Functional Annotation of an Enzyme Family by Integrated Strategy Combining Bioinformatics with Microanalytical and Microfluidic Technologies

Pavel Vanacek[1,2,#], Michal Vasina[1,2,#], Jiri Hon[2,3], David Kovar[1], Hana Faldynova[1], Antonin Kunka[1,2], Tomas Buryska[1], Christoffel P. S. Badenhorst[4], Stanislav Mazurenko[1,2], David Bednar[1,2], Uwe Bornscheuer[4], Jiri Damborsky[1,2,*], Zbynek Prokop[1,2,*]

[1] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Brno, Czech Republic

[2] International Clinical Research Centre, St. Ann's Hospital, Brno, Czech Republic

[3] IT4Innovations Centre of Excellence, Faculty of Information Technology, Brno University of Technology, 612 66 Brno, Czech Republic

[4] Department of Biotechnology & Enzyme Catalysis, Institute of Biochemistry, Greifswald University, Greifswald 17487, Germany

# P.V. and M.V. contributed equally

* Authors for correspondence: jiri@chemi.muni.cz; zbynek@chemi.muni.cz

ABSTRACT

Next-generation sequencing technologies enable doubling of the genomic databases every 2.5 years. Collected sequences represent a rich source of novel biocatalysts. However, the rate of accumulation of sequence data exceeds the rate of functional studies, calling for acceleration and miniaturization of biochemical assays. Here, we present an integrated platform employing bioinformatics, microanalytics, and microfluidics and its application for exploration of unmapped sequence space, using haloalkane dehalogenases as model enzymes. First, we employed bioinformatic analysis for identification of 2,905 putative dehalogenases and rational selection of 45 representative enzymes. Second, we expressed and experimentally characterized 24 enzymes showing sufficient solubility for microanalytical and microfluidic testing. Miniaturization increased the throughput to 20,000 reactions per day with 1000-fold lower protein consumption compared to conventional assays. A single run of the platform doubled dehalogenation toolbox of family members characterized over three decades. Importantly, the dehalogenase activities of nearly one-third of these novel biocatalysts far exceed that of most published HLDs. Two enzymes showed unusually narrow substrate specificity, never before reported for this enzyme family. The strategy is generally applicable to other enzyme families, paving the way towards the acceleration of the process of identification of novel biocatalysts for industrial applications but also for the collection of homogenous data for machine learning. The automated *in silico* workflow has been released as a user-friendly web-tool EnzymeMiner: https://loschmidt.chemi.muni.cz/enzymeminer/.

Keywords: enzyme mining; enzyme diversity; novel biocatalysts; microscale; microfluidics

INTRODUCTION

Nature is an experienced "protein engineer" since it launched its bioengineering experiments billions of years ago.[1] It has evolved a fascinating diversity of biocatalysts. This enormous natural heritage represents an immense treasure trove for the life sciences, which strive for biocatalysts fulfilling demands for new and higher quality products, and societal demands for 'greener' technologies. We are still scratching the surface of potential biocatalytic applications as the pool of available enzymes catering to biomedical, pharmaceutical, and industrial applications is still very limited.[2] The advent of next-generation sequencing technologies has revolutionized genomic research, filling public databases with DNA and protein sequences. There are currently almost 300 million non-redundant protein sequences in genomic databases.[3] Despite the enormous promise for biological and biotechnological discovery, the rate of sequence data accumulation far exceeds the speed of functional studies. In theory, systematic functional characterization integrating *in vitro* biochemical with *in vivo* physiological approaches would be the preferred basis of functional annotation of novel proteins. In practice, a low ratio of explored to unexplored sequences reflects the poor efficiency of the low-throughput biochemical techniques, in contrast to high-throughput next-generation sequencing technologies.

The magnitude of the "big data" problem in biology is magnified by the fact that a large portion of functional annotations contain vague, indirect, or even incorrect descriptions.[4] Extrapolation from sequence to protein function is not trivial and often proves to be incorrect when tested experimentally. Therefore, the challenge now lies in exploring this new sequence information to mine desired biocatalysts effectively and, concurrently, designing experimental efforts to accelerate biocatalyst characterization. Several strategies have been developed to exploit the large source of enzyme sequences contained in genome and metagenome databases and to discover novel biocatalysts.[5,6] Efficient exploration of the millions of uncharacterized enzymes in public databases can be achieved by computational approaches, which offer an adequate capacity for the screening of large pools of sequences.[7] Our sequence-based strategy identifies putative members of the known enzyme families and facilitates their prioritization and well-informed selection for experimental characterization.[8] The main benefit of such a genome mining approach is effective identification of thousands of putative hits among millions of sequence entries and rational selection of a restricted set of attractive targets.[9,10] The selected representative candidates need to be experimentally characterized, which is the rate-limiting process. Conventional techniques used for the collection of experimental biochemical data are time-demanding, cost-ineffective, and low-throughput. The development of miniaturized technologies and their implementation in high-performance experimental pipelines will accelerate the process of discovering new biocatalysts.[11,12]

Here, we describe an integrated workflow for the effective mining of novel family members using computer-assisted genome mining and high-throughput experimental characterization. It was applied to the haloalkane dehalogenases (HLDs, EC 3.8.1.5) as model enzymes. Initially, we applied an automated bioinformatics workflow for the identification of putative family members and selection of promising candidates. Next, we experimentally characterized selected hits from the screening. Employing microanalytical and microfluidic methods for experimental data acquisition increased the analytical throughput to >20,000 reactions per day and reduced protein consumption by three orders of magnitude to only hundreds of micrograms. We have almost doubled the number of experimentally characterized HLDs by a single run of the platform and obtained a pool of new biocatalysts with industrially attractive characteristics: high catalytic efficiencies, unique substrate specificities, high enantioselectivities, and broadened temperature tolerance.

RESULTS

I. *In Silico* Screening using an Automated Workflow

Model Enzymes. The automated database-mining protocol was tested with the HLDs (**Fig. 1**). Three decades of intensive research on HLDs has made them the benchmark enzymes for studying the structure-function relationships of the >100,000 members of the α/β-hydrolase superfamily[13] and the development of novel concepts in the field of protein engineering.[14–16] Members of this enzyme family have been employed in several practical applications: (i) biocatalytic preparation of optically pure building-blocks for organic synthesis, (ii) recycling of by-products from chemical processes, (iii) bioremediation of toxic environmental pollutants, (iv) decontamination of chemical warfare agents, (v) biosensing of environmental pollutants, and (vi) protein tagging for cell imaging and protein analysis.[17]

Database Mining. The genomic databases are doubling every ~2.5 years, therefore periodic mining of novel genes is highly advisable. We reran the *in silico* screening with the same input sequence as in the initial round[8] in 2018, using the current version of NCBI nr database and recently developed tools for automated database mining.[18] In comparison to the initial version, the *in silico* workflow has been expanded by: (i) application of EFI-EST[19] and Cytoscape[20] for calculation and visualization of the sequence similarity network, (ii) extraction of the biotic relationships and disease annotations of the source organisms from the BioProject database,[21] and (iii) the quantitative assessment of the quality of all homology models by MolProbity.[22] Sequence database searches using four known HLDs as query sequences generated 24,594 hits sharing minimal sequence similarity to at least one of the query sequences. The putative HLD sequences containing the target HLD domain were automatically recognized using global pairwise sequence identities and average-link hierarchical clustering. Artificial

protein sequences annotated by the terms "artificial", "synthetic construct", "vector", "vaccinia virus", "plasmid", "HaloTag", or "replicon", were excluded.

Clustering and Filtering. The remaining 2,905 protein sequences were clustered into four subfamilies: HLD-I (915), HLD-II (1058), HLD-III/IIIb (910), and HLD-IV (22), based on sequence identity and the composition of their catalytic pentads. Despite having identical catalytic pentads, HLD-III and HLD-IIIb were clustered separately based on differences in their sequence similarity. The incomplete and degenerated sequences were filtered out by the construction of multiple sequence alignments of individual subfamilies. Sequence-similarity networks were used to visualize functional relationships among putative HLD sequences (**Fig. 2**). The clearest defining features are clustering in the distinct HLD subfamilies, implying that the sequence-similarity networks might provide a framework for identifying HLDs that act on specific compound classes and for surveying regions of sequence space where substrate preference is unknown. To diversify HLD sequence space, redundant sequences with $\geq 90\%$ sequence identity to the set of characterized dehalogenase sequences were filtered out.

Annotation and Homology Modelling. The remaining 2,578 putative HLD sequences were subjected to an annotation step consisting of the retrieval of information from biological databases and structure predictions. The annotation step revealed that the identified HLDs span over a broad range of sequence and host diversity, including bacterial, archaeal, and eukaryotic proteins. The overall accuracy of annotation, judged by assignment to HLD family, was 63% but varied significantly among each of the HLD subfamilies. The majority of sequences in HLD-I (73%) and HLD-II (86%) subfamilies were annotated correctly. In contrast, the portion of correctly annotated sequences was reduced to 31% for HLD-III and to 56% for HLD-IIIb (**Table S1**). The majority of members from the putative HLD-IIIb subfamily were annotated as HLDs, despite their low sequence identity to experimentally characterized HLDs or other HLD subfamily members. The annotation revealed 4 putative dehalogenases from psychrophilic organisms, 35 novel proteins from moderate halophilic organisms and 4 protein templates with known tertiary structure. Reliable homology models could be constructed for the majority of members from subfamily HLD-I and HLD-II, but for only a limited number of HLD-III members and for no HLD-IIIb members. The predicted volumes of catalytic pockets ranged from 60 $Å^3$ to 2170 $Å^3$ (**Fig. S1**).

Prioritization and Selection of Targets. Rational selection of hits for experimental characterization was carried out to maximize functional diversity of the studied protein family (**Supporting Data Set)**. The dataset of 2,578 putative HLDs was summarized in 17 datasheets focused on different annotations or computed properties. Hits represented by homology models with MolProbity scores >3.0 were removed from the datasheets summarizing the annotations based on the predicted homology structure,

i.e., active site volume and tunnel properties. To make the selection as diverse as possible, a few sequences were picked from each datasheet (**Table S2**). The sequences with higher predicted solubility and higher-quality homology models were preferred. Simultaneously, we tried to balance a reasonable number of sequences from each haloalkane dehalogenase subfamily (HLD-I. HLD-II and HLD-III). The only exception was the HLD-IIIb subfamily, which contains multi-domain protein sequences derived from eukaryotic organisms. We avoided sequences with additional Pfam domains, as they are usually poorly expressible in bacterial host systems. Altogether 45 representative sequences were selected as targets for experimental characterization (**Table S3, Table S4**).

