Advanced Database Mining of Efficient Biocatalysts by Sequence and Structure Bioinformatics and Microfluidics

Michal Vasina[1,2#], Pavel Vanacek[1,2#], Jiri Hon[2,3], David Kovar[1,2], Hana Faldynova[1], Antonin Kunka[1,2], Tomas Buryska[1], Christoffel P. S. Badenhorst[4], Stanislav Mazurenko[1,2], David Bednar[1,2], Stavros Stavrakis[5], Uwe T. Bornscheuer[4], Andrew deMello[5], Jiri Damborsky[1,2], Zbynek Prokop[1,2*]

[1] Loschmidt Laboratories, Department of Experimental Biology and RECETOX, Faculty of Science, Masaryk University, Kotlarska 2, Brno, Czech Republic

[2] International Clinical Research Centre, St. Ann's Hospital, 656 91 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

[5] Institute for Chemical and Bioengineering, ETH Zürich, 8093 Zürich, Switzerland

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

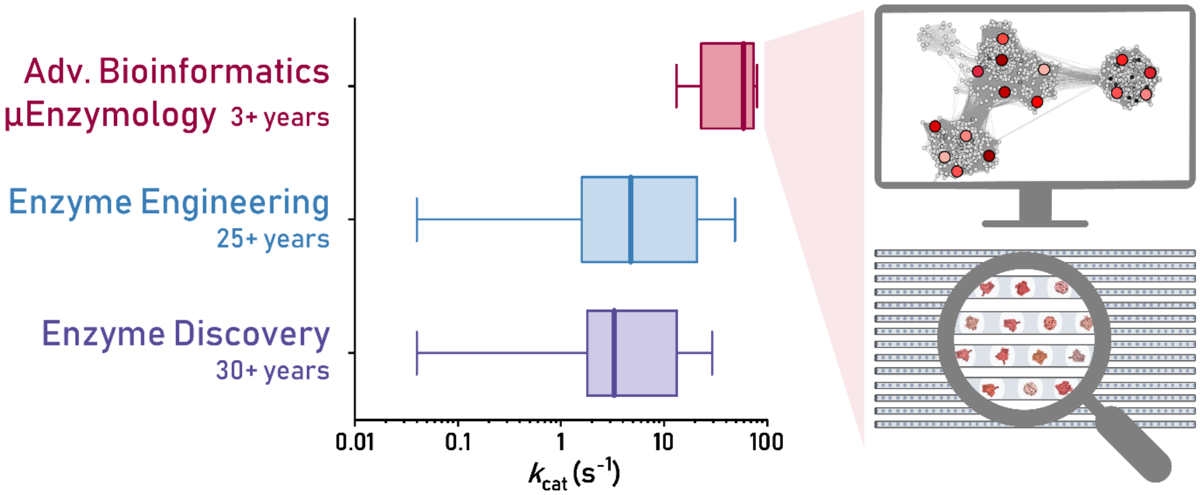
* Authors for correspondence: zbynek@chemi.muni.cz, ORCID 0000-0001-9358-4081

SUMMARY

Next-generation sequencing doubles genomic databases every 2.5 years. The accumulation of sequence data provides a unique opportunity to identify interesting biocatalysts directly in the databases without tedious and time-consuming engineering. Herein, we present a pipeline integrating sequence and structural bioinformatics with microfluidic enzymology for bioprospecting of efficient and robust haloalkane dehalogenases. The bioinformatic part identified 2,905 putative dehalogenases and prioritized a "small-but-smart" set of 45 genes, yielding 40 active enzymes, 24 of which were biochemically characterized by microfluidic enzymology techniques. Combining microfluidics with modern global data analysis provided precious mechanistic insights related to the high catalytic efficiency of selected enzymes. Overall, we have doubled the dehalogenation "toolbox" characterized over three decades, yielding biocatalysts that surpass the efficiency of currently available wild-type and engineered enzymes. This pipeline is generally applicable to other enzyme families and can accelerate the identification of efficient biocatalysts for industrial use.

Keywords: enzyme mining; enzyme diversity; biocatalysts; microfluidics; bioinformatics; global data analysis, haloalkane dehalogenases

GRAPHICAL ABSTRACT



INTRODUCTION

Nature relies heavily on enzymes, which enable virtually every biosynthetic and biodegradation process in all life forms. Humanity recognized the power of enzymes and harnessed them in a wide range of industrial sectors, such as food-, textile-, agro-, chemical- or pharma-industry.[1] Despite the successful application of a range of enzymes, their properties often do not match the application requirements for high catalytic efficacy. For decades, scientists have asked themselves how to find better biocatalysts. Shall we explore the natural sequence space (discover new enzymes), or rather the artificial diversity (improve existing enzymes)?[2]

Thanks to the genomic revolution, the avalanche of protein sequences filling the genomic databases at an unprecedented pace represents an outstanding achievement, but it also brings new challenges to its effective exploration and practical utilization. So far, only a negligible fraction of genes deposited in databases have been experimentally characterized. Moreover, incorrect automatic annotations are quite frequent and tend to percolate, leading to error accumulation.[3,4] Thus, without advanced bioinformatics expertise, relying only on database annotations, many efforts dedicated to discovering new biocatalysts do not succeed, even after large investments and the application of high-throughput screening campaigns.[5] This can lead to underestimating the potential of natural diversity hidden in sequence databases.

In parallel, the success of many protein engineering studies arises from applying modern directed evolution strategies, combined with *in silico* identification of hot spots and followed by an experimental screening of smaller “smart” libraries.[6] Similarly, applying advanced bioinformatic methods for “smart” prioritization of a smaller list of candidates towards a “focused” experimental characterization represents a promising strategy for identifying suitable biocatalysts from the rich sequence information accessible in databases.[7] A critical requirement is the availability of bioinformatic tools, especially for non-expert users, enabling wide and effective exploration of the natural diversity hidden in sequence databases for scientific and industrial communities.[8]

Herein, we present a pipeline integrating advanced sequence and structural bioinformatics with microfluidic enzymology for bioprospecting of efficient and robust biocatalysts. We doubled the number of experimentally characterized members of a model enzyme family in a single run of this workflow. At the same time, the obtained enzymes catalytically surpass the previously known variants, whether discovered or engineered. The experimental pipeline relies heavily on two in-house microfluidic platforms, MicroPEX and KinMAP, where the latter is introduced in this study. By subjecting the multidimensional data from KinMAP to modern global data analysis, unique mechanistic insights were obtained for the enzymes with the highest overall activity.

A model enzyme family, haloalkane dehalogenases (HLDs), were used as the case study. Three decades of intensive research on HLDs has made them benchmark enzymes for studying the structure-function relationships of the >100,000 members of the α/β-hydrolase fold superfamily[9] and the development of novel concepts in the field of protein engineering.[10] Thanks to the long-term, extensive research on HLDs, we were also able to conceptually compare the variants obtained by the advanced database mining with enzymes previously isolated by classical enzymological approaches,[11] and variants systematically constructed for more than 20 years by various protein engineering strategies. These strategies include optimizing[12] and introducing *de novo* access tunnels,[13] active site remodeling,[14,15] engineering dynamical protein loops,[16] targeting mutations which enhance thermostability,[17–19] or resurrecting HLDs by ancestral sequence reconstruction.[20] We believe the current study presents an interesting conceptual view of current approaches used in biocatalysts development, which should not underestimate the potential of structural and functional diversity found in nature.

RESULTS

The bioprospecting of efficient and robust biocatalysts from sequence databases was performed as follows. First, we applied an automated *in silico* workflow (**Fig. 1**) to identify putative family members and select promising candidates. Next, we experimentally characterized the prioritized hits from the *in silico* screening by employing small-scale expression, followed by in-depth microfluidic characterization (**Fig. 2**).

I. Automated *in silico* workflow

The *in silico* bioprospecting workflow is composed of three steps: (i) database search and sequence processing using a previously developed sequence bioinformatics pipeline (available as the web tool EnzymeMiner[21]), (ii) structure prediction and its systematic analysis using various computational tools within a newly developed structural bioinformatics pipeline, and (iii) prioritization of hits from both sequence and structural bioinformatics pipelines and selection of a “small-but-smart” set of proteins for experimental characterization (**Fig. 1**).

Database search. We reran the *in silico* screening with the same four input sequences as previously[22] using the current version of the NCBI nr database and a recently developed tool for automated database mining.[21] The previously used workflow has been significantly expanded by: (i) application of EFI-EST[23] and Cytoscape[24] 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,[25] and (iii) the quantitative assessment of the quality of all homology models by MolProbity.[26] 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, alignment, and filtering. The remaining 2,905 protein sequences were clustered into four subfamilies: HLD-I (915), HLD-II (1058), HLD-III (910), and HLD-IV (22), based on the sequence identity and the composition of their catalytic pentads.[27] Despite having identical catalytic pentads, HLD-III and HLD-IV were clustered separately based on differences in their sequences. Incomplete and degenerated sequences were filtered out by constructing multiple sequence alignments of individual subfamilies. Sequence-similarity networks were constructed to visualize relationships among putative HLD sequences (**Fig. 3**). The most apparent defining features were clustered in the distinct HLD subfamilies, implying that the sequence-similarity networks might provide a framework for identifying HLDs of similar structural and functional properties and surveying regions of sequence space with high diversity. To diversify HLD sequence space, redundant sequences with ≥ 90% sequence identity to the set of 22 characterized dehalogenase sequences (**Table S1**) were filtered out.

Structure prediction, active site analysis, transport path analysis, and substrate binding. The remaining 2,578 putative HLD sequences were subjected to an annotation step consisting of information retrieval from biological databases and structure predictions. The annotation step revealed that the identified HLDs span a broad range of sequence and host diversity, including bacterial, archaeal, and eukaryotic proteins. The overall accuracy of annotation, judged by assignment to the HLD family, was 63% but varied significantly among each of the HLD subfamilies. Most 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 56% for HLD-IV (**Table S2**). Most members from the putative HLD-IV subfamily were annotated as HLDs, despite their low sequence identity to the experimentally characterized HLDs or other subfamily members. The annotation revealed four putative dehalogenases from psychrophilic organisms, 35 novel proteins from moderate halophilic organisms, and four proteins with known tertiary structures. Reliable homology models could be constructed for most subfamily HLD-I and HLD-II members, but only a limited number of HLD-III members and none of the HLD-IV members. The predicted volumes of catalytic pockets ranged from 50 Å^3 to 3,950 Å^3 (**Fig. S1**). Putative transport pathways were analyzed by predicting access tunnels connecting a buried active site with a protein surface. The molecular docking simulations were employed to probe potential binding modes of representative halogenated compounds (**Table S3)**.

Prioritization and selection of targets. Rational selection of hits for experimental characterization was carried out to maximize the functional diversity of the studied protein family. 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. A few sequences were selected from each datasheet to make the selection as diverse as possible (**Supplemental dataset, Table S4**). The sequences with a higher predicted solubility and higher-quality homology models were prioritized. Simultaneously, we tried to balance the number of sequences from each haloalkane dehalogenase subfamily (HLD-I, HLD-II, and HLD-III). The only exception was the HLD-IV subfamily, which contains multi-domain protein sequences derived from eukaryotic organisms. We avoided sequences with additional Pfam domains, as they were previously poorly expressible in bacterial host systems.[22] A "small-but-smart" set of 45 diverse sequences was selected as experimental characterization targets (**Table S5**, **Table S6**).

