

Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: K. Ji, S. S. Barnes, C. Ziegler, M. Nikpey, A. de la Paz, E. Pack, N. Kvasovs, F. Morcos and S. Dodani, *Chem. Sci.*, 2026, DOI: 10.1039/D6SC02671K.



This is an Accepted Manuscript, which has been through the Royal Society of Chemistry peer review process and has been accepted for publication.

Accepted Manuscripts are published online shortly after acceptance, before technical editing, formatting and proof reading. Using this free service, authors can make their results available to the community, in citable form, before we publish the edited article. We will replace this Accepted Manuscript with the edited and formatted Advance Article as soon as it is available.

You can find more information about Accepted Manuscripts in the [Information for Authors](#).

Please note that technical editing may introduce minor changes to the text and/or graphics, which may alter content. The journal's standard [Terms & Conditions](#) and the [Ethical guidelines](#) still apply. In no event shall the Royal Society of Chemistry be held responsible for any errors or omissions in this Accepted Manuscript or any consequences arising from the use of any information it contains.

Decoding Cryptic Defluorinases through a Latent Generative Sequence Landscape

View Article Online
DOI: 10.1039/C6CY02671K

Ke Ji^{+a}, Sydney S. Barnes^{+a}, Cheyenne Ziegler^{+b}, Marjan Nikpey^c, Alberto de la Paz^b, Elizabeth K. Pack^a, Nikita Kvasovs^a, Faruck Morcos^{*b,c,d,e} and Sheel C. Dodani^{*a}

^aDepartment of Chemistry and Biochemistry, The University of Texas at Dallas, Richardson, TX 75080

^bDepartment of Biological Sciences, The University of Texas at Dallas, Richardson, TX 75080

^cDepartment of Bioengineering, The University of Texas at Dallas, Richardson, TX 75080

^dDepartment of Physics, The University of Texas at Dallas, Richardson, TX 75080

^eCenter for Systems Biology, The University of Texas at Dallas, Richardson, TX 75080

*sheel.dodani@utdallas.edu; faruckm@utdallas.edu

+These authors contributed equally to this work

Abstract

The special nature of the fluorine atom imparts remarkable strength and unique physical properties to chemical bonds. Unlike man-made fluorochemicals, fluorinated natural products remain rare due to low bioavailability and toxicity of fluoride. Despite this, defluorinases have evolved in nature to cleave carbon-fluorine bonds, with the hydrolytic fluoroacetate dehalogenase being one of the most well-characterized examples. These enzymes are of fundamental interest and hold unrealized biotechnological potential, yet the scope of this unique chemistry remains underexplored in the biosphere. Here, we trained and applied a machine learning-based framework, termed Latent Generative Landscapes (LGLs), to map the functional sequence space of the α/β -hydrolase superfamily. This approach identified 3,014 putative defluorinases that were previously not annotated or plausibly misannotated. Experimental validation of selected candidates led to the reclassification of five novel defluorinases, all exhibiting high thermal stability ($T_m > 70$ °C) and diverse catalytic efficiencies with conserved enantioselectivity on the model substrate 2-fluoro-2-phenylacetate. Notably, the enzyme A0A4Z0BVY8 exhibited 2.7-fold greater defluorination activity than the current state-of-the-art enzyme Q6NAM1. Our results establish that LGL modeling is a powerful strategy to decode cryptic carbon-fluorine bond chemistry in nature, enabling the future discovery and engineering of defluorination biocatalysts.



Introduction

View Article Online
DOI: 10.1039/D6SC02671K

Of the halogens, fluorine is the most abundant in the Earth's crust.^{1–3} However, it is largely hidden from the biological world, sequestered in minerals in the form of fluoride.^{2,4} Only in rare cases is fluorine incorporated into natural products, with fluoroacetate being the simplest known example.^{4–6} While carbon-fluorine bond formation is mediated by a single enzyme class, naturally occurring enzymes that can cleave carbon-fluorine bonds have evolved across multiple classes, accessing a wider range of mechanistic solutions, in some cases through promiscuous activity.^{7–30} Among these, hydrolytic defluorination mediated by fluoroacetate dehalogenases from the α/β -hydrolase superfamily (ABH; Pfam: PF00561) represents a well-characterized strategy.^{31–33} Efforts have primarily focused on understanding these enzymes through computational modeling, mechanistic enzymology, and structural biology, with the fluoroacetate dehalogenase from *Rhodopseudomonas palustris* (UniProt Accession Number: Q6NAM1) serving as the most extensively studied member (Figure 1A).^{14,34–40}

In Q6NAM1, the entry of fluoroacetate is gated by W185, and recognition in the active site relies on carboxylate anchoring residues R111 and R114 (Figure 1B).¹⁴ The defluorination mechanism proceeds via an S_N2 mechanism, in which a nucleophilic attack by D110 displaces fluoride to form a transient covalent intermediate, followed by hydrolysis mediated by D134 and H280 to release hydroxyacetate and a proton (Figure 1C).^{14,35,37–39,41} Fluoride stabilization is facilitated by H155, W156, and Y219, and mutations at these positions can not only tune catalytic efficiency but also shift halide selectivity.^{14,42,43} Q6NAM1 can also transform α -fluoro-carboxylic acids such as 2-fluoro-2-phenylacetate with stereochemical preference (Figure 1C).^{37,43} More recently, Q6NAM1 and other fluoroacetate dehalogenases have attracted attention as biocatalysts for kinetic resolution of α -fluoro-carboxylic acids and degradation of anthropogenic organofluorine compounds, underscoring their biotechnological potential.^{22,25,44,45} At the same time, microbial genome mining and homology-based annotation suggest that hydrolytic defluorination may be more widespread than currently recognized.^{23,46–50}

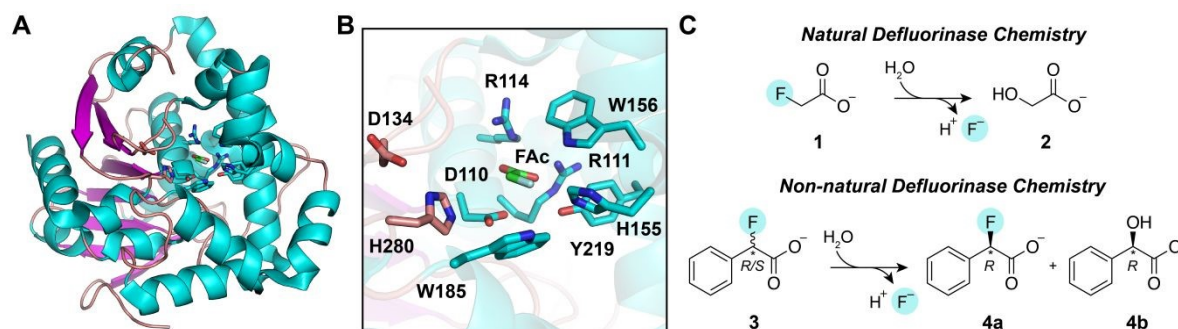


Figure 1. (A) Crystal structure of the fluoroacetate dehalogenase from *Rhodopseudomonas palustris* (UniProt Accession Number: Q6NAM1; PDB ID: 3R3V). For this representation, N110 was reverted to D110 in PyMOL. (B) The enzyme active site with fluoroacetate (FAc) bound. Select residues within 4 Å of FAc are shown. (C) Q6NAM1 can defluorinate (top) the natural product fluoroacetate (compound 1) into hydroxyacetate (compound 2) and (bottom) the non-natural substrate racemic 2-fluoro-2-phenylacetate (compound 3) into (*R*)-2-fluoro-2-phenylacetate (compound 4a) and (*R*)-2-hydroxy-2-phenylacetate (compound 4b).



Despite this growing interest, functional assignment across the ABH superfamily remains challenging because of the large number of members (982,505; Swiss-Prot and TrEMBL 2025 release) that share a high degree of sequence and structural similarity yet span diverse and even promiscuous chemistries.^{51,52} As a result, scalable and accurate annotation methods are needed to systematically map sequence space to defluorination activity. Recent advances in computational and machine learning approaches provide new opportunities to address this gap by integrating large-scale sequence information with experimental datasets to build predictive models.^{53–57}

Latent Generative Landscapes (LGLs) provide one such approach as they are maps created from protein sequence data that illustrate how sequence diversity connects to function.⁵⁸ By integrating Variational Autoencoders (VAE) alongside Direct Coupling Analysis (DCA), LGLs learn not only which individual residues are key for functional and phylogeny classification but also how important interactions between residues are indicative of features within the family that define functional and fitness relationships.^{58–62} This helps researchers investigate protein diversity, predict functional integrity, and create new variants while maintaining the networks crucial for protein activity. LGLs have been used to study thermal adaptation in cytosolic malate dehydrogenases, and its generative properties have produced viable functional sequences encoding multidomain ATPase metal transporters.^{63,64} By capturing higher-order relationships in sequence space beyond simple sequence identity, this approach can distinguish among the closely related, structurally similar ABH sequences while providing an interpretable framework to explore and guide hypotheses and experimental inquiries. Here, we trained and applied the LGL framework to classify carbon-halogen bond chemistries and uncover novel defluorinases with experimental validation from the visualized sequence space of the ABH superfamily.

