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

Maria Bzówka[1#], Karolina Mitusińska[1#], Agata Raczyńska[1], Tomasz Skalski[1], Aleksandra Samol[1], Weronika Bagrowska[1], Tomasz Magdziarz[1], Artur Góra[1*]

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

[#] These Authors contributed equally to this work

[*] Corresponding Author: a.gora@tunnelinggroup.pl

Abstract

The evolutionary variability of a protein's residues is highly dependent on protein region and protein function. Solvent-exposed residues, excluding those at interaction interfaces, are more variable than buried residues. Active site residues are considered to be conserved as they ensure an enzyme's activity and selectivity. The abovementioned rules apply also to α/β-hydrolase fold proteins - an example of enzymes with buried active sites equipped with tunnels linking the reaction site with the exterior. We hypothesised two scenarios: (1) tunnels are lined by mostly variable residues, allowing adaptation to the evolutionary pressures of a changeable environment; or (2) tunnels are lined by mostly conserved amino acids, and are equipped with a number of specific variable residues that are able to respond to evolutionary pressure. We also wanted to check if evolutionary analysis can help distinguish functional and non-functional tunnels. Soluble epoxide hydrolases (sEHs) represent a good case study for the analysis of the evolution of tunnels in an α/β-hydrolase fold family due to their size and architecture. Here, we propose methods for the comparison of tunnels detected in both crystal structures and molecular dynamics simulations, as well as the assignment of tunnel functionality, and we identify critical steps for careful tunnel inspection. We also compare the entropy values of the tunnel-lining residues and system-specific compartments in seven selected sEHs from different clades. We present three different cases of entropy distribution among tunnel-lining residues. As a result, we propose a 'perforation' model for tunnel evolution *via* the merging of internal cavities or surface perforations. We also report an approach for the identification of highly variable tunnel-lining residues as potential targets to be used for the fine-tuning of selected enzymes.

Author Summary

So far a very little is known about proteins tunnels evolution. The goal of this study is evaluation of tunnels evolution in the family of epoxide hydrolases representatives of numerous α/β-

hydrolase fold enzymes. The rigorous approach assigns potential functionality to identified tunnels providing access to the active sites of enzymes. As the results the ‘perforation’ model was proposed which 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).

As for our knowledge, tunnels in enzymes are not usually divided into functional and non-functional. Description of a tunnel's functionality – even *in silico* – is a complex process that requires lots of computational efforts, such as molecular dynamics simulations combined with tracking of small molecules through the tunnels. Therefore, in this study, we were looking for a simple approach that could be used to determine a tunnel's functionality based on crystal structures and evolutionary analysis. We hypothesised that functional 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 questions: i) are tunnels conserved due to their conformational constraints (since they connect a conserved

active site with the environment, and they run through a conserved core) and ii) are we able to distinguish functional and non-functional tunnels based on the evolutionary analysis. We need to emphasize 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, dimerisation or quaternary protein structure organisation.

Verification of our hypothesis requires a precise definition of functional tunnels, and the identification of tunnel-lining residues. Despite the existence of several tools for tunnel detection in proteins(35), it is a non-trivial task to assign the significance of the detected tunnels to their functionality(36). Enzymes can utilize one or more tunnels connecting the buried active site with the bulk solvent or with different catalytic sites(28); therefore, tunnel identification in some structures might be challenging. Moreover, assigning a tunnel's functionality requires knowledge of its shape and size, and of the physico-chemical properties of the tunnel-lining residues. In this paper, we combined tunnel identification with assessment of functionality by comparing the results obtained from geometry-based tunnel detection software(37) with those coming from a small-molecule-tracking approach implemented in AQUA-DUCT software(38). AQUA-DUCT uses small molecules as molecular probes to identify the most commonly used pathways for solvent transportation, thus assessing a tunnel's functionality. We defined the functional tunnels as the ones used for water molecules transport, while in the non-functional tunnels no water molecules pathways were identified. Such analysis sheds light on the function-evolution relationship of tunnels. Finally, we conducted an in-depth analysis of three different tunnels in order to demonstrate the applicability of our study.

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(39–41). We chose soluble

epoxide hydrolase due to the fact that crystal structures of different clade members (mammals, plants, fungi and bacteria) were available. Moreover, sEHs catalyse 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. In this paper, we focus only on the tunnel-lining residues and the tunnels themselves, present in the experimentally solved protein structure of soluble epoxide hydrolases.

Results

For this study, we chose only the unique and complete structures of soluble epoxide hydrolases deposited in the Protein Data Bank (PDB)(42). 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

Epoxide hydrolases 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(43). The active site of the epoxide hydrolases is buried inside the main domain, and therefore the

transportation of substrates and products is facilitated by tunnel (either singly 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(44). We analysed the compartments listed in **Supplementary Table S1** (i.e. residues forming the active site; buried and exposed residues; main and cap domains; NC-loop; cap-loop; and α-helices, loops and β-strands). In order to determine the compartments' variability, we calculated the median distance as a difference between medians of Schneider entropy values(45) of selected compartments and the remaining positions of the multiple sequence alignment (MSA) (**Figure 1**). Negative values indicate compartments with lower variability, and positive values indicate compartments with higher variability in comparison to the remaining positions in the MSA. For quantitative statistical analysis, we compared the calculated Schneider entropy values of these compartments with the remaining positions of the MSA using the Epps-Singleton test(46). 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.

Based on these median distances, the active site residues were classified as conserved, i.e. with lower entropy scores (the active site was the only compartment with negative values of median distance). The exposed residues (classified as surface residues according to the NetSurfP server(47)) were classed as the most variable. The median distance values for these exposed residues were found to be the largest among all structures analysed in our study, i.e. these should be considered variable residues. Entropy analysis showed that the variability of the protein core was substantially lower than the variability of the surface residues (**Figure 1, Supplementary Table 2**). These results are in agreement with the general findings mentioned previously(3,16,48). With regards to the structural elements specific to epoxide hydrolases, all compartments (cap

domain, main domain, cap-loop, and NC-loop) were classified as variable among all the selected EHs. In all analysed proteins, the secondary structures (α-helices, loops and β-strands) were also found to be variable, with α-helices being the most and β-strands the least variable (**Figure 1, Supplementary Table S2**).

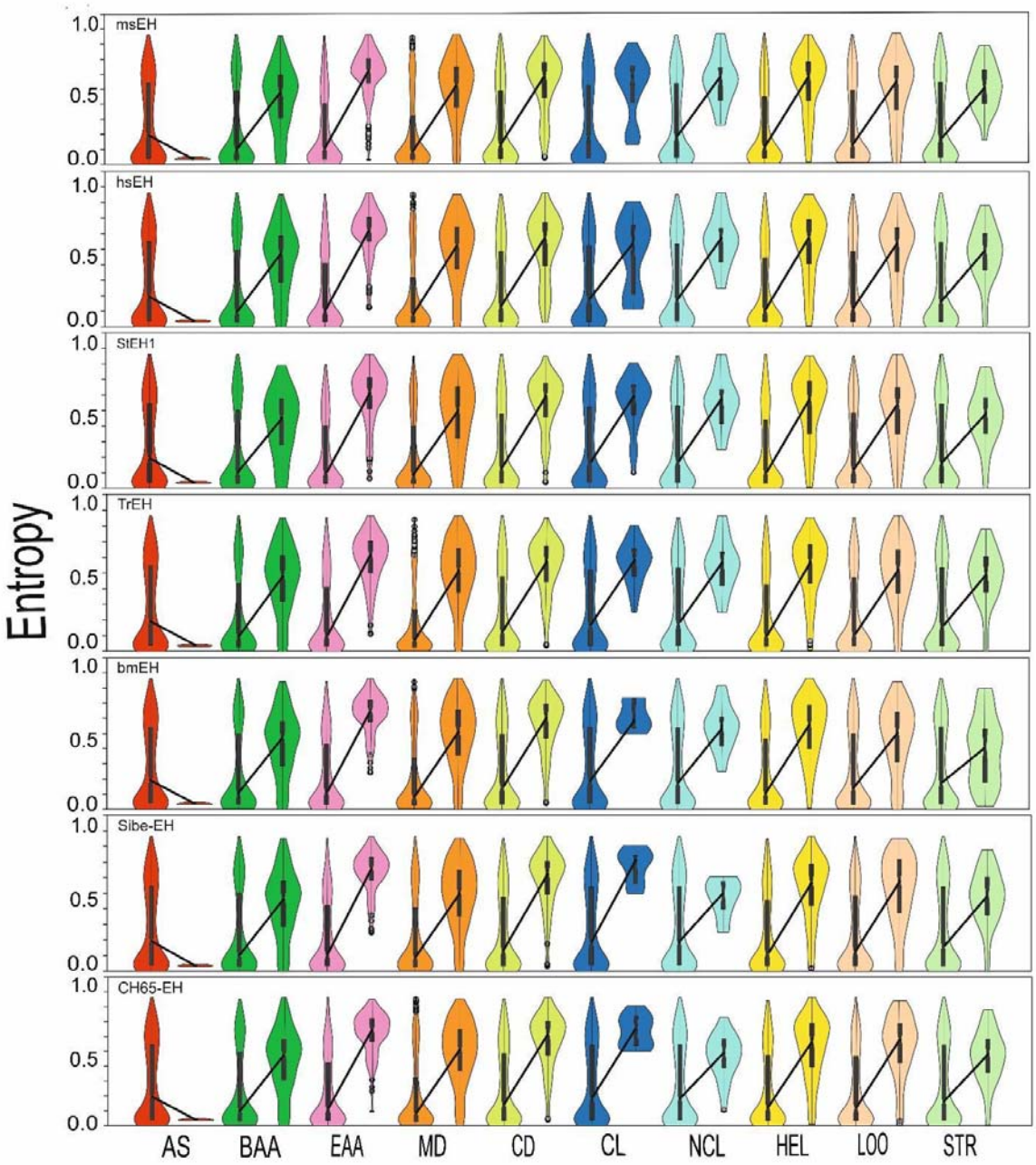

Figure 1. Distribution of the entropy values and the median entropy values of referential compartments, and the remaining positions of the MSA (violin and box plot with notches and outliers). Figure represents data shown in **Supplementary Tables S2**. All pairwise differences