II. Small-Scale Protein Expression

The representative set of 45 HLD genes was subjected to small-scale expression in *Escherichia coli* in 96-deep well square plates (**Table S5**) and screening of HLD activity. Overall, 40 out of 45 (89%) genes could be overexpressed, while 30 out of 45 (67%) genes yielded proteins in soluble form (**Fig. S3A**). The comparison of the *in silico* prediction of soluble expression with experiments showed a poor correlation (Pearson's correlation coefficient 0.263) and 66.7% prediction accuracy. Specifically, the *in silico* solubility predictions resulted in 8 true negatives, 22 true positives, 4 false-negative and 14 false-positives (**Table S5**). A further thorough analysis of solubility profiles revealed that the majority of proteins belonging to HLD-I (73%) and HLD-II (71%) were expressed in a soluble form, while a minority of HLD-III (40%) proteins were soluble. We have probed the expressibility of all 45 HLD genes using a reconstituted cell-free transcription and translation (PURExpress) system. Overall, 41 of 45 (91%) genes were overexpressed and 29 (64%) proteins were obtained in soluble form (**Fig. S3B**). Application of the cell-free system did not result in a desired improvement of solubility for the difficult-to-express HLDs. The screening of the HLD activity was performed using a newly developed ultrasensitive fluorescence assay [23], which strikingly confirmed the HLD activity of all soluble proteins.

III. Characterization using Microscale and Microfluidic Techniques

The small-scale expression screening in *E. coli* resulted in the selection of 24 HLDs possessing sufficient solubility for biochemical characterization (**Fig. S4**). The experimental pipeline comprised microanalytical technologies employing commercial instruments and *in-house* microfluidic devices, which led to 1000-fold lower consumption of protein and increased throughput to 22,000 reactions per day (**Table S6**). These methods provided experimental data on protein stability, catalytic activity, temperature profiles, substrate specificity, and enantioselectivity (**Table 1**).

Protein Stability. The stability of the novel HLDs was analyzed by monitoring changes in extrinsic (SYPRO orange dye) and intrinsic (tryptophan) fluorescence during temperature scanning experiments in a high-throughput manner, using thermal shift assay (TSA) and differential scanning fluorimetry

(DSF), respectively. The midpoint of the denaturation curve (apparent melting temperature -T_m^{app}) was used for the stability evaluation. The accuracy of the fast-screening methods was benchmarked against conventional circular dichroism (CD) spectroscopy and showed significant correlation (**Table S7, Fig. S5**). In comparison to conventional CD spectroscopy, the fluorescence methods benefit from reduced sample consumption and increased analytical throughput. Concerning the sufficient accuracy, any of the tested microanalytical methods can be used alone. The values of apparent melting temperatures (T_m) obtained reflect the mostly mesophilic origins of the novel HLDs (40-55 °C). DsmA and DppsA exhibited T_m^{app} values correlating with their psychrophilic origin (35.7 and 38.1 °C, respectively). The most stable variant identified was DspoA with T_m^{app} value close to 60°C.

Temperature Profiling. Temperature profiling was performed using a capillary-based droplet microfluidic system.[24] All novel HLDs showed activity over a wide range of temperatures. DmaA showed activity even at 5 °C and DspoA temperature optimum close to 60 °C (**Fig. S6**). Interestingly, DmaA retained more than 60% dehalogenase activity even at 5 °C. A positive correlation was observed between the temperature of the highest observed activity (T_{max}) and the temperature at which protein denaturation starts (T_{onset}) obtained from temperature scanning experiments (**Table S8**). The temperature profiles were used to set the suitable temperatures (5-10 °C bellow T_{max} value) for high-throughput substrate specificity analysis, providing a suitable balance between activity and stability.

Substrate Specificity Profiling. The substrate specificity profiling towards a set of 27 representative substrates was also conducted using a capillary-based droplet microfluidic system (**Table S9**). This set reflects the established application of HLDs in environmental technologies and includes compounds that are environmentally important (**Table S10)**. The raw data of specific activities and the log-transformed data are summarized in **Table S11** and **Table S12**, respectively. The HLDs preferred halogenated substrates in the following order: brominated > iodinated ≫ chlorinated. The optimal substrates of the novel HLDs are linear alkyl-chains of 2-4 carbon atoms (**Fig. 3A)**. Substrates that can be converted by the majority of the enzymes are simultaneously the ones converted with the highest efficiency (**Fig. 3B).** Based on this observation, we suggest a set of "universal" substrates: 1-bromobutane (#18), 1-iodopropane (#28), 1-iodobutane (#29), and 1,2-dibromoethane (#47). The substrate specificity profiling also identified several "recalcitrant" substrates: 1,2-dichloroethane (#37), 1,2-dichloropropane (#67), 1,2,3-trichloropropane (#80), bis(2-chloroethyl)ether (#111), and chlorocyclohexane (#115), which is in good agreement with previous studies.[25] Half of the novel enzymes possess broad substrate specificity and convert >22 of 27 tested substrates (**Fig. S7**). Interestingly, two novel enzymes, DstA and DthA, showed unusually narrow substrate specificity. DstA effectively converted a single substrate, 1-bromohexane (#20), with five-fold higher activity compared to all other substrates. Similarly, DthA exhibits considerable activity for only two substrates, 1,2-dibromoethane (#47) and 1-bromo-2-chloroethane (#137).

Principal Component Analysis (PCA). The first PCA was carried out using the untransformed specificity data of 8 benchmark and 24 newly identified HLDs. This analysis aimed to compare the enzymes according to their score along with the first principal component (t_1), quantifying the overall activity (**Fig. 4**). Surprisingly, 11 of the 24 newly characterized HLDs showed elevated overall activity compared to the benchmark HLDs (**Fig. S8**). The second PCA was performed with log-transformed and weighted substrate specificity data (**Fig. S9**). The benchmark HLDs (DbjA, LinB, DmbA, DhlA and DhaA), in line with the previously reported clustering of HLDs into substrate-specificity groups,[25] remained in proximity. Two of the novel variants, DstA and DthA, are far away from the other enzymes due to their extremely narrow substrate specificity profiles.

Hierarchical Clustering. The log-transformed specificity data were subjected to hierarchical clustering to identify patterns for enzymes (**Fig. 5A**) and substrates (**Fig. 5B**). The data were also plotted as a heatmap with hierarchical clustering dendrograms (**Fig. 5C**). The analysis clustered the substrates into two main groups. The first group (gold in **Fig 5B**) is comprised of frequently converted substrates, mostly iodinated compounds with a chain length of 3-4 carbon atoms. The second group (blue in **Fig 5B**) includes moderately and poorly convertible (mostly chlorinated) substrates. The clustering of enzymes divided HLDs into two major groups. The first group (green in **Fig 5A**) consists of highly active and broad-specificity enzymes, including the benchmark enzymes DhlA, DhaA, DbjA, LinB, and DmbA, capable of converting the majority of the substrates. The second group of enzymes (red in **Fig 5A**) is almost entirely composed of newly identified enzymes (except DatA) with more complex specificity profiles and varied activities. The enzymes forming the second group are barely active with 1,2-dibromopropane (#72), 4-bromobutyronitrile (#141), and 1,2,3-tribromopropane (#154), unlike enzymes from the first group.

Enantioselectivity. Enantioselectivity was assessed by determining the kinetic resolution of representatives of two distinct groups of racemic substrates: β-brominated alkanes (2-bromopentane) and β-brominated esters (ethyl 2-bromopropionate). Individual HLDs show variable enantioselectivity in the reaction with racemic 2-bromopentane. High enantioselectivity was identified for DeaA and DthA, exhibiting E-values of >200 and 156, respectively (**Fig. S10**). Most of the novel HLDs exhibited a preference for the (*R*)- over the (*S*)-enantiomer of 2-bromopentane. Interestingly, the enzymes DmmarA, DspoA, DphxA and DhxA showed the opposite enantiopreference. To date, only two HLD family enzymes (DsvA and eHLD-B) have been reported to possess such unique enantiopreference.[26,27] High enantioselectivity (E-value > 200) towards ethyl 2-bromopropionate was observed in case of DprxA, DthA, and DhxA (**Fig. S11**).

Secondary and Quaternary Structure. Finally, we analyzed the secondary and quaternary structure using far-UV CD spectroscopy and size-exclusion chromatography, respectively, as a validation and

quality control of the protein material. All HLDs exhibited CD spectra with one positive peak at 195 nm and two negative minima at 208 and 222 nm, characteristic for proteins with an α/β-hydrolase fold (**Fig. S12**). Newly identified HLDs were mostly monomeric, similar to the previously characterized HLD members. Several variants can form dimers apart from the monomers (**Table S13**). The exceptions are DmmarA, which exists as a dimer, and DprxA, which exists as a mixture of monomer, dimer, and higher oligomeric states, respectively (**Fig. S13**). DstA forms dimers under oxidative conditions and its oligomeric state thus depends on the oxidation/reduction potential of the environment.

DISCUSSION

The field of biotechnology employing enzymes represents a billion-dollar industry, putting constant pressure on speeding up the process of identification, acquisition, and characterization of novel biocatalysts. The avalanche of newly available sequences from next-generation sequencing represents enormous potential, but also a significant challenge for the practical discovery of novel enzymes. The application of rational genome mining can facilitate effective management of the large quantity of complex sequence data [28]. It is currently not feasible to characterize all sequences being deposited in sequence databases. *In silico* screening and prioritization, followed by miniaturized high-throughput characterization, appears to be an attractive approach. In this study, we have integrated computational genome mining with high-throughput microanalytical and microfluidic techniques. Only 63% of the identified putative HLDs were labelled correctly as dehalogenating enzymes in genomic databases. While missannotations were rare, many proteins annotated as "α/β-hydrolase" or "hypothetical protein" would have been missed by a simple text-based search. Proteins from the α/β-hydrolase superfamily are well-known for their catalytic promiscuity and tendency to catalyze diverse reactions using the same catalytic machinery.[29] Substrates are not known for 35% of enzymes annotated as α/β-hydrolases and thus their functions remain unclear.[30] The current mining identified more than 2,578 putative HLDs, that is nearly five times more hits than in the previous *in silico* screening conducted in 2013 (530 putative HLDs). Current screening missed only 97 sequences out of the original set and identified 2,145 new sequences.