Experimental

General. All reagents and supplies used in this study were acquired from AmBeed, Research Products International, Sigma-Aldrich, Thermo Fisher Scientific, USA Scientific, and VWR International, with exceptions noted.

Multiple Sequence Alignment (MSA) generation for training the LGL. An overview of the workflow for the training and application of the LGL is shown in Figure S1. The α/β -hydrolase (ABH) MSA was obtained by running a *hmmsearch* (hmmer.org) against the UniProt Knowledgebase (Swiss-Prot and TrEMBL, release prior to 2024), with the corresponding Pfam Hidden Markov Model (HMM) profile (Pfam ID: PF00561, Accessed 2023) using the Ganymede 2 High Performance Computing at The University of Texas at Dallas resulting in 977,130 sequences.^{40,51,65} Sequences were aligned with HMMER and aligned sequences with more than 20 contiguous gaps were excluded. The training MSA consisted of the subset of sequences annotated in Gene Ontology (GO:0016824) as hydrolases acting on acid-halide bonds (2,753 sequences) using the QuickGO API (Figure S2).^{66,67}

Latent Generative Landscape (LGL). Latent Generative Landscapes (LGL) provide a framework for analyzing the organization of protein sequence space by combining deep generative modeling with coevolutionary analysis.⁵⁸ In this approach, a variational



autoencoder (VAE) is trained on a MSA to learn a two-dimensional latent representation that captures the statistical structure of a protein family. The latent space is then systematically sampled, and sequences generated from each latent coordinate are evaluated using a direct coupling analysis (DCA) derived Potts Hamiltonian score, which reflects evolutionary constraints between amino acid positions.⁵⁹ Mapping these scores across the latent coordinates produces a latent generative landscape in which low Hamiltonian scores (lower energy) form dark blue basins corresponding to groups of sequences sharing similar evolutionary and functional characteristics, typically resulting in more stable and functional proteins. High Hamiltonian scores (higher energy) create bright green barriers that separate distinct regions of sequence space, with sequences less likely to be stable and less explored by the extant dataset. This landscape representation enables the visualization and analysis of polygenetic relationships, functional diversity and evolutionary trajectories within protein family. For this work, the LGL framework was used to train the MSA from above with a VAE, and a Hamiltonian score assigned fitness as a statistical energy to each sequence in the landscape. The training sequences with their corresponding predicted molecular function annotation haloalkane dehalogenases (GO:0018785) or haloacetate dehalogenases (GO:0018786) are shown in the LGL map of the annotated training data (Figure 2A), and the unannotated training sequences are mapped in the LGL map in Figure S3.

Plasmid design and preparation. The plasmids encoding the genes with the following UniProt accession numbers were synthesized by GenScript: A0A5P9PYN8, A0A5P9CV11, A0A960H4U9, A0A4V6IMR0, Q6NAM1, A0ABF7PGA4, A0A840N899, K8NY92, and A0A4Z0BVY8. Each gene was codon-optimized for expression in *Escherichia coli* and cloned between the NdeI and Sall restriction sites in the pET-28a(+) vector, resulting in a N-terminally polyhistidine-tagged enzyme. The nucleotide and amino acid sequences are shown in Figures S4–S12.

Each plasmid was prepared by reconstituting 4 μg of freeze-dried stock in 20 μL of autoclaved water. A 1 μL aliquot of this reconstituted plasmid was then diluted into 39 μL of autoclaved water to achieve a final concentration of 5 $\text{ng}/\mu\text{L}$. Subsequently, 1 μL of the diluted plasmid was transformed into *E. coli* 10G ELITE Competent Cells (Lucigen) using an electroporator (MicroPulser, Bio-Rad Laboratories). The resulting transformation mixture was diluted with pre-warmed Super Optimal Broth with Catabolite Repression (SOC) media and incubated for 60 min at 37 $^{\circ}\text{C}$ with shaking at 225 rpm (Benchtop Shaking Incubator 6790, Corning). This mixture was then plated onto Miller's Luria Broth (LB) agar containing 50 $\mu\text{g}/\text{mL}$ kanamycin and incubated for 14 h at 37 $^{\circ}\text{C}$ (Model 12-140 Incubator, Quincy Lab). Individual colonies were selected and inoculated into 5 mL of LB media supplemented with 50 $\mu\text{g}/\text{mL}$ kanamycin, then incubated for 16 h at 30 $^{\circ}\text{C}$ with shaking at 230 rpm (New Brunswick Innova 42R, Eppendorf). Finally, 2 mL of the culture was harvested via centrifugation at 8,000g (5424 R Centrifuge, Eppendorf) for 3 min at room temperature, and the plasmid was isolated using the QIAprep Spin Miniprep Kit (Qiagen).

Protein expression and purification. Protein expression and purification were performed by adapting and modifying a previously established protocol, with specific expression conditions for each protein detailed in Table S3.⁶⁸ In general, the commercial plasmid was transformed into *E. coli* EXPRESS BL21 (DE3) (Lucigen) cells via electroporation, and the recovery and



plating process followed the procedure described above. For expression, single colonies were inoculated into 10 mL of either LB or Terrific Broth (TB) containing 50 $\mu\text{g/mL}$ kanamycin and incubated at 30 °C for 14 h with shaking at 230 rpm. The following day, 2.5 mL of the overnight culture was diluted into 600 mL of LB or TB media (supplemented with 50 $\mu\text{g/mL}$ kanamycin) and incubated at 37 °C for 3 h with shaking at 250 rpm (New Brunswick Innova 44R, Eppendorf). Once the culture reached an Optical Density (OD_{600}) of 2, the shaker was gradually cooled to 30 °C or 18 °C while shaking for 15–30 min. Following this cooling period, protein expression was induced by adding 300 μL isopropyl β -D-1-thiogalactopyranoside (IPTG) from a 1 M stock solution or by relying on leaky expression, and the culture was incubated for an additional 21 h at 250 rpm. Finally, the culture was harvested by centrifugation at 3,200g (5810 R Centrifuge, Eppendorf) for 30 min at 4 °C. The resulting cell pellet was resuspended in pre-chilled 20 mM Tris buffer at pH 7.5 with 200 mM NaCl, 4 °C and stored at -20 °C.

Prior to purification, cell pellets were thawed overnight at 4 °C and resuspended by vortexing. Cell lysis was carried out by sonication in an ice-water bath at 35% amplitude, with a 15 s pulse on and 45 s off pulse cycle for a total time of 5 min (Q500, QSonica). The resulting lysates were clarified via ultracentrifugation at 37,000g for 35 min at 4 °C (Optima XPN-80 Ultracentrifuge, Beckman Coulter). The clarified supernatant was loaded onto a pre-equilibrated 5 mL HisTrap Ni-NTA column (Cytiva Life Sciences) using a sample pump operating at 15 °C (NGC Chromatography System, Bio-Rad Laboratories). Following sample loading, the column was washed with 20 column volumes (CV) of running buffer (20 mM Tris buffer at pH 7.5, 15 °C with 200 mM NaCl and 30 mM imidazole). The polyhistidine-tagged protein was then eluted using a two-step method: first with 3 CV of 5% elution buffer (20 mM Tris buffer at pH 7.5, 15 °C with 200 mM NaCl and 500 mM imidazole) mixed with 95% running buffer, followed by a linear gradient of elution buffer from 5% to 100% over 20 CV. Affinity elution fractions with absorbance signal at 280 nm were pooled and applied to a HiPrep 26/10 desalting column (Cytiva Life Sciences) equilibrated with 5 CV of 20 mM Tris buffer at pH 7.5, 15 °C with 150 mM NaCl. Desalted fractions exhibiting 280 nm absorbance were subsequently pooled and concentrated using an Amicon Ultra-15 Centrifugal Filter Unit with a 10 kDa molecular weight cut-off (MWCO) (EMD Millipore). Finally, the concentrated enzyme samples were buffer exchanged twice (final dilution of 1:100) into a storage buffer of 20 mM Tris buffer at pH 7.5, 4 °C with 50 mM NaCl using 10 kDa MWCO centrifugal filters. The final purified samples were stored at -20 °C. For every enzyme, two enzyme preparations were independently expressed and purified.