are statistically significant. Abbreviations: AS – active site; BAA – buried amino acids; EAA – exposed amino acids; MD – main domain; CD – cap domain; CL – cap-loop; NCL – NC-loop; HEL – helices; LOO – loops; STR – strands.

Tunnel identification and comparison

As already mentioned, tunnels can maintain different functions. In the case of sEHs, the most important tunnels' functions are: 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. Therefore, it is essential to distinguish such tunnels from non-functional ones (e.g. tunnel-like structures contributing only to protein stability or resulting from the local protein conformational changes). For the identification of functional and non-functional tunnels, we used a two-step approach (see the **Methods** section). First, we identified tunnels providing access to the active site using a geometry-based approach implemented in CAVER software for both crystal structures and in molecular dynamics (MD) simulations, and then compared their geometries (for details see the **Methods** section). Then, we used AQUA-DUCT to determine their functionality, based on the identified water molecules pathways. We defined functional tunnels as those which are capable of transporting water molecules, while in the non-functional tunnels no water molecules pathways were identified. This approach was called the 'top-down' approach. In the second step, the 'bottom-up' approach we used all tunnels identified by AQUA-DUCT as functional and searched for those which could not be combined with tunnels identified by CAVER neither in crystal structures nor in MD simulations. Since part of the tunnels can be seen as short-lasting cavities which rarely connect with other internal voids, they are difficult to identify using the CAVER geometry-based approach. The selection of water molecules as a molecular probe for tunnel detection is supported by comparison of the MD simulations with

NMR study in the work of Persson and Hale in which they revealed that water molecules gain access even to the internal sites by a transient water transport mechanism(49).

In the first step, the 'top-down' approach, 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 (**Supplementary Table S3**), 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. When we compared tunnels detected by CAVER during MD simulations, with the actual pathways identified for water molecules by AQUA-DUCT, most tunnels were assigned as functional **(Figure 2, Supplementary Table S3**). We suggest that the non-functional tunnels might not be utilised by water molecules due to their length (too long tunnels), or their physico-chemical properties. 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).

Using the 'top-down' approach, we identified seven tunnels in the main domain, six in the cap domain, and three at the border between those domains. Most functional tunnels were found in the main domain, namely Tm1, Tm2, Tg, and Tside (**Figure 2;** all functional tunnels identified during the 'top-down' approach are marked by a black star), and in the cap domain, namely Tcap3, and Tcap4. On the border between the cap and main domains, the Tc/m tunnel (**Figure 2**) was found. It should be pointed out that the Tc/m tunnel was identified as multiple tunnels by CAVER (Tc/m1, Tc/m2, and Tc/m3), whereas AQUA-DUCT detected only one entry/exit in this region. This issue is related to the asymmetric shape of the Tc/m tunnel, which makes it difficult to classify in a geometry-based approach (**Supplementary Figure S1**). We identified four non-

functional tunnels in the cap domain (Tcap2a, Tcap5, Tcap6, and Tcap7), two in the main domain, (Tm5 and Tback), and one in the border between the cap and main domains, (Tc/m_side) (**Figure 2**). Among the identified tunnels, four tunnels were found as functional in one or more structures and non-functional in other(s). For example, Tcap4 was identified as functional in msEH, TrEH, and Sibe-EH, and as non-functional in hsEH (**Supplementary Table S4**).

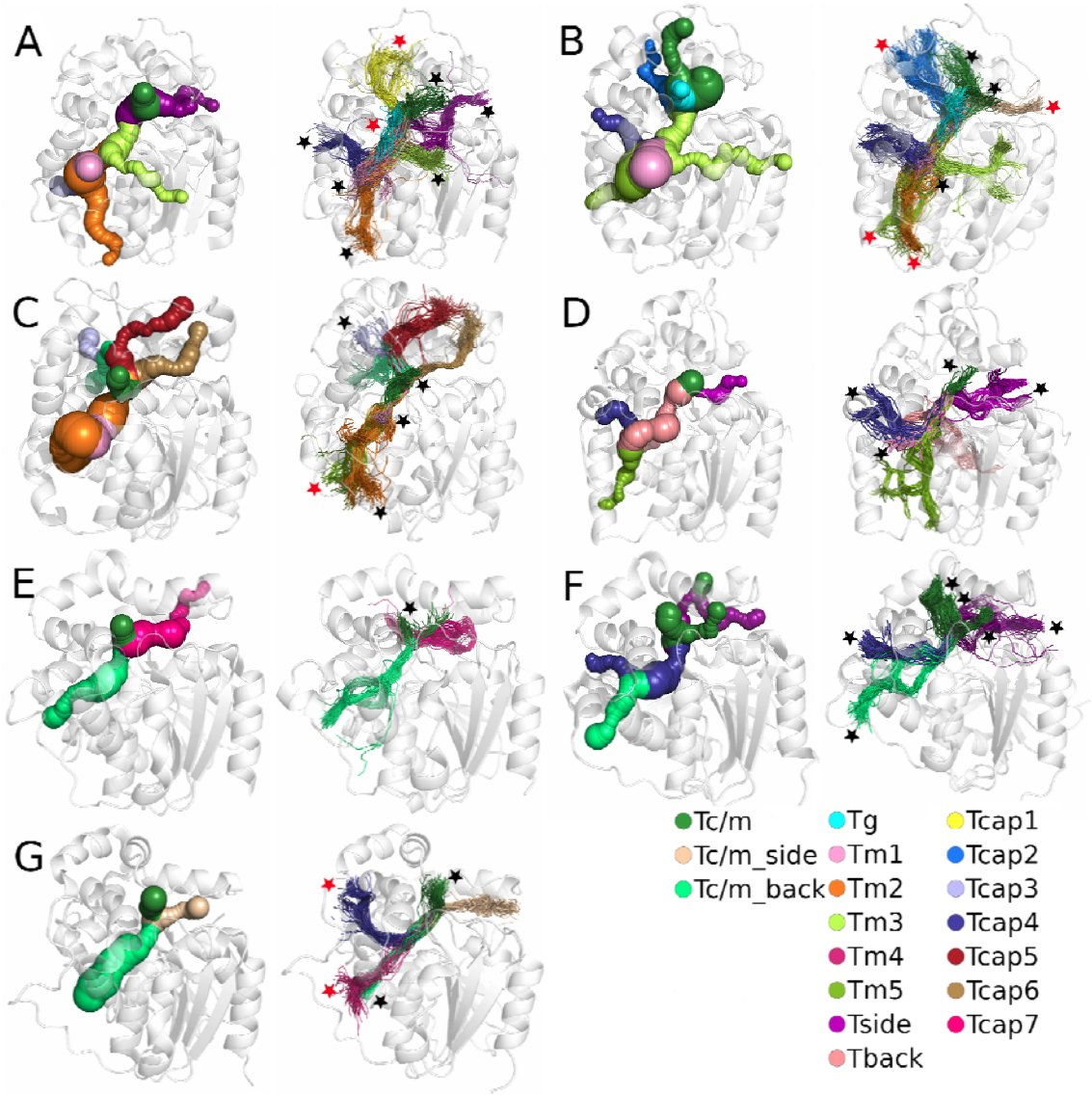

Figure 2. Comparison of tunnels identified in the crystal structure (left) and the results after molecular dynamics (MD) simulation (right) for each system: A) *M. musculus* soluble epoxide hydrolase (msEH), B) *H. sapiens* soluble epoxide hydrolase (hsEH), C) *S. tuberosum* soluble epoxide hydrolase (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). Functional tunnels identified during the 'top-down' approach are marked by a black star (*) and those identified during the 'bottom-up' approach are marked by a red star (*). Note that the Tm5 tunnel of the StEH1 enzyme was identified during the 'bottom-up' approach even though it seems like it has its counterpart in the crystal structure.

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. Indeed, the results of AQUA-DUCT analysis showed that, in some cases, the area of entry/exit for water molecules can correspond to more than one tunnel identified by CAVER during MD simulations (**Supplementary Figure S1**).

