Functional and Mechanistic Characterization of an Enzyme Family Combining Bioinformatics and High-Throughput Microfluidics

Michal Vasina
Masaryk University
Pavel Vanacek
Loschmidt Laboratories, Department of Experimental Biology and Centre for Toxic Compounds in the Environment RECETOX, Faculty of Science, Masaryk University, 625 00 Brno
Jiri Hon
Brno University Technology
David Kovar
Masaryk University
Hanka Faldynova
Masaryk University
Antonin Kunka
Masaryk University
Christoffer Badenhorst
University Greifswald
Tomas Buryska
Masaryk University
Stanislav Mazurenko
Masaryk University
David Bednar
Masaryk University
Stavros Stavros
ETH Zurich https://orcid.org/0000-0002-0888-5953
Uwe Bornscheuer
nstitute of Biochemistry https://orcid.org/0000-0003-0685-2696
Andrew deMello
ETH Zurich
Jiri Damborsky (jiri@chemi.muni.cz)
Masaryk University https://orcid.org/0000-0002-7848-8216
Zbynek Prokop
Masaryk University

Article

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

Posted Date: November 8th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1027271/v1

Functional and Mechanistic Characterization of an Enzyme Family Combining Bioinformatics and High-Throughput 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, 625 00 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

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

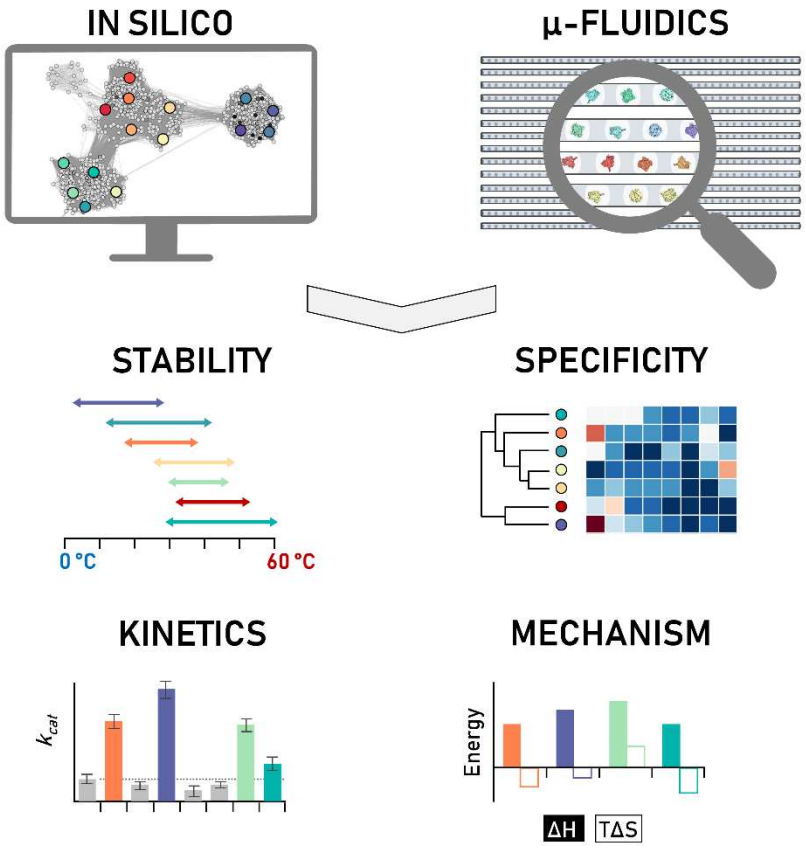
* Authors for correspondence: jiri@chemi.muni.cz, ORCID, 0000-0002-7848-8216; zbynek@chemi.muni.cz, ORCID 0000-0001-9358-4081

ABSTRACT

Next-generation sequencing doubles genomic databases every 2.5 years. The accumulation of sequence data raises the need to speed up functional analysis. Herein, we present a pipeline integrating bioinformatics and microfluidics and its application for high-throughput mining of novel haloalkane dehalogenases. We employed bioinformatics to identify 2,905 putative dehalogenases and selected 45 representative enzymes, of which 24 were produced in soluble form. Droplet-based microfluidics accelerates subsequent experimental testing up to 20,000 reactions per day while achieving 1,000-fold lower protein consumption. This resulted in doubling the dehalogenation “toolbox" characterized over three decades, yielding biocatalysts surpassing the efficiency of currently available enzymes. Combining microfluidics with modern global data analysis provided precious mechanistic information related to the high catalytic efficiency of new variants. This pipeline applied to other enzyme families can accelerate the identification of biocatalysts for industrial applications as well as the collection of high-quality data for machine learning.

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

GRAPHICAL ABSTRACT



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 pool of biodiversity represents an inventory of valuable biocatalysts fulfilling the demands for new and higher-quality products and services. We still only scratch the surface of potential biocatalytic applications as the current pool of available enzymes catering to biomedical, pharmaceutical, and industrial applications remain 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] The rate of sequence data accumulation far exceeds the speed of functional studies. In practice, the low ratio of explored to unexplored protein sequences reflects the low-throughput efficiency of existing biochemical techniques compared to high-throughput next-generation sequencing technologies.

The magnitude of the "big data" generated in biology is enlarged by the fact that a large portion of functional annotations contains vague, indirect, or even incorrect descriptions.[4] Extrapolation from sequence to protein function is not trivial and often proves incorrect when tested experimentally. Thus, the challenge now is to explore how this new sequence information can be harnessed to obtain new biocatalysts and what modern techniques can be used to characterize them rapidly. Several strategies have been developed to exploit the vast number of enzyme sequences in genome and metagenome databases 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 screening large pools of sequences.[7] In a previous study, we have developed a sequence-based strategy that identifies putative members of known enzyme families and facilitates their prioritization for experimental characterization.[8] The main benefit of such a genome mining approach is the effective identification of thousands of putative hits among millions of sequence entries and the rational selection of a restricted set of attractive targets.[9,10] Subsequent experimental characterization of selected representative candidates is a crucial but rate-limiting process. Conventional techniques used to collect experimental biochemical data are labour and time-demanding, cost-ineffective, and low-throughput. Given the many attractive targets, analytical tools capable of collecting functional and mechanistic data with high throughput are particularly attractive to accelerate the discovery of new biocatalysts.[11,12] In this context, implementing microfluidic technology in experimental workflows offers key advantages, including low sample consumption, the possibility of parallelization, and integration with optical detection.

Herein, we describe an integrated workflow for effectively mining novel protein family members using computer-assisted genome mining and high-throughput experimental characterization (**Fig. 1**). First, we applied an automated bioinformatics workflow to identify putative family members and select

promising candidates (**Fig. 1a**). Using homology modelling, we predicted tertiary structures of all target proteins, analyzed them for cavities and access tunnels, and conducted molecular docking simulations with a set of representative substrates. Next, we experimentally characterized the selected hits from the *in silico* screening by employing small-scale expression followed by microfluidic and microanalytical analysis (**Fig. 1b**). Achieving an analytical throughput of up to 20,000 reactions per day and reduced protein consumption to only tens of micrograms, three orders of magnitude less than conventional approaches, droplet-based microfluidics enabled fast collection of combinatorial biochemical data, as well as detailed kinetic and thermodynamic properties providing valuable mechanistic information. In a single run of this workflow, we achieved doubling the number of experimentally characterized haloalkane dehalogenases (HLDs). We obtained a pool of new biocatalysts with industrially attractive characteristics such as high catalytic efficiencies, unique substrate specificities, high enantioselectivities, and broad temperature tolerance.

