catalysts

MDPI

Article

Conformational Landscapes of Halohydrin Dehalogenases and Their Accessible Active Site Tunnels

Miquel Estévez-Gay [1], Javier Iglesias-Fernández [1,*,†] and Sílvia Osuna [1,2,*]

[1] CompBioLab Group, Institut de Química Computacional i Catàlisi (IQCC) and Departament de Química, Universitat de Girona, c/Maria Aurèlia Capmany 69, 17003 Girona, Catalonia, Spain; miquel.estevez@udg.edu

[2] ICREA, Passeig Lluís Companys 23, 08010 Barcelona, Catalonia, Spain

* Correspondence: jiglesiasfrn@gmail.com (J.I.-F.); silvia.osuna@udg.edu (S.O.)

† Present address: Nostrum Biodiscovery, Carrer de Baldiri Reixac, 10–12, 08028 Barcelona, Catalonia, Spain.

Received: 12 November 2020; Accepted: 28 November 2020; Published: 1 December 2020

Abstract: Halohydrin dehalogenases (HHDH) are industrially relevant biocatalysts exhibiting a promiscuous epoxide-ring opening reactivity in the presence of small nucleophiles, thus giving access to novel carbon–carbon, carbon–oxygen, carbon–nitrogen, and carbon–sulfur bonds. Recently, the repertoire of HHDH has been expanded, providing access to some novel HHDH subclasses exhibiting a broader epoxide substrate scope. In this work, we develop a computational approach based on the application of linear and non-linear dimensionality reduction techniques to long time-scale Molecular Dynamics (MD) simulations to study the HHDH conformational landscapes. We couple the analysis of the conformational landscapes to CAVER calculations to assess their impact on the active site tunnels and potential ability towards bulky epoxide ring opening reaction. Our study indicates that the analyzed HHDHs subclasses share a common breathing motion of the halide binding pocket, but present large deviations in the loops adjacent to the active site pocket and N-terminal regions. Such conformational differences affect the available tunnels for epoxide binding to the active site. The superior activity of the HHDH G subclass towards bulkier substrates is explained by the additional structural elements delimiting the active site region, its rich conformational heterogeneity, and the substantially wider and frequently observed active site tunnels. This study therefore provides key information for HHDH promiscuity and engineering.

Keywords: Halohydrin dehalogenases; conformational dynamics; active site tunnels; molecular dynamics simulations

1. Introduction

Enzymes are highly efficient in accelerating the chemical reactions under biologically controlled conditions, and can provide synthetically useful building blocks with high selectivity and specificity. The ability of enzymes of accelerating additional side reactions, i.e., they present catalytic promiscuity, is thought to play a key role in the evolution of enzymes towards new functions [1,2]. The appearance of novel enzyme functionalities through evolution has been attributed to the fine-tuning of the conformational ensemble present in solution, whose relative stabilities can be tuned by mutations [3–7]. Many of these pre-existing conformations can play a key role in recognizing and binding the substrate and/or releasing the product, in conferring the enzyme the catalytic promiscuity, and in some cases in regulating the operating allosteric communication. These additional conformations of the enzyme can present deviations in the available tunnels for accessing the active site, thus playing a role in the enzyme catalytic activity. Indeed, the engineering of some flexible loops gating substrate access to

the active site and contributing to product release were shown to be key for boosting the catalytic activity of some enzymes [8,9]. The enzyme conformational landscape therefore plays a crucial role in its function, promiscuity, regulation, and evolution.

Halohydrin dehalogenases (HHDHs) perform a cofactor independent dehalogenation reaction for degrading halogenated compounds. They are highly valuable biocatalysts as they exhibit promiscuous epoxide ring-opening catalytic activity in the presence of small nucleophiles, thus giving access to novel carbon–carbon, carbon–oxygen, carbon–nitrogen, and carbon–sulfur bonds [10,11]. Some biocatalytic examples of HHDH-catalyzed reactions include their application for obtaining statin side chains precursors more efficiently, enantiopure epihalohydrins, oxazolidinones, tertiary and beta-substituted alcohols [12–18]. As shown by the solved X-ray structures, the active site of HHDH is composed by a binding site for the epoxide and a spacious halide binding pocket that can accommodate linear monovalent anions as nucleophiles (see Figure 1) [19,20]. HHDHs feature a conserved catalytic triad composed by Ser-Tyr-Arg, which catalyzes epoxide formation and subsequent halide release [19,20]. The promiscuous epoxide ring-opening reaction usually occurs at the less-hindered carbon via S_N2 mechanism [20], and the range of epoxides accepted by HHDHs is usually limited to terminal epoxides [21], although some recent examples of HHDH accepting sterically more demanding epoxides have been reported [22] (see Scheme 1). In a recent paper by the Schallmey lab, novel HHDHs were identified following a database mining approach, with six phylogenetic subtypes of HHDH ranging from A through G characterized [23]. Particularly useful is HheG, as it represents the first example of HHDH able to accept with synthetically useful activities bulky cyclic epoxides as substrates [22,23]. Interestingly, HheG has also been recently found to exhibit high activity towards sterically demanding di-substituted epoxides, whereas the A-F subtypes present activity only towards methyl-disubstituted epoxide substrates [24] (see Scheme 1). Unfortunately, structure (i.e., conformational dynamics)–activity relationships are not available for this family of enzymes.