Although *in silico* screening strategies for identifying novel enzymes are being used profitably,[31–33] automation of the selection of a limited number of promising hits for characterization is challenging. Several sequence-based analysis tools have been developed for the prediction of key protein characteristics, e.g., thermostability,[34] optimum pH[31] or protein solubility.[35–38] Other computational tools help to analyze, filter and visualize the large sets of identified hits.[19,20] Automated *in silico* workflow optimized within this study has been released as the user-friendly web tool EnzymeMiner (https://loschmidt.chemi.muni.cz/enzymeminer/), making analysis described in this article accessible to the scientific and industrial communities.[18] Prediction of protein solubility proved to be one of the

limitations. Despite the application of a solubility prediction tool,[35] our comprehensive expression analysis of the whole set of 45 selected putative HLDs revealed only 67 % success rate in terms of soluble proteins. A similar result was achieved by the previous screening of novel HLDs where only 60% of the constructed variants could be expressed in soluble form. Protein production in *E. coli* can be improved by optimizing genetic constructs and medium engineering, but combinatorial variation is impractical for such a large set of proteins. Therefore, the production of soluble proteins remains a hit-or-miss affair and currently represents the biggest bottleneck in the functional and structural characterization of novel proteins. Improvement of the *in silico* solubility prediction is paramount for the increased success rate of protein characterization pipelines.[14,39]

An important component of the experimental workflow is the application of time- and material-efficient microscale methods. These techniques can be miniaturized and parallelized, allowing high throughput with low demands on the amount of biological material.[40] The comparison of conventional methods with applied microscale and droplet-based microfluidic techniques presented here demonstrate high accuracy and reliability of the miniaturized assays. The gradual replacement of the conventional methods by their miniaturized versions is inevitable. Our study also enriched the toolbox of HLDs available for biotechnological applications.[17] Homology modelling followed by calculation of the active site volumes is a powerful approach for identification of enzymes with high catalytic activities: 11 of 24 characterized HLDs showed activities higher than most previously published enzymes.[25] The small, occluded and desolvated active site is favorable for dehalogenation reactions catalyzed by S_N2 reaction mechanism.[41] Moreover, HLDs selective towards one or two substrates (DstA and DthA) were described for the first time. Several novel enzymes showed high enantioselectivity towards representative β-brominated alkane and β-brominated ester substrates. The enzymes DmmarA, DspoA, DphxA and DhxA showed a rare (*S*)-enantiopreference with 2-bromopentane. Temperature profiling identified DmaA, which retained >60% of residual activity at 5°C, which is an attractive property for application of HLDs as environmental biosensors.[42] DspoA is an example of a biocatalyst possessing an attractive combination of industrially relevant properties such as high activity, stability and selectivity.

CONCLUSIONS

We describe an integrated workflow for the identification of novel family members using computer-assisted genome mining and high-throughput experimental characterization. The proposed *in silico* pipeline employs sequence similarity searches accompanied by annotation and structural bioinformatics analyses. The experimental characterization was accelerated and scaled-down by applying high-throughput microanalytical and microfluidic methods. These miniaturized techniques are suitable for characterization of smaller to medium sets of novel enzymes (tens to hundreds). The integration of the high-throughput techniques for small-scale protein production and microscale characterization of enzymes resulted in a robust workflow. An interesting perspective appears to be the application of cell-free methods for protein production and its integration into microfluidic systems, which will lead to a simplification of the workflow and a further increase in its throughput. Cell-free protein production did not improve the acquisition of soluble variants in our case. Therefore, the development of more reliable prediction tools to enrich proteins with good solubility remains an important challenge. In this study, we illustrated that repetitive database mining provides a variety of novel enzymes with valuable industrial properties, while it also enables collection of consistent experimental data. High-quality datasets are attractive for statistical and machine learning methods which may in the future provide an understanding of sequence-function relationships and contribute to the development of a new generation of tools in protein engineering.

EXPERIMENTAL SECTION

In Silico Screening. A previously developed *in silico* pipeline for identification and characterization of putative HLDs was employed.[8] To automate and improve the *in silico* protocol, several innovations were introduced to the original pipeline. Briefly, the sequences of three experimentally characterized HLDs [LinB (accession number to NCBI BAA03443), DhlA (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[43] v2.6.0 searches against the NCBI nr database (version 2017/02) with E-value thresholds of 10^{-20}. A multiple sequence alignment of all putative full-length HLD sequences was constructed by Clustal Omega v1.2.0.[44] Sequence similarity networks of putative HLDs were calculated and visualized by EFI-EST[19] and Cytoscape v3.6.1,[20] respectively. The obtained SSN was subjected to the Enzyme Function Initiative Genome Neighborhood Tool analysis to obtain genome neighborhood diagrams. Information about the source organisms of all putative HLDs was collected from the NCBI Taxonomy and BioProject databases (version 2017/02).[21] The homology modelling was performed using MODELLER v9.18.[45] The quality of generated homology models was assessed by MolProbity v4.3.1.[22] Pockets in each homology model were calculated and measured using the CASTp program[46] with a probe radius of 1.4 Å. The CAVER v3.02 program[47] was then used to calculate tunnels in the

ensemble of all homology models. The probability of soluble expression in *E. coli* of each protein was predicted based on the revised Wilkinson-Harrison solubility model.[48]

Gene Synthesis and DNA Manipulation. The codon-optimized genes encoding 45 HLDs were designed and commercially synthesized (BaseClear B.V., The Netherlands). The synthetic genes were subcloned individually into the expression vector pET24a(+) between NdeI/XhoI restriction sites. Competent *E. coli* DH5α cells were transformed with individual constructs for plasmid propagation using a heat-shock method. The correct insertions of target HLD genes into recombinant plasmids were verified by restriction analysis of the re-isolated plasmids (**Fig. S2**) and DNA sequencing.

Small-Scale Protein Expression and Purification. Competent *E. coli* BL21(DE3) cells were transformed with pET24a(+)::*HLD*x (x = 45 different HLDs candidates) plasmid DNA using a heat shock method, plated on 2x lysogeny broth agar plates with kanamycin (50 μg.ml^{-1}) and grown overnight at 37 °C. The next day, single kanamycin-resistant colonies were transferred into wells of a 2 ml 96-deep well square plate containing 500 µl of 2xLB medium with 50 μg.ml^{-1} kanamycin and cells were grown at 37 °C for 5-6 hours with shaking at 300 rpm. For all HLD gene candidates, plates were inoculated in duplicate. Next, using 96-deep well plates, 40 µl of the culture was inoculated in 450 µl of 2xLB medium supplemented with 50 μg.ml^{-1} kanamycin. The cultures were cultivated at 37 °C for 1.5-2 hours with shaking at 300 rpm until $O.D_{600}$ reached 0.4 – 0.6. Expression was induced by addition of IPTG to a final concentration of 0.5 mM and cultivation was continued at 22 °C for 24 h. The cells were harvested by centrifugation (3,500 rpm, 10,000 *g*, 20 min at 4 °C). The cells were washed three times with 200 µl PB buffer (40mM K_2HPO_4, 10mM KH_2PO_4, pH 7.5). Cells were disrupted by three cycles of ultrasonication (3 min with 50% frequency and 50% amplitude) using an Elma Ultrasonic Cleaner S100H (Elma Schmidbauer GmbH, Germany). Lysate was clarified by centrifugation at 10,000 *g* at 4°C for 1 h. Protein expression was analyzed by SDS-PAGE using 12.5% polyacrylamide gels. Proteins were visualized using a Coomassie Brilliant Blue staining solution (1% w/v α-cyclodextrin, 4.25% w/v phosphoric acid, and 0.5x Roti® -Nanoquant). The rest of the sample was subjected to high-throughput affinity purification using the MagneHis Protein Purification System (Promega, USA) according to the manufacturer's manual, with minor modifications. The washing and elution buffers were supplemented with 500 mM NaCl. The elution was performed by 100 µl of an elution buffer containing 250 mM imidazole. Finally, a desalting plate (Merck KGaA, Germany) was used (3x times at 3,700 rpm, 10,000 *g*, for 10 min) to exchange from the elution buffer to the storage buffer PB. 100 µl of PB was added to each well between centrifugation steps. The enzymes were dissolved in 100 µl of PB. The presence of HLDs was proven by SDS-PAGE using 15% polyacrylamide gels stained by Coomassie Brilliant Blue R-250 dye (Fluka, Switzerland).

Dehalogenase Activity Screening. The reactions were 200 µL in volume and contained 50 mM PBO buffer (40mM K_2HPO_4, 10mM KH_2PO_4, pH 7.5 with 1 mM orthovanadate), 10 mM H_2O_2, 5 $U.ml^{-1}$ *Curvularia inaequalis* chloroperoxidase, 4 µg purified protein, 12.5 µM APF and 1 mM of substrate. The reactions in HOX assay[23] were started by addition of purified protein and measured in a plate reader at 525 nm (488 nm excitation) and 30 °C overnight.

Cell-Free Protein Synthesis. The Cell-free protein synthesis (CFPS) of 45 selected HLDs was performed using the PURExpress kit (NEB, USA) according to the manufacturer's instructions.[49] The recommended 250 ng of DNA template per reaction was used. The CFPS reactions were incubated at 37 °C for 2.5 h. To maintain precise reaction conditions, a thermocycler was used for temperature control. The total fractions of HLDs were detected by SDS-PAGE stained by Coomassie Brilliant Blue R-250 and by silver staining (SilverQuest, Fermentas, USA). Subsequently, the total fractions of HLDs were centrifuged at 10,000 *g* at 4 °C for 1 h. The rest of the sample was dialyzed using Slide-A-Lyzer MINI Dialysis Devices (ThermoFisher Scientific, Germany) into the PB buffer used for the screening of HLD activity using the HOX assay. [23]

Large-Scale Protein Expression and Purification. Selected mutant enzymes were overproduced in *E. coli* BL21(DE3). A single colony was used to inoculate 10 ml of LB medium with kanamycin (to a final concentration of 50 $\mu g.ml^{-1}$) and cells were grown at 37 °C for 4.5-5 hours. The preculture was used to inoculate 1 L of LB medium with kanamycin (50 $\mu g.ml^{-1}$). Cells were cultivated at 37 °C for 1.5-2 hours until $O.D_{600}$ reached 0.4 – 0.6. The expression was induced with IPTG to a final concentration of 0.5 mM. Cells were then cultivated at 20 °C overnight. At the end of cultivation, biomass was harvested by centrifugation (20 min; 3,500 *g*, 4 °C) and immediately resuspended in the purification buffer A (20 mM K_2HPO_4/ KH_2PO_4, pH 7.5, 500 mM NaCl, 10 mM imidazole). DNaseI was added to the final concentration of 1.25 $\mu g.ml^{-1}$ of cell suspension. Cells in suspension were disrupted by ultrasonication using a Hielscher UP200S ultrasonic processor (Teltow, Germany) with 0.3 s pulses and 70 % amplitude. The cell lysates were centrifuged for 1 h at 21,000 *g* at 4 °C. The crude extracts were decanted and total protein concentration was determined using Bradford solution (Sigma-Aldrich, USA).

Overexpressed HLDs were purified using single-step nickel affinity chromatography. The cell-free extract was applied to a 5 ml Ni-nitrilotriacetic acid (Ni-NTA) Superflow column charged with Ni^{2+} ions (Qiagen, Germany) in the equilibrating buffer (purification buffer A). Target proteins were eluted with an increasing two-step gradient. First, unbound and weakly bound proteins were washed out with a 10% gradient of purification buffer B (20 mM potassium phosphate, pH 7.5, 500 mM NaCl, 300 mM imidazole). Subsequently, the target proteins were eluted with a 60% gradient of purification buffer B. Enzymes eluted by 300 mM imidazole (60% gradient) by metal-affinity chromatography were loaded

on an ÄKTA FPLC system (GE Healthcare) equipped with a UV_{280} detector and a HiLoad 16/600 Superdex 200 prep grade column (GE Healthcare, Uppsala, Sweden) equilibrated in 50 mM potassium phosphate (pH 7.5). Elution was done using the same purification buffer at a constant flow rate of 1 $ml.min^{-1}$. The protein purity was checked by SDS-PAGE using 15% polyacrylamide gels stained with Coomassie Brilliant Blue R-250 (Fluka, Switzerland). The molecular weights were estimated using the Unstained Protein Molecular Weight Marker (Thermo Scientific, USA). The total enzyme concentration was determined by measuring absorbance at 280 nm using a NanoDrop (ThermoFischer Scientific, USA).

