Enzymatic carbon-fluorine bond cleavage by human gut microbes

Silke I. Probst[1], Florian D. Felder[1], Victoria Poltorak[1,2], Ritesh Mewalal[3], Ian K. Blaby[3], Serina L. Robinson[1*]

[1]Department of Environmental Microbiology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), Dübendorf, Switzerland; [2]Department of Biology, Institute of Microbiology, Eidgenössische Technische Hochschule (ETH) Zürich, Zürich, Switzerland; [3]U.S. Department of Energy Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
*corresponding author: serina.robinson@eawag.ch

Significance

Humans have introduced carbon-fluorine bonds into numerous manufactured compounds, including pharmaceuticals, leading to the formation of toxic fluorinated byproducts. While the human gut microbiota is known for its ability to metabolize drugs, its encoded capacity to break the strong carbon-fluorine chemical bond was previously unknown. Here we discovered that human gut microbial enzymes are capable of cleaving carbon-fluorine bonds. We developed a 96-well colorimetric fluoride assay amenable to bacterial culture-based screening. We additionally conducted whole-protein alanine scanning mutagenesis and identified through machine learning that flexible C-terminal loop residues were predictive of defluorination. Taken in the context of flexible regions of other enzyme families known to perform fluorine chemistry, this work supports using convergent structural features to predict defluorination specificity.

Abstract

Fluorinated compounds are used for agrochemical, pharmaceutical, and numerous industrial applications, resulting in global contamination. In many molecules, fluorine is incorporated to enhance the half-life and improve bioavailability. Fluorinated compounds enter the human body through food, water, and xenobiotics including pharmaceuticals, exposing gut microbes to these substances. The human gut microbiota is known for its xenobiotic biotransformation capabilities, but it was not previously known whether gut microbial enzymes could break carbon-fluorine bonds, potentially altering the toxicity of these compounds.

Here, through the development of a rapid, miniaturized fluoride detection assay for whole-cell screening, we discovered active gut microbial defluorinases. We biochemically characterized enzymes from diverse human gut microbial classes including Clostridia, Bacilli, and Coriobacteriia, with the capacity to hydrolyze (di)fluorinated organic acids and a fluorinated amino acid. Whole-protein alanine scanning, molecular dynamics simulations, and chimeric protein design enabled the identification of a disordered C-terminal protein segment involved in defluorination activity. Domain swapping exclusively of the C-terminus conferred defluorination activity to a non-defluorinating dehalogenase. To advance our understanding of the structural and sequence differences between defluorinating and non-defluorinating dehalogenases, we trained machine learning models which identified protein termini as important features. Models trained on 41-amino acid segments from protein C-termini alone predicted defluorination activity with 83% accuracy (compared to 95% accuracy based on full-length protein features). This work is relevant

for therapeutic interventions and environmental and human health by uncovering specificity-determining signatures of fluorine biochemistry from the gut microbiome.

Introduction

Over 20% of pharmaceuticals and 50% of agrochemicals released in the past 25 years contain one or more fluorine atom(s) (1, 2). 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 (3). 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- (4, 5) and cardiotoxicity (6, 7).

The gut microbiota is known to modify and thereby inactivate or increase the toxicity of various xenobiotics (8, 9) including fluorinated compounds. Gut microbes increase the cytotoxicity of the fluorinated antineoplastic prodrug fludarabine (10) and modify the fluorinated chemotherapeutic gemcitabine through acetylation (10) or deamination (11). Anti-cancer fluoropyrimidine drugs including capecitabine and 5-FU are inactivated by gut bacteria through a pathway that mimics host metabolism (9, 12). Microbiome-dependent pharmacokinetics (9) have been linked to interindividual differences in the composition and metabolic capabilities of the human gut microbiota (13). This could impact drug activity by increasing toxicity and side effects and may potentially influence the significant interpatient variability in responses to fluorinated compounds. 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 microbes in the human gut (14). Although there are reports of defluorinating enzymes from bacteria isolated from soil (15), ruminants (16), and activated sludge (17), none were previously associated with the human gut microbiota.

Earlier efforts to identify microbes and enzymes with defluorination activity focused on activity-guided enrichment and isolation-based strategies (16, 18). In addition to the high strength of the carbon-fluorine bond, the product of defluorination activity (fluoride anion) is also toxic to many microbes, presenting evolutionary hurdles to positive selection during enrichment and isolation (19, 20). Even when defluorination activity is detected, isolation of the responsible microbes and identification of the responsible enzyme(s) remains difficult. With the rapid growth of sequencing databases, targeted sequence-guided (15) or structure-guided (21) metagenome mining techniques are emerging as alternative approaches to identify defluorinating enzymes.

Although several different enzyme classes show promiscuous activity for fluorinated substrates (20), only hydrolases are known to cleave fluorinated compounds as their native substrates (22). Enzymes capable of hydrolytic defluorination (23) belong to two different superfamilies: the α/β hydrolase superfamily (24) and, more recently, the haloacid dehalogenase-like (HAD) superfamily (25). Although the majority of HADs actually act on a wide range of organic phosphate substrates (26), a subset of HADs are capable of regio- and stereospecific hydrolysis of halogenated (including fluorinated) carboxylic acids with a broader substrate specificity than the α/β hydrolase superfamily (25). The few defluorinating HADs described, primarily from soil, catalyze carbon-

84 halide bond cleavage via a S_N2-like mechanism through nucleophilic attack by an aspartate to
85 displace the halide (25, 27). The active sites of defluorinating and dechlorinating HADs are highly
86 similar (25), thus it remains challenging to mine metagenomic datasets to accurately identify
87 HADs with specificity for fluorinated compounds (21). Previously, molecular dynamics (MD)
88 showed that HADs capable of defluorination sampled more compact halide-binding pocket
89 conformations compared to non-defluorinating HADs (25). Here,e we tested this MD-based
90 classification method and found that it did not generalize well to a larger set of defluorinating and
91 non-defluorinating HADs, which us motivated to search for alternative accessible features to
92 predict metabolic defluorination potential from microbiome data.

93 In this study, we combined metagenome mining with in vitro experimentation to test the activity of
94 over twenty new dehalogenases from the gut microbiome. To overcome the limitations of time-
95 consuming ion chromatography (IC) measurements for fluoride, we developed an easy-to-use
96 microtiter plate screen for defluorination activity. Using this approach, we screened large alanine
97 scanning mutagenesis libraries to identify key amino acids important for defluorination. Notably,
98 we successfully converted a dechlorinating HAD into an active defluorinating HAD by substituting
99 its terminal 41 amino acids with those from a naturally-occurring defluorinating HAD. Molecular
100 dynamics simulations, alanine scanning, and machine learning results converged to support the
101 previously unknown role of the protein carboxyl (C)-terminal region of HADs in defluorination.
102 Swapping this C-terminal region conferred defluorination capability to a non-defluorinating
103 enzyme, thereby providing new molecular features to predict enzyme-substrate specificity for
104 fluorinated compounds from microbiome data.

105 **Results**
106
107 ***The human gut microbiota encodes HAD-like enzymes***
108
109 To investigate the potential for defluorination by human gut microbial enzymes, we first tested
110 defluorination activity for fluoroacetate as a simple fluorinated substrate and a known 5-FU
111 degradation product (28). Notably, prior studies identified the soil bacterium, *Rhodococcus jostii*
112 RHA1, as a degrader of fluoroacetate along with the discovery of its responsible defluorinating
113 HAD superfamily enzyme classified as a L-2-haloacid dehalogenase (15). These enzymes
114 primarily dehalogenate small chlorinated alkanoic acids with only a few demonstrating the
115 capability to cleave the more recalcitrant carbon-fluorine bond (25). We therefore used the
116 *Rhodococcus jostii* RHA1 HAD (Rha0230) for a protein homology search in the HumGut database
117 (29). To our surprise, we detected over 500 HAD homologs encoded in human gut metagenomes
118 (Fig. S1, Table S1).
119
120 The identified HAD-encoding genes were predominantly found in bacteria belonging to the phylum
121 Bacillota, formerly named the Firmicutes (30). However, homologs were also prevalent in the
122 Actinomycetota, Pseudomonadota, and Bacteroidota (Fig. 1A, Fig. S1). To further elucidate the
123 relationships among these proteins, we constructed a protein sequence similarity network (Fig.
124 S2), which revealed several distinct HAD clusters (29, 31). Two of these clusters contained
125 characterized members of the L-2-haloacid dehalogenase subfamily (25) within the larger HAD
126 superfamily (26). Other clusters could be assigned to other HAD superfamily family functions such