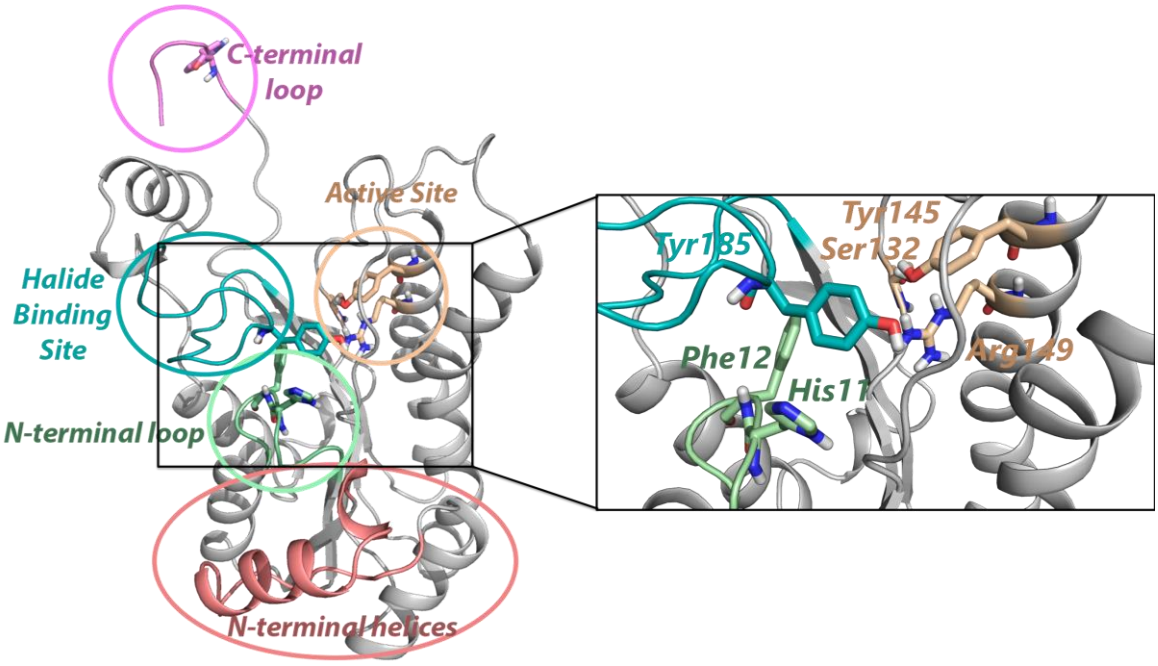
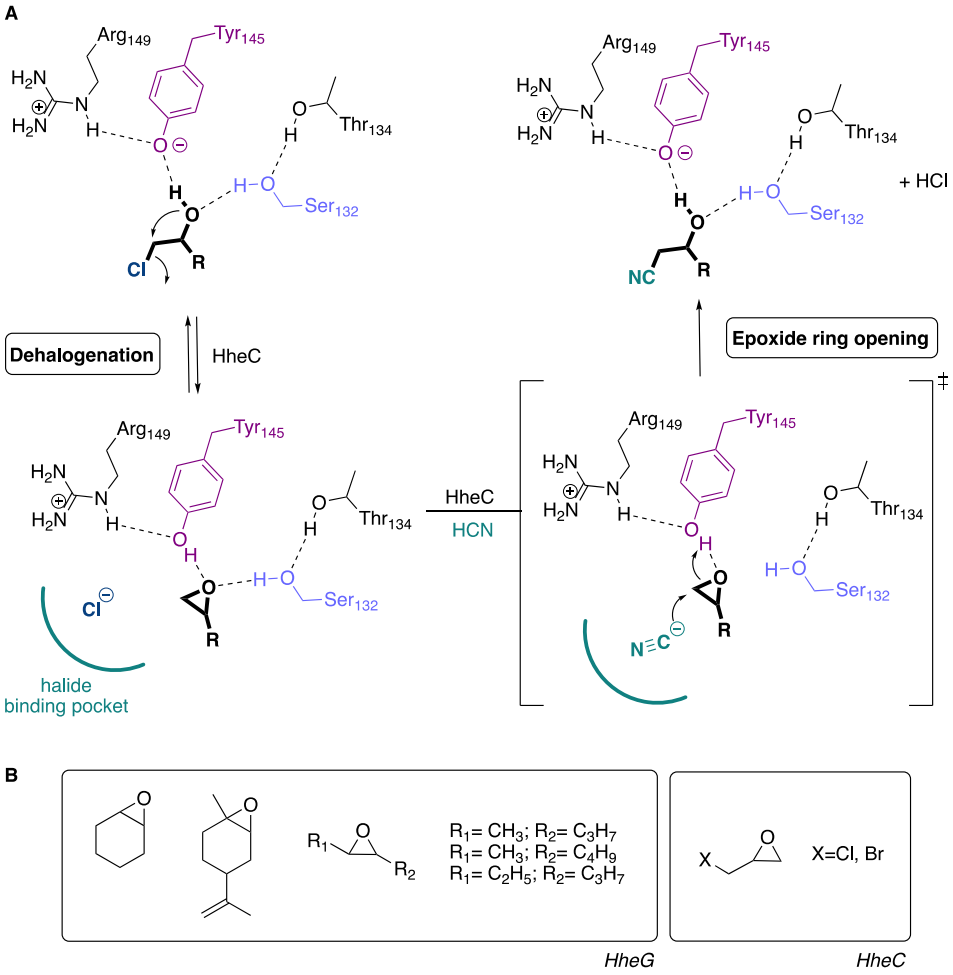


Figure 1. Halohydrin dehalogenase (HHDH) distinct structural elements and zoom of the active site and halide binding pockets based on the HheC structure. Active site residues are highlighted in wheat color, halide-binding site in teal, N-terminal loop in light green, C-terminal loop in purple and N-terminal 6–7 helices in salmon. In the active site zoom, potential residues blocking the accessible active site tunnels are depicted using the same color scheme.



Scheme 1. (**A**). Reaction scheme of the two-step HHDH catalyzed enzymatic reaction: (1, left) dehalogenation for epoxide formation, followed by the promiscuous (2, right) enantioselective epoxide-ring opening reaction by a nucleophile. Numbering of the residues is based on HheC. (**B**). Representative bulky epoxide substrates accepted by the HheG variant: cyclohexene oxide, limonene epoxide [22], and racemic di-substituted trans-epoxides [24]. Examples of epoxide substrates accepted by HheC are also displayed: epichlorohydrin and epibromohydrin [11].

Computational techniques and, in particular, molecular dynamics (MD) simulations are particularly useful in elucidating the ensemble of thermally accessible enzyme conformations by integrating Newton's laws of motion [25]. This enables the reconstruction of the enzyme conformation landscape and assess how this is shifted by ligand binding, sequence differences between protein family members, and/or the introduction of mutations in the enzyme active site or at distal positions [7,26]. Recovery of time-dependent dynamical descriptors, such as volume cavities, solvent-accessible surface area, or changes in internal tunnels/channels is also possible by post-processing the highly dimensional MD datasets [4,27]. Particularly useful is the application of dimensionality reduction techniques for automatically identifying key enzymatic states from MD simulations and account for as much information as possible. These methods can be broadly classified into linear and non-linear dimensionality reduction techniques and have been successfully used to identify key states in MD simulations [28–33]. Combinations of linear and non-linear methods have also been proposed to take advantage of both approximations, with the time-lagged t-Distributed Stochastic Neighbor Embedding (t-SNE) [33] as a clear example [34]. In this direction, we have previously developed a computational protocol based on the combination of the linear time-Independent Component Analysis (tICA) [35]

and t-SNE [33] for elucidating the conformational ensemble of *Candida rugosa* lipase and its accessible tunnels for substrate binding to the active site [27].

In this work, intrigued by the observed differences in the catalytic activity and substrate scope of the different subclasses of HHDHs, we characterized the intrinsic conformational landscapes of HHDHs and correlate them with changes in available tunnels for substrate binding and product release in A2, B, C, D2, and G of HHDH subclasses. Extensive MD simulations, followed by dimensionality reduction techniques and tunnel analysis with the CAVER package, provide a clear rationalization of substrate preferences of the studied HHDH subclasses. Our protocol based on projecting MD data into a linear and non-linear reduced space ensures an extensive characterization of the HHDH dynamical ensemble and elucidates how it is modified in the different subclasses. This approach not only provides clear insights of how the available tunnels for substrate binding and product release are altered, but also derives interesting data for HHDH evolution and engineering. This is the first comprehensive study that evaluates the conformational dynamics and associated changes in tunnel accessibility to the active site in different phylogenetic subclasses of HHDH.

2. Results

2.1. Conformational Landscapes of Halohydrin Dehalogenases (HHDHs)

Our study starts with the evaluation of the conformational landscapes of the different Halohydrin dehalogenase (HHDH) variants from the subclasses A2, B, C, D2, and G, followed by a rationalization of how dynamics affect their promiscuity towards epoxide ring opening [22–24]. All analyzed HHDH subclasses are tetrameric, they share the catalytic triad (Ser, Tyr and Arg) and present the halide binding residues located in the loop that connects the 6th β-strand and 9th α-helix (see Figure 1). They also exhibit some structural differences: two α-helices (2nd and 3rd, residues 32–55 according to HheC numbering) are found near the N-terminal part of the protein, with the exception of B and D2; and in G the 2nd α-helix is disordered. HheG also presents the distinctive feature of presenting an additional α-helix in the halide binding site loop, which might potentially broaden the active site entrance tunnel. In the particular case of HheC, the C-terminal part of the opposite monomer is positioned close to the active site and halide binding pockets, which interacts with some halide binding site sidechain residues. Such large structural differences among the subclasses studied might lead to substantial deviations in the HHDH conformational landscapes.