Thermostability. Thermal unfolding was analyzed by four independent methods: (i) circular dichroism spectroscopy, (ii) DSF (thermal shift assay), (iii) nano DSF (UNcle), (iv) nano DSF (Prometheus). Circular dichroism spectroscopy (CD) was employed as a well-established technique. The unfolding of an enzyme (0.2 $mg.ml^{-1}$ in 50 mM phosphate buffer, pH 7.5) was monitored by the change in the ellipticity at a wavelength with the highest difference in ellipticity over the temperature range from 15 to 90 °C at a 1 $°C.min^{-1}$ scan rate. The thermal denaturation curves recorded were 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. Thermal shift assay was conducted in MicroAmp Fast Optical 96-well Reaction Plates (Thermo Fisher Scientific). Each reaction mixture of the final volume of 25 µl was composed of 2 µl of SYPRO Orange Protein Gel Stain (Thermo Fisher Scientific), enzyme (1 $mg.ml^{-1}$ dialyzed in 50 mM potassium phosphate buffer, pH 7.5) and 50 mM potassium phosphate buffer of pH 7.5. The assay was performed using a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific) from 20 to 90 °C at 1 $°C.min^{-1}$ scan rate. The T_m values were determined from obtained data using Protein Thermal Shift software (Thermo Fisher Scientific). Two nano differential scanning fluorimetry (nanoDSF) techniques based on tryptophan or tyrosine fluorescence were employed. Nano DSF (UNcle, Unchained labs) measured the temperature-induced denaturation of enzymes (1 $mg.ml^{-1}$ in 50 mM phosphate buffer of pH 7.5) by monitoring changes in fluorescence spectra (excitation at 266 nm) from 15 to 90 °C at 1 $°C.min^{-1}$ scan rate. Thermostability was determined from the midpoint of the barycentric mean fluorescence (BCM) curve. A Prometheus NT.48 scanning fluorometer (NanoTemper Technologies, GmbH) measured temperature-induced denaturation of enzymes (1 $mg.ml^{-1}$ in 50 mM potassium phosphate buffer, pH 7.5) by monitoring changes in fluorescence signal at 335 and 350 nm from 15 to 90 °C at 1 $°C.min^{-1}$ scan rate. The ratio of fluorescence intensities at both excitation wavelengths (corresponding to the "redshift" of the tryptophan fluorescence upon protein unfolding) was plotted as a function of temperature and the point of inflexion of the resulting curve was used as a thermostability parameter.

Secondary Structure. 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 nm.min^{-1}, with 1-s response time and 1-nm bandwidth, using a 0.1-cm quartz cuvette containing the 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 (ΘMRE). Secondary structure determination and analysis were carried out on measured ellipticity from 190 to 250 nm using the BeStSel online tool with default settings.[50]

Quaternary Structure. The protein quaternary structures were investigated using analytical gel filtration chromatography using a Superdex 200 10/300 GL column (GE Healthcare Life Sciences). The ÄKTA FPLC system (GE Healthcare Life Sciences) was initially equilibrated with a mobile phase composed of 50 mM potassium phosphate buffer and 150 mM NaCl (pH 7.5). The protein sample (100 µl at 1 mg.ml^{-1}) was injected onto the column and separated at a constant flow rate of 0.5 ml.min^{-1} using the mobile phase described above. The void volume was determined by loading blue dextran (100 µl at 1 mg.ml^{-1}). Two gel filtration calibration mixtures were applied for molecular weight determination (GE Healthcare Life Sciences). The mixture A of standard proteins contained aldolase (158,000 Da), ovalbumin (44,000 Da), ribonuclease A (13,700 Da), and aprotinin (6,500 Da). The mixture B of standard proteins contained ferritin (440,000 Da), conalbumin (75,000 Da), carbonic anhydrase (29,000 Da), and ribonuclease A (13,700 Da).

Enantioselectivity. Kinetic resolution experiments were performed at 20 °C. The reaction mixtures 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 detailed description is provided in Vanacek and co-workers[8]. The kinetic resolution data were fitted globally using KinTek Explorer software (KinTek Corporation, USA). The competitive steady-state model (***Scheme 1***) was used to obtain estimates of specificity constants k_{cat}/K_m[51] for both *R* and *S* enantiomers during the conversion of the racemic mixture, where *E* is an enzyme, S_i and P_i are substrate and product, respectively. The index *i* depicts the (*S*)-enantiomer or (*R*)-enantiomer, respectively. The estimate for enantioselectivity of the enzymatic reaction, defined as the ratio of specificity constants for the conversion of *S* and *R* enantiomers (*E*-value, ***Equation 1***), was obtained during the fitting procedure by fixing the ratio between individual values of k_{cat}/K_m for *R* and *S* enantiomers.

$$E + S_R \xrightarrow{\left(\frac{k_{cat}}{K_m}\right)_R} ES_R \xrightarrow{(k_{cat})_R} E + P_S$$

$$E + S_S \xrightarrow{\left(\frac{k_{cat}}{K_m}\right)_S} ES_S \xrightarrow{(k_{cat})_S} E + P_R$$

Scheme 1

$$E - value = \frac{\left(\frac{k_{cat}}{K_m}\right)_R}{\left(\frac{k_{cat}}{K_m}\right)_S}$$

Equation 1

The spontaneous conversion of the substrate (reaction without enzyme) was included in the model and analyzed globally to obtain a specific effect of enzyme selectivity. Numerical integration of rate equations searching a set of kinetic parameters that produce a minimum $\chi 2$ value was performed using the Bulirsch–Stoer algorithm with adaptive step size, and nonlinear regression to fit data was based on the Levenberg–Marquardt method. To account for fluctuations in experimental data, enzyme or substrate concentrations were slightly adjusted at a boundary ± 5 % to derive best fits. Residuals were normalized by sigma value for each data point. The standard error (S.E.) was calculated from the covariance matrix during nonlinear regression.[52]

Substrate Specificity Profiles and Temperature Profiles. Both substrate specificity and temperature profiles were measured on the capillary-based droplet microfluidic platform, enabling the characterization of specific enzyme activity within droplets for typically 6-10 variants in one run. The temperature profiles were measured towards either 1,2-dibromoethane or 1-bromohexane in 5-degree increments in the range of 5 °C to 55 °C. The temperatures for individual enzymes were chosen based on their T_m and T_{onset} values (determined by Prometheus) so that the activities at 7-9 temperatures were measured for each enzyme. The substrate specificity of individual enzyme variants was measured towards 27 representative halogenated substrates, which were previously chosen for validation of the microfluidic platform. Each enzyme was assayed at the temperature closest to its T_{max} value or where it retains > 90% of activity at that temperature. A detailed description of the microfluidic method was provided previously.[53] Briefly, the droplets were generated using Mitos Dropix (Dolomite, UK). A custom sequence of droplets (150 nL aqueous phase, 300 nL oil spacing) was generated using negative pressure (microfluidic pump) and the droplets were guided through a polythene tubing to the incubation chamber. Within the incubation chamber, the halogenated substrate was delivered to the droplets via a combination of microdialysis and partitioning between the oil (FC-40) and the aqueous phase. The reaction solution consisted of a weak buffer (1 mM HEPES, 20 mM Na_2SO_4, pH 8.2) and a complementary fluorescent indicator 8-hydroxypyrene-1,3,6-trisulfonic acid (50 μM HPTS). The buffer exchange of enzyme samples was carried out using the standard spin protocol of PD Minitrap™ G-25 (GE Healthcare, USA), where 2 centrifugation steps were applied, each at 1,000 g for 2 min. The fluorescence signal was obtained by using an optical setup with excitation laser (450 nm), a dichroic mirror with a cut-off at 490 nm filtering the excitation light and a Si-detector (GE Healthcare, USA). By employing a pH-based fluorescence assay, small changes in the pH were observed, enabling monitoring of the enzymatic activity. Reaction progress was analyzed as an end-point measurement

recorded after passing of 10 droplets/sample through the incubation chamber. The reaction time was 4 min. The raw signal was processed by a droplet detection script written in MATLAB 2017b (MathWorks, USA) to obtain the specific activities. The raw signal of every single measurement was at first processed by a LabView-based code (National Instruments, USA) developed in-house. Using this software, the peaks were assigned to the particular sample and the mean signal was calculated for them. The output XLS file gathering mean signal values for every sample type (calibration, enzyme activity, buffer, blank buffer and blank enzyme) for a particular dataset (e.g., 6 enzymes measured in one temperature with all 27 substrates) served as an input for the MATLAB script (MathWorks, USA) calculating the specific activities using the same principle described previously.[24,53]

Principal Component Analysis and Hierarchical Clustering. The matrix containing the activity data of 24 novel HLDs and 8 previously characterized HLDs towards 27 halogenated substrates was analyzed by Principal Component Analysis (PCA) to uncover the relationships among individual HLDs (objects) based on their activities towards the set of halogenated substrates (variables). Two PCA models were constructed to visualize systematic trends in the dataset. The first one was done on the raw data, which as a result ordered the enzymes according to their total activity. The second PCA was carried out on the log-transformed data. Each specific activity needed to be incremented by 1 to avoid the logarithm of zero values. The resulting values were then divided by the sum of the values for a particular enzyme, and weighted values were estimated. These transformed data were used to calculate principal components, and the components explaining the highest variability in the data were then plotted for identification of substrate specificity groups. Additionally, the hierarchical clustering analysis was performed on the log-transformed data using MATLAB (MathWorks, USA).

ACKNOWLEDGEMENTS

We would like to thank to Simon Godehard and Mark Dörr (University Greifswald, Germany) for fruitful discussions of the experimental design of cell-free expression experiments. Authors would like to acknowledge funding from the Czech Ministry of Education (CZ.02.1.01/0.0/0.0/17_043/0009632, CZ.02.1.01/0.0/0.0/16_026/0008451, LM2018121 and LM2015047). This project has received funding from the European Union's Horizon 2020 research and Innovation programme (857560 and 814418). The article reflects the author's view and the Agency is not responsible for any use that may be made of the information it contains.

ORCID:

Christoffel P. S. Badenhorst: 0000-0002-5874-4577

Stanislav Mazurenko: 0000-0003-3659-4819

David Bednar: 0000-0002-6803-0340

Uwe Bornscheuer: 0000-0003-0685-2696

Jiri Damborsky: 0000-0002-7848-8216

Zbynek Prokop: 0000-0001-9358-4081

REFERENCES

(1) Fernández-Arrojo, L.; Guazzaroni, M. E.; López-Cortés, N.; Beloqui, A.; Ferrer, M. Metagenomic Era for Biocatalyst Identification. *Current Opinion in Biotechnology*. 2010. https://doi.org/10.1016/j.copbio.2010.09.006.

(2) Truppo, M. D. Biocatalysis in the Pharmaceutical Industry: The Need for Speed. *ACS Medicinal Chemistry Letters* **2017**, *8* (5), 476–480. https://doi.org/10.1021/acsmedchemlett.7b00114.