127 as phosphonate and azetidine hydrolysis (26, 32). For subsequent analysis, we focused on
128 clusters containing the 2-haloacid dehalogenases and curated the dataset through a multiple
129 sequence alignment to eliminate redundant sequences and confirmed the presence of the
130 catalytic residues (Fig. S3). We selected seven new representatives for in vitro experiments (Fig.
131 S2, Table S2). The selected sequences from the human gut microbiota showed low amino acid
132 sequence identity to previously characterized HADs (18% - 46%; mean: 29%; Table S1). We also
133 investigated the genomic context by extracting ± six genes on either side of the query HADs
134 (Table S3). No CLC^F or Fluc (*crcB*-encoded) fluoride ion transporters or other relevant protein
135 families were detected. Therefore, the role of these HADs in defluorination could not be inferred
136 based on genomic co-localization. The lack of functional prediction based on sequence similarity
137 or genomic context therefore motivated experimental validation of these unknown HADs.
138
139 ***Human gut microbial dehalogenases are active in carbon-fluorine bond cleavage***
140
141 To screen for in vitro defluorination activity, we heterologously expressed, purified, and tested
142 seven selected gut microbial dehalogenases with ten different halogenated substrates (nine
143 fluorinated organic acids and one chlorinated substrate, chloroacetate, Fig. 1, Fig. S4 Table S4).
144 The HAD from the soil bacterium *Rhodococcus jostii* RHA1 (P5, Protein Data Bank (PDB) ID:
145 3UMG) was included as a positive defluorination control as described previously. All proteins
146 tested except for one (*Enterocloster aldenensis*, P1, Fig. 1C) showed dehalogenase activity with
147 at least one substrate (Fig. S5-S12). Two enzymes hydrolyzed chloroacetate but were not active
148 with any fluorinated substrates (P7 and P8, Fig. 1C). The remaining four enzymes hydrolyzed
149 fluoroacetate, producing fluoride ion and glycolate stoichiometrically, as measured by ion
150 chromatography (Fig. 1B). Compounds with perfluorinated methyl ($-CF_3$) or perfluorinated
151 methylene ($-CF_2-$) groups including trifluoroacetate, difluoropropionate, trifluoropropionate,
152 trifluoropentanoate, and heptafluorobutanoate were not accepted by any of the eight HADs tested
153 (Fig. 1C). Four of the enzymes, including the control were active with known transformation
154 products of fluorinated xenobiotics, namely, 2-fluoropropionate, α-fluoro-β-alanine, and the
155 polyfluorinated compound, difluoroacetate. These active defluorinating enzymes originate from
156 diverse gut bacterial families including Eggerthellaceae (*Gordonibacter*, P2 and P3),
157 Moraxellaceae (*Acinetobacter*, P4), and Christensellaceae (*Guopingia*, P6).
158
159 We further examined the impact of oxygen on one defluorinating enzyme P3 as a representative
160 for our gut microbial HADs. As expected, based on the established L-2-haloacid dehalogenase
161 mechanism (24–26), P3 cleaved fluoroacetate to produce glycolate and fluoride ions under both
162 aerobic and anaerobic conditions (Fig. S13). To test whether we could detect the defluorination
163 activity of the gut bacterium *Gordonibacter pamelaeae* anaerobically in gut microbial medium
164 (GMM) with fluoroacetate and measured fluoride concentrations using a fluoride ion electrode
165 (Table S5, Methods). Defluorination activity was not detectable, perhaps due to challenges in
166 expression, transport, or the preferential use of other substrates by *G. pamelaeae*. Due to the
167 limited genetic tractability of the native gut strains encoding these enzymes, we instead focused
168 efforts on in vitro enzyme characterization. We measured the steady-state enzyme kinetics for P3
169 with fluoroacetate and compared our results to other defluorinating enzymes including a L-2-
170 haloacid dehalogenase (DeHa2) and fluoroacetate dehalogenases (DeHa4) from *Delftia*

171 *acidovorans* D4B (33) and the fluoroacetate dehalogenase from *Rhodopseudomonas palustris*
172 (RPA1163, (34)). Briefly, P3 has a K_m of 59.8 ± 7.7 mM approximately one order of magnitude
173 higher than the K_m values of DeHa2 and DeHa4 (Table S6), although lower than the K_m of
174 RPA1163 (86.8 ± 13.7 mM)(34). The k_{cat} of 0.25 ± 0.02 s^{-1} is about fourfold lower than the k_{cat} of
175 RPA1163 (Table S6). The weak substrate binding and correspondingly high off rate suggested
176 by the *Km* may be a contributing factor to the overall low catalytic efficiency.
177

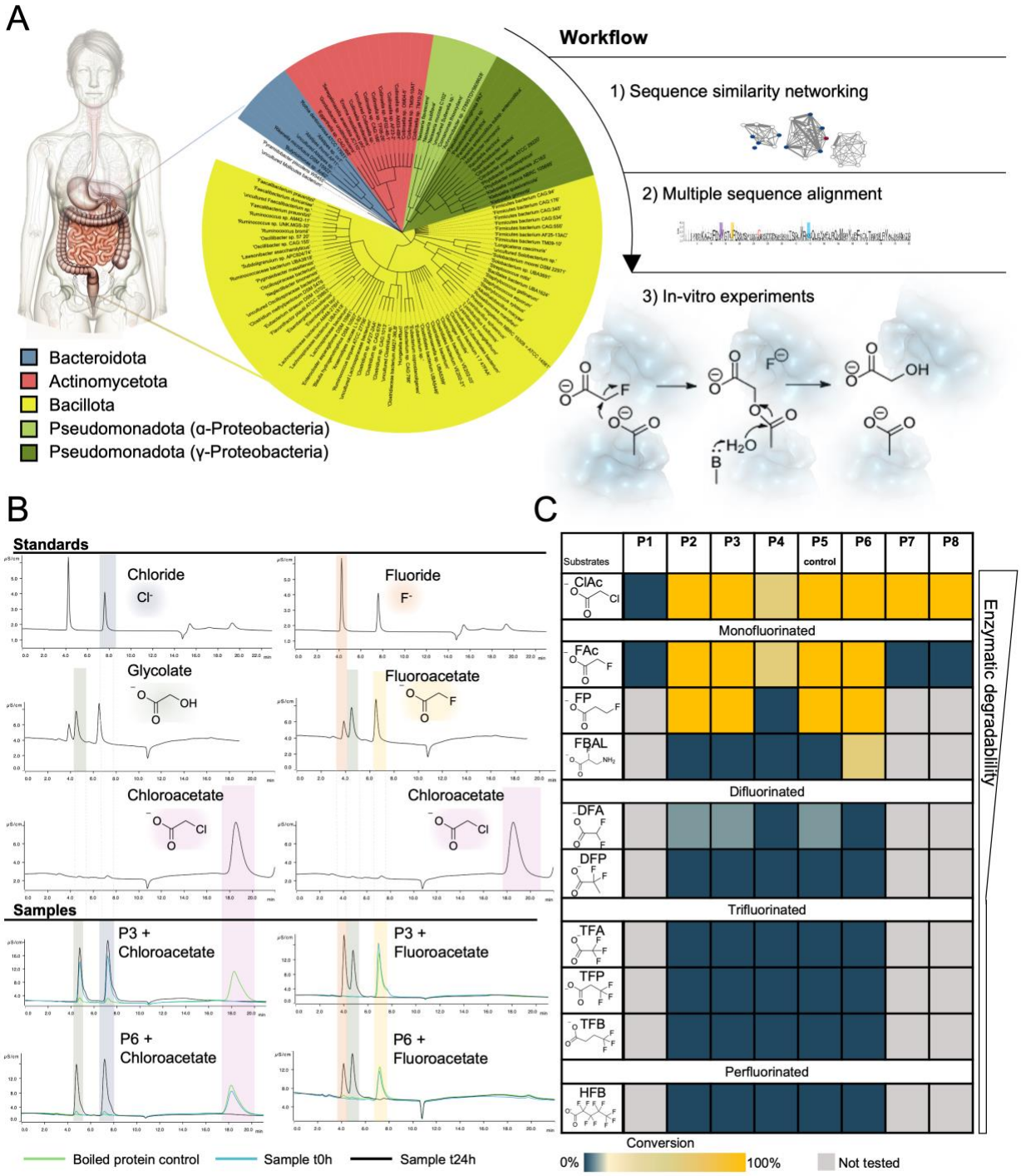

Figure 1. Haloacid dehalogenases are widespread across human gut bacterial phyla and exhibit broad substrate specificity. A) Overview of the identification pipeline for the selected haloacid dehalogenases. We reduced initial homology search hits using sequence similarity networking and multiple sequence alignment, resulting in eight candidates. B) Ion chromatography traces of standards colored for distinction. Blue shade: Chloride; Green shade: Glycolate; Pink shade: Chloroacetate; Red shade: Fluoride; Yellow shade: Fluoroacetate. The last four traces show protein 3 and protein 6 tested with fluoroacetate and chloroacetate. Green line: boiled protein control; Blue line: sample

at t0; Black: sample after 24h incubation. The different standard peaks are color coded and dotted lines are extending the colored shading towards the sample peaks. C) Substrates specificity of ten different substrates with the eight selected proteins. The substrate conversion rate ranges from dark blue (0% conversion) to Yellow (100% conversion) (Fig. S5-S12) gray: not tested. ClAc: Chloroacetate; FAc: Fluoroacetate, FP: Fluoropropionate; FBAL: α-fluoro-β-alanine; DFA: Difluoroacetate; TFA: Trifluoroacetate; DFP:Difluoropropionate; TFP: Trifluoropropionate; TFB: Trifluorbutanate; HFB: Heptafluorobutanoate. Predicted enzymatic degradability of the substrates based on literature is indicated by the broadness of the bar.

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 (25). Since the HADs identified here from the human gut microbiota were not tested in earlier studies, we first replicated the molecular dynamics (MD) simulations as described (25). In a subsequent round of simulations, we then increased the simulation time from 300 ns to 1 µs using the Amber99SB-ILDN force field instead of OPLS/A because of better performance for proteins (Fig. S14) (35, 36). Within the past few years, major advancements have been made in protein structure prediction using AlphaFold (AF) (37, 38), even enabling accurate MD simulations using relaxed AF models (39). Thus, for proteins where crystal structures were not available, we used AF to generate models (Fig. S15). In a benchmark comparison with Chan et al. (25), we obtained comparable results in MD simulations of the available X-ray crystal structures and their respective AF models (Table S7, Fig. S16). 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. S17; Table S8).

We used MD simulations to further investigate the differences between enzymes capable of defluorinating difluoroacetate (DFA). Previous research on DFA defluorination indicated a smaller active site volume in the fluoroacetate dehalogenases with higher DFA activity (24). Therefore, we compared the size of the halide-binding site in the three enzymes (P2, P3, and P5) capable of defluorinating DFA with those in P4 and P6 (Fig. S18). We observed no significant difference in the distribution of halide-binding pocket sizes that could explain the altered substrate preference. We note, however, that the distributions of Euclidean distances for some of the HADs were multimodal, suggesting distinct conformational states which are likely to include both catalytically active and inactive states. This requires more in-depth investigation but does not change the overall conclusions that the MD simulations alone are insufficient to distinguish defluorinating from non-defluorinating HADs.

