

Supporting Information

Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization

Pavel Vanacek^{1,2#}, Eva Sebestova^{1#}, Petra Babkova^{1,2}, Sarka Bidmanova^{1,2}, Lukas Daniel^{1,2}, Pavel Dvorak¹, Veronika Stepankova^{1,2,3}, Radka Chaloupkova^{1,2,3}, Jan Brezovsky^{1,2}, Zbynek Prokop^{1,2,3*}, Jiri Damborsky^{1,2*}

Affiliations

¹ Loschmidt Laboratories, Department of Experimental Biology and Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, 625 00 Brno, Czech Republic

² International Clinical Research Center, St. Anne's University Hospital, Pekarska 53, 656 91 Brno, Czech Republic

³ Enantis Ltd., Biotechnology Incubator INBIT, Kamenice 34, 625 00 Brno, Czech Republic

authors contributed equally; * authors for correspondence: Zbynek Prokop, zbynek@chemi.muni.cz, phone +420-5-4949-6667; Jiri Damborsky, jiri@chemi.muni.cz, phone +420-5-4949-3467

Table of Contents

Supporting Methods

Supporting Tables

Table S1: List of known HLDs used for definition of HLD clusters

Table S2: Catalytic residues in putative degenerated HLD sequences

Table S3: Annotations of putative HLD sequences

Table S4: List of the chemical formulas of 34 halogenated compounds used for screening

Table S5: Overview of the selection process with respective criteria

Table S6: Accession numbers of putative HLDs

Table S7: Overview of expression analysis of 20 putative HLDs

Table S8: Robotic screening of enzymatic activity

Table S9: The parameters related to size, shape and oligomeric state of HLDs

Table S10: Melting temperatures of HLDs by CD spectroscopy and robotic DSC

Table S11: Specific activities of HLDs towards 30 halogenated substrates

Table S12: Steady-state kinetic parameters of HLDs

Table S13: Enantioselectivities of HLDs towards 2-bromopentane and ethyl-2-bromopropionate

Table S14: Activities towards of HLDs warfare agents and environmental pollutants

Supporting Figures

Figure S1: The representative 3D structure of HLD

Figure S2: Availability of homology modelling templates in HLD subfamilies

Figure S3: Distributions of predicted volumes of catalytic pockets in HLD subfamilies

Figure S4: Distributions of predicted bottleneck radii of p1, p2 and p3 tunnels in HLD subfamilies

Figure S5: Distributions of mechanism-based geometric criteria for reactivity in HLD subfamilies

Figure S6: Expression analysis of the set of 20 putative HLDs

Figure S7: Far-UV circular dichroism spectra of HLDs

Figure S8: Temperature profiles

Figure S9: pH profiles

Figure S10: The set of thirty halogenated substrates

Figure S11: Kinetic resolution of 2-bromopentane

Figure S12: Kinetic resolution of ethyl 2-bromopropionate

Supporting References

Supporting Methods

***In silico* screening.**

Identification of sequences.

Sequence database search. The sequences of three experimentally characterized HLDs, LinB (NCBI accession number BAA03443), Dh1A (P22643) and DrbA (NP_869327), and a putative HLD from *Aspergillus niger* (EHA28085, residues 90-432), were used as queries for two iterations of PSI-BLAST v2.2.28+¹ searches against the NCBI nr database (version 25-9-2013)², which was downloaded from the NCBI FTP site. PSI-BLAST searching was performed with *E*-value thresholds of 10^{-20} for both the initial BLAST search and inclusion of the sequence in the position specific matrix. In order to recognize the target HLD domain, all sequences identified were aligned with queries using the Needleman-Wunsch global pairwise alignment algorithm³ implemented in the USEARCH v.6.0.307 program⁴. Each sequence identified was then shortened so that it did not exceed any query by more than 20 amino acids at either end.

Sequence clustering. The sequence identity of each pair of shortened sequences in the dataset was derived from their pairwise global alignment constructed using the USEARCH implementation of the Needleman-Wunsch algorithm. The pairwise distance of each pair of sequences was then calculated as $1-i$, where i is the sequence identity of a given sequence pair. The average-link hierarchical clustering implemented in the MC-UPGMA program⁵ was used to construct a tree hierarchy of sequences based on their pairwise distances. The tree was cut at varying cut-offs, and the cut-offs at which all 22 known HLD sequences (**Table S1**) were grouped into four clusters, *i.e.* subfamilies HLD-I, HLD-II and HLD-III and a cluster consisting of the putative HLD from *A. niger* (“HLD-IIIb”), were further analyzed. For each such cut-off c , the difference between the total number of sequences comprising HLD clusters at the cut-off $c+1$ and the number at the given cut-off c was calculated. The cut-off value of 74, *i.e.* the highest cut-off value among all those values giving the smallest difference in the overall sequence count, was selected as the final cut-off for cutting the hierarchical sequence tree.

Multiple sequence alignment. A multiple sequence alignment (MSA) of all putative full-length HLD sequences was constructed by Clustal Omega v1.2.0⁶. First, each of the four HLD clusters was aligned individually, then profile-profile alignment of the two closest clusters was constructed in a stepwise manner until all clusters were aligned. All artificial sequences (*i.e.* sequences containing the term artificial, synthetic construct, vector, vaccinia virus, plasmid, halotag or replicon in the header line), as well as incomplete or degenerate sequences, were removed from the dataset. A sequence was classified as incomplete or degenerate if it had an incomplete catalytic pentad due to, respectively, deletion or substitution. The criteria for a complete catalytic pentad were as follows: (i) Asp in the position of the

nucleophile (the position corresponding to Asp108 of LinB), (ii) Asp or Glu in at least one of the catalytic acid positions (those corresponding to Asp260 of Dh1A or Glu132 of LinB), (iii) His in the position of the base (position His272 of LinB) and (iv) Asn, Gln, Trp, Tyr or His in at least two halide-binding positions (position Asn38 or Trp109 of LinB, or Trp175 of Dh1A). A new MSA comprising only sequences with a complete catalytic pentad was then constructed in the same way as described above, *i.e.* by producing the cluster-specific MSAs followed by step-wise profile-profile alignments. The newly constructed alignment was again checked for incomplete and degenerate sequences and after they had been removed the final HLD dataset was created.

Homology modelling.

Identification of potential templates. Those HLD sequences with known 3D structures (“PDB sequences”), which were identified based on the description in the sequence header line, were converted into BLAST database format using the makeblastdb program from the BLAST+ v2.2.28+ package. To prevent problems arising in the construction of homology models, the N- and C-terminal ends of long sequences (e.g. sequences with multiple domains) were removed in the following way. In the final MSA, the consensus beginning and end positions of the alignment were assigned as the first and last positions of the MSA where at least 50% of PDB sequences had a non-gap character. This consensus beginning and end of the alignment were then shifted by 20 positions towards the N- and C-terminal ends, respectively, and all positions outside the specified range were removed from the MSA. For each sequence, the information about the number of amino acids cut off from its N- and C-terminal ends was stored. Each sequence from the truncated alignment was then used as a query for one iteration of PSI-BLAST searching against the BLAST database prepared as described above. The search was performed with an E -value threshold of 10^{-3} and all hits with an E -value of 0 were considered to be potential templates for homology modelling of a given sequence. If no such hit existed, the best hit, and all hits with an E -value such that $E\text{-value}_{\text{best hit}}/E\text{-value} \geq 10^{-15}$, were selected as potential templates for a given sequence. All potential template structures were downloaded from the RCSB PDB database⁷.

Selection of templates. For each structure, information about its resolution and R_{Free} factor were extracted from the mmCIF file. Structures with $R_{Free} > 0.4$ or a missing R_{Free} value were considered to be unreliable and removed from the list of potential templates. Each template was penalized based on its resolution: $P_{\text{resolution}} = \text{resolution} \times 20$. Modeller v9.11⁸ was then used to construct a pairwise alignment of each individual target sequence with each of its potential templates and the alignment was penalized in the following way. First, we checked whether each catalytic pentad residue in the target was aligned with the corresponding catalytic pentad residue of the template. If so, and if the residues in the template and target were found to be identical, a given residue was subsequently treated as fixed during the homology modelling procedure. If the positions corresponded, but the residues were not identical, a given residue was penalized by $P_{\text{substituted}} = 30$. If the position of some catalytic pentad

residues did not correspond (e.g. the location of one halide-binding residue was different in HLD-I compared to the other two subfamilies), a given residue was treated as missing and penalized by $P_{\text{missing}} = 100$. If the missing residue was a nucleophile, base or catalytic acid (e.g. the location of the catalytic acid was different in HLD-II compared to the other two subfamilies), a template was removed from the list of potential templates for a given sequence. Note that if the target sequence contained two catalytic acids and/or three halide-stabilizing residues, correspondence for only, respectively, one catalytic acid and/or two halide-stabilizing residues was required. To penalize deletions and insertions, only that part of the alignment within the range specified by the first and last residues of the target sequence was evaluated. All gap characters found in the first 20 or the last 20 positions of the evaluated alignment were classified as “outer gap” characters. Each outer gap character was penalized by $P_{\text{o_short}} = 1$. In the case of gaps longer than 9 gap characters, the tenth consecutive outer gap character and any other consecutive outer gap character in a row were each penalized by $P_{\text{o_long}} = 2$. All other gap characters found in the alignment being evaluated were classified as “inner gap” characters. Each inner gap character which was part of a gap of up to 4 gap characters in length was penalized by $P_{\text{i_short}} = 5$. Each inner gap character which was part of a gap longer than 4 gap characters was penalized by $P_{\text{o_long}} = 15$. The overall template score of each potential template was then calculated as $S = P_{\text{resolution}} + \sum P_{\text{substituted}} + \sum P_{\text{missing}} + \sum P_{\text{o_short}} + \sum P_{\text{o_long}} + \sum P_{\text{i_short}} + \sum P_{\text{i_long}}$ and for each sequence, the structure with the lowest score was selected as a template for homology modelling. The penalty scores were set in a way to strongly penalize all unreliable resolution of template, particularly insertions and deletions.

Model construction. Homology models were constructed only for sequences with a template score ≤ 100 . The homology modelling was performed using Modeller with the following non-default settings: slow schedule and a maximum of 300 iterations for the model optimization step, by the variable target function method with conjugate gradients and slow schedule for molecular dynamics with simulated annealing to refine the model. The whole optimization cycle was repeated twice. As mentioned above, identical catalytic pentad residues were treated as fixed, *i.e.* they were not refined, during the optimization procedure.

Theoretical characterization.

Sequence characteristics. For each putative HLD, GI numbers, accession numbers, and annotations were extracted from the sequence header line. Information about the catalytic pentad residues was extracted from the MSA and subfamily membership was assigned based on the results of clustering. The Needleman–Wunsch algorithm implemented in the USEARCH program was used to calculate pairwise global alignment for each pair of full-length sequences from the final dataset, and for each sequence, the most closely related known HLD, as well as closely related proteins (*i.e.* those with $\geq 90\%$ sequence identity to a given protein), were identified. Information about the source organisms of all putative HLDs was collected from the NCBI Taxonomy and Bioproject databases² downloaded from the NCBI FTP

site. The chance for soluble expression in *E. coli* of each protein was predicted based on the revised Wilkinson-Harrison solubility model^{9,10}. Knowledge of the number of residues removed during the creation of the truncated MSA (described above) was used to predict whether a given sequence might possibly contain multiple domains (*i.e.* > 60 removed residues). Information about the domain composition of individual proteins was obtained by sequence searches of Pfam database¹¹ via the Pfam RESTful interface. Potential transmembrane regions in individual sequences were predicted using TOPCONS¹².

Structural characteristics. All the homology models constructed were superimposed on chain A of the DmmA structure (PDB-ID 3U1T) using the jCE v.20130412 program¹¹. Pockets in each homology model were calculated and measured using the CASTp program^{13,14} with a probe radius of 1.4 Å. For each pocket identified, the total number of catalytic atoms (*i.e.* atoms in a catalytic pentad residue) was counted, as was as the number of the following selected catalytic atoms: CG, OD1 and OD2 atoms of the nucleophile and NE1 and CD1 atoms of the halide-stabilizing Trp (the position following the nucleophile, *i.e.* the position corresponding to Trp109 of LinB). The pocket containing the greatest number of selected catalytic atoms was designated the catalytic pocket. In the case of two or more pockets having equal numbers, the number of all catalytic atoms forming a given pocket and the volume of the pocket volume were used as the second and the third criterion respectively. The CAVER v. 3.01 program¹⁵ was then used to calculate tunnels in the ensemble of all homology models. The initial calculation starting point was placed at the center of gravity of the CG atom of the nucleophile and the NE1 atom of the halide-stabilizing Trp (the position following the nucleophile) and it was allowed to be moved for a maximum of 5 Å during the optimization procedure. The calculation was conducted using both the shell probe and a shell depth of 4 Å to define the protein surface. The tunnel search was performed using a probe radius of 1 Å. The probe size was defined based on our previous practical experience with analysis of tunnels in native proteins, taking into account protein dynamics. The tunnels identified in all homology models were clustered according to the pairwise distances of tunnels using a clustering threshold of 8.