(3) Copp, J. N.; Akiva, E.; Babbitt, P. C.; Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **2018**, *57* (31), 4651–4662. https://doi.org/10.1021/acs.biochem.8b00473.

(4) Furnham, N.; Garavelli, J. S.; Apweiler, R.; Thornton, J. M. Missing in Action: Enzyme Functional Annotations in Biological Databases. *Nature chemical biology* **2009**, *5*, 521–525. https://doi.org/10.1038/nchembio0809-521.

(5) Mak, W. S.; Tran, S.; Marcheschi, R.; Bertolani, S.; Thompson, J.; Baker, D.; Liao, J. C.; Siegel, J. B. Integrative Genomic Mining for Enzyme Function to Enable Engineering of a Non-Natural Biosynthetic Pathway. *Nature Communications* **2015**, *6*, 10005. https://doi.org/10.1038/ncomms10005.

(6) Marshall, J. R.; Yao, P.; Montgomery, S. L.; Finnigan, J. D.; Thorpe, T. W.; Palmer, R. B.; Mangas-Sanchez, J.; Duncan, R. A. M.; Heath, R. S.; Graham, K. M.; Cook, D. J.; Charnock, S. J.; Turner, N. J. Screening and Characterization of a Diverse Panel of Metagenomic Imine Reductases for Biocatalytic Reductive Amination. *Nature Chemistry* **2020**, 1–9. https://doi.org/10.1038/s41557-020-00606-w.

(7) Lobb, B.; Doxey, A. C. Novel Function Discovery through Sequence and Structural Data Mining. *Current Opinion in Structural Biology* **2016**, *38*, 53–61. https://doi.org/10.1016/j.sbi.2016.05.017.

(8) Vanacek, P.; Sebestova, E.; Babkova, P.; Bidmanova, S.; Daniel, L.; Dvorak, P.; Stepankova, V.; Chaloupkova, R.; Brezovsky, J.; Prokop, Z.; Damborsky, J. Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catalysis* **2018**, *8* (3), 2402–2412. https://doi.org/10.1021/acscatal.7b03523.

(9) Zhou, N.; Jiang, Y.; Bergquist, T. R.; Lee, A. J.; Kacsoh, B. Z.; Crocker, A. W.; Lewis, K. A.; Georghiou, G.; Nguyen, H. N.; Hamid, M. N.; Davis, L.; Dogan, T.; Atalay, V.; Rifaioglu, A. S.;

Dalkıran, A.; Cetin Atalay, R.; Zhang, C.; Hurto, R. L.; Freddolino, P. L.; Zhang, Y.; Bhat, P.; Supek, F.; Fernández, J. M.; Gemovic, B.; Perovic, V. R.; Davidović, R. S.; Sumonja, N.; Veljkovic, N.; Asgari, E.; Mofrad, M. R. K.; Profiti, G.; Savojardo, C.; Martelli, P. L.; Casadio, R.; Boecker, F.; Schoof, H.; Kahanda, I.; Thurlby, N.; McHardy, A. C.; Renaux, A.; Saidi, R.; Gough, J.; Freitas, A. A.; Antczak, M.; Fabris, F.; Wass, M. N.; Hou, J.; Cheng, J.; Wang, Z.; Romero, A. E.; Paccanaro, A.; Yang, H.; Goldberg, T.; Zhao, C.; Holm, L.; Törönen, P.; Medlar, A. J.; Zosa, E.; Borukhov, I.; Novikov, I.; Wilkins, A.; Lichtarge, O.; Chi, P.-H.; Tseng, W.-C.; Linial, M.; Rose, P. W.; Dessimoz, C.; Vidulin, V.; Dzeroski, S.; Sillitoe, I.; Das, S.; Lees, J. G.; Jones, D. T.; Wan, C.; Cozzetto, D.; Fa, R.; Torres, M.; Warwick Vesztrocy, A.; Rodriguez, J. M.; Tress, M. L.; Frasca, M.; Notaro, M.; Grossi, G.; Petrini, A.; Re, M.; Valentini, G.; Mesiti, M.; Roche, D. B.; Reeb, J.; Ritchie, D. W.; Aridhi, S.; Alborzi, S. Z.; Devignes, M.-D.; Koo, D. C. E.; Bonneau, R.; Gligorijević, V.; Barot, M.; Fang, H.; Toppo, S.; Lavezzo, E.; Falda, M.; Berselli, M.; Tosatto, S. C. E.; Carraro, M.; Piovesan, D.; Ur Rehman, H.; Mao, Q.; Zhang, S.; Vucetic, S.; Black, G. S.; Jo, D.; Suh, E.; Dayton, J. B.; Larsen, D. J.; Omdahl, A. R.; McGuffin, L. J.; Brackenridge, D. A.; Babbitt, P. C.; Yunes, J. M.; Fontana, P.; Zhang, F.; Zhu, S.; You, R.; Zhang, Z.; Dai, S.; Yao, S.; Tian, W.; Cao, R.; Chandler, C.; Amezola, M.; Johnson, D.; Chang, J.-M.; Liao, W.-H.; Liu, Y.-W.; Pascarelli, S.; Frank, Y.; Hoehndorf, R.; Kulmanov, M.; Boudellioua, I.; Politano, G.; Di Carlo, S.; Benso, A.; Hakala, K.; Ginter, F.; Mehryary, F.; Kaewphan, S.; Björne, J.; Moen, H.; Tolvanen, M. E. E.; Salakoski, T.; Kihara, D.; Jain, A.; Šmuc, T.; Altenhoff, A.; Ben-Hur, A.; Rost, B.; Brenner, S. E.; Orengo, C. A.; Jeffery, C. J.; Bosco, G.; Hogan, D. A.; Martin, M. J.; O'Donovan, C.; Mooney, S. D.; Greene, C. S.; Radivojac, P.; Friedberg, I. The CAFA Challenge Reports Improved Protein Function Prediction and New Functional Annotations for Hundreds of Genes through Experimental Screens. *Genome Biology* **2019**, *20* (1), 244. https://doi.org/10.1186/s13059-019-1835-8.

(10) Li, Y.; Wang, S.; Umarov, R.; Xie, B.; Fan, M.; Li, L.; Gao, X. DEEPre: Sequence-Based Enzyme EC Number Prediction by Deep Learning. *Bioinformatics* **2018**, *34* (5), 760–769. https://doi.org/10.1093/bioinformatics/btx680.

(11) Colin, P.-Y.; Kintses, B.; Gielen, F.; Miton, C. M.; Fischer, G.; Mohamed, M. F.; Hyvönen, M.; Morgavi, D. P.; Janssen, D. B.; Hollfelder, F. Ultrahigh-Throughput Discovery of Promiscuous Enzymes by Picodroplet Functional Metagenomics. *Nature Communications* **2015**, *6*, 10008. https://doi.org/10.1038/ncomms10008.

(12) Beneyton, T.; Thomas, S.; Griffiths, A. D.; Nicaud, J.-M.; Drevelle, A.; Rossignol, T. Droplet-Based Microfluidic High-Throughput Screening of Heterologous Enzymes Secreted by the Yeast Yarrowia Lipolytica. *Microb Cell Fact* **2017**, *16*. https://doi.org/10.1186/s12934-017-0629-5.

(13) Kokkonen, P.; Koudelakova, T.; Chaloupkova, R.; Daniel, L.; Prokop, Z.; Damborsky, J. Structure-Function Relationships and Engineering of Haloalkane Dehalogenases. In *Aerobic Utilization of Hydrocarbons, Oils and Lipids*; Rojo, F., Ed.; Handbook of Hydrocarbon and Lipid Microbiology; Springer International Publishing: Cham, 2017; pp 1–21. https://doi.org/10.1007/978-3-319-39782-5_15-1.

(14) Musil, M.; Konegger, H.; Hon, J.; Bednar, D.; Damborsky, J. Computational Design of Stable and Soluble Biocatalysts. *ACS Catal.* **2019**, *9* (2), 1033–1054. https://doi.org/10.1021/acscatal.8b03613.

(15) Beerens, K.; Mazurenko, S.; Kunka, A.; Marques, S. M.; Hansen, N.; Musil, M.; Chaloupkova, R.; Waterman, J.; Brezovsky, J.; Bednar, D.; Prokop, Z.; Damborsky, J. Evolutionary Analysis As a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catal.* **2018**, *8* (10), 9420–9428. https://doi.org/10.1021/acscatal.8b01677.

(16) Brezovsky, J.; Babkova, P.; Degtjarik, O.; Fortova, A.; Gora, A.; Iermak, I.; Rezacova, P.; Dvorak, P.; Smatanova, I. K.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. Engineering a de Novo Transport Tunnel. *ACS Catal.* **2016**, *6* (11), 7597–7610. https://doi.org/10.1021/acscatal.6b02081.

(17) Koudelakova, T.; Bidmanova, S.; Dvorak, P.; Pavelka, A.; Chaloupkova, R.; Prokop, Z.; Damborsky, J. Haloalkane Dehalogenases: Biotechnological Applications. *Biotechnology Journal* **2013**, *8* (1), 32–45. https://doi.org/10.1002/biot.201100486.

(18) Hon, J.; Borko, S.; Stourac, J.; Prokop, Z.; Zendulka, J.; Bednar, D.; Martinek, T.; Damborsky, J. EnzymeMiner: Automated Mining of Soluble Enzymes with Diverse Structures, Catalytic Properties and Stabilities. *Nucleic Acids Research* **2020**, *48* (W1), W104–W109. https://doi.org/10.1093/nar/gkaa372.

(19) Gerlt, J. A.; Bouvier, J. T.; Davidson, D. B.; Imker, H. J.; Sadkhin, B.; Slater, D. R.; Whalen, K. L. Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A Web Tool for Generating Protein Sequence Similarity Networks. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **2015**, *1854* (8), 1019–1037. https://doi.org/10.1016/j.bbapap.2015.04.015.

(20) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res* **2003**, *13* (11), 2498–2504. https://doi.org/10.1101/gr.1239303.

(21) Barrett, T.; Clark, K.; Gevorgyan, R.; Gorelenkov, V.; Gribov, E.; Karsch-Mizrachi, I.; Kimelman, M.; Pruitt, K. D.; Resenchuk, S.; Tatusova, T.; Yaschenko, E.; Ostell, J. BioProject and BioSample Databases at NCBI: Facilitating Capture and Organization of Metadata. *Nucleic Acids Research* **2012**, *40* (D1), D57–D63. https://doi.org/10.1093/nar/gkr1163.

(22) Williams, C. J.; Headd, J. J.; Moriarty, N. W.; Prisant, M. G.; Videau, L. L.; Deis, L. N.; Verma, V.; Keedy, D. A.; Hintze, B. J.; Chen, V. B.; Jain, S.; Lewis, S. M.; Arendall, W. B.; Snoeyink, J.; Adams, P. D.; Lovell, S. C.; Richardson, J. S.; Richardson, D. C. MolProbity: More and Better Reference Data for Improved All-Atom Structure Validation. *Protein Sci* **2018**, *27* (1), 293–315. https://doi.org/10.1002/pro.3330.