We conducted additional experiments to confirm that the fluoride ion release is not solely due to fluoroacetate (FAc) contamination in the DFA (Fig. S19). However, DFA activity is weak, and it remains unknown whether the enzyme itself contributes to the second defluorination step, as the intermediate gem-fluoro alcohol is likely to undergo spontaneous HF elimination to form a ketone (40).

230 ***Phylogenetic analysis does not differentiate defluorinating and non-defluorinating***
231 ***enzymes***
232
233 To further investigate methods to differentiate defluorinating and non-defluorinating enzymes, we
234 generated a phylogenetic tree of the HAD-like family protein sequences (Fig. S20, Table S9). The
235 dehalogenating enzymes originated from eight diverse bacterial and archaeal phyla (Fig. S21)
236 which did not group based on the taxonomy of the source organisms. The tree topology revealed
237 that the defluorinating and non-defluorinating enzymes did not form separate or easily
238 distinguishable clades.
239
240 ***Rational active site engineering converts an inactive enzyme to a dechlorinating enzyme***
241
242 As neither MD simulations nor phylogenetic analysis yielded conclusive results to distinguish
243 defluorinating and non-defluorinating HADs, we next sought to experimentally test the role of
244 specific residues in the active site. We used two of the seven proteins tested in vitro sharing the
245 highest sequence and structural similarity but distinct activities from each other (Fig. S21) (54%
246 sequence identity, 68% sequence similarity). P1, a non-dehalogenating enzyme, and P6, a
247 defluorinating enzyme. Based on comparison between the active sites of the two enzymes, there
248 were only four residues that were significantly different in the active site vicinity including the
249 halide-binding pocket and substrate tunnel (Fig. S21). We predicted these residues may provide
250 substrate specificity for defluorination, therefore we first aimed for rational engineering of the non-
251 dehalogenating enzyme P1 (*E. aldenensis*) via site-directed mutagenesis to generate single,
252 double, and triple variant combinations as well as the full quadruple variant P1 (F54I, T57M, H74F,
253 F187S) (Fig. S22-S25).
254
255 Among the single variants tested, only one (F54I) showed dehalogenation activity. The F54I
256 variant hydrolyzed chloroacetate but did not hydrolyze fluorinated substrates. Based on structural
257 modeling, F54 is positioned at the entrance channel to the substrate binding pocket. In the non-
258 dehalogenating protein P1 (*E. aldenensis*) the larger phenylalanine may exclude chloroacetate at
259 the binding pocket entrance. In contrast, the dehalogenating protein P6 (*Guopingia tenuis*)
260 features an isoleucine at this position, potentially enabling access for chloroacetate. While we did
261 achieve a functional dehalogenase through the single-point mutation in P1 (*E. aldenensis*), the
262 individual active site changes did not singularly confer defluorination activity. We hypothesized
263 epistatic interactions may promote defluorination. We therefore used the F54I variant as the
264 template for generating double, triple, and quadruple active site variants. In the P1 (F54I, H74F)
265 double variant and all triple variants we achieved improved dechlorination activity (Fig. 2B).
266 Surprisingly, we did not observe defluorination activity in any of the rationally engineered variants,
267 even though the active site residues were 100% identical to the defluorinating reference enzyme
268 P6 (*G. tenuis*). This result suggests that additional second shell interactions or other
269 conformational changes (41) not localized in the active site may be critical for defluorination
270 activity.

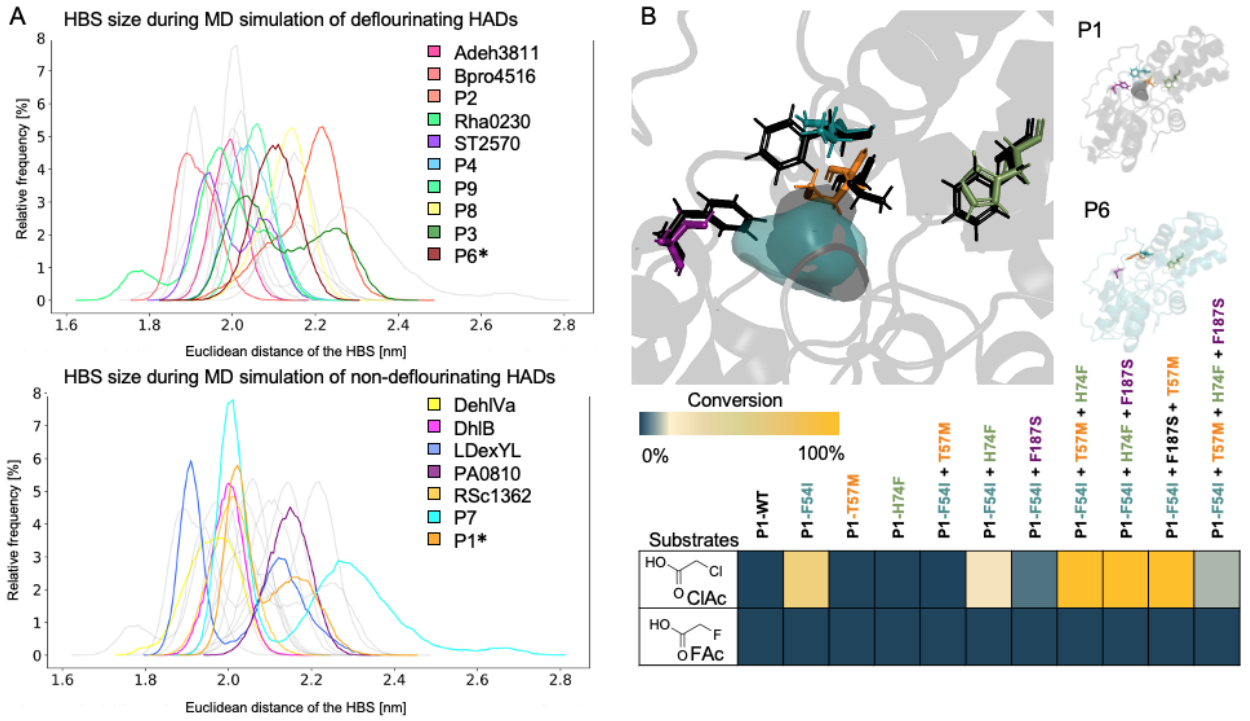

Figure 2. Defluorination capability does not correlate with the active site compactness or amino acid sequence identity. A) Halide binding site (HBS) size in nm during molecular simulations. Distribution of the Euclidean distances of the 3 C_α atoms of the HBS during simulation from 500 ns to 1000 ns. *Asterisks indicate the two enzymes used for protein engineering in this study B) Visualization of the active site of the inactive protein P1 (WP_118709078, *E. aldenensis*) black and the defluorinating protein P6 (WP_178618037, *G. tenuis*) teal. The variants introduced in P1 (*E. aldenensis*) 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).

Development of a new microtiter well fluoride detection assay for alanine scanning

In order to test the contributions of a wider set of amino acids in defluorination e.g., peripheral residues in addition to active site residues, we used alanine scanning mutagenesis to generate full alanine variant libraries of the proteins: P6 *(G. tenuis*) and P1 (F54I, T57M, H74F, F187S) (*E. aldenensis*). In total, the library consisted of 474 single alanine variants. To screen the alanine library, we devised a microtiter well plate colorimetric screening method for defluorination to avoid lengthy ion chromatography steps for analysis. Previously, a lanthanum-alizarin assay for microtiter well plates (42) and a xylenol orange assay for an agar plate-based method (43) were reported. We tested the lanthanum-alizarin assay (42) and additionally developed a miniaturized liquid adaptation of the xylenol orange screening method. Both assays yielded a clear change in absorbance in the presence of free fluoride ions as validated by ion chromatography results (Table S10). Due to its enhanced color intensity, we report here the results of screening using the miniaturized xylenol orange assay. As the xylenol orange assay has not been described previously for use in liquid, microtiter well plate format, a detailed methods section and list of interfering substances is included (Fig. S26-S27, Table S11). In our assay, sodium phosphate monobasic was the compound with the highest observed interference in the present fluoride

assay which can be reduced to below the limit-of-detection with washing steps (Figure S28, Methods).

To test our miniaturized screening method, we selected eleven candidate dehalogenases spanning eleven additional gut microbial genera (Table S12). Based on the colorimetric screen, we detected an active gut microbial defluorinase from *Priestia megaterium* sharing less than 35% amino acid identity to our reference defluorinase (PDB ID: 3UMG). We purified the *P. megaterium* enzyme to homogeneity and validated its defluorination activity using ion chromatography (Fig. S29). This highlights the potential of microtiter-well assays to rapidly identify defluorinating enzymes.

With our new microtiter defluorination assay in hand, we next screened the alanine scanning library with a dual-assay approach for both defluorination and dechlorination activity (Fig. 3A). The screening was conducted in two independent experimental campaigns each with biological replicates as further described in the Methods section, with and without dilution, and normalization for protein concentration (Experiment 1 and Experiment 2, respectively). Across independent experiments and biological replicates of the alanine scanning library (Methods), 39.8-41.7% of alanine variants lost defluorination activity. A slightly lower proportion of variants (33.0-33.5%) lost dechlorination activity in Experiment 1. In Experiment 2, 31%-37% of alanine variants lost defluorination activity and 39-41% lost dechlorination activity. We identified the top 19 alanine variants which showed dechlorination activity but lost defluorination activity in both experiments (Fig. S32-S34, Table S14-15). Despite variability among replicates within and across Experiments 1 and 2 caused by inherent biological differences in growth and expression—remaining even after protein normalization correction (Fig. S30-31)—both experiments identified the same important regions and residues linked to defluorination (Fig. S35, Table S13). Subsequent machine learning results are reported here for Experiment 1 for clarity (Exp. 2 results are reported in Table S16 and Fig. S36).