We also examined tunnel rankings from both the crystal structures and the MD simulations, but we did not find any correlation between the two. Tunnels identified by CAVER were ranked according to their priority, i.e. 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. 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. For example, tunnels identified for the TrEH structure in both

the crystal structure and MD simulation present almost the same ranking, while for other analysed structures the rankings differ substantially. Most tunnels identified in the hsEH crystal structure were identified as tunnels with lower priority, i.e. they had low ranking (below 8th position). However, when we compared the maximal radii of bottlenecks, i.e. the radii of the widest regions of analysed tunnels, we found a correlation between the corresponding tunnels (r=0.85) (**Supplementary Table S5, Supplementary Figure S2**). 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.

In the second step – the 'bottom-up' approach – we compared all tunnels identified by AQUA-DUCT with those identified by CAVER in both the crystal structures and in MD simulations (**Supplementary Table S6**). While in the 'top-down' approach we focused more on the tunnels identified by CAVER and assigned their functionality using AQUA-DUCT, in the 'bottom-up' approach we inverted the workflow. We searched for tunnels with assigned functionality which were not detected by CAVER. We suggest that functional tunnels might not be identified by CAVER in two situations: i) when a tunnel is very narrow and its bottleneck radius value is below the CAVER detection radius, or ii) when a tunnel comprises at least two joined cavities whose opening is not synchronised. We call those situations 'rare events' or 'tunnel leakage' and we previously discussed that situation in the case of StEH1(50). Such situations were observed several times in the analysed sEHs; we identified two additional tunnels located in the cap domain (Tcap1 and Tcap2b), one tunnel in the main domain (Tm4), and one tunnel located on the border between the cap and main domain, namely Tc/m_side (**Figure 2**; all additional tunnels identified during the 'bottom-up' approach are marked by a red star). In one case – the Tm5 of the StEH1 enzyme – the tunnel was identified in this step, even though it seemed that it had its corresponding counterpart in the crystal structure. However, due to its branching, it was difficult

to match this tunnel with its counterpart during the 'top-down' approach. Only in three structures (TrEH, bmEH, and Sibe-EH) no additional functional tunnels were found.

For all functional tunnels except one (Tc/m_side tunnel of msEH), we were able to combine the identified entries/exits with corresponding tunnels identified during MD simulation. None of those tunnels were identified in crystal structures, which might suggest that the protein structure requires a potential conformational change to open such a tunnel/pathway, or that some functional tunnels have transient character. For the Tc/m_side tunnel of msEH, we were unable to combine its entrance/exit with a previously identified tunnel (in both crystal structure and during MD simulation). This suggests that some regions of a protein structure are more prone to have a leakage than others. We need to emphasise that due to the mechanism of the reaction catalysed by sEH even tunnels opened temporarily for water molecules transport can contribute to the enzyme's functionality.

Using both approaches, in total, we identified eight tunnels located in the cap domain, three located on the border between the cap and main domains, and eight tunnels located in the main domain (**Supplementary Table S4 and Supplementary Table S6**). We identified nine unique functional tunnels in msEH, eight tunnels in hsEH, six in Sibe-EH, five in StEH1, four in TrEH and CH65-EH, and one in the bmEH structure. We also identified non-functional tunnels in all analysed proteins except the msEH and Sibe-EH structures. We found three non-functional tunnels in hsEH and StEH1, two in TrEH and in bmEH, and one in CH65-EH. The Tc/m tunnel was identified in all analysed structures as a functional tunnel. The second most frequently identified functional tunnel, the Tcap4 tunnel, was identified as functional in four structures, and in one structure it was identified as non-functional. Tunnels Tcap1, Tcap2b, Tcap3, and Tm4 were unique and identified as functional tunnels, while Tcap2a, Tcap5, Tcap6, Tcap7, and Tback were identified as non-functional. The list of structures in which the tunnels were identified is

shown in **Supplementary Table S4**. For the investigation of tunnel evolution, we decided to analyse the tunnel-lining residues from the MD simulations (**Supplementary Tables S7–S13**).

Tunnel evolutionary analysis

As we confirmed during the referential compartments' evolutionary analysis, exposed residues are more variable than buried residues. Indeed, **Figure 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 added newly identified tunnels, coloured according to their frequency of detection 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 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).

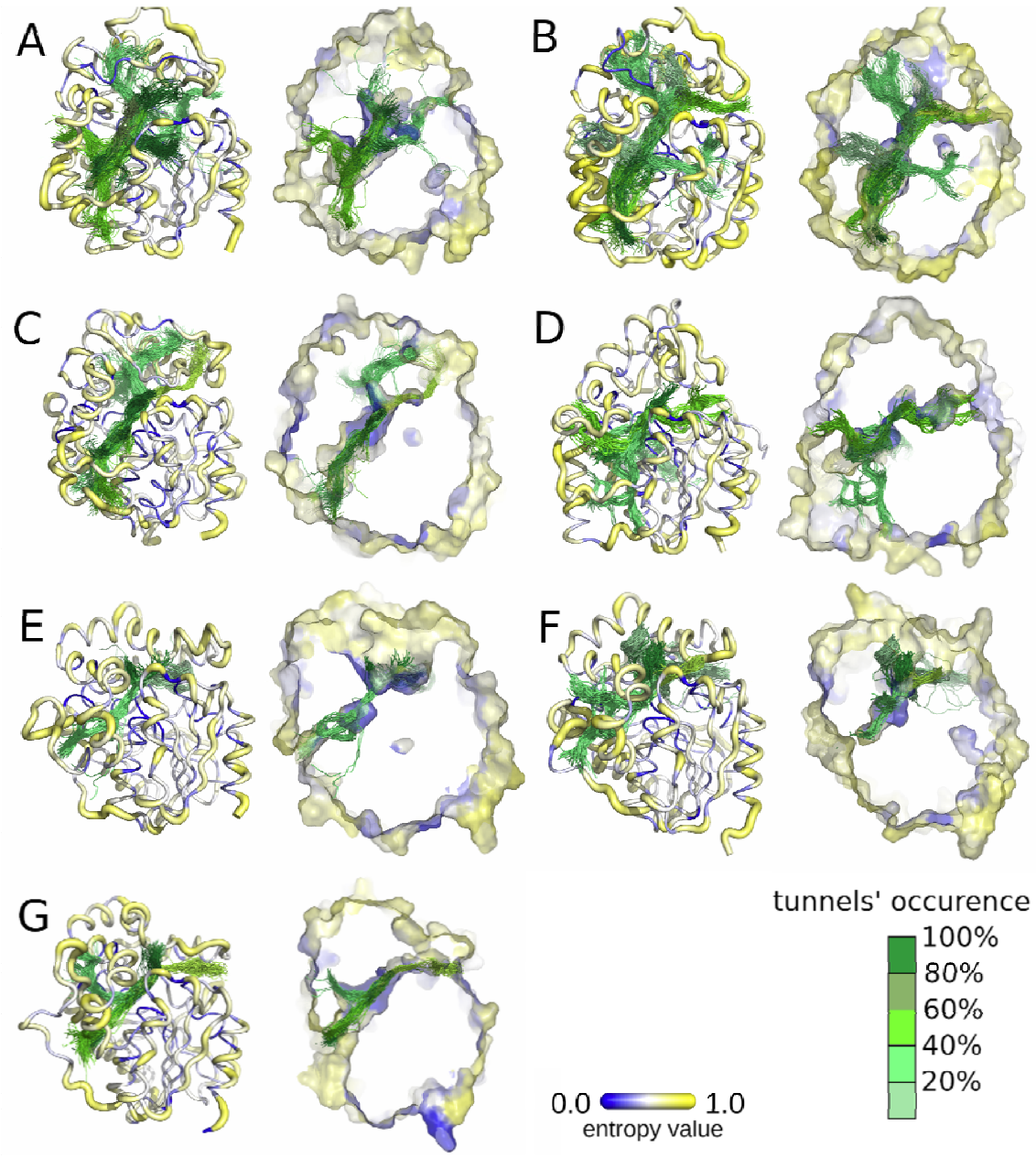

Figure 3. Visualisation of the entropy score of each protein residue (right), and frequency of tunnels identified with CAVER during molecular dynamics (MD) simulations (left) for each system: A) *M. musculus* soluble epoxide hydrolase (msEH), B) *H. sapiens* soluble epoxide hydrolase (hsEH), C) *S. tuberosum* 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 frequency of their occurrence during MD simulations: 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.

We identified the residues lining these particular tunnels during the MD simulations. However, during MD simulations, the protein is not a rigid body and the residues gain some level of 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 exposed residues. Therefore, we decided to perform an entropy analysis of: i) all tunnel-lining residues; ii) exposed tunnels-lining residues; and iii) tunnel-lining residues without the exposed residues. An evolutionary analysis of the tunnel-lining residues without the exposed residues is presented in **Figure 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 **Supplementary Tables S7–S13**. A detailed analysis of the epoxide hydrolase tunnels is shown in **Supplementary Figures S3–S9**. Median distances of all analysed proteins are listed in **Supplementary Table S14**.