We evaluated the conformational landscapes of A–D, G HHDHs by performing Molecular Dynamics (MD) simulations with an accumulated simulation time of 1.25 microseconds for each system, in explicit water solvent using AMBER (see Methods) [36]. The obtained high dimensional MD dataset was then further analyzed by combining linear and non-linear dimensionality reduction techniques (see Figure 2). We first applied the linear method time-Independent Component Analysis (tICA) [35], which focuses on those motions that most rarely occur, i.e., the slowest kinetically relevant processes. Similarly to the linear Principal Component Analysis (PCA) method [28], tICA allows for a direct identification of relevant descriptors with limited complexity on the dynamics represented, thus requiring more data dimensions to represent the same data variance. To capture more data variance with a reduced number of descriptors, we combined tICA with the non-linear t-Distributed Stochastic Neighbor Embedding (t-SNE) [33] method. This method is used to represent high dimensional data into a 2D or 3D visually appealing low dimensional space by approximating the probability distribution of points in the high dimensional data into the reduced space. In this way, similar data points are placed together with high probability in the reduced space whereas dissimilar data is located distant. By following this tICA-t-SNE methodology, the most relevant conformational states sampled along the MD simulations for each HHDH subclass were revealed (see Figure 2 and Figure S1).

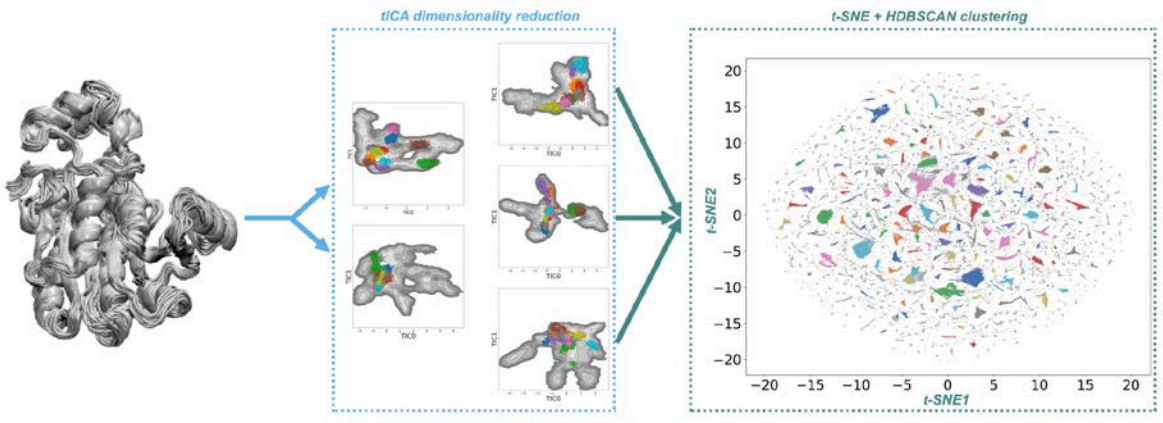


Figure 2. Computational protocol used to reconstruct the conformational landscapes of the different HHDH subclasses. It is based on a two-step process consisting of: first applying to the MD dataset the linear time-Independent Component Analysis (tICA) [35], followed by the application of the non-linear t-distributed Stochastic Neighbor Embedding (t-SNE) [33] method. In this fashion, the high dimensional MD dataset is reduced into a 2D space, that is subsequently clustered using HDBScan [37].

The evaluation of the enzyme conformational dynamics by means of the mentioned tICA-t-SNE methodology indicated that all HHDH subclasses have a moderate to high degree of flexibility, exploring conformations outside the main energy minima (see Figure 3). This is especially true for HheA2 and HheG, which display the most flexible behavior among variants. It is worth mentioning that for all analyzed HHDH subclasses, the projection of the t-SNE most populated clusters onto the corresponding tICA space agrees with well-defined energy minima, confirming that the t-SNE dimensionality reduction method faithfully represents protein dynamics (Figure 3). For all HHDH subclasses explored, the slowest conformational change corresponds to a 'breathing' motion of the protein, involving a coordinated conformational change of both catalytic and halide binding sites. The impact of this breathing motion into substrate binding or product release will be evaluated below.

Analysis of the structural differences observed for HheA2 clusters indicates that conformational changes mainly occur in the halide binding site (Residues 170–210), and the loop located close to the catalytic Tyr146 (Residues 80–95) (Figure 3). The slowest conformational changes (as represented by tIC1, tIC2 and tIC3) correspond to different conformations of the α-helix contained in the halide binding region, and loop motions within the active site (tIC2). The conformational changes observed in the case of the HheD2 variant are similar to those observed for HheA2, however, the catalytic (130–150) and halide binding (170–190) residues display a lower degree of flexibility compared to HheA2.

As mentioned earlier, a distinctive feature of HheC and HheG is the presence of a flexible region close to the N-terminal part of the protein, comprised by an α-helix (Residues 32–36) in HheC and a disordered loop (Residues 30–50) in HheG. The HheC most populated conformations mainly involve coordinated motions of the N-terminal flexible α-helix and the halide binding pocket region. The HheG disordered character of the loop region comprised by residues 30 to 50, which is close to the halide binding residue Tyr13, is the slowest conformational change (as described by tIC0 and tIC1 dimensions). The large structural variability of this protein region is likely involved in substrate accessibility and the presence/absence of lateral access channels, which most likely confer the enzyme the ability to accept bulkier epoxide substrates (see below). As explained before for the D2 and A2 subclasses, HheG also presents a 'breathing' or coordinated motion of the catalytic and halide binding domains.

Contrary to the previous HHDH variants, our MD simulations indicate that HheB displays a quite different conformational behavior, with a tight packing of the protein structure and only minor rearrangements of the α-helix located above the halide binding site (residues 170–190). Most populated conformations only display minor rearrangements on the halide binding region. Interestingly, most visited t-SNE clusters for HheB fall into a narrow region of the tICA space, thus explaining the observed conformational rigidity.