Molecular docking. The three-dimensional structures of 34 halogenated compounds (**Table S1**), which are environmental pollutants, artificial sweeteners, chemical warfare agents or their surrogates and disinfectants, were constructed in Avogadro¹⁶. The partial atomic charges on them were derived by the R.E.D. server web tool¹⁷. Input geometries were optimized by the Gaussian 2009 D.01 program interfaced with this server and a multi-orientation RESP fit was performed with the RESP-A1A charge model. The output files, which were in Sybyl mol2 format, were converted into AutoDock4.0¹⁸ compliant pdbqt format by MGLTools¹⁸. Hydrogen atoms and Gasteiger charges were subsequently assigned to the homology models by MGLTools. The catalytic pockets of superimposed homology models were selected as targets for molecular docking using AutoDock 4.2.3. The region was

represented by a set of atomic and electrostatic maps calculated by AutoGrid4.0¹⁸. The grid maps, which were of dimensions 20.0 Å × 21.5 Å × 21.5 Å, and were centred at the catalytic nucleophile so as to cover the whole active site and the access tunnels. The energy of the unbound system was estimated as the internal energy of the unbound extended conformation determined from a Lamarckian Genetic Algorithm search. 250 docking calculations were performed for each compound using the following parameters for the genetic algorithm: initial population size 300, maximum of 27,000 generations, elitism value 1, mutation rate 0.02 and cross-over rate 0.8. The maximum number of energy evaluations was set to $(1.5 \cdot 105 \cdot N_{\text{tor}}^2) + 1.5 \cdot 10^6$, where N_{tor} is the number of torsional degrees of freedom of a docked compound. Each local search was based on the pseudo Solis and Wets algorithm with a maximum of 300 iterations per search¹⁹. Final orientations resulting from every docking were clustered by setting the tolerance for the root-mean-square positional deviation at 2 Å.

Reactivity estimation. The largest cluster of each docked compound was evaluated for proper positioning of the halogen atom between the two halogen-stabilizing residues using a cutoff distance of 4.5 Å. Compounds with the halogen properly stabilized were investigated for possible S_N2 displacement reactions. The —COO— · · · C—Cl reactive distance (R_D) and —COO— · · · C · · · Cl reactive angle (R_A) parameters were recorded. The cut-off values selected for reactivity were $R_D < 3.3$ Å and $R_A > 140^\circ$ ²⁰.

Expression analysis and activity screening.

Gene synthesis and subcloning. Codon-optimized genes encoding 20 putative HLDs were designed and commercially synthesized (Bio Basic Inc, Canada). The synthetic genes were subcloned individually into the expression vector pET21b (Novagen, USA) between the *NdeI/XhoI* restriction sites.

Gene overexpression.

Overexpression in E. coli BL21(DE3) Initially, *E. coli* BL21(DE3) competent cells were transformed with DNA constructs using the heat-shock method, plated on agar plates with ampicillin and grown overnight at 37 °C. Single colonies were used to inoculate 10 ml of lysogeny broth (LB) medium (Sigma-Aldrich, USA) with ampicillin, and cells were grown overnight at 37 °C. Each overnight culture was used to inoculate 1000 ml of LB medium containing ampicillin. Cells were cultivated at 37 °C until an optical density (OD) of 0.4-0.6 at 600 nm was attained. Overexpression was induced with isopropyl β-D-1-thiogalactopyranoside (IPTG) (Sigma-Aldrich, USA) at a final concentration of 0.5 mM. Cells were then cultivated overnight at 20 °C. Biomass was harvested at the end of the cultivation, washed with purification buffer A (20 mM di-potassium hydrogen phosphate and potassium dihydrogen phosphate, pH 7.5, 0.5 M NaCl, 10 mM imidazole). In order to avoid aggregation of DmsaA, the protein was washed with 50 mM Tris-SO₄ buffer, pH 8. The harvested biomass was frozen at -80 °C. DNase I (New England Biolabs, USA) was added to a final concentration of 1.25 μg.ml⁻¹ of cell suspension.

Cells in suspension were disrupted using a One Shot model cell homogenizer (Constant System Ltd., United Kingdom). The cell lysate was centrifuged for 1 hour at 21,000 x g. The crude extract was decanted and total protein concentration was determined by the method of Bradford (Sigma-Aldrich, USA). The expression levels of particular proteins were estimated from the trace densities of corresponding bands determined using a calibrated densitometer GS-800 (Bio-Rad Laboratories Inc., USA). Molecular weights were verified by using Unstained Protein Molecular Weight Marker (Thermo Scientific, USA).

Overexpression in E. coli ArcticExpress(DE3) cells. *E. coli* ArcticExpress(DE3) cells were transformed with 50 ng DNA constructs using heat-shock, plated on agar plates with ampicillin and grown overnight at 37 °C. Single colonies were used to inoculate 10 ml of LB medium with ampicillin and gentamycin. Cells were subsequently grown overnight at 37 °C. Overnight culture (1 ml) was used to inoculate 50 ml of LB medium containing no antibiotics. Cells were incubated at 30 °C. After 3 hours of incubation, the culture was equilibrated by 10 min incubation at 12 °C. The expression was induced with IPTG to a final concentration of 0.5 mM. Cells were then cultivated at 12 °C for 24 hours. Sampling, harvesting, and disruption of biomass were performed by the same procedure as described above.

Overexpression in E. coli BL21(DE3) under modified conditions. *E. coli* BL21(DE3) competent cells were transformed with DNA constructs using heat-shock, plated on agar plates with ampicillin and grown overnight at 37 °C. Single colonies were used to inoculate 10 ml of LB medium with ampicillin and cells were grown overnight at 37 °C. The overnight culture was used to inoculate 1000 ml of LB medium with ampicillin. Cells were cultivated at 37 °C until OD of 0.4-0.6 at 600 nm. The overexpression was induced with IPTG to a final concentration of 0.1 mM. Cells were then cultivated for 4 hours at 30 °C. Sampling, harvesting, and disruption of biomass were performed by the same procedure as described above.

Overexpression in E. coli Rosetta-gami B(DE3) pLysS. *E. coli* Rosetta-gami B(DE3) pLysS cells were transformed with DNA constructs. Transformants were cultured in 50 ml of LB medium containing antibiotics (tetracycline, kanamycin, chloramphenicol and ampicillin) at 37 °C overnight. Fresh 1000 ml of pre-warmed LB medium containing the four above-mentioned antibiotics, was then inoculated with 50 ml of the overnight culture and grown at 37 °C until the OD_{600 nm} reached 0.5 after which the culture was cooled on ice for 30 min. Inductor IPTG was then added to a final concentration of 1 mM, and the culture was grown at 20 °C for overnight. Sampling, harvesting, and disruption of biomass were performed by the same procedure as described above.

Overexpression in E. coli BL21(DE3) (Enbase medium). The cultivations were performed according to manufacturer's instructions (Biosilta, Finland). Inoculum was prepared by picking single colonies of transformed *E. coli* carrying the gene coding a target protein to 1 ml of LB medium with 2 g.l⁻¹ of glucose and shaken for 6 hours at 37 °C. Prior to the cultivation, white-bag tablets were dissolved in 50 ml of sterile water in Ultra Yield Flask. The cultivation was initialized by addition of respective antibiotics (ampicillin; gentamycin, and tetracycline), 0.5 mL of pre-culture, 25 µl of Reagent A (glucoamylase) and 100 µl.l⁻¹ of antifoam agent Struktol SB2020 (Sigma-Aldrich, USA). The culture was grown for 18 hours before induction at 30 °C. At induction time, black-bag booster tablet, 75 µl of Reagent A and 0.5 mM final IPTG was added to the shake flasks. The induced cultures were incubated for 24 hours in 30 °C. Sampling, harvesting, and disruption of biomass were performed by the same procedure as described above.

Activity screening.

Robotic activity screening. Enzyme activity towards 1-iodobutane, 1,2-dibromoethane, and 4-bromobutyronitrile substrates was robotically screened using a Hamilton MICROLAB STARlet robot (Hamilton Robotics, Switzerland). The reactions were performed in 2 ml glass vials (VWR, USA) containing 1 ml of 100 mM glycine buffer, pH 8.6 and 1 µl of the halogenated substrate at 10, 37 and 55 °C. The reaction was initiated by addition of the enzyme in the form of a cell-free extract. The progress of the reaction was monitored by periodically withdrawing samples from the reaction mixture and immediately mixing these samples with 35% (v/v) nitric acid to terminate the reaction. The release of the halide ion product was analyzed spectrophotometrically using the end-point assay developed by Iwasaki and co-workers²¹. Dehalogenation activities were quantified as the rate of product formation over time.

Protein purification.

Affinity protein purification. Expressed histidine-tagged recombinant proteins were purified using single step nickel affinity chromatography. Cell-free extract was applied on a Ni-nitrilotriacetic acid (Ni-NTA) Superflow column (5 ml) charged with Ni²⁺ ions (Qiagen, Germany) in equilibrating buffer (20 mM potassium phosphate buffer, pH 7.5, containing 10 mM imidazole, and 0.5 M sodium chloride). The target protein was eluted with a buffer containing 300 mM imidazole. The eluted protein was dialyzed overnight against 50 mM phosphate buffer, pH 7.5 at 4 °C. In order to avoid aggregation of DmsaA, the protein was purified in 50 mM Tris-SO₄ buffer, pH 8, containing 20% glycerol and appropriate concentration of imidazole. Purified protein was dialyzed against 100 mM glycine buffer, pH 8.6 at 4 °C. The purity of the protein was checked by SDS-PAGE on a 15% polyacrylamide gel stained with Coomassie Brilliant Blue R-250 dye (Fluka, Switzerland). The molecular weight was

verified using Unstained Protein Molecular Weight Marker (Thermo Scientific, USA). Finally, all the purified enzymes were lyophilized using an Alpha 1-2 LD (Martin Christ, Germany) freeze dryer.

Biochemical and biophysical characterization.

Secondary structure analysis. Circular dichroism (CD) spectra were recorded at room temperature using a Chirascan CD Spectrometer (Applied Photophysics, UK) equipped with a Peltier thermostat (Applied Photophysics, UK). Data were collected from 185 to 260 nm, at $100 \text{ nm}\cdot\text{min}^{-1}$, with 1-s response time and 1-nm bandwidth, using a 0.1-cm quartz cuvette containing the studied enzymes. Each spectrum shown is the average of five individual scans and has been corrected for the buffer's absorbance. Collected CD data were expressed in terms of the mean residue ellipticity ($[\Theta\text{MRE}]^{22}$).

Quaternary structure analysis. Size exclusion chromatography with static light scattering, refractive index, ultraviolet and differential viscometer detectors was used to analyze protein quaternary structure using the Viscotec 305 TDA instrument (Malvern, UK) and Zenix SEC-300 column (Sepax Technologies, USA). The device was equilibrated with 50 mM phosphate buffer, pH 7.5. The system was calibrated using thyroglobulin as a protein standard. Proteins were injected onto the column and separated at a constant flow rate of $0.5 \text{ ml}\cdot\text{min}^{-1}$ with an elution buffer consisting of 50 mM phosphate buffer, pH 7.5. Retention volumes, molecular weights, hydrodynamic radius, and intrinsic viscosities were evaluated by the OmniSec software package (Malvern, UK).

Thermostability. Thermal unfolding was analyzed by two methods (i) circular dichroism spectroscopy and (ii) robotic differential scanning calorimetry. In the first method, the unfolding was monitored by change in the ellipticity at 222 nm over the temperature range 20 to 90 °C, with a resolution of 0.1 °C and a heating rate of $1 \text{ }^\circ\text{C}\cdot\text{min}^{-1}$. The thermal denaturation curves recorded were roughly normalized such that signal changes were represented in a range between approximately 1 and 0 and fitted to sigmoidal curves using Origin8 software (OriginLab, USA). Melting temperatures (T_m) were evaluated from the collected data as the midpoints of the normalized thermal transitions. In the second method, thermal unfolding of a $1 \text{ mg}\cdot\text{ml}^{-1}$ enzyme solution was followed by monitoring the heat capacity using a MicroCal VP-Capillary DSC (GE Healthcare Life Science, USA) equipped with an integrated autosampler (GE Healthcare Life Science, USA). Measurements were performed at temperatures from 20 to 80 °C with a heating rate of $1 \text{ }^\circ\text{C}\cdot\text{min}^{-1}$. The melting temperatures (T_m) were evaluated by baseline subtraction and concentration normalization using the software package Origin 8.0 (OriginLab, USA).