II. Small-scale protein expression

This representative set of 45 HLD genes was subjected to a small-scale expression in *Escherichia coli* in 96-deep well square plates and screening of HLD activity in whole cells (**Fig. 3**) using the halide oxidation (HOX) assay.[28] Overall, 40 out of 45 genes (89%) could be overexpressed. Although 30 out of 45 genes (67%) yielded soluble proteins (**Fig. S3a**), only 24 of them (53%) showed sufficient expression and solubility for downstream biochemical characterization (**Fig. S4**). Comparison of the *in silico* prediction of soluble expression with experimental data showed a poor correlation (Pearson's correlation coefficient 0.263) and only 66.7% prediction accuracy. Specifically, the *in silico* solubility predictions resulted in 22 true positives, 8 true negatives, 4 false negatives, and 11 false positives (**Table S7**). A further thorough analysis of solubility profiles revealed that most of the proteins belonging to HLD-I (73%) and HLD-II (71%) sub-families were expressed in a soluble form, while a less than half of HLD-III (40%) proteins were soluble. We then probed the expressibility of all 45 HLD genes using a reconstituted cell-free transcription and translation system (PURExpress, NEB). Overall, 41 of 45 genes (91%) were overexpressed, and 29 proteins (64%) were obtained in soluble form (**Fig. S3b**). Application of the cell-free PURExpress system did not result in the desired improvement of solubility for the "difficult-to-produce" HLDs suggesting that *in vivo* toxicity has little effect on the production of these proteins. In addition, the number of active variants detectable in whole cells (40 out of 45) is higher than that of finally purified proteins (24 out of 45), indicating problems with protein stability/solubility related to the purification process (**Fig. 3**, **Table S8**). The activity analysis in whole cells showed that the success rate of target activity prediction is at least 90%.

III. Microfluidic enzymology

The experimental pipeline (**Fig. 2**) comprised commercial microfluidic instruments and two custom-made microfluidic platforms. Combining these modern technologies led to an efficient yet in-depth biochemical characterization of the selected 24 HLDs (**Table 1**). The results of individual characterization steps provided key parameters for the experimental design of the subsequent step within the workflow (**Fig. 2**). First, thermostability measurements helped estimate the temperature ranges for each temperature profile. Second, temperature profiles provided the optimum temperature for the subsequent substrate specificity characterization. Finally, based on the overall catalytic activity from substrate specificity measurements, the best variants were characterized in terms of steady-state kinetics and reaction thermodynamics, providing further mechanistic insights.

Thermostability. After protein purification, the thermostability of the novel HLDs was analyzed in a high-throughput manner by monitoring changes in extrinsic (SYPRO orange dye) and intrinsic (tryptophan) fluorescence during thermal denaturation experiments, using the thermal shift assay (TSA) and microscale differential scanning fluorimetry (DSF), respectively. The thermostability measurements provided the temperature at which protein denaturation starts (onset temperature, T_{onset}) and the midpoint of the denaturation curve (apparent melting temperature, $T_{\mathrm{m}}^{\mathrm{app}}$), where the latter was used for comparison of individual thermostability methods. The results of the microscale methods showed an excellent agreement (R^2 0.79 and 0.93 for TSA and capillary DSF, respectively) with conventional circular dichroism (CD) spectroscopy (**Table S9**, **Fig. S5**). The apparent melting temperature ($T_{\mathrm{m}}^{\mathrm{app}}$) values (**Table S9**) primarily reflect the mesophilic origins of the novel HLDs (40-60 °C). Exceptions are the DsmA and DppsA, which exhibited $T_{\mathrm{m}}^{\mathrm{app}}$ values at 35.7 and 38.1 °C, respectively, correlating with their psychrophilic origin. It is worth noting that the most stable protein identified was DspoA, with a $T_{\mathrm{m}}^{\mathrm{app}}$ value of 60 °C.

Temperature Profiling. Temperature profiling was performed using the first custom-made microfluidic profile explorer (MicroPEX), utilizing pH-based fluorescence assay in droplets, as described previously.[29] The new dehalogenases obtained in this study showed activity over a wide temperature range (**Fig. 4B**, **Fig. S6**, **Table S10).** DmaA was especially unique, as it retained more than 65% dehalogenase activity at 5 °C. This dehalogenase performed equally well at this low temperature compared to benchmark dehalogenases at their temperature optima (30-45 °C).[30] A positive correlation was observed between the temperature of the highest observed activity (T_{max}) and T_{onset} obtained from thermal denaturation experiments (**Fig. S7**).

Substrate Specificity Profiling. Substrate specificity profiling towards 27 representative substrates was conducted using the same analytical assay as temperature profiling on MicroPEX (**Table S11**). This structurally diverse set of substrates reflects the application of HLDs, including environmentally important compounds (**Table S12)**. The raw data of specific activities (**Table S13)** showed that HLDs exhibited better activities with the following order of preference: brominated > iodinated ≫ chlorinated. Analysis of the substrate preferences showed that the optimal substrates of the newly discovered HLDs have linear alkyl chains of 2-4 carbon atoms (**Fig. S8a)** and that the majority of the HLDs can convert this type of substrate with the highest efficiency (**Fig. S8b).** Based on these observations, we suggest a set of "universal" substrates: 1-bromobutane (#18), 1-iodopropane (#28), 1-iodobutane (#29), 1,2-dibromoethane (#47) and 1,3-dibromopropane (#48). The substrate specificity profiling also identified a set of "recalcitrant" substrates: 1,2-dichloroethane (#37), 1,2-dichloropropane (#67), 1,2,3-trichloropropane (#80), the analog of warfare-agent yperite bis(2-chloroethyl)ether (#111), and chlorocyclohexane (#115), which is in good agreement with previous studies.[30,31] It is worth noting that two-thirds of the newly discovered enzymes possess broad substrate specificity and convert > 80% of the substrates tested (**Table S14**). Interestingly, two new enzymes, DstA and DthA, showed a previously undescribed narrow specificity. Specifically, DstA effectively converted one specific substrate, 1-bromohexane (#20), with five-fold higher activity than any other substrate. Similarly, DthA exhibited considerable debromination activity for only two substrates, 1,2-dibromoethane (#47) and 1-bromo-2-chloroethane (#137).

Principal Component Analysis (PCA). First, we conducted PCA analysis using the untransformed specificity data of 8 benchmarks[29] and 24 newly identified HLDs. This analysis aimed to compare the enzymes according to their score with the first principal component (t_1), thus quantifying their global activity against the set of substrate activities (**Fig. 4D**). Surprisingly, 11 of the 24 newly characterized HLDs showed significantly higher global activity than the known benchmark HLDs. This result was validated using conventional activity measurements with an overall well-converted substrate, 1,3-dibromopropane (**Fig. S9**). Six out of these 11 highly active enzymes exhibited outstanding overall activity, and therefore, they were chosen to characterize their steady-state kinetics and reaction thermodynamics using Kinetic Microfluidic Autonomous Platform (KinMAP) (**Fig. 4D**). The second PCA was performed with log-transformed and weighted activity data allowing a direct comparison of the specific profiles of individual enzymes unbiased by the different levels of their global activity (**Fig. S10**). The benchmark HLDs (DbjA, LinB, DmbA, DhlA, and DhaA) were clustered in agreement with the previously reported substrate specificity groups of HLDs.[31] In this analysis, two of the newly discovered variants, DstA and DthA, were separated from other enzymes due to their unusually narrow substrate specificity.

Hierarchical Clustering. The log-transformed specificity data were subjected to hierarchical clustering to identify similarity in preferred substrates or selectivity of enzymes; both were plotted as a double dendrogram heatmap (**Fig. 4C**). Our analysis clustered the substrates into three main groups. The first group (yellow in **Fig. 4C**) comprises frequently converted substrates, mostly iodinated compounds with a chain length of 3-4 carbon atoms. The second group (green in **Fig. 4C**) includes moderately and poorly convertible (mainly chlorinated) substrates. The third group (brown in **Fig. 4C**) contains only three structurally similar substrates preferred over other tested substrates by most enzymes. Clustering of the specificity profiles divided analyzed HLD variants into two major groups. The first group (purple in **Fig. 4C**) 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 (orange in **Fig. 4C**) is almost entirely composed of newly identified enzymes (except for DatA), which preferentially convert the more frequently converted substrates (the first and the third group of substrates - yellow and brown in **Fig. 4C**, respectively) over the second group of substrates (green in **Fig. 4C**). 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. The third group (teal in **Fig. 4C**) contains four enzymes possessing the narrow substrate specificity profiles, e.g., DrbA towards 1,2-dibromo-3-chloropropane (#155) or DsmA towards 3-chloro-2-methylpropene (#209).

Steady-State Kinetics and Reaction Thermodynamics. Inspired by technology for kinetic analysis of nanoparticle synthesis,[32] we developed a microfluidic device for kinetic and thermodynamic analysis enzyme reaction called the Kinetic Microfluidic Autonomous Platform (KinMAP) (**Fig. 5A,** see details in **Supplemental experimental procedures 1.5, Table S16, Fig. S12-S16**). KinMAP operates autonomously thanks to the software MAPit, integrating control over all hardware units and providing fully automated calibration, data acquisition, and signal processing for a wide range of conditions with minimal user involvement (**Supplemental experimental procedures 1.5, Fig. S15-S16**).

KinMAP was used to determine steady-state kinetics and reaction thermodynamics parameters for selected highly active enzymes (DspoA, DexA, DeaA, DprxA, DphxA, and DhxA) (**Fig. 4D**). Multidimensional data, including concentration and temperature dependence of the reaction, were collected by monitoring the conversion progress at six different substrate concentrations (0-1 mM 1,3-dibromopropane), and each of them at six different temperatures from 25 to 50 °C in 5-degree increments (**Fig. 5B**). The global numerical fitting of such a complex dataset provided estimates for the kinetic constants, namely specificity constant (k_{cat}/K_m), turnover number (k_{cat}), the equilibrium constant for enzyme-product complex dissociation (K_P), and the corresponding thermodynamic parameters (**Fig. 5C, Fig. S11, Table S15**). Following new standards for collecting and fitting steady-state kinetic data,[33] we estimated k_{cat}/K_m directly instead of K_m. Unlike K_m, which has no mechanistic meaning, k_{cat}/K_m can be interpreted as the apparent second-order rate constant for substrate binding and quantifies enzyme specificity, efficiency, and proficiency. Moreover, there are smaller errors in the fitting process to derive k_{cat}/K_m directly rather than calculating the ratio of k_{cat} and K_m derived independently (see details in **Supplemental experimental procedures 1.6**).

All six selected enzymes showed one of the highest turnover numbers (13 to 80 s^{-1}) ever observed within the HLD family compared to previously isolated wild-type and engineered variants (**Fig. 6)**. The highest previously reported turnover number for a dehalogenase, k_{cat} of 57 s^{-1}, was determined for LinB86 in converting 1,2-dibromoethane (#47) **(Fig. S8)**. This four-point mutant with an introduced *de novo* access tunnel was obtained by several cycles of computer modeling and rational engineering.[13] Three new biocatalysts identified in this study (DprxA, DhxA, and DexA) exhibited higher k_{cat} (80, 74, and 64 s^{-1}, respectively) than LinB86 (**Fig. 5C**, **upper left**), which makes them the fastest HLDs ever reported.