Machine learning highlights important protein regions for defluorination

To analyze our alanine scanning data, we used the random forest algorithm to train a machine learning model to predict defluorination activity. Random forest is suitable for both regression and classification tasks even under low-data regimes (n=474 data points from the alanine scanning experiments). Our regression model trained on 80% of the alanine scanning data was capable of predicting quantitative changes in defluorination activity as the result of single alanine point mutations with the remaining 20% of the data with a root-mean square error (RMSE) of 0.2 ± 0.02 enzyme activity units (on a min-max normalized activity scale between 0-1 defluorination activity). Based on the relative feature importance, the most predictive feature was the position of each amino acid within the protein sequence rather than the size, charge, identity, codon diversity, or polarity of each amino acid converted to alanine (Fig. 3B). These results showing position carried the highest importance for defluorination activity corresponds with the visible ‘defluorination’ hotspots on the protein structure (Fig. 3C).

342 To train a more generalizable model with applications beyond our selected alanine scanning
343 libraries P1 (F54I, T57M, H74F, F187S) and P6, we used random forest to classify defluorinating
344 vs. non-defluorinating enzymes based on general amino acid features extendable to untested
345 HAD sequences (Fig. 3D). We obtained 95% training and test set classification accuracy (F1-
346 score = 0.928 ± 0.028, Table S16). To test the generalizability of our classification model, we used
347 the model to predict defluorination activity for a new set of five uncharacterized HAD homologs
348 (Methods, Fig. S35). We obtained a defluorination probability score of 0.9 for a HAD
349 (WP_139652913.1) encoded in the genome of *Raoultibacter phocaeensis* Marseille-P8396^{T}
350 isolated from human stool samples (44). We heterologously expressed and confirmed positive
351 defluorination activity in the *R. phocaeensis* HAD. The other four HADs in the dataset expressed
352 well in *E. coli* but were not active in defluorination. Our models predicted a low defluorination
353 probability of 0.5 or less (Fig. S37, Table S17), supporting the ability of our models to discern
354 high-probability defluorinating HADs from others.
355
356 Although recognizing the limited scope of this model to HAD-like family enzymes, this
357 classification analysis pinpointed residues that were important for defluorination activity (Fig. S35-
358 36). Strikingly, the top five most important residues (Experiment 1) for defluorination activity
359 prediction accuracy were localized in the C-terminus (final 41 amino acids). A simplified model
360 trained on only the 41 C-terminal residues as features still achieved 83% testing and training set
361 prediction accuracy (average F1-score = 0.728 ± 0.129, Experiment 1, Table S16).
362
363 Additionally, we computed the root mean square fluctuations (RMSF) from our molecular
364 dynamics (MD) simulations of P6 (*G. tenuis*), allowing us to compare the spatial fluctuations of
365 different regions within the protein structural model. Our analysis identified two regions exhibiting
366 high RMSF values. Both regions are situated at the channel entry leading to the active site (Fig.
367 S38), with one of them located at the C-terminus (Fig. 3E). This finding aligns with the residues
368 implicated in defluorination, as determined by both machine learning predictions and experimental
369 evidence.

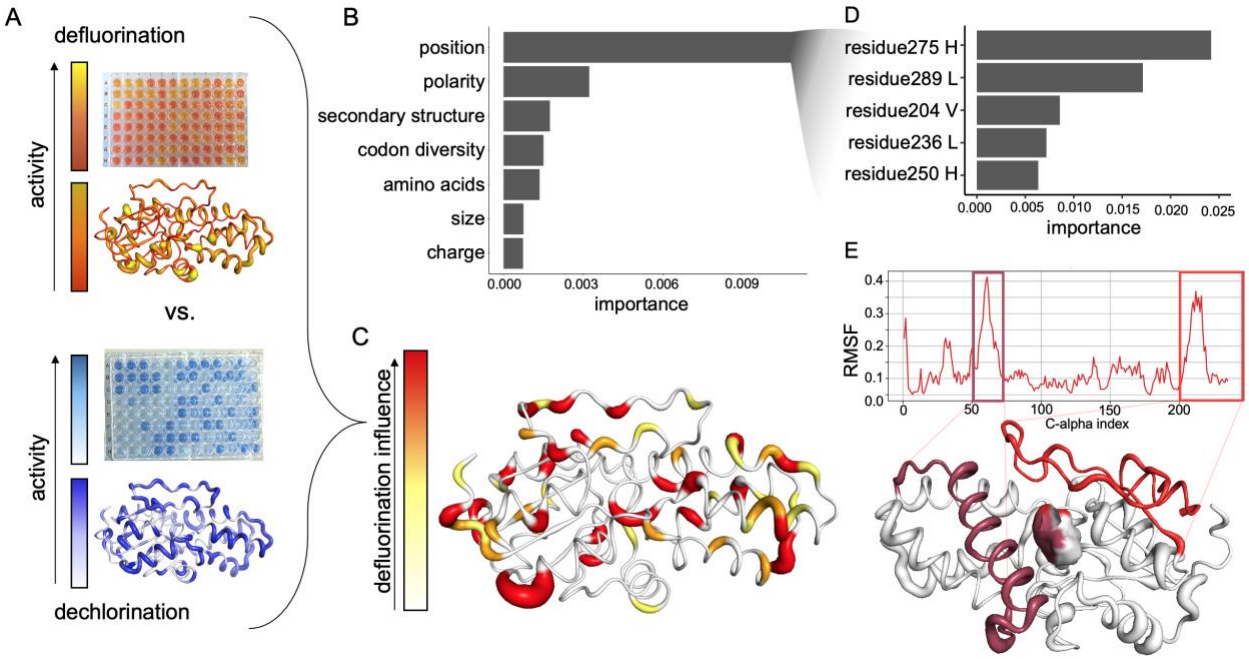

Figure 3. The intrinsically disordered C-terminal region is important for defluorination activity. A) Results of Experiment 1 colorimetric assays for defluorination and dechlorination activity, respectively, mapped onto the P6 protein model. The color gradient from white (low activity) to yellow (medium 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 correspond 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 C_α atoms indicates high movement in the two regions of the protein. The two boxes highlight the more flexible regions in the proteins and lines connect the corresponding amino acids in the protein visualization. Maroon = amino acids 36-69, Red = amino acids 197-235. The surface of the binding pocket location is shaded by proximity to the two regions.

Non-defluorinating enzyme gains defluorination function by C-terminus swap

To test the relevance of the C-terminal region experimentally, we generated chimeric proteins by combining different regions from the quad variant non-defluorinating protein P1 (*E. aldenensis* F54I-T57M-F74H-F187S) with the defluorinating enzyme P6 (*G. tenuis*) (Fig. S39). Strikingly, the single P1 quad variant (*E. aldenensis* F54I-T57M-F74H-F187S) 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. S39). 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 sufficient to install defluorination activity when fused to a non-defluorinating homologous enzyme.

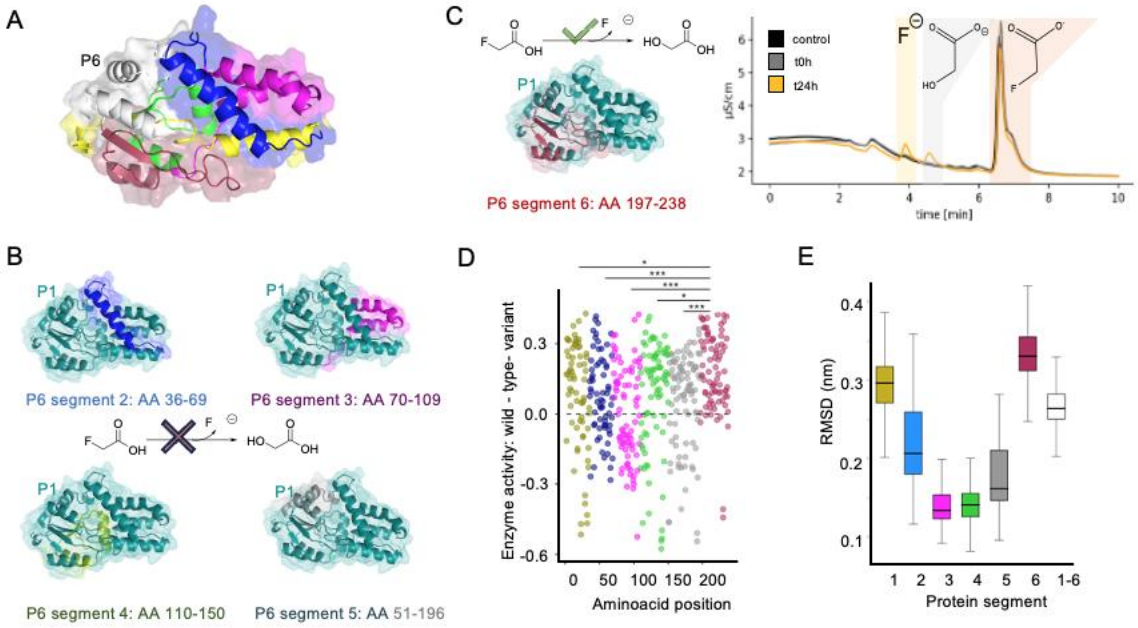

Figure 4. Swapping the C-terminus of a non-defluorinating HAD with the C-terminus of a defluorinating homologous enzyme enables defluorination activity. A) Representation of the defluorinating protein P6. The colored parts represent the exchanged segments used for generating the chimeric protein variants with P1. B) The protein variants with the colored segment from P6 showed no defluorination activity. C) The engineered chimeric protein with a swapped C-terminal segment enables weak conversion of fluoroacetate to glycolate and fluoride ion. D) Protein defluorination activity of the alanine scanning library with all data points shown for two independent biological replicates (Experiment 2 results are visualized, both Experiment 1 and Experiment 2 results are provided in Table S13-14, Table S18). The x-axis indicates the position in P6 and the y-axis indicates the change in min-max normalized activity for each alanine scanning variant relative to wild-type activity (dotted gray line). 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). E) Boxplot of root mean square deviation (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 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 HADs sourced from the human gut microbiota capable of catalyzing this reaction. Despite sharing sequence and structural similarity, the proteins in this study exhibited variable substrate specificity including activity with fluoroacetate and α-fluoro-β-alanine. Both of these are known toxic transformation products (6, 7) of 5-fluorouracil (5-FU), a widely-used cancer medication on the World Health Organization's list of Essential Medicines (45). The anticancer mechanism of 5-FU is primarily based on the inhibition of thymidylate synthase (TS) by 5-fluoro-2'-deoxyuridine monophosphate (FdUMP). FdUMP competes with deoxyuridine monophosphate by forming a covalent ternary complex with TS and methylenetetrahydrofolate (46). This mechanism-based inactivation has been a significant driving force behind therapeutic innovations targeting not only human, but also bacterial enzymes, such as α-fluoromethyl ketone bile acid

