Enzymatic carbon-fluorine bond cleavage by human gut microbes

Silke I. Probst[1], Florian D. Felder[1], Victoria Poltorak[1], Ritesh Mewalal[2], Ian K. Blaby[2], Serina L. Robinson[1*]

[1]Department of Environmental Microbiology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), 8600 Dübendorf, Switzerland
[2]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

*corresponding author: serina.robinson@eawag.ch

Keywords: Haloacid dehalogenase, Defluorination, Human gut microbiota, 5-Fluorouracil, Fluoroacetate, Molecular dynamics, Protein carboxyl-terminus, Alanine scanning, Protein engineering

Abstract

The human gut microbiota is known for its pharmaceutical biotransformation capabilities. Many pharmaceuticals are fluorinated; however, it was previously unknown whether gut microbes were capable of cleaving carbon-fluorine bonds. Here we show that dehalogenases encoded by human gut microbes, including Clostridia, Bacilli, and Coriobacteriia, are active on fluorinated amino and organic acids. Through the development of a rapid, 96-well fluoride detection assay, we screened and characterized new microbial gut defluorinases. Using alanine scanning and chimeric protein design, we identified a 41 amino acid carboxyl-terminal region which is necessary and sufficient to enable defluorination activity when fused to a non-defluorinating dehalogenase. Based on this carboxyl-terminal region alone, we were able to predict the defluorination or non-defluorination activity of microbial dehalogenases with >83% accuracy. By uncovering the hidden potential of human gut microbes to catalyze defluorination reactions, our findings lay the foundation for therapeutic interventions relevant to human and gut microbiome health.

Introduction

Over 20% of pharmaceuticals and 50% of agrochemicals released in the past 25 years contain one or more fluorine atom(s)[1,2]. Fluorinated compounds are also widespread in everyday items spanning food packaging textiles, face masks, personal care products, sports equipment, and firefighting foams, leading to frequent human exposure[3,4]. Fluorination often improves bioavailability and extends compound half-life. However, fluorine also has a 'dark side' due to the production of fluorinated degradation products which can disrupt host metabolism[5]. For example, key degradation products of the widely-used chemotherapeutic 5-fluorouracil (5-FU) including fluorocitrate, α-fluoro-β-alanine, and fluoroacetate, inhibit the citric acid cycle and cause severe neuro- and cardiotoxicity[6].

The gut microbiota is known to modify and thereby inactivate or increase the toxicity of various xenobiotics[7] including fluorinated compounds. *Escherichia coli* hydrolyzes the antineoplastic prodrug fludarabine, enhancing its cytotoxic activity[8]. The chemotherapeutic gemcitabine is modified through deamination[9] or acetylation[8]. The efficacy of 5-FU was shown to be altered by microbial metabolism, leading to inactivation by degradation[10]. Given the increased use of

fluorinated pharmaceuticals and the ingestion of fluorinated xenobiotics from the environment, we hypothesized that microbial carbon-fluorine bond cleavage may also be catalyzed by commensals in the human gut. This could impact drug activity by increasing toxicity and side effects and may potentially influence the significant interpatient variability in responses to fluorinated compounds, which has been associated with inter-individual differences in the composition and metabolic capabilities of the human gut microbiota[11].

To date, there is a limited number of enzymes capable of defluorination, likely due to the scarcity of fluorinated molecules in nature[12], the toxic effects of fluoride on microorganisms[13], and the strength of the carbon-fluorine bond[14]. Haloacid dehalogenase (HAD) family enzymes are known to catalyze hydrolytic dehalogenation and, in some cases, can defluorinate[15,16]. This makes them one of the few enzymes capable of breaking carbon-fluorine bonds without requiring oxygen, cofactors or the presence of other enzymes, which has potential applications in medical and environmental remediation.

Although the few defluorinating HADs reported have been studied in detail[15,17], it remains challenging to rapidly screen metagenomic datasets such as from the human gut for HADs which have defluorination capabilities. Previously, molecular dynamics (MD) simulations identified a correlation between the frequency with which protein conformations samples a compact active site configuration with the ability of HADs to defluorinate[15]. Here we uncovered that MD analysis did not generalize to a larger set of proteins, which motivated our search for alternative and accessible features governing defluorination activity encoded in the human gut microbiome.

1. Identification and selection of HADs from the HumGut Database

We retrieved protein sequences of known environmental HADs (2-haloacid dehalogenase subfamily) previously characterized for their defluorination activity[15] from NCBI (WP_011593529.1 and WP_011481495.1). The sequences were used as queries for BLAST searches within the HumGut human gut genome collection (Version: February 2023)[18]. A BLASTdb was created using the 30,691 HumGut genomes and a tblastn search was performed to identify homologs of the WP_011593529.1 and WP_011481495.1 proteins. The 500 sequence hits with bitscores greater than 40 were subjected for a reverse blast search against the NCBI database. Hits that could be mapped to NCBI database entries with 100% sequence identity and 90-100% query coverage were further curated for redundancies. The resulting list of 114 protein sequences (Table S1) were aligned using Clustal Omega[19] and the presence of the catalytic active residues were verified. Code and query sequences are available at https://github.com/MSM-group/biodefluorination-paper.
Subsequently, we used the EFI-EST webtool to generate a sequence similarity network (SSN) with default parameters[20]. We visualized the 95% amino acid identity representative network using Cytoscape[21] (Fig. S2) and seven SSN cluster representatives were selected for experimental validation. Genes were synthesized as described below for in vitro work. In a second 'validation' round, 11 new proteins were selected based on high sequence similarity alignment scores for additional experimental validation (Table S2, S7).

2. Cloning and heterologous expression of candidate enzymes

The first batch of full-length dehalogenases (n=7) were codon-optimized for *E. coli* codon usage using the Integrated DNA Technologies Codon Optimization Tool. The sequences were synthesized as gBlocks by Integrated DNA Technologies. The gBlocks were cloned individually by Gibson Assembly into multiple cloning site 1 of pCFDuet-1 vectors (spectinomycin/streptomycin resistance marker) with tobacco etch virus (TEV) cleavage sites and C-terminal hexahistidine tags (Table S5). Constructs were transformed into *E. coli* DH5α cells and verified by Sanger sequencing (Microsynth, Balgach, Switzerland). Sequence-verified constructs were transformed into *E. coli* T7 Express cells for protein production (New England Biolabs Frankfurt, Germany, NEB C2566I). The second batch of synthesized genes (n=11) were cloned by the U.S. Department of Energy Joint Genome Institute. These were constructed identically to the first batch with the exception of using the build-optimization software, BOOST[22] for codon optimization and expression in T7 Express *E. coli*. Codon-optimized sequences are available at: https://github.com/MSM-group/biodefluorination-paper/tree/main/protein_sequences_info.

Starter overnight cultures (5 mL) were grown in LB with 100 µg/mL spectinomycin overnight at 37°C. From the starter culture, 0.5 mL were added to 200 mL Terrific Broth with 100 µg/mL spectinomycin in a 1 L baffled Erlenmeyer flask and grown to an optical density of 0.5-0.8 at 37°C. Protein expression was induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and cultures were incubated for 24h at 16°C.

For scale up and purification, induced cultures were poured into 50 mL Falcon tubes and centrifuged at 4000 RCF for 30 min. The supernatant was decanted, and cell pellets were resuspended in 4 mL buffer A (20 mM Tris-HCl, 500 mM NaCl, 20 mM imidazole) per gram cell pellet, 10% glycerol, pH = 8) per gram of pellet weight. Cells were lysed by 3 cycles of sonication using a 6 mm needle, 30% amplitude and 10 x 10 second pulses with a rest time of 10 seconds in between (total sonication time per cycle = 3 minutes). Lysate was centrifuged at 20,000 RCF at 4°C for 45 min. For benchtop purification, supernatants were applied to a column packed with Ni-NTA agarose beads (Qiagen) pre-equilibrated in buffer A. The column was subsequently washed with buffer A and then subjected to two washes with buffer B (40 mM Tris-HCl, 500 mM NaCl, 10% glycerol and 40 mM imidazole). Enzymes were eluted from the column using buffer C (40 mM Tris-HCl, 500 mM NaCl, 10% glycerol, and 500 mM imidazole) The fractions with the highest protein concentration was determined qualitatively by Bradford assay[23]. Imidazole and NaCl were removed by running the sample over a PD-10 desalting column (Cytiva) and eluting in buffer (100 mM Tris-SO_4, pH = 8,5). The concentration of the purified proteins were determined by Qubit™ protein assay and the proteins were flash frozen using liquid nitrogen in 500 µL aliquots and subsequently stored at -80°C.