RESULTS

I. *In Silico* Screening Using an Automated Workflow

Model Enzyme Family. The automated database-mining protocol was tested with the HLDs (**Fig. 1a**). 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[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 rate generation of genomic databases is doubling every ~2.5 years. Therefore, periodic mining of novel genes is highly required. We rerun the *in silico* screening with the same input sequence as previously[8] using the current version of the NCBI nr database and a recently developed tool for automated database mining.[18] In comparison to the initial version, the presented *in silico* workflow has been significantly 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 (910), and HLD-IV (22), based on the sequence identity and the composition of their catalytic pentads. Despite having identical catalytic pentads, HLD-III and HLD-IV were clustered separately based on differences in their sequence similarity. Incomplete and degenerated sequences were filtered out by the construction of multiple sequence alignments of individual subfamilies. Sequence-similarity networks were used to visualize relationships among putative HLD sequences (**Fig. 2**). The most apparent defining features are 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 were filtered out.

Annotation and Homology Modelling. 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 S1**). Most members from the putative HLD-IV subfamily were annotated as HLDs, despite their low sequence identity to experimentally characterized HLDs or other HLD 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 members from subfamily HLD-I and HLD-II but only a limited number of HLD-III members and none of the HLD-IV members. The predicted volumes of catalytic pockets ranged from 60 $Å^3$ to 2,170 $Å^3$ (**Fig. S1**).

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 (**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. A few sequences were picked from each datasheet to make the selection as diverse as possible (**Table S2**). 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 are usually poorly expressible in bacterial host systems. Altogether 45 diverse sequences were selected as targets for experimental characterization (**Table S3**, **Table S4**).

II. Small-Scale Protein Expression

A 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. 2**). 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 stability 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 S5**). 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 minority 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). 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-express" HLDs. Activity measurements with whole cells showed that 40 out of 45 proteins (89%) had HLD activity with at least one of the five substrates tested (**Fig. 2**, **Table S6**). This confirms that most of the identified genes code for proteins with target activity. These include not only those 24 well-expressed genes yielding soluble proteins listed as sufficient for biochemical characterization but also 12 genes coding for HLDs with low or poor solubility and 4 genes showing weak expression in the small–scale expression analysis (**Fig. S3a**).

III. Microfluidic and Microanalytical Characterization

The experimental pipeline comprised commercial microanalytical instruments (e.g., capillary and microcuvette differential scanning fluorimetry) and custom-made microfluidic platforms leading to 1000-fold lower consumption of protein and increased throughput up to 20,000 reactions per day compared to 96 well plates. The methodology provided experimental data on protein stability, catalytic activity, temperature profiles, substrate specificity, kinetics, and thermodynamics (**Table 1**).

Protein Stability. The stability of the novel HLDs was analyzed in a high-throughput manner by monitoring changes in extrinsic (SYPRO orange dye) and intrinsic (tryptophan) fluorescence during temperature scanning experiments, using thermal shift assay (TSA) and microscale differential scanning fluorimetry (DSF), respectively. The midpoint of the denaturation curve (apparent melting temperature, T_m^{app}) was used for the stability evaluation. The results of fast microscale methods showed a very good agreement (R^2 0.79 and 0.93 for TSA and micro-DSF, respectively) with conventional circular dichroism (CD) spectroscopy (**Table S7**, **Fig. S5**). The apparent melting temperatures (T_m^{app}) values, shown in **Table S7**, reflect the mesophilic origins of the novel HLDs (40-55 °C) primarily. Exceptions are the DsmA and DppsA, which exhibited T_m^{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_m^{app} value of 60 °C.

Temperature Profiling. Temperature profiling was performed using a capillary-based droplet microfluidic system described elsewhere.[23] The new dehalogenases obtained in this study showed activity over a wide temperature range (**Fig. 3b**, **Fig. S6**, **Table S8).** DmaA was especially unique, as it retained more than 65 % dehalogenase activity at 5 °C. This dehalogenase performs equally well at this low temperature compared to benchmark dehalogenases at their temperature optima (30-45 °C).[17] 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 (**Fig. S7**).

Substrate Specificity Profiling. Profiling of substrate specificity towards a set of 27 representative substrates was also conducted using the capillary-based droplet microfluidic system (**Table S9**). This structurally diverse set of substrates reflects the application of HLDs, including environmentally important compounds (**Table S10)**. The raw data of specific activities (**Table S11)** 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 are linear alkyl chains of 2-4 carbon atoms (**Fig. S8a)** and that the majority of the HLDs can convert this type of substrates 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), and 1,2-dibromoethane (#47). The substrate specificity profiling also identified a set of "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.[24] It is worth noting that two-thirds of the newly discovered enzymes possess broad substrate specificity and convert > 80 % of the substrates tested (**Table S12**). 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 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[23] and 24 newly identified HLDs. This analysis aimed to compare the enzymes according to their score along with the first principal component (t_1), thus quantifying their global activity against the set of substrates activity (**Fig. 3d**). 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 a selected, overall well-converted substrate, 1,3-dibromopropane (**Fig. S9**). Six out of these 11 highly active enzymes were then chosen to characterize steady-state kinetics and reaction thermodynamics using the microfluidic approach (**Fig. 3d**). 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.[24] 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 for identifying similarity in preferred substrates or selectivity of enzymes; both plotted as a double dendrogram heatmap (**Fig. 3c**). Our analysis clustered the substrates into three main groups. The first group (yellow in **Fig. 3c**) comprises frequently converted substrates, mostly iodinated compounds with a chain length of 3-4 carbon atoms. The second group (green in **Fig. 3c**) includes moderately and poorly convertible (mainly chlorinated) substrates. The third group (brown in **Fig. 3c**) contains only three structurally similar substrates preferred over other tested substrates by most enzymes. Clustering of the specificity profiles divided analyzed HLDs variants into two major groups. The first group (purple in **Fig. 3c**) 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. 3c**) is almost entirely composed of newly identified enzymes (except DatA), which preferably convert the more frequently converted substrates (the first and the third group of substrates) over the second group of substrates. 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. 3c**) 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. Steady-state kinetics and reaction thermodynamics parameters were determined for selected highly active enzymes, DspoA, DexA, DeaA, DprxA, DphxA, and DhxA, using a combination of droplet-based microfluidics and global numerical analysis of kinetic data (**Fig. 4a**). Complex data, including concentration and temperature dependence of the reaction, were collected by monitoring the conversion progress starting at 6 different substrate concentrations (0-1 mM 1,3-dibromopropane) each at 6 different temperatures from 25 to 50 °C in 5-degree increments (**Fig. 4b**). The global numerical fitting of such a complex dataset provided unique estimates for the specificity constant (k_{cat}/K_m), turnover number (k_{cat}), the equilibrium constant for enzyme-product complex dissociation (K_P), and the corresponding energy barriers (**Fig. 4c, Fig. S11, Table S13**). There were two reasons to follow a new methodology in fitting steady-state data and calculating directly k_{cat}/K_m instead of K_m. First, unlike K_m, which has no mechanistic meaning, k_{cat}/K_m is interpreted as the apparent second-order rate constant for substrate binding and quantifies enzyme specificity, efficiency, and proficiency.[25] Second, 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 **Supporting Information**).

All six selected enzymes showed higher values of k_{cat} in comparison with the hitherto known enzyme from the HLDs family, including engineered variants with typical k_{cat} below 10 s^{-1}.[17] The highest known turnover number for a dehalogenase, k_{cat} of 56 s^{-1}, was determined for LinB86 in the conversion of 1,2-dibromoethane. This multiple-point mutant with an introduced *de novo* access tunnel was obtained by several cycles of computer modelling and rational engineering.[16] Despite the high k_{cat}, it still showed a relatively low specific constant (k_{cat}/K_m = 24 $mM^{-1}.s^{-1}$). Interestingly, LinB wild type in the reaction with 1-chlorohexane exhibited a high specific constant (k_{cat}/K_m = 160 $mM^{-1}.s^{-1}$), but in this case, associated with lower k_{cat} = 2.6 s^{-1}. Significantly, the new biocatalysts identified in this study exceeded the currently available enzymes concerning high values in turnover numbers ranging from 13 to 80 s^{-1} (**Fig. 4c**, **upper left**) without the need for laborious engineering procedures and, in addition, these new enzymes showed an advantageous combination of high values of both, turnover numbers and specificity constants (**Fig. 4c**), making them superior to any known haloalkane dehalogenase.