430 analogs which covalently inhibit gut bacterial bile salt hydrolases (47). Studies on fluorinated
431 substrate analogs that inhibit pyridoxal phosphate (PLP)-dependent alanine racemases (48) as
432 well as other enzyme classes (49), have advanced the discovery and characterization of enzyme
433 catalytic mechanisms and their physiological roles. A notable example is the discovery of a *(S)*-
434 α-fluoromethyltyrosine inhibitor of a microbial PLP-dependent tyrosine decarboxylase. This
435 previously unidentified microbial enzyme converts levodopa, a medication used to treat
436 Parkinson's disease, into dopamine, thereby contributing to interindividual variation in drug
437 efficacy (50). Collectively, these findings and our study contribute to a growing mechanistic
438 understanding of how diverse gut microbial enzymes interact with fluorinated substrates.

439 In this study, we also identified four new HADs active on DFA, doubling the number of hydrolytic
440 dehalogenases experimentally shown to defluorinate -CF_2 groups (24). DFA is a transformation
441 product of various xenobiotics including the refrigerant R1113, chlorotrifluoroethylene, (51) and
442 the anesthetic 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-ethane (52). DFA is also a byproduct of
443 the photodegradation of penoxsulam, a widely used pesticide (53), thereby exposure to the gut
444 microbiota may occur via the consumption of pesticide residues on food. In general, fluorinated
445 substances seep into groundwater and drinking water, where they are subsequently consumed
446 by humans, thus coming in contact with the gut microbiota, although there is a lack of quantitative
447 exposure data to fully assess the extent and impact of this interaction.Additionally, this study is
448 limited in its scope to overexpressed recombinant HADs sourced from the gut microbiota. In vivo
449 and pure culture experiments remain inconclusive due to unknown conditions for gene expression
450 and interference from complex matrices, requiring further investigation.
451
452 Specificity-determining residues or structural features that differentiate defluorinating enzymes
453 from their non-defluorinating counterparts in the L-2-haloacid dehalogenases have remained
454 elusive (25). Alanine scanning proved useful to identify protein variants that retained
455 dechlorination activity but lost their defluorination capability and to pinpoint several amino acids
456 critical for defluorination. Surprisingly, these amino acids were primarily located in the peripheral
457 regions of the protein with a higher percentage of significant residues located in the C-terminal
458 segment of the protein. These observations, along with results from more recent studies (54),
459 support the growing complex understanding of protein function beyond the simplistic lock-and-
460 key model (55). Consequently, our protein engineering efforts shifted towards engineering
461 chimeric proteins by replacing segments of non-defluorinating protein variants with segments from
462 a defluorinating enzyme. We observed chimeric proteins with weak, but robust and reproducible
463 defluorination activity only when the C-terminal segment was exchanged.
464
465 The C-terminal regions of proteins have been extensively studied in eukaryotic cells (56, 57),
466 where 'short linear motifs' within intrinsically disordered regions (58) have been associated with
467 a broad spectrum of functions, including substrate recognition (56, 59), signaling (60), and
468 disease risk (56, 61). This functional diversity is attributed to the solvent exposure and higher
469 degree of disorder of C-termini, which facilitate protein-protein interactions (62) and
470 oligomerization facilitated through natural C-domain swapping (63). Previously, in a HAD
471 superfamily phosphatase, the flexible C-terminal tail was shown to interface with adjacent
472 monomers to facilitate product release (41). Our MD simulations suggested gut microbial HAD C-

termini exhibit greater flexibility than other protein regions. Combined with a 'gain-of-function' through the design of a C-terminal chimera, we propose the C-terminal region of HADs plays a significant and previously unreported role in defluorination. This knowledge lays the foundation to engineer gut microbial proteins with expanded substrate ranges for carbon-fluorine bonds, opening new applications in human and environmental health. It adds a prokaryotic dimension to the 'C-terminome' to more accurately predict and modulate enzyme functions in microbiomes and across the domains of life.

Data and code availability

Scripts and data associated with analyses presented in this manuscript are available at: https://github.com/MSM-group/gut_microbe_defluorination_paper

Acknowledgements

We acknowledge Michael Zimmermann, Mitchell Levesque, and Larry Wackett for their thoughtful discussions and input. We acknowledge René Gall for his technical support with analytics and Till Epprecht for his support in computational analysis. We thank the entire Joint Genome Institute team including Miranda Harmon-Smith, Yasuo Yoshikuni, and many more 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 postdoctoral funding from the Peter und Traudl Engelhorn Stiftung. V.P. and S.L.R. acknowledge funding from the Swiss National Science Foundation (grant no. PZPGP2_209124), the Uniscientia Foundation, and the Helmut Horten Foundation (2022-YIG-090).

Author contribution

Conceptualization, S.L.R and S.I.P.; Methodology, S.L.R, S.I.P. F.F., V.P., R.M. and I.K.B.; Investigation, S.L.R, S.I.P. F.F., V.P., R.M, and I.K.B.; Writing – Original Draft, S.L.R., S.I.P., V.P., and F.F.; Writing – Review & Editing, S.L.R., F.F., I.K.B., R.M., V.P., and S.I.P.; Funding Acquisition, S.L.R. and I.K.B.; Supervision, S.I.P., I.K.B., and S.L.R.

Declaration of interests

The authors declare no competing interests.

Methods

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 (25) from the National Center for Biotechnology Information (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) (29). A BLASTdb was created using the 30,691 HumGut genomes and

517 a tblastn search was performed to identify homologs of the WP_011593529.1 and
518 WP_011481495.1 proteins. The 500 sequence hits with bitscores greater than 40 were subjected
519 for a reverse blast search against the NCBI database. Hits that could be mapped to NCBI
520 database entries with 100% sequence identity and 90-100% query coverage were further curated
521 for redundancies. The resulting list of 114 protein sequences (Table S1) were aligned using
522 Clustal Omega within the EMBL-EBI Framework (64) and the presence of the catalytic active
523 residues were verified. Genomic context for the query sequences were extracted using the
524 genome mining tool RODEO v2.0 (65). Code and query sequences are available at
525 https://github.com/MSM-group/gut_microbe_defluorination_paper. Subsequently, we used the
526 EFI-EST webtool to generate a sequence similarity network (SSN) with default parameters (66).
527 We visualized the 95% amino acid identity representative network using Cytoscape (67) (Fig. S2)
528 and seven SSN cluster representatives were selected for experimental validation (Table S2).
529 Genes were synthesized as described below for in vitro work. In a second 'validation' round, 11
530 new proteins were selected based on high sequence similarity alignment scores for additional
531 experimental validation (Table S12).
532
533 **2. Cloning and heterologous expression of candidate enzymes**
534 The first batch of full-length dehalogenases (n=7) were codon-optimized for *E. coli* codon usage
535 using the Integrated DNA Technologies (IDT) Codon Optimization Tool. The sequences were
536 synthesized as gBlocks by Integrated DNA Technologies. The gBlocks were cloned individually
537 by Gibson Assembly into multiple cloning site 1 of pCFDuet-1 vectors
538 (spectinomycin/streptomycin resistance marker) with tobacco etch virus (TEV) cleavage sites and
539 C-terminal hexahistidine tags (Table S19-S21). Constructs were transformed into *E. coli* DH5α
540 cells and verified by Sanger sequencing (Microsynth, Balgach, Switzerland). Sequence-verified
541 constructs were transformed into *E. coli* T7 Express cells for protein production (New England
542 Biolabs Frankfurt, Germany, NEB C2566I). The second batch of synthesized genes for
543 colorimetric assay validation (n=11) were cloned by the U.S. Department of Energy Joint Genome
544 Institute. These were constructed identically to the first batch with the exception of using the build-
545 optimization software, BOOST (68) for codon optimization and expression in T7 Express *E. coli*.
546 All codon-optimized sequences are available at:
547 github.com/MSM-group/gut_microbe_defluorination_paper/tree/main/protein_sequences_info
548
549 Starter overnight cultures (5 mL) were grown in LB with 100 µg/mL spectinomycin overnight at
550 37°C. From the starter culture, 0.5 mL were added to 200 mL Terrific Broth with 100 µg/mL
551 spectinomycin in a 1 L baffled Erlenmeyer flask and grown to an optical density of 0.5-0.8 at 37°C.
552 Protein expression was induced by adding 0.1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG)
553 and cultures were incubated for 24h at 16°C.
554
555 For scale up and purification, induced cultures were poured into 50 mL Falcon tubes and
556 centrifuged at 4000 RCF for 30 min. The supernatant was decanted, and cell pellets were
557 resuspended in 4 mL buffer A (20 mM Tris-HCl, 500 mM NaCl, 20 mM imidazole) per gram cell
558 pellet, 10% glycerol, pH = 8) per gram of pellet weight. Cells were lysed by 3 cycles of sonication
559 using a 6 mm needle, 30% amplitude and 10 x 10 second pulses with a rest time of 10 seconds
560 in between (total sonication time per cycle = 3 minutes). Lysate was centrifuged at 20,000 RCF