Despite the high k_{cat} of LinB86, its specificity constant was relatively low (k_{cat}/K_m = 24 mM^{-1}.s^{-1}). On the contrary, LinB wild type in the reaction with 1,3-dibromopropane (**Fig. S8)** exhibited a high specificity constant (k_{cat}/K_m = 165 mM^{-1}.s^{-1}) yet a lower k_{cat} = 6.6 s^{-1}.[34] A rare example of an HLD exhibiting high values of both k_{cat} and k_{cat}/K_m was the engineered variant DmxA Q/N.[35] This single-point mutant, engineered from DmxA originating from the psychrophilic bacterium *Marinobacter* sp. ELB17, shows k_{cat} of 31 s^{-1} and k_{cat}/K_m = 244 mM^{-1}.s^{-1} with 1,3-dibromopropane. Such a rather rare combination of high k_{cat} and k_{cat}/K_m values was observed for three newly identified enzymes (**Fig. 5C)**, namely DprxA, DhxA, and DphxA. Remarkably, DphxA with k_{cat} of 54 s^{-1} and k_{cat}/K_m = 290 mM^{-1}.s^{-1} shows the best combination of turnover number and catalytic efficiency ever reported (**Fig. 6C)**.

The temperature dependences analyzed for the catalytic rate (k_{cat}) indicated that the free energy of activation is predominantly determined by a positive enthalpy, or a combination of both entropy and enthalpy, in the case of DprxA and DhxA. Interestingly, DspoA, DexA, and DphxA showed a favorable entropic contribution in lowering the activation energy of the catalytic turnover (**Fig. 5C**). The temperature dependences of k_{cat}/K_m indicated that the efficiency of substrate binding is similarly influenced predominantly by enthalpy (DeaA, DprxA, and DhxA) or a combination of positive enthalpy and unfavorable loss of entropy (DspoA and DhxA). The other two interesting cases are DexA, with its specificity constant dominated by unfavorable entropy, and DeaA, with a favorable positive entropy compensating activation enthalpy and reducing the overall free energy of activation (**Fig. 5C**). The mechanistic information derived from the differences in the thermodynamic profiles provides an excellent starting point for rational design[36] and further analysis using machine learning.[37]

IV. Additional biochemical characteristics

Enantioselectivity. Enantioselectivity was assessed by determining the kinetic resolution of *rac*-2-bromopentane and *rac*-ethyl 2-bromopropionate, representing two distinct groups of chiral substrates (β-brominated alkanes and esters, respectively). Individual HLDs showed variable enantioselectivity in the reaction with the racemic substrate 2-bromopentane. More specifically, high enantioselectivity was identified for DeaA and DthA, exhibiting E-values of > 200 and 156, respectively (**Fig. S17**). Most of the novel HLDs preferred 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.[38,39] High enantioselectivity (E-value > 200) towards the second representative substrate, ethyl 2-bromopropionate, was observed in the case of DprxA, DthA, and DhxA (**Fig. S18**).

Secondary and Quaternary Structure. We also analyzed the secondary and quaternary structure using far-UV-CD spectroscopy and size-exclusion chromatography. All HLDs exhibited CD spectra with one positive peak at 195 nm and two negative minima at 208 and 222 nm, characteristic of proteins with an α/β-hydrolase fold (**Fig. S19**).[40] Newly identified HLDs were mostly monomeric, similar to the previously characterized HLD members (**Table S17**). Exceptions were DmmarA, which exists as a dimer, and DprxA, which exists as a mixture of monomer, dimer, and higher oligomeric states (**Fig. S20**). Interestingly, native PAGE revealed that DstA was sensitive to the oxidation/reduction potential of the environment and formed dimers only under oxidative conditions (**Fig. S21**).

DISCUSSION

The biotechnology field employing enzymes as catalysts represents a billion-dollar industry, putting constant pressure on speeding up the identification and characterization of novel biocatalysts.[41] The avalanche of newly available sequences from next-generation sequencing represents an enormous potential but, at the same time, a significant challenge for the practical aspects of efficient search and throughput for experimental functional characterization. The application of genome mining can provide a potential solution to managing a large quantity of complex sequence data effectively.[42] Currently, it is not feasible to characterize all sequences being deposited in sequence databases. Instead, *in silico* screening and prioritizing a narrower selection of targeted sequences based on advanced bioinformatic analyses, followed by microfluidic high-throughput characterization, appears to be an attractive approach.

We have used such a strategy to identify novel variants of the model enzyme family – haloalkane dehalogenases, which have been thoroughly investigated for more than thirty years. Our results show that only 63% of the identified putative HLDs were labeled correctly as dehalogenase enzymes in genomic databases. While miss-annotations 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 fold superfamily are well-known for their catalytic promiscuity and tendency to catalyze diverse reactions using the same catalytic machinery.[43,44] Substrates are currently not known for 35% of enzymes annotated as α/β-hydrolases, and thus their functions remain unclear.[45] The current mining approach identified more than 2,578 putative HLDs. The number of hits increased nearly five times compared to the previous *in silico* screening.[22] The current screening approach missed only 97 sequences out of the original set and identified 2,145 new sequences.

The sequence mining analysis presented in this study is available as a user-friendly web tool, EnzymeMiner, making at least part of our *in silico* pipeline widely accessible to the scientific and industrial communities.[21] Although other computational tools help automatically analyze, filter, and visualize large sets of identified hits,[23,24] EnzymeMiner remains, to the best of our knowledge, the only available web tool for automated selection of promising candidates from the genomic databases. In addition to the prediction of tertiary structures that can be achieved using AlphaFold2,[46] analysis of cavities and access tunnels and modeling of enzyme-substrate complexes will be implemented in the future. The structural bioinformatics part of this study, including homology modeling followed by molecular docking of halogenated substrates, has been proven to be a powerful approach to identifying enzymes with high catalytic activities: 11 of 24 characterized HLDs showed higher activity levels than those reported previously (**Fig. 4D**).[31] Particularly, molecular docking of halogenated substrates turned out to be a promising selection criterion (**Table S3, Table S5**). All five enzymes (DeaA, DhxA, DphxA, DprxA, and DspxA) showing overall high dehalogenase activity were selected based on the positive docking of the warfare-agent yperite[47] (**Fig. 4D, Table S3**).

The major limitation of *in silico* analysis is the prediction of protein solubility. Despite applying the recent solubility prediction tool SoluProt,[48] our comprehensive expression analysis of the whole set of 45 selected putative HLDs revealed a 67% success rate in terms of soluble proteins, which is a slight improvement in comparison with the previously achieved 60%.[22] Protein production in *E. coli* can be improved by optimizing genetic constructs or expression conditions. However, the related combinatorial variation or a switch to other expression hosts such as yeasts or *Bacillus* species 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 most significant bottleneck toward the functional characterization of novel proteins. Improving the *in silico* solubility prediction is paramount for the increased success rate of protein characterization pipelines.[10,37] Nevertheless, 90% of the selected candidates were active dehalogenases (**Fig 3**), some limited to working in whole cells due to sub-optimal *in vitro* solubility.

An essential component of our experimental workflow is the application of time- and biological material-efficient microfluidic methods. First, the Microfluidic Profile Explorer (MicroPEX) was used to characterize HLD variants in terms of temperature profiles and substrate specificity.[29] Although state-of-the-art microfluidic systems can characterize >1,000 enzymes in a run,[49] they are limited to water-soluble substrates since the hydrophobic substrates tend to leak to the oil phase.[50] MicroPEX overcomes this limitation by microdialysis and oil-water partitioning[29] and thus enables determination of activities also towards hydrophobic substrates, such as haloalkanes. In comparison with conventional methods for HLD activity characterization, MicroPEX provides up to 1,000-fold lower protein consumption and 100-fold higher throughput.[29]

Second, the Kinetic Microfluidic Autonomous Platform (KinMAP) enables the measurement of temperature-dependent steady-state kinetics and extraction of the energetic and entropic contributions. This combination of kinetic and thermodynamic analysis was applied to characterize six HLD variants superior to currently available enzymes. These experiments revealed thermodynamic parameters driving their catalytic activity. Such valuable mechanistic information is rarely collected for multiple catalysts during protein discovery campaigns due to the time-consuming experiments, requirements of large amounts of purified proteins, and complex data analysis. Enzymes possessing a differential mix of enthalpy and entropy contributions to the catalytic activity provide unique starting points for laboratory evolution, targeting active sites,[14,15] access tunnels[12], or dynamical protein loops.[16] Accordingly, developing an automated droplet-based microfluidic device will open up new opportunities for optimal data collection employing back-loops and machine learning algorithms.[37] We are currently working on expanding the range of KinMAP automation to include adaptive dynamic ranges of substrate and enzyme concentrations during the scan to increase the precision of the kinetic parameter estimations.

Overall, this study doubled the "toolbox" of HLD biocatalysts available for various biotechnological applications by combining advanced bioinformatics with microfluidics. Several discovered enzymes exhibited the highest turnover numbers and catalytic efficiencies ever reported for HLDs. Moreover, unique substrate specificity and unusual enantioselectivity combined with a wide range of operational temperatures make these enzymes industrially relevant. We believe that further development of bioinformatic algorithms and microfluidic enzymology technologies will facilitate database mining for a variety of novel enzymes. Such advances will provide a deeper understanding of sequence-function relationships and contribute to developing a new generation of tools in protein engineering and data-driven prediction of enzyme function.[51]

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead contact, Zbynek Prokop (zbynek@chemi.muni.cz).

Materials availability

This study did not generate new unique reagents.

Data and code availability

Supplemental data set with summarized bioinformatic results for the selected enzymes is provided in the **Supplemental Information**. Other datasets supporting the current study have not been deposited in a public repository but are available from the corresponding author on request.

In silico bioprospecting

The automated *in silico* bioprospecting was based on a protocol described previously.[22] Briefly, representative HLD sequences, including three experimentally characterized HLDs [LinB (NCBI accession number 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[52] 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.[53] Sequence similarity networks (SSN) of putative HLDs were calculated and visualized by EFI-EST[23] and Cytoscape v3.6.1,[24] respectively and further subjected to the EFI-GNT[54] 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).[25] The homology modeling was performed using MODELLER v9.18.[55] The quality of the generated homology models was assessed by MolProbity v4.3.1.[26] Pockets in each homology model were calculated and measured using the CASTp program[56] with a probe radius of 1.4 Å. The CAVER v3.02 program[57] 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.[58] The molecular docking simulations with selected halogenated substrates were conducted using AutoDock Vina[59] with default settings.

Gene synthesis and DNA manipulation

Codon-optimized genes encoding 45 selected HLDs were designed and commercially synthesized (BaseClear B.V., The Netherlands). The synthetic genes were subcloned individually into the expression vector pET-24a(+) between the NdeI and XhoI restriction sites. For plasmid propagation, competent *E. coli* DH5α cells were transformed with individual constructs 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

E. coli cell transformation with plasmid DNA, cultivation in 96-deep well plates, harvesting, SDS-PAGE analysis, and high-throughput affinity purification using the MagneHis Protein Purification System (Promega, USA) are described in detail in **Supplemental experimental procedures 1.1**.