The purity of enzyme from each enzyme preparation was evaluated using Sodium Dodecyl Sulfate-Polyacrylamide Gel Electrophoresis (SDS-PAGE) based on our previously published study (Figure S13).⁶⁹

Protein concentration determination. The enzyme concentrations were calculated based on our previously published study.⁷⁰ The extinction coefficients for the full-length protein including the polyhistidine-tag of A0A5P9PYN8, A0A5P9CV11, A0A960H4U9, A0A4V6IMR0, Q6NAM1, A0ABF7PGA4, A0A840N899, K8NY92, and A0A4Z0BVIY8 were determined to be 52,370 $\text{M}^{-1}\text{cm}^{-1}$, 62,910 $\text{M}^{-1}\text{cm}^{-1}$, 47,440 $\text{M}^{-1}\text{cm}^{-1}$, 62,910 $\text{M}^{-1}\text{cm}^{-1}$, 65,890 $\text{M}^{-1}\text{cm}^{-1}$, 65,890 $\text{M}^{-1}\text{cm}^{-1}$



1cm^{-1} , $67,380\text{ M}^{-1}\text{cm}^{-1}$, $64,400\text{ M}^{-1}\text{cm}^{-1}$, and $58,900\text{ M}^{-1}\text{cm}^{-1}$, respectively, using the ProtParam tool in ExPASy.⁷¹

View Article Online
DOI: 10.1039/D6SC02671K

Colorimetric pH assay for enzyme activity. A colorimetric pH assay with phenol red was used to qualitatively determine haloalkane and haloacetate dehalogenase activity by proton release for the following enzymes: A0A5P9PYN8, A0A5P9CV11, A0A960H4U9, A0A4V6IMR0, and Q6NAM1. All experimental methods and substrates were based on previously reported procedures.^{22,46,72} For haloalkane dehalogenase activity, 1,2-dibromoethane, 4-bromobutyronitrile, and 1-iodopropane were tested, and for haloacetate dehalogenase activity, fluoroacetate, chloroacetate, and 2-fluoro-2-phenyl acetic acid were tested.^{43,46,73} To prepare substrate stock solutions, haloalkane substrates were diluted to 200 mM in isopropyl alcohol (IPA), and haloacetate substrates were diluted to 287.7 mM in 25 mM MOPS, buffered to pH 7.5 at 30 °C. These substrate stock solutions were combined with a 1 mg/mL phenol red stock in water to make a substrate/phenol red reaction stock solution, resulting in a final reaction concentration of 20 $\mu\text{g}/\text{mL}$ phenol red, 10 mM substrate, 875 μM MOPS buffer, and 5% IPA (for haloalkane reactions only). For haloalkane reaction stocks, the reaction stock contained 800 μL of 1 mg/mL phenol red, 2 mL of 200 mM substrate stock in IPA, and 1.4 mL of 25 mM MOPS pH 7.5 at 30 °C. For haloacetate reaction stocks, the reaction stock contained 800 μL 1 mg/mL phenol red and 1.4 mL 287.7 mM substrate stock. Enzyme stocks consisted of 4 μM enzyme diluted in water from an 80 μM enzyme stock in storage buffer (20 mM Tris buffer at pH 7.5, 4 °C with 50 mM NaCl), with the storage buffer being diluted 20-fold for the control reactions. Enzyme reactions contained a final concentration of 0.5 μM enzyme and 125 μM Tris buffer, while control reactions contained only 125 μM Tris buffer. Each 1 mL reaction took place in a 1.5 mL microcentrifuge tube and contained: i) 770 μL water for haloalkane reactions or 820 μL water for haloacetate reactions; and for positive control reactions, 20 μL water for haloalkane reactions or 70 μL water for haloacetate reactions; ii) 105 μL haloalkane phenol red stock or 55 μL haloacetate phenol red stock; iii) 125 μL 4 μM enzyme stock, or for control reactions, 125 μL 1 mM Tris buffer at pH 7.5, 4 °C with 2.5 mM NaCl; iv) for the positive control, 750 μL 1 mM HCl was added for a final reaction concentration of 750 μM HCl. Reactions were performed for two preparations per enzyme, with three reaction replicates per enzyme preparation and per substrate. Reactions proceed for 24 h at 30 °C (Thermo Scientific 2-Block Digital Dry Bath). Upon completion of the reaction, three 200 μL aliquots from each reaction were transferred to a Greiner 96-well clear, PS, F-bottom microplate, and the absorbance at 560 nm (Tecan Spark Microplate Reader) was measured in triplicate and averaged for each reaction. Subsequently, the reaction replicates were then averaged for each enzyme preparation, and the values from the two enzyme preparations were finally averaged to obtain the reported raw absorbance, along with the propagated standard deviation. To calculate the relative activity for each reaction, the average absorbance was normalized to the negative control of the corresponding substrate. The values used to generate Figure 2C can be found in Figure S14.

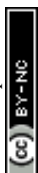
Identifying defluorinases by LGL from untrained, extant sequences. To obtain a set of non-annotated sequences with information excluded from the training data, we downloaded the entire ABH Pfam family (PF00561; 543,513 sequences) (Figure S1). This was retrieved on August 12, 2025 using the InterPro API.⁷⁴ After removing training set sequences, the remaining 540,933 sequences were aligned using HMMER, eliminating sequences with more



than 5% contiguous gaps (12 consecutive gaps), resulting in 296,598 sequences (Figure S15).⁷⁵ Using the QuickGO API, sequences were annotated with their Gene Ontology (GO) predicted function; this annotation was retrieved on August 13, 2025.⁶⁶ The resulting sequences were visualized on the LGL. All sequences in the haloacetate dehalogenase region were selected (3,014 total sequences) and the frequency of their respective GO annotations were analyzed (Table S4). These sequences were further aligned via Clustal Omega with respect to Q6NAM1.⁷⁶ The residues corresponding to the Q6NAM1 active site were extracted and visualized with WebLogo.^{77,78} Based on the three most frequently occurring GO annotations, sequences with conserved active site residues, less than three cysteines, and an isoelectric point (pI) less than 6.9 were selected for expression and characterization (Table S4). These included i) A0A840N899 and K8NY92 (GO: 0016787; hydrolase activity), ii) A0A4Z0BVY8 (GO:0047372; monoacylglycerol lipase activity, and iii) A0ABF7PGA4 (no GO annotation). After being identified for initial evaluation, A0ABF7PGA4 was subsequently annotated with hydrolase activity by QuickGO (January 28, 2026). The MSA for these selected sequences was generated using Clustal Omega (Figure S16).⁷⁶ A summary of pairwise identities pertaining to the training set annotated as haloacetate dehalogenases, the 3014 sequences selected from the haloacetate dehalogenase region and the enzymes tested in this study are reported in Figure S17 using the MSA for plotting onto the LGL (HMMER alignment with 5% (12) consecutive gaps removed).

Enzyme unfolding temperature (T_m) determination. Differential scanning fluorimetry (DSF) methods were adapted from a previous study.⁶⁹ In brief, the purified enzymes were adjusted to a concentration of 20 μ M using a buffer containing 20 mM Tris buffer at pH 7.5, 4 $^{\circ}$ C with 50 mM NaCl. Separately, a dye solution was prepared by diluting 3 μ L of 5000X SYPRO Orange Protein Gel Stain into 1.5 mL of the identical buffer. To set up the assay, 15 μ L of the prepared enzyme and 15 μ L of the diluted SYPRO Orange solution were combined into the wells of a MicroAmp Fast Optical 96-Well reaction plate. This yielded a final assay mixture consisting of 10 μ M enzyme and 5X SYPRO Orange dye within the 20 mM Tris buffer at pH 7.5, 4 $^{\circ}$ C with 50 mM NaCl. After sealing with a MicroAmp Optical Adhesive Film, the plate was centrifuged at 800g for 2 min at room temperature (5810 R Centrifuge, Eppendorf). Thermal shift assays were then performed on a QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems). The instrument was configured to use ROX as the reporter dye, with both the quencher and passive reference disabled. Measurements were recorded across a temperature range from 25 $^{\circ}$ C to 99 $^{\circ}$ C at a heating rate of 0.05 $^{\circ}$ C/s over 30 min. Post-run data analysis was conducted using Protein Thermal Shift Software v1.4. Raw fluorescence readings were normalized to facilitate sample comparison, and the unfolding temperature (T_m) for each enzyme was calculated via the Boltzmann fit method. For each experiment, three technical replicates were averaged to determine the value for a single enzyme preparation, and the final T_m values are reported as the average of two such enzyme preparations with the propagated standard deviation (Figure S18).