3. Assessment of enzyme-substrate specificity

Stock solutions containing 100 mM of the test substrates listed in Table S3 dissolved in water were prepared. The concentration of the Ni-NTA purified proteins were determined by using the Qubit™ protein assay and the proteins were diluted to a final concentration of 0.5 mg/mL in 100 mM Tris-SO_4 buffer, pH 8.5. As controls the proteins were boiled for 10 min at 100°C. The test

substrates were added to a final concentration of 1 mM and the samples were incubated at 37°C. 50 µL samples were taken for the different timepoints and mixed with 50 µL acetonitrile to quench the reaction. The samples were centrifuged at 13 000 RPM for 5 min and the supernatant transferred into plastic vials for ion chromatography (IC).

An IC system (Metrohm, Switzerland) was used consisting of an IC sampling center 889 coupled to an ion chromatograph (930 Compact IC Flex) with a conductivity detector. Samples were incubated at 10°C in the sample center with 10 µL injection volumes first onto a precolumn (Metrosep A Supp 16, Guard, 4 mm) and subsequently onto a Metrosep A Supp 16 column (250 x 4.0 mm, 5 µm). The temperature in the column was kept constant at 30°C. The eluent consisted of 2.0 mM Na_2CO_3 and 5.0 mM $NaHCO_3$. The flow rate was set to 0.8 mL/min. IC data processing and quantification was performed using MagIC Net Professional (version 3.3).

4. AlphaFold and structural comparison

Structural models for the seven initial dehalogenase sequences selected from the HumGut databases, were generated using AlphaFold2 (AF) ColabFold[24]. Models with the highest rank for each dehalogenase were prepared with relaxation using the Amber99SB force field[25]. Additionally, we used AF to generate models for the HADs PA0810, RSC1362, ST2570 and RHA0230. The accuracy of AF models was compared to known X-ray crystal structures, where available[15].

5. Molecular dynamics simulations

Molecular dynamics simulations were performed using the GROMACS package version 2022.5[26,27] compiled with CUDA support and run on NVIDIA GeForce RTX 2080 Ti graphic cards. The crystal structures for HADs PA0810, RSC1362, ST2570 and RHA0230 were downloaded from RCSB PDB (PDB IDs: 3UMC, 3UMB, 2W43 and 3UMG, respectively. For PDB files which contained selenomethionine, selenomethionine was exchanged with methionine. Relaxed AF models were also subjected to simulations. The system was prepared with GROMACS tools, parametrized with the force field Amber99SB-ILDN[28], and solvated in explicit water using the default water model TIP3P[29] in a cubic box ~1.2 nm bigger than the protein in each direction. The protein's net charge was neutralized with Na^+ or Cl^- ions. Conjugate gradient energy minimization (cg) was performed with a minimum of 50 steps and a maximum of 1000 steps or until the energy of the system reaches below 1000 kJ/mol/nm. After a short (200 ps) NVT equilibration and (200 ps) NPT equilibration, the production runs were carried out. The simulations were performed under periodic boundary conditions for a total simulation length of 0.1 µs with ten replicates and 1 µs with two replicates using the leap-frog algorithm with time steps of 2 fs for each simulation. The bonds involving hydrogen atoms were constrained with the LINCS algorithm[30]. The cutoff scheme for nonbonded interactions was Verlet with a cutoff of 1.2 nm (for all AMBER force fields). The Coulomb interactions were calculated using the particle mesh Ewald (PME) algorithm[31] and a Fourier spacing of 0.14 nm. The temperature was controlled using a modified Berendsen thermostat[32] with a coupling time of 0.1 ps (protein & non-protein coupling groups) and a reference temperature of 300K. The pressure was controlled using the Berendsen barostat[33] an isotropic

pressure coupling in x-y-z directions, a coupling time of 1.0 ps, a reference pressure of 1.0 bar, and an isothermal compressibility of 4.5 x 10^{-5} bar^{-1}. The variance of the 100 ns simulations with 10 replicates ranged between 0.015 and 0.08 nm for all simulated proteins. For the 1 µs runs the variance between the two replicates ranged between 0.0016 nm and 0.031 nm. The RMSD of the simulated proteins relative to the starting structure (Cα and all atoms) was measured using gmx_rms for both whole proteins and for the protein segments with the necessary index files generated with make_ndx. The three active site residues of the halogen binding site (aspartate, asparagine and arginine) were identified and the Euclidean distance between the three Cα atoms was tracked during the simulation after an equilibration period of 500 ns with the built-in tool gmx distance[34]. To measure the Root Mean Square Fluctuation of all Cα atoms during the simulation the built-in function gmx_rmsf was used.

6. Sequence based comparison of HADs

Multiple sequence alignment and maximum-likelihood phylogenetic analysis (Fig. S17-S18) were performed to gain a better perspective on the sequence features of the HAD-like enzymes. We performed these two analyses with a data set of the entire HAD-like family available in UniProt[35] for 248 sequences available as of August 2023 with experimental evidence characterized at the protein level. A multiple sequence alignment of the resulting 248 proteins was performed using the ClustalOmega algorithm[36] and the phylogenetic tree (Fig. S17-S18) was generated using the DistanceTreeConstructor method from the biopython library[37].

7. Rational active site engineering

To identify potentially significant amino acids involved in the defluorination process, we used two of seven in vitro tested proteins sharing high structural and sequential similarity but distinct activities from each other (Fig. S19) (54% sequence identity, 68% sequence similarity). WP_1187090781.1, a non-defluorinating enzyme, and WP_1786180371.1, a defluorinating enzyme. We targeted amino acids within WP_1187090781.1, particularly those situated in or near the active site, for substitution with their counterparts from WP_1786180371.1. Four single variants were constructed: F54I, T57M, F74H, and F187S and their combinations resulting in 14 variants encoded by point mutations in codons for up to four different active site residues. The mutant gene constructs were generated by using mismatch primers (Table S5). A 3 µL Kinase, Ligase & DpnI (KLD) Treatment Reaction containing 0.5 µL of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes at 37°C.

8. Colorimetric assay for fluoride detection

A colorimetric fluoride screening method was developed here for 96-well plate format based on the agar-plate based colony screening method initially described by Davis et al.[38] Xylenol orange stock solution was prepared with a concentration of 20 mM. To prepare the zirconium oxychloride octahydrate stock solution, 0.3022 g of zirconium(IV)-oxychloride octahydrate was dissolved in 925 µL water and 1.85 mL 12.1 M HCl added and then boiled for 5 min to achieve final

concentrations of 500 mM fully depolarized zirconium and 8 M HCl. After cooling, the stock solution was diluted to a concentration of 50 mM zirconium. Shortly before the experiment, the working solution was prepared with 7.8 M HCl, 400 µM xylenol orange and the complex was initiated by the addition of 250 µM zirconium. A list of potential interfering substances and buffers were tested and results are listed in Table S6.

E. coli T7 expression strains carrying the corresponding plasmids were grown in TB medium in a 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 µg/mL spectinomycin until a optical density of 0.8 was reached The cells were induced with 0.1 mM IPTG and incubated for 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant was discarded and the cells were washed twice with 2 mL 20 mM HEPES (pH 8). After centrifugation at 4000 RCF for 10 min the cells were resuspended in 1 mL 20mM HEPES (pH 8) and fluoroacetate was added to a final concentration of 1mM prior to incubation at 37°C for 24 h. The cells were centrifuged at 4000 RCF for 15 min and 200 µL of the supernatant was transferred to a new 96 well plate and 40 µL of the zirconium/xylenol orange working solution was added. The final concentrations in each well were 1.3 M HCl, 66.7 µM xylenol orange, and 41.7 µM zirconium.