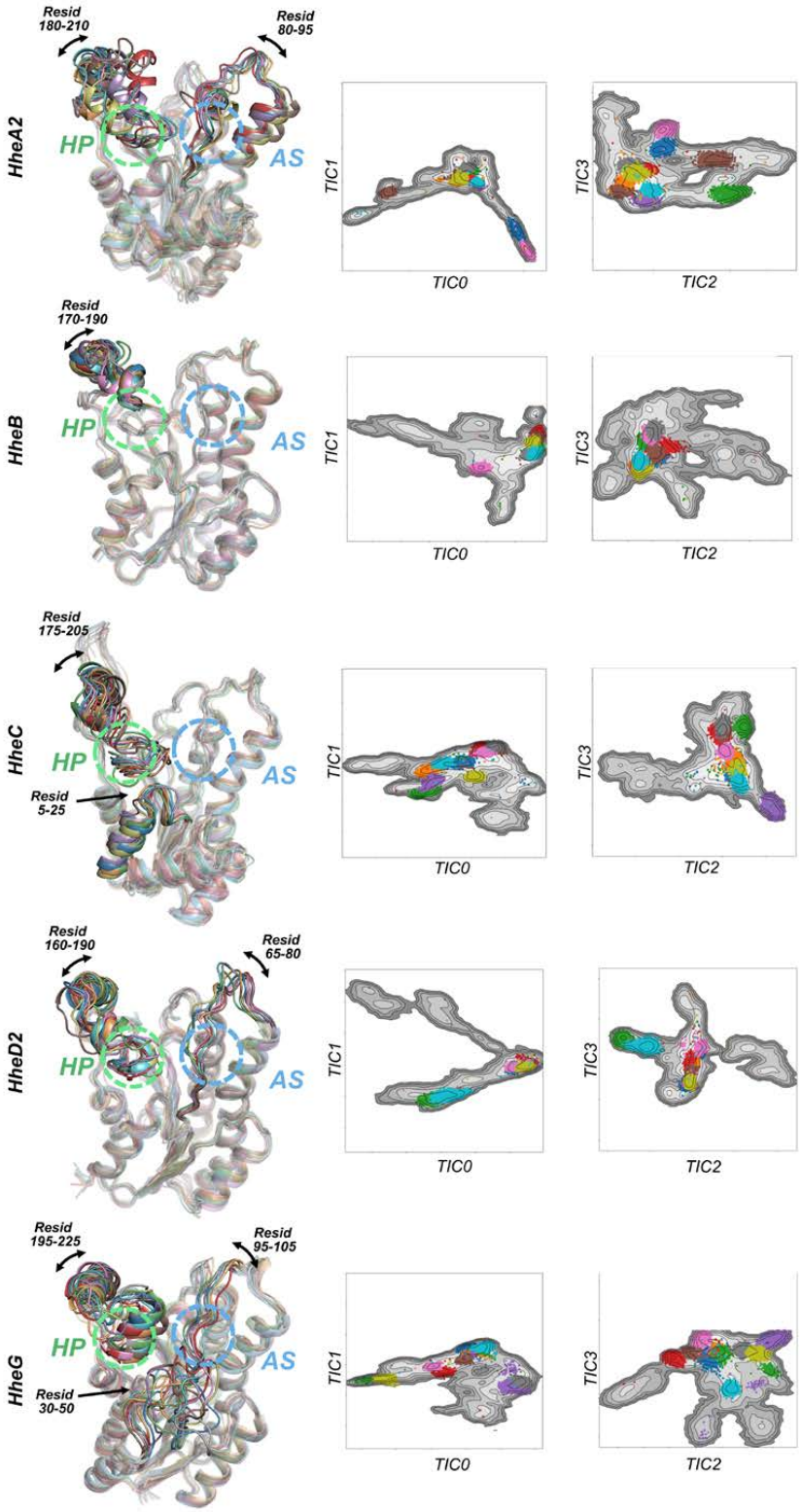


Figure 3. Representation of the 10 most populated MD conformations as described by the t-SNE technique for the different HHDH subclasses analyzed: HheA2, HheB, HheC, HheD2, and HheG. The 10 different conformations (each one colored differently) are projected on the tICA conformational landscapes. The most flexible parts of the enzymes are marked and numbered accordingly. The location of the active site (AS) and halide binding pocket (HP) are marked with a green and blue discontinuous circle, respectively.

2.2. Active Site Accessibility Tunnels of Halohydrin Dehalogenases (HHDHs)

The analysis of the conformational landscapes of the A–D, G HHDHs has revealed major differences among subclasses in the flexibility of the halide binding site region and loops located at the vicinity of the active site pocket. Such conformational changes may impact the available tunnels for substrate accessibility to the active site, thus regulating the enzyme ability to accept bulky epoxide substrates for the industrially relevant promiscuous reactivity. Tunnel analysis with the CAVER software [38] was performed for all t-SNE clusters of each HHDH studied system (see Figures 2 and 3). The average bottleneck radius of the computed tunnel (BR, i.e., narrower region of the tunnel) for each cluster was calculated (see Tables 1 and 2).

Table 1. Mean tunnel bottleneck radius (BR, in Å) for each HHDH system computed on a representative structure of each cluster center (see Methods).

HHDH	Tunnel T1	Tunnel T2	Tunnel T3
HheA2	1.8 ± 0.4	1.6 ± 0.6	n.d. [1]
HheB	1.9 ± 0.6	1.8 ± 0.8	n.d. [1]
HheC	2.0 ± 0.3	1.3 ± 0.2	1.0 ± 0.02
HheD2	1.8 ± 0.5	1.7 ± 0.4	n.d. [1]
HheG	2.2 ± 0.4	1.9 ± 0.5	1.8 ± 0.5 Å

[1] Not detected.

Table 2. Computed tunnel frequency for each HHDH subclass (see Methods).

HHDH	Tunnel T1	Tunnel T2	Tunnel T3
HheA2	92.4%	12.3%	n.d. [1]
HheB	97.6%	25.7%	n.d. [1]
HheC	96.9%	77.5%	36.2%
HheD2	88.0%	71.1%	n.d. [1]
HheG	97.6%	91.8%	65.8%

[1] Not detected.

One major tunnel (named T1) was identified in all analyzed HHDH variants (see Figure 4). In some cases, two additional tunnels (T2–T3) were also found, although with different probabilities. T1 defines the direct vertical path from the active site to the bulk solvent, and is shaped by the active and halide binding sites. This tunnel was found in 92.4%, 97.6%, 96.9%, 88.0%, and 97.6% of the clustered MD frames for HHDH subclasses A2, B, C, D2, and G, respectively. Contrary to what we initially expected, the C-terminal part from the neighbor chain in HheC does not affect T1 formation, thus not blocking substrate accessibility to the active site (see Figure S2). The analysis of the average bottleneck radius (BR) of T1 in the different HHDH subclasses indicates that in A2, B, C, and D it is ca. 1.90 Å, whereas it is slightly larger in the case of HheG (ca. 2.2 Å). This larger BR for T1 observed in G together with its high frequency (98%) explains the higher catalytic activity of HheG with substantially bigger epoxide substrates [22]. The BR of T1 is located close to the active site in all HHDHs, and thus contrary to what one might have initially expected, T1 is not directly affected by the 'breathing' motions of the α-helix contained in the halide binding region and the loop close to the catalytic residues described above.

T2 and T3 are lateral tunnels delimitated by the position of the sidechains of some blocking residues (H11, F12, I84, Y185, F186 and the N-terminal loop mentioned above, numbering according to HheC). Some deviations in the T2 exit to the bulk solvent can be found depending on the position of the N-terminal part and the H11, and F12 sidechain conformations. If the N-terminal part is not blocking the direct exit of T2, the shape and exit of T2 is regulated by the rotation of H11 and F12 sidechains. Alternatively, the longer tunnel T2' located under the halide binding site loop (residues 180–183 in HheC) can be formed. Tunnel T2 was found in 12.3%, 25.7%, 77.5%, 71.1% and 91.8% of the MD frames for HHDH subclasses A2, B, C, D2, and G, respectively. T2 is therefore hardly found in subclasses A2 and B, more visited in C and D2, and highly frequent in HheG. The analysis of the mean BR

reveals that in A2, T2 is hardly present and when formed has a narrow mean BR value of ca. 1.6 Å. In B2, the average BR is expanded to ca. 1.8 Å although with a large variability and low frequency (26%). In contrast, T2 is highly frequent in C (78%) but it is likely too narrow to allow the access of bulky substrates (mean BR of 1.3 Å). The high frequency of T2 in HheD2 and HheG (72% and 92%, see above), is combined with larger average BR values of ca. 1.7 Å in the case of D2, and ca. 1.9 Å in G. These findings of a high frequency of T2 combined with wider BR are in line with the higher activity of HheD2 and HheG with larger di-substituted epoxide substrates [24].