Temperature profile. The effect of temperature on enzymatic activity was determined towards 1,3-diiodopropane at pH 8.6. The activity measurement was performed in triplicates at the temperature ranging from 5 to 80 °C by the robot Hamilton MICROLAB STARlet (Hamilton Robotics, Switzerland).

pH profile. The effect of pH on enzymatic activity was determined towards 1,3-dibromopropane at the pH ranging from 4 to 12. The activity measurement was performed in triplicates at 10 °C (DpaA, DpaB,

and DgpA), 37 °C (DmsaA, DsxA, DtaA, and DsbA) or 55 °C (DhmeA) by the robotic Hamilton MICROLAB STARlet (Hamilton Robotics, Switzerland).

Substrate specificity. Enzymatic activity towards a set of 30 halogenated substrates was measured by the robotic Hamilton MICROLAB STARlet (Hamilton Robotics, Switzerland). All liquid handling steps were performed with a robotic liquid handling system. The substrate specificity was analyzed at either 10, 37, or 55 °C as described above (Robotic screening of enzymatic activity). Abiotic control was measured with all tested substrates at temperatures above 35 °C and subtracted from the activity data.

Principal component analysis. The specific activity data set for 13 previously and 8 newly examined HLDs towards 30 substrates was analyzed by Principal Component Analysis (PCA)²³ to reveal the relationships among individual HLDs (objects) based on their activities towards the set of halogenated substrates (variables) as described elsewhere²⁴. Statistical analysis was conducted using the Statistica 12.0 software package (StatSoft, USA). Two PCAs were constructed to visualize systematic trends in the dataset. In the first, the raw dataset of the specific activities of individual enzymes towards particular substrates was used as the primary input data. In the second, the specific activities were log-transformed and weighted prior to PCA, giving a measure of the activity for a specific substrate and enzyme combination relative to the sum of the enzyme's activities for all the other substrates considered.

Steady-state kinetics. The steady-state kinetics of the novel HLDs towards 1,2-dibromoethane were measured using a VP-ITC isothermal titration calorimeter (MicroCal, USA) at either 10, 37, or 55 °C. The enzyme DsxA was additionally assayed towards 1,3-dibromopropane at 37 °C. Substrate concentration was verified using a gas chromatograph (Trace 2000 Thermo Finnigan, USA) equipped with MS detection (EM640 GC-MS, Bruker, USA). The enzyme was titrated with increasing amounts of the substrate at 150-s intervals in the reaction mixture vessel, while pseudo-first-order conditions were maintained. Every injection increased the substrate concentration, leading to a further increase in the enzyme reaction rate until the enzymatic reaction was saturated. A total of 28 injections were carried out during the titration. The reaction rates reached after each injection (expressed in units of thermal power) were recalculated to determine enzyme turnover. The calculated enzyme turnover plotted against the actual concentration of the substrate after each injection was then fitted to kinetic models by nonlinear regression using Origin 8.0 (OriginLab, USA).

Enantioselectivity. Kinetic resolution experiments were performed at 20 °C. The reaction mixture consisted of 1 ml glycine buffer (100 mM, pH 8.6) and 1 µl of a racemic mixture of 2-bromopentane or ethyl 2-bromopropionate. The enzymatic reaction was initiated by addition of enzyme solution. The enzymatic conversion of the racemate mixture was analyzed using gas chromatography with flame ionization detection (Agilent 7890, Agilent Technologies, USA). The racemic mixture was separated

using a chiral capillary column, Astec Chiraldex B-DM (50 m x 0.25 mm x 0.12 μ l film thickness, Sigma-Aldrich, USA). The enantiomeric ratio (*E*-value), defined as the ratio between the specificity constants ($k_{\text{cat}}/K_{\text{m}}$) for the two enantiomers, was calculated using the following equation:

$$E = \frac{k_{\text{cat}}^{\text{R}}/K_{\text{m}}^{\text{R}}}{k_{\text{cat}}^{\text{S}}/K_{\text{m}}^{\text{S}}}$$

where k_{cat} and K_{m} represent the Henri-Michaelis-Menten parameters of the two enantiomers. Kinetic parameters were derived by fitting the reaction progress curves obtained from kinetic resolution experiments onto the competitive Henri-Michaelis-Menten kinetics²⁵ by employing nonlinear least-squares regression using the software DynaFit (BioKin, Ltd., USA)²⁶.

Activity towards warfare agents and environmental pollutants. Enzymatic activity towards chemical warfare agent bis(2-chloroethyl)sulfide (usually called sulfur mustard) was measured using an in-house fluorescent assay based on the ion-paired pH indicator 8-hydroxypyrene-1,3,6-trisulfonic acid. The reaction mixture was composed of 50 μ l enzyme dialyzed in 1 mM HEPES buffer, pH 8.2, 10 μ l HPTS-IP (2 $\mu\text{g}\cdot\text{ml}^{-1}$), 40 μ l substrate in 50% DMSO with eight replicates. The final concentration of sulfur mustard in the reaction mixture was determined using a gas chromatograph (GC Trace 2000 Thermo Finnigan, USA) coupled with a mass spectrometer (EM640 GC-MS, Bruker, USA). Increasing concentrations of reaction products (protons) were monitored as changes in fluorescence of HPTS-IP over time at 25 $^{\circ}\text{C}$. The fluorescence was read at 520 nm after excitation at 410 and 485 nm using a fluorescence reader FLUOstar OPTIMA (BMG LABTECH, Germany). The specific activities of individual enzymes were determined by estimation of the initial slope from non-linear data using the software package Origin 6.1 (OriginLab, USA). The degradation of a selected environmental pollutants, 1,3-dichloropropene, γ -hexachlorocyclohexane, hexabromocyclododecane was analyzed using GC-MS with automatized robotic arm. The reactions were performed in 2 ml glass vials fitted with a magnetic screw cap (VWR, USA) containing 1 ml of 100 mM glycine buffer, pH 8.6 and 1 μ l of the halogenated substrate at 10, 37 and 55 $^{\circ}\text{C}$. The enzymatic reaction was initiated by the addition of the enzyme. The consumption of substrates was quantified using gas chromatograph Trace 1300 (Thermo Scientific, USA) equipped with capillary column TG-SQC, 30m x 0.25mm x 0.25 μm (Thermo Scientific, USA) and connected with mass spectrometer ISQTM LT Single Quadrupole (Thermo Scientific, USA). The initial rate was analyzed by non-linear regression of time course of substrate depletion time course using software Origin 8.0 (OriginLab, USA).

Supporting Tables

Table S1: List of known HLDs used for definition of subfamily clusters.

| NCBI accession | Protein | Subfamily | Organism | Reference |
|----------------|--------------|-----------|---|-------------|
| P22643 | DhlA | HLD-I | <i>Xanthobacter autotrophicus</i> | 27 |
| WP_003872427 | DhmA | HLD-I | <i>Mycobacterium avium</i> complex | 28 |
| NP_216812 | DmbB | HLD-I | <i>Mycobacterium tuberculosis</i> H37Rv | 29 |
| WP_011513586 | DpcA | HLD-I | <i>Psychrobacter cryohalolentis</i> | 30 |
| WP_006972606 | DppA | HLD-I | <i>Plesiocystis pacifica</i> | 31 |
| Q8U671 | DatA | HLD-II | <i>Agrobacterium fabrum</i> C58 | 32 |
| BAJ23986 | DbeA | HLD-II | <i>Bradyrhizobium elkanii</i> USDA 94 | 33 |
| NP_767727 | DbjA | HLD-II | <i>Bradyrhizobium diazoefficiens</i> USDA 110 | 34 |
| P0A3G2 | DhaA | HLD-II | <i>Rhodococcus rhodochrous</i> | 35 |
| WP_003413363 | DmbA | HLD-II | <i>Mycobacterium tuberculosis</i> complex | 29 |
| Q98C03 | DmlA | HLD-II | <i>Mesorhizobium loti</i> MAFF303099 | 34 |
| 3U1T_A | DmmA | HLD-II | unidentified - marine microbial consortium | 36 |
| AAL17946 | DmsA | HLD-II | <i>Mycobacterium smegmatis</i> MC2 155 | unpublished |
| WP_007349233 | DmxA | HLD-II | <i>Marinobacter</i> sp. ELB17 | unpublished |
| XP_794172 | DpsA | HLD-II | <i>Strongylocentrotus purpuratus</i> | 37 |
| ABD55537 | Jann2620 | HLD-II | <i>Jannaschia</i> sp. CCS1 | 38 |
| BAA03443 | LinB | HLD-II | <i>Sphingobium japonicum</i> | 39 |
| WP_010986202 | Sav4779 | HLD-II | <i>Streptomyces avermitilis</i> | 38 |
| WP_011399283 | DhcA | HLD-III | <i>Hahella chejuensis</i> | unpublished |
| NP_216349 | DmbC | HLD-III | <i>Mycobacterium tuberculosis</i> H37Rv | 40 |
| NP_869327 | DrbA | HLD-III | <i>Rhodopirellula baltica</i> SH 1 | 40 |
| EHA28085 | putative HLD | HLD-IIIb | <i>Aspergillus niger</i> ATCC 1015 | 37 |

Table S2: Catalytic residues in putative degenerated HLD sequences.

| Nucleophile | Acid | Base | Halide-stabilizing 1 | Halide-stabilizing 2 | Nb. of sequences |
|--------------------|-------------|-------------|-----------------------------|-----------------------------|-------------------------|
| yes | yes | yes | yes | no | 110 |
| yes | yes | yes | no | no | 28 |
| no | yes | yes | yes | yes | 9 |
| no | yes | yes | yes | no | 14 |
| yes | yes | no | yes | yes | 7 |
| yes | yes | no | yes | no | 2 |
| yes | no | no | yes | yes | 2 |
| no | yes | no | yes | no | 6 |

Table S3: Annotations of putative HLD sequences.

| Annotation | Number of sequences | | | |
|---|----------------------------|---------------|----------------|-----------------|
| | HLD-I | HLD-II | HLD-III | HLD-IIIb |
| haloalkane dehalogenase | 126 | 184 | 64 | 5 |
| dehalogenase | 1 | 0 | 0 | 0 |
| luciferase | 0 | 6 | 0 | 0 |
| renilla-luciferin 2-monooxygenase-like | 0 | 7 | 0 | 0 |
| alpha/beta hydrolase | 12 | 9 | 111 | 0 |
| hydrolase | 0 | 2 | 24 | 0 |
| hypothetical protein/unnamed protein | 24 | 26 | 22 | 1 |
| CMP deaminase/tRNA-specific adenosine deaminase | 5 | 0 | 0 | 0 |
| acyl-CoA synthetase/fatty-acid-CoA ligase/AMP-dependent synthetase/AMP-ligase/AMP-binding protein | 0 | 0 | 27 | 0 |
| CurN | 0 | 1 | 0 | 0 |
| epoxide hydrolase | 0 | 0 | 0 | 1 |

Table S4: List of the chemical formulas of 34 halogenated compounds used for screening.

| Environmental pollutants | | | |
|---------------------------------|------------------------------------|------------------------------------|----------------------|
| α -hexachlorocyclohexane | α -hexabromocyclododecane | chlordane | 1,2-dichloroethane |
| β -hexachlorocyclohexane | β -hexabromocyclododecane | chlordecone | 1,2-dibromoethane |
| γ -hexachlorocyclohexane | γ -hexabromocyclododecane | 3-chloro-2(chloromethyl)-1-propene | 1-chlorobutane |
| δ -hexachlorocyclohexane | δ -hexabromocyclododecane | 1,3-dichlorprop-1-en | 1-bromobutane |
| heptachlorocyclohexane | aldrin | 1,2-dichloropropane | 1-iodopropane |
| heptachlor | dieldrin | 1,2,3-trichloropropane | |
| bis(2-chloroethyl)ether | endrin | 1-iodobutane | |
| Chemical warfare agents | Chemical warfare surrogates | Artificial sweetener | Desinfectants |
| bis(2-chloroethyl)sulfide | 2-chloroethyl methyl sulfide | sucralose | trichloramine |
| bis(2-chloroethyl)ethylamine | 2-chloroethyl phenyl sulfide | | |
| tris(2-chloroethyl)amine | | | |
| bis(2-chloroethyl)methylamine | | | |

