 “Diamond Grant” program [Project Number: DI2018 014148; Agreement Number:
507 0141/DIA/2019/48]. The work of KS, TB, and AG was partially supported by the European
508 Commission - EIC Pathfinder program under grant agreement No. 101046815.

509 We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC
510 Centers: ACK Cyfronet AGH) for providing computer facilities and support within
511 computational grant no PLG/2021/014751 and PLG/2022/015577. The computations were
512 also partially performed at the Poznan Supercomputing and Networking Center.

513

514

515

516 **References**

517
518

- 519 1. Iwasaki, A. & Medzhitov, R. Regulation of Adaptive Immunity by the Innate Immune
520 System. *Science (80-.)*. **327**, 291–295 (2010).
- 521 2. Fitzgerald, K. A. & Kagan, J. C. Toll-like Receptors and the Control of Immunity. *Cell*
522 **180**, 1044–1066 (2020).
- 523 3. Li, D. & Wu, M. Pattern recognition receptors in health and diseases. *Signal*
524 *Transduct. Target. Ther.* **6**, 291 (2021).
- 525 4. Bzówka, M., Bagrowska, W. & Góra, A. Recent Advances in Studying Toll-like
526 Receptors with the Use of Computational Methods. *J. Chem. Inf. Model.* **63**, 3669–
527 3687 (2023).
- 528 5. Ewald, S. E. *et al.* The ectodomain of Toll-like receptor 9 is cleaved to generate a
529 functional receptor. *Nature* **456**, 658–662 (2008).
- 530 6. Park, B. *et al.* Proteolytic cleavage in an endolysosomal compartment is required for
531 activation of Toll-like receptor 9. *Nat. Immunol.* **9**, 1407–1414 (2008).
- 532 7. Sepulveda, F. E. *et al.* Critical Role for Asparagine Endopeptidase in Endocytic Toll-
533 like Receptor Signaling in Dendritic Cells. *Immunity* **31**, 737–748 (2009).
- 534 8. Ewald, S. E. *et al.* Nucleic acid recognition by Toll-like receptors is coupled to
535 stepwise processing by cathepsins and asparagine endopeptidase. *J. Exp. Med.* **208**,
536 643–651 (2011).
- 537 9. Hipp, M. M. *et al.* Processing of Human Toll-like Receptor 7 by Furin-like Proprotein
538 Convertases Is Required for Its Accumulation and Activity in Endosomes. *Immunity*
539 **39**, 711–721 (2013).
- 540 10. Tanji, H., Ohto, U., Shibata, T., Miyake, K. & Shimizu, T. Structural Reorganization
541 of the Toll-Like Receptor 8 Dimer Induced by Agonistic Ligands. *Science (80-.)*. **339**,
542 1426–1429 (2013).
- 543 11. Tanji, H. *et al.* Autoinhibition and relief mechanism by the proteolytic processing of
544 Toll-like receptor 8. *Proc. Natl. Acad. Sci.* **113**, 3012–3017 (2016).
- 545 12. Ishii, N., Funami, K., Tatematsu, M., Seya, T. & Matsumoto, M. Endosomal
546 Localization of TLR8 Confers Distinctive Proteolytic Processing on Human Myeloid
547 Cells. *J. Immunol.* **193**, 5118–5128 (2014).
- 548 13. Hosaka, M. *et al.* Arg-X-Lys/Arg-Arg motif as a signal for precursor cleavage
549 catalyzed by furin within the constitutive secretory pathway. *J. Biol. Chem.* **266**,
550 12127–30 (1991).
- 551 14. Henrich, S. *et al.* The crystal structure of the proprotein processing proteinase furin

- 552 explains its stringent specificity. *Nat. Struct. Mol. Biol.* **10**, 520–526 (2003).
- 553 15. Braun, E. & Sauter, D. Furin-mediated protein processing in infectious diseases and
554 cancer. *Clin. Transl. Immunol.* **8**, (2019).
- 555 16. Hedstrom, L. Serine Protease Mechanism and Specificity. *Chem. Rev.* **102**, 4501–4524
556 (2002).
- 557 17. Rungrotmongkol, T. *et al.* Combined QM/MM mechanistic study of the acylation
558 process in furin complexed with the H5N1 avian influenza virus hemagglutinin's
559 cleavage site. *Proteins Struct. Funct. Bioinforma.* **76**, 62–71 (2009).
- 560 18. Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv* (2021).
561 doi:10.1101/2021.10.04.463034
- 562 19. Klamt, A. Conductor-like Screening Model for Real Solvents: A New Approach to the
563 Quantitative Calculation of Solvation Phenomena. *J. Phys. Chem.* **99**, 2224–2235
564 (1995).
- 565 20. Bím, D. *et al.* Predicting Effects of Site-Directed Mutagenesis on Enzyme Kinetics by
566 QM/MM and QM Calculations: A Case of Glutamate Carboxypeptidase II. *J. Phys.*
567 *Chem. B* **126**, 132–143 (2022).
- 568 21. Motlová, L. *et al.* Comprehensive Mechanistic View of the Hydrolysis of Oxadiazole-
569 Based Inhibitors by Histone Deacetylase 6 (HDAC6). *ACS Chem. Biol.* **18**, 1594–1610
570 (2023).
- 571 22. Magdziarz, T. *et al.* AQUA-DUCT 1.0: structural and functional analysis of
572 macromolecules from an intramolecular voids perspective. *Bioinformatics* **36**, (2020).
- 573 23. Mitusińska, K., Raczyńska, A., Wojśa, P., Bzówka, M. & Góra, A. AQUA-DUCT:
574 Analysis of Molecular Dynamics Simulations of Macromolecules with the use of
575 Molecular Probes [Article v1.0]. *Living J. Comput. Mol. Sci.* **2**, (2020).
- 576 24. Elsässer, B. & Goettig, P. Mechanisms of Proteolytic Enzymes and Their Inhibition in
577 QM/MM Studies. *Int. J. Mol. Sci.* **22**, 3232 (2021).
- 578 25. Rodríguez, A., Oliva, C., González, M., van der Kamp, M. & Mulholland, A. J.
579 Comparison of Different Quantum Mechanical/Molecular Mechanics Boundary
580 Treatments in the Reaction of the Hepatitis C Virus NS3 Protease with the NS5A/5B
581 Substrate. *J. Phys. Chem. B* **111**, 12909–12915 (2007).
- 582 26. Lima, M. C. P. & Seabra, G. M. Reaction mechanism of the dengue virus serine
583 protease: a QM/MM study. *Phys. Chem. Chem. Phys.* **18**, 30288–30296 (2016).
- 584 27. Jean, F., Boudreault, A., Basak, A., Seidah, N. G. & Lazure, C. Fluorescent Peptidyl
585 Substrates as an Aid in Studying the Substrate Specificity of Human Prohormone
586 Convertase PC1 and Human Furin and Designing a Potent Irreversible Inhibitor. *J.*
587 *Biol. Chem.* **270**, 19225–19231 (1995).