Dehalogenase whole-cell activity screening

The reactions were 200 μL in volume and contained 50 mM PBO buffer (40 mM K_2HPO_4, 10 mM KH_2PO_4, pH 7.5 with 1 mM orthovanadate), 10 mM H_2O_2, 5 $U.mL^{-1}$ *Curvularia inaequalis* chloroperoxidase, 10 μL of whole cells with OD_{600} approximately 5, 12.5 μM aminophenyl fluorescein and 10 mM of a halogenated substrate. The reactions in the HOX assay[23] were started by adding whole cells. The measurement was conducted overnight in a plate reader (30 °C) by measuring fluorescence at 525 nm (488 nm excitation).

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.[60] 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 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 PBO buffer used for the screening of HLD activity using the HOX assay.[23]

Large-scale protein expression and purification

Selected mutant enzymes were expressed in *E. coli* BL21(DE3). Cultivation, harvesting, purification by affinity chromatography, SDS-PAGE analysis, and protein concentration determination are described in detail in **Supplemental experimental procedures 1.2**.

Thermostability

Thermal unfolding was analyzed independently by four methods: (i) microcuvette DSF (UNcle, Unchained labs), (ii) capillary DSF (Prometheus NT.48, NanoTemper Technologies, GmbH), (iii) thermal shift assay (using SYPRO Orange Protein Gel Stain (Thermo Fisher Scientific) in a StepOnePlus Real-Time PCR System (Thermo Fisher Scientific)), and (iv) circular dichroism spectroscopy as a well-established technique (using a Chirascan CD Spectrometer (Applied Photophysics, UK). All methods are described in detail in **Supplemental experimental procedures 1.3**.

Temperature profiles and substrate specificity profiles

Both temperature and substrate specificity profiles were measured using the previously described droplet-based microfluidic profile explorer (MicroPEX),[29] 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 microscale DSF) 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, previously chosen to validate the microfluidic device.[29] Each enzyme was assayed at the temperature closest to its T_{max} value (0-10 °C below T_{max}). A detailed protocol of the microfluidic method was provided previously,[2] and a brief description is available in **Supplemental experimental procedures 1.4**.

Principal component analysis and hierarchical clustering

The matrix containing the activity data of 24 newly identified HLDs and eight previously characterized HLDs towards 27 halogenated substrates (all measured on MicroPEX) was analyzed by principal component analysis (PCA) in MATLAB (MathWorks, USA) 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 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. These transformed data were used to calculate principal components, and the components explaining the highest variability in the data were then plotted to identify substrate specificity groups. Additionally, the hierarchical clustering analysis was performed on the log-transformed data using MATLAB (MathWorks, USA).

Conventional dehalogenase activity measurement

The specific activity of all 24 newly identified HLDs was validated by the conventional method of Iwasaki et al.[61] Dehalogenation reactions were performed at temperatures close to the optimal temperature of each enzyme (**Table S10**) in 25-mL Reacti Flasks closed by Mininert Valves. The reaction mixture was composed of 10 mL of glycine buffer (pH 8.6) and 10 μL of the substrate (1,3-dibromopropane). By adding 0.2 mL of enzyme solution to the mixture, the reaction was initiated. The reaction progress was monitored by periodically withdrawing 1 mL samples from the reaction mixture. Finally, the reaction was stopped by adding 0.1 mL of 35% nitric acid. The reagents with mercuric thiocyanate and ferric ammonium sulfate, employed for detection of halides, were subsequently added to the collected samples, and absorbance of the final mixture was measured spectrophotometrically at 460 nm using microplate reader SUNRISE (Tecan, Austria). Dehalogenase activities were quantified as the rate of product formation with time.

Steady-state kinetics and reaction thermodynamics

A newly introduced droplet-based Microfluidic Autonomous Platform for kinetic analysis (KinMAP), adopted from a previous technology for nanoparticle synthesis,[32] determined the steady-state kinetics and reaction thermodynamics parameters for a selected set of enzymes. The pH-based fluorescence assay to determine HLD kinetics and substrate delivery via a substrate partition between oil and aqueous phase (**Fig. S12**) was the same as in the MicroPEX operation described above.[29] Within one run, the steady-state kinetics of a single enzyme variant was measured with 1,3-dibromopropane in the temperature range of 25-50 °C in 5-degree increments. The HLD enzymatic rate was determined for each temperature and six substrate concentrations. The device and method is described in detail in **Supplemental experimental procedures 1.5**, including **Fig. S12-S16** and **Table S16**.

Global numerical integration of rate equations

The datasets consisting of temperature and concentration dependence of reaction rates were fit globally based on numerical integration of rate equations using KinTek Explorer software 10 (KinTek Corporation, USA),[62] which includes the capability to fit temperature-dependent rate constants.[63] A detailed description of the data fitting is provided in **Supplemental experimental procedures 1.6**.

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 glycine buffer was selected to maintain sufficient buffering capacity in the mildly alkaline pH range corresponding with the pH profiles for most characterized HLDs. A detailed description is provided by Vanacek et al.[8] The kinetic resolution data were fitted globally using KinTek Explorer software (KinTek Corporation, USA), in detail described in **Supplemental experimental procedures 1.7**.

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 nm 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 corrected for the buffer absorbance. Collected CD data were expressed in terms of the mean residue ellipticity (Θ_{MRE}). Secondary structure determination and analysis were carried out on measured ellipticity from 190 nm to 250 nm using the BeStSel online tool with default settings.[64]

Quaternary Structure

The quaternary protein 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). NaCl was supplemented to minimize secondary interactions of the sample components with the resin following the supplier's instructions. 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). Mixture A of standard proteins contained aldolase (158,000 Da), ovalbumin (44,000 Da), ribonuclease A (13,700 Da), and aprotinin (6,500 Da). 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).

Native PAGE

The separation of DstA was investigated by native PAGE. 10 μL of protein sample (0.5 mg.ml^{-1} to 1.5 mg.ml^{-1}) was mixed with 30 μL of 4x loading buffer (3.5 ml 100% glycerol, 2.5 ml 1M Tris-HCl pH 6.8, 4 mg Bromophenol Blue and 4 ml water) and 13 μL of the mix was loaded to the native gel.

Electrophoresis was performed in Tris-glycine electrophoretic buffer at 110V and 4 °C. According to the supplier's protocol, the protein bands of polyacrylamide gels were stained by InstantBlue Protein Stain and checked by GS-800 Calibrated Densitometer (Bio-RAD, USA).

ACKNOWLEDGEMENTS

We would like to thank Simon Godehard and Mark Dörr (Greifswald University, Germany) for fruitful discussions of the experimental design of cell-free protein synthesis experiments. The authors would like to acknowledge funding from the Czech Ministry of Education (grants INBIO CZ.02.1.01/0.0/0.0/16_026/0008451, ESFRI RECETOX LM2018121, and ESFRI ELIXIR LM2018131). This project has received funding from the European Union’s Horizon 2020 research and Innovation program (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. Michal Vasina acknowledges the financial support of his doctoral study by the scholarship Brno Ph.D. Talent.

AUTHOR CONTRIBUTIONS

D.B., J.D., and Z.P. conceived the project and designed the research plan. J.H., D.B., and J.D. performed the *in silico* bioprospecting and selected enzymes for production. P.V. performed small- and large-scale protein expression and dehalogenase whole-cell activity screening, P.V. with C.P.S.B conducted the cell-free protein synthesis experiments. P.V. performed the enzyme characterization in terms of secondary and quaternary structure, enantioselectivity, and thermostability. M.V and T.B. performed the substrate specificity and temperature profile characterization on MicroPEX, while M.V. and S.M. did the multivariate data analysis. D.K., H.F., Z.P. and S.S. developed the microfluidic platform KinMAP and its methodology, D.K. and M.V. performed the kinetic experiments, D.K. and S.M developed code for automated data processing, and Z.P performed the global numerical analysis of thre kinetic data. M.V, P.V., J.D., and Z.P. drafted the manuscript to which all authors contributed. All authors have approved the final version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

ORCID

Michal Vasina: 0000-0002-1504-9929

Pavel Vanacek: 0000-0002-9046-2983

Jiri Hon: 0000-0002-3321-9629

David Kovar: 0000-0002-5550-6143

Hana Faldynova: 0000-0002-8232-5524

Antonin Kunka: 0000-0002-1170-165X

Tomas Buryska: 0000-0003-3740-1679

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

Stanislav Mazurenko: 0000-0003-3659-4819

David Bednar: 0000-0002-6803-0340

Stavros Stavrakis: 0000-0002-0888-5953

Uwe Bornscheuer: 0000-0003-0685-2696

Andrew DeMello: 0000-0003-1943-1356

Jiri Damborsky: 0000-0002-7848-8216

Zbynek Prokop: 0000-0001-9358-4081

REFERENCES

1. Badenhorst, C.P.S., and Bornscheuer, U.T. (2018). Getting Momentum: From Biocatalysis to Advanced Synthetic Biology. Trends Biochem. Sci. *43*, 180–198.

2. Vasina, M., Vanacek, P., Damborsky, J., and Prokop, Z. (2020). Chapter Three - Exploration of enzyme diversity: High-throughput techniques for protein production and microscale biochemical characterization. In Methods in Enzymology Enzyme Engineering and Evolution: General Methods., D. S. Tawfik, ed. (Academic Press), pp. 51–85.

3. Gilks, W.R., Audit, B., de Angelis, D., Tsoka, S., and Ouzounis, C.A. (2005). Percolation of annotation errors through hierarchically structured protein sequence databases. Math. Biosci. *193*, 223–234.

4. Schnoes, A.M., Brown, S.D., Dodevski, I., and Babbitt, P.C. (2009). Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput. Biol. *5*, e1000605.

5. Rembeza, E., and Engqvist, M.K.M. (2021). Experimental and computational investigation of enzyme functional annotations uncovers misannotation in the EC 1.1.3.15 enzyme class. PLOS Comput. Biol. *17*, e1009446.

6. Gargiulo, S., and Soumillion, P. (2021). Directed evolution for enzyme development in biocatalysis. Curr. Opin. Chem. Biol. *61*, 107–113.

7. Bell, E.L., Finnigan, W., France, S.P., Green, A.P., Hayes, M.A., Hepworth, L.J., Lovelock, S.L., Niikura, H., Osuna, S., Romero, E., et al. (2021). Biocatalysis. Nat. Rev. Methods Primer *1*, 1–21.

8. Vasina, M., Velecký, J., Planas-Iglesias, J., Marques, S.M., Skarupova, J., Damborsky, J., Bednar, D., Mazurenko, S., and Prokop, Z. (2022). Tools for computational design and high-throughput screening of therapeutic enzymes. Adv. Drug Deliv. Rev. *183*, 114143.

9. Kokkonen, P., Koudelakova, T., Chaloupkova, R., Daniel, L., Prokop, Z., and Damborsky, J. (2017). Structure-Function Relationships and Engineering of Haloalkane Dehalogenases. In Aerobic Utilization of Hydrocarbons, Oils and Lipids Handbook of Hydrocarbon and Lipid Microbiology., F. Rojo, ed. (Springer International Publishing), pp. 1–21.

10. Musil, M., Konegger, H., Hon, J., Bednar, D., and Damborsky, J. (2019). Computational Design of Stable and Soluble Biocatalysts. ACS Catal. *9*, 1033–1054.