HPLC calibration curves. To quantify the enzyme activities, a calibration curve of the product (2-hydroxy-2-phenylacetic acid) and internal standard (IS, methyl benzoate) was generated (Figures S19–S20). Specifically, a stock solution of 100 mM 2-hydroxy-2-phenylacetic acid in 50 mM MOPS at pH 7 was first diluted to 30 mM in the same buffer, then serially diluted to 5, 10, 15, 20, and 25 mM. Next, 200 μ L of each diluted solution was transferred to a 1.5 mL



microcentrifuge tube containing 300 μL of 6.67 mM methyl benzoate in acetonitrile (HPLC grade). The product-IS solutions were mixed by vortexing and then clarified by centrifugation at room temperature for 15 min at 21,130g (5424 R Centrifuge, Eppendorf). After centrifugation, 450 μL of the product-IS mixture for each standard was transferred to a 2 mL HPLC screw cap vial. The HPLC analysis were performed using a Shim-pack Velox C18 column (2.7 μm , 3.0 \times 100 mm) with a Shim-pack Velox C18 EXP Guard Column Cartridge (2.7 μm , 3.0 \times 5 mm) on Shimadzu Prominence-i LC-2030C 3D. The mobile phase A was water, adjusted to pH 2.7 with formic acid, and the mobile phase B was acetonitrile. A gradient of 5% to 95% mobile phase B was applied over 4 min, and then held at 95% mobile phase B for 3 min, using a flow rate of 0.4 mL/min. An absorbance wavelength of 254 nm was used for detection of the product and internal standard. A background injection was performed using a mixture of 200 μL water and 300 μL acetonitrile. During data analysis, the background trace was subtracted from each sample trace using the Shimadzu LabSolutions software, and then all peaks were manually integrated. For each sample, three injections were conducted, and their average peak area constituted one technical replicate. Three such technical replicates were analyzed per product concentration, and the final linear regression was constructed by plotting the ratio of the product peak area (PDT) to internal standard peak area (IS) (2-hydroxy-2-phenylacetate/methyl benzoate) on the y-axis with standard deviation versus the known product concentrations on the x-axis. The origin (0, 0) was included in the dataset, and the linear regression was forced through this point. This calibration curve reports the product concentration prior to the addition of IS.

Optimization of enzyme reaction conditions. To optimize the reaction pH, reactions were performed in 50 mM sodium acetate buffer at pH 5, 50 mM MES buffer at pH 6, 50 mM MOPS buffer at pH 7, 50 mM bicine buffer at pH 8, and 50 mM CHES buffer at pH 9 at 30 $^{\circ}\text{C}$. To remain within the calibration curve range due to its high reactivity, the A0A4Z0BVY8 reaction was limited to 10 min. All other enzyme reactions for pH optimization were conducted for 120 min. To prepare the substrate solutions, racemic 2-fluoro-2-phenylacetic acid was dissolved in each buffer at 25 mM. To account for temperature-dependent shifts, the pH of the substrate solution was adjusted and confirmed at each reaction temperature. Each enzyme was diluted to 20 μM using a buffer containing 20 mM Tris buffer at pH 7.5 with 50 mM NaCl and all the following reactions were taken from the same enzyme stock solution. In each reaction, 5 μL of 20 μM diluted enzyme solution was added to 195 μL of the 25 mM substrate solution in each buffer, resulting in a final concentration of 0.5 μM enzyme and 24.375 mM substrate in a final 200 μL reaction volume. A control reaction was created for each condition containing 195 μL of the substrate solution in each buffer with 5 μL buffer containing 20 mM Tris buffer at pH 7.5 with 50 mM NaCl. The reactions proceeded for 10 min with A0A4Z0BVY8 or 120 min with the rest of the enzymes. All reactions were quenched with the addition of 300 μL acetonitrile containing 6.67 mM methyl benzoate as the internal standard. The quenched reaction was vortexed and clarified by centrifugation at room temperature for 15 min at 21,130g (5424 R Centrifuge, Eppendorf). After centrifugation, 450 μL of the reaction mixture was transferred to a 2 mL HPLC screw cap vial. A background injection was performed using a mixture of 200 μL water and 300 μL acetonitrile. During data analysis, the background trace was subtracted from each sample trace using the Shimadzu LabSolutions software, and then all peaks were manually integrated. For each reaction, two injections were performed and the average peak areas of these injections constituted one technical replicate. Product



concentration was determined for each technical replicate using the calibration curve described above with the technical replicate internal standard and product peak areas. Three technical enzyme replicates were conducted for each enzyme preparation at a given reaction condition with one technical control reaction for the same conditions. To account for background hydrolysis, the calculated product concentration produced by the technical control replicate was subtracted from that of each of the enzyme technical replicates. The three enzyme technical replicates are averaged for each enzyme preparation, and the final total turnover number (TTN) are reported as the average of two independent enzyme preparations, along with the propagated standard deviation.

The resulting averaged product concentration for each enzyme preparation was then used to calculate the TTN using the following equation:

$$TTN = \frac{[Product]}{[Enzyme]}$$

The calculated product concentration and resulting TTN for pH optimization of technical replicates can be found in Tables S5–S10, with the compiled data shown in Figure S21.

To optimize the reaction temperature, reactions were performed in 50 mM bicine buffer at pH 8. For A0A4V6IMR0, Q6NAM1, and A0ABF7PGA4, reactions were performed at 70 °C, 80 °C, and 90 °C, for A0A4Z0BVY8, K8NY92, and A0A840N899 reactions were performed at 60 °C, 70 °C, and 80 °C. To prepare the substrate solution, racemic 2-fluoro-2-phenylacetic acid was dissolved in 50 mM bicine buffer at pH 8 at 25 mM. To account for temperature-dependent shifts, the pH of the substrate solution was adjusted and confirmed at each reaction temperature. To prepare the enzyme solutions, A0A4Z0BVY8 was diluted to 4 μM and Q6NAM1 was diluted to 8 μM from 20 μM stock using a buffer containing 20 mM Tris buffer at pH 7.5, 4 °C with 50 mM NaCl. All other enzymes used 20 μM solution for temperature optimization reactions. In each reaction, a 5 μL diluted enzyme solution (4 μM for A0A4Z0BVY8, 8 μM for Q6NAM1, and 20 μM for the rest of the enzymes) was added to 195 μL of the substrate solution in each buffer, with a final concentration of 0.1 μM for A0A4Z0BVY8, 0.25 μM for Q6NAM1, and 0.5 μM for the rest of the enzymes and 24.375 mM substrate in a final 200 μL reaction volume.

The reaction work-up and data analysis were processed as described above. The calculated product concentrations and resulting TTNs for the temperature optimization can be found in Tables S11–S16, with the compiled data shown in Figure S22.

Enzyme enantioselectivity determination. To determine the stereoselective preference, a 10 μL diluted enzyme solution (20 μM) was added to 195 μL of the substrate solution (125 mM), resulting in a final concentration of 1 μM enzyme and 118.75 mM racemic 2-fluoro-2-phenylacetic acid in 200 μL of 100 mM bicine buffer at pH 8 at 60 °C. To account for temperature-dependent shifts, the pH of the substrate solution was adjusted and confirmed at the reaction temperature. A control reaction was prepared without enzyme with buffer and substrate only. The reaction times were 1 h for A0A4Z0BVY8, 4 h for Q6NAM1, A0ABF7PGA4, K8NY92, and 8 h for A0A4V6IMR0 and A0A840N899. The reaction time was 8 h for the no



enzyme control. All reactions were quenched first with 20 μL of 12 M hydrochloric acid and then extracted with ethyl acetate three times (2 x 500 μL , 1 x 300 μL). The extracted organic layers were transferred to a 1.5 mL microcentrifuge tube, dried with anhydrous sodium sulfate, and then centrifuged for 1 min at 21,130g at room temperature (5424 R Centrifuge, Eppendorf). The clarified solution was transferred to a 20 mL glass vial, and the solvent was removed by a rotary evaporator. Next, 1 mL of HPLC grade methanol was added to the vial, and the solution was transferred to a 1.5 mL microcentrifuge tube and clarified for 1 min at 21,130g at room temperature (5424 R Centrifuge, Eppendorf). The resulting clarified solution was transferred to a 2 mL HPLC screw cap vial for analysis.

The enantioselectivity of the enzyme reactions were determined by normal phase HPLC (Agilent 1260 Infinity II LC System) with a Lux Amylose-1 column (3 μm , 250 \times 4.6 mm) at 23 $^{\circ}\text{C}$. The compounds were eluted with 95:5 hexane/isopropanol and 0.5% trifluoroacetic acid with a flow rate of 1 mL/minute. The injection volume was 5 μL for each sample and the absorbance wavelength was monitored at 210 nm. The racemic 2-fluoro-2-phenylacetic acid, (*R*)-2-fluoro-2-phenylacetic acid, racemic 2-hydroxy-2-phenylacetic acid, (*R*)-2-hydroxy-2-phenylacetic acid, and (*S*)-2-hydroxy-2-phenylacetic acid were used as references. One control reaction and two enzyme reactions (one from each enzyme preparation) were evaluated, and all peaks were manually integrated. All parameters, including Retention Time (min), Area, Height, Width, Area%, and Symmetry, were analyzed with the Agilent ChemStation software (Tables S17–S24; Figures S23–S29). Enantiomeric excess (*ee*) was calculated for each compound and enzyme preparation by subtracting the Area % for the two enantiomers. The average *ee* with standard deviation for each enzyme preparation is reported in Table 1 and Table S25.