Table S5: Overview of the selection process and applied selection criteria.

| Main category | Target property | Target # | No. | 1 st criteria | No. | 2 nd criteria | No. | 3 rd criteria | No. | Selected candidates |
|--------------------------|---|----------------|-----|---|-----|---|-----|-------------------------------|-----|---------------------|
| Taxonomy | Archaea | 1 | 5 | Halophile | 3 | Solubility | 2 | Lowest identity to known HLDs | 1 | DhmeA |
| Phylogenetics | HLD-IIIb subfamily | 1 | 6 | TM helices | 5 | 2 catalytic acids | 1 | | | DacA |
| Extremophilic origin | Thermophile | 1 | 3 | Solubility | 1 | Predicted activity | 2 | DcaA | | |
| | Psychrophile/cryophile | 2 | 16 | Moderate halophile | 3 | | | DpaB, DpaA | | |
| | Extreme halophile | 1 | 4 | Predicted activity | 1 | | | DmsaA | | |
| Multi-domain HLDs | AMP binding domain | 1 | 60 | Solubility | 6 | Random, no additional useful criteria available | 1 | DmtA | | |
| | CMP/dCMP deaminase domain | 1 | 12 | Predicted activity | 1 | | | DlxA | | |
| | Radical SAM domain | 1 | 1 | | | | | DdaA | | |
| Catalytic residues | 3 halide-stabilizing residues | 4* | 30 | Psychrophile | 2 | Lowest identity to known HLDs | 1 | DssA, DgpA | | |
| | | | | Solubility | 6 | | | DmgA | | |
| Volume of active site | Large active site >1500 [Å ³] | 2 | 5 | Predicted activity | 3 | Limit HLDs III subfamily – as it is often problematic | 2 | DncA, DamA | | |
| | Small active site < 200 [Å ³] | 1 | 2 | Predicted activity | 1 | | | DsxA | | |
| Molecular docking | Likely active HLDs (Predicted activity with high confidence) | 3 | 8 | Solubility | 3 | Limit HLDs III subfamily – as it is often problematic | | DmmaA, DtaA, DsbA | | |
| | Likely active with warfare agents (Predicted activity towards five yperite analogs) | 2 ^o | 2 | | 1 | | | DadA | | |
| Prediction of solubility | Predicted as soluble with high probability | 1 | 5 | Limit HLDs III subfamily – as it is often problematic | 3 | Random, no additional useful criteria available | 1 | DcsA | | |

Number of required hits for target property; * excluding already selected DlxA; ^o excluding already selected DsbA

Table S6: Accession numbers of putative HLDs.

| Name | Accession number |
|-------------|-------------------------|
| DacA | XP_001275122 |
| DssA | WP_012142124 |
| DcaA | YP_001637106 |
| DdaA | WP_012609180 |
| DamA | YP_003768533 |
| DmtA | WP_013583627 |
| DtaA | EHK43615 |
| DsbA | WP_014179778 |
| DhmeA | WP_004059800 |
| DadA | WP_014994594 |
| DmgA | EEB78108 |
| DpaB | WP_008302912 |
| DpaA | WP_008304783 |
| DmsaA | WP_008939131 |
| DgpA | WP_009022765 |
| DcsA | WP_009541349 |
| DsxA | WP_017667204 |
| DlxA | WP_019429073 |
| DncA | WP_020543687 |
| DmmaA | EPQ72372 |

Table S7: Overview of expression analysis of 20 putative HLDs.

| Expression strain | BL21 (DE3) | ArcticExpress (DE3) | BL21 (DE3) | Rosetta-gami 2 (DE3) | BL21 (DE3) | Predicted solubility [%] | |
|--|---------------|---------------------|------------|----------------------|------------|--------------------------|----|
| Temperature of cultivation/expression [°C] | 37/20 | 37/12 | 37/30 | 37/20 | 37/20 | | |
| Time of expression [hours] | 24 | 24 | 4 | 24 | 24 | | |
| Medium | Luria-Bertani | | | EnBase | | | |
| Proteins | DpaA | 63 | | | | 75 | |
| | DlxA | 3 | 16 | | | 34 | |
| | DmsaA | 40 | | | | 27 | |
| | DpaB | 48 | | | | 50 | |
| | DsxA | 45 | | | | 60 | |
| | DgpA | 47 | | | | 73 | |
| | DssA | | | | | 36 | |
| | DcaA | 11 | 2 | 13 | 4 | 9 | 39 |
| | DdaA | | | | | | 24 |
| | DhmeA | 26 | | | | | 80 |
| | DmtA | 5 | 6 | 4 | | 4 | 52 |
| | DadA | | | | | | 19 |
| | DtaA | 19 | | | | | 54 |
| | DcsA | 8 | | 15 | | | 81 |
| | DamA | 8 | | 7 | 3 | 9 | 68 |
| | DsbA | 48 | | | | | 71 |
| | DmmaA | | | | | | 56 |
| | DncA | 24 | | | | | 55 |
| | DmgA | 35 | | | | | 67 |
| | DacA | | | | | | 18 |

Black boxes indicate that the protein was obtained in a soluble form (at least 15 % of total intracellular proteins) judged by a visible band with the correct molecular weight in SDS-PAGE compared to a control sample. Grey boxes indicate expression but no or low solubility of protein. Pale grey boxes indicate no expression. The percentage content of soluble proteins in cell free extract was quantified by calibrated densitometer GS-800 (Bio-Rad, USA).

Table S8. Robotic screening of enzymatic activity.

| | Substrate | 1-iodobutane | | | 1,2-dibromoethane | | | 4-bromobutyronitrile | | |
|--------|-----------|------------------|-----|----|-------------------|-----|----|----------------------|-----|----|
| | | Temperature (°C) | 10 | 37 | 55 | 10 | 37 | 55 | 10 | 37 |
| Enzyme | DpaA | - | - | - | - | - | - | * | - | - |
| | DlxA | - | - | - | - | - | - | - | - | - |
| | DmsaA | ** | *** | - | * | *** | - | * | *** | * |
| | DpaB | * | - | - | * | - | - | ** | - | - |
| | DsxA | * | *** | * | ** | *** | - | ** | *** | * |
| | DgpA | - | * | - | * | * | - | * | - | - |
| | DhmeA | - | - | - | - | * | ** | - | * | * |
| | DtaA | - | - | - | * | *** | - | - | ** | - |
| | DcsA | - | - | - | - | - | - | - | * | - |
| | DsbA | - | * | - | - | - | - | - | - | * |
| | DncA | - | - | - | - | - | - | - | - | - |
| | DmgA | - | - | - | - | - | - | - | - | - |

The activity values of crude extracts are related to the total concentration of soluble proteins in cell-free extracts determined by a densitometer. Enzyme activity is expressed as a semi-quantitative measure of specific activity, where *** represents activity > 10.0 nmol.s⁻¹.mg⁻¹; ** activity 5.1-10.0 nmol.s⁻¹.mg⁻¹; * activity 0.1-5.0 nmol.s⁻¹.mg⁻¹; - no activity.

Table S9: The parameters for size, shape and oligomeric state of HLDs.

| Enzyme | Mw (kDa) | IV (dl/g) | Rh (nm) | Oligomeric state |
|---------------|-----------------|------------------|----------------|---------------------------------|
| DpaA | 70.639 | 0.029 | 2.820 | dimer |
| DmsaA | n.d. | n.d. | n.d. | n.d. |
| DpaB | 35.044 | 0.044 | 2.730 | monomer |
| DsxA | 34.070 | 0.043 | 2.130 | monomer |
| DgpA | 36.885 | 0.047 | 2.420 | monomer |
| DhmeA | 2 277.000 | 0.066 | 13.070 | multimer |
| DtaA | 34.414 | 0.046 | 2.290 | monomer |
| DsbA | 31.167 | 0.032 | 2.50 | monomer (97.6 %), dimer (2.4 %) |

n.d. - not determined

Table S10: Melting temperatures of HLDs by CD spectroscopy and robotic DSC.

| | CD | DSC |
|-------|------------------------------|-----------------|
| | T_m (°C) | |
| DpaA | 39.0 ± 0.53 | 38.2 ± 0.25 |
| DmsaA | n.d. | 51.4 ± 0.41 |
| DpaB | 40.4 ± 1.02 | 37.3 ± 0.63 |
| DsxA | 51.9 ± 0.34 | 48.8 ± 0.11 |
| DgpA | 36.6 ± 0.67 | 35.2 ± 0.16 |
| DhmeA | 69.8 ± 0.24 | 70.6 ± 0.34 |
| DtaA | 50.5 ± 0.18 | 47.2 ± 0.45 |
| DsbA | 58.1 ± 0.23 | 54.6 ± 0.50 |

n.d. – not determined

Table S11: Specific activities of HLDs towards 30 halogenated substrates.

| Substrate Code | Substrate | Specific activity (nmol.s ⁻¹ .mg ⁻¹) | | | | | | | |
|----------------|-----------------------------|---|-------|------|-------|------|-------|------|------|
| | | DpaA | DmsaA | DpaB | DsxA | DgpA | DhmeA | DtaA | DsbA |
| 4 | 1-chlorobutane | 0.0 | 0.0 | 0.7 | 6.5 | 0.0 | 0.0 | 0.9 | 0.0 |
| 6 | 1-chlorohexane | 1.1 | 6.5 | 10.4 | 119.4 | 2.0 | 0.0 | 0.3 | 0.0 |
| 18 | 1-bromobutane | 13.0 | 106.2 | 36.1 | 493.7 | 6.7 | 1.3 | 2.2 | 0.8 |
| 20 | 1-bromohexane | 8.8 | 36.0 | 14.7 | 356.3 | 8.7 | 1.1 | 2.8 | 1.1 |
| 28 | 1-iodopropane | 7.0 | 52.7 | 22.2 | 104.3 | 2.4 | 1.6 | 3.6 | 0.9 |
| 29 | 1-iodobutane | 4.6 | 40.6 | 30.5 | 176.0 | 2.2 | 2.1 | 1.5 | 0.5 |
| 31 | 1-iodohexane | 1.2 | 17.4 | 18.4 | 274.0 | 0.0 | 0.7 | 2.0 | 0.6 |
| 37 | 1,2-dichloroethane | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 38 | 1,3-dichloropropane | 0.0 | 0.0 | 0.0 | 8.9 | 0.0 | 1.2 | 8.3 | 0.0 |
| 40 | 1,5-dichloropentane | 5.8 | 17.9 | 7.5 | 114.4 | 3.5 | 0.5 | 2.9 | 0.0 |
| 47 | 1,2-dibromoethane | 3.9 | 13.1 | 10.7 | 99.4 | 2.8 | 8.1 | 62.2 | 0.2 |
| 48 | 1,3-dibromopropane | 15.6 | 251.7 | 39.8 | 661.2 | 6.8 | 3.7 | 11.4 | 3.7 |
| 52 | 1-bromo-3-chloropropane | 9.7 | 103.9 | 20.9 | 444.6 | 3.3 | 2.9 | 13.8 | 1.4 |
| 54 | 1,3-diiodopropane | 5.6 | 47.5 | 21.8 | 44.2 | 1.9 | 4.2 | 5.4 | 1.2 |
| 64 | 2-iodobutane | 2.1 | 25.7 | 5.7 | 56.9 | 1.5 | 2.2 | 3.9 | 0.8 |
| 67 | 1,2-dichloropropane | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 72 | 1,2-dibromopropane | 0.5 | 15.3 | 1.6 | 148.3 | 0.6 | 3.5 | 16.5 | 0.0 |
| 76 | 2-bromo-1-chloropropane | 0.7 | 12.4 | 1.1 | 106.7 | 0.9 | 2.1 | 16.2 | 0.0 |
| 80 | 1,2,3-trichloropropane | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.4 | 0.0 |
| 111 | bis(2chloroethyl)ether | 0.0 | 0.0 | 1.6 | 1.8 | 0.0 | 0.4 | 3.5 | 0.0 |
| 115 | chlorocyclohexane | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 |
| 117 | bromocyclohexane | 0.0 | 0.0 | 0.0 | 2.1 | 0.0 | 0.0 | 1.1 | 0.0 |
| 119 | (1-bromomethyl)cyclohexane | 0.7 | 9.8 | 3.6 | 21.1 | 0.6 | 0.4 | 1.2 | 0.2 |
| 137 | 1-bromo-2-chloroethane | 1.9 | 6.8 | 3.4 | 27.8 | 1.1 | 7.1 | 51.2 | 0.0 |
| 138 | chlorocyclopentane | 0.8 | 0.0 | 0.0 | 14.7 | 0.0 | 0.3 | 0.8 | 0.0 |
| 141 | 4-bromobutyronitrile | 5.2 | 37.0 | 14.2 | 167.7 | 2.8 | 6.9 | 28.5 | 0.7 |
| 154 | 1,2,3-tribromopropane | 0.3 | 10.3 | 0.6 | 23.0 | 0.3 | 6.1 | 24.2 | 0.0 |
| 155 | 1,2-dibromo-3-chloropropane | 0.0 | 8.1 | 0.0 | 13.1 | 0.0 | 5.4 | 25.8 | 0.0 |
| 209 | 3-chloro-2-methylpropene | 2.7 | 4.9 | 4.0 | 15.2 | 1.5 | 0.8 | 2.8 | 0.0 |
| 225 | 2,3-dichloropropene | 1.0 | 1.2 | 1.3 | 20.9 | 0.0 | 3.1 | 13.5 | 0.0 |