The absorbances used to calculate fluoride ion concentrations were measured at 440 nm and 540 nm and the isosbestic points were measured at 370 nm and 480 nm using an Eon spectrophotometer (Biotek). The approximate fluoride concentration is calculated using the following formula: x=(abs(540 nm))/(abs(440 nm)) where a ratio above 2 indicates a [F^-] greater than 100 µM. Strongly positive samples (more than 100 µM [F^-]) appear yellow (x>2) and negative samples appear red (less than 10 µM [F^-]) (x<1.8). Concentrations between 10-100 µM appear in different shades of orange (2>x>1.8). The assay shows a linear relationship between fluoride concentration and x from 0 µM to 125 µM and a logarithmic relationship from 50 µM to 2000 µM.

9. Chloride detection assay

A chloride detection assay kit (QuantiChrom™ Chloride Assay Kit) was obtained from Huber lab and assays were run according to the manufacturer's instructions with the following changes: Briefly, *E. coli* T7 expression strains carrying the corresponding plasmids were grown in TB medium in a 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 µg/mL spectinomycin until an optical density of 0.8 was reached. The cells were induced with 0.1mM IPTG and incubated for 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant was discarded and the cells were washed twice with 2 mL 20 mM HEPES (pH 8). After centrifugation at 4000 RCF for 10 min the cells were resuspended in 1 mL 20 mM HEPES (pH 8) and chloroacetate was added to a final concentration of 30 mM prior to incubation at 37°C for 24 h. The cells were centrifuged at 4000 RCF for 15 min and 2.5 µL of the supernatant was transferred to a new 96 well plate and 100 µL of the colorimetric working solution was added. The chloride concentration is measured at 610 nm with an Eon spectrophotometer (Biotek). The limit-of-quantification was 0.2 mM.

To increase the comparability between the fluoride and chloride assay results, the values of both assays were 0 to 1 linearized and thresholds used to decide whether they are positive for their

respective activity. The normalization formula used is as follows: (raw value - lowest value)/(highest value - lowest value). For the chloride assay values below 0.15 were counted as negative for dechlorination and for the defluorination assay values below 0.4 were counted as negative for defluorination. Differences in the chloride and fluoride values greater than 0.275 were counted as exhibiting different dechlorination vs. defluorination activity.

10. Whole-gene alanine-scanning mutagenesis and chimeragenesis

High-throughput alanine-scanning mutagenesis was performed using the Q5® Site-Directed Mutagenesis Protocol (New England Biolabs Inc.) with the following amendments. Briefly, back-to-back forward and reverse primers were designed using NEBaseChanger (https://nebasechanger.neb.com/), where the alanine codon substitution was contained in the center of the forward mutagenic primer. Note that a histidine codon was substituted in place of any native alanine codons. Primers were ordered from Integrated DNA Technologies. The complete set of alanine (n=475) are available on https://github.com/MSM-group/biodefluorination-paper/tree/main/Alanine_scanning_primers. A 6 µL PCR reaction, containing 1x Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs), 0.5 µM of the forward and reverse primer, and 10 ng template DNA, was carried out for each variant. WP_178618037.1 (*Guopingia tenuis*) and WP_118709078.1 (*Enterocloster aldenensis*) cloned into the pCDFDuet-1 vector was used as template DNA. PCR cycles were 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 65°C for 20 seconds and 72°C for 2 minutes, followed by a 2-minute final extension at 72°C. A 3 µl KLD Treatment Reaction containing 0.5 µl of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes at 37°C. The entire reaction was transformed into NEB® DH5α competent *E. coli* using the manufacturer's recommended transformation protocol and sequence verified on the Pacific Biosciences Revio platform (Pacific Biosciences, CA) and analyzed using custom pipelines. Plasmid DNA of sequence-verified mutant variants were isolated and transformed into T7 Express cells (New England Biolabs), which was once again sequence verified (Pacific Biosciences). We engineered five chimeric proteins by combining longer sequences from WP_1786180371.1 and WP_1187090781.1 (Table S5). Gibson assembly (primers in Table S5) was used for the construction of the plasmids.

Results

The human gut microbiota encodes HAD-like enzymes

To investigate the potential for defluorination by human gut microbial enzymes, we first tested defluorination activity fluoroacetate as a simple fluorinated substrate and known 5-FU degradation product[39]. Notably, prior studies identified the soil bacterium, *Rhodococcus jostii* RHA1, as a degrader of fluoroacetate, along with the discovery of its responsible HAD family enzyme[16]. These enzymes dehalogenate small chlorinated alkanoic acids with only a few demonstrating the capability to cleave the more recalcitrant carbon-fluorine bond[15]. We therefore used the *Rhodococcus jostii* RHA1 HAD (Rha0230) for a protein homology search in the HumGut database[18]. To our surprise, we detected over 500 HAD homologs encoded in human gut metagenomes (Fig S1, Table S1).

The identified HAD-encoding genes were predominantly found in bacteria belonging to the phylum Bacillota, formerly named the Firmicutes[40]. However, homologs were also prevalent in the Actinomycetota, Pseudomonadota, and Bacteroidota (Fig. 1A, Fig. S1). To further elucidate the relationships among these proteins, we constructed a protein sequence similarity network (Fig. S2), which revealed several distinct clusters[18,41]. Two of these clusters contained characterized members of the 2-haloacid dehalogenase subfamily[15] within the larger HAD-like superfamily[42]. Other smaller clusters could be assigned to the HAD-family phosphatases capable of phosphoryl transfer chemistry and family members with other functions (e.g., azetidine hydrolysis)[42,43]. For subsequent analysis, we focused on clusters containing the 2-haloacid dehalogenases and curated the dataset through a multiple sequence alignment to eliminate redundant sequences and confirmed the presence of the catalytic residues (Fig. S3). The selected sequences showed low amino acid sequence identity to reference 2-haloacid-dehalogenases (18% - 46%; Table S1), from which we selected seven new representatives for in vitro experiments (Fig. S2, Table S2).

Human gut microbial dehalogenases are active in carbon-fluorine bond cleavage

To screen for in vitro defluorination activity, we heterologously expressed, purified, and tested seven selected gut microbial dehalogenases with ten different halogenated substrates (nine fluorinated organic acids and one chlorinated substrate, chloroacetate, Fig. 1, Table S3). The HAD from *Rhodococcus jostii* (P5, Protein Data Bank ID: 3UMG) was included as a positive defluorination control as described previously. All proteins tested except for one (P1, *E. aldenensis,* Fig. 1C) showed dehalogenase activity with at least one substrate (Fig. S4-S11).
Two enzymes hydrolyzed chloroacetate but were not active with any fluorinated substrates.
The remaining four enzymes hydrolyzed fluoroacetate, producing fluoride ion and glycolate stoichiometrically, as measured by ion chromatography (Fig. 1B). Two of the enzymes were additionally active with other small fluorinated organic acids which are known transformation products of fluorinated xenobiotics, namely difluoroacetate, 2-fluoropropionate and α-fluoro-β-alanine.

To further investigate the activity of proteins under anaerobic environments, such as the gut, we tested the anaerobic defluorination capabilities of P3 (*Gordonibacter urolithinfaciens*). P3 effectively cleaved fluoroacetate to produce glycolate and fluoride ions under both aerobic and anaerobic conditions, as would be expected for hydrolytic enzymes (Fig. S12).