561 at 4°C for 45 min. For benchtop purification, supernatants were applied to a column packed with
562 Ni-NTA agarose beads (Qiagen) pre-equilibrated in buffer A. The column was subsequently
563 washed with buffer A and then subjected to two washes with buffer B (40 mM Tris-HCl, 500 mM
564 NaCl, 10% glycerol and 40 mM imidazole). Enzymes were eluted from the column using buffer C
565 (40 mM Tris-HCl, 500 mM NaCl, 10% glycerol, and 500 mM imidazole) The fractions with the
566 highest protein concentration was determined qualitatively by Bradford assay(69). Imidazole and
567 NaCl were removed by running the sample over a PD-10 desalting column (Cytiva) and eluting in
568 buffer (100 mM Tris-SO_4, pH = 8.5). The concentration of the purified proteins were determined
569 by Qubit™ protein assay and the proteins were flash frozen using liquid nitrogen in 500 µL aliquots
570 and subsequently stored at -80°C.
571
572 **3. Assessment of enzyme-substrate specificity**
573 Stock solutions containing 100 mM of the test substrates listed in Table S5 dissolved in water
574 were prepared. The concentration of the Ni-NTA purified proteins were determined by using the
575 Qubit™ protein assay and the proteins were diluted to a final concentration of 0.5 mg/mL in 100
576 mM Tris-SO_4 buffer, pH 8.5. As controls the proteins were boiled for 10 min at 100°C. The test
577 substrates were added to a final concentration of 1 mM and the samples were incubated at 37°C.
578 50 µL samples were taken for the different timepoints and mixed with 50 µL acetonitrile to quench
579 the reaction. The samples were centrifuged at 13 000 RPM for 5 min and the supernatant
580 transferred into plastic vials for ion chromatography (IC).
581
582 An IC system (Metrohm, Switzerland) was used consisting of an IC sampling center 889 coupled
583 to an ion chromatograph (930 Compact IC Flex) with a conductivity detector. Samples were
584 incubated at 10°C in the sample center with 10 µL injection volumes first onto a precolumn
585 (Metrosep A Supp 16, Guard, 4 mm) and subsequently onto a Metrosep A Supp 16 column (250
586 x 4.0 mm, 5 µm). The temperature in the column was kept constant at 30°C. The eluent consisted
587 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
588 and quantification was performed using MagIC Net Professional (version 3.3).
589 All experiments were conducted at least in triplicates.
590
591 **4. AlphaFold and structural comparison**
592 Structural models for the seven initial dehalogenase sequences selected from the HumGut
593 databases, were generated using AlphaFold2 (AF) ColabFold(70). Models with the highest rank
594 for each dehalogenase were prepared with relaxation using the Amber99SB force field(71).
595 Additionally, we used AF to generate models for the HADs PA0810, RSC1362, ST2570 and
596 RHA0230. The accuracy of AF models was compared to known X-ray crystal structures, where
597 available (25).
598
599 **5. Molecular dynamics simulations**
600 Molecular dynamics simulations were performed using the GROMACS package version 2022.5
601 (72, 73) compiled with CUDA support and run on NVIDIA GeForce RTX 2080 Ti graphic cards.
602 The crystal structures for HADs PA0810, RSC1362, ST2570 and RHA0230 were downloaded
603 from RCSB PDB (PDB IDs: 3UMC, 3UMB, 2W43 and 3UMG, respectively. For PDB files which
604 contained selenomethionine, selenomethionine was exchanged with methionine. Relaxed AF

605 models were also subjected to simulations. The system was prepared with GROMACS tools,
606 parameterized with the force field Amber99SB-ILDN (74), and solvated in explicit water using the
607 default water model TIP3P (75) in a cubic box ~1.2 nm bigger than the protein in each direction.
608 The protein's net charge was neutralized with Na^+ or Cl^- ions. Conjugate gradient energy
609 minimization was performed with a minimum of 50 steps and a maximum of 1000 steps or until
610 the energy of the system reaches below 1000 kJ/mol/nm. After a short (200 ps) NVT equilibration
611 and (200 ps) NPT equilibration, the production runs were carried out. The simulations were
612 performed under periodic boundary conditions for a total simulation length of 0.1 μs with ten
613 replicates and 1 μs with two replicates using the leap-frog algorithm with time steps of 2 fs for
614 each simulation. The bonds involving hydrogen atoms were constrained with the LINCS algorithm
615 (76). The cutoff scheme for nonbonded interactions was Verlet with a cutoff of 1.2 nm (for all
616 AMBER force fields). The Coulomb interactions were calculated using the particle mesh Ewald
617 algorithm (77) and a Fourier spacing of 0.14 nm. The temperature was controlled using a modified
618 Berendsen thermostat (78) with a coupling time of 0.1 ps (protein & non-protein coupling groups)
619 and a reference temperature of 300K. The pressure was controlled using the Berendsen barostat
620 (79) an isotropic pressure coupling in x-y-z directions, a coupling time of 1.0 ps, a reference
621 pressure of 1.0 bar, and an isothermal compressibility of 4.5×10^{-5} bar^{-1}. The variance of the 100
622 ns simulations with 10 replicates ranged between 0.015 and 0.08 nm for all simulated proteins.
623 For the 1 μs runs the variance between the two replicates ranged between 0.0016 nm and 0.031
624 nm. The root mean square deviation (RMSD) of the simulated proteins relative to the starting
625 structure (C_α and all atoms) was measured using gmx_rms for both whole proteins and for the
626 protein segments with the necessary index files generated with make_ndx. The three active site
627 residues of the halogen binding site (aspartate, asparagine and arginine) were identified and the
628 Euclidean distance between the three C_α atoms was tracked during the simulation after an
629 equilibration period of 500 ns with the built-in tool gmx_distance (73).. To measure the Root Mean
630 Square Fluctuation (RMSF) of all C_α atoms during the simulation the built-in function gmx_rmsf
631 was used.
632

633 **6. Sequence based comparison of HADs**

634 Multiple sequence alignment and maximum-likelihood phylogenetic analysis (Fig. S20-S21, Table
635 S9) were performed to gain a better perspective on the sequence features of the HAD-like
636 enzymes. We performed these two analyses with a data set of the entire HAD-like family available
637 in UniProt (80) for 248 sequences available as of August 2023 with experimental evidence
638 characterized at the protein level. A multiple sequence alignment of the resulting 248 proteins
639 was performed using the Clustal Omega algorithm (81) with a Biopython wrapper (82) and the
640 phylogenetic tree (Fig. S20-21) was generated using the IQ-TREE algorithm (83, 84). The model
641 selection was performed by IQ-TREE model finder function (83). The best model was selected
642 (substitution model Q.pfam+R8) and a maximum-likelihood tree was constructed with 1000
643 bootstrap replicates and visualized using ETE 3 (85) .
644

645 **7. Rational active site engineering**

646 To identify potentially significant amino acids involved in the defluorination process, we used two
647 of seven in vitro tested proteins sharing high structural and sequential similarity but distinct
648 activities from each other (Fig. S21) (54% sequence identity, 68% sequence similarity).

649 WP_1187090781.1, a non-defluorinating enzyme, and WP_1786180371.1, a defluorinating
650 enzyme. We targeted amino acids within WP_1187090781.1, particularly those situated in or near
651 the active site, for substitution with their counterparts from WP_1786180371.1. Four single
652 variants were constructed: F54I, T57M, F74H, and F187S and their combinations resulting in 14
653 variants encoded by point mutations in codons for up to four different active site residues. The
654 mutant gene constructs were generated by using mismatch primers (Table S21). A 3 µL Kinase,
655 Ligase & DpnI (KLD) Treatment Reaction containing 0.5 µL of the unpurified PCR product, 1x
656 KLD Reaction Buffer and 1x KLD Enzyme Mix (New England Biolabs) was incubated at room
657 temperature for 10 minutes followed by 10 minutes at 37°C.
658
659 **8. Colorimetric assay for fluoride detection**
660 A colorimetric fluoride screening method was developed here for 96-well plate format based on
661 the agar-plate based colony screening method initially described by Davis et al. (43) Xylenol
662 orange stock solution was prepared with a concentration of 20 mM. To prepare the zirconium
663 oxychloride octahydrate stock solution, 0.3022 g of zirconium(IV)-oxychloride octahydrate was
664 dissolved in 925 µL water and 1.85 mL 12.1 M HCl added and then boiled for 5 min to achieve
665 final concentrations of 500 mM fully depolarized zirconium and 8 M HCl. After cooling, the stock
666 solution was diluted to a concentration of 50 mM zirconium. Shortly before the experiment, the
667 working solution was prepared with 7.8 M HCl, 400 µM xylenol orange and the complex was
668 initiated by the addition of 250 µM zirconium. A list of potential interfering substances and buffers
669 were tested, and results are listed in Table S11. Sodium phosphate monobasic (hereafter
670 phosphate) interferes with the formation of the xylenol orange-fluoride complex at phosphate
671 concentrations above 100 µM. The time dependence of phosphate interference was analyzed
672 over a range of phosphate concentrations between 100 and 1000 µM (Fig. S28).
673
674 *E. coli* T7 expression strains carrying the corresponding plasmids were grown in TB medium in a
675 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 µg/mL spectinomycin until a
676 optical density of 0.8 was reached The cells were induced with 0.1 mM IPTG and incubated for
677 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant was decanted and the
678 cells were washed twice with 2 mL 20 mM HEPES (pH 8). After centrifugation at 4000 RCF for
679 10 min the cells were resuspended in 1 mL 20mM HEPES (pH 8) and fluoroacetate was added
680 to a final concentration of 1mM prior to incubation at 37°C for 24 h. The cells were centrifuged at
681 4000 RCF for 15 min and 200 µL of the supernatant was transferred to a new 96 well plate and
682 40 µL of the zirconium/xylenol orange working solution was added. The final concentrations in
683 each well were 1.3 M HCl, 66.7 µM xylenol orange, and 41.7 µM zirconium. Two independent
684 experiments were performed. The first experiment was performed in duplicate for all protein
685 variants and assessed for defluorination activity. The second experiment was conducted in
686 triplicate for all protein variants and tested for defluorination but with normalization to protein
687 concentrations. For the second experiment, protein concentrations were determined using the
688 Pierce™ 660nm Protein Assay kit (Thermo Fisher Scientific, Waltham, MA, USA) following the
689 manufacturer's instructions.
690
691 The absorbances used to calculate fluoride ion concentration was measured at 540 nm and the
692 isosbestic point was measured at 480 nm using an Eon spectrophotometer (Biotek). The