Table S12: Steady-state kinetic parameters of HLDs.

| Substrate | Enzyme | T (°C) | k_{cat} (s ⁻¹) | $K_{0.5}$ (mM) | n | $k_{\text{cat}}/K_{0.5}$ (s ⁻¹ .mM ⁻¹) |
|--------------------|--------|--------|-------------------------------------|----------------|--------------|---|
| 1,2-dibromoethane | DpaA | 10 | >0.06 | >3.70 | 2.34 | >0.02 |
| | DmsaA | 37 | 0.17 | 1.75 | ^a | 0.10 |
| | DpaB | 10 | 0.09 | 1.49 | 1.60 | 0.06 |
| | DsxA | 37 | >0.87 | >2.73 | 1.71 | >0.32 |
| | DgpA | 10 | >0.04 | >1.61 | 1.85 | >0.02 |
| | DhmeA | 55 | 0.22 | 0.99 | 1.34 | 0.22 |
| | DtaA | 37 | 1.54 | 1.18 | 1.29 | 1.31 |
| | DsbA | 37 | >0.01 | >3.59 | 1.40 | >0.003 |
| 1,3-dibromopropane | DsxA | 37 | 16.46 | 0.17 | 1.66 | 96.82 |

All parameters had standard errors less than 10%. k_{cat} , catalytic constant; $K_{0.5}$, concentration of substrate at half maximal velocity; n, Hill coefficient. ^a not applicable

Table S13: Enantioselectivities of HLDs towards 2-bromopentane and ethyl 2-bromopropionate.

| Enzyme | <i>E</i> -value | |
|--------|-----------------|-------------------------|
| | 2-bromopentane | Ethyl 2-bromopropionate |
| DpaA | 85 | >200 |
| DmsaA | 1 | 54 |
| DpaB | 13 | >200 |
| DsxA | 6 | >200 |
| DgpA | 1 | 3 |
| DhmeA | n.a. | 79 |
| DtaA | 12 | >200 |
| DsbA | 18 | 1 |

n.a., no activity

Table S14: Activities of HLDs towards warfare agent and environmental pollutants.

| | γ -HCH | HBCD | 1,3-dichloropropene [nmol.s ⁻¹ .mg ⁻¹] | bis(2-chloroethyl)sulfide [nmol.s ⁻¹ .mg ⁻¹] |
|-------|---------------|------|--|--|
| DpaA | - | - | 7.8 | - |
| DmsaA | - | - | 14.1 | - |
| DpaB | - | - | 10.7 | - |
| DsxA | - | - | 149.3 | - |
| DgpA | - | - | 9.1 | - |
| DhmeA | - | - | - | - |
| DtaA | - | - | 15.2 | 1.46 |
| DsbA | - | - | 0.5 | - |

-, no activity

Supporting Figures

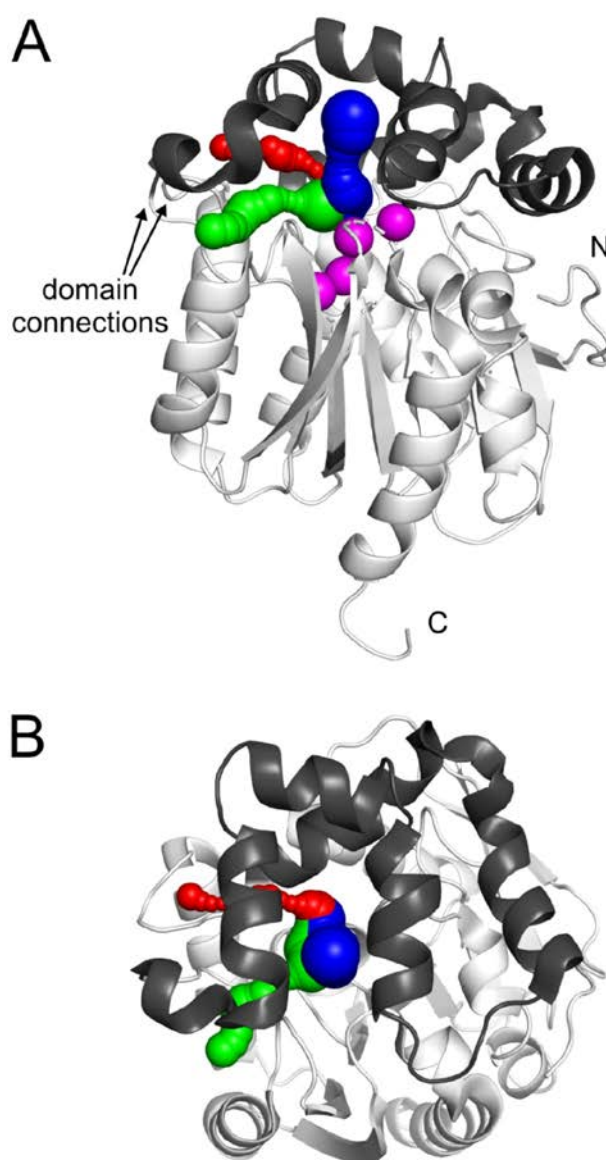


Figure S1: Positioning of the three investigated tunnels in dehalogenase structure using DhaA enzyme (PDB code 1CQW) as representative. α/β -hydrolase fold domain (white) and the specificity-determining cap domain (black) are distinguished. The p1, p2, and p3 tunnels are shown as blue, green, and red spheres, respectively. The catalytic residues are displayed as pink balls. A) Front view. B) Top view.

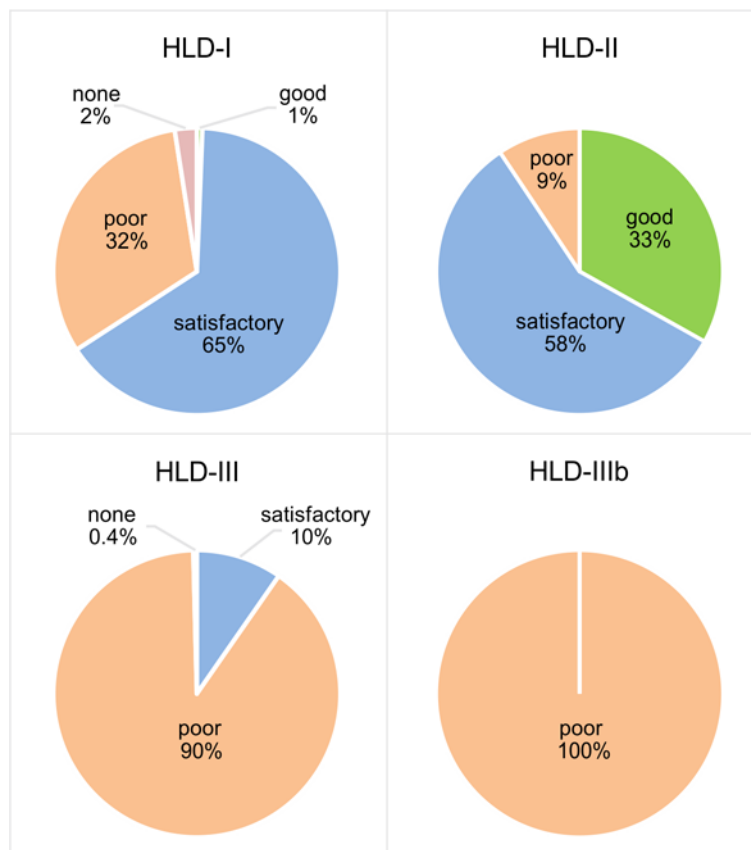


Figure S2: Availability of homology modelling templates in HLD subfamilies. Each template was classified based on its score as good (score ≤ 50), satisfactory (score ≤ 100) or poor (score > 100).

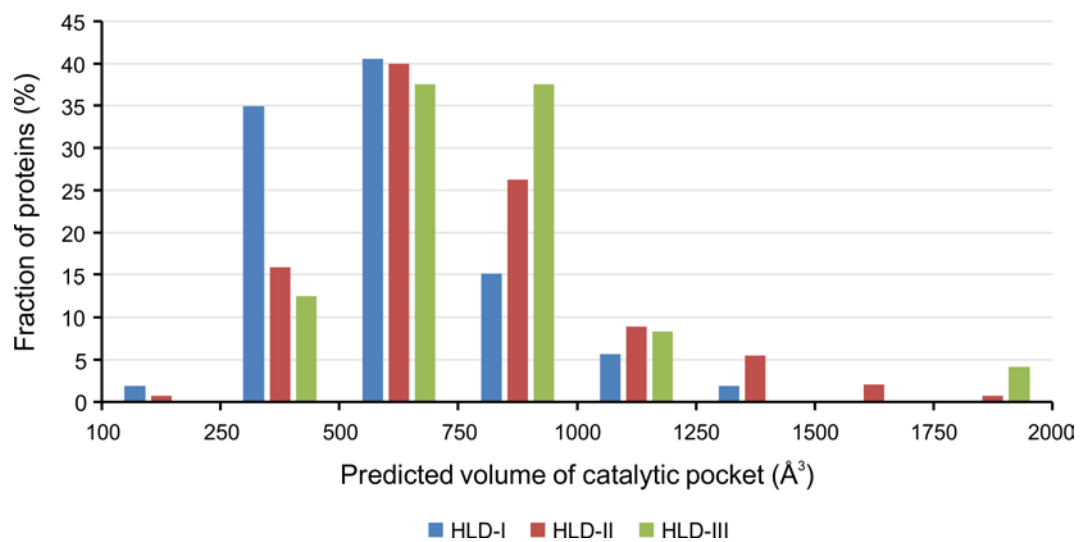


Figure S3: Distributions of predicted volumes of catalytic pockets in HLD subfamilies.

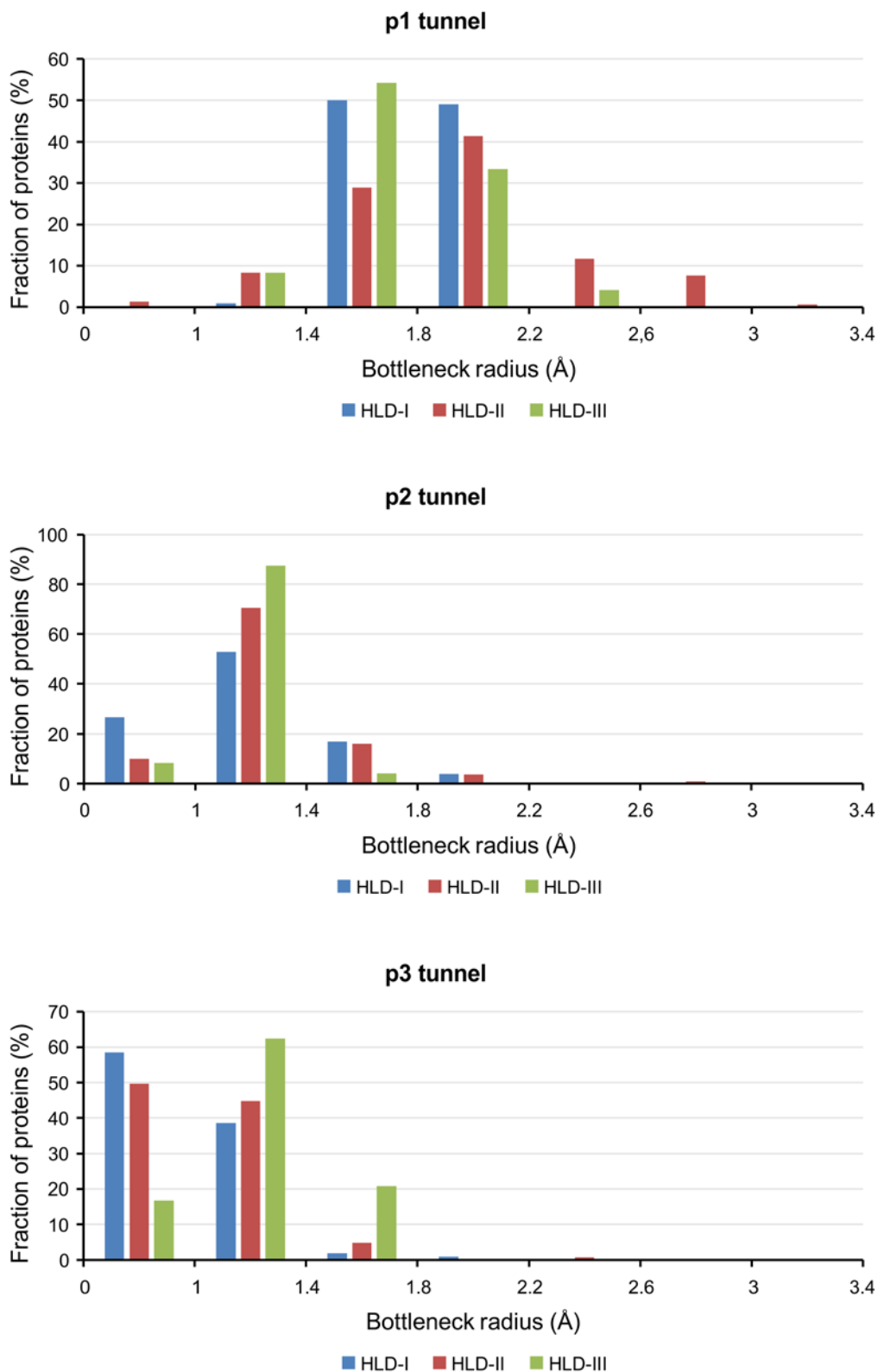


Figure S4: Distributions of predicted bottleneck radii of p1, p2 and p3 tunnels in HLD subfamilies.