(23) Aslan-Üzel, A. S.; Beier, A.; Kovář, D.; Cziegler, C.; Padhi, S. K.; Schuiten, E. D.; Dörr, M.; Böttcher, D.; Hollmann, F.; Rudroff, F.; Mihovilovic, M. D.; Buryška, T.; Damborský, J.; Prokop, Z.; Badenhorst, C. P. S.; Bornscheuer, U. T. An Ultrasensitive Fluorescence Assay for the Detection of Halides and Enzymatic Dehalogenation. *ChemCatChem* **2019**. https://doi.org/10.1002/cctc.201901891.

(24) Buryska, T.; Vasina, M.; Gielen, F.; Vanacek, P.; van Vliet, L.; Jezek, J.; Pilat, Z.; Zemanek, P.; Damborsky, J.; Hollfelder, F.; Prokop, Z. Controlled Oil/Water Partitioning of Hydrophobic Substrates Extending the Bioanalytical Applications of Droplet-Based Microfluidics. *Anal. Chem.* **2019**, *91* (15), 10008–10015. https://doi.org/10.1021/acs.analchem.9b01839.

(25) Koudelakova, T.; Chovancova, E.; Brezovsky, J.; Monincova, M.; Fortova, A.; Jarkovsky, J.; Damborsky, J. Substrate Specificity of Haloalkane Dehalogenases. *Biochem J* **2011**, *435* (2), 345–354. https://doi.org/10.1042/BJ20101405.

(26) Chmelova, K.; Sebestova, E.; Liskova, V.; Beier, A.; Bednar, D.; Prokop, Z.; Chaloupkova, R.; Damborsky, J. A Haloalkane Dehalogenase from Saccharomonospora Viridis Strain DSM 43017, a Compost Bacterium with Unusual Catalytic Residues, Unique (S)-Enantiopreference, and High Thermostability. *Appl. Environ. Microbiol.* **2020**, *86* (17). https://doi.org/10.1128/AEM.02820-19.

(27) Kotik, M.; Vanacek, P.; Kunka, A.; Prokop, Z.; Damborsky, J. Metagenome-Derived Haloalkane Dehalogenases with Novel Catalytic Properties. *Appl Microbiol Biotechnol* **2017**, *101* (16), 6385–6397. https://doi.org/10.1007/s00253-017-8393-3.

(28) Zaparucha, A.; de Berardinis, V.; Vaxelaire-Vergne, C. Chapter 1 Genome Mining for Enzyme Discovery. In *Modern Biocatalysis: Advances Towards Synthetic Biological Systems*; The Royal Society of Chemistry, 2018; pp 1–27. https://doi.org/10.1039/9781788010450-00001.

(29) Marchot, P.; Chatonnet, A. Enzymatic Activity and Protein Interactions in Alpha/Beta Hydrolase Fold Proteins: Moonlighting Versus Promiscuity. *Protein & Peptide Letters* **2012**, *19* (2), 132–143. https://doi.org/10.2174/092986612799080284.

(30) Rauwerdink, A.; Kazlauskas, R. J. How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: The Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. *ACS Catal.* **2015**, *5* (10), 6153–6176. https://doi.org/10.1021/acscatal.5b01539.

(31) Foroozandeh Shahraki, M.; Ariaeenejad, S.; Fallah Atanaki, F.; Zolfaghari, B.; Koshiba, T.; Kavousi, K.; Salekdeh, G. H. MCIC: Automated Identification of Cellulases From Metagenomic Data and Characterization Based on Temperature and PH Dependence. *Front Microbiol* **2020**, *11*, 567863. https://doi.org/10.3389/fmicb.2020.567863.

(32) Cai, X.; Seitl, I.; Mu, W.; Zhang, T.; Stressler, T.; Fischer, L.; Jiang, B. Combination of Sequence-Based and in Silico Screening to Identify Novel Trehalose Synthases. *Enzyme and Microbial Technology* **2018**, *115*, 62–72. https://doi.org/10.1016/j.enzmictec.2018.04.012.

(33) Barriuso, J.; Martínez, M. J. In Silico Metagenomes Mining to Discover Novel Esterases with Industrial Application by Sequential Search Strategies. *J Microbiol Biotechnol* **2015**, *25* (5), 732–737. https://doi.org/10.4014/jmb.1406.06049.

(34) Mahmoudi, M.; Arab, S. S.; Zahiri, J.; Parandian, Y. An Overview of the Protein Thermostability Prediction: Databases and Tools. *J Nanomed Res* **2016**, *Volume 3* (Issue 6). https://doi.org/10.15406/jnmr.2016.03.00072.

(35) Hon, J.; Marusiak, M.; Martinek, T.; Kunka, A.; Zendulka, J.; Bednar, D.; Damborsky, J. SoluProt: Prediction of Soluble Protein Expression in Escherichia Coli. *Bioinformatics* **2021**, *XX* (XX), XXX–XXX. https://doi.org/(in press).

(36) Khurana, S.; Rawi, R.; Kunji, K.; Chuang, G.-Y.; Bensmail, H.; Mall, R. DeepSol: A Deep Learning Framework for Sequence-Based Protein Solubility Prediction. *Bioinformatics* **2018**, *34* (15), 2605–2613. https://doi.org/10.1093/bioinformatics/bty166.

(37) Raimondi, D.; Orlando, G.; Fariselli, P.; Moreau, Y. Insight into the Protein Solubility Driving Forces with Neural Attention. *PLOS Computational Biology* **2020**, *16* (4), e1007722. https://doi.org/10.1371/journal.pcbi.1007722.

(38) Bhandari, B. K.; Gardner, P. P.; Lim, C. S. Solubility-Weighted Index: Fast and Accurate Prediction of Protein Solubility. *Bioinformatics* **2020**, *36* (18), 4691–4698. https://doi.org/10.1093/bioinformatics/btaa578.

(39) Mazurenko, S.; Prokop, Z.; Damborsky, J. Machine Learning in Enzyme Engineering. *ACS Catal.* **2020**, *10* (2), 1210–1223. https://doi.org/10.1021/acscatal.9b04321.

(40) Bunzel, H. A.; Garrabou, X.; Pott, M.; Hilvert, D. Speeding up Enzyme Discovery and Engineering with Ultrahigh-Throughput Methods. *Current Opinion in Structural Biology* **2018**, *48*, 149–156. https://doi.org/10.1016/j.sbi.2017.12.010.

(41) 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. *Nature Chemical Biology* **2009**, *5* (10), 727–733. https://doi.org/10.1038/nchembio.205.

(42) Bidmanova, S.; Kotlanova, M.; Rataj, T.; Damborsky, J.; Trtilek, M.; Prokop, Z. Fluorescence-Based Biosensor for Monitoring of Environmental Pollutants: From Concept to Field Application. *Biosensors and Bioelectronics* **2016**, *84*, 97–105. https://doi.org/10.1016/j.bios.2015.12.010.

(43) 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 Research* **1997**, *25* (17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389.

(44) 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. *Molecular Systems Biology* **2011**, *7* (1), 539. https://doi.org/10.1038/msb.2011.75.

(45) Webb, B.; Sali, A. Comparative Protein Structure Modeling Using MODELLER. *Current Protocols in Bioinformatics* **2016**, *54* (1), 5.6.1-5.6.37. https://doi.org/10.1002/cpbi.3.

(46) Tian, W.; Chen, C.; Lei, X.; Zhao, J.; Liang, J. CASTp 3.0: Computed Atlas of Surface Topography of Proteins. *Nucleic Acids Research* **2018**, *46* (W1), W363–W367. https://doi.org/10.1093/nar/gky473.

(47) Pavelka, A.; Sebestova, E.; Kozlikova, B.; Brezovsky, J.; Sochor, J.; Damborsky, J. CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules. *IEEE/ACM Trans. Comput. Biol. and Bioinf.* **2016**, *13* (3), 505–517. https://doi.org/10.1109/TCBB.2015.2459680.

(48) Wilkinson, D. L.; Harrison, R. G. Predicting the Solubility of Recombinant Proteins in Escherichia Coli. *Bio/Technology* **1991**, *9* (5), 443. https://doi.org/10.1038/nbt0591-443.

(49) Shimizu, Y.; Inoue, A.; Tomari, Y.; Suzuki, T.; Yokogawa, T.; Nishikawa, K.; Ueda, T. Cell-Free Translation Reconstituted with Purified Components. *Nature Biotechnology* **2001**, *19* (8), 751–755. https://doi.org/10.1038/90802.

(50) Micsonai, A.; Wien, F.; Bulyáki, É.; Kun, J.; Moussong, É.; Lee, Y.-H.; Goto, Y.; Réfrégiers, M.; Kardos, J. BeStSel: A Web Server for Accurate Protein Secondary Structure Prediction and Fold Recognition from the Circular Dichroism Spectra. *Nucleic Acids Research* **2018**, *46* (W1), W315–W322. https://doi.org/10.1093/nar/gky497.

(51) Johnson, K. A. New Standards for Collecting and Fitting Steady State Kinetic Data. *Beilstein J Org Chem* **2019**, *15*, 16–29. https://doi.org/10.3762/bjoc.15.2.

(52) Johnson, K. A.; Simpson, Z. B.; Blom, T. Global Kinetic Explorer: A New Computer Program for Dynamic Simulation and Fitting of Kinetic Data. *Analytical Biochemistry* **2009**, *387* (1), 20–29. https://doi.org/10.1016/j.ab.2008.12.024.

(53) Vasina, M.; Vanacek, P.; Damborsky, J.; Prokop, Z. Chapter Three - Exploration of Enzyme Diversity: High-Throughput Techniques for Protein Production and Microscale Biochemical Characterization. In *Methods in Enzymology*; Tawfik, D. S., Ed.; Enzyme Engineering and Evolution: General Methods; Academic Press, 2020; Vol. 643, pp 51–85. https://doi.org/10.1016/bs.mie.2020.05.004.

Table 1. Summary of biochemical properties of HLDs.