11. Janssen, D.B. (2004). Evolving haloalkane dehalogenases. Curr. Opin. Chem. Biol. *8*, 150–159.

12. Pavlova, M., Klvana, M., Prokop, Z., Chaloupkova, R., Banas, P., Otyepka, M., Wade, R.C., Tsuda, M., Nagata, Y., and Damborsky, J. (2009). Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. Nat. Chem. Biol. *5*, 727–733.

13. Brezovsky, J., Babkova, P., Degtjarik, O., Fortova, A., Gora, A., Iermak, I., Rezacova, P., Dvorak, P., Smatanova, I.K., Prokop, Z., et al. (2016). Engineering a de Novo Transport Tunnel. ACS Catal. *6*, 7597–7610.

14. Sykora, J., Brezovsky, J., Koudelakova, T., Lahoda, M., Fortova, A., Chernovets, T., Chaloupkova, R., Stepankova, V., Prokop, Z., Smatanova, I.K., et al. (2014). Dynamics and hydration explain failed functional transformation in dehalogenase design. Nat. Chem. Biol. *10*, 428–430.

15. Liskova, V., Stepankova, V., Bednar, D., Brezovsky, J., Prokop, Z., Chaloupkova, R., and Damborsky, J. (2017). Different Structural Origins of the Enantioselectivity of Haloalkane Dehalogenases toward Linear β-Haloalkanes: Open–Solvated versus Occluded–Desolvated Active Sites. Angew. Chem. Int. Ed. *56*, 4719–4723.

16. Schenkmayerova, A., Pinto, G.P., Toul, M., Marek, M., Hernychova, L., Planas-Iglesias, J., Daniel Liskova, V., Pluskal, D., Vasina, M., Emond, S., et al. (2021). Engineering the protein dynamics of an ancestral luciferase. Nat. Commun. *12*, 3616.

17. Beerens, K., Mazurenko, S., Kunka, A., Marques, S.M., Hansen, N., Musil, M., Chaloupkova, R., Waterman, J., Brezovsky, J., Bednar, D., et al. (2018). Evolutionary Analysis As a Powerful Complement to Energy Calculations for Protein Stabilization. ACS Catal. *8*, 9420–9428.

18. Bednar, D., Beerens, K., Sebestova, E., Bendl, J., Khare, S., Chaloupkova, R., Prokop, Z., Brezovsky, J., Baker, D., and Damborsky, J. (2015). FireProt: Energy- and Evolution-Based Computational Design of Thermostable Multiple-Point Mutants. PLOS Comput. Biol. *11*, e1004556.

19. Floor, R.J., Wijma, H.J., Colpa, D.I., Ramos-Silva, A., Jekel, P.A., Szymański, W., Feringa, B.L., Marrink, S.J., and Janssen, D.B. (2014). Computational Library Design for Increasing Haloalkane Dehalogenase Stability. ChemBioChem *15*, 1660–1672.

20. Chaloupkova, R., Liskova, V., Toul, M., Markova, K., Sebestova, E., Hernychova, L., Marek, M., Pinto, G.P., Pluskal, D., Waterman, J., et al. (2019). Light-Emitting Dehalogenases: Reconstruction of Multifunctional Biocatalysts. ACS Catal. *9*, 4810–4823.

21. Hon, J., Borko, S., Stourac, J., Prokop, Z., Zendulka, J., Bednar, D., Martinek, T., and Damborsky, J. (2020). EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res. *48*, W104–W109.

22. Vanacek, P., Sebestova, E., Babkova, P., Bidmanova, S., Daniel, L., Dvorak, P., Stepankova, V., Chaloupkova, R., Brezovsky, J., Prokop, Z., et al. (2018). Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. ACS Catal. *8*, 2402–2412.

23. Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen, K.L. (2015). Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for generating protein sequence similarity networks. Biochim. Biophys. Acta BBA - Proteins Proteomics *1854*, 1019–1037.

24. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. Genome Res. *13*, 2498–2504.

25. Barrett, T., Clark, K., Gevorgyan, R., Gorelenkov, V., Gribov, E., Karsch-Mizrachi, I., Kimelman, M., Pruitt, K.D., Resenchuk, S., Tatusova, T., et al. (2012). BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. Nucleic Acids Res. *40*, D57–D63.

26. 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., et al. (2018). MolProbity: More and better reference data for improved all-atom structure validation. Protein Sci. Publ. Protein Soc. *27*, 293–315.

27. Chovancová, E., Kosinski, J., Bujnicki, J.M., and Damborský, J. (2007). Phylogenetic analysis of haloalkane dehalogenases. Proteins Struct. Funct. Bioinforma. *67*, 305–316.

28. Aslan-Üzel, A.S., Beier, A., Kovář, D., Cziegler, C., Padhi, S.K., Schuiten, E.D., Dörr, M., Böttcher, D., Hollmann, F., Rudroff, F., et al. (2020). An Ultrasensitive Fluorescence Assay for the Detection of Halides and Enzymatic Dehalogenation. ChemCatChem *12*, 2032–2039.

29. Buryska, T., Vasina, M., Gielen, F., Vanacek, P., van Vliet, L., Jezek, J., Pilat, Z., Zemanek, P., Damborsky, J., Hollfelder, F., et al. (2019). Controlled Oil/Water Partitioning of Hydrophobic Substrates Extending the Bioanalytical Applications of Droplet-Based Microfluidics. Anal. Chem. *91*, 10008–10015.

30. Koudelakova, T., Bidmanova, S., Dvorak, P., Pavelka, A., Chaloupkova, R., Prokop, Z., and Damborsky, J. (2013). Haloalkane dehalogenases: Biotechnological applications. Biotechnol. J. *8*, 32–45.

31. Koudelakova, T., Chovancova, E., Brezovsky, J., Monincova, M., Fortova, A., Jarkovsky, J., and Damborsky, J. (2011). Substrate specificity of haloalkane dehalogenases. Biochem. J. *435*, 345 LP – 354.

32. Lignos, I., Stavrakis, S., Kilaj, A., and deMello, A.J. (2015). Millisecond-Timescale Monitoring of PbS Nanoparticle Nucleation and Growth Using Droplet-Based Microfluidics. Small *11*, 4009–4017.

33. Johnson, K.A. (2019). New standards for collecting and fitting steady state kinetic data. Beilstein J. Org. Chem. *15*, 16–29.

34. Kmunícek, J., Hynková, K., Jedlicka, T., Nagata, Y., Negri, A., Gago, F., Wade, R.C., and Damborský, J. (2005). Quantitative Analysis of Substrate Specificity of Haloalkane Dehalogenase LinB from Sphingomonas paucimobilis UT26. Biochemistry *44*, 3390–3401.

35. Chrast, L., Tratsiak, K., Planas-Iglesias, J., Daniel, L., Prudnikova, T., Brezovsky, J., Bednar, D., Kuta Smatanova, I., Chaloupkova, R., and Damborsky, J. (2019). Deciphering the Structural Basis of High Thermostability of Dehalogenase from Psychrophilic Bacterium Marinobacter sp. ELB17. Microorganisms *7*, 498.

36. Planas-Iglesias, J., Marques, S.M., Pinto, G.P., Musil, M., Stourac, J., Damborsky, J., and Bednar, D. (2021). Computational design of enzymes for biotechnological applications. Biotechnol. Adv. *47*, 107696.

37. Mazurenko, S., Prokop, Z., and Damborsky, J. (2020). Machine Learning in Enzyme Engineering. ACS Catal. *10*, 1210–1223.

38. Chmelova, K., Sebestova, E., Liskova, V., Beier, A., Bednar, D., Prokop, Z., Chaloupkova, R., and 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. *86*, e02820-19.

39. Kotik, M., Vanacek, P., Kunka, A., Prokop, Z., and Damborsky, J. (2017). Metagenome-derived haloalkane dehalogenases with novel catalytic properties. Appl. Microbiol. Biotechnol. *101*, 6385–6397.

40. Li, F., Luan, Z., Chen, Q., Xu, J., and Yu, H. (2016). Rational selection of circular permutation sites in characteristic regions of the α/β-hydrolase fold enzyme RhEst1. J. Mol. Catal. B Enzym. *125*, 75–80.

41. Truppo, M.D. (2017). Biocatalysis in the Pharmaceutical Industry: The Need for Speed. ACS Med. Chem. Lett. *8*, 476–480.

42. Zaparucha, A., de Berardinis, V., and Vaxelaire-Vergne, C. (2018). Chapter 1 Genome Mining for Enzyme Discovery. In Modern Biocatalysis: Advances Towards Synthetic Biological Systems (The Royal Society of Chemistry), pp. 1–27.

43. Marchot, P., and Chatonnet, A. (2012). Enzymatic Activity and Protein Interactions in Alpha/Beta Hydrolase Fold Proteins: Moonlighting Versus Promiscuity. Protein Pept. Lett. *19*, 132–143.

44. Schuiten, E.D., Badenhorst, C.P.S., Palm, G.J., Berndt, L., Lammers, M., Mican, J., Bednar, D., Damborsky, J., and Bornscheuer, U.T. (2021). Promiscuous Dehalogenase Activity of the Epoxide Hydrolase CorEH from Corynebacterium sp. C12. ACS Catal. *11*, 6113–6120.

45. Rauwerdink, A., and Kazlauskas, R.J. (2015). How the Same Core Catalytic Machinery Catalyzes 17 Different Reactions: the Serine-Histidine-Aspartate Catalytic Triad of α/β-Hydrolase Fold Enzymes. ACS Catal. *5*, 6153–6176.

46. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature *596*, 583–589.

47. Prokop, Z., Koudelakova, T., Bidmanova, S., and Damborsky, J. (2018). Chapter 18:Enzymes for Detection and Decontamination of Chemical Warfare Agents. In Modern Biocatalysis, pp. 539–565.

48. Hon, J., Marusiak, M., Martinek, T., Kunka, A., Zendulka, J., Bednar, D., and Damborsky, J. (2021). SoluProt: prediction of soluble protein expression in Escherichia coli. Bioinformatics *37*, 23–28.

49. Markin, C.J., Mokhtari, D.A., Sunden, F., Appel, M.J., Akiva, E., Longwell, S.A., Sabatti, C., Herschlag, D., and Fordyce, P.M. (2021). Revealing enzyme functional architecture via high-throughput microfluidic enzyme kinetics. Science *373*, eabf8761.

50. Neun, S., Zurek, P.J., Kaminski, T.S., and Hollfelder, F. (2020). Chapter Thirteen - Ultrahigh throughput screening for enzyme function in droplets. In Methods in Enzymology Enzyme Engineering and Evolution: General Methods., D. S. Tawfik, ed. (Academic Press), pp. 317–343.

51. Mokhtari, D.A., Appel, M.J., Fordyce, P.M., and Herschlag, D. (2021). High throughput and quantitative enzymology in the genomic era. Curr. Opin. Struct. Biol. *71*, 259–273.

52. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402.

53. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

54. Zallot, R., Oberg, N., and Gerlt, J.A. (2019). The EFI Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. Biochemistry *58*, 4169–4182.

55. Webb, B., and Sali, A. (2016). Comparative Protein Structure Modeling Using MODELLER. Curr. Protoc. Bioinforma. *54*, 5.6.1-5.6.37.