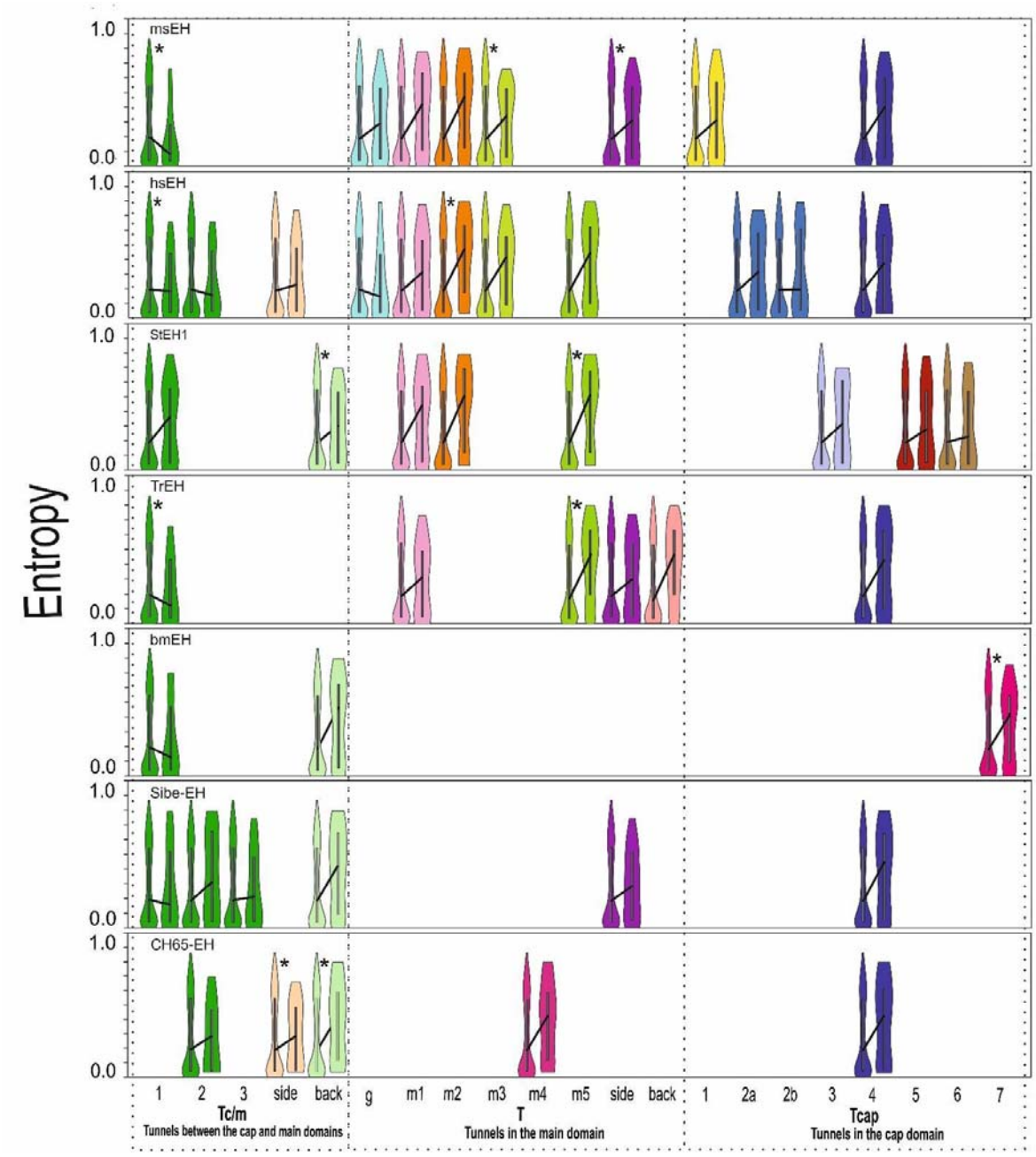

Figure 4. Distribution of the entropy values and median entropy values of tunnel-lining residues without the exposed residues, and the remaining positions of the MSA (violin plots), for all analysed soluble epoxide hydrolase structures. Figure represents the data shown in **Supplementary Table S14**. Statistically significant pairwise differences in median distances are marked by a star (*).

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 variable. Following exclusion of the exposed residues from the tunnel-lining residues, median values decreased. Importantly, median distances for the Tc/m tunnel in the case of msEH, hsEH, and TrEH proteins had negative values, suggesting this tunnel's evolutionary preserved structure (**Supplementary Table S14**). It is clear that the exposed tunnel-lining residues generally reach higher entropy values than the other analysed tunnel-lining residues (**Supplementary Figures S3–S9**). After comparing entropy values using the Epps-Singleton statistical test, it appears that exposed tunnel-lining residues are significantly different from the remaining positions in the MSA, and can be characterised as variable. However, there were no differences between tunnel mouths and the exposed residues.

Detailed analysis of selected tunnels

As shown above, only one tunnel in soluble epoxide hydrolases was found in all analysed crystal structures and during MD simulations (the functional tunnel Tc/m). Interestingly, this is also the only tunnel which, in some of the structures, can be considered as conserved. Vast majority of tunnels should be considered as variable (both functional and non-functional tunnels). This finding stays in disagreement with our hypothesis. At first glance it may suggest that the identification of the evolutionarily preserved and crucial tunnels in soluble epoxide hydrolases can be performed using just crystal structures. However, the problem presented in our study is more complex. First of all, the analyses of the tunnels ranking and their geometrical parameters show that even well-defined α/β-hydrolase fold proteins undergo significant conformational changes which suggest that the shape and size of the main tunnel can substantially differ relative with the crystal structure. Secondly, the mechanism of the reaction catalysed by epoxide hydrolases requires precise control of water flow(51), which can be achieved also *via* narrow

tunnels controlling the solvent access and precisely delivering the catalytic water molecule. Separation of the substrate and/or product transport and the water molecules delivery is of great importance considering the predominantly used tunnel(s) which need to encompass long-chain hydrophobic substrates. Finally, a few excellent examples of the protein engineering (presented in the **Discussion** section) has proved the importance of the additional tunnels for the catalysis.

To gain a deeper understanding of the function-evolution relationship, we focused on 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 functional tunnels, and the Tc/m_back tunnel of bmEH represents a tunnel which was identified as functional or non-functional tunnel depending on the examined sEH. The entropy values of the tunnel-lining residues are presented in **Supplementary Table S15**.