Enzyme	Yield (mg. L^{-1})	Specific activity* ($\mu mol.s^{-1}.mg^{-1}$)	T_{onset} (°C)	T_m (°C)	T_{max} (°C)	E value	
						2-bromopentane	ethyl 2-bromopropionate
DstA	70	0.0030±0.0001	30.9±0.2	43.4±0.1	30	1.27±0.01	2.59±0.04
DfxA	10	0.0030±0.0002	30.5±0.6	40.6±0.5	35	n.a.	n.a.
DlaA	40	0.0040±0.0001	35.6±1.2	48.1±0.6	30	n.a.	n.a.
DaxA	120	0.0010±0.0002	42.9±0.2	48.7±0.1	35	16.4±0.3	n.a.
DsmA	120	0.086±0.001	27.6±0.1	35.7±0.1	25	1.60±0.01	81±1
DmmarA	20	0.0060±0.0001	32.3±0.1	42.1±0.2	30	6.33±0.04	1.22±0.01
DathA	60	0.0060±0.0001	38.1±0.6	46.4±0.1	35	27.3±0.4	45 ± 1
DmaA	30	0.211±0.005	32.5±0.1	40.2±0.3	35	2.13±0.01	49.8±0.4
DspoA	80	0.86±0.02	50.8±0.2	58.7±0.6	50	9.755±0.083	128±1
DexA	120	0.57±0.01	43.4±1.1	47.5±0.4	45	5.46±0.04	152±2
DppsA	100	0.029±0.001	24.7±0.2	38.1±0.2	35	3.32±0.03	84±1
DeaA	70	0.41±0.01	45.3±0.1	52.2±0.2	45	>200	113 ± 2
DmgaA	100	0.006±0.001	38.2±1.6	44.7±0.9	40	n.a.	n.a.
DprxA	150	0.63±0.01	44.3±1.7	51.8±0.3	45	3.23±0.02	>200
DrgA	20	0.0020±0.0002	36.8±0.4	44.2±0.4	35	n.a.	n.a.
DmbaA	10	0.132±0.002	36.8±0.3	46.6±0.2	45	5.54±0.04	22.2±0.2
DthA	90	0.031±0.001	40.4±0.3	49.9±0.9	35	155.9±0.7	>200
DphxA	30	0.596±0.007	47.0±0.6	55.4±0.2	35	1.82±0.01	26.0±0.2
DthB	20	0.122±0.004	44.8±0.6	53.4±0.4	45	2.98±0.02	15.9±0.1
DnbA	90	0.006±0.002	37.3±0.1	47.8±0.4	40	14.1±0.3	n.a.
DhxA	120	0.611±0.001	44.1±0.4	53.1±0.3	35	1.574±0.011	>200
DspxA	30	0.082±0.001	44.2±0.3	53.3±0.2	35	42.1±0.5	156±3
DchA	20	0.143±0.005	47.0±0.1	55.2±0.8	40	2.52±0.02	27.7±0.3
Dcta	10	0.0050±0.0002	31.6±0.1	39.8±0.6	35	n.a.	187±2

*Specific activity towards 1,3-dibromopropane was determined in 1 mM HEPES buffer at pH 8.2 and temperature close to the optimal temperature (Table S8); T_{onset} – unfolding onset temperature; T_m – melting temperature by nanoDSF; T_{max} – maximum HLD activity; n.a. – no activity

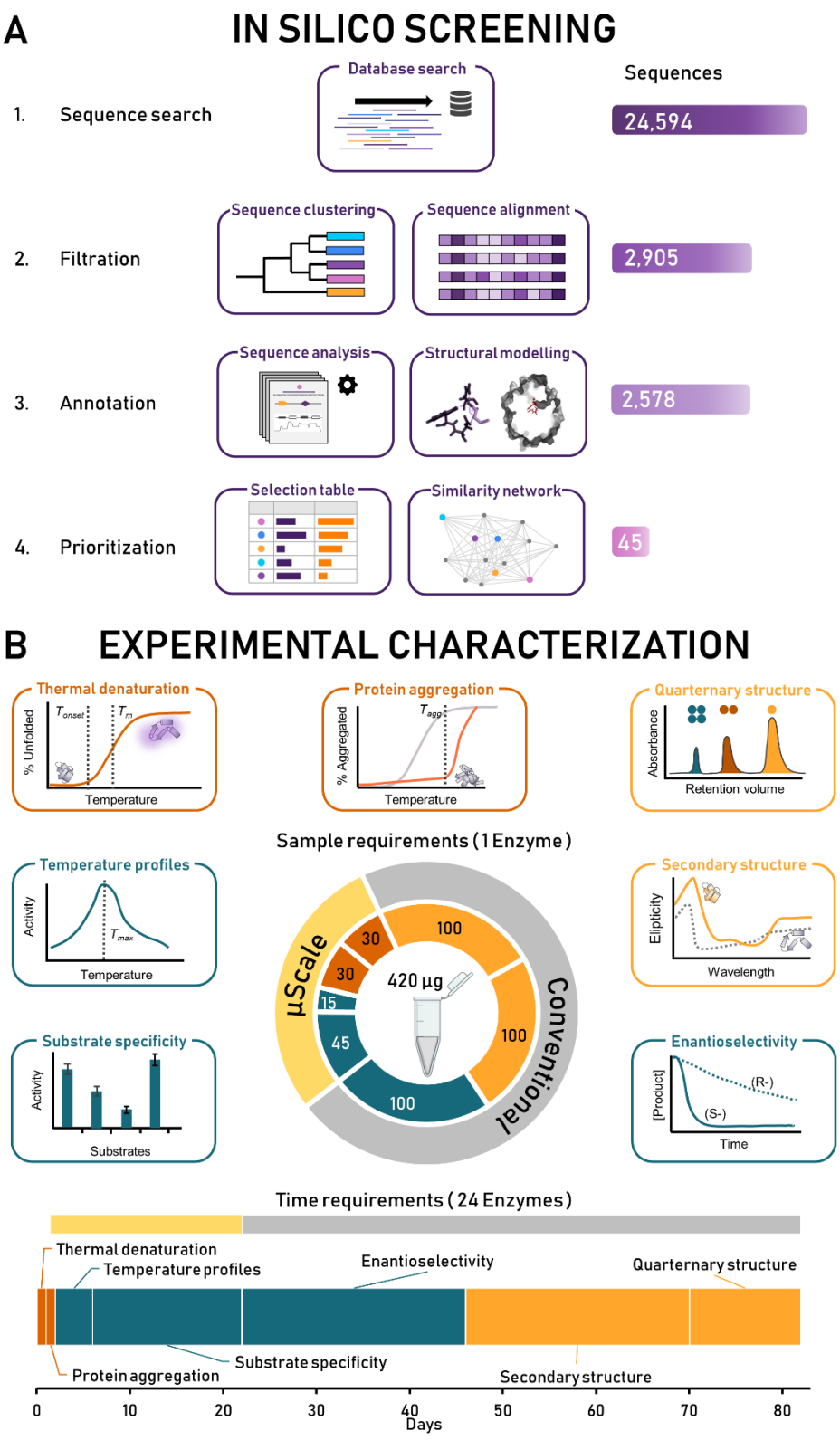


Figure 1. Integrated theoretical-experimental strategy for high-throughput exploration of unmapped sequence space. (A) *In silico* screening and (B) structural and functional characterization applying microscale and microfluidics techniques.

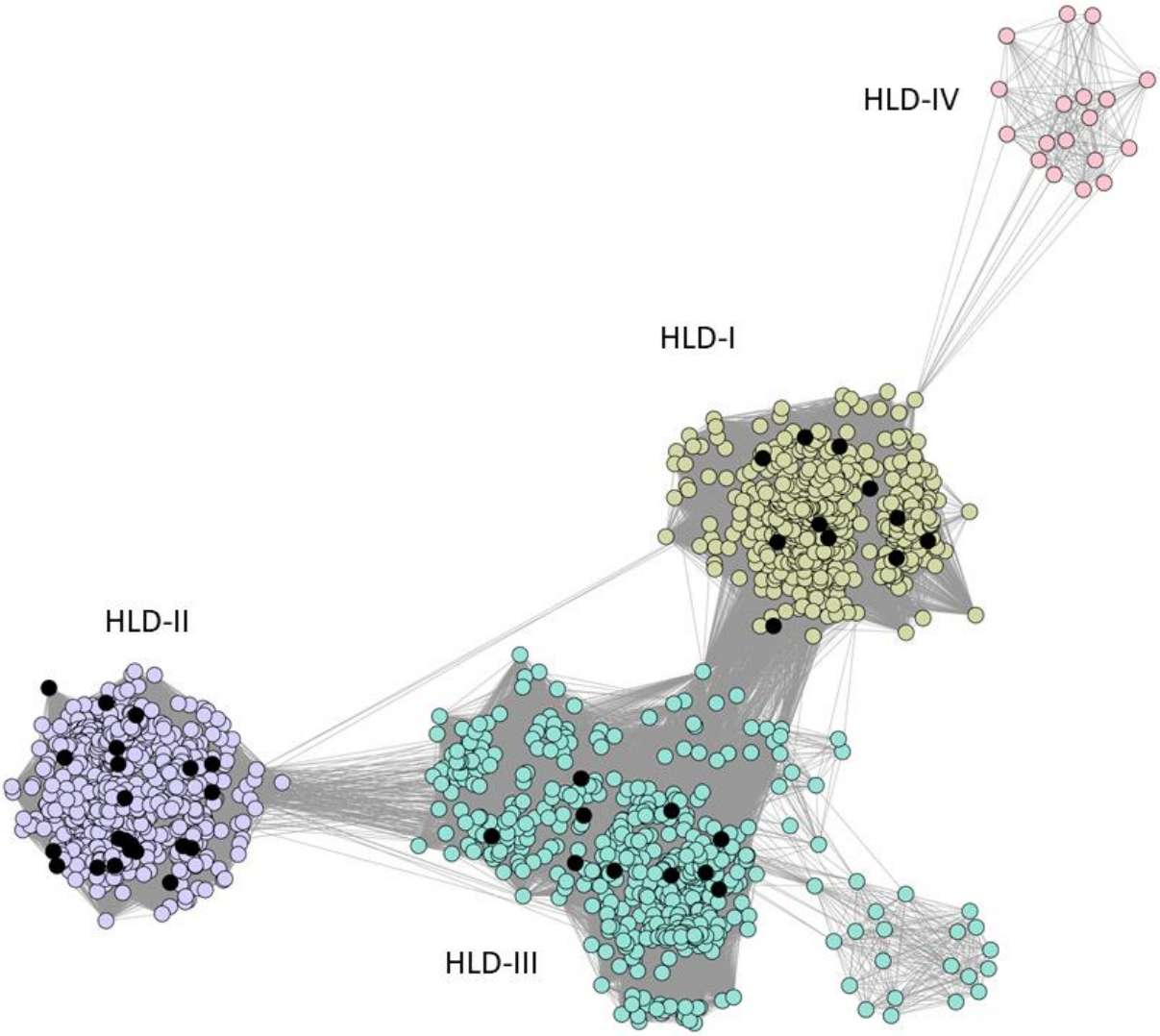
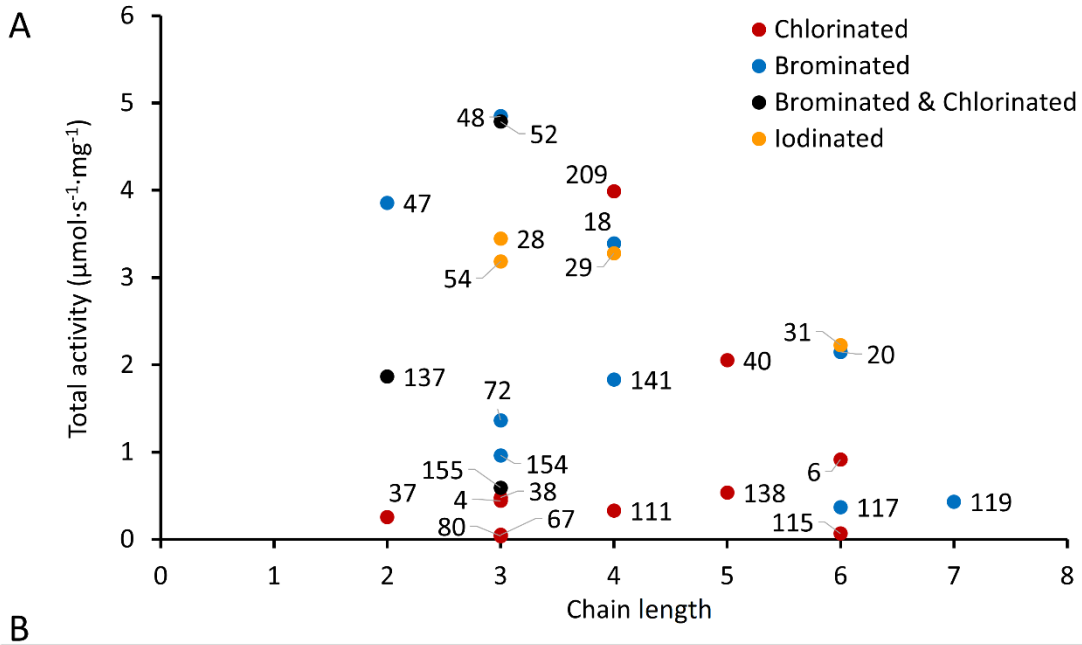