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 combination of both contributions, entropy, and enthalpy, in the case of DprxA and DhxA. Interestingly, DspoA, DexA, and DphxA showed a favourable entropic contribution lowering the activation energy of the catalytic turnover (**Fig. 4c**). The temperature dependences of k_{cat}/K_m indicated that the efficiency of substrate binding is similarly influenced predominantly by enthalpy (DeaA, DprxA, DhxA) or a combination of positive enthalpy and unfavourable loss of entropy (DspoA, DhxA). The other two interesting cases are DexA, with its specificity constant dominated by unfavourable entropy, and DeaA, with a favourable positive entropy

compensating activation enthalpy and reducing the overall free energy of activation (**Fig. 4c**). The mechanistic information derived from the differences in the thermodynamic profiles provides an excellent starting point for rational design[26] and further analysis using machine learning.[27]

IV. Additional Biochemical Characteristics

Enantioselectivity. Enantioselectivity was assessed by determining the kinetic resolution of 2-bromopentane and ethyl 2-bromopropionate representing two distinct groups of racemic 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. S13**). 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.[28,29] High enantioselectivity (E-value > 200) towards the second representative substrate, ethyl 2-bromopropionate, was observed in the case of DprxA, DthA, and DhxA (**Fig. S14**).

Secondary and Quaternary Structure. Additionally, we analyzed the secondary and quaternary structure using far-UV-CD spectroscopy and size-exclusion chromatography, respectively. 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. S15**). Newly identified HLDs were mostly monomeric, similar to the previously characterized HLD members (**Table S15**). Exceptions were DmmarA, which exists as a dimer, and DprxA, a mixture of monomer, dimer, and higher oligomeric states, respectively (**Fig. S16**). Interestingly, DstA showed sensitivity to the environment's oxidation/reduction potential and formed dimers only under oxidative conditions.

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.[2] 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 rational genome mining can provide a potential solution to managing a large quantity of complex sequence data effectively.[30] 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, followed by miniaturized high-throughput characterization, appears to be an attractive approach. This study has integrated computational genome mining with high-throughput microfluidic and microanalytical techniques to

identify novel variants of the model enzyme family, haloalkane dehalogenases. Our results show that only 63 % of the identified putative HLDs were labelled 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.[31] Substrates are not currently known for 35 % of enzymes annotated as α/β-hydrolases, and thus their functions remain unclear.[32] The current mining approach identified more than 2,578 putative HLDs, nearly five times more hits than in the previous *in silico* screening (530 putative HLDs.[8] The current screening approach 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,[33–35] there is no report providing automated protocols for rational selection of a limited number of promising hits for further characterization. In this project, several sequence-based analysis tools have been developed to predict key protein characteristics, e.g., thermostability,[36] optimal pH[33], or protein solubility.[37–40] Other computational tools help automatically analyze, filter and visualize large sets of identified hits.[19,20] The *in silico* part of sequence mining analysis presented in this study is available as a user-friendly web tool EnzymeMiner (https://loschmidt.chemi.muni.cz/enzymeminer/), making the analysis widely accessible to the scientific and industrial communities.[18] In addition to the prediction of tertiary structures that can be achieved using the recently released AlphaFold 2,[41] analysis of cavities and access tunnels, modelling of enzyme-substrate complexes will be implemented in a future version of the presented web tool. The major limitation of *in silico* analysis is the prediction of protein solubility. Despite applying the recent solubility prediction tool SoluProt,[37], our comprehensive expression analysis of the whole set of 45 selected putative HLDs revealed only a 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.[8] Protein production in *E. coli* can be improved by optimizing genetic constructs or expression conditions. However, related 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 bottlenecktoward the functional and structural characterization of novel proteins. Improvement of the *in silico* solubility prediction tool is paramount for the increased success rate of protein characterization pipelines.[14,27]

An essential component of the experimental workflow is the application of time- and material-efficient microscale and microfluidic methods. These miniaturized techniques can be parallelized, allowing high throughput with low demands on the amount of biological material.[42] The droplet microfluidics system described in this study provides a high-throughput platform for activity screening, temperature profiling, and recording kinetic data of different HLD variants in a miniaturized fashion.

Since each droplet is isolated and independent from the other droplets, large numbers of discrete experiments or assays can be performed on short timescales. The current droplet screening setup allows for the measurement of enzymatic activity of 24 HLDs against 27 substrates in a few days representing a marked improvement from state-of-the-art systems. Compared with plate-based screening approaches, our system provides enhanced levels of material efficiency, 1,000-fold lower protein consumption, and throughputs up to 20,000 reactions per day.

The temperature-controlled droplet-based platform described within this study was used to measure temperature-dependent kinetics and extract energetic and entropic contributions. This combination of kinetic and thermodynamic analysis identified variants that are superior to currently available enzymes concerning their catalytic efficiency and exhibiting notable variations in thermodynamic parameters, which essentially drive their catalytic force. Enzymes possessing a differential mix of enthalpy and entropy contributions to the catalytic activity provide unique starting points for laboratory evolution, targeting active sites,[43,44] access tunnels[45], or dynamical protein loops.[46] Such valuable mechanistic information is rarely collected for multiple catalysts during protein discovery campaigns due to the time-consuming nature of these experiments and complex data analysis.

This study significantly enriched the toolbox of model HLDs biocatalysts available for various biotechnological applications.[17] Homology modelling followed by the calculation of the active site volumes is a powerful approach for identifying enzymes with high catalytic activities: 11 of 24 characterized HLDs showed activities higher than the reported previously.[24] Moreover, the high selectivity described for two enzymes, DstA and DthA, where these prefer only one or two molecules from a wide set of representative substrates, has not been described so far. Several novel enzymes showed high enantioselectivity, including (*S*)-enantiopreference unusual for this family of enzymes. Temperature profiling experiments identified DmaA, which retained > 65 % of its maximal activity at 5 °C, an attractive property for applying HLDs as environmental biosensors.[47] It is worth noting that some newly discovered enzymes combine several industrially attractive properties, such as high activity, stability, and selectivity.

CONCLUSIONS

In this study, we demonstrated how a combination of automated *in silico* screening and high throughput miniaturized experimental techniques can be used to obtain highly efficient catalysts from the large unexplored diversity of an enzyme family hidden in genome databases. A set of variants with catalytic efficiencies an order of magnitude higher compared to currently available family members, unique substrate specificity, and selectivity in combination with a wide range of operational temperature was obtained without the need for demanding engineering or high-throughput experimental screening. The hereby-used *in silico* pipeline employs automated sequence similarity search accompanied by

annotation and structural bioinformatics analyses. The experimental characterization of new variants (tens to hundreds) was accelerated by applying high-throughput microfluidic and microanalytical methods. The droplet-based microfluidics enabled minimized consumption of protein samples and high throughput acquisition of substrate specificity, activity, and steady-state kinetics data at various temperatures. The latter provided valuable mechanistic information and the extraction of thermodynamic parameters, entropy, and enthalpy, contributing to the catalysis. Accordingly, developing an automated droplet-based microfluidic device will open up new opportunities for future optimal data collection employing back-loops and machine learning algorithms. Another interesting perspective appears to be the application of cell-free methods for protein production and its integration into microfluidic systems, which will simplify the experimental workflow. Contrary to expectations, cell-free protein production did not increase the yield of soluble variants in this case. Therefore, developing more reliable prediction tools to refine the selection of protein candidates with good solubility remains a significant challenge. We also demonstrated that repetitive database mining generates a variety of novel enzymes with valuable industrial properties and enables the consistent collection of experimental data. High-quality datasets are attractive for 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 the 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[48] 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.[49] 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 neighbourhood 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.[50] The quality of the generated homology models was assessed by MolProbity v4.3.1.[22] Pockets in each homology model were calculated and measured using the CASTp program[51] with a probe radius of 1.4 Å. The CAVER v3.02 program[52] 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.[53]

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 pET-24a(+) between the NdeI and XhoI restriction sites. For plasmid propagation, competent *E. coli* DH5α cells were transformed using a heat-shock method with individual constructs. 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 pET-24a(+)::*HLD*x (x = 45 different HLDs candidates) plasmid DNA using a heat shock method, plated on 2x lysogeny broth agar plates with kanamycin (50 $\mu g.mL^{-1}$) 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 $\mu g.mL^{-1}$ 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 $\mu g.mL^{-1}$ 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 adding 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). The 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). According to the manufacturer's manual, the rest of the sample was subjected to high-throughput affinity purification using the MagneHis Protein Purification System (Promega, USA), 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 (3 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 50 mM PB was added to each well between centrifugation steps. The enzymes were dissolved in 100 µL of 50 mM PB.