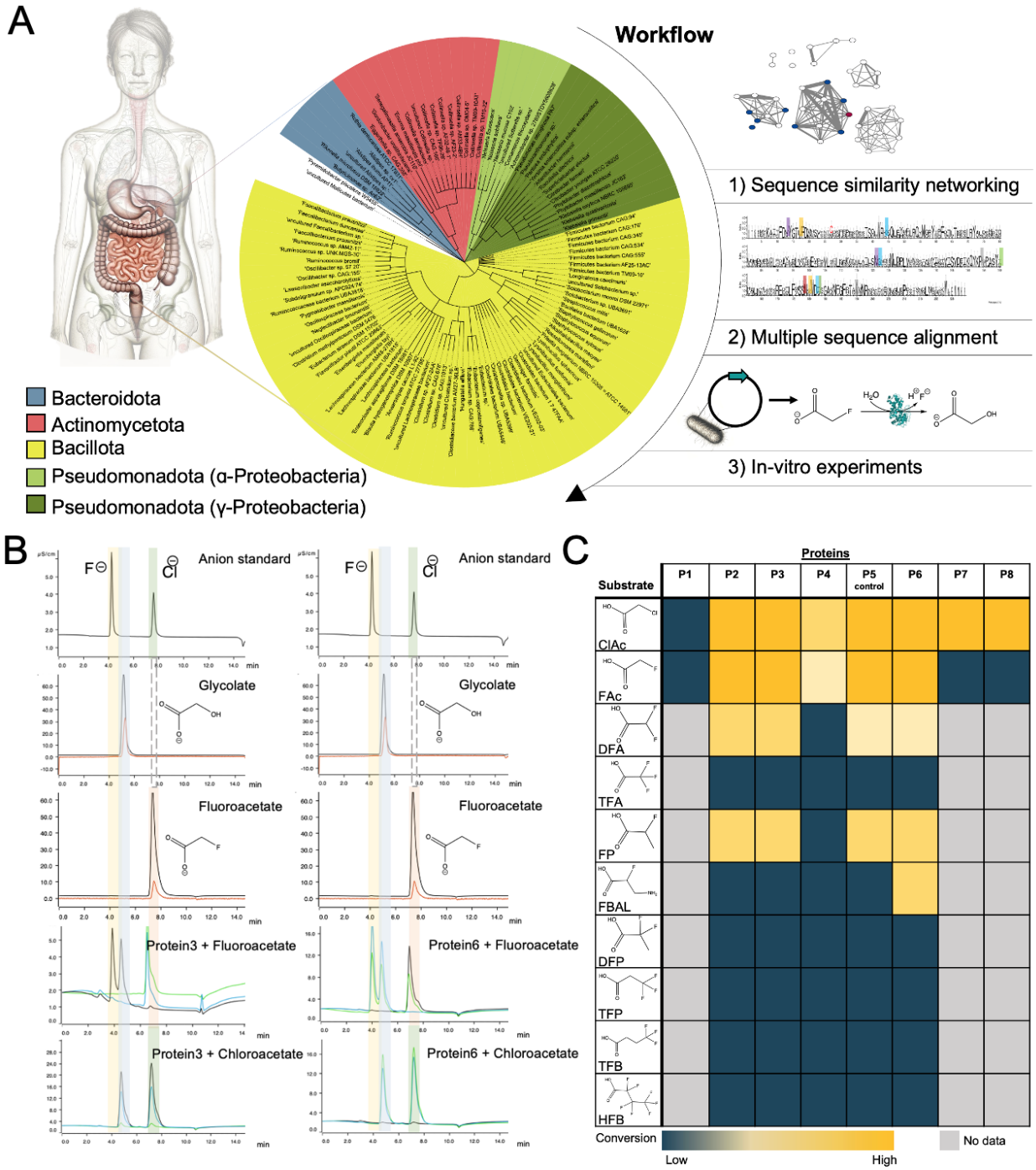

Figure 1. A) Overview of the identification pipeline for the selected haloacid dehalogenases. We reduced initial homology search hits by sequence similarity networking, multiple sequence alignment to eight candidates. B) Ion chromatography traces of standards and the halogenated organic acids tested with protein 3 and protein 6. C) Substrates specificity of ten different substrates with the eight selected proteins. Yellow: active; light yellow: medium

active; blue: non-active; gray: not tested. ClAc: Chloroacetate; FAc: Fluoroacetate, DFA: Difluoroacetate; TFA: Trifluoroacetate; FP: Fluoropropionate; FBAL: α-fluoro-β-alanine; DFP: Difluoropropionate; TFP: Trifluoropropionate; TFpent: Trifluoropentonate; HFB: Heptafluorobutanoate

Molecular dynamics of defluorinating and non-defluorinating enzymes reveals range of conformational compactness

Previously, MD simulations showed that HADs with defluorination activity sampled more compact conformations in the halide binding pocket relative to non-defluorinating enzymes[15]. Since the HADs identified here from the human gut microbiota were not tested in this modeling scheme, we first replicated the molecular dynamics (MD) simulations previously described[15]. In a subsequent round of simulations, we then increased the simulation time from 300 ns to 1 μs using the force field Amber99SB-ILDN instead of OPLS/A because of better performance with protein structures (Fig. S13)[44,45]. Since the original MD publication[15], major advancements have been made in protein structure prediction using AlphaFold (AF)[46,47], even enabling accurate MD simulations using relaxed AF models[48]. Thus, for proteins where crystal structures were not available, we used AF to generate models (Fig. S14). In a benchmark comparison with Chan et al., we obtained generally comparable results in MD simulations of the available X-ray crystal structures and their respective AF models (Fig. S15). Simulating all 11 available defluorinating dehalogenases and 7 non-defluorinating dehalogenases in water for 1 μs, while tracking the size of the halide binding site, did not yield any significant differences between the two groups (Fig. 2A). Further analysis on all conserved amino acids of HADs and their relative distances between each other also showed no significant differences (Fig. S16; Table S4).

Phylogenetics and alignment-independent motifs do not differentiate defluorinating and non-defluorinating enzymes

To further investigate methods to differentiate defluorinating and non-defluorinating enzymes, we generated a phylogenetic tree of the HAD-like family protein sequences (Fig. S17). The tree topology revealed that the defluorinating and non-defluorinating enzymes did not form separate or easily-distinguishable clades. The dehalogenating enzymes originated from eight diverse bacterial and archaeal phyla (Fig. S18) which do not group based on the taxonomy of the source organisms.

Active site variants convert inactive enzyme to dechlorinating enzyme

As neither MD simulations nor phylogenetic analysis yielded conclusive results to distinguish defluorinating from non-defluorinating HADs, we next sought to experimentally test the role of specific residues in the active site. These residues were identified by comparing the non-defluorinating enzyme P1 (*E. aldenensis)* with the active defluorinase from P6 (*G. tenuis)*. Based on comparison between the active sites of the two enzymes (68% sequence similarity, bitscore 268), there were only four residues which were significantly different in the general active site and binding pocket (Fig. S19). We predicted these residues may provide substrate specificity for defluorination, therefore we first aimed for rational engineering of the non-defluorinating enzyme P1 (*E. aldenensis*) via site-directed mutagenesis to generate single, double, and triple variant

combinations as well as the full quadruple variant P1 (F541, T57M, H74F, F187S) (Fig. S20-S23, Table S5).

Among the single variants tested, only one (F54I) showed dehalogenation activity. The F54I variant hydrolyzed chloroacetate but did not hydrolyze fluorinated substrates. Based on structural modeling, F54 is positioned at the entrance channel to the active pocket. In the non-dehalogenating protein P1 (*E. aldenensis*) the larger phenylalanine likely blocks the entrance to the active pocket. In contrast, the dehalogenating protein P6 (*G. tenuis*) features a smaller isoleucine at this position, potentially enabling access for chloroacetate. While we did achieve a functional dehalogenase through the single-point mutation in P1 (*E. aldenensis*), the individual active site changes did not singularly confer defluorination activity. We hypothesized epistatic interactions may promote defluorination. We therefore used the F54I variant as the template for generating double, triple, and quadruple active site variants. In the P1 (F54I, H74F) double variant and all triple variants we achieved improved dechlorination activity (Fig. 2C). Surprisingly, we did not observe defluorination activity in any of the rationally-engineered variants, even though the active site residues were 100% identical to the defluorinating reference enzyme P6 (*G. tenuis*). This result suggests that additional second shell interactions or other conformational changes[49] not localized in the active site may be critical for defluorination activity.