AI Disclosure. Claude Sonnet 4.5 was used to support generation of Python code for sequence extraction/annotation and data visualization.

Results and Discussion

The LGL framework learns sequence features from a multiple sequence alignment (MSA) of extant proteins (Figure S1).^{58,75} From the MSA, the VAE learns to encode sequences into a two-dimensional latent space, from which 250,000 grid points can then be decoded back into sequences.⁶¹ In parallel, DCA of coevolved residues in the MSA is then used to infer a Potts Hamiltonian model that assigns a statistical energy (Hamiltonian) score to each sequence, visualized on a color scale in a third dimension (Figure 2A).^{58,59} Sequences that better satisfy the statistical constraints of the MSA are ranked with lower (favorable energy) Hamiltonian values and are located in dark blue regions of the map.⁵⁸ In contrast, sequences less explored by the extant dataset and more likely to be unstable have higher (unfavorable energy) Hamiltonian scores and fall in the bright green regions. Within the landscape, sequence location ultimately reflects phylogenetic and functional relationships, while Hamiltonian scores are an estimate of relative sequence fitness that delimit separable functional regions within the family.



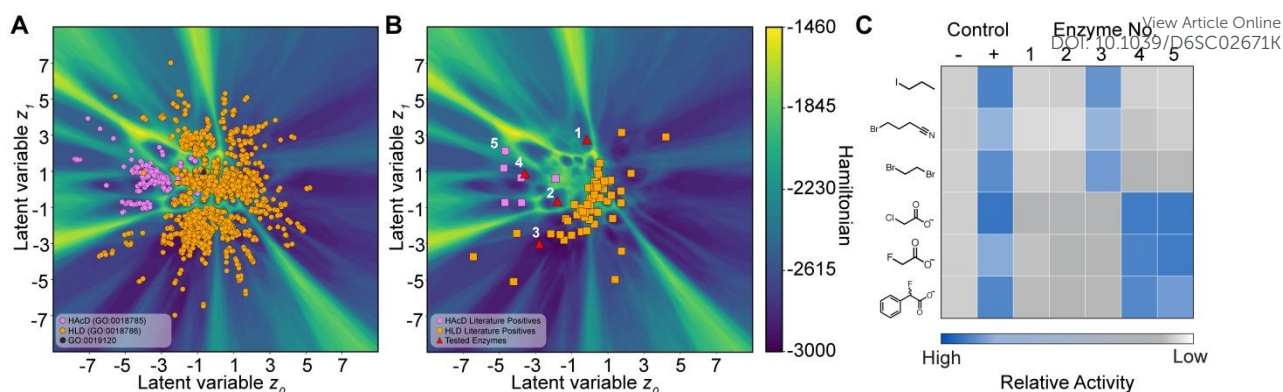


Figure 2. (A) Latent generative landscape (LGL) trained with sequences of the α/β hydrolase (ABH) superfamily acting on halide bonds (GO:0016824), with mapped haloacetate dehalogenases (HAcD; GO:0018785; pink circles), haloalkane dehalogenases (HLD; GO:0018786; orange circles), and one sequence with hydrolase activity acting on halide bonds in C-halide compounds (GO:0019120; brown circle). (B) Known HAcDs (pink squares) and HLDs (orange squares), along with sequences of unknown function selected for testing (red triangles), mapped onto the LGL. The numbers in panels B and C correspond to the enzyme entries in Table 1 below. (C) Qualitative phenol red activity for the negative (-) and positive (+) control reactions without enzymes along with the select enzymes in panel B for each substrate. Activity is calculated by normalizing the average absorbance value at 560 nm to that of the negative control for the corresponding substrate. The normalized values and raw absorbance data with standard deviations are shown in Figure S14.

To construct an LGL to discover fluoroacetate dehalogenases, we first trained the model on an aligned subset of 2,753 sequences from the ABH superfamily annotated in Gene Ontology (GO) as hydrolases acting on acid halide bonds (GO:0016824) (Figures S2–S3).^{66,79} Within this molecular function, two child terms define more specialized activities: diisopropyl-fluorophosphatase activity (GO:0047862) and hydrolase activity acting on acid halide bonds in C-halide compounds (GO:0019120). Because no sequences in the training set were annotated with GO:0047862, the functional space captured by the LGL reflects GO:0019120 and was dominated by the lower-level child terms associated with the ABH superfamily: haloacetate dehalogenase activity (GO:0018785) and haloalkane dehalogenase activity (GO:0018786) (Figure S2). We used this curated dataset to characterize sequence variability with a degree of functional annotation.

Mapping the GO-annotated molecular functions of the training set onto the LGL revealed that sequences mostly clustered into distinct regions, with high Hamiltonian score barriers separating the two activities (Figure 2A). To confirm this organization, we compiled and mapped experimentally validated haloalkane and haloacetate dehalogenases from the literature (Figure 2B; Tables S1–S2). The locations of these sequences agreed with the respective GO-annotated regions, with only a few exceptions at the origin of the LGL, which typically exhibit unfavorable or higher Hamiltonian scores, containing sequences less represented by the training MSA.

Next, we sought to experimentally assess the functional organization of the LGL. We used the LGL as a guide to choose enzymes from various parts of the cluster. Based on regions with specific GO molecular function annotations, we selected A0A5P9PYN8 (haloalkane dehalogenase; enzyme 1) and A0A5P9CV11 (haloacetate dehalogenase; enzyme 2) from regions lacking previously validated enzymes; A0A960H4U9 (haloalkane dehalogenase;



enzyme 3) and A0A4V6IMR0 (haloacetate dehalogenase; enzyme 4) from regions containing previously validated enzymes (Figure 2B; Table 1; Figures S4–S7). Q6NAM1 (enzyme 5) was included as a positive control for comparison in all assays (Figure 2B; Figure S8). All enzymes were expressed in *Escherichia coli* with an N-terminal polyhistidine tag and purified by nickel-affinity chromatography (Table S3; Figure S13). Enzymatic activity toward representative haloalkane substrates (4-bromobutyronitrile, 1-iodopropane, and 1,2-dibromoethane) and haloacetate substrates (chloroacetate, fluoroacetate, and 2-fluoro-2-phenylacetate) was evaluated using a colorimetric pH-shift assay with phenol red to detect proton release upon hydrolysis (Figure 2C; Figure S14).^{22,43,46,72,73} Notably, A0A960H4U9 and A0A4V6IMR0 displayed activity toward their predicted substrate classes, whereas A0A5P9PYN8 and A0A5P9CV11 showed no measurable activity, suggesting either that the assay conditions were not optimal or that the actual functions may differ from the currently assigned annotations.

Finally, we asked whether the functional organization of the LGL could be applied more broadly to reannotate extant sequences with incorrect or ambiguous GO-term assignments. To do this, we collated and curated the 543,513 sequences from the ABH superfamily (InterPro; August 12, 2025, Figure S1).^{74,80} After removing training-set sequences (GO:0016824) and those containing more than 5% contiguous gaps, a total of 296,598 sequences were mapped onto the LGL (Figure S15). Interestingly, 3,014 sequences were in the haloacetate dehalogenase region (Figure 3A). Furthermore, multiple sequence alignment indicated that residues needed for fluoroacetate recognition and catalysis in Q6NAM1 are conserved across this subset of sequences, suggesting potential defluorinase activity (Figure 3B).^{76–78}

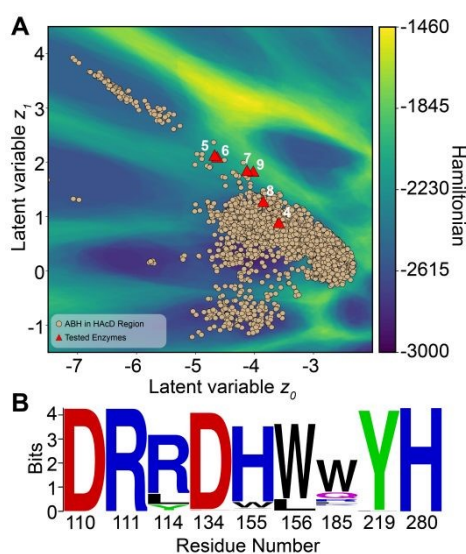


Figure 3. (A) ABH sequences outside of the training set (tan circles) mapped within the HAcD region of the LGL. Sequences selected for testing are indicated with numbered red triangles that correspond to the enzyme entries in Table 1 below. (B) WebLogo analysis of the active site residues in Q6NAM1 for the ABH sequences in the HAcD region shown in panel A.