Figure 2. Sequence similarity network for HLD family. The putative HLDs are clustered into four subfamilies: HLD-I (yellow), HLD-II (violet), HLD-III/IIIb (green) and HLD-IV (pink). The sequences were first clustered at 50% identity to reduce the number of nodes and edges. Sequences with greater identity are consolidated into a single node. Edge lengths indicate sequence similarity between representative sequences of the connected nodes. Black nodes contain at least one selected target.



B

No.	Substrate name	Convertibility (%)	No.	Substrate name	Convertibility (%)	No.	Substrate name	Convertibility (%)
47	1,2-dibromoethane	100	137	1-bromo-2-chloroethane	91	119	(1-bromomethyl)cyclohexane	78
20	1-bromohexane	100	52	1-bromo-3-chloropropane	91	38	1,3-dichloropropane	78
48	1,3-dibromopropane	97	4	1-chlorobutane	91	37	1,2-dichloroethane	75
18	1-bromobutane	97	31	1-iodohexane	91	138	chlorocyclopentane	75
29	1-iodobutane	97	40	1,5-dichloropentane	88	155	1,2-dibromo-3-chloropropane	66
209	3-chloro-2-methylpropene	94	141	4-bromobutyronitrile	84	111	bis(2-chloroethyl)ether	59
6	1-chlorohexane	94	117	bromocyclohexane	84	115	chlorocyclohexane	59
28	1-iodopropane	94	154	1,2,3-tribromopropane	81	67	1,2-dichloropropane	41
66	1,2-dibromoethane	91	54	1,3-diidopropane	81	80	1,2,3-trichloropropane	41

Figure 3. Substrate preference of HLDs. (**A**) Dependence of total activity (sum of specific activities of 24 newly characterized and 8 benchmark HLDs) on the chain length of the halogenated substrate. The individual colors of the data points stand for the type of substrate: chlorinated (red), brominated (blue), brominated & chlorinated (black) and iodinated (orange). (**B**) Convertibility of the substrate quantifies the fraction of HLDs which convert a particular substrate.

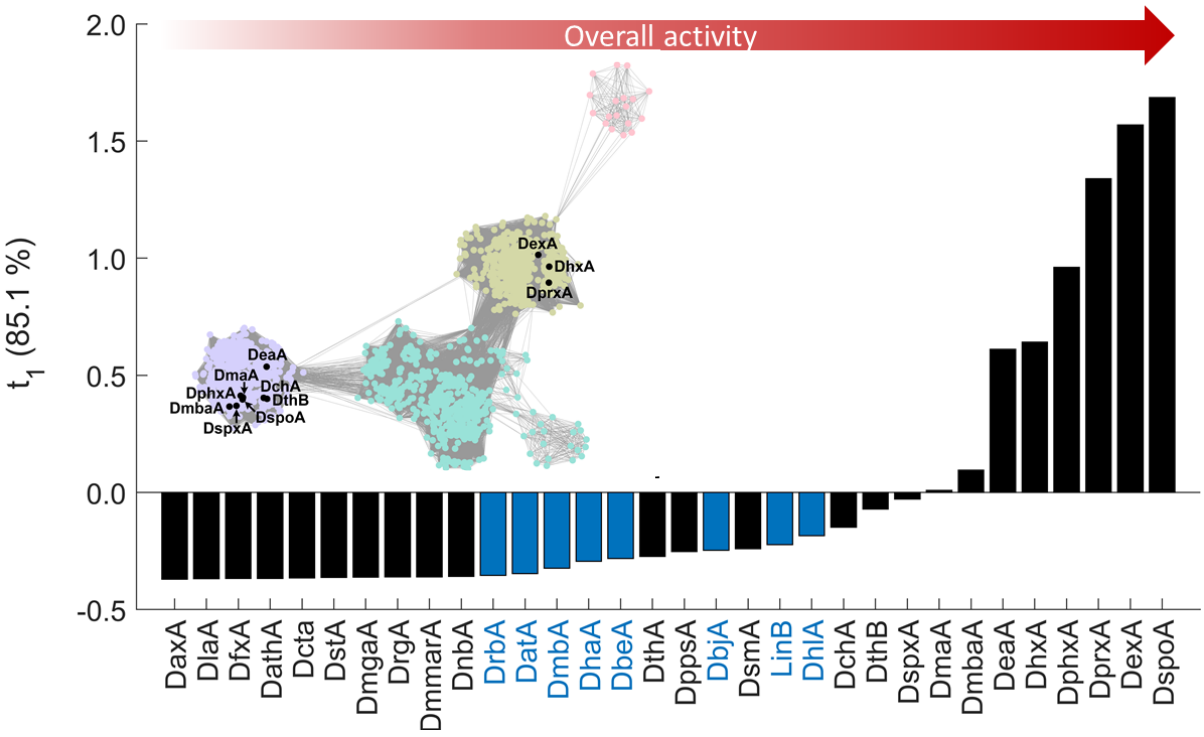


Figure 4. Multivariate analysis of catalytic activity. The score-contribution plot t_1 compares the enzymes in terms of their overall activity with 27 substrates and explains 85.1 % of the variance in the untransformed data set. Known HLDs are depicted by blue color and newly identified HLDs by black color. Inset: the sequence similarity network of putative HLDs clustered into four subfamilies: HLD-I in yellow, HLD-II in violet, HLD-III/IIIb in green and HLD-IV in pink. The highlighted nodes in black represent 11 novel enzymes with elevated activity.

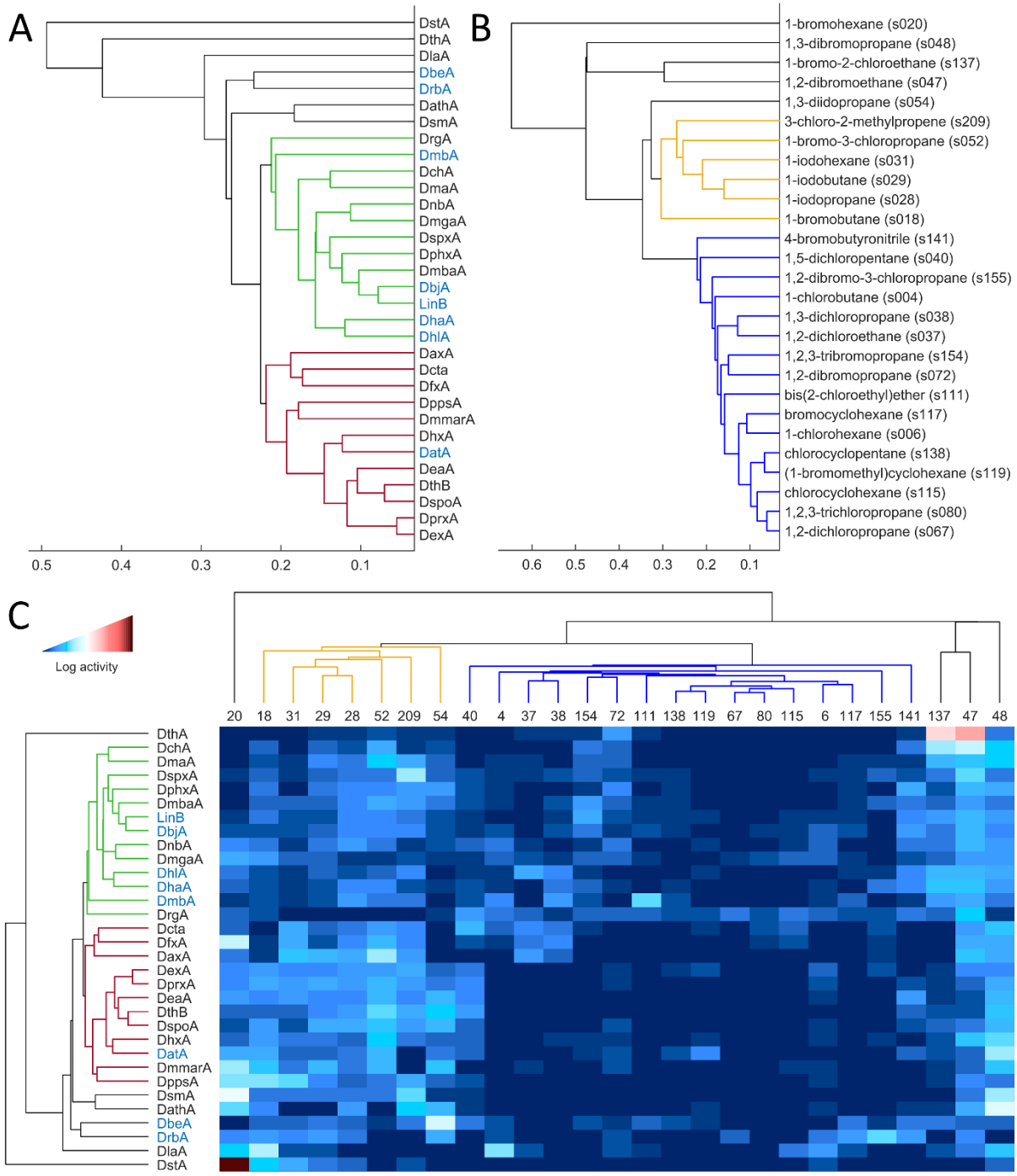


Figure 5. Multivariate analysis of substrate specificity. A heat map with hierarchical clustering dendrograms for the similarity of enzyme activity (A) and halogenated substrate conversion (B). Substrate functional groups and halide detected in the reaction are color-coded with bars at the tips of the dendrograms. Only reactions with >8% conversion were shown to remove false positives. (**C**) Detailed hierarchical clustering of enzymes and halogenated substrates shown in a double-dendrogram. Major groups are highlighted with the same color. The colormap displays the log-transformed data.