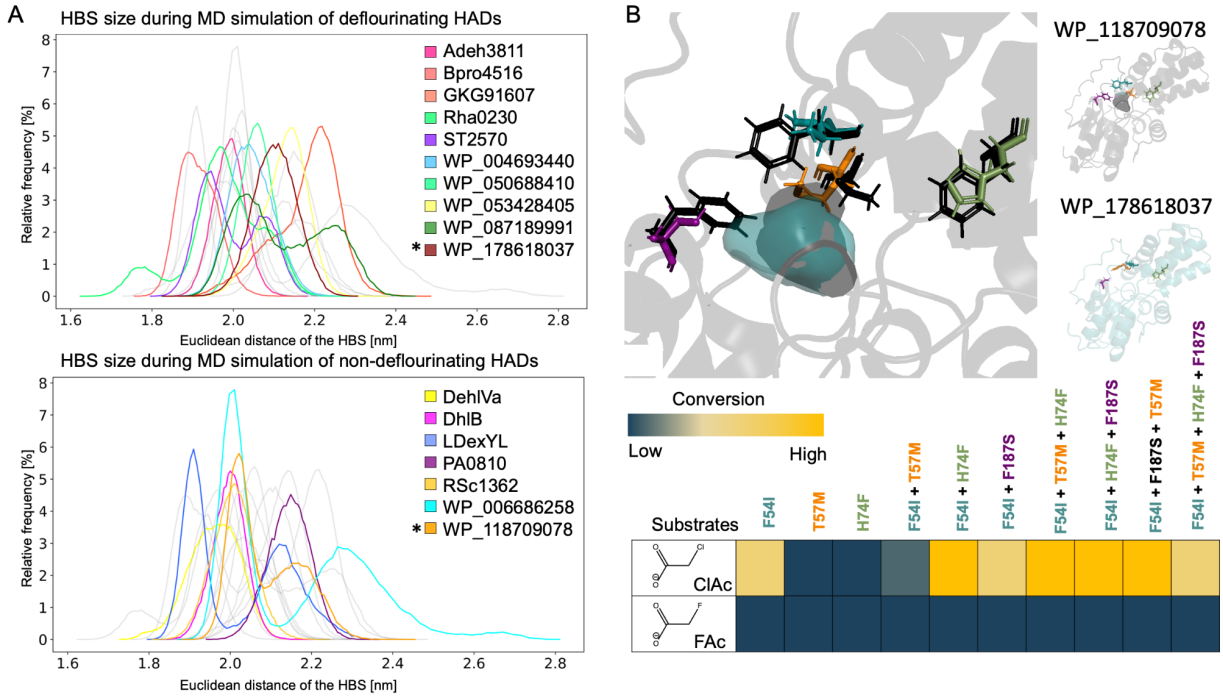

Figure 2. A) Halide binding site (HBS) size during molecular simulation. Distribution of the Euclidean distances of the 3 Cα atoms of the HBS during simulation from 500 ns to 1000 ns. *Asterisks indicate the two enzymes used for protein engineering in this study B) Visualization of the active pocket of the inactive protein P1 (*E. aldensis*) black and the defluorinating protein P6 (WP_178618037, *G. tenuis*) teal. The variants introduced in P1 (*E. aldensis*) are F54I (blue) T57M (orange), H74F (green) and F187S (purple). Substrate specificity of the different protein variants with the two substrates chloroacetate (ClAc) and fluoroacetate (FAc).

431 ***Alanine scanning and a new colorimetric 96-well plate defluorination screening method***
432
433 In order to test the contributions of a wider set of amino acids in defluorination e.g., peripheral
434 residues in addition to active site residues, we used alanine scanning mutagenesis to generate
435 full alanine variant libraries of the proteins: P6 *(G. tenuis*) and P1 (F541, T57M, H74F, F187S) (*E.*
436 *aldensis*). In total, the library consisted of 474 single alanine variants. To screen the alanine
437 library, we devised a miniaturized colorimetric screening method for defluorination to avoid
438 lengthy ion chromatography steps for analysis. Previously, a lanthanum-alizarin assay for
439 microtiter well plates[50] and a xylenol orange assay for an agar plate-based method[38] for colony
440 screening were reported. We tested both the lanthanum-alizarin assay[50] and the new 96-well
441 adaptation of the xylenol orange screening method. Both assays yielded a clear change in
442 absorbance in the presence of free fluoride ions as validated by ion chromatography results. Due
443 to its enhanced color intensity, we report here the results of screening using the xylenol orange
444 assay[38]. As the xylenol orange assay has not been described previously for use in liquid, microtiter
445 well-plate format, a detailed methods section and list of interfering substances is included (Fig.
446 S24-S25, Table S6).
447
448 For method testing and validation, we selected eleven new dehalogenases spanning eleven
449 additional gut microbial genera (Table S7) in addition to the seven dehalogenases initially
450 included in this study. Through this rapid screening assay, we detected a new gut microbial
451 defluorinase from *Priestia megaterium* shared less than 35% amino acid identity to the nearest
452 validated reference enzyme (PDB ID: 3UM9). We purified the *P. megaterium* enzyme to
453 homogeneity and validated its defluorination activity using ion chromatography (Fig. S26). This
454 highlights the potential of the 96-well colorimetric assays to screen for new defluorinating
455 enzymes.
456
457 With our new 96-well defluorination assay in hand, we screened the alanine scanning library with
458 a dual-assay approach for both defluorination and dechlorination activity (Fig. 3A). Across
459 independent biological replicates of the alanine scanning library (Methods), 39.8-41.7% of alanine
460 variants lost defluorination activity. A slightly lower proportion of variants (33.0-33.5%) lost
461 dechlorination activity. We identified 64 alanine variants which showed dechlorination activity but
462 lost defluorination activity (Fig. S27-S29).
463
464 ***Statistical machine learning highlights the importance of the C-terminal region***
465
466 To analyze our alanine scanning data, we used the random forest algorithm to train statistical
467 machine learning models to predict defluorination activity. This algorithm is suitable for low-data
468 regimes and was employed for both regression and classification tasks. Our regression model
469 trained on 80% of the alanine scanning data was capable of predicting quantitative changes in
470 defluorination activity as the result of single alanine point mutations with the remaining 20% of the
471 data with a root-mean square error RMSE of 0.2 ± 0.02 enzyme activity units (on a normalized
472 activity scale between 0-1 defluorination activity). Based on the relative feature importance, the
473 most predictive feature was the relative position of each amino acid within the protein sequence
474 rather than the size, charge, identity, codon diversity, or polarity of each amino acid converted to

alanine (Fig. 3B). These results suggested that the amino acid location within the protein had the highest importance for defluorination activity corresponding to 'hotspots' visible on the protein structure (Fig. 3C).

To train a more generalizable model with applications beyond our selected alanine scanning libraries P1 (F541, T57M, H74F, F187S) and P6, we next trained a random forest classifier to predict defluorinating vs. non-defluorinating enzymes based on general amino acid features extendable to untested HAD sequences (Fig. 3D). We obtained 95% training and test set classification accuracy (average F1-score = 0.93, Table S8). This classification analysis further enabled us to pinpoint residues within the entire HAD family that were most important for defluorination activity (Fig. S30). Strikingly, the top five most important residues for defluorination activity prediction accuracy were localized in the C-terminus (final 41 amino acids). Moreover, a simplified model trained on only the 41 C-terminal residues as features still achieved 83% testing and training set prediction accuracy (average F1-score = 0.73, Table S8).

Additionally, we utilized our molecular dynamics (MD) simulation of P6 *(G. tenuis*) to compute the root mean square fluctuations (RMSF), allowing us to compare the spatial fluctuations of the critical residues within the protein structure. Our analysis identified two regions exhibiting high RMSF values. Both regions are situated at the channel entry leading to the active site (Fig. S31), with one of them located at the C-terminus (Fig. 3E). This finding aligns with the residues implicated in defluorination, as determined by both machine learning predictions and experimental evidence.