To investigate whether these sequences indeed encode defluorinases, we selected representative sequences for experimental characterization belonging to the three most frequently occurring GO annotations: hydrolase activity, monoacylglycerol lipase activity, and



no annotation (Table S4). We hypothesized that we could increase the likelihood of identifying active enzymes by focusing on sequences located near A0A4V6IMR0 and Q6NAM1 (Figure 3A), while also considering biochemical properties relevant to expression and purification (i.e., cysteine content, isoelectric point). The selected candidates were A0ABF7PGA4, A0A840N899, K8NY92, and A0A4Z0BVY8 (enzymes 6–9, respectively) (Table 1; Figures S9–S12, S16, S17). After expressing and purifying these candidates as described above, we measured the unfolding temperature (T_m) of all enzymes to determine enzyme stability and establish optimal reaction conditions for further testing of the putative defluorinases (Figures S13, S18; Table S3).⁸¹ Q6NAM1, in line with prior data, and its close homolog A0ABF7PGA4 (98% identity) exhibited exceptionally high T_m values (> 90 °C).⁴³ Despite sharing modest sequence identity with Q6NAM1 (49–79%), A0A4V6IMR0, A0A840N899, K8NY92, and A0A4Z0BVY8 were also thermostable ($T_m > 70$ °C) (Table 1; Figure S17).^{43,76}

Lastly, we quantified the defluorination activity of these enzymes together with Q6NAM1 and A0A4V6IMR0 using 2-fluoro-2-phenylacetate (compound 3) as the substrate (Figure 1C; Table 1). Reaction conditions were screened across pH and temperature and analyzed by reverse-phase high-performance liquid chromatography (HPLC) (Figures S19–S22; Tables S5–S16). All enzymes shared the same pH optimum at pH 8, but their optimal reaction temperatures ranged from 60 °C to 90 °C, consistent with their measured unfolding temperatures (Table 1). Overall, the total turnover numbers (TTNs) ranged from 5671 to 59828 in 10 min (Table 1; Figure S22). Notably, A0A4Z0BVY8 exhibited approximately 2.7-fold greater activity than Q6NAM1, the current state-of-the-art enzyme. Each enzyme maintained a preference for the (*S*)-enantiomer substrate, generating the (*R*)-enantiomer product with high enantioselectivity ($> 93\%$) (Table 1; Figures S23–S28; Tables S17–S24).^{37,43} This conserved stereochemical outcome is in line with the S_N2 mechanism known for Q6NAM1.^{14,35,39}



Table 1. Bioinformatic annotations, melting temperature (T_m), and catalytic performance of the enzymes tested on racemic 2-fluoro-2-phenylacetate (compound 3). Data are reported as the average with propagated standard deviation of two independent enzyme preparations, each derived from the average of three technical replicates.

Enzyme Number	UniProt Accession Number	GO Molecular Function	T_m (°C)	Reaction Temperature (°C)	TTN ^[1]	4a ee (%) ^[2]	4b ee (%) ^[3]
1	A0A5P9PYN8	haloalkane dehalogenase	67.85 ± 0.08	- ^[4]	-	-	-
2	A0A5P9CV11	haloacetate dehalogenase	54.28 ± 0.07	-	-	-	-
3	A0A960H4U9	haloalkane dehalogenase	50.40 ± 0.17	-	-	-	-
4	A0A4V6IMR0	haloacetate dehalogenase	88.03 ± 0.26	70	5671 ± 128	87 ± 0	93 ± 0
5	Q6NAM1	haloacetate dehalogenase	92.14 ± 0.15	90	22176 ± 1102	91 ± 0	94 ± 0
6	A0ABF7PGA4	no annotation	90.74 ± 0.15	90	18129 ± 485	91 ± 0	94 ± 0
7	A0A840N899	hydrolase	78.85 ± 0.11	70	8911 ± 388	92 ± 1	93 ± 0
8	K8NY92	hydrolase	81.42 ± 0.16	70	10752 ± 837	89 ± 1	94 ± 0
9	A0A4Z0BVY8	monoacylglycerol lipase	78.61 ± 0.11	60	59828 ± 1241	99 ± 0	95 ± 0

[1] TTN, total turnover number.

[2] Percent ee, enantiomeric excess, of compound 4a ((*R*)-2-fluoro-2-phenylacetate).

[3] Percent ee of compound 4b ((*R*)-2-hydroxy-2-phenylacetate).

[4] -, not determined.

Conclusions

Motivated by the chemical significance and biotechnological potential of enzymatic carbon-fluorine bond cleavage, we applied our machine learning-based LGL framework to uncover hydrolytic defluorinases within the ABH superfamily. By training this model on GO-annotated hydrolases acting on acid halide bonds, the LGL revealed that distinct regions of latent space could be associated with particular enzyme functions. Specific examples from the literature, together with our experimental validation, confirmed the accuracy of the LGL designations for haloalkane and haloacetate dehalogenases, highlighting the potential to reannotate extant sequences with incorrect or ambiguous GO-term assignments. In this pursuit, we identified a total of 3,014 putative haloacetate dehalogenases and discovered five new fluoroacetate dehalogenases with a range of defluorination activities. A notable example is A0A4Z0BVY8 that surpassed the activity of the current state-of-the-art enzyme Q6NAM1 on 2-fluoro-2-phenylacetic acid.



Our results establish that the LGL framework can resolve functionally distinct carbon-halogen bond chemistries within the ABH sequence space, highlighting how LGLs can help reveal enzymes that are not readily accessible through traditional similarity-based annotations. While the annotated training sequences and the identified defluorinases cluster within a confined region of the LGL, they originate from diverse microbial species and span a wide range of sequence identities ruling out traditional phylogenetic relationships (Figure S17, Table S25). Moreover, the experimentally characterized defluorinases share identical active site residues, yet exhibit markedly different catalytic efficiencies. We conclude that although LGL regions delimited by latent coordinates are predictive of similar functionalities, their Hamiltonian scores are insufficient to predict specific values of catalytic activities (Table S25). Instead, global sequence context and structural dynamics are likely critical factors for function and will be important directions for future mechanistic investigations. Finally, the presence of thousands of putative defluorinases, despite the rarity of fluorinated natural products, suggests that this chemistry may be more deeply embedded in the ABH sequence space than previously recognized, raising questions regarding the natural roles of these enzymes and the evolutionary pressures that shaped them. To close, our approach lays the foundation for unraveling how natural and engineered protein sequence space encodes for rare and challenging carbon-fluorine bond chemistry in the ABH superfamily and beyond, opening new opportunities to advance defluorinases for sustainable synthesis and bioremediation applications.

View Article Online
DOI: 10.1039/D6SC02671K

Supporting Information

The data that support the findings of this study are available in the main text and supplementary material of this article. Sequence FASTA files and associated LGL files are available in a Zenodo repository (DOI: 10.5281/zenodo.15485612). The corresponding authors can be contacted for additional requests.

Author Contributions

K.J.: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing

S.S.B.: Data curation, Formal analysis, Investigation, Methodology, Validation, Visualization, Writing – original draft, Writing – review & editing

C.Z.: Formal analysis, Investigation, Methodology, Software, Validation, Visualization

M.N.: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing

A.D.L.P.: Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing

E.K.P.: Formal analysis, Investigation, Writing – review & editing

N.K.: Methodology



F.M.: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

View Article Online

DOI:10.1039/D6SC02671K

S.C.D: Conceptualization, Funding acquisition, Supervision, Writing – review & editing

Acknowledgements

Research support was provided to S.C.D. by the UT Dallas Sustainability Seed Program for Interdisciplinary Research and the Welch Foundation grant AT-2060-20240404; F.M. by the NIH grant R35GM133631 and NSF CAREER Award MCB-1943442. The authors acknowledge the UTD Center for High-Throughput Reaction Discovery and Synthesis for assistance with chiral chromatography analysis; Genome Center at UT Dallas for access to qPCR instrumentation; the Office of Information Technology Cyberinfrastructure Research Computing (CIRC) at The University of Texas at Dallas for providing for providing HPC and storage. This study does not represent the views of the supporting agencies and is the responsibility of the authors.

Conflicts of Interest

A provisional patent application has been filed on behalf of K.J., S.S.B., C.Z., F.M., and S.C.D. The remaining authors have no conflicts of interest.