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, 10 µL of whole cells with OD600 approximately 5, or with 4 µg purified protein, 12.5 µM aminophenyl fluorescein and 10 mM of a halogenated substrate. The reactions in HOX assay[23] were started by adding whole cells or purified protein. 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.[54] 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 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 the 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 equilibration buffer (purification buffer A). Target proteins were eluted with an increasing two-step gradient. First, unbound and weakly bound proteins were washed out with 10 % of purification buffer B (20 mM K_2HPO_4/ KH_2PO_4, pH 7.5, 500 mM NaCl, 300 mM imidazole). Subsequently, the target proteins were eluted with a 60 % purification buffer B. Enzymes eluted by 180 mM imidazole 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 PB buffer (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 protein concentration was determined by measuring absorbance at 280 nm using a DS-11 series Spectrophotometer (De Novix, USA) with the extinction coefficients calculated using the ProtParam tool.[55]

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 PB, 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 assays were 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 PB buffer, pH 7.5), and 50 mM PB 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 PB 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. The inflexion point of the resulting curve was used as a thermostability parameter.

Substrate Specificity Profiles and Temperature Profiles. Both substrate specificity and temperature profiles were measured using the capillary-based droplet microfluidic device, 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, previously chosen to validate the microfluidic device. Each enzyme was assayed at the temperature closest to its T_{max} value (0-10 °C below T_{max}). A detailed description of the microfluidic method was provided previously.[56] Briefly, the droplets were generated using Mitos Dropix (Dolomite, UK). A custom sequence of droplets (150 nL aqueous phase, 300 nL spacing oil) was generated using negative pressure generated by a syringe pump. 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 (3M, USA) + 0.5 % PicoSurf 1 (Sphere Fluidics, UK)) and the aqueous phase. The reaction solution consisted of a weak buffer (1 mM HEPES, 20 mM Na_2SO_4, pH 8.2) and fluorescent indicator 8-hydroxypyrene-1,3,6-trisulfonic acid (HPTS, 50 μM). The composition of buffer was optimized to maintain both droplet stability and provide a weak buffering system for the sensitivity of pH assay. 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 an excitation laser (450 nm), a dichroic mirror with a cut-off at 490 nm filtering the excitation light, and a Si-detector (Thorlabs, Germany). 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 10 droplets/sample passed 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. The peaks were assigned to the particular sample using this software, and the mean signal was calculated. 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.[23,56] The activities were classified as "not determined" whenever the measured product concentration was below the limit of detection (LOD – 3 times noise signal). Each substrate had a different calibration curve, so the LOD product concentration was in the range of 10-100 μM).

Principal Component Analysis and Hierarchical Clustering. The matrix containing the activity data of 24 newly-identified 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 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 to identify substrate specificity groups. Additionally, the hierarchical clustering analysis was performed on the log-transformed data using MATLAB (MathWorks, USA).

Steady-state Kinetics and Reaction Thermodynamics. The steady-state kinetics and reaction thermodynamics parameters for a selected set of enzyme variants were determined using a capillary-based droplet microfluidic device. The continuous phase was fluorous oil FC-40 (Dolomite Microfluidics, Blacktrace Holdings Ltd, UK). When performing an enzymatic reaction, 1,3-dibromopropane was dissolved in an additional sample of FC-40. The pH-based fluorescence assay to determine HLD kinetics was the same as described above for the determination of substrate specificities and temperature profiles. Within one run, the steady-state kinetics of a single enzyme variant was measured with 1,3-dibromopropane in the temperatures range of 25-50 °C in 5-degree increments. For each temperature, the HLD enzymatic rate was determined for six substrate concentrations.

Low-pressure Nemesys microfluidic pumps (Cetoni, Germany) were used for precise flow control operating with 10-mL and 25-mL Luer-lock Gas Tight ® syringes (Hamilton, USA) in the positive pressure mode. These high-precision syringe pumps equipped with glass syringes motivate all the solutions through high purity perfluoralkoxy (PFA) tubing (OD 1/16 inch, ID 0.02 inch, IDEX Health & Science, USA) towards a mixing junction. After mixing, the oil and aqueous phases were connected via fluidic connectors (F-127; Upchurch Scientific, Germany) into a Y-junction (P- 512, IDEX Health & Science, USA) for the formation of droplets. The ratios of solutions and velocity of droplet movement were controlled by the relative flow rates of the individual pumps driving the syringes. The buffer of the enzyme samples was exchanged the same way as for the microfluidic activity characterization. The enzyme was then diluted into a reaction buffer with a similar enzyme concentration as during substrate specificity profiling. To introduce substrate into the system, a second Y-Assembly PEEK connector was joined upstream to allow the confluence of two streams of FC-40, with and without substrate, before droplet generation. In this manner, the final substrate concentration in the microfluidic system was directly controlled by adjusting the flow rate ratio of the two oil streams.

The formed droplets containing the reaction mixture were subsequently directed through a tubing coiled around a copper-heating rod (diameter 1.5 cm, length 8 cm) which allows moving along the longer axis and heating up to 180 °C. The heating rod was then placed on top of a motorized rotation stage (CR1/M-Z7E, Thorlabs, Inc., USA), simultaneously mounted on a motorized linear translation stage (MTS25/M-28E, Thorlabs, Inc., USA). The axial and rotational movement of the rod was controlled using Kinesis ® software (Thorlabs, Inc., USA). By moving the heating rod to different positions, fluorescence measurements can be performed at various points along with the tubing, each corresponding to a unique reaction time. The temperature of the copper rod is controlled using a heating cartridge (6.5 x 40 mm, 100 W, Farnell, Switzerland) embedded inside the heating rod. The temperature is monitored using a thermocouple (Sensor, Thermoelement Type K - 0.5 mm, Farnell, Switzerland), inserted into the copper block close to the surface. Temperature control is realized using a PID controller (CN7800, Omega, USA), with an observed temperature variation from the set point of 0.1 °C. The detection of fluorescence emission was performed by an optical system composed of a light-emitting diode (M470L3 Thorlabs) and a QE Pro spectrometer (OceanOptics, USA).