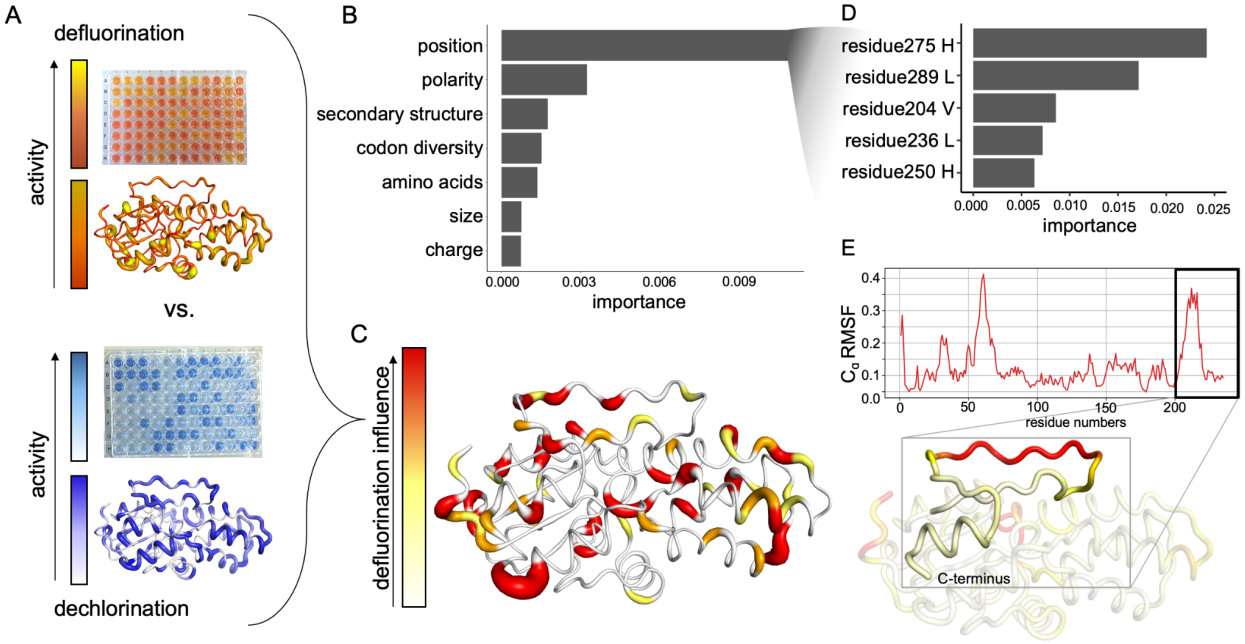

Figure 3. A) Results of the colorimetric assays for defluorination and dechlorination activity, respectively, mapped onto the P6 protein model. The color gradient from white (low activity) to yellow (medium activity) to red (high activity) and line thickness is used to visualize activity differences based on location in the wild-type protein structure. B) Mean relative feature importance plot extracted from random forest regression model using alanine scanning data to predict

defluorination activity. C) Protein 'hotspots' important for defluorination and dechlorination based on the alanine scanning data, where line thickness and coloring corresponds to a higher importance for defluorination alone. D) Top five residue positions with the highest mean gini importance in random forest classification for defluorinating vs. non-defluorinating proteins E) Root mean square fluctuation (RMSF reported in nm) of the Cα atoms indicates high movement in the C-terminal region of the protein.

Chimeric defluorinases indicate importance of the C-terminal region

To test the relevance of the C-terminal region experimentally, we generated chimeric proteins by combining different regions from the non-defluorinating protein P1 (*E. aldenensis)* with the defluorinating enzyme P6 (*G. tenuis)* (Fig. S32). Strikingly, the single P1 (*E. aldensis*) chimera generated by swapping in the 41 amino acid C-terminus of the P6 (*G. tenuis)* enzyme showed slight defluorination activity (Fig. 4A). All other chimeric proteins were active in dechlorination, but not defluorination (Fig. 4A, Fig. S32). Notably, the C-terminal region has not been previously reported for its importance in dehalogenation or fluorinated compound substrate specificity. Analyzing our alanine scanning data based on our chimeric segments (Fig. 4B), we observed a pattern in the activity distribution of defluorination activity (Fig. 4B). Based on our MD simulations, we observed that the C-terminal chimeric segment also had the highest overall movement relative to other protein chimeric segments (Fig. 4C). Overall, the combination of alanine scanning and chimeric mutagenesis indicated that the C-terminal 41 amino acid segment was both necessary and sufficient to install defluorination activity in a non-defluorinating enzyme.

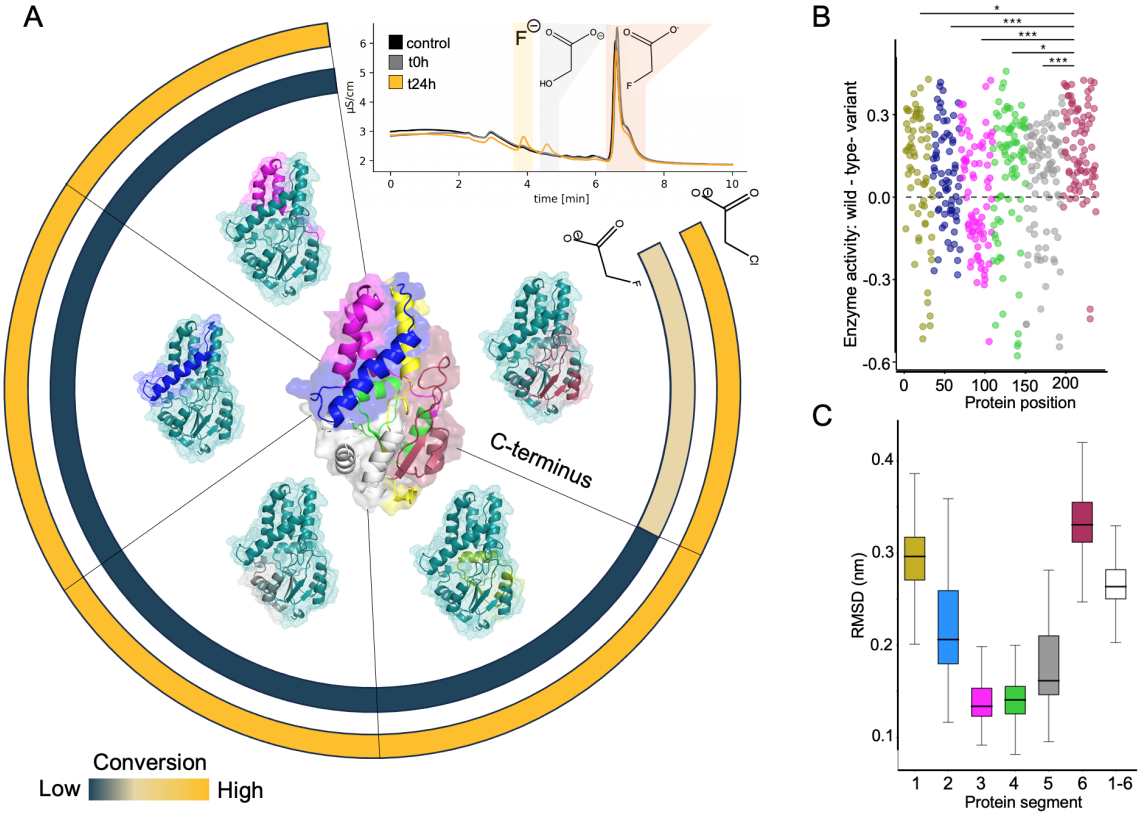

Figure 4. A) Activity of the chimeras generated from low conversion (blue) to high conversion (gold) of two substrates chloroacetate (outer ring) and fluoroacetate (inner ring). Inset of the ion chromatogram shows an engineered chimeric protein with a 'swapped' C-terminal segment enables weak conversion of fluoroacetate to glycolate and fluoride ion.