56. Tian, W., Chen, C., Lei, X., Zhao, J., and Liang, J. (2018). CASTp 3.0: computed atlas of surface topography of proteins. Nucleic Acids Res. *46*, W363–W367.

57. Pavelka, A., Sebestova, E., Kozlikova, B., Brezovsky, J., Sochor, J., and Damborsky, J. (2016). CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules. IEEE/ACM Trans. Comput. Biol. Bioinform. *13*, 505–517.

58. Wilkinson, D.L., and Harrison, R.G. (1991). Predicting the Solubility of Recombinant Proteins in *Escherichia coli*. Bio/Technology *9*, 443.

59. Trott, O., and Olson, A.J. (2010). AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J. Comput. Chem. *31*, 455–461.

60. Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K., and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nat. Biotechnol. *19*, 751–755.

61. Iwasaki, I., Utsumi, S., and Ozawa, T. (1952). New Colorimetric Determination of Chloride using Mercuric Thiocyanate and Ferric Ion. Bull. Chem. Soc. Jpn. *25*, 226–226.

62. Johnson, K.A., Simpson, Z.B., and Blom, T. (2009). Global Kinetic Explorer: A new computer program for dynamic simulation and fitting of kinetic data. Anal. Biochem. *387*, 20–29.

63. Li, A., Ziehr, J.L., and Johnson, K.A. (2017). A new general method for simultaneous fitting of temperature and concentration dependence of reaction rates yields kinetic and thermodynamic parameters for HIV reverse transcriptase specificity. J. Biol. Chem. *292*, 6695–6702.

64. Micsonai, A., Wien, F., Bulyáki, É., Kun, J., Moussong, É., Lee, Y.-H., Goto, Y., Réfrégiers, M., and Kardos, J. (2018). BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. Nucleic Acids Res. *46*, W315–W322.

TABLES AND FIGURES

Table 1. Summary of biochemical properties of HLDs.

Enzyme	Yield (mg. L^{-1})	Specific activity* (nmol.s^{-1}.mg^{-1})	T_{onset} (°C)	T_m^{app} (°C)	T_{max} (°C)	E value	
						2-bromopentane	ethyl 2-bromo-propionate
DstA	70	2.5±0.1	30.9±0.2	43.4±0.1	30	1.27±0.01	2.59±0.04
DfxA	10	2.9±0.2	30.5±0.6	40.6±0.5	35	n.a.	n.a.
DlaA	40	3.9±0.1	35.6±1.2	48.1±0.6	30	n.a.	n.a.
DaxA	120	1.1±0.2	42.9±0.2	48.7±0.1	35	16.4±0.3	n.a.
DsmA	120	86.1±0.4	27.6±0.1	35.7±0.1	25	1.60±0.01	81±1
DmmarA	20	5.9±0.1	32.3±0.1	42.1±0.2	30	6.33±0.04	1.22±0.01
DathA	60	5.7±0.2	38.1±0.6	46.4±0.1	35	27.3±0.4	45 ± 1
DmaA	30	211.1±4.7	32.5±0.1	40.2±0.3	35	2.13±0.01	49.8±0.4
DspoA	80	860.7±16.8	50.8±0.2	58.7±0.6	50	9.755±0.083	128±1
DexA	120	572.7±10.1	43.4±1.1	47.5±0.4	45	5.46±0.04	152±2
DppsA	100	29.0±0.1	24.7±0.2	38.1±0.2	35	3.32±0.03	84±1
DeaA	70	405.0±7.6	45.3±0.1	52.2±0.2	45	>200	113 ± 2
DmgaA	100	6.1±0.1	38.2±1.6	44.7±0.9	40	n.a.	n.a.
DprxA	150	630.1±14.3	44.3±1.7	51.8±0.3	45	3.23±0.02	>200
DrgA	20	1.8±0.2	36.8±0.4	44.2±0.4	35	n.a.	n.a.
DmbaA	10	132.5±1.7	36.8±0.3	46.6±0.2	45	5.54±0.04	22.2±0.2
DthA	90	31.3±0.7	40.4±0.3	49.9±0.9	35	155.9±0.7	>200
DphxA	30	595.7±7.0	47.0±0.6	55.4±0.2	35	1.82±0.01	26.0±0.2
DthB	20	121.8±4.4	44.8±0.6	53.4±0.4	45	2.98±0.02	15.9±0.1
DnbA	90	6.5±0.2	37.3±0.1	47.8±0.4	40	14.1±0.3	n.a.
DhxA	120	610.8±0.9	44.1±0.4	53.1±0.3	35	1.574±0.011	>200
DspxA	30	81.9±0.1	44.2±0.3	53.3±0.2	35	42.1±0.5	156±3
DchA	20	143.3±5.2	47.0±0.1	55.2±0.8	40	2.52±0.02	27.7±0.3
Dcta	10	5.0±0.1	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 a temperature close to the optimal temperature (**Table S10**); T_{onset} – unfolding onset temperature by capillary DSF; T_m^{app} – apparent melting temperature by capillary DSF; T_{max} – maximum HLD activity; n.a. – no activity

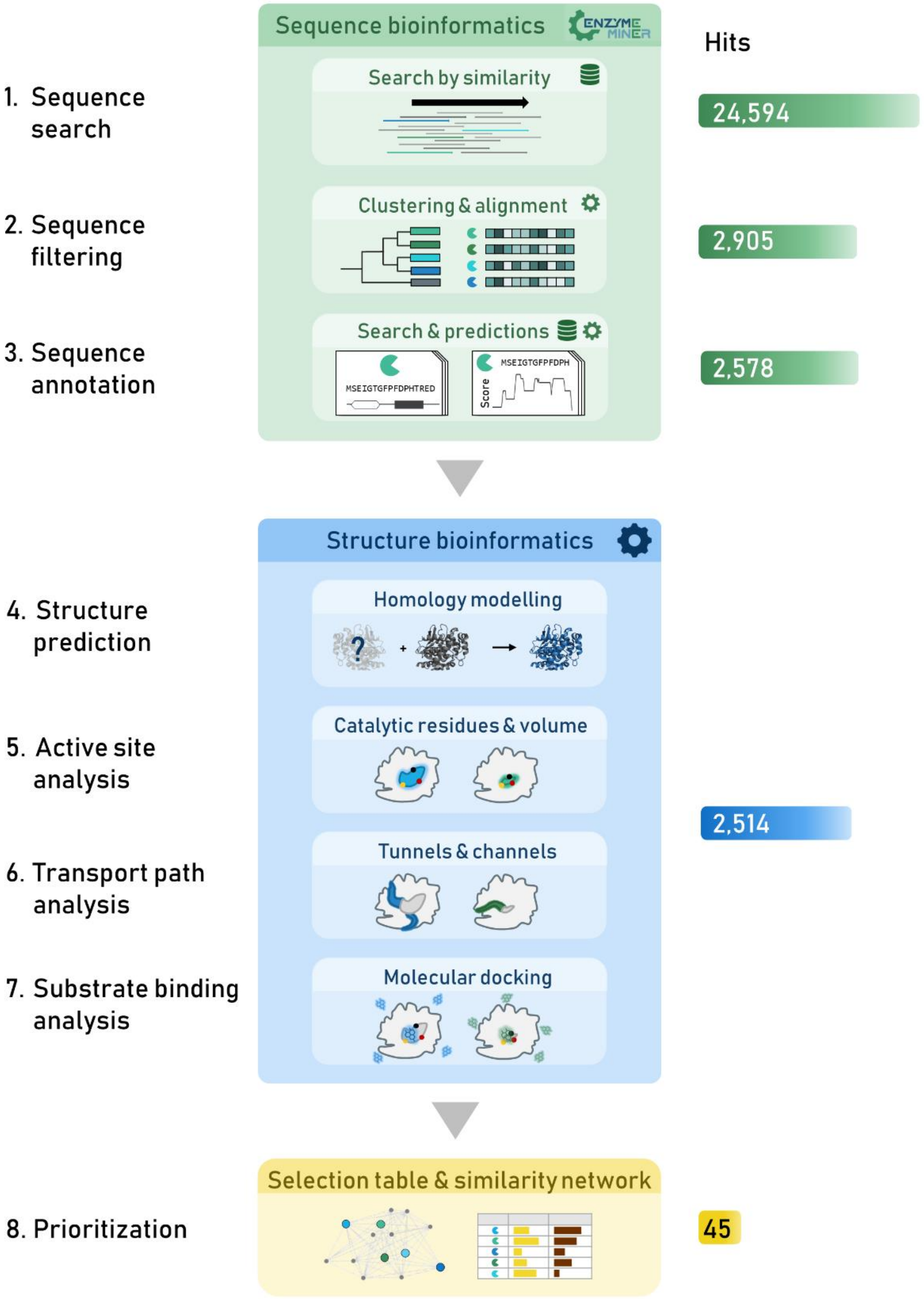


Figure 1 Bioinformatics workflow enabling selection of "small-but-smart" set of enzymes for thorough experimental characterization. The sequence bioinformatics pipeline (green) has been previously implemented as a web-based software tool EnzymeMiner.[21] Automated sequence analysis has been complemented by a structural bioinformatics pipeline (blue), providing additional high-quality annotations for prioritization and selection of a "small-but-smart" set of proteins (yellow) for experimental characterization. The individual steps are illustrated in the middle panel and labeled. The numbers of hits achieved in every step are highlighted on the right side. The symbols located in the upper right corner distinguish the steps utilizing database or software tools.