The pH-based activity assay was calibrated using three buffer solutions containing the enzyme, pH of the individual solution was adjusted by a defined concentration of hydrochloric acid (HCl). The pH of all aqueous solutions was checked before loading into syringes by pH-electrode. The linear dependence of pH on increasing product concentration (HCl) was determined for individual enzymes outside the microfluidic device via titration of HCl into the enzyme solution. The substrate was delivered into the reaction droplet through oil-buffer partitioning, similarly as by microfluidic activity characterization[23]. The oil buffer partitioning is quantified by a partition coefficient calculated as the logarithm of the compound concentration ratio in fluorinated oil and water (***Equation 1***):

$$\mathrm{Log}P = \log \frac{[\mathrm{compound_{oil}}]}{[\mathrm{compound_{buffer}}]}$$

Equation 1

The dependence of partition coefficients on temperature was determined in the temperature range from 20 to 60 °C in 10-degree increments (**Fig. S12**). The partition coefficients of 1,3-dibromopropane were analyzed by monitoring the substrate concentration in a two-phase system, including reaction buffer excluding HPTS and FC-40 oil. Initially, 1 µL of the halogenated substrate was dissolved in 1 mL of the oil and continuously shaken for 30 min at room temperature. Then 1 mL of reaction buffer excluding HPTS was added to create the two-phase system similar to droplet-based microfluidics. After 30 min incubation in a water bath with continuous shaking at the target temperature, 30 µL samples of both the oil and the aqueous phase were extracted with 300 µL acetone containing 4.7 mM 1,2,3-trichloropropane as internal standard. In both phases, the concentration of a particular substrate was quantified using an

Agilent 5975C mass spectrometer coupled to an Agilent 7890A (Agilent Technologies, USA) gas chromatographer equipped with a ZB-5 capillary column (30 m x 0.25 mm x 0.25 μm, Phenomenex, USA). The samples were taken by Automatic Liquid Sampler, and 1 μL was injected into the Split-Splitless inlet at 250 °C, with a split ratio of 1:20. The temperature program was isothermal at 50 °C for 1 min, followed by an increase to 190 °C at 25 °C/min. The flow of carrier gas (He) was 2.2 $ml.min^{-1}$. The concentration of 1,3-dibromopropane in reaction oil was determined using GC-MS similarly to the above-described measurement of partition coefficients. The extraction was performed using 50 μL of oil with substrate extracted in 250 or 500 μL of acetone with internal standard.

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 version 10 (KinTek Corporation, USA),[57] which includes the capability to fit temperature-dependent rate constants.[58] The software allows for the input of a given kinetic model via a simple text description, and the program then automatically derives the differential equations needed for numerical integration. An updated form of the steady-state model applying new standards for fitting kinetic data (***Scheme 1***) was used to obtain the values of turnover number (k_{cat}) and specificity constant k_{cat}/K_m directly. By setting $k_{-1} = 0$, after substrate binds, it is always converted to the product, so k_{cat}/K_m is defined by the value of k_1.[25] By setting $k_{-2} = 0$, the value of k_{cat} is defined by k_2. The Michaelis constant was derived from the two primary steady-state kinetic parameters ($K_m = k_{cat} / k_{cat}/K_m$). Numerical analysis of full progress curve kinetics provided accurate parameter estimates, including product inhibition ($K_P = k_3/k_{-3}$, when k_{-3} was varied as a fitted parameter and $k_3 = 1{,}000\ s^{-1}$ was used as a fixed value to satisfy $k_3 >> k_2$ assumption).

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} EP \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} E + P$$

$$k_{cat} = k_2 \ (k_{-2} = 0)$$

$$k_{cat}/K_m = k_1 \ (k_{-1} = 0)$$

Scheme 1

Numerical integration of rate equations searching a set of kinetic parameters that produce a minimum χ^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, the halogenated substrate concentrations were allowed to be slightly adjusted (± 5 %) to derive the 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. In addition to S.E. values, a more rigorous analysis of the variation of the kinetic parameters was accomplished by confidence contour analysis using FitSpace Explorer (KinTek Corporation, USA).[59] In these analyses, the lower

and upper limits for each parameter were derived from the confidence contour by setting the χ^2 threshold at 0.95. The standard error estimates in the fitted parameters were propagated to obtain estimates of the error in the calculated value of K_m. The global kinetic model was used to analyze the temperature dependence of obtained kinetic parameters using the Eyring equation, $\ln(k_{cat}/T) = -\Delta H^{\ddagger}/(R.T) + \ln(k_B/h) + \Delta S^{\ddagger}/R$, to obtain estimates for the enthalpy ($\Delta H^{\ddagger}$) and entropy of activation ($\Delta S^{\ddagger}$), where R is the universal gas constant, k_B is the Boltzmann constant and h is Planck's constant. The Gibbs free energy ($\Delta G^{\ddagger}$) was defined as $\Delta G^{\ddagger} = \Delta H^{\ddagger} - T.\Delta S^{\ddagger}$ at the reference temperature 310.15 K.

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). Applying new standards for collecting and fitting steady-state kinetic data,[25] the alternative form of competitive steady-state model (***Scheme 2***) was used to obtain direct estimates of specificity constants k_{cat}/K_m 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 2***), 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 2

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

Equation 2

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.[57]

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 corrected for the buffer´s 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 to 250 nm using the BeStSel online tool with default settings.[60]

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). 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).

ACKNOWLEDGEMENTS

We would like to thank Simon Godehard and Mark Dörr (Greifswald University, Germany) for fruitful discussions of the experimental design of cell-free expression experiments. Michal Vasina acknowledges the financial support of his doctoral study by the scholarship Brno Ph.D. Talent and IGA MU. The 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 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.

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. Fernández-Arrojo, L., Guazzaroni, M.-E., López-Cortés, N., Beloqui, A. & Ferrer, M. Metagenomic era for biocatalyst identification. *Curr. Opin. Biotechnol.* **21**, 725–733 (2010).

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

3. Copp, J. N., Akiva, E., Babbitt, P. C. & Tokuriki, N. Revealing Unexplored Sequence-Function Space Using Sequence Similarity Networks. *Biochemistry* **57**, 4651–4662 (2018).

4. Furnham, N., Garavelli, J. S., Apweiler, R. & Thornton, J. M. Missing in action: enzyme functional annotations in biological databases. *Nat. Chem. Biol.* **5**, 521–525 (2009).

5. Mak, W. S. *et al.* Integrative genomic mining for enzyme function to enable engineering of a non-natural biosynthetic pathway. *Nat. Commun.* **6**, 10005 (2015).

6. Marshall, J. R. *et al.* Screening and characterization of a diverse panel of metagenomic imine reductases for biocatalytic reductive amination. *Nat. Chem.* 1–9 (2020) doi:10.1038/s41557-020-00606-w.

7. Lobb, B. & Doxey, A. C. Novel function discovery through sequence and structural data mining. *Curr. Opin. Struct. Biol.* **38**, 53–61 (2016).

8. Vanacek, P. *et al.* Exploration of Enzyme Diversity by Integrating Bioinformatics with Expression Analysis and Biochemical Characterization. *ACS Catal.* **8**, 2402–2412 (2018).

9. Zhou, N. *et al.* The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome Biol.* **20**, 244 (2019).

10. Li, Y. *et al.* DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics* **34**, 760–769 (2018).

11. Colin, P.-Y. *et al.* Ultrahigh-throughput discovery of promiscuous enzymes by picodroplet functional metagenomics. *Nat. Commun.* **6**, 10008 (2015).

12. Beneyton, T. *et al.* Droplet-based microfluidic high-throughput screening of heterologous enzymes secreted by the yeast Yarrowia lipolytica. *Microb. Cell Factories* **16**, 18 (2017).

13. Kokkonen, P. *et al.* Structure-Function Relationships and Engineering of Haloalkane Dehalogenases. in *Aerobic Utilization of Hydrocarbons, Oils and Lipids* (ed. Rojo, F.) 1–21 (Springer International Publishing, 2017). doi: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.* **9**, 1033–1054 (2019).

15. Beerens, K. *et al.* Evolutionary Analysis As a Powerful Complement to Energy Calculations for Protein Stabilization. *ACS Catal.* **8**, 9420–9428 (2018).

16. Brezovsky, J. *et al.* Engineering a de Novo Transport Tunnel. *ACS Catal.* **6**, 7597–7610 (2016).

17. Koudelakova, T. *et al.* Haloalkane dehalogenases: Biotechnological applications. *Biotechnol. J.* **8**, 32–45 (2013).