693 approximate fluoride concentration is calculated using the following formula: x=(abs(480 nm)) /
694 (abs(540 nm)) where a ratio above 1.8 indicates a [F^-] greater than 100 µM. Strongly positive
695 samples (>100 µM [F^-], x>1.8) appear yellow and negative samples appear red (<10 µM [F^-],
696 x<1.2). Fluoride calibration curves were prepared by spiking sodium fluoride amounts between 0
697 µM and 100 µM into cell supernatants taken from negative control samples. The assay shows a
698 2nd degree polynomial relationship for fluoride concentrations between 0 µM and 100 µM. Fluoride
699 concentrations above 100 µM were diluted 1:3 (Fig. S26) and quantified using a 1:3 diluted
700 calibration curve. It is recommended to measure samples in a span of 15 min after adding the
701 reagent mix, as this time span. Is shown to only slightly alter measurement data even at
702 phosphate concentrations of 100 µM. It is advised to perform the calibration in the sample matrix
703 (negative control) to get reliable concentration data. More information and a spreadsheet to
704 prepare diluted and undiluted calibration curves is available https://github.com/MSM-
705 group/gut_microbe_defluorination_paper/tree/main/Dehalogenation_colorimetric_assays/Experi
706 ment2.
707
708 **9. Chloride detection assay**
709 A chloride detection assay kit (QuantiChrom™ Chloride Assay Kit) was obtained from Huber lab
710 and assays were run according to the manufacturer's instructions with the following changes:
711 Briefly, *E. coli* T7 expression strains carrying the corresponding plasmids were grown in TB
712 medium in a 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 µg/mL
713 spectinomycin until an optical density of 0.8 was reached. The cells were induced with 0.1mM
714 IPTG and incubated for 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant
715 was discarded, and the cells were washed twice with 2 mL 20 mM HEPES (pH 8). After
716 centrifugation at 4000 RCF for 10 min the cells were resuspended in 1 mL 20 mM HEPES (pH 8)
717 and chloroacetate was added to a final concentration of 30 mM prior to incubation at 37°C for 24
718 h. The cells were centrifuged at 4000 RCF for 15 min and 2.5 µL of the supernatant was
719 transferred to a new 96 well plate and 100 µL of the colorimetric working solution was added. The
720 chloride concentration was measured at 610 nm with an Eon spectrophotometer (Biotek). The
721 limit-of-quantification was 0.2 mM.Two independent experiments were performed. The first
722 experiment was performed in duplicate for all protein variants and assessed for dechlorination
723 activity. The second experiment was conducted in triplicate for all protein variants and tested for
724 dechlorination but with normalization to protein concentrations.
725
726 **10. Whole-gene alanine-scanning mutagenesis and chimeragenesis**
727 High-throughput alanine-scanning mutagenesis was performed using the Q5® Site-Directed
728 Mutagenesis Protocol (New England Biolabs Inc.) with the following amendments. Briefly, back-
729 to-back forward and reverse primers were designed using NEBaseChanger
730 (https://nebasechanger.neb.com/), where the alanine codon substitution was contained in the
731 center of the forward mutagenic primer. Note that a histidine codon was substituted in place of
732 any native alanine codons. Primers were ordered from Integrated DNA Technologies. The
733 complete set of alanine (n=474) are available on https://github.com/MSM-group/biodefluorination-
734 paper/tree/main/Alanine_scanning_primers. A 6 µL PCR reaction, containing 1x Q5 Hot Start
735 High-Fidelity 2X Master Mix (New England Biolabs), 0.5 µM of the forward and reverse primer,
736 and 10 ng template DNA, was carried out for each variant. P6, WP_178618037.1 (*Guopingia*

737 *tenuis*) and P1 quadruple variant, WP_118709078.1 F54I-T57M-F74H-F187S (*Enterocloster*
738 *aldenensis*) cloned into the pCDFDuet-1 vector was used as template DNA. PCR cycles were
739 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 65°C for 20 seconds and
740 72°C for 2 minutes, followed by a 2-minute final extension at 72°C. A 3 µl KLD Treatment Reaction
741 containing 0.5 µl of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix
742 (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes
743 at 37°C. The entire reaction was transformed into NEB® DH5α competent *E. coli* using the
744 manufacturer's recommended transformation protocol and sequence verified on the Pacific
745 Biosciences Revio platform (Pacific Biosciences, CA) and analyzed using custom pipelines.
746 Plasmid DNA of sequence-verified variants were isolated and transformed into T7 Express cells
747 (New England Biolabs), which was once again sequence verified (Pacific Biosciences). We
748 engineered five chimeric proteins by combining longer sequences from P6, WP_1786180371.1
749 and P1 quad variant, WP_1187090781.1 F54I-T57M-F74H-F187S (Table S5). Gibson assembly
750 (primers in Table S5) was used for the construction of the plasmids.

751 **11. Machine learning**

752 Random forest regression and classification was performed based on scripts and featurization
753 methods published previously (86, 87) with the random forest algorithm as implemented in the
754 ranger package (88). Briefly, for activity prediction using regression based on the alanine
755 scanning data, each point variant (n=474) was featurized using the method of Atchley et al. (89)
756 to a vector representing the position index, size, charge, identity, codon diversity, and polarity of
757 each amino acid which was converted to alanine. These features were subsequently used to
758 predict the change or delta in defluorination enzyme activity from wild-type as a result of each
759 alanine substitution. For the random forest classification task, the sequences of all HAD
760 superfamily enzymes which have been characterized in this study and previously in the literature
761 were aligned using a protein structure-aware alignment method (90). Invariant residues in the
762 alignment were trimmed and the remaining residues were one-hot encoded and used as features
763 to classify HAD sequences as 1: defluorinating or 0: non-defluorinating. For both classification
764 and regression tasks, the data were split by stratified random sampling into 80% training and 20%
765 testing sets. Grid search was used to tune model hyperparameters by 10-fold cross validation
766 repeated in triplicate. All forests were grown to a size of 1000 trees. Hyperparameters that were
767 tuned during cross-validation included the number of variables randomly sampled as candidates
768 at each split, the minimum size of terminal nodes and the splitting rule methods. The variability in
769 training and testing set prediction performances were examined with 1000 different random seeds
770 for training-test splits. To assess the specific importance of the C-terminal region for classification
771 prediction, only the final 41 residues in the HAD alignment were used as features and model
772 performance was compared to models trained on the full-length alignment. Scripts for the machine
773 learning analysis are available at:
774 https://github.com/MSM-group/gut_microbe_defluorination_paper/tree/main/machine_learning/.

775 **12. Experimental validation of machine learning model predictions**

776 Random forest classification models were used to predict defluorination probability for HAD
777 homologs in NCBI RefSeq and assembled coding sequences from human gut metagenomes in
778 the Short Read Archive (BioProject ID: PRJNA268964). Five diverse HAD sequences (denoted
779 P20-P24) with defluorination prediction probabilities ranging between 0.38 and 0.85 (Fig. S35)

780 were selected for experimental validation. Genes were codon optimized and synthesized as
781 described above by IDT and cloned by Gibson Assembly into multiple cloning site 1 of pCFDuet-
782 1 vectors. Codon optimized sequences are available here: https://github.com/MSM-
783 group/gut_microbe_defluorination_paper/tree/main/protein_sequences_info/. Constructs were
784 transformed into *E. coli* DH5α cells and verified by Sanger sequencing (Microsynth, Balgach,
785 Switzerland). Sequence-verified constructs were transformed into *E. coli* T7 Express cells for
786 protein production (New England Biolabs Frankfurt, Germany, NEB C2566I). Defluorination
787 activity was then assessed using the colorimetric assay method described above and compared
788 to machine learning predictions for defluorination.
789
790 **13. In vivo validation of P2 from *Gordonibacter pamelaeae* DSM 110924**
791 *Gordonibacter pamelaeae* was inoculated from a cryostock into 8 mL gut microbiota medium
792 (GMM: tryptone peptone 2 g/L; yeast extract 1 g/L; raffinose 10 g/L; L-cysteine 0.5 g/L; meat
793 extract 5 g/L; KH_2PO_4 100 mM; $MgSO_4$-$7H_2O$ 0.002 g/L; $NaHCO_3$ 0.4 g/L; NaCl 0.08 g/L; $CaCl_2$ 8
794 mg/L, menadione 1 mg/L; $FeSO_4$ 0.4 mg/L ; hematin 1.2 mg/L; histidine 0.2 mM; Tween 80 0.5
795 mL/L; ATCC vitamin mix 10 mL/L; ATCC trace mineral mix 10 mL/L; resazurin 1 mg/L) in triplicates
796 and incubated anaerobically for 8 days until it reached an OD of 0.1. Fluoroacetate was added to
797 a final concentration of 10 mM and the cultures were incubated overnight. The culture was
798 centrifuged and the supernatant used for fluoride determination before transfer of 2 mL of
799 supernatant into a small beaker with 10 mL TISAB-II-solution (ammonium acetate 77.1 g/L;
800 Titriplex IV 4 g/L; ammonium chloride 26.75 g/L; acetic acid 96% 59 mL/L) and filled up to 40 mL.
801 The fluoride concentration was measured by using an ISE (fluoride-ion sensitive electrode Orion
802 type No. 9-09-00 and a reference electrode Orion type No. 90-02-00) by adding stepwise a 10
803 mg/L fluoride standard. The values obtained in mV were used for fluoride quantification (Table
804 S4).
805
806 **14. Steady-state kinetics experiments**
807 Steady-state kinetic parameters for P3 were measured at a protein concentration of 0.5 mg/mL.
808 The substrate fluoroacetate was added to final concentration of 0.4 mM; 0.8 mM; 1.6 mM; and
809 3.0 mM in 100mM Tris-SO_4 (pH 8.5) and 20 mM; 50 mM; 100 mM in 300mM Tris-SO_4 (pH 8.5).
810 The reaction mixtures were incubated at 37 °C for the duration of the reaction period. Samples
811 were quenched with acetonitrile at timepoints: 0 min, 10 min, 20 min, 40 min, 50 min, 60 min, 75
812 min, 90 min, 105 min, 120 min, 6h for the sample containing 0.4 mM-3 mM fluoroacetate and 0
813 min, 10 min, 30min, 60 min 90 min and 120 min for the samples containing 20-100 mM
814 fluoroacetate. Following quenching, samples were centrifuged at 15000 RCF for 5 min and the
815 supernatant was transferred into plastic vials for subsequent fluoride concentrations using
816 identical methods as described in the above IC section. For each substrate concentration, the
817 slope was calculated over the linear range of this curve, giving the velocity of product formation
818 (mM fluoride/min). All the experiments were performed with three independent biological
819 replicates and three technical replicates per biological replicate and per substrate concentration.
820 Michaelis-Menten equations were used to calculate the kinetic parameters K_m, k_{cat}, and k_{cat}/K_m. A
821 non-linear regression analysis was used to fit the data to the Michaelis–Menten equation using
822 the R package "nlstools" (91) with 1000 non-parametric bootstrap resamplings to estimate the
823 values and standard errors for the kinetic parameters (Table S6).