B) Protein defluorination activity of the alanine scanning library with all data points shown for two independent biological replicates. The x-axis indicates the position in the protein and the y-axis indicates the change in normalized activity for each alanine scanning variant relative to wild-type. Points are colored by chimeric segments as depicted in other figure panels. The C-terminal chimeric segment exhibits a significantly different distribution of active to inactive alanine variants from the other chimeric segments of the protein ($p<0.05$, Kruskal-Wallis test). Only one variant (L231A) in the C-terminal segment displays higher activity than wild-type. C) Boxplot of root mean squared distance (RMSD) in nm for different protein segments derived from MD simulations with the C-terminal chimeric segment exhibiting the highest overall movement.

Discussion

Here we demonstrated that HAD family enzymes capable of carbon-fluorine bond cleavage are widely distributed across bacterial phyla common in the human gut. To our knowledge, this is the first report of enzymes sourced from the human gut microbiota capable of catalyzing this reaction. Despite sharing sequence and structural similarity, the proteins in this study exhibited different substrate specificity including defluorination activity for fluoroacetate and α-fluoro-β-alanine. These are known degradation products of 5-FU, a widely-used cancer medication on the World Health Organization's list of Essential Medicines[51]. Fluorocitrate, fluoroacetate, and α-fluoro-β-alanine resulting from the metabolism of 5-FU were previously shown to drive the variable and severe side effects of 5-FU treatment, including neurotoxicity[6,52] and cardiotoxicity[53,54].

In this study, we also identified four new HADs active on difluoroacetate, doubling the number of hydrolytic dehalogenases experimentally shown to defluorinate -CF_2 groups[55]. Difluoroacetate is a transformation product of various xenobiotics, such as the refrigerant chlorotrifluoroethylene (R1113)[56] or the anesthetic 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-ethane[57]. Difluoroacetate can also enter our system through the ingestion of food[58]. It is a byproduct of the photodegradation of penoxsulam, a widely used pesticide[58]. These substances enter our environment and can seep into groundwater, where they are subsequently consumed by humans, thus coming in contact with the gut microbiome. It has been shown that long-term ingestion of per- and polyfluoroalkyl substances (PFAS) including pharmaceuticals containing -CF_3 groups alter the host microbiome[17]. It is tempting to speculate that long-term ingestion of fluorinated compounds may also contribute to the abundance and activity of HADs we identified encoded in the human gut microbiome.

Defining residues or structural features that differentiate defluorinating enzymes from their non-defluorinating counterparts in the HAD family has remained elusive[15]. Phylogenetic analysis and molecular dynamic simulations of an extended HAD set yielded inclusive evidence to distinguish between the two groups. We identified two HADs sharing 68% sequence similarity yet one was active in defluorination and dechlorination while the other was completely inactive. We engineered the inactive enzyme to dechlorinate by introducing a single point mutation at the entrance to the catalytic pocket but no additional point mutations conferred defluorination activity.

We next utilized alanine scanning to identify protein variants that retained dechlorination activity but lost their defluorination capability and identified several amino acids critical to defluorination. Surprisingly, these amino acids were primarily located in the peripheral regions of the protein with

a higher percentage of critical residues located in the C-terminal segment of the protein. These observations, along with results from more recent studies[59], support a more complex and dynamic understanding of protein function than the simplistic and traditional lock-and-key model[60]. Consequently, our protein engineering efforts shifted towards engineering chimeric proteins by replacing segments of non-defluorinating protein variants with a defluorinating enzyme. Exclusively when the C-terminal segment was swapped, we observed chimeric proteins with weak, but robust and reproducible defluorination activity.

The C-terminal regions of proteins have been extensively studied in eukaryotic cells[61], where they have been shown to be responsible for a broad spectrum of functions, including induced fit of substrates[61,62] and protein-protein interactions[63]. This functional diversity is attributed to their solvent exposure and higher degree of disorder, which facilitate a more dynamic and flexible secondary structure[64]. In a prokaryotic HAD family phosphatase, the flexible C-terminal tail was previously shown to interact with the active site of an adjacent monomer and play a critical regulatory role in product release[49]. Our MD simulations suggest the C-termini exhibits greater flexibility compared to the rest of the regions in the protein structure in the HAD-like defluorinases. Taken together with the observed defluorination exclusively in the C-terminal chimera, we propose the C-terminal region of HAD-family dehalogenases also play a significant and previously unreported role in defluorination activity. This knowledge lays the foundation to engineer proteins with expanded substrate ranges for carbon-fluorine bond cleavage applications in human and environmental health.

Data and code availability

Scripts associated with analysis presented in this manuscript are available at: https://github.com/MSM-group/biodefluorination-paper.

Acknowledgements

We thank the entire Joint Genome Institute team including Miranda Harmon-Smith, Yasuo Yoshikuni, and co-authors Ian Blaby and Ritesh Mewalal for their exceptional support. The work (proposal: 10.46936/10.25585/60008420) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. S.P. was supported by funding from the Helmut Horten Foundation. V.P. and S.L.R. acknowledge funding from the Swiss National Science Foundation (PZPGP2_209124). We acknowledge René Gall for his technical support with analytics. We acknowledge Larry Wackett for thoughtful comments and discussions on the manuscript.

619 **References**

620 1. Current Contributions of Organofluorine Compounds to the Agrochemical Industry. *iScience*
621 **23**, 101467 (2020).
622 2. Inoue, M., Sumii, Y. & Shibata, N. Contribution of Organofluorine Compounds to
623 Pharmaceuticals. *ACS Omega* **5**, 10633–10640 (2020).
624 3. Kotthoff, M., Müller, J., Jürling, H., Schlummer, M. & Fiedler, D. Perfluoroalkyl and
625 polyfluoroalkyl substances in consumer products. *Environ. Sci. Pollut. Res.* **22**, 14546–14559
626 (2015).
627 4. Phelps, D. W., Parkinson, L. V., Boucher, J. M., Muncke, J. & Geueke, B. Per- and
628 Polyfluoroalkyl Substances in Food Packaging: Migration, Toxicity, and Management
629 Strategies. *Environ. Sci. Technol.* **58**, 5670–5684 (2024).
630 5. Pan, Y. The Dark Side of Fluorine. *ACS Med. Chem. Lett.* **10**, 1016–1019 (2019).
631 6. Okeda, R. *et al.* Experimental neurotoxicity of 5-fluorouracil and its derivatives is due to
632 poisoning by the monofluorinated organic metabolites, monofluoroacetic acid and alpha-
633 fluoro-beta-alanine. *Acta Neuropathol.* **81**, 66–73 (1990).
634 7. Koppel, N., Maini Rekdal, V. & Balskus, E. P. Chemical transformation of xenobiotics by the
635 human gut microbiota. *Science* **356**, (2017).
636 8. Lehouritis, P. *et al.* Local bacteria affect the efficacy of chemotherapeutic drugs. *Sci. Rep.* **5**,
637 1–12 (2015).
638 9. Geller, L. T. *et al.* Potential role of intratumor bacteria in mediating tumor resistance to the
639 chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).
640 10. Spanogiannopoulos, P. *et al.* Host and gut bacteria share metabolic pathways for anti-cancer
641 drug metabolism. *Nat Microbiol* **7**, 1605–1620 (2022).
642 11. Jennings, B. A. *et al.* Evaluating predictive pharmacogenetic signatures of adverse events in
643 colorectal cancer patients treated with fluoropyrimidines. *PLoS One* **8**, e78053 (2013).
644 12. Walker, M. C. & Chang, M. C. Y. Natural and engineered biosynthesis of fluorinated natural

645 products. *Chem. Soc. Rev.* **43**, 6527–6536 (2014).

646 13. Wackett, L. P. Evolutionary obstacles and not C-F bond strength make PFAS persistent.
647 *Microb. Biotechnol.* **17**, e14463 (2024).

648 14. Key, B. D., Howell, R. D. & Criddle, C. S. Fluorinated organics in the biosphere. *Environ. Sci.*
649 *Technol.* **31**, 2445–2454 (1997).

650 15. Chan, P. W. Y. *et al.* Defluorination Capability of l-2-Haloacid Dehalogenases in the HAD-
651 Like Hydrolase Superfamily Correlates with Active Site Compactness. *Chembiochem* **23**,
652 e202100414 (2022).