18. Hon, J. *et al.* EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. *Nucleic Acids Res.* **48**, W104–W109 (2020).

19. Gerlt, J. A. *et al.* 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 (2015).

20. Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **13**, 2498–2504 (2003).

21. Barrett, T. *et al.* BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. *Nucleic Acids Res.* **40**, D57–D63 (2012).

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

23. Buryska, T. *et al.* Controlled Oil/Water Partitioning of Hydrophobic Substrates Extending the Bioanalytical Applications of Droplet-Based Microfluidics. *Anal. Chem.* **91**, 10008–10015 (2019).

24. Koudelakova, T. *et al.* Substrate specificity of haloalkane dehalogenases. *Biochem. J.* **435**, 345 LP – 354 (2011).

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

26. Planas-Iglesias, J. *et al.* Computational design of enzymes for biotechnological applications. *Biotechnol. Adv.* **47**, 107696 (2021).

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

28. Chmelova, K. *et al.* 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**, (2020).

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

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

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

32. 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.* **5**, 6153–6176 (2015).

33. Foroozandeh Shahraki, M. *et al.* MCIC: Automated Identification of Cellulases From Metagenomic Data and Characterization Based on Temperature and pH Dependence. *Front. Microbiol.* **11**, 567863 (2020).

34. Cai, X. *et al.* Combination of sequence-based and in silico screening to identify novel trehalose synthases. *Enzyme Microb. Technol.* **115**, 62–72 (2018).

35. Barriuso, J. & Martínez, M. J. In silico metagenomes mining to discover novel esterases with industrial application by sequential search strategies. *J. Microbiol. Biotechnol.* **25**, 732–737 (2015).

36. Mahmoudi, M., Arab, S. S., Zahiri, J. & Parandian, Y. An Overview of the Protein Thermostability Prediction: Databases and Tools. *J. Nanomedicine Res.* **3**, 00072 (2016).

37. Hon, J. *et al.* SoluProt: prediction of soluble protein expression in Escherichia coli. *Bioinformatics* **37**, 23–28 (2021).

38. Khurana, S. *et al.* DeepSol: a deep learning framework for sequence-based protein solubility prediction. *Bioinformatics* **34**, 2605–2613 (2018).

39. Raimondi, D., Orlando, G., Fariselli, P. & Moreau, Y. Insight into the protein solubility driving forces with neural attention. *PLOS Comput. Biol.* **16**, e1007722 (2020).

40. Bhandari, B. K., Gardner, P. P. & Lim, C. S. Solubility-Weighted Index: fast and accurate prediction of protein solubility. *Bioinformatics* **36**, 4691–4698 (2020).

41. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).

42. Bunzel, H. A., Garrabou, X., Pott, M. & Hilvert, D. Speeding up enzyme discovery and engineering with ultrahigh-throughput methods. *Curr. Opin. Struct. Biol.* **48**, 149–156 (2018).

43. Sykora, J. *et al.* Dynamics and hydration explain failed functional transformation in dehalogenase design. *Nat. Chem. Biol.* **10**, 428–430 (2014).

44. Liskova, V. *et al.* 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 (2017).

45. Pavlova, M. *et al.* Redesigning dehalogenase access tunnels as a strategy for degrading an anthropogenic substrate. *Nat. Chem. Biol.* **5**, 727–733 (2009).

46. Schenkmayerova, A. *et al.* Engineering the protein dynamics of an ancestral luciferase. *Nat. Commun.* **12**, 3616 (2021).

47. Bidmanova, S. *et al.* Fluorescence-based biosensor for monitoring of environmental pollutants: From concept to field application. *Biosens. Bioelectron.* **84**, 97–105 (2016).

48. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

49. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).

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

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

52. Pavelka, A. *et al.* CAVER: Algorithms for Analyzing Dynamics of Tunnels in Macromolecules. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **13**, 505–517 (2016).

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

54. Shimizu, Y. *et al.* Cell-free translation reconstituted with purified components. *Nat. Biotechnol.* **19**, 751–755 (2001).

55. Gasteiger, E. *et al.* Protein Identification and Analysis Tools on the ExPASy Server. in *The Proteomics Protocols Handbook* (ed. Walker, J. M.) 571–607 (Humana Press, 2005). doi:10.1385/1-59259-890-0:571.

56. 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* (ed. Tawfik, D. S.) vol. 643 51–85 (Academic Press, 2020).

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

58. Li, A., Ziehr, J. L. & Johnson, K. A. 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 (2017).

59. Johnson, K. A., Simpson, Z. B. & Blom, T. FitSpace Explorer: An algorithm to evaluate multidimensional parameter space in fitting kinetic data. *Anal. Biochem.* **387**, 30–41 (2009).

60. Micsonai, A. *et al.* BeStSel: a web server for accurate protein secondary structure prediction and fold recognition from the circular dichroism spectra. *Nucleic Acids Res.* **46**, W315–W322 (2018).

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 (°C)	T_{max} (°C)	E value	
						2-bromopentane	ethyl 2-bromopropionate
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 temperature close to the optimal temperature (**Table S8**); T_{onset} – unfolding onset temperature; T_m^{app} – apparent melting temperature by nanoDSF; T_{max} – maximum HLD activity; n.a. – no activity