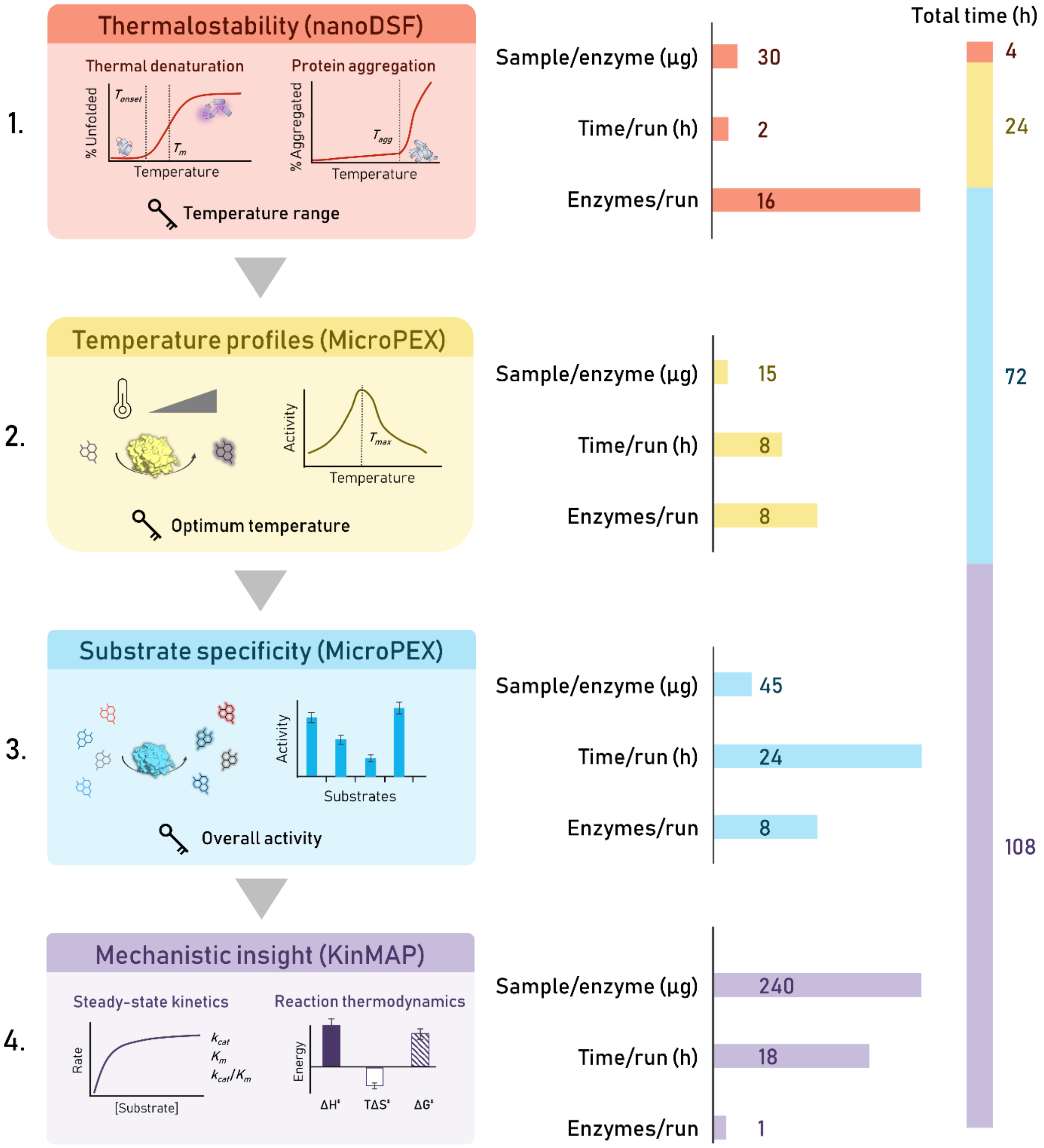


Figure 2 Experimental workflow for efficient and thorough characterization of well-expressed enzymes. The left side captures individual characterization steps with the microfluidic techniques in brackets. The key parameters of each characterization necessary for the experimental design of the following step (key symbol) are at the bottom of each frame. The right side describes the characteristics of each technique with respect to sample requirements per enzyme in µg, time requirements per run in hours, and the possible number of enzymes measured in parallel. The timescale for characterization of the 24 discovered dehalogenases (steps 1-3) and 6 selected enzymes (step 4) is shown in hours of measurement on the very right.

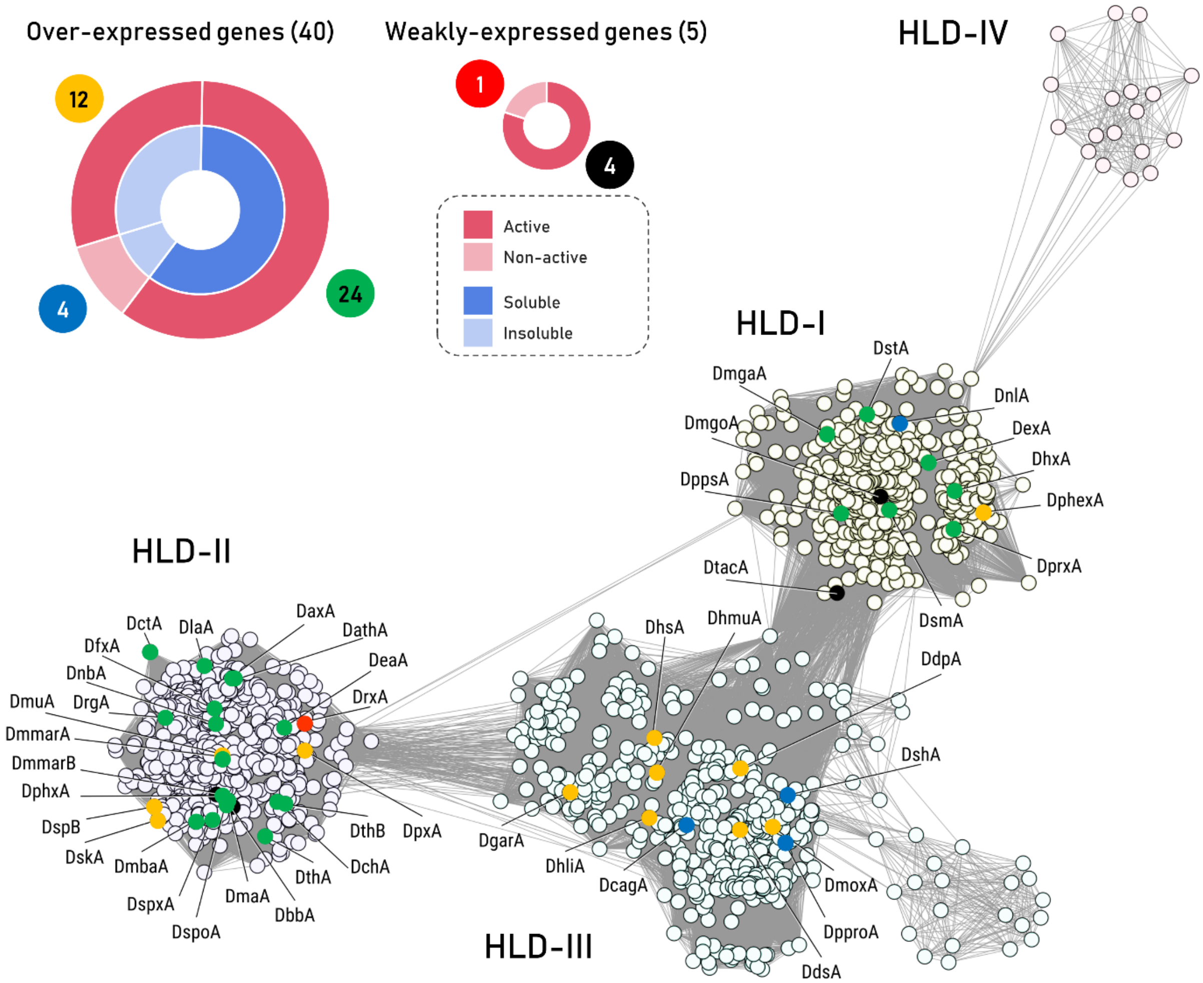


Figure 3 Sequence similarity network for HLDs categorized by their expression, solubility, and activity. The putative HLDs are clustered into four subfamilies: HLD-I, HLD-II, HLD-III, and HLD-IV. The sequences were first clustered at 50% identity to reduce the number of nodes and edges. The sequences with higher identity are consolidated into a single node. Edge lengths indicate sequence similarity between representative sequences of the connected nodes. Sequence similarity networks of putative HLDs were calculated and visualized by EFI-EST[23] and Cytoscape v3.6.1[24]. The results from expression, solubility and activity analysis are shown in the doughnut graphs (upper left). Enzymes were assigned to five distinct groups of enzymes based on their expressibility, solubility, and activity, indicated by different colors in doughnut graphs and the sequence similarity network. A set of 24 well-soluble and active enzymes (green) was subjected to systematic biochemical characterization. Four weakly expressed genes (black) and twelve over-expressed genes providing proteins with low solubility (yellow) were tested positive with at least one of the five halogenated substrates in the whole-cell activity screening assay (**Table S8**). Four over-expressed genes providing insoluble proteins (blue) and one weakly-expressed gene (red) led to proteins that did not exhibit any activity in our tests.

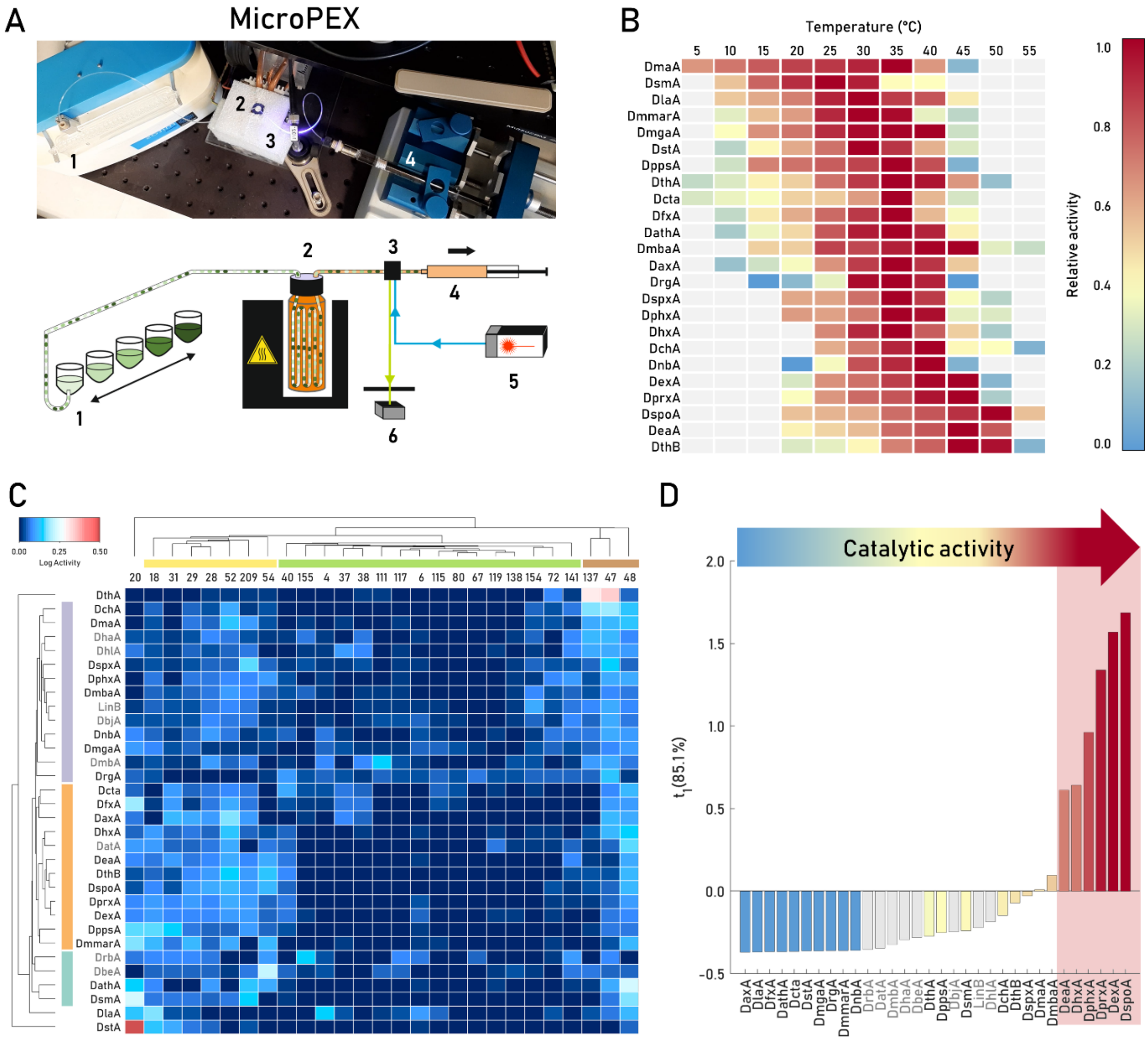


Figure 4 Temperature profiles and substrate specificity by droplet-based microfluidics. (A) Photograph and scheme of the droplet-based microfluidic profile explorer (MicroPEX) to determine temperature profiles and substrate specificity. Depicted are the main parts of the device, including the droplet generator (1), incubation chamber for substrate delivery under temperature control (2), detection cell (3), microfluidic pump (4), fluorescence excitation laser (5), and a photodetector (6). **(B)** Temperature profiles. The heat map represents the relative activity of individual enzymes. **C,** Multivariate analysis of substrate specificity. A double-dendrogram heat map of log-transformed data depicts the similarity of enzyme activity (vertical axis) and conversion of halogenated substrates (horizontal axis). Major groups of enzymes and substrates are highlighted with the same color. **(D)** Multivariate analysis of catalytic activity. The score plot t_1 compares the enzymes in terms of their overall activity with 27 substrates and explains 85.1% of the data variance. The light red frame highlights new enzymes with an outstanding catalytic activity, which were characterized by steady-state kinetics and reaction thermodynamics (**Fig. 5**). The previously characterized HLDs are colored grey in **C** and **D**. The heat maps (**B, C**) and bars (**D**) are color-coded by enzymatic activity from low activity (blue) to medium activity (yellow) and high activity (red).