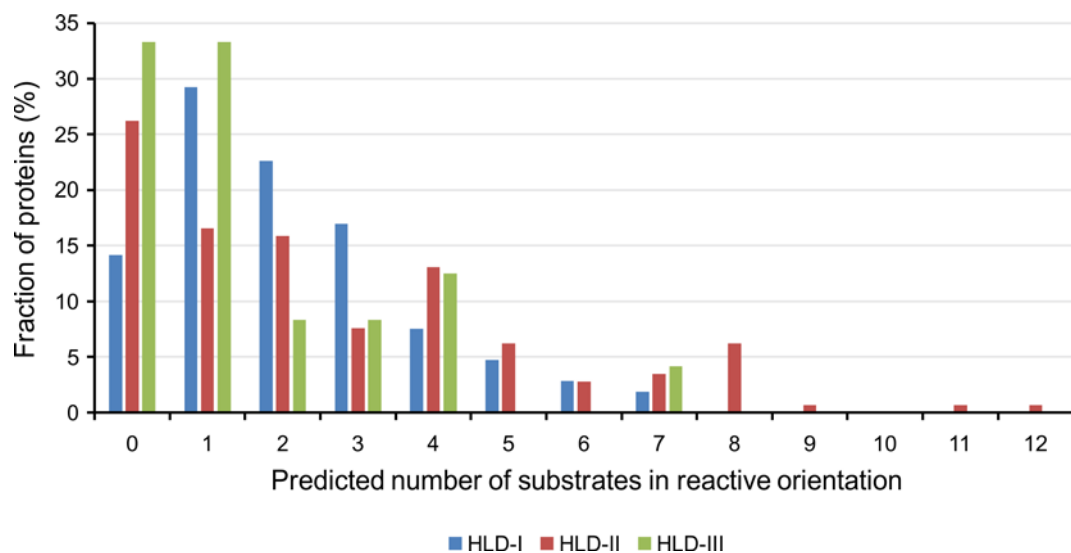


Figure S5: Distributions of mechanism-based geometric criteria for reactivity in HLD subfamilies.

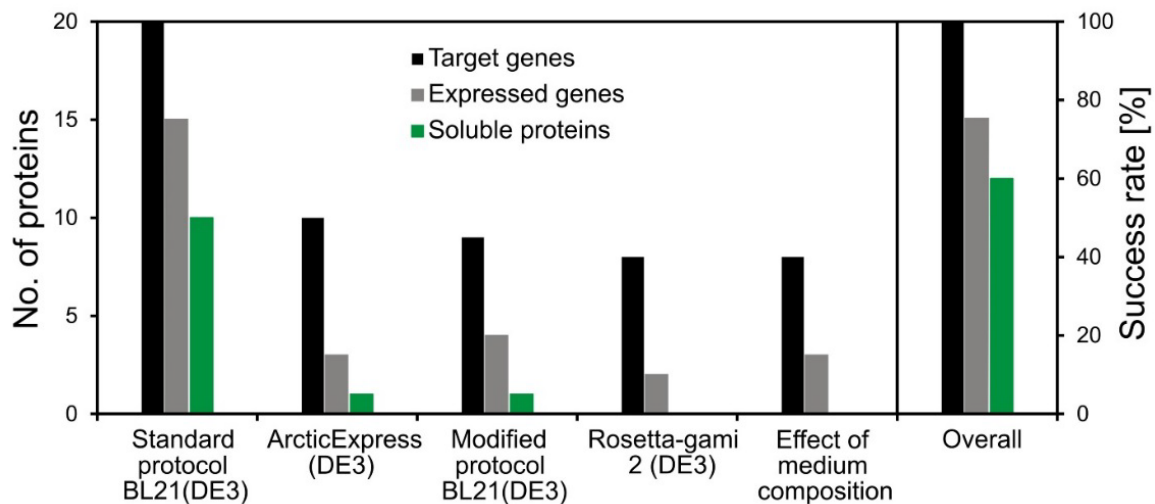


Figure S6: Expression analysis of the set of 20 putative HLDs. The plot shows the number of genes expressed and soluble proteins produced for each of five expression strategies as well as the overall success rate.

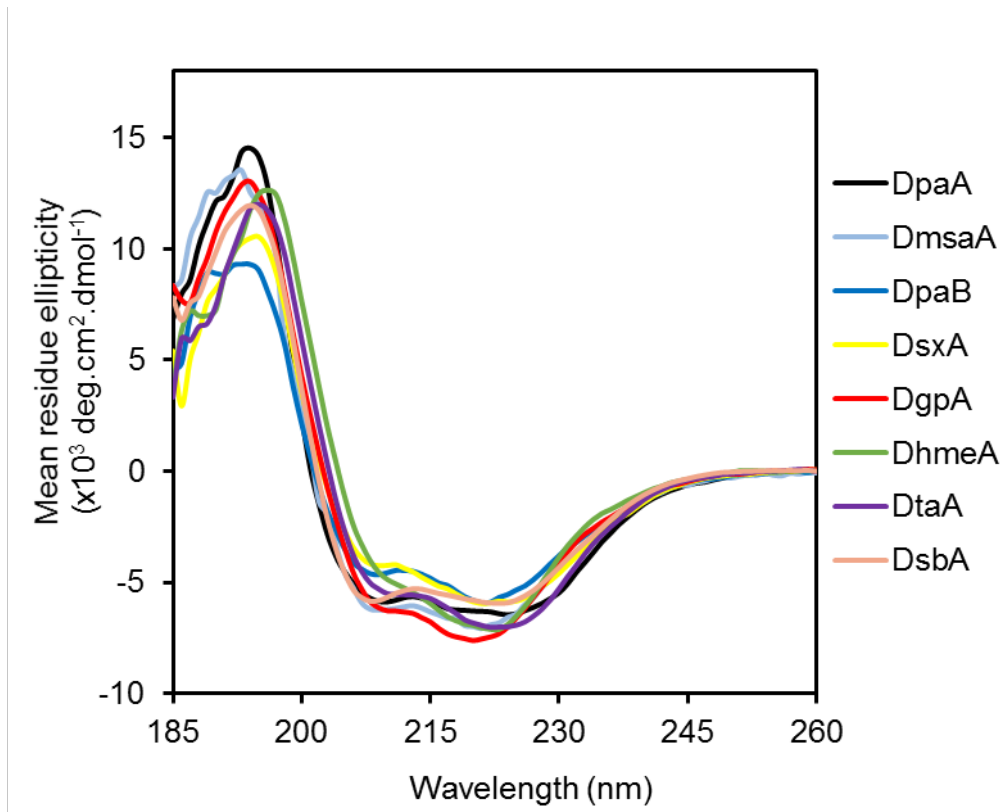


Figure S7: Far-UV circular dichroism spectra of novel HLDs.

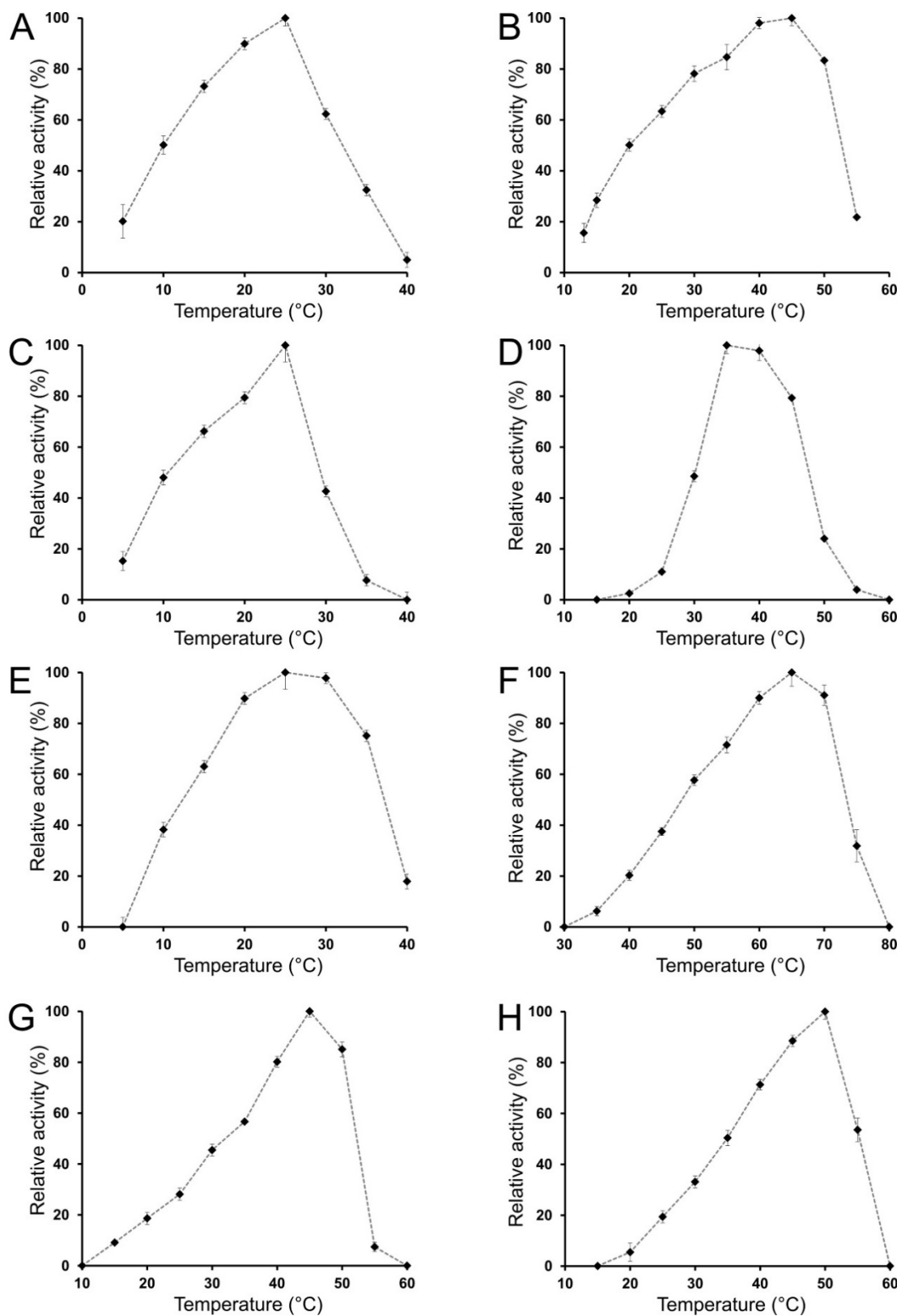


Figure S8: Temperature profiles of (A) DpaA, (B) DmsaA, (C) DpaB, (D) DsxA, (E) DgpA, (F) DhmeA, (G) DtaA, and (H) DsbA. The highest activity towards 1,3-diiodopropane at pH 8.6 was set to 100 %. Error bars represent the standard deviation from triplicated experiments.