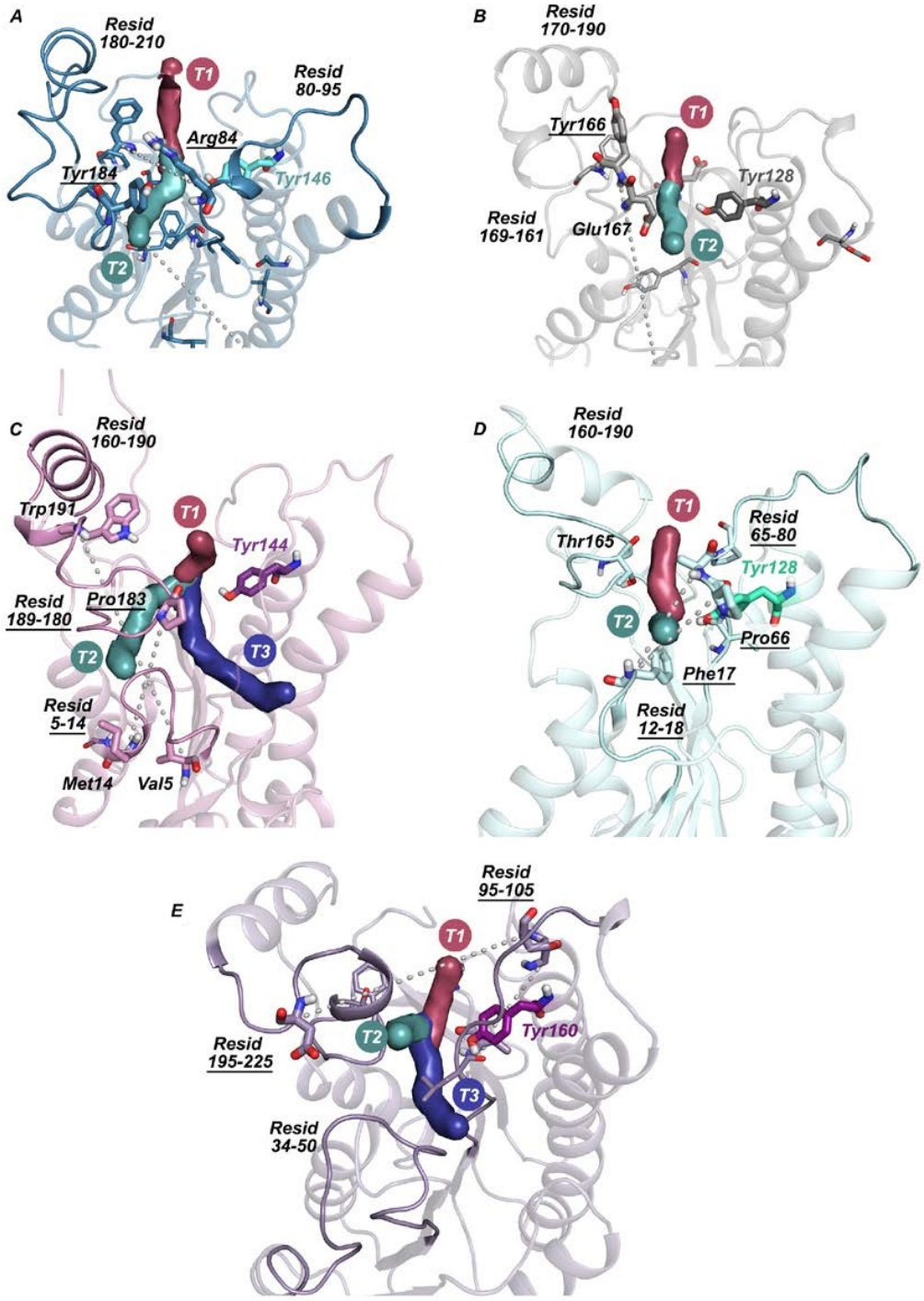


Figure 4. Representation of the three major tunnels that exist in (**A**) HheA2, (**B**) HheB, (**C**) HheC, (**D**) HheD2, and (**E**) HheG: T1 shown in raspberry, T2 in teal, and T3 in dark blue. The key elements that determine T2 formation in the different subclasses are highlighted.

We hypothesized that these large differences in the prevalence of T2 in the analyzed HHDH classes might be related to its associated conformational changes and/or structural variations. To that end, we relied on random forest classifiers [39] to elucidate the key heavy atom distances that modulate T2 formation (see Figure 4, Figures S3–S8). In A2, T2 formation is directly affected by the side-chain conformation of Tyr184 located at the halide binding loop (180–210) and Arg84 contained in the loop close to the catalytic Tyr146 (Residues 80–95). As shown in Figure 3, these regions correspond to the most flexible parts of HheA2, and thus T2 formation is directly related to the enzyme conformational dynamics. In B, T2 formation depends on the side-chain conformation of Tyr166 located at the halide binding loop (see Figure 4B). In fact, T2 is only observed when Tyr166 is displaced out of the active site pocket (see Figure 4B). HheB is the most rigid HHDH analyzed (see Figure 3), and the halide binding region containing Tyr166 does not exhibit a high flexibility, thus explaining the low frequency of T2 in B.

In HheC, T2 is substantially more frequently observed (78%), however, as shown in Figure 4C, T2 follows a slightly different path. In C, T2 formation is dependent on the conformation of the halide binding loop containing Pro183 (189–180) and the positioning of the N-terminal loop 5–14. These regions also correspond to the most flexible parts of the enzyme (see Figure 3), and therefore the formation of T2 in C is also related to its conformational dynamics. The high frequency of T2 in HheD2 is explained by its different location with respect to the previously mentioned cases (see Figure 4D). T2 is rather short and depends on the side-chain conformation of Phe17 at the N-terminal loop 12–18, and the adjacent loop close to the catalytic Tyr128 (65–80). The latter loop exhibits a quite high flexibility (see Figure 3), however, the N-terminal loop is rather rigid. As observed in Figure 4D, T2 in HheG is located between the halide binding loop 195–225, and the loop located close to the catalytic Tyr160 (95–105). However, given the wider active site pocket of HheG a longer distance between the latter loops is observed (this distance is 7.6 ± 2.3 Å in HheG, whereas 5.0 ± 2.4, 6.4 ± 3.2, 4.6 ± 1.4, 6.6 ± 2.3 for HheA2, HheB, HheC, HheD2, respectively), which makes T2 less dependent on their adopted conformation thus explaining the high prevalence (92%) of this tunnel in G. Thanks to the additional α-helix in the halide binding loop in HheG, T2 is not hampered by the N-terminal loop as observed in the other HHDH subclasses.

Tunnel T3 is only observed in those HHDH subclasses with a rather flexible N-terminal region (HheC and HheG, see Figures 3 and 4). The presence of T3 is therefore related to the positioning of the N-terminal loop and its associated conformational dynamics. T3 was found only in 36.19% of the HheC MD clusters, whereas it was found in 65.8% of the HheG MD dataset in line with the higher flexibility of the N-terminal loop observed in the G subclass. The analysis of the average BR indicates that although in HheC T3 is observed in 37% of the analyzed structures it is too narrow (ca. 1.0 Å) to allow the access of the epoxide substrate to the active site. In contrast, T3 is highly frequent in HheG and presents a substantially larger average BR of ca. 1.8 Å. This observation is again in line with the higher ability of HheG to accommodate bulkier epoxide substrates for the industrially relevant promiscuous reactivity [22].