653 16. Chan, W. Y. *et al.* Sequence- and activity-based screening of microbial genomes for novel
654 dehalogenases. *Microb. Biotechnol.* **3**, 107–120 (2010).

655 17. Rashid, F., Dubinkina, V., Ahmad, S., Maslov, S. & Irudayaraj, J. M. K. Gut Microbiome-Host
656 Metabolome Homeostasis upon Exposure to PFOS and GenX in Male Mice. *Toxics* **11**,
657 (2023).

658 18. Hiseni, P., Rudi, K., Wilson, R. C., Hegge, F. T. & Snipen, L. HumGut: a comprehensive
659 human gut prokaryotic genomes collection filtered by metagenome data. *Microbiome* **9**, 165
660 (2021).

661 19. Madeira, F. *et al.* Search and sequence analysis tools services from EMBL-EBI in 2022.
662 *Nucleic Acids Res.* **50**, W276–W279 (2022).

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

666 21. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular
667 interaction networks. *Genome Res.* **13**, 2498–2504 (2003).

668 22. Oberortner, E., Cheng, J.-F., Hillson, N. J. & Deutsch, S. Streamlining the Design-to-Build
669 Transition with Build-Optimization Software Tools. *ACS Synth. Biol.* **6**, 485–496 (2017).

670 23. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of

671 protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).

672 24. Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat. Methods* **19**, 679–
673 682 (2022).

674 25. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for
675 molecular dynamics. *PLoS Comput. Biol.* **13**, e1005659 (2017).

676 26. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-
677 level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).

678 27. Bauer, P., Hess, B. & Lindahl, E. GROMACS 2022.5 Manual.
679 https://doi.org/10.5281/ZENODO.7586765 (2023).

680 28. Lindorff-Larsen, K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein
681 force field. *Proteins* **78**, 1950–1958 (2010).

682 29. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison
683 of simple potential functions for simulating liquid water. *J. Chem. Phys.* **79**, 926–935 (1983).

684 30. Hess, B., Bekker, H., Berendsen, H. J. C. & Fraaije, J. G. E. LINCS: A linear constraint solver
685 for molecular simulations. *J. Comput. Chem.* **18**, 1463–1472 (1997).

686 31. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–8593
687 (1995).

688 32. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J.*
689 *Chem. Phys.* **126**, 014101 (2007).

690 33. Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R.
691 Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690 (1984).

692 34. GROMACS 2024.1 Manual. doi:10.5281/zenodo.10721192.

693 35. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids*
694 *Res.* **51**, D523–D531 (2023).

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

697 37. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular
698 biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
699 38. Davis, C. K., Denman, S. E., Sly, L. I. & McSweeney, C. S. Development of a colorimetric
700 colony-screening assay for detection of defluorination by micro-organisms. *Lett. Appl.*
701 *Microbiol.* **53**, 417–423 (2011).
702 39. Lemaire, L., Malet-Martino, M. C., de Forni, M., Martino, R. & Lasserre, B. Cardiotoxicity of
703 commercial 5-fluorouracil vials stems from the alkaline hydrolysis of this drug. *Br. J. Cancer*
704 **66**, 119–127 (1992).
705 40. Oren, A. & Garrity, G. M. Valid publication of the names of forty-two phyla of prokaryotes. *Int.*
706 *J. Syst. Evol. Microbiol.* **71**, 005056 (2021).
707 41. Gerlt, J. A. *et al.* Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool
708 for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **1854**, 1019–
709 1037 (2015).
710 42. Burroughs, A. M., Allen, K. N., Dunaway-Mariano, D. & Aravind, L. Evolutionary genomics of
711 the HAD superfamily: understanding the structural adaptations and catalytic diversity in a
712 superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.* **361**, 1003–1034 (2006).
713 43. Gross, C., Felsheim, R. & Wackett, L. P. Genes and enzymes of azetidine-2-carboxylate
714 metabolism: detoxification and assimilation of an antibiotic. *J. Bacteriol.* **190**, 4859–4864
715 (2008).
716 44. Lindorff-Larsen, K. *et al.* Systematic validation of protein force fields against experimental
717 data. *PLoS One* **7**, e32131 (2012).
718 45. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with
719 respect to force field parameterization? *Biophys. J.* **100**, L47–9 (2011).
720 46. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold
721 3. *Nature* (2024) doi:10.1038/s41586-024-07487-w.
722 47. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,

723 583–589 (2021).

724 48. Guo, H.-B. *et al.* AlphaFold2 modeling and molecular dynamics simulations of an intrinsically
725 disordered protein. *PLoS One* **19**, e0301866 (2024).

726 49. Biswas, T. *et al.* The tail of KdsC: conformational changes control the activity of a haloacid
727 dehalogenase superfamily phosphatase. *J. Biol. Chem.* **284**, 30594–30603 (2009).

728 50. Bygd, M. D., Aukema, K. G., Richman, J. E. & Wackett, L. P. Microwell Fluoride Screen for
729 Chemical, Enzymatic, and Cellular Reactions Reveals Latent Microbial Defluorination
730 Capacity for -CF Groups. *Appl. Environ. Microbiol.* **88**, e0028822 (2022).

731 51. Miura, K. *et al.* 5-fu metabolism in cancer and orally-administrable 5-fu drugs. *Cancers* **2**,
732 1717–1730 (2010).

733 52. Yamashita, K., Yada, H. & Ariyoshi, T. Neurotoxic effects of alpha-fluoro-beta-alanine (FBAL)
734 and fluoroacetic acid (FA) on dogs. *J. Toxicol. Sci.* **29**, 155–166 (2004).

735 53. Muneoka, K., Shirai, Y., Yokoyama, N., Wakai, T. & Hatakeyama, K. 5-Fluorouracil
736 cardiotoxicity induced by alpha-fluoro-beta-alanine. *Int. J. Clin. Oncol.* **10**, 441–443 (2005).

737 54. Franck, C., Malfertheiner, P. & Venerito, M. Safe administration of S-1 after 5-fluorouracil-
738 induced cardiotoxicity in a patient with colorectal cancer. *BMJ Case Rep.* **2017**, (2017).

739 55. Khusnutdinova, A. N. *et al.* Structural insights into hydrolytic defluorination of difluoroacetate
740 by microbial fluoroacetate dehalogenases. *FEBS J.* (2023) doi:10.1111/febs.16903.

741 56. Dorgerloh, U., Becker, R. & Kaiser, M. Evidence for the Formation of Difluoroacetic Acid in
742 Chlorofluorocarbon-Contaminated Ground Water. *Molecules* **24**, (2019).

743 57. Herbst, J., Köster, U., Kerssebaum, R. & Dekant, W. Role of P4502E1 in the metabolism of
744 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-ethane. *Xenobiotica* **24,** 507–516 (1994).

745 58. Bhat, A. P., Pomerantz, W. C. K. & Arnold, W. A. Finding Fluorine: Photoproduct Formation
746 during the Photolysis of Fluorinated Pesticides. *Environ. Sci. Technol.* **56**, 12336–12346
747 (2022).

748 59. Uversky, V. N. & Kulkarni, P. Intrinsically disordered proteins: Chronology of a discovery.

Biophys. Chem. **279**, 106694 (2021).

60. James, L. C. & Tawfik, D. S. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).
61. Sharma, S. & Schiller, M. R. The carboxy-terminus, a key regulator of protein function. *Crit. Rev. Biochem. Mol. Biol.* **54**, 85–102 (2019).
62. Sargeant, D. P. *et al.* Secondary structure, a missing component of sequence-based minimotif definitions. *PLoS One* **7**, e49957 (2012).
63. Beuming, T., Skrabanek, L., Niv, M. Y., Mukherjee, P. & Weinstein, H. PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics* **21**, 827–828 (2005).
64. Lobanov, M. Y., Furletova, E. I., Bogatyreva, N. S., Roytberg, M. A. & Galzitskaya, O. V. Library of disordered patterns in 3D protein structures. *PLoS Comput. Biol.* **6**, e1000958 (2010).