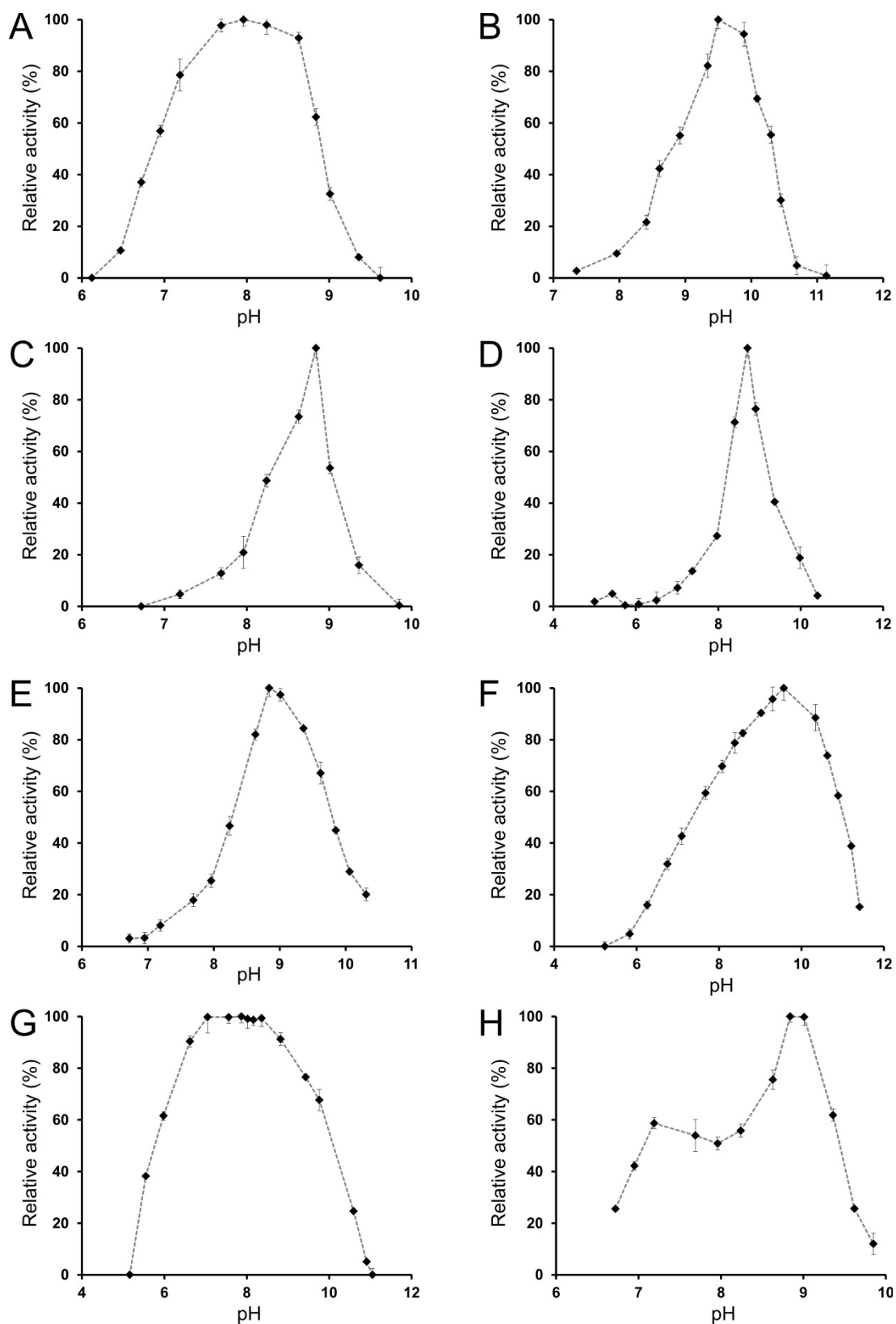


Figure S9: pH profiles of by (A) DpaA, (B) DmsaA, (C) DpaB, (D) DsxA, (E) DgpA, (F) DhmeA, (G) DtaA, and (H) DsbA. The highest activity towards 1,3-dibromopropane was set to 100 %. Error bars represent the standard deviation from triplicated experiments.

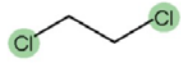
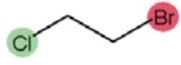
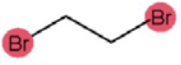
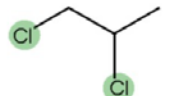
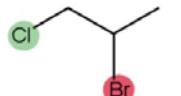

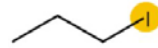
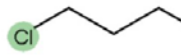
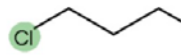
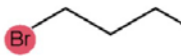

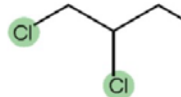
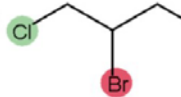
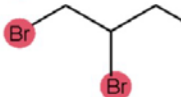



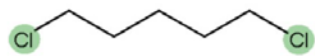
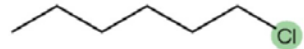






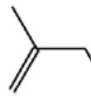
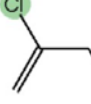
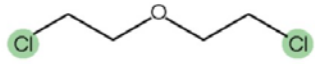

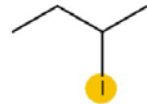
| | Chlorinated | Chlorinated & brominated | Brominated | Iodinated |
|-----------------------|---|---|---|--|
| Ethanes | 37  | 137  | 47  | |
| Propanes | 67  | 76  | 72  | 28  |
| | 38  | 52  | 48  | 54  |
| | 80  | 155  | 154  | |
| Butanes | 4  | | 18  | 29  |
| Pentanes | 40  | | | |
| Hexanes | 6  | | 20  | 31  |
| Cyclic alkanes | 138  | 115  | 117  | 119  |
| | 209  | 225  | | |
| Miscellaneous alkanes | 111  | | 141  | 64  |

Figure S10: The set of thirty halogenated substrates used to describe the substrate specificity of the HLDs. The numerical coding of the substrates is based on previously established nomenclature.

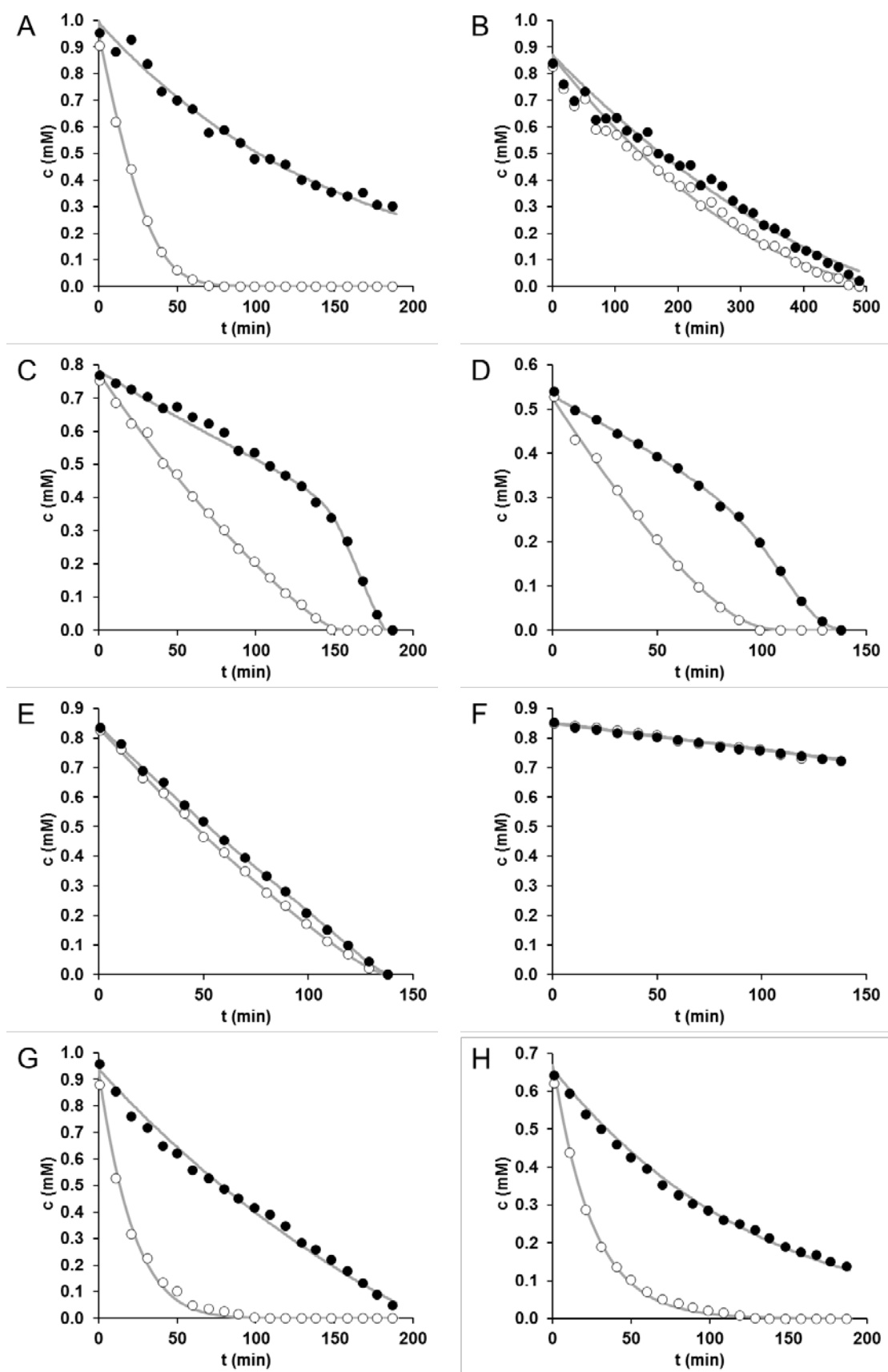


Figure S11: Kinetic resolution of 2-bromopentane catalyzed by (A) DpaA, (B) DmsaA, (C) DpaB, (D) DsxA, (E) DgpA, (F) DhmeA, (G) DtaA, and (H) DsbA. \circ : (*R*)-enantiomer, \bullet : (*S*)-enantiomer.

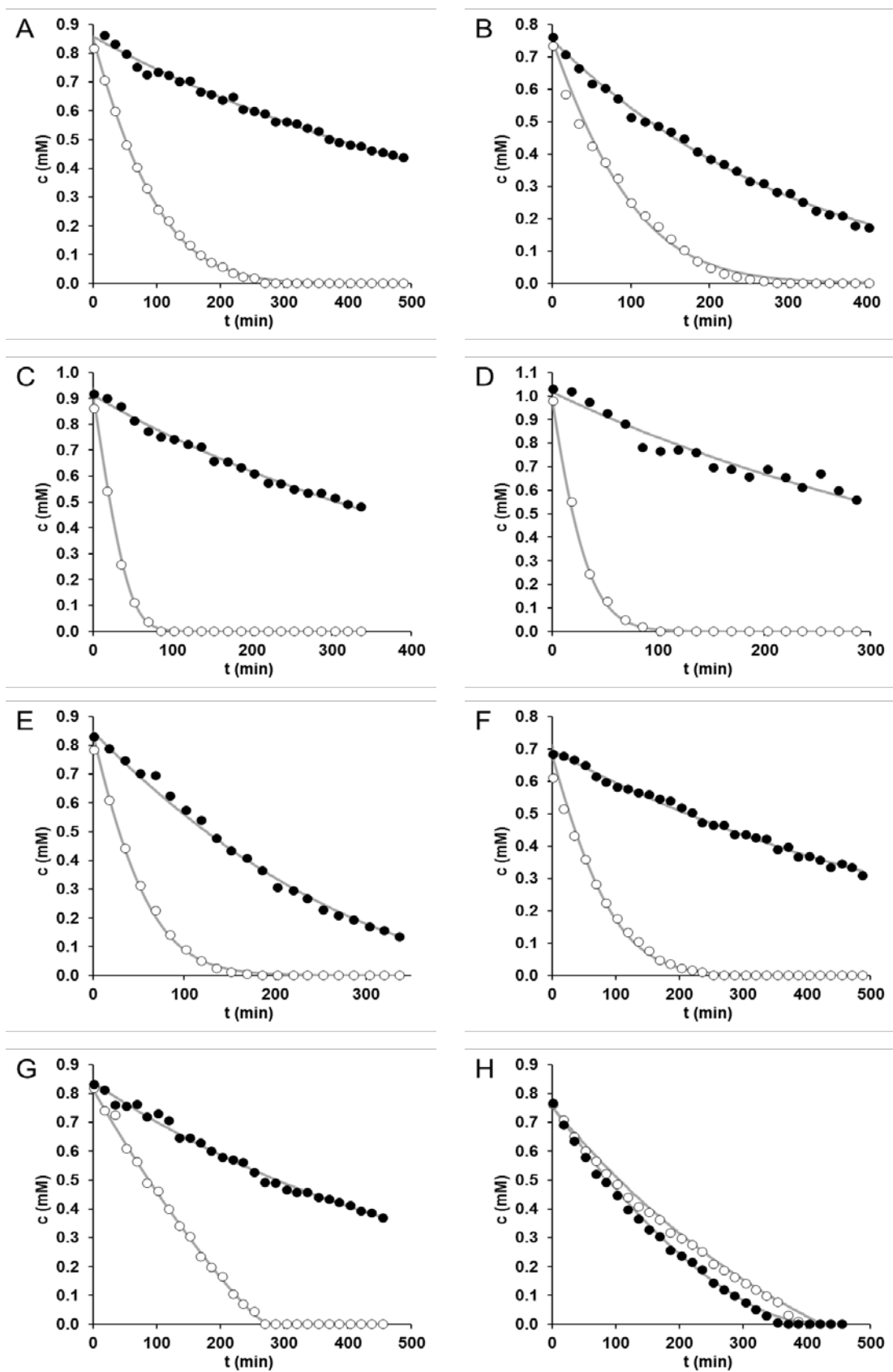


Figure S12: Kinetic resolution of ethyl 2-bromopropionate catalyzed by (A) DpaA, (B) DmsaA, (C) DpaB, (D) DsxA, (E) DgpA, (F) DhmeA, (G) DtaA, and (H) DsbA. \circ : (*R*)-enantiomer, \bullet : (*S*)-enantiomer.