3. Discussion

The repertoire of Halohydrin dehalogenases (HHDH) has been recently expanded, which provides access to some novel HHDH subclasses. These novel enzymes present substantial structural similitudes, although large differences especially in loops and alpha-helices located at the vicinity of the active site of the enzyme are also present. The analysis of the conformational landscapes of HHDH by means of linear and non-linear dimensionality reduction techniques has revealed that a common feature of all analyzed HHDH is a high flexibility of the alpha-helix and loop containing the halide binding pocket, i.e., 'breathing motion'. Apart from this common motion, there are some conformational differences among the analyzed subclasses: HheA2 and HheD2 exhibit a high flexibility of the loop located close to the catalytic Tyr; HheC has instead a high mobility of the N-terminal loop, and HheG is the most

conformationally rich HHDH displaying a large mobility of the N-terminal loop, the loop located adjacent to the catalytic Tyr and the already mentioned halide binding pocket.

The characterization of the accessible tunnels at the ensemble of conformations explored by means of CAVER has evidenced some relevant deviations in the available active site tunnels, some of them clearly influenced by the conformational dynamics of the HHDH subclass. All analyzed enzymes present a major tunnel (named T1) that vertically connects the active site pocket to the bulk solvent through the cavity formed between the loop located close to the catalytic Tyr and the halide binding pocket. In contrast to what we originally expected, the formation of T1 is independent from the above-mentioned breathing motion of the halide binding pocket and has a high occurrence in all HHDH (which ranges from ca. 88–98%). The bottleneck radius (BR) of T1 is located close to the active site residues, thus not being substantially affected by the halide binding pocket conformation. The computed BR for T1 is ca. 1.9 Å for all HHDH, except in HheG that is broadened to ca. 2.2 Å. Interestingly, dramatic differences are observed in the case of the secondary tunnel T2. The length, the exact positioning, BR values, and the frequency of T2 is subclass-dependent. In A2 and B, T2 is hardly present and is mostly dependent on the conformation of a Tyr residue (185 in A2 and 167 in B) contained in the halide binding pocket. In HheC, T2 has a high frequency and is located between the halide binding pocket (179–190) and the N-terminal loops (5–14), which present a rather high flexibility. T2 is substantially shorter in HheD2 and is situated between the rather rigid N-terminal and the substantially more flexible loop situated close to the catalytic Tyr. Thanks to the additional α-helix in the halide binding loop in HheG, T2 formation is less affected by the conformation of the halide binding region, the N-terminal and the loop situated close to the catalytic Tyr. The BR of T2 ranges from 1.3 Å for HheC to 1.9 Å for HheG. Finally, a third secondary tunnel (T3) is also present in the case of HheC and HheG that present a substantially more flexible N-terminal region. T3 is, however, less observed in HheC and exhibits a substantially narrower BR value (1.0 Å for C and 1.8Å for G).

4. Materials and Methods

Systems Set-Up. MD simulations were carried out on the structures of A2, B, C, and G HHDH subclasses with protein database (PDB) codes 1ZMO, 4ZD6, 1ZMT, and 5O30, respectively, as initial structures. The X-ray structure for D2 is not released yet (made available by the Schallmey lab). Protonation states of enzyme residues were assigned based on pKa values provided by the H++ server (http://biophysics.cs.vt.edu/H++) [40]. The enzymes were then solvated in a pre-equilibrated cubic box with a 10 Å buffer of transferable intermolecular potential with 3 points (TIP3P) [41] water molecules, resulting in the addition of approximately 27,000 solvent molecules per protein variant. The systems were neutralized by the addition of approximately 32 explicit counter ions (Na+). All simulations were done using the Amber 99SB force field (ff99SB-ildn) [42].

MD Simulations. The graphics processing unit (GPU) version of pmemd in Amber16 was used for the MD simulations. A two-stage geometry optimization was performed, first minimizing the positions of solvent molecules and ions, by imposing harmonic positional restraints of 500 kcal mol^{-1} $Å^{-2}$ on solute molecules, followed by an unrestrained minimization. Afterwards, a gradual heating of the systems was performed by increasing the temperature 50 K along six 20 ps sequential MD simulations (0–300 K) under constant volume and periodic boundary conditions. Harmonic restraints of 10 kcal/mol were applied to the solute, and the Langevin equilibration scheme was used to control and equalize the temperature. The time step was kept at 1 fs during the heating stages, allowing potential inhomogeneities to self-adjust. Each system was then equilibrated without restraints for 2 ns with a 2 fs time step at a constant pressure of 1 atm and temperature of 300 K. After equilibration in the isothermal-isobaric ensemble (NPT), 5 replicas of 250 ns were run for each system (i.e., 1.25 μs per HHDH subclass) in the canonical ensemble (NVT) yielding a total MD simulation time for all systems of 6.25 μs.

MD dimensionality reduction. MD simulation trajectories were post-process with the pyemma2 software package [43]. C-alpha coordinates of the aligned protein subclasses at each nanosecond of MD

simulation were used as initial features, resulting in 182,250,000, 168,000,000, 189,000,000, 168,000,000, 192,750,000 extracted values (features x frames x replicas) for the A2, B, C, D2, and G HHDH subclasses, making the statistical analysis unfeasible. Subsequently, the time-lagged Independent Component Analysis (t-ICA) [35], with a lag time τ set to obtain the minimum number of reduced dimensions, was applied to reduce the dimensionality of the initial MD features. The linear method t-ICA focuses on those motions that most rarely occur, i.e., the slowest kinetically relevant processes. Conversely to linear methods, non-linear techniques have the advantage of capturing more data variance with less descriptors, although at the cost making the biophysical interpretations of such reduced descriptors challenging. These methods include the recently proposed variational autoencoders [31,32], and the t-Distributed Stochastic Neighbor Embedding (t-SNE) [33], among others. After applying t-ICA, we further reduced the dimensionality of the data by applying the t-SNE method to the 20 most informative t-ICA dimensions. These 20 most informative t-ICA dimensions describe the 25% of the total variance. The resulting 2D t-SNE space was clustered with the HDBSCAN algorithm [37], with a minimum cluster size of 200 and other default parameters, resulting in 133, 126, 134, 124, 119 clusters for the A2, B, C, D2, and G variants, respectively. By applying the t-SNE dimensionality reduction, less than 75% of the variance was lost.

Tunnel analysis of MD simulation. CaverAnalyst [44] was used to compute substrate entry channels for the 10 most populated HDBSCAN clusters of each HHDH variant. For each t-SNE cluster, the nearest MD snapshot was extracted with the Mdtraj software [45] for the analysis of accession tunnels, thus spanning the whole dynamical space of the enzyme. The parameters used for the tunnel search were 4 Å shell depth, 2.5 Å shell radius, clustering threshold value of 3.5 and a 1 Å minimum probe radius were used as tunnel search parameters.

Decision trees and feature importance. Decision trees are supervised learning methods and, therefore, require a set of training examples for which the output or label is known. They infer relations from training instances by asking a series of questions about the input in a tree-shaped hierarchy. For categorical data, yes/no questions can be asked regarding the presence/absence of a particular input, whereas for real-valued features, such as atomic distances, threshold values are normally used. Here, we defined as input features all possible minimum distances between residues defining the shape of the corresponding tunnel and the presence/absence of the studied tunnel as a target feature. We used a Python pipeline to standardize the input data and select the best Random Forest parameters for the classification. MD data was randomly split into a training set (80%) and test set (20%). We used Python packages Numpy [46], Pandas, Scikit-Learn [47], and Matplotlib for data manipulation, machine-learning, and visualization. Pyemma2 [43] was used for feature extraction from MD simulations, PCA, and tICA dimensionality reduction.