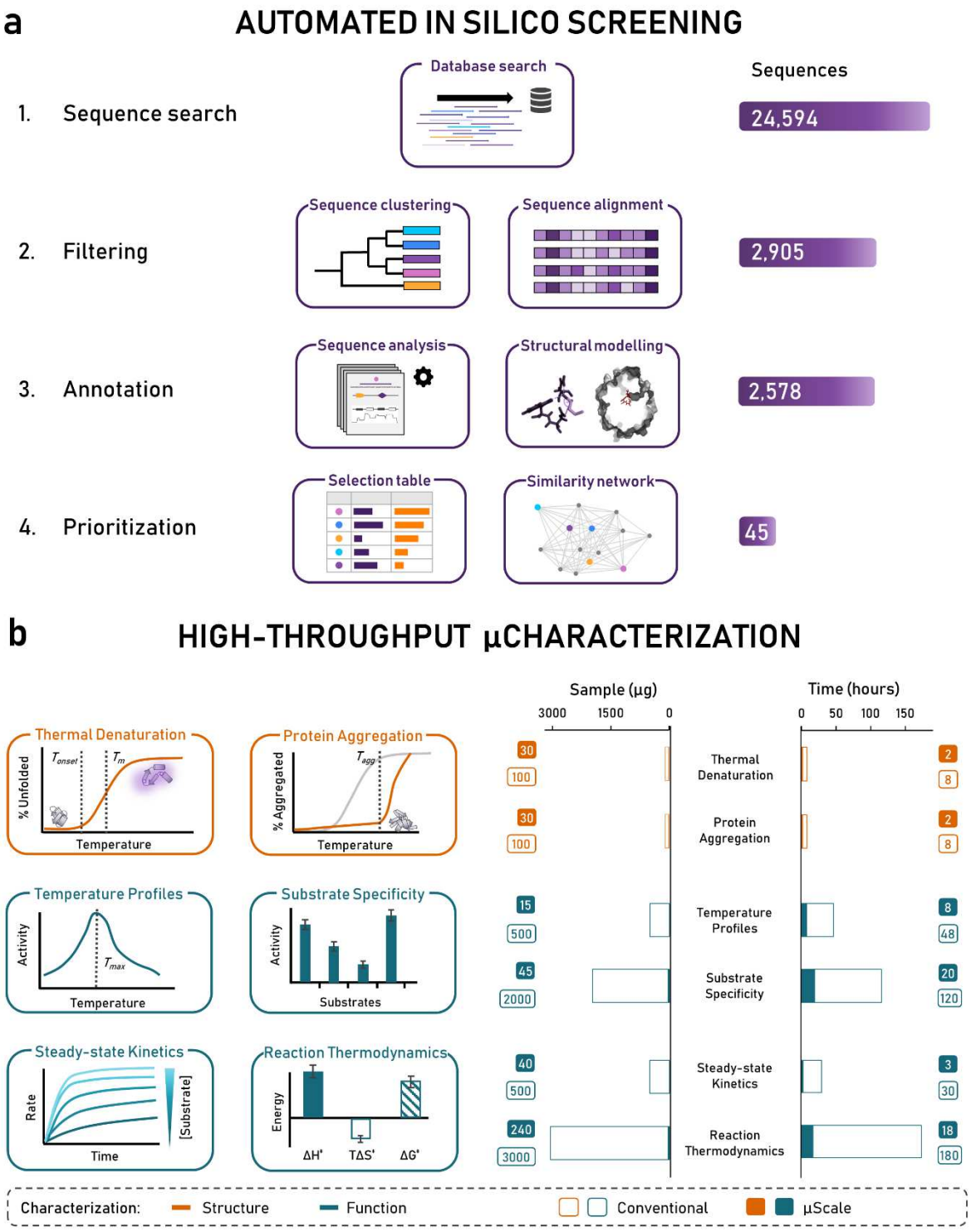


Figure 1. Integrated theoretical-experimental strategy for high-throughput exploration of unmapped sequence space. a, Workflow of the automated i*n silico* screening. The bars on the right show the number of protein sequences resulting from each step of the workflow. **b**, Overview of the functional and mechanistic characterization of well-expressed enzymes applying microscale and microfluidics techniques. The bar graph on the right compares sample size and time requirements per enzyme using conventional (empty) and microscale (filled) techniques.

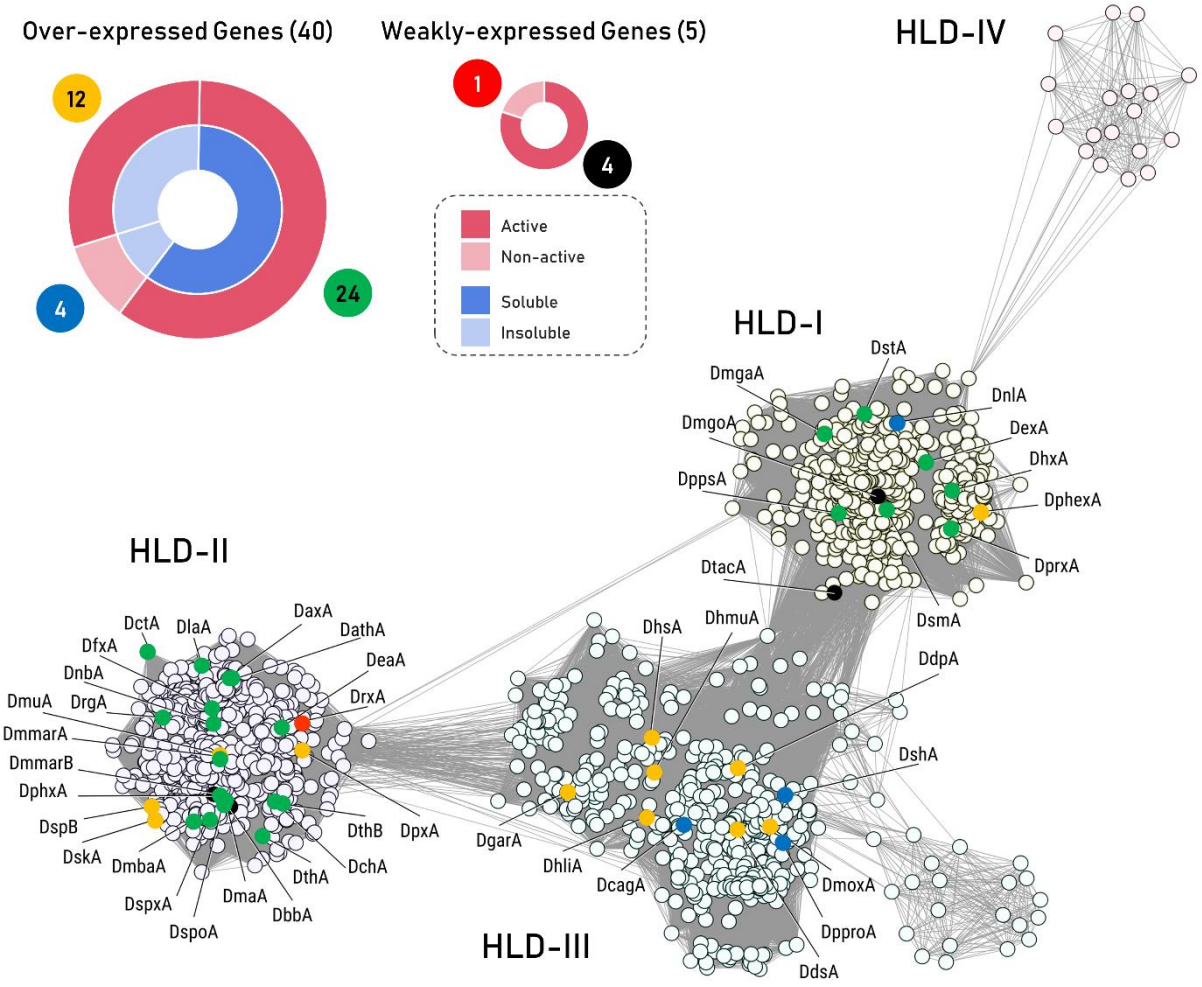


Figure 2. Sequence similarity network for haloalkane dehalogenases categorized by their expression, solubility, and activity. The putative haloalkane dehalogenases (HLDs) are clustered into four subfamilies: HLD-I, HLD-II, HLD-III, and HLD-IV. The sequences were first clustered at 50 % identity to reduce 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[19] and Cytoscape v3.6.1[20]. The results from expression, solubility, and activity analysis is 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 colours in doughnut graphs and the sequence similarity network. 24 well-soluble and active enzymes (green) were 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 S6**). Four over-expressed genes providing insoluble proteins (blue) and one weakly-expressed gene (red) led to proteins that did not exhibit any activity with the substrates tested.