Supporting References

- (1) Altschul, S. F.; Madden, T. L.; Schäffer, A. A.; Zhang, J.; Zhang, Z.; Miller, W.; Lipman, D. J. Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs. *Nucleic Acids Res.* **1997**, *25*, 3389–3402.
- (2) Coordinators, N. R. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res.* **2013**, *41*, D8–D20.
- (3) Needleman, S. B.; Wunsch, C. D. A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *J. Mol. Biol.* **1970**, *48*, 443–453.
- (4) Edgar, R. C. Search and Clustering Orders of Magnitude Faster than BLAST. *Bioinformatics* **2010**, *26*, 2460–2461.
- (5) Loewenstein, Y.; Portugal, E.; Fromer, M.; Linial, M. Efficient Algorithms for Accurate Hierarchical Clustering of Huge Datasets : Tackling the Entire Protein Space. *Bioinformatics* **2008**, *24*, 41–49.
- (6) Sievers, F.; Wilm, A.; Dineen, D.; Gibson, T. J.; Karplus, K.; Li, W.; Lopez, R.; McWilliam, H.; Remmert, M.; Söding, J.; Thompson, J. D.; Higgins, D. G. Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments Using Clustal Omega. *Mol. Syst. Biol.* **2011**, *7*, 539.
- (7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (8) Sali, A.; Blundell, T. L. Comparative Protein Modelling by Satisfaction of Spatial Restraints. *J. Mol. Biol.* **1993**, *234*, 779–815.
- (9) Harrison, R. G. Expression of Soluble Heterologous Proteins via Fusion with NusA Protein. *Innovations* **2000**, *11*, 4–7.
- (10) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in *Escherichia coli*. *Nat. Biotechnol.* **1991**, *9*, 443–448.
- (11) Finn, R. D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R. Y.; Eddy, S. R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; Sonnhammer, E. L. L.; Tate, J.; Punta, M. Pfam: The Protein Families Database. *Nucleic Acids Res.* **2014**, *42*, D222–D230.
- (12) Tsirigos, K. D.; Peters, C.; Shu, N.; Käll, L.; Elofsson, A. The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides. *Nucleic Acids Res.* **2015**, *43*, W401–W407.
- (13) Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed Atlas of Surface Topography of Proteins with Structural and Topographical Mapping of Functionally Annotated Residues. *Nucleic Acids Res.* **2006**, *34*, W116–W118.
- (14) Liang, J.; Edelsbrunner, H.; Woodward, C. Anatomy of Protein Pockets and Cavities: Measurement of Binding Site Geometry and Implications for Ligand Design. *Protein Sci.* **1998**, *7*, 1884–1897.
- (15) Chovancova, E.; Pavelka, A.; Benes, P.; Strnad, O.; Brezovsky, J.; Kozlikova, B.; Gora, A.; Sustr, V.; Klvana, M.; Medek, P.; Biedermannova, L.; Sochor, J.; Damborsky, J. CAVER 3.0: A Tool for the Analysis of Transport Pathways in Dynamic Protein Structures. *PLoS Comput. Biol.* **2012**, *8*, e1002708.
- (16) Hanwell, M. D.; Curtis, D. E.; Lonie, D. C.; Vandermeersch, T.; Zurek, E.; Hutchison, G. R.

- Avogadro: An Advanced Semantic Chemical Editor, Visualization, and Analysis Platform. *J. Chem. Informatics* **2012**, *4*, 1–17.
- (17) Vanquelef, E.; Simon, S.; Marquant, G.; Garcia, E.; Klimerak, G.; Delepine, J. C.; Cieplak, P.; Dupradeau, F. Y. R.E.D. Server: A Web Service for Deriving RESP and ESP Charges and Building Force Field Libraries for New Molecules and Molecular Fragments. *Nucleic Acids Res.* **2011**, *39*, W511–W517.
- (18) Morris, G. M.; Huey, R.; Lindstrom, W.; Sanner, M. F.; Belew, R. K.; Goodsell, D. S.; Olson, A. J. AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility. *J. Comput. Chem.* **2009**, *30*, 2785–2791.
- (19) Solis, F. J.; Wets, R. J.-B. Minimization by Random Search Techniques. *Math. Oper. Res.* **1981**, *6*, 19–30.
- (20) Daniel, L.; Buryška, T.; Prokop, Z.; Damborsky, J.; Brezovsky, J. Mechanism-Based Discovery of Novel Substrates of Haloalkane Dehalogenases Using in Silico Screening. *J. Chem. Inf. Model.* **2015**, *55*, 54–62.
- (21) Iwasaki, I.; Utsumi, S.; Ozawa, T. New Colorimetric Determination of Chloride Using Mercuric Thiocyanate and Ferric Ion. *Bull. Chem. Soc. Jpn.* **1952**, *25*, 226.
- (22) Woody, R. W. *Circular Dichroism and the Conformational Analysis of Biomolecules*, 1st ed.; Fasman, G. D., Ed.; Plenum Press: New York, 1996; Vol. 1.
- (23) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.
- (24) Koudelakova, T.; Chovancova, E.; Brezovsky, J.; Monincova, M.; Fortova, A.; Jarkovsky, J.; Damborsky, J. Substrate Specificity of Haloalkane Dehalogenases. *Biochem. J.* **2011**, *435*, 345–354.
- (25) Spelberg, J. H. L.; Rink, R.; Kellogg, R. M.; Janssen, D. B. Enantioselectivity of a Recombinant Epoxide Hydrolase from *Agrobacterium Radiobacter*. *Tetrahedron Asymmetry* **1998**, *9*, 459–466.
- (26) Kuzmic, P. Dynafit-A Software Package for Enzymology. *Methods Enzymol.* **2009**, *467*, 247–280.
- (27) Keuning, S.; Janssen, D. B.; Witholt, B. Purification and Characterization of Hydrolytic Haloalkane Dehalogenase from *Xanthobacter autotrophicus* GJ10. *J. Bacteriol.* **1985**, *163*, 635–639.
- (28) Jesenska, A.; Bartos, M.; Czernekova, V.; Rychlik, I.; Pavlik, I.; Damborský, J. Cloning and Expression of the Haloalkane Dehalogenase Gene *dhmA* from *Mycobacterium avium* N85 and Preliminary Characterization of DhmA. *Appl. Environ. Microbiol.* **2002**, *68*, 3724–3730.
- (29) Jesenska, A.; Pavlova, M.; Strouhal, M.; Chaloupkova, R.; Tesinska, I.; Monincova, M.; Bartos, M.; Pavlik, I.; Rychlik, I.; Möbius, P.; Nagata, Y.; Damborsky, J. Cloning, Biochemical Properties, and Distribution of Mycobacterial Haloalkane Dehalogenases. *Appl. Environ. Microbiol.* **2005**, *71*, 6736–6745.
- (30) Drienovska, I.; Chovancova, E.; Koudelakova, T.; Damborsky, J.; Chaloupkova, R. Biochemical Characterization of a Novel Haloalkane Dehalogenase from a Cold-Adapted Bacterium. *Appl. Environ. Microbiol.* **2012**, *78*, 4995–4998.
- (31) Hesseler, M.; Bogdanović, X.; Hidalgo, A.; Berenguer, J.; Palm, G. J.; Hinrichs, W.; Bornscheuer, U. T. Cloning, Functional Expression, Biochemical Characterization, and

- Structural Analysis of a Haloalkane Dehalogenase from *Plesiocystis pacifica* SIR-1. *Appl. Microbiol. Biotechnol.* **2011**, *91*, 1049–1060.
- (32) Hasan, K.; Fortova, A.; Koudelakova, T.; Chaloupkova, R.; Ishitsuka, M.; Nagata, Y.; Damborsky, J.; Prokop, Z. Biochemical Characteristics of the Novel Haloalkane Dehalogenase DatA, Isolated from the Plant Pathogen *Agrobacterium tumefaciens* C58. *Appl. Environ. Microbiol.* **2011**, *77*, 1881–1884.
- (33) Chaloupkova, R.; Prudnikova, T.; Rezacova, P.; Prokop, Z.; Koudelakova, T.; Daniel, L.; Brezovsky, J.; Ikeda-Ohtsubo, W.; Sato, Y.; Kutý, M.; Nagata, Y.; Smetanova, I. K.; Damborsky, J. Structural and Functional Analysis of a Novel Haloalkane Dehalogenase with Two Halide-Binding Sites. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **2014**, *70*, 1884–1897.
- (34) Sato, Y.; Monincova, M.; Chaloupkova, R.; Prokop, Z.; Ohtsubo, Y.; Minamisawa, K.; Tsuda, M.; Damborsky, J.; Nagata, Y. Two Rhizobial Strains, *Mesorhizobium loti* MAFF303099 and *Bradyrhizobium japonicum* USDA110, Encode Haloalkane Dehalogenases with Novel Structures and Substrate Specificities. *Appl. Environ. Microbiol.* **2005**, *71*, 4372–4379.
- (35) Kulakova, A. N.; Larkin, M. J.; Kulakov, L. A. The Plasmid-Located Haloalkane Dehalogenase Gene from *Rhodococcus rhodochrous* NCIMB 13064. *Microbiology* **1997**, *143*, 109–115.
- (36) Gehret, J. J.; Gu, L.; Geders, T. W.; Brown, W. C.; Gerwick, L.; Gerwick, W. H.; Sherman, D. H.; Smith, J. L. Structure and Activity of DmmA, a Marine Haloalkane Dehalogenase. *Protein Sci.* **2012**, *21*, 239–248.
- (37) Fortova, A.; Sebestova, E.; Stepankova, V.; Koudelakova, T.; Palkova, L.; Damborsky, J.; Chaloupkova, R. DspA from *Strongylocentrotus purpuratus*: The First Biochemically Characterized Haloalkane Dehalogenase of Non-Microbial Origin. *Biochimie* **2013**, *95*, 2091–2096.
- (38) Chan, W. Y.; Wong, M.; Guthrie, J.; Savchenko, A. V.; Yakunin, A. F.; Pai, E. F.; Edwards, E. A. Sequence- and Activity-Based Screening of Microbial Genomes for Novel Dehalogenases. *Microb. Biotechnol.* **2010**, *3*, 107–120.
- (39) Nagata, Y.; Miyauchi, K.; Damborsky, J.; Manova, K.; Ansorgova, A.; Takagi, M. Purification and Characterization of a Haloalkane Dehalogenase of a New Substrate Class from a Gamma-Hexachlorocyclohexane - Degrading Bacterium, *Sphingomonas paucimobilis* UT26. *Appl. Environ. Microbiol.* **1997**, *63*, 3707–3710.
- (40) Jesenska, A.; Monincova, M.; Koudelakova, T.; Hasan, K.; Chaloupkova, R.; Prokop, Z.; Geerlof, A.; Damborsky, J. Biochemical Characterization of Haloalkane Dehalogenases DrbA and DmbC, Representatives of a Novel Subfamily. *Appl. Environ. Microbiol.* **2009**, *75*, 5157–5160.