References

- 1 R. Fuge, *Environ. Geochem. Health*, 1988, **10**, 51–61.
- 2 K. T. Koga and E. F. Rose-Koga, *Comptes Rendus. Chimie*, 2018, **21**, 749–756.
- 3 W. H. Schlesinger, E. M. Klein and A. Vengosh, *Global Biogeochem. Cycles*, 2020, **34**, e2020GB006722.
- 4 J. J. Petkowski, S. Seager and W. Bains, *Sci. Rep.*, 2024, **14**, 15575.
- 5 B. D. Key, R. D. Howell and C. S. Criddle, *Environ. Sci. Technol.*, 1997, **31**, 2445–2454.
- 6 D. B. Harper, D. O'Hagan and C. D. Murphy, in *Natural Production of Organohalogen Compounds*, ed. G. Gribble, Springer Berlin Heidelberg, Berlin, Heidelberg, 2003, pp. 141–169.
- 7 C. Walsh, in *Advances in Enzymology and Related Areas of Molecular Biology*, ed. A. Meister, John Wiley & Sons, Inc., Hoboken, New Jersey, 1983, pp. 197–289.
- 8 D. S. Fournoy and P. A. Frey, *Biochemistry*, 1986, **25**, 6036–6043.
- 9 S. Li and L. P. Wackett, *Biochemistry*, 1993, **32**, 9355–9361.
- 10 D. J. T. Porter, J. A. Harrington, M. R. Almond, W. G. Chestnut, G. Tanoury and T. Spector, *Biochem. Pharmacol.*, 1995, **50**, 1475–1484.
- 11 I. P. Solyanikova, O. V. Moiseeva, S. Boeren, M. G. Boersma, M. P. Kolomytseva, J. Vervoort, I. M. C. M. Rietjens, L. A. Golovleva and W. J. H. van Berkel, *Appl. Environ. Microbiol.*, 2003, **69**, 5636–5642.
- 12 F. Bellezza, A. Cipiciani, G. Ricci and R. Ruzziconi, *Tetrahedron*, 2005, **61**, 8005–8012.
- 13 L. Williams, T. Nguyen, Y. Li, T. N. Porter and F. M. Raushel, *Biochemistry*, 2006, **45**, 7453–7462.
- 14 P. W. Y. Chan, A. F. Yakunin, E. A. Edwards and E. F. Pai, *J. Am. Chem. Soc.*, 2011, **133**, 7461–7468.
- 15 D. O'Hagan and H. Deng, *Chem. Rev.*, 2015, **115**, 634–649.
- 16 V. Agarwal, Z. D. Miles, J. M. Winter, A. S. Eustáquio, A. A. El Gamal and B. S. Moore, *Chem. Rev.*, 2017, **117**, 5619–5674.
- 17 Y. Wang and A. Liu, *Chem. Soc. Rev.*, 2020, **49**, 4906–4925.
- 18 L. Wu and H. Deng, *Org. Biomol. Chem.*, 2020, **18**, 6236–6240.
- 19 M. García-Ramos, A. Cuetos, W. Kroutil, G. Grogan and I. Lavandera, *ChemCatChem*, 2021, **13**, 3967–3972.



- 20 P. W. Y. Chan, N. Chakrabarti, C. Ing, O. Halgas, T. K. W. To, M. Wälti, A. Petit, C. Tran, A. Savchenko, A. F. Yakunin, E. A. Edwards, R. Pomès and E. F. Pai, *ChemBioChem*, 2022, **23**, e202100414. New Article Online
DOI: 10.1039/D2SC02671K
- 21 Q. Huang, X. Zhang, Q. Chen, S. Tian, W. Tong, W. Zhang, Y. Chen, M. Ma, B. Chen, B. Wang and J. Wang, *ACS Catal.*, 2022, **12**, 265–272.
- 22 A. N. Khusnutdinova, K. A. Batyrova, G. Brown, T. Fedorchuk, Y. S. Chai, T. Skarina, R. Flick, A. Petit, A. Savchenko, P. Stogios and A. F. Yakunin, *FEBS J.*, 2023, **290**, 4966–4983.
- 23 M. F. Khan, S. Chowdhary, B. Kokscha and C. D. Murphy, *Environ. Sci. Technol.*, 2023, **57**, 9762–9772.
- 24 Y. Yu, F. Xu, W. Zhao, C. Thoma, S. Che, J. E. Richman, B. Jin, Y. Zhu, Y. Xing, L. Wackett and Y. Men, *Sci. Adv.*, 2024, **10**, eado2957.
- 25 L. P. Wackett and S. L. Robinson, *Biochem. J.*, 2024, **481**, 1757–1770.
- 26 C. M. Smorada, M. W. Sima and P. R. Jaffé, *Curr. Opin. Biotechnol.*, 2024, **88**, 103170.
- 27 L. Hendricks, C. R. Reinhardt, T. Green, L. Kunczynski, A. J. Roberts, N. Miller, N. Rafalin, H. J. Kulik, J. T. Groves and R. N. Austin, *J. Am. Chem. Soc.*, 2025, **147**, 9085–9090.
- 28 K. M. Cunningham, W. Shin and Z. J. Yang, *ChemPhysChem*, 2025, **26**, e202401130.
- 29 Y. Zhang, Y. Cao, C. V. Sastri and S. P. de Visser, *ACS Catal.*, 2025, **15**, 3898–3912.
- 30 X. Chen, Y. Zhang, J. Fu, Z. Chen, X. Zhou and X. Chu, *Commun. Chem.*, 2025, **8**, 396.
- 31 P. Goldman, *Journal of Biological Chemistry*, 1965, **240**, 3434–3438.
- 32 D. L. Ollis, E. Cheah, M. Cygler, B. Dijkstra, F. Frolow, S. M. Franken, M. Harel, S. J. Remington, I. Silman, J. Schrag, J. L. Sussman, K. H. G. Verschueren and A. Goldman, *Protein Eng. Des. Sel.*, 1992, **5**, 197–211.
- 33 M. Holmquist, *Curr. Protein Pept. Sci.*, 2000, **1**, 209–235.
- 34 K. Jitsumori, R. Omi, T. Kurihara, A. Kurata, H. Mihara, I. Miyahara, K. Hirotsu and N. Esaki, *J. Bacteriol.*, 2009, **191**, 2630–2637.
- 35 S. Miranda-Rojas and A. Toro-Labbé, *J. Chem. Phys.*, 2015, **142**, 194301.
- 36 T. Nakayama, T. Kamachi, K. Jitsumori, R. Omi, K. Hirotsu, N. Esaki, T. Kurihara and K. Yoshizawa, *Chem. Eur. J.*, 2012, **18**, 8392–8402.
- 37 J. Wang, A. Ilie, S. Yuan and M. T. Reetz, *J. Am. Chem. Soc.*, 2017, **139**, 11241–11247.
- 38 P. Mehrabi, E. C. Schulz, R. Dsouza, H. M. Müller-Werkmeister, F. Tellkamp, R. J. D. Miller and E. F. Pai, *Science*, 2019, **365**, 1167–1170.
- 39 Y. Yue, J. Fan, G. Xin, Q. Huang, J. Wang, Y. Li, Q. Zhang and W. Wang, *Environ. Sci. Technol.*, 2021, **55**, 9817–9825.
- 40 A. Bateman, M.-J. Martin, S. Orchard, M. Magrane, A. Adesina, S. Ahmad, E. H. Bowler-Barnett, H. Bye-A-Jee, D. Carpentier, P. Denny, J. Fan, P. Garmiri, L. J. da C. Gonzales, A. Hussein, A. Ignatchenko, G. Insana, R. Ishtiaq, V. Joshi, D. Jyothi, S. Kandasamy, A. Lock, A. Luciani, J. Luo, Y. Lussi, J. S. M. Marin, P. Raposo, D. L. Rice, R. Santos, E. Speretta, J. Stephenson, P. Totoo, N. Tyagi, N. Urakova, P. Vasudev, K. Warner, S. Wijerathne, C. W.-H. Yu, R. Zaru, A. J. Bridge, L. Aimò, G. Argoud-Puy, A. H. Auchincloss, K. B. Axelsen, P. Bansal, D. Baratin, T. M. Batista Neto, M.-C. Blatter, J. T. Bolleman, E. Boutet, L. Breuza, B. C. Gil, C. Casals-Casas, K. C. Echioukh, E. Coudert, B. Cucho, E. de Castro, A. Estreicher, M. L. Famiglietti, M. Feuermann, E. Gasteiger, P. Gaudet, S. Gehant, V. Gerritsen, A. Gos, N. Gruaz, C. Hulo, N. Hyka-Nouspikel, F. Jungo, A. Kerhornou, P. Le Mercier, D. Lieberherr, P. Masson, A. Morgat, S. Paesano, I. Pedruzzi, S. Pilbout, L. Pourcel, S. Poux, M. Pozzato, M. Pruess, N. Redaschi, C. Rivoire, C. J. A. Sigrist, K. Sonesson, S. Sundaram, A. Sveshnikova, C. H. Wu, C. N. Arighi, C. Chen, Y. Chen, H. Huang, K. Laiho, M. Lehvaslaiho, P. McGarvey, D. A. Natale, K. Ross, C. R. Vinayaka, Y. Wang and J. Zhang, *Nucleic Acids Res.*, 2025, **53**, D609–D617.
- 41 J.-Q. Liu, T. Kurihara, S. Ichiyama, M. Miyagi, S. Tsunasawa, H. Kawasaki, K. Soda and N. Esaki, *J. Biol. Chem.*, 1998, **273**, 30897–30902.
- 42 S. Wang, Z. Cheng, Y. Xu, L. Yang, J.-B. Wang, Z. Tian and X. Qu, *Green Synth. Catal.*, 2020, **1**, 60–65.
- 43 H. Zhang, S. Tian, Y. Yue, M. Li, W. Tong, G. Xu, B. Chen, M. Ma, Y. Li and J. Wang, *ACS Catal.*, 2020, **10**, 3143–3151.
- 44 M. Hu and C. Scott, *Appl. Environ. Microbiol.*, 2024, **90**, e00157-24.
- 45 S. C. Jansen, P. van Beers and C. Mayer, *Angew. Chem. Int. Ed.*, 2026, **65**, e24234.
- 46 W. Y. Chan, M. Wong, J. Guthrie, A. V. Savchenko, A. F. Yakunin, E. F. Pai and E. A. Edwards, *Microb. Biotechnol.*, 2010, **3**, 107–120.