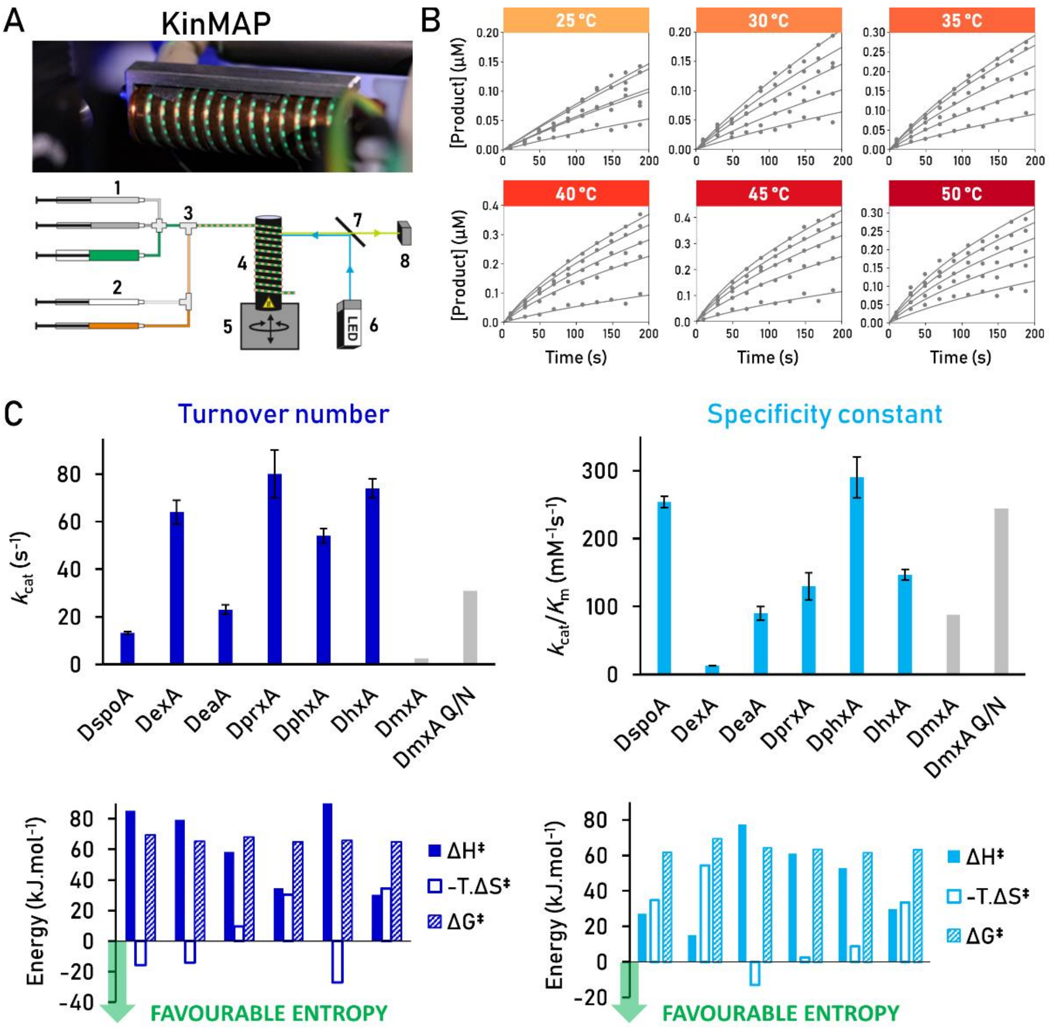


Figure 5 Mechanistic analysis by droplet-based microfluidics and global numerical integration. (**A)** The droplet-based Kinetic Microfluidic Autonomous Platform (KinMAP) enables kinetic and thermodynamic measurements. A photograph (top) illustrates the reaction droplets traveling through the incubation zone with temperature control. The schematic of the device (bottom) depicts syringe pumps for aqueous solutions of reactants (1) and oil phase (2), droplet generator (3), reaction zone with temperature control (4), motorized stage (5), excitation light source (6), dichroic mirror (7) and detection of emitted light (8). **(B)** Example of the kinetic and thermodynamic data collected for DspoA by monitoring the enzymatic conversion under different substrate concentrations (0-1 mM 1,3-dibromopropane) at different temperatures (25-50 °C) in 1 mM HEPES buffer (pH 8.2). Each data point represents an average of 20 repetitions; the solid lines represent the best global fit. The data for all selected enzymes, parameter estimates, and statistics are summarized in **Fig. S11** and **Table S15**. **(C)** The kinetic parameters (top figures), turnover number (k_{cat}), and specificity constant (k_{cat}/K_m) were obtained by global fitting complex kinetic and thermodynamic data (values at reference temperature 310.15 K, 37 °C). The error bars represent standard errors. Grey columns represent previously reported values for the reaction of wild-type DmxA and its engineered single-point mutant DmxA Q/N, both with 1,3-dibromopropane (100 mM glycine buffer, pH 8.6, 37 °C).[35] The contributions of activation enthalpy (ΔH‡) and entropy (-T.ΔS‡) to the Gibbs free energy of activation (ΔG‡) derived from the temperature dependence of catalytic turnover (k_{cat}) and specificity constant (k_{cat}/K_m) for the reference temperature 310.15 K (bottom figures). The green arrows show favorable entropy values lowering the activation barrier.

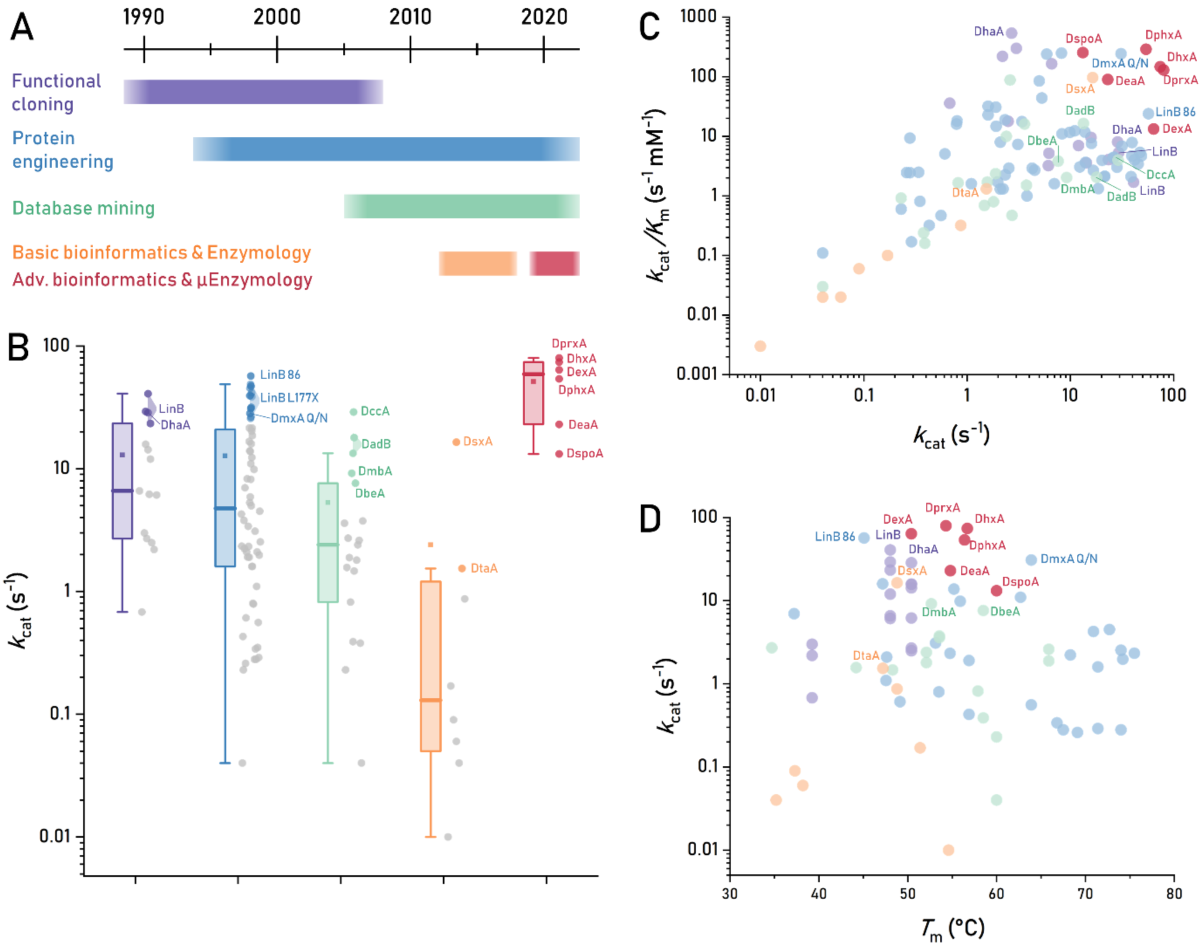


Figure 6 Functional characteristics of the HLD family members. (A) Five strategies were used to obtain catalytically efficient HLDs: Functional cloning (purple), protein engineering (blue), database mining (green), basic bioinformatics & enzymology (orange), and advanced bioinformatics & µEnzymology (red, present study). **(B)** Box chart comparing turnover numbers for enzyme variants obtained by respective strategies with data points to the right of the boxes. The box shows median (line), mean (small square), quartiles, minima, and maxima. The 25% best data points are highlighted in color, while the remaining data points are grey. Selected best variants are labeled. **(C)** The dependence of catalytic efficiency on turnover numbers provides the complex catalytic evaluation of enzymes. **(D)** The dependence of turnover numbers on the melting temperature of each variant provides activity-stability relationships. The data were collected from published research. The kinetic data gathered here were measured with the best substrates for HLDs, 1,2-dibromoethane and 1,3-dibromopropane at 37 °C or lower temperatures. Plot **D** does not contain all the data points from **B** and **C** since thermostability data are unavailable for some variants.