- 588 28. Krysan, D. J., Rockwell, N. C. & Fuller, R. S. Quantitative Characterization of Furin
589 Specificity. *J. Biol. Chem.* **274**, 23229–23234 (1999).
- 590 29. Basak, A., Zhong, M., Munzer, J. S., Chretien, M. & Seidah, N. G. Implication of the
591 proprotein convertases furin, PC5 and PC7 in the cleavage of surface glycoproteins of
592 Hong Kong, Ebola and respiratory syncytial viruses: a comparative analysis with
593 fluorogenic peptides. *Biochem. J.* **353**, 537 (2001).
- 594 30. Izidoro, M. A. *et al.* A study of human furin specificity using synthetic peptides
595 derived from natural substrates, and effects of potassium ions. *Arch. Biochem.*
596 *Biophys.* **487**, 105–114 (2009).
- 597 31. Evans, M. G. & Polanyi, M. Some applications of the transition state method to the
598 calculation of reaction velocities, especially in solution. *Trans. Faraday Soc.* **31**, 875
599 (1935).
- 600 32. Daggett, V., Schroeder, S. & Kollman, P. Catalytic pathway of serine proteases:
601 classical and quantum mechanical calculations. *J. Am. Chem. Soc.* **113**, 8926–8935
602 (1991).
- 603 33. Ishida, T. & Kato, S. Theoretical Perspectives on the Reaction Mechanism of Serine
604 Proteases: The Reaction Free Energy Profiles of the Acylation Process. *J. Am. Chem.*
605 *Soc.* **125**, 12035–12048 (2003).
- 606 34. Bravaya, K. *et al.* Molecular Modeling the Reaction Mechanism of Serine-Carboxyl
607 Peptidases. *J. Chem. Theory Comput.* **2**, 1168–1175 (2006).
- 608 35. Perona, J. J., Craik, C. S. & Fletterick, R. J. Locating the Catalytic Water Molecule in
609 Serine Proteases. *Science (80-.)*. **261**, 620–622 (1993).
- 610 36. Warshel, A. & Russell, S. Theoretical correlation of structure and energetics in the
611 catalytic reaction of trypsin. *J. Am. Chem. Soc.* **108**, 6569–6579 (1986).
- 612 37. Topf, M. & Richards, W. G. Theoretical Studies on the Deacylation Step of Serine
613 Protease Catalysis in the Gas Phase, in Solution, and in Elastase. *J. Am. Chem. Soc.*
614 **126**, 14631–14641 (2004).
- 615 38. Kamerlin, S. C. L., Chu, Z. T. & Warshel, A. On Catalytic Preorganization in
616 Oxyanion Holes: Highlighting the Problems with the Gas-Phase Modeling of
617 Oxyanion Holes and Illustrating the Need for Complete Enzyme Models. *J. Org.*
618 *Chem.* **75**, 6391–6401 (2010).
- 619 39. Meyer, E. Internal water molecules and H-bonding in biological macromolecules: A
620 review of structural features with functional implications. *Protein Sci.* **1**, 1543–1562
621 (1992).
- 622 40. Case, D. A. *et al.* AMBER 2018. (2018).
- 623 41. TURBOMOLE. University of Karlsruhe and Forschungszentrum Karlsruhe GmbH.

- 624 (2019).
- 625 42. Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.;
626 Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; Li, X.;
627 Caricato, M.; Marenich, A. V.; Bloino, J.; Janesko, B. G.; Gomperts, R.; Mennucci, B.;
628 Hratch, D. J. Gaussian 16, Revision C.01. (2016).
- 629 43. Case, D. A. *et al.* Amber 2022. (2022).
- 630 44. Case, D. A. *et al.* AmberTools. *J. Chem. Inf. Model.* **63**, 6183–6191 (2023).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [BzowkaSI.docx](#)