- 47 S. Farajollahi, N. V. Lombardo, M. D. Crenshaw, H.-B. Guo, M. E. Doherty, T. R. Davison, J. Steel, E. A. Almand, V. A. Varaljay, C. Suei-Hung, P. A. Mirau, R. J. Berry, N. Kelley-Loughnane and P. B. Dennis, *ACS Omega*, 2024, **9**, 28546–28555. View Article Online
DOI: 10.1039/D3SC02671K
- 48 C. Badel, E. Bocconetti, R. Khodr, C. Husser, M. Ryckelynck and S. Vuilleumier, *Microbiol. Resour. Announc.*, 2025, **14**, e00812-24.
- 49 S. I. Probst, F. D. Felder, V. Poltorak, R. Mewalal, I. K. Blaby and S. L. Robinson, *Proc. Natl. Acad. Sci. U.S.A.*, 2025, **122**, e2504122122.
- 50 A. Forouzandeh, M. S. L. Schlosser, A. N. Vernon and T. K. Nielsen, *bioRxiv preprint*, 2025, DOI:10.1101/2025.07.23.666183.
- 51 A. Bairoch, *Nucleic Acids Res.*, 2000, **28**, 45–48.
- 52 F. Ozhelvaci and K. Steczkiewicz, *Proteins: Struct., Funct., Bioinf.*, 2025, **93**, 855–870.
- 53 T. Yu, H. Cui, J. C. Li, Y. Luo, G. Jiang and H. Zhao, *Science*, 2023, **379**, 1358–1363.
- 54 P. Kohout, M. Vasina, M. Majerova, V. Novakova, J. Damborsky, D. Bednar, M. Marek, Z. Prokop and S. Mazurenko, *JACS Au*, 2025, **5**, 838–850.
- 55 B. Norton-Baker, E. Komp, J. E. Gado, M. C. R. Denton, I. I. Mathews, N. P. Murphy, E. Erickson, O. O. Storment, R. Sarangi, N. P. Gauthier, J. E. McGeehan and G. T. Beckham, *ACS Catal.*, 2025, **15**, 16070–16083.
- 56 T. Lambert, A. Tavakoli, G. Dharuman, J. Yang, V. Bhethanabotla, S. Kaur, M. Hill, A. Ramanathan, A. Anandkumar and F. H. Arnold, *Nat. Commun.*, 2026, **17**, 1680.
- 57 F. Moorhoff, Y. Zhang, S. Qiu, W. Dong, D. Medina-Ortiz, J. Zhao and M. D. Davari, *ACS Catal.*, 2026, **16**, 12–30.
- 58 C. Ziegler, J. Martin, C. Sinner and F. Morcos, *Nat. Commun.*, 2023, **14**, 2222.
- 59 F. Morcos, A. Pagnani, B. Lunt, A. Bertolino, D. S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa and M. Weigt, *Proc. Natl. Acad. Sci. U.S.A.*, 2011, **108**, E1293–E1301.
- 60 A. J. Riesselman, J. B. Ingraham and D. S. Marks, *Nat. Methods*, 2018, **15**, 816–822.
- 61 P. K. Diederik and W. Max, *arXiv preprint*, 2019, **12**, DOI: <https://doi.org/10.48550/arXiv.1906.02691>.
- 62 X. Ding, Z. Zou and C. L. Brooks III, *Nat. Commun.*, 2019, **10**, 5644.
- 63 F. Montalvillo Ortega, F. Hossain, V. V. Volobouev, G. Meloni, H. Torabifard and F. Morcos, *ACS Cent. Sci.*, 2025, **11**, 1452–1466.
- 64 D. Shukla, J. Martin, F. Morcos and D. A. Potoyan, *J. Chem. Theory Comput.*, 2025, **21**, 3277–3287.
- 65 J. Mistry, S. Chuguransky, L. Williams, M. Qureshi, G. A. Salazar, E. L. L. Sonnhammer, S. C. E. Tosatto, L. Paladin, S. Raj, L. J. Richardson, R. D. Finn and A. Bateman, *Nucleic Acids Res.*, 2021, **49**, D412–D419.
- 66 D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan and R. Apweiler, *Bioinformatics*, 2009, **25**, 3045–3046.
- 67 R. P. Huntley, T. Sawford, M. J. Martin and C. O'Donovan, *Gigascience*, 2014, **3**, 4.
- 68 K. Ji, E. K. Pack, C. Maydew, K. A. Alberto, S. Abeyrathna, R. L. E. Villones, H. Gull, G. Meloni, S. O. Nielsen and S. C. Dodani, *Commun. Chem.*, 2025, **8**, 275.
- 69 K. Ji, K. Baek, W. Peng, K. A. Alberto, H. Torabifard, S. O. Nielsen and S. C. Dodani, *Chem. Commun.*, 2022, **58**, 965–968.
- 70 W. S. Y. Ong, K. Ji, V. Pathirana, C. Maydew, K. Baek, R. L. E. Villones, G. Meloni, A. R. Walker and S. C. Dodani, *Angew. Chem. Int. Ed.*, 2023, **62**, e202302304.
- 71 E. Gasteiger, A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch, *Nucleic Acids Res.*, 2003, **31**, 3784–3788.
- 72 P. Holloway, J. T. Trevors and H. Lee, *J. Microbiol. Methods*, 1998, **32**, 31–36.
- 73 T. Koudelakova, E. Chovancova, J. Brezovsky, M. Monincova, A. Fortova, J. Jarkovsky and J. Damborsky, *Biochem. J.*, 2011, **435**, 345–354.
- 74 M. Blum, A. Andreeva, L. C. Florentino, S. R. Chuguransky, T. Grego, E. Hobbs, B. L. Pinto, A. Orr, T. Paysan-Lafosse, I. Ponamareva, G. A. Salazar, N. Bordin, P. Bork, A. Bridge, L. Colwell, J. Gough, D. H. Haft, I. Letunic, F. Llinares-López, A. Marchler-Bauer, L. Meng-Papaxanthos, H. Mi, D. A. Natale, C. A. Orengo, A. P. Pandurangan, D. Piovesan, C. Rivoire, C. J. A. Sigrist, N. Thanki, F. Thibaud-Nissen, P. D. Thomas, S. C. E. Tosatto, C. H. Wu and A. Bateman, *Nucleic Acids Res.*, 2025, **53**, D444–D456.
- 75 S. C. Potter, A. Luciani, S. R. Eddy, Y. Park, R. Lopez and R. D. Finn, *Nucleic Acids Res.*, 2018, **46**, W200–W204.



- 76 F. Madeira, N. Madhusoodanan, J. Lee, A. Eusebi, A. Niewielska, A. R. N. Tivey, R. Lopez and S. Butcher, *Nucleic Acids Res.*, 2024, **52**, W521–W525. View Article Online
DOI: 10.1039/D6SC02671K
- 77 T. D. Schneider and R. M. Stephens, *Nucleic Acids Res.*, 1990, **18**, 6097–6100.
- 78 G. E. Crooks, G. Hon, J.-M. Chandonia and S. E. Brenner, *Genome Res.*, 2004, **14**, 1188–1190.
- 79 R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin and C. O'Donovan, *Nucleic Acids Res.*, 2015, **43**, D1057–D1063.
- 80 T. Paysan-Lafosse, A. Andreeva, M. Blum, S. R. Chuguransky, T. Grego, B. L. Pinto, G. A. Salazar, M. L. Bileschi, F. Llinares-López, L. Meng-Papaxanthos, L. J. Colwell, N. V. Grishin, R. D. Schaeffer, D. Clementel, S. C. E. Tosatto, E. Sonnhammer, V. Wood and A. Bateman, *Nucleic Acids Res.*, 2025, **53**, D523–D534.
- 81 F. H. Niesen, H. Berglund and M. Vedadi, *Nat. Protoc.*, 2007, **2**, 2212–2221.



The data that support the findings of this study are available in the main text and supplementary material of this article. Sequence FASTA files and associated LGL files are available in a Zenodo repository (DOI: 10.5281/zenodo.15485612). The corresponding authors can be contacted for additional requests.