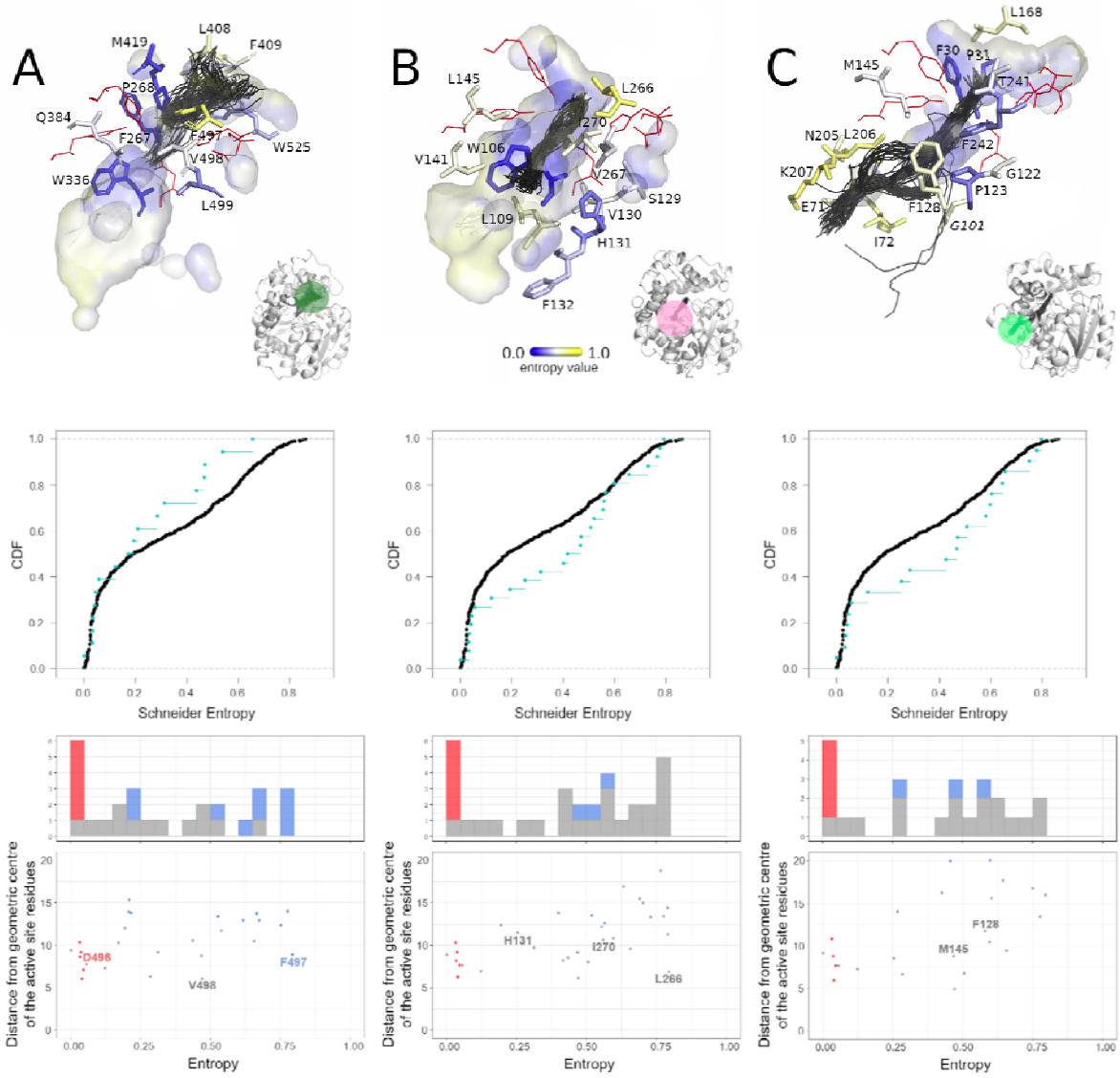

Figure 5. Analysis of the selected tunnels of soluble epoxide hydrolases (sEHs). A) The functional Tc/m tunnel of the *H. sapiens* soluble epoxide hydrolase (hsEH) structure, B) the functional Tc/m tunnel of the S. *tuberosum* soluble epoxide hydrolase (StEH1) structure, and C) the non-functional Tc/m_back 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 exposed 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; exposed residues – blue; buried – grey).

The Tc/m tunnel of hsEH whose mouth is located between the main and cap domains was identified as a functional tunnel in all analysed epoxide hydrolases, and in some cases should be considered as a conserved structural feature. This tunnel (**Figure 5, panel A**) 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. Close inspection of the tunnel revealed a variable residue (i.e. with higher entropy value) – F497 (Schneider entropy value 0.7946) – located 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). We suspect that the F497 residue might act as a molecular gate, since its position in several other crystal structures differs substantially (**Supplementary Figure S10**). The remaining variable residues are located mostly close to the surface, or on the interface of the cap domain.

The Tm1 tunnel of StEH1 is the shortest identified in this structure (**Figure 5, panel B**). Similar tunnels were identified in three other analysed epoxide hydrolases: 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 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 (Mitusińska et al. 2018), we identified three residues, namely P188 (Schneider entropy value 0.5117) (not shown at Figure 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.

Most tunnels running through the main domain were identified as functional tunnels. However, here, we would like to present a non-functional Tc/m_back tunnel for bmEH (**Figure 5, panel C**). The third tunnel identified during MD simulation 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. A 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(52). Additionally, when we ran more repetitions of the MD simulations of the bmEH structure, we identified a leakage of one or more water molecules through this tunnel (data not shown).

Perforation hypothesis

Scatterplots (similar to those at **Figure 5, panel A-C**) can be used to draw further conclusions regarding the distribution of entropy values of tunnel-lining residues. They can also be used to

identify the most variable and conserved residues. In general, the analysed examples (**Figure 5**) show three cases of entropy distribution among tunnel-lining residues (excluding the active site residues and exposed residues): i) the flat distribution of the entropy values (**Figure 5, panel A**); ii) the overrepresentation of residues with higher entropy values (**Figure 5, panel B**); and iii) the quasi-sigmoidal distribution (**Figure 5, panel C**; most of the residues have values of the entropy in the range of 0.4–0.8).

We have shown that most tunnels in epoxide hydrolases are variable structural features. The only exception (in few sEHs) is the mostly conserved Tc/m tunnel which runs between the main and cap domains, which was also present in all analysed structures. The other tunnels are distributed throughout the members of all clades. Since all sEHs share a conserved active site, we can speculate that the variability of the tunnel network allows the enzyme to adapt and catalyse different substrates. We have also presented two cases (**Figure 5, panel B and C** of variable tunnels, which were lined by specific, mostly variable residues, contributing significantly to tunnel functionality.

Based on the fact that entropy values of tunnel-lining residues usually range from 0.4 to 0.8 (**Supplementary Table S15**), we hypothesized that new tunnels providing access to the protein interior can appear as a result of a mutation occurring in the surface layer, of protein or at the border of large cavities affecting surface integrity (**Figure 6**). We showed 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.

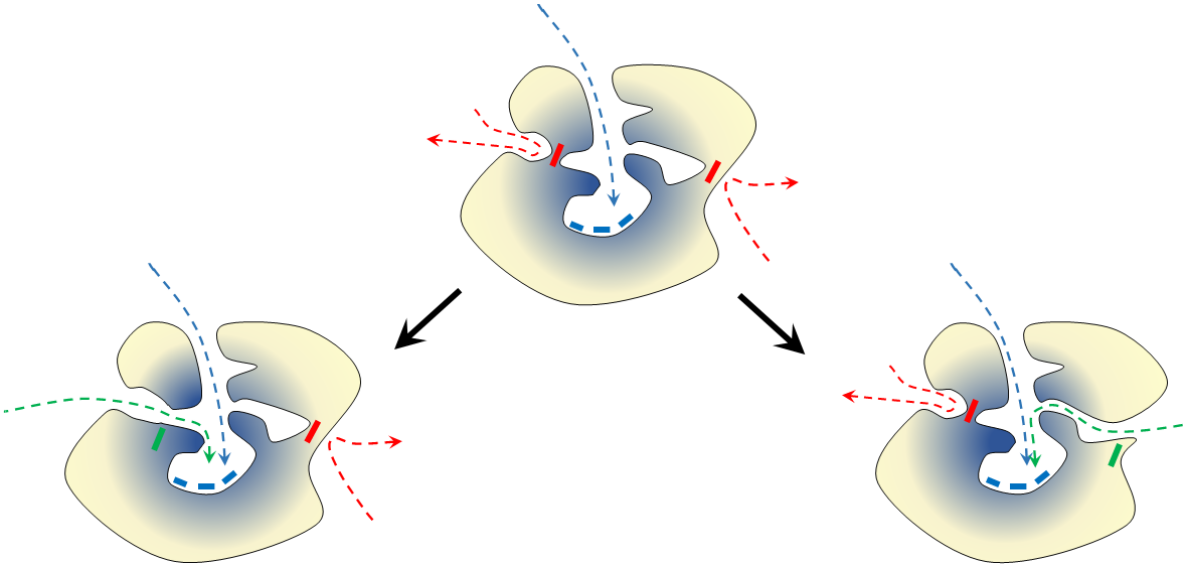

Figure 6. Schematic representation of the 'perforation' model of protein tunnel evolution. The ancestor protein (middle) and two modification pathways leading to a new enzyme by merging internal cavities (left) or by surface perforation (right). Yellow – variable residues; blue – conserved residues. 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.

Discussion

In our study, we focused on epoxide hydrolases, 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(48), which seems to be also the oldest(53) and most stable(54) fold used by one the largest groups of enzymes(55). Structural and evolutionary analyses of epoxide hydrolases have been reported systematically(43,56–58), 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 epoxide hydrolase 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(43). However, our analysis suggests that their analysis might be incomplete.

Here, we analysed 1455 epoxide hydrolase sequences and seven crystal structures representing different clades (animals, plants, fungi and bacteria). By calculating a median distance – defined as a difference between medians of Schneider entropy values of a selected compartment and the remaining positions of the 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 exposed residues than by buried residues(3,16,48). However, in contrast to the work of Barth *et al.*(43), we showed that the cap-loop and NC-loop should be considered as variable features. In all analysed proteins, the secondary structures (α-helices, loops, and β-strands) were found to be variable, with α-helices being the most and β-strands the least variable (**Supplementary Table 1**), which was also shown previously for other systems elsewhere(18). Further, since we were able to identify structural compartments of the seven epoxide hydrolases analysed, observations regarding the modularity of EHs is still applicable(41). Subsequently, we decided to focus our analysis on the tunnels connecting the conserved active site with the environment.