Formula for computing the tunnel frequency (f) for each HHDH subclass (Table 2):

$$f = \frac{\sum_{i=1}^{n} \delta_i p_i}{M} \cdot 100 \quad \delta_i \Rightarrow \begin{array}{c} 0 \; if \; tunnel \; not \; present \\ 1 \; if \; tunnel \; present \end{array}$$

where M is the total number of frames, p_i is the number of frames in the cluster and n is the number of clusters of each system.

5. Conclusions

The exploration of the conformational landscape of the different HHDH subclasses coupled to the active site tunnel calculations has indicated that the superior activity of HheG towards bulky epoxide substrates is due to the presence of some additional structural elements adjacent to the active site pocket, its higher conformational heterogeneity, and the presence of highly prevalent active site tunnels that present bottleneck radius of ca. 1.9 Å. This is unique to the G subclass, as the rest of the analyzed HHDH are conformationally more restricted and present a reduced number of narrower active site tunnels. Altogether, our study has shown how the HHDH structural dissimilarities influence their conformational landscape, thus impacting their associated active site tunnels, and in turn, their catalytic

promiscuity. By means of extensive MD simulations and CAVER analysis, this work has provided key information for rationalizing HHDH promiscuity and for further engineering.

Supplementary Materials: The following are available online at http://www.mdpi.com/2073-4344/10/12/1403/s1, Figure S1: tSNE and HDBSCAN representations, Figure S2: T1 representation as tetramer, Figures S3–S7: Random Forest Classifier for HheA2, HheB, HheC, HheD2, and HheG, Figure S8: Most important contacts computed using Random Forest Classifier.

Author Contributions: M.E.-G. has performed all simulations and analysis, M.E.-G., J.I.-F. and S.O. have discussed the results, and written the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: We thank the Generalitat de Catalunya for the emerging group CompBioLab (2017 SGR-1707) and Spanish MINECO for project PGC2018-102192-B-I00. J.I.F. was supported by the European Community for Marie Curie fellowship (H2020-MSCA-IF-2016-753045) and Juan de la Cierva-Incorporación fellowship (IJCI-2017-34129). S.O. is grateful to the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (ERC-2015-StG-679001).

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

References

1. Tokuriki, N.; Tawfik, D.S. Protein Dynamism and Evolvability. *Science* **2009**, *324*, 203–207. [CrossRef] [PubMed]
2. Tawfik, O.K.; Dan, S. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **2010**, *79*, 471–505. [CrossRef] [PubMed]
3. Campbell, E.C.; Correy, G.J.; Mabbitt, P.D.; Buckle, A.M.; Tokuriki, N.; Jackson, C.J. Laboratory evolution of protein conformational dynamics. *Curr. Opin. Struct. Biol.* **2018**, *50*, 49–57. [CrossRef] [PubMed]
4. Maria-Solano, M.A.; Serrano-Hervás, E.; Romero-Rivera, A.; Iglesias-Fernández, J.; Osuna, S. Role of conformational dynamics in the evolution of novel enzyme function. *Chem. Commun.* **2018**, *54*, 6622–6634. [CrossRef]
5. Petrović, D.; Risso, V.A.; Kamerlin, S.C.L.; Sanchez-Ruiz, J.M. Conformational dynamics and enzyme evolution. *J. R. Soc. Interface* **2018**, *15*. [CrossRef]
6. Jiménez-Osés, G.; Osuna, S.; Gao, X.; Sawaya, M.R.; Gilson, L.; Collier, S.J.; Huisman, G.W.; Yeates, T.O.; Tang, Y.; Houk, K.N. The role of distant mutations and allosteric regulation on LovD active site dynamics. *Nat. Chem. Biol.* **2014**, *10*, 431–436. [CrossRef]
7. Romero-Rivera, A.; Garcia-Borràs, M.; Osuna, S. Role of Conformational Dynamics in the Evolution of Retro-Aldolase Activity. *ACS Catal.* **2017**, *7*, 8524–8532. [CrossRef]
8. Nestl, B.M.; Hauer, B. Engineering of Flexible Loops in Enzymes. *ACS Catal.* **2014**, *4*, 3201–3211. [CrossRef]
9. Pavlova, M.; Klvana, M.; Prokop, Z.; Chaloupkova, R.; Banas, P.; Otyepka, M.; Wade, R.C.; Tsuda, M.; Nagata, Y.; Damborsky, J. Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.* **2009**, *5*, 727. [CrossRef]
10. de Vries, E.J.; Janssen, D.B. Biocatalytic conversion of epoxides. *Curr. Opin. Biotechnol.* **2003**, *14*, 414–420. [CrossRef]
11. Hasnaoui-Dijoux, G.; Majerić Elenkov, M.; Lutje Spelberg, J.H.; Hauer, B.; Janssen, D.B. Catalytic Promiscuity of Halohydrin Dehalogenase and its Application in Enantioselective Epoxide Ring Opening. *ChemBioChem* **2008**, *9*, 1048–1051. [CrossRef] [PubMed]
12. Fox, R.J.; Davis, S.C.; Mundorff, E.C.; Newman, L.M.; Gavrilovic, V.; Ma, S.K.; Chung, L.M.; Ching, C.; Tam, S.; Muley, S.; et al. Improving catalytic function by ProSAR-driven enzyme evolution. *Nat. Biotechnol.* **2007**, *25*, 338–344. [CrossRef] [PubMed]
13. Schallmey, A.; Schallmey, M. Recent advances on halohydrin dehalogenases-from enzyme identification to novel biocatalytic applications. *Appl. Microbiol. Biotechnol.* **2016**, *100*, 7827–7839. [CrossRef] [PubMed]
14. Schallmey, M.; Floor, R.J.; Hauer, B.; Breuer, M.; Jekel, P.A.; Wijma, H.J.; Dijkstra, B.W.; Janssen, D.B. Biocatalytic and Structural Properties of a Highly Engineered Halohydrin Dehalogenase. *ChemBioChem* **2013**, *14*, 870–881. [CrossRef] [PubMed]
15. Wan, N.-W.; Liu, Z.-Q.; Huang, K.; Shen, Z.-Y.; Xue, F.; Zheng, Y.-G.; Shen, Y.-C. Synthesis of ethyl (R)-4-cyano-3-hydroxybutyrate in high concentration using a novel halohydrin dehalogenase HHDH-PL from Parvibaculum lavamentivorans DS-1. *RSC Adv.* **2014**, *4*, 64027–64031. [CrossRef]