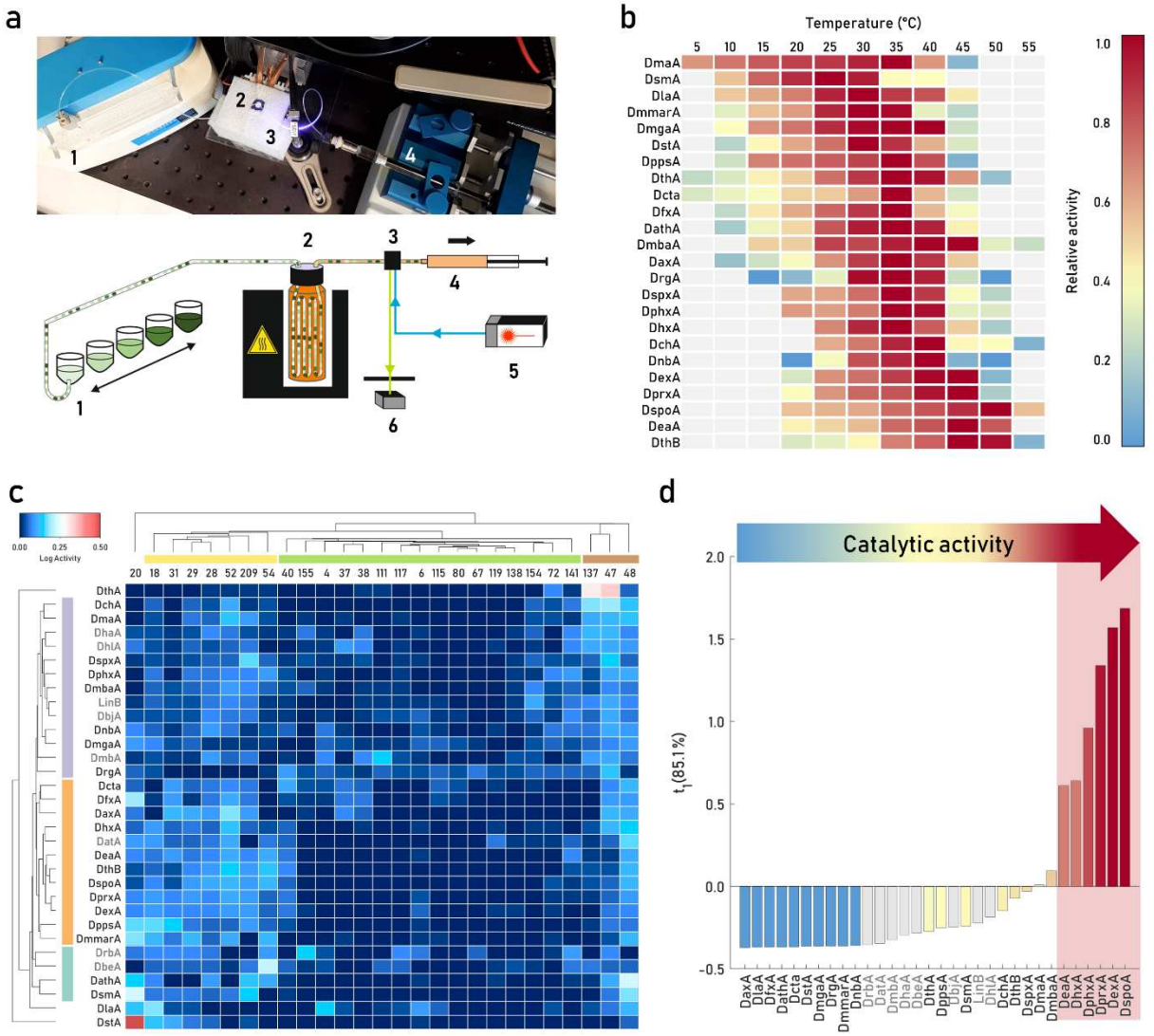


Figure 3. Temperature profiles and substrate specificity by droplet-based microfluidics. **a,** Photograph and scheme of the droplet-based microfluidic device for determination of temperature profiles and substrate specificity. Depicted are the main parts of the device, including the droplet generator (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 colour. **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 data variance. The light red frame highlights new enzymes with an outstanding catalytic activity, which were characterized for their steady-state kinetics and reaction thermodynamics (**Fig. 4**). The previously characterized HLDs are coloured grey in **c, d**. The heat maps (**b, c**) and bars (**d**) are colour-coded by enzymatic activity from low activity (blue) to medium activity (yellow) and high activity (red).

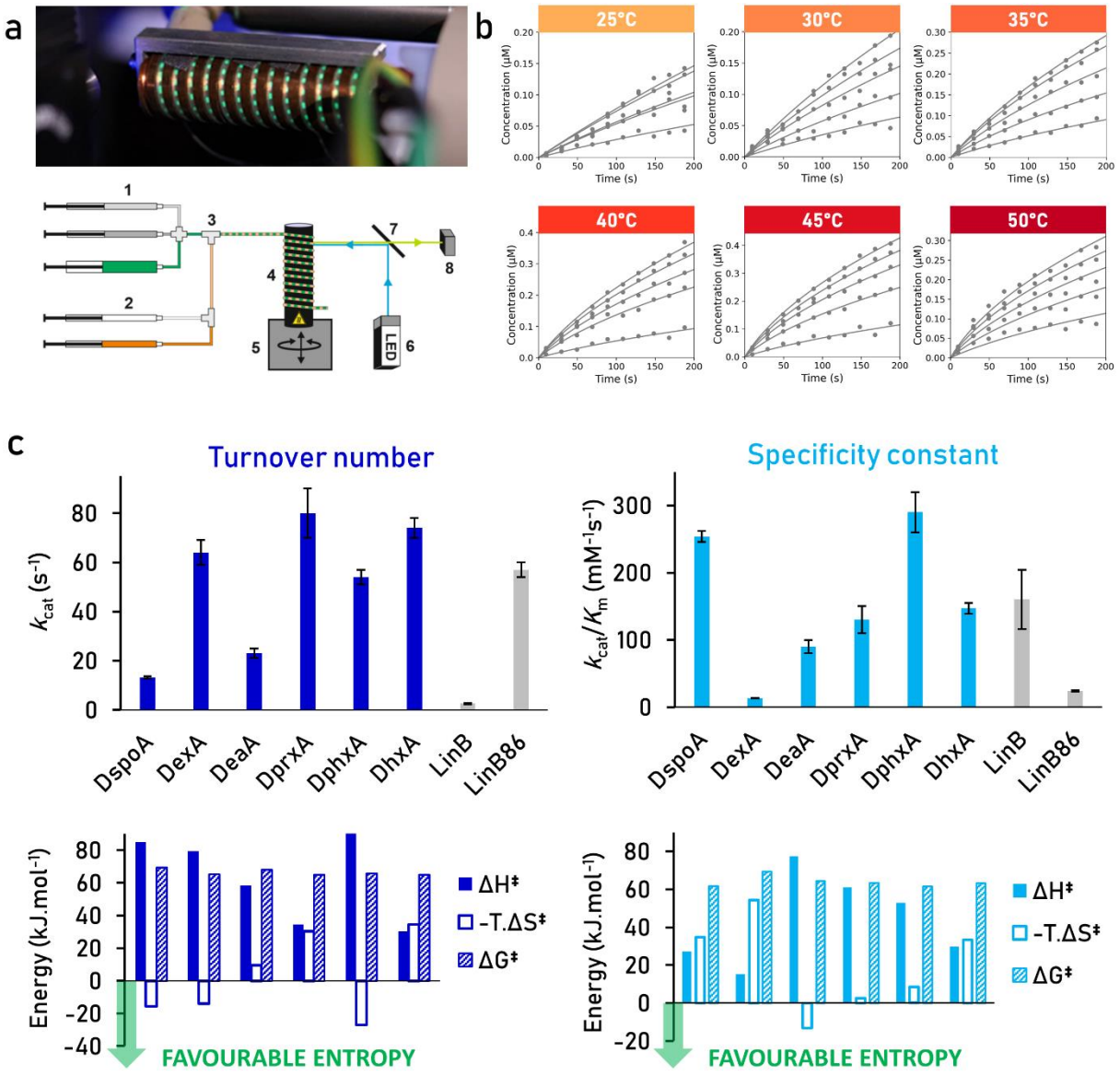


Figure 4. Mechanistic analysis by droplet-based microfluidics and global numerical integration. **a,** The droplet-based microfluidic device enabling kinetic and thermodynamic measurements. A photograph (top) illustrates the reaction droplets travelling through the incubation zone with temperature control. The scheme 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 and 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 S13**. **c**, The kinetic parameters (top figures), turnover number (k_{cat}), and specificity constant (k_{cat}/K_m), obtained by global fitting complex kinetic and thermodynamic data (values at reference temperature 310.15 K, 30 °C). The error bars represent standard errors. Grey columns represent previously reported values for the reaction of wild type LinB with 1-chlorohexane and engineered LinB86 with 1,2-dibromoethane (100 mM glycine buffer pH 8.6 at 37 °C). The

contributions of activation enthalpy ($\Delta H^{\ddagger}$) and entropy ($-T.\Delta S^{\ddagger}$) to the Gibbs free energy of activation ($\Delta G^{\ddagger}$) 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 favourable entropy values lowering the activation barrier.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupportingDataSet.xls
- VasinaNatCatSI.pdf