The identification of tunnels in the crystal structures of proteins has been performed computationally since the beginning of the 21st century(35), however the assessment of their functionality still remains a non-trivial task. This capability was soon extended with the introduction of MD simulations(37,59). However, the vast majority of previously investigated

tunnels have been described on the basis of solely crystal structures, and the consequent inaccuracies of such a simplification are mostly unknown. The knowledge and time required to perform additional MD simulations is probably the main barrier to wider usage of more advanced techniques. Therefore, we compared the tunnels identified in both crystal structures and during MD simulations, using an automated approach to identify corresponding tunnels. Our results suggest that most tunnels identified during MD simulations have 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 (**Supplementary Figure S2**, **Supplementary Table S5**). Differences may be attributable to packing inaccuracies or poor resolutions of crystal structures(60–63). 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 (**Supplementary Figure S11**). Crystal structures with poor resolutions are prone to be inaccurate, especially within the most flexible regions, such as loops(62,64) or gating residues within tunnels(28,65). We compared tunnels identified in both crystal structures and during MD simulations using the geometry-based approach implemented in CAVER. In most cases the number of tunnels identified during MD simulations was higher relative to the crystal structures. This suggests that during MD simulation proteins undergo conformational changes leading to opening additional tunnels, which also could be utilised by water molecules. Therefore, in some cases a simple analysis of crystal structure leads to an incomplete picture of the enzyme's tunnel network. Moreover, we compared the tunnels ranking provided by CAVER for the tunnels identified in crystal structures and during MD simulations and in some cases revealed substantial differences. For example, tunnels identified for the TrEH structure in both the crystal structure and MD simulation present almost the same ranking, while most tunnels identified in the hsEH crystal structure were identified as tunnels with lower positions. Therefore, we suggest that one should consider using MD simulations instead of single-structure analyses. The high correlation

between measured bottlenecks may suggest that the shape and size of tunnels present in a crystal structure are preserved despite the potential conformational changes which may affect overall protein structure. However, longer or more advanced molecular dynamics simulations might be required to sample alternative protein conformations to identify tunnels which are very rarely utilised. Once the access of the solvent or small molecules is essential for the catalysis, the molecular tracking approach applied for MD simulations study can bring additional benefits providing a holistic picture of tunnel network. It is worth noting that we compared tunnels with relatively narrow bottlenecks, mostly of 1.0–2.0 Å radii. While conformational changes in amino acids side chains may cause opening and closing of narrow tunnels, their effect can be ignored in cases of wider tunnels, channels, or pores.

A detailed analysis of the evolution of tunnel-lining residues from soluble epoxide hydrolases has brought several unexpected discoveries, including insights into the utility of tunnel-detecting software, as well as into tunnel origin and evolution. Important conclusions arose from our analysis of the tunnel functionality. Using a geometry-based approach, we found several tunnels observable in both crystal structures and MD simulations but whose functionality was not confirmed with the use of a water-tracking approach i.e. they were unable to transport even a single water molecule during the course of the whole MD simulation. This observation may be a consequence of neglecting the physico-chemical properties of tunnel-lining residues during tunnel detection. Further, when we used a water-tracking approach to determine tunnel functionality, we observed that, in some cases, the area of entry/exit for water molecules may correspond to more than one tunnel identified during MD simulations by geometry-based tools; this was mostly due to tunnel asymmetry (**Supplementary Figure S1**). This issue could be solved by modification of the clustering threshold parameter in CAVER, however it often caused joining of other clearly separated tunnels (data not shown). We found that particular tunnels which can be identified by analysing water pathways were omitted by the geometry-based approach, even when applied to

an analysis of the same MD simulation. Such a situation could easily be explained by the transient character of narrow tunnels, which may be considered a sequence of small cavities rather than permanently open tunnels. During our analysis of small molecule transport, such tunnels can be easily identified, whereas geometry-based approaches are unable to identify those pathways. Taking into consideration the mechanism of catalysed reaction, the role of such tunnels cannot be neglected. The described behaviour of small cavities merging into a tunnel supports the observations presented in a novel study on *Mycobacterium tuberculosis* dUTPase, which revealed that the substrate-binding requires small loop movements, which only partially uncover the active site(66). Therefore, findings obtained with a geometry-based tunnel identification approach should be carefully interpreted, even in cases where such tunnels were identified during MD simulations. A detailed analysis is essential during the protein reengineering process. It can provide information about the residues blocking a potential new pathway, whose opening can promote significant modification of protein activity(52,67,68) or selectivity(23,29,69,70).

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 soluble epoxide hydrolases. Therefore, our results can be applied mostly to the epoxide hydrolases, and – with some minor adjustments – to other members of the α/β-hydrolases. 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 exposed 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. We identified only one tunnel which in some structures is in line with our hypothesis (i.e. was identified as a conserved feature), namely the Tc/m tunnel. It was found in the all analysed structures, and, in

some cases, should be considered a conserved feature. In contrast, the majority of the identified tunnels were in disagreement with the hypothesis, and should be considered as relatively variable structural features of soluble epoxide hydrolases.

Given the fact that in most cases more than one tunnel was identified as functional, it is clear that even though the enzymes from the α/β-hydrolases superfamily share similar structures, they can catalyse different substrates and present different stereo- and regioselectivity due to the differences in their tunnel network(71). It should be pointed out that in different members of the α/β-hydrolase superfamily different tunnels might be crucial for maintaining the reagents transport, and thus different tunnels might be preserved. The only tunnel which was found in all analysed structures 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(48). Both domains interact, creating a hydrogen bond network(72); 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(73). 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(43,74–77). Moreover, it is thought that the capability to switch their conformation may significantly expand accessibility of the active sites to larger substrates(30,77). Our results suggest that access to the active site of the ancestor of the epoxide hydrolase family (and probably of the α/β hydrolases members) was achieved *via* the tunnel between the cap and main domains, and that this feature has been preserved in most of the family members in all clades. Indeed, analysis of the α/β hydrolase evolution performed by Barth *et al.* showed that the position of cap domain is conserved in all GX, GGGX and Y-types of α/β hydrolases(41,43).

As we pointed above, the insertion of the cap domain has created the buried active site cavity and the Tc/m tunnel ensuring access to that cavity. 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 epoxide hydrolases involves proton shuttling (deprotonation of the nucleophile in the hydrolysis step) and water attack, it requires precise transportation of water molecules. New tunnels could significantly improve the enzyme's performance by separating the substrate/products transport pathways from water delivery tracks. Additionally, since the active site is preserved in all analysed structures, the shape and length of the tunnels can be seen as a main factor of significant modification of the substrate specificity among different clades. Perhaps we were unable to find any correlation between tunnel evolution and tunnel functionality due to that issue. Comparing tunnel networks between different organisms we observed a dramatic increase in the number of tunnels between representatives of bacterial and mammalian soluble epoxide hydrolases and differences in the tunnel network pattern in members of each clade. This might suggest that in some cases the substrate specificity profile has changed and that those enzymes may utilise different tunnels for the transport of substrate and/or product and other tunnels deliver water molecules.

Based on the results of our analysis we proposed a perforation model describing mechanism of new tunnels evolution. The transformation of the non-functional Tc/m_back tunnel of bmEH into functional one by Kong *et al.* (52) can be seen as a perfect example of the mimicking of the proposed surface perforation mechanism in the tunnel engineering process in order to improve the enzyme's functionality. We showed that 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 higher entropy values than their neighbours. Therefore, we hypothesised that mutations of such variable residues could 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.* (78), 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(79).

We confirmed that the surface residues are more variable than the buried residues and that there is no statistically significant difference between the variability of surface residues and the residues lining tunnels' mouths. This may seem counterintuitive at a first glance, however it might be also seen as a consequence of surface residues being more variable than those comprising the rest of the protein and can suggest that tunnels' mouths do not take part in the process of the substrate recognition in the case of soluble epoxide hydrolases. Indeed, all analysed soluble epoxide hydrolases exhibit a broad substrate range and in their tunnels might maintain a control of the reaction environment and/or proper substrates positioning.

We were also able to verify that part of our hypothesis in which we claim that even conserved tunnels can be lined with more variable residues. 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 different conformation in crystal structures (**Supplementary Figure S11**). This might indicate a potential role for F497 as a gate, controlling access through this tunnel. 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)(80,81).

Entropy analysis of particular amino acid positions has become a widely applied method in the search for feasible key amino acids (hot-spots). 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(82). 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(83). 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 this paper, we detailed three cases of entropy distribution among the tunnel-lining residues (excluding active site residues and exposed residues): i) the overrepresentation of variable residues with higher entropy values; ii) the flat distribution of entropy values along these tunnels; and iii) the quasi-sigmoidal distribution of the entropy values themselves (most are in the range of 0.45–0.80). Moreover, our results suggest that, based on the entropy values of particular tunnel-lining residues, we are unable to distinguish tunnel functionality. Other approaches, such as using water molecules as a specific probe to test tunnel permeability, may be effective(38). The tunnels described in our findings which we consider conserved have provided substantial information about their origin, and about the evolution of enzymes' families more broadly. On the other hand, our results suggest that the 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 of 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 a MSA comprising sequences of epoxide hydrolases. The identified residues were then used as input for an evolutionary analysis using the BALCONY software. The tunnel identification stage was divided into two steps, in order to include all functional tunnels. During the 'top-down' approach, tunnels were identified in both crystal structures and during MD simulations, and then compared with each other to find their corresponding counterparts. Their functionality was then assessed based on a small-molecule-tracking approach, i.e. overlaying the paths of water molecules and their entries/exits, and movements through the system on top of the positions of identified tunnels. In the second step – the 'bottom-up' approach – the same water molecules pathways and entries/exits were matched with previously identified tunnels. Finally, the tunnel-lining residues, the exposed tunnel-lining residues, and the tunnel-lining residues without exposed residues were used for the evolutionary analysis using BALCONY software (**Figure 7**).