16. Assis, H.M.S.; Bull, A.T.; Hardman, D.J. Synthesis of Chiral Epihalohydrins Using Haloalcohol Dehalogenase A from Arthrobacter Erithii H10a. *Enzyme Microb. Technol.* **1998**, *22*, 545–551. [CrossRef]
17. Elenkov, M.M.; Tang, L.; Meetsma, A.; Hauer, B.; Janssen, D.B. Formation of Enantiopure 5-Substituted Oxazolidinones through Enzyme-Catalysed Kinetic Resolution of Epoxides. *Org. Lett.* **2008**, *10*, 2417–2420. [CrossRef]
18. Molinaro, C.; Guilbault, A.-A.; Kosjek, B. Resolution of 2,2-Disubstituted Epoxides via Biocatalytic Azidolysis. *Org. Lett.* **2010**, *12*, 3772–3775. [CrossRef]
19. de Jong, R.M.; Tiesinga, J.J.W.; Rozeboom, H.J.; Kalk, K.H.; Tang, L.; Janssen, D.B.; Dijkstra, B.W. Structure and mechanism of a bacterial haloalcohol dehalogenase: A new variation of the short-chain dehydrogenase/reductase fold without an NAD(P)H binding site. *EMBO J.* **2003**, *22*, 4933–4944. [CrossRef]
20. de Jong, R.M.; Tiesinga, J.J.W.; Villa, A.; Tang, L.; Janssen, D.B.; Dijkstra, B.W. Structural Basis for the Enantioselectivity of an Epoxide Ring Opening Reaction Catalyzed by Halo Alcohol Dehalogenase HheC. *J. Am. Chem. Soc.* **2005**, *127*, 13338–13343. [CrossRef]
21. Elenkov, M.M.; Hauer, B.; Janssen, D.B. Enantioselective Ring Opening of Epoxides with Cyanide Catalysed by Halohydrin Dehalogenases: A New Approach to Non-Racemic β-Hydroxy Nitriles. *Adv. Synth. Catal.* **2006**, *348*, 579–585. [CrossRef]
22. Koopmeiners, J.; Diederich, C.; Solarczek, J.; Voß, H.; Mayer, J.; Blankenfeldt, W.; Schallmey, A. HheG, a Halohydrin Dehalogenase with Activity on Cyclic Epoxides. *ACS Catal.* **2017**, *7*, 6877–6886. [CrossRef]
23. Schallmey, M.; Koopmeiners, J.; Wells, E.; Wardenga, R.; Schallmey, A. Expanding the Halohydrin Dehalogenase Enzyme Family: Identification of Novel Enzymes by Database Mining. *Appl. Environ. Microbiol.* **2014**, *80*, 7303–7315. [CrossRef] [PubMed]
24. Calderini, E.; Wessel, J.; Süss, P.; Schrepfer, P.; Wardenga, R.; Schallmey, A. Selective Ring-Opening of Di-Substituted Epoxides Catalysed by Halohydrin Dehalogenases. *ChemCatChem* **2019**, *11*, 2099–2106. [CrossRef]
25. Orozco, M. A theoretical view of protein dynamics. *Chem. Soc. Rev.* **2014**, *43*, 5051–5066. [CrossRef]
26. Osuna, S. The challenge of predicting distal active site mutations in computational enzyme design. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2020**, e1502. [CrossRef]
27. Wang, L.; Marciello, M.; Estévez-Gay, M.; Rodriguez, P.E.D.S.; Morato, Y.L.; Iglesias-Fernández, J.; Huang, X.; Osuna, S.; Filice, M.; Sanchez, S. Enzyme Conformation Influences the Performance of Lipase-powered Nanomotors. *Angew. Chem. Int. Ed.* **2020**, *59*, 21080–21087. [CrossRef]
28. Mu, Y.; Nguyen, P.H.; Stock, G. Energy landscape of a small peptide revealed by dihedral angle principal component analysis. *Proteins* **2005**, *58*, 45–52. [CrossRef]
29. Ferguson, A.L.; Panagiotopoulos, A.Z.; Kevrekidis, I.G.; Debenedetti, P.G. Nonlinear dimensionality reduction in molecular simulation: The diffusion map approach. *Chem. Phys. Lett.* **2011**, *509*, 1–11. [CrossRef]
30. Ceriotti, M.; Tribello, G.A.; Parrinello, M. Simplifying the representation of complex free-energy landscapes using sketch-map. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 13023–13028. [CrossRef]
31. Hernández, C.X.; Wayment-Steele, H.K.; Sultan, M.M.; Husic, B.E.; Pande, V.S. Variational encoding of complex dynamics. *Phys. Rev. E* **2018**, *97*, 062412. [CrossRef] [PubMed]
32. Mardt, A.; Pasquali, L.; Wu, H.; Noé, F. VAMPnets for deep learning of molecular kinetics. *Nat. Commun.* **2018**, *9*, 5. [CrossRef] [PubMed]
33. Zhou, H.; Wang, F.; Tao, P. t-Distributed Stochastic Neighbor Embedding Method with the Least Information Loss for Macromolecular Simulations. *J. Chem. Theory Comput.* **2018**, *14*, 5499–5510. [CrossRef] [PubMed]
34. Spiwok, V.; Kříž, P. Time-Lagged t-Distributed Stochastic Neighbor Embedding (t-SNE) of Molecular Simulation Trajectories. *Front. Mol. Biosci.* **2020**, *7*. [CrossRef] [PubMed]
35. Pérez-Hernández, G.; Paul, F.; Giorgino, T.; De Fabritiis, G.; Noé, F. Identification of slow molecular order parameters for Markov model construction. *J. Chem. Phys.* **2013**, *139*, 015102. [CrossRef]
36. Case, D.A.; Darden, T.A.; Cheatham, T.E.; Simmerling, C.L.; Wang, J.; Duke, R.E.; Luo, R.; Crowley, M.; Walker, R.C.; Zhang, W.; et al. *AMBER 16, University of California, San Francisco*; United States of America: Washington, DC, USA, 2016.
37. Campello, R.J.G.B.; Moulavi, D.; Sander, J. *Density-Based Clustering Based on Hierarchical Density Estimates*; Springer: Berlin/Heidelberg, Germany, 2013; pp. 160–172.

38. Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; et al. CAVER 3.0: A tool for the analysis of transport pathways in dynamic protein structures. *PLoS Comput. Biol.* **2012**, *8*, e1002708. [CrossRef]
39. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
40. Anandakrishnan, R.; Aguilar, B.; Onufriev, A.V. H++ 3.0: Automating pK prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations. *Nucleic Acids Res.* **2012**, *40*, W537–W541. [CrossRef]
41. Jorgensen, W.L.; Chandrasekhar, J.; Madura, J.D.; Impey, R.W.; Klein, M.L. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* **1983**, *79*, 926–935. [CrossRef]
42. Maier, J.A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K.E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713. [CrossRef]
43. Scherer, M.K.; Trendelkamp-Schroer, B.; Paul, F.; Pérez-Hernández, G.; Hoffmann, M.; Plattner, N.; Wehmeyer, C.; Prinz, J.-H.; Noé, F. PyEMMA 2: A Software Package for Estimation, Validation, and Analysis of Markov Models. *J. Chem. Theory Comput.* **2015**, *11*, 5525–5542. [CrossRef] [PubMed]
44. Jurcik, A.; Bednar, D.; Byska, J.; Marques, S.M.; Furmanova, K.; Daniel, L.; Kokkonen, P.; Brezovsky, J.; Strnad, O.; Stourac, J.; et al. CAVER Analyst 2.0: Analysis and visualization of channels and tunnels in protein structures and molecular dynamics trajectories. *Bioinformatics* **2018**, *34*, 3586–3588. [CrossRef] [PubMed]
45. McGibbon, R.T.; Beauchamp, K.A.; Harrigan, M.P.; Klein, C.; Swails, J.M.; Hernández, C.X.; Schwantes, C.R.; Wang, L.-P.; Lane, T.J.; Pande, V.S. MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophys. J.* **2015**, *109*, 1528–1532. [CrossRef] [PubMed]
46. Harris, C.R.; Millman, K.J.; van der Walt, S.J.; Gommers, R.; Virtanen, P.; Cournapeau, D.; Wieser, E.; Taylor, J.; Berg, S.; Smith, N.J.; et al. Array programming with NumPy. *Nature* **2020**, *585*, 357–362. [CrossRef] [PubMed]
47. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.