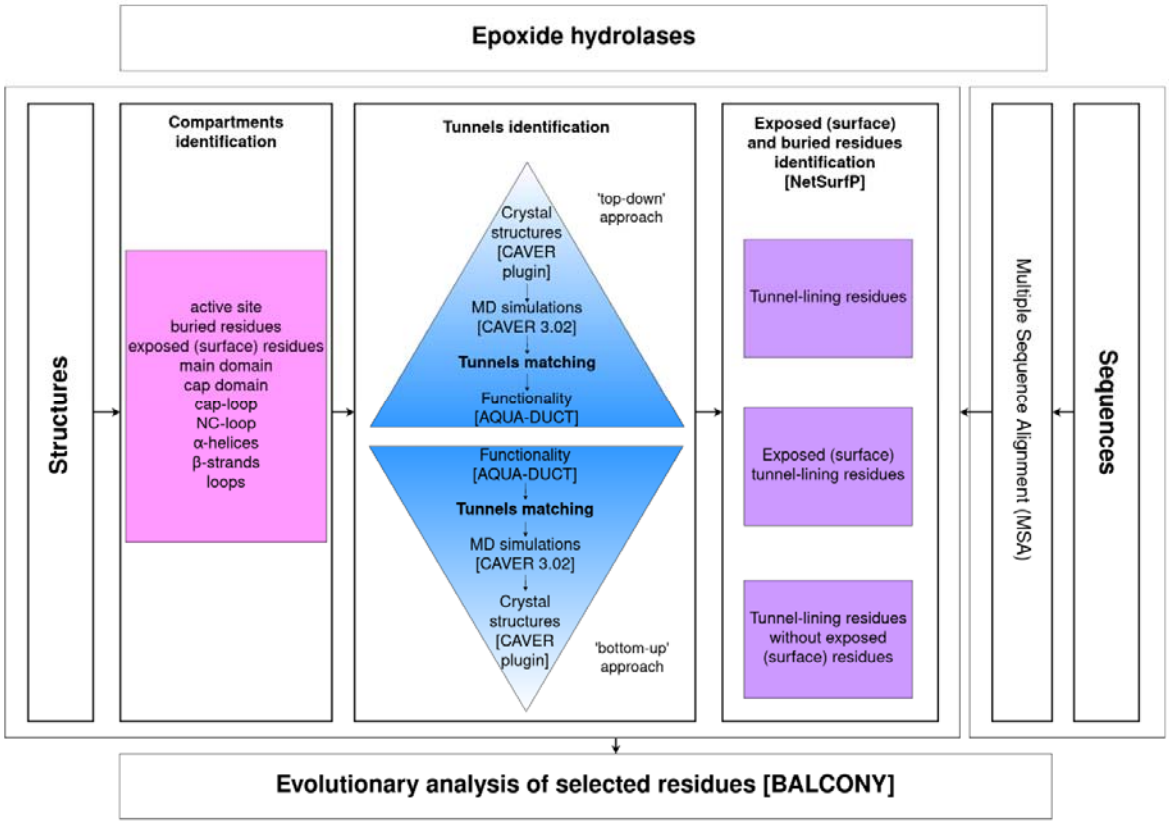

Figure 7. Research workflow.

Obtaining protein structures for analysis

Seven unique and complete crystal structures were downloaded from the Protein Data Bank (PDB). The selected structures all belong to the α/β hydrolase superfamily, share the same core fold scheme(48), and consist of a main domain and a cap domain (Barth et al., 2004). 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, 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 analysis (see **Supplementary Table S2**): the active site; buried and exposed 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.*(43) and of Smit and Labuschagne(84). The NetSurfP service was used to identify both buried and exposed residues. Tunnels identified by CAVER software were also selected for further analysis.

MD simulations

The H++ server(85) was used to protonate the analysed structures using standard parameters at a pH optimal for enzyme activity (**Supplementary Table S12**). Additionally, counterions were added to the structures. Water molecules were placed using a combination of 3D-RISM(86) and the Placevent algorithm(87). Water molecules were added to fill internal protein cavities and pockets(36). The Amber14 tLEaP package(88) was used to immerse the models in a truncated octahedral box of a 10 Å radius with TIP3P water molecules, and the ff14SB force field(89) was used for the parametrization of each system. The PMEMD CUDA package of the AMBER 14 software was used to run a set of 50 ns MD simulations of the selected EHs (one simulation per structure). To improve conformation sampling, the starting geometry for each system was maintained, 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

× mol^{-1} × $Å^{-2}$) and a final minimisation with no constraints of conjugate gradient energy minimisation. 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} under periodic boundary conditions with constant volume. An equilibration stage was run using the periodic boundary conditions with constant pressure as per the time noted in **Supplementary Table S16**, at 1 fs time step using Langevin dynamics, with a frequency collision of 1 ps^{-1} to maintain constant temperature. The 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. Coordinates were saved at intervals of 1 ps. The number of added water molecules and ions is shown in **Supplementary Table S15**.

CAVER analysis

Tunnel identification and analysis in each system was carried out using CAVER 3.02 software in two steps: i) the crystal structure of the enzyme was analysed by the CAVER plugin for PyMOL(90); ii) tunnels were identified and analysed in 50,000 snapshots of multiple molecular dynamics simulations by the standalone CAVER 3.02 software. The parameters used for both steps are shown in **Supplementary Table S17**. The tunnels found in MD simulations and in crystal structures were ranked and numbered on the basis of their throughput value(37).

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: Cα, backbone atoms, and side-chain atoms were considered to be of the same importance. For MD simulations' results, occurrence was defined differently: as the number of MD snapshots in which a 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 of the list of residues that form tunnels, i.e. by applying certain threshold values. In this study, the threshold was a number in 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(91):

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

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. In the case of tunnels identified in MD simulations but for which no corresponding counterpart was found in the crystal structures, their tunnel-lining residues were selected based on the cutoff threshold of 0.65. This value was chosen on the basis of quantile calculations for the tunnels identified in MD which had their counterparts in the crystallographic structures.

AQUA-DUCT analysis

AQUA-DUCT software was used to identify functional tunnels (defined as those tunnels which are used for water exchange between the active site and the protein's surroundings), and to compare them with the results of the geometry-based tunnel identification approach implemented in CAVER. AQUA-DUCT traced all water molecules which visited the active site cavity (called the *Object*) within the protein's interior (called the *Scope*). Regions where water molecules entered or left the *Scope* are marked by an inlet. All inlets were then clustered to reflect the entrances to the enzyme's active site. Amino acids which were used to define the *Object* cavity are listed in **Supplementary Table S17**.

Obtaining protein sequences, and multiple sequence alignment

Each of the amino acid sequences of the selected EHs (PDB IDs: 1S8O, 1CQZ, 2CJP, 4NZZ, 5URO, 5NFQ, 5NG7) was used as a separate query for a BLAST(92) search of similar protein sequences. The obtained results were merged and duplicates were removed, providing 1484 unique sequences (including those primarily selected). The outlying sequences (12) were detected and individually checked in the Uniprot database(93). 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.(57) 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. In order 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(94) using default parameters.

BALCONY analysis

BALCONY (Better ALignment CONsensus analYsis), 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(45) was calculated for each alignment position. Selected structures of *M. musculus*, *H. sapiens*, *S. tuberosum*, *T. reesei*, and *B. megaterium*, soluble epoxide hydrolases, 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 functional compartments/tunnels as shown in **Supplementary Table 1** and **Supplementary Tables S4–S11**. The compartment/tunnel residues were then appropriately mapped with MSA, and Schneider entropy values were collected.

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 remaining positions of 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 remaining positions of the MSA, the Epps–Singleton two-sample test(46) was used. The comparison analysis was performed using the es.test() function from GitHub repository(95). In an attempt to visualise the variability of selected tunnels (**Figures 4–6**), the collected entropy values of selected tunnel-lining residues without the exposed (surface) residues and the remaining MSA positions were sorted separately, and cumulative distribution functions (CDF) were calculated. For each position in the selected tunnel, a paired one from the rest of the MSA was found, based on the minimal CDF. Plots of CDF as a function of entropy score were prepared.

Acknowledgement

This work was supported by the National Science Centre, Poland (grant number DEC-2013/10/E/NZ1/00649) and the Own Scholarship Fund of the Silesian University of Technology in the year 2019/2020 (grant number 919/RN2/RR4/2019).

References

1. Kimura M, Ohta T. On Some Principles Governing Molecular Evolution. Proc Natl Acad Sci U S A. 1974;71(7):2848–52.
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.
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.
4. Jäckel C, Kast P, Hilvert D. Protein Design by Directed Evolution. Annu Rev Biophys. 2008;37:153–73.
5. Damborsky J, Brezovsky J. Computational tools for designing and engineering enzymes. Curr Opin Chem Biol. 2014;19:8–16.
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.
8. Siddiq MA, Hochberg GK, Thornton JW. Evolution of protein specificity: insights from ancestral protein reconstruction. Curr Opin Struct Biol. 2017;47:113–22.
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.
11. Guney E, Tuncbag N, Keskin O, Gursoy A. HotSprint: database of computational hot spots in protein interfaces. Nucleic Acids Res. 2008;36(Database):D662–6.
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.

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.

14. Martinez R, Schwaneberg U. A roadmap to directed enzyme evolution and screening systems for biotechnological applications. Biol Researcg. 2013;46(4).

15. Amaurys 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.

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

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

18. Sitbon E, Pietrokovski S. Occurrence of protein structure elements in conserved sequence regions. BMC Struct Biol [Internet]. 2007 Jan 9;7:3. Available from: http://www.ncbi.nlm.nih.gov/pubmed/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

20. Goldman N, Thorne JL, Jones DT. Assessing the Impact of Secondary Structureand Solvent Accessibility on Protein Evolution. Genetics. 1998;149:445–58.

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

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

23. Subramanian K, Mitusińska K, Raedts J, Almourfi F, Joosten H-J, 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

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 [Internet]. 2018 Dec 3;9(1):1314. Available from: http://www.nature.com/articles/s41467-018-03562-9

25. Petrović D, Risso VA, Kamerlin SCL, Sanchez-Ruiz JM. Conformational dynamics and enzyme evolution. J R Soc Interface [Internet]. 2018 Jul 18;15(144):20180330. Available from: https://royalsocietypublishing.org/doi/10.1098/rsif.2018.0330

26. Carlson GM, Fenton AW. What Mutagenesis Can and Cannot Reveal About Allostery. Biophys J [Internet]. 2016 May;110(9):1912–23. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0006349516300728

27. Weinkam P, Chen YC, Pons J, Sali A. Impact of Mutations on the Allosteric Conformational Equilibrium. J Mol Biol [Internet]. 2013 Feb;425(3):647–61. Available from: https://linkinghub.elsevier.com/retrieve/pii/S002228361200914X

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

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

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

31. Nakamura A, Yao M, Chimnaronk 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

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.

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

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

35. 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 [Internet]. 2013 Jan;31(1):38–49. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0734975012000390

36. Mitusińska K, Raczyńska A, Bzówka M, Bagrowska W, Góra A. Applications of water

molecules for analysis of macromolecule properties. Comput Struct Biotechnol J [Internet]. 2020;18:355–65. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2001037019304556

37. 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. Prlic A, editor. PLoS Comput Biol [Internet]. 2012 Oct 18;8(10):e1002708. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002708

38. Magdziarz T, Mitusińska K, Bzówka M, Raczyńska A, Stańczak A, Banas M, et al. AQUA-DUCT 1.0: structural and functional analysis of macromolecules from an intramolecular voids perspective. Elofsson A, editor. Bioinformatics [Internet]. 2020 Apr 15;36(8):2599–601. Available from: https://academic.oup.com/bioinformatics/article/36/8/2599/5682414

39. 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

40. 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

41. 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

42. 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

43. 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

44. Płuciennik A, Stolarczyk M, Bzówka M, Raczyńska A, Magdziarz T, Góra A. BALCONY: an R package for MSA and functional compartments of protein variability analysis. BMC Bioinformatics [Internet]. 2018 Dec 14;19(1):300. Available from: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2294-z

45. 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

46. 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

47. 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

48. 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

49. Persson F, Halle B. Transient Access to the Protein Interior: Simulation versus NMR. J Am Chem Soc [Internet]. 2013 Jun 12;135(23):8735–48. Available from: https://pubs.acs.org/doi/10.1021/ja403405d

50. Mitusińska K, Magdziarz T, Bzówka M, Stańczak A, Gora 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

51. Fink MJ, Syrén P-O. Redesign of water networks for efficient biocatalysis. Curr Opin Chem Biol [Internet]. 2017 Apr;37:107–14. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1367593117300261

52. Kong X-D, Yuan S, Li L, Chen S, Xu J-H, Zhou J. Engineering of an epoxide hydrolase for efficient bioresolution of bulky pharmaco substrates. Proc Natl Acad Sci [Internet]. 2014 Nov 4;111(44):15717–22. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.1404915111

53. 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

54. 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

55. 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

56. Heikinheimo P, Goldman A, Jeffries C, Ollis DL. Of barn owls and bankers: a lush variety of α/β hydrolases. Structure [Internet]. 1999 Jun;7(6):R141–6. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0969212699800793

57. van Loo B, Kingma J, Arand M, Wubbolts MG, Janssen DB. Diversity and Biocatalytic Potential of Epoxide 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

58. 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

59. Sehnal D, Svobodová Vařeková R, Berka K, Pravda L, Navrátilová V, Banáš P, et al. MOLE 2.0: advanced approach for analysis of biomacromolecular channels. J Cheminform [Internet]. 2013 Dec 16;5(1):39. Available from: https://jcheminf.biomedcentral.com/articles/10.1186/1758-2946-5-39

60. DePristo MA, de Bakker PI., Blundell TL. Heterogeneity and Inaccuracy in Protein Structures Solved by X-Ray Crystallography. Structure [Internet]. 2004 May;12(5):831–8. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0969212604001145

61. Furnham N, Blundell TL, DePristo MA, Terwilliger TC. Is one solution good enough? Nat Struct Mol Biol [Internet]. 2006 Mar 1;13(3):184–5. Available from: http://www.nature.com/articles/nsmb0306-184

62. 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 [Internet]. 2009 Jun 30;106(26):10505–10. Available from: http://www.pnas.org/cgi/doi/10.1073/pnas.0812152106

63. Lamb AL, Kappock TJ, Silvaggi NR. You are lost without a map: Navigating the sea of protein structures. Biochim Biophys Acta - Proteins Proteomics [Internet]. 2015 Apr;1854(4):258–68. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1570963914003379

64. Djinovic-Carugo K, Carugo O. Missing strings of residues in protein crystal structures. Intrinsically Disord Proteins [Internet]. 2015 Jan 23;3(1):e1095697. Available from: https://www.tandfonline.com/doi/full/10.1080/21690707.2015.1095697

65. Gora A, Brezovsky J, Damborsky J. Gates of Enzymes. Chem Rev [Internet]. 2013 Aug

14;113(8):5871–923. Available from: https://pubs.acs.org/doi/10.1021/cr300384w

66. Lopata A, Leveles I, Bendes ÁÁ, Viskolcz B, Vértessy BG, Jójárt B, et al. A Hidden Active Site in the Potential Drug Target Mycobacterium tuberculosis dUTPase Is Accessible through Small Amplitude Protein Conformational Changes. J Biol Chem [Internet]. 2016 Dec;291(51):26320–31. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0021925820344549

67. Hamre AG, Frøberg EE, Eijsink VGH, Sørlie M. Thermodynamics of tunnel formation upon substrate binding in a processive glycoside hydrolase. Arch Biochem Biophys [Internet]. 2017 Apr;620:35–42. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0003986117301029

68. Chaplin VD, Valliere MA, Hangasky JA, Knapp MJ. Investigations on the role of a solvent tunnel in the α-ketoglutarate dependent oxygenase factor inhibiting HIF (FIH). J Inorg Biochem [Internet]. 2018 Jan;178:63–9. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0162013416302835

69. Li G, Yao P, Gong R, Li J, Liu P, Lonsdale R, et al. Simultaneous engineering of an enzyme's entrance tunnel and active site: the case of monoamine oxidase MAO-N. Chem Sci [Internet]. 2017;8(5):4093–9. Available from: http://xlink.rsc.org/?DOI=C6SC05381E

70. Gorshkova IN, Mei X, Atkinson D. Arginine 123 of apolipoprotein A-I is essential for lecithin:cholesterol acyltransferase activity. J Lipid Res [Internet]. 2018 Feb;59(2):348–56. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0022227520342371

71. Rauwerdink A, Kazlauskas RJ. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. ACS Catal [Internet]. 2015 Oct 2;5(10):6153–76. Available from: https://pubs.acs.org/doi/10.1021/acscatal.5b01539

72. 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

73. Jeon J, Nam H-J, Choi YS, Yang J-S, 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

74. 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

75. 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

76. 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

77. 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

78. 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

79. 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

80. 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

81. 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

82. 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

83. 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

84. 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&spage=1145

85. 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 [Internet]. 2012 Jul 1;40(W1):W537–41. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks375

86. 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 [Internet]. 2010 Mar 9;6(3):607–24. Available from: https://pubs.acs.org/doi/10.1021/ct900460m

87. 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 [Internet]. 2012 Jul 5;33(18):1536–43. Available from: http://doi.wiley.com/10.1002/jcc.22984

88. Case DA, Babin V, Berryman JT, Betz RM, Cai Q, Cerutti DS, et al. AMBER 14. University of California, San Francisco; 2014.

89. 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 [Internet]. 2015 Aug 11;11(8):3696–713. Available from: https://pubs.acs.org/doi/10.1021/acs.jctc.5b00255

90. Schrödinger. The PyMOL Molecular Graphic Systems. Schrödinger, LLC;

91. Gilbert G. Distance between Sets. Nature [Internet]. 1972 Sep;239(5368):174–174. Available from: http://www.nature.com/articles/239174c0

92. 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

93. 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

94. 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

95. Ileppane. https://github.com/ileppane/statistics/blob/master/EStest.R. 2014.