824 **References**
825

826 1. Current Contributions of Organofluorine Compounds to the Agrochemical Industry. *iScience*
827 **23**, 101467 (2020).
828
829 2. M. Inoue, Y. Sumii, N. Shibata, Contribution of Organofluorine Compounds to
830 Pharmaceuticals. *ACS Omega* **5**, 10633–10640 (2020).

831 3. Y. Pan, The Dark Side of Fluorine. *ACS Med. Chem. Lett.* **10**, 1016–1019 (2019).

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

834 5. R. Okeda, *et al.*, Experimental neurotoxicity of 5-fluorouracil and its derivatives is due to
835 poisoning by the monofluorinated organic metabolites, monofluoroacetic acid and alpha-
836 fluoro-beta-alanine. *Acta Neuropathol.* **81**, 66–73 (1990).

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

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

841 8. N. Koppel, V. Maini Rekdal, E. P. Balskus, Chemical transformation of xenobiotics by the
842 human gut microbiota. *Science* **356** (2017).

843 9. B. Javdan, *et al.*, Personalized Mapping of Drug Metabolism by the Human Gut
844 Microbiome. *Cell* **181**, 1661–1679.e22 (2020).

845 10. P. Lehouritis, *et al.*, Local bacteria affect the efficacy of chemotherapeutic drugs. *Sci. Rep.*
846 **5**, 1–12 (2015).

847 11. L. T. Geller, *et al.*, Potential role of intratumor bacteria in mediating tumor resistance to the
848 chemotherapeutic drug gemcitabine. *Science* **357**, 1156–1160 (2017).

849 12. P. Spanogiannopoulos, *et al.*, Host and gut bacteria share metabolic pathways for anti-
850 cancer drug metabolism. *Nat Microbiol* **7**, 1605–1620 (2022).

851 13. B. A. Jennings, *et al.*, Evaluating predictive pharmacogenetic signatures of adverse events
852 in colorectal cancer patients treated with fluoropyrimidines. *PLoS One* **8**, e78053 (2013).

853 14. P. Lehouritis, *et al.*, Local bacteria affect the efficacy of chemotherapeutic drugs. *Sci. Rep.*
854 **5**, 1–12 (2015).

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

857 16. C. K. Davis, R. I. Webb, L. I. Sly, S. E. Denman, C. S. McSweeney, Isolation and survey of
858 novel fluoroacetate-degrading bacteria belonging to the phylum Synergistetes. *FEMS*
859 *Microbiol Ecol* **80**, 671–684 (2012).

860 17. S. Che, *et al.*, Structure-Specific Aerobic Defluorination of Short-Chain Fluorinated

861 Carboxylic Acids by Activated Sludge Communities. *Environ Sci Technol Lett* **8**, 668–674
862 (2021).

863 18. M. Yan, *et al.*, Defluorination of monofluorinated alkane by *Rhodococcus* sp. NJF-7 isolated
864 from soil. *AMB Express* **14**, 65 (2024).

865 19. R. B. Stockbridge, L. P. Wackett, The link between ancient microbial fluoride resistance
866 mechanisms and bioengineering organofluorine degradation or synthesis. *Nature*
867 *Communications* **15**, 1–12 (2024).

868 20. L. P. Wackett, S. L. Robinson, A prescription for engineering PFAS biodegradation.
869 *Biochem J* **481**, 1757–1770 (2024).

870 21. S. L. Robinson, Structure-guided metagenome mining to tap microbial functional diversity.
871 *Curr. Opin. Microbiol.* **76**, 102382 (2023).

872 22. P. W. Y. Chan, A. F. Yakunin, E. A. Edwards, E. F. Pai, Mapping the Reaction Coordinates
873 of Enzymatic Defluorination. (2011). https://doi.org/10.1021/ja200277d.

874 23. L. P. Wackett, S. L. Robinson, The ever-expanding limits of enzyme catalysis and
875 biodegradation: polyaromatic, polychlorinated, polyfluorinated, and polymeric compounds.
876 *Biochem. J* **477**, 2875–2891 (2020).

877 24. A. N. Khusnutdinova, *et al.*, Structural insights into hydrolytic defluorination of
878 difluoroacetate by microbial fluoroacetate dehalogenases. *FEBS J.* (2023).
879 https://doi.org/10.1111/febs.16903.

880 25. P. W. Y. Chan, *et al.*, Defluorination Capability of L-2-Haloacid Dehalogenases in the HAD-
881 Like Hydrolase Superfamily Correlates with Active Site Compactness. *Chembiochem* **23**,
882 e202100414 (2022).

883 26. A. M. Burroughs, K. N. Allen, D. Dunaway-Mariano, L. Aravind, Evolutionary genomics of
884 the HAD superfamily: understanding the structural adaptations and catalytic diversity in a
885 superfamily of phosphoesterases and allied enzymes. *J. Mol. Biol.* **361**, 1003–1034 (2006).

886 27. F. Rashid, V. Dubinkina, S. Ahmad, S. Maslov, J. M. K. Irudayaraj, Gut Microbiome-Host
887 Metabolome Homeostasis upon Exposure to PFOS and GenX in Male Mice. *Toxics* **11**
888 (2023).

889 28. L. Lemaire, M. C. Malet-Martino, M. de Forni, R. Martino, B. Lasserre, Cardiotoxicity of
890 commercial 5-fluorouracil vials stems from the alkaline hydrolysis of this drug. *Br. J. Cancer*
891 **66**, 119–127 (1992).

892 29. P. Hiseni, K. Rudi, R. C. Wilson, F. T. Hegge, L. Snipen, HumGut: a comprehensive human
893 gut prokaryotic genomes collection filtered by metagenome data. *Microbiome* **9**, 165
894 (2021).

895 30. A. Oren, G. M. Garrity, Valid publication of the names of forty-two phyla of prokaryotes. *Int.*
896 *J. Syst. Evol. Microbiol.* **71**, 005056 (2021).

897 31. J. A. Gerlt, *et al.*, Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool
898 for generating protein sequence similarity networks. *Biochim. Biophys. Acta* **1854**, 1019–
899 1037 (2015).

900 32. C. Gross, R. Felsheim, L. P. Wackett, Genes and enzymes of azetidine-2-carboxylate
901 metabolism: detoxification and assimilation of an antibiotic. *J. Bacteriol.* **190**, 4859–4864
902 (2008).

903 33. S. Farajollahi, *et al.*, Defluorination of Organofluorine Compounds Using Dehalogenase
904 Enzymes from *Delftia acidovorans* (D4B). *ACS Omega* (2024).
905 https://doi.org/10.1021/acsomega.4c02517.

906 34. P. Mehrabi, *et al.*, Substrate-Based Allosteric Regulation of a Homodimeric Enzyme.
907 *Journal of the American Chemical Society* (2019). https://doi.org/10.1021/jacs.9b03703.

908 35. K. Lindorff-Larsen, *et al.*, Systematic validation of protein force fields against experimental
909 data. *PLoS One* **7**, e32131 (2012).

910 36. S. Piana, K. Lindorff-Larsen, D. E. Shaw, How robust are protein folding simulations with
911 respect to force field parameterization? *Biophys. J.* **100**, L47–9 (2011).

912 37. J. Abramson, *et al.*, Accurate structure prediction of biomolecular interactions with
913 AlphaFold 3. *Nature* (2024). https://doi.org/10.1038/s41586-024-07487-w.

914 38. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
915 583–589 (2021).

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

918 40. L. P. Wackett, Confronting PFAS persistence: enzymes catalyzing C-F bond cleavage.
919 *Trends Biochem Sci* **50**, 71–83 (2025).

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

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

925 43. C. K. Davis, S. E. Denman, L. I. Sly, C. S. McSweeney, Development of a colorimetric
926 colony-screening assay for detection of defluorination by micro-organisms. *Lett. Appl.*
927 *Microbiol.* **53**, 417–423 (2011).

928 44. A. Yacouba, *et al.*, *Raoultibacter phocaeensis* sp. nov., A New Bacterium Isolated from a
929 Patient with Recurrent *Clostridioides difficile* Infection. *Curr Microbiol* **79**, 263 (2022).

930 45. K. Miura, *et al.*, 5-FU metabolism in cancer and orally-administrable 5-FU drugs. *Cancers*
931 **2**, 1717–1730 (2010).

932 46. G. J. Peters, *et al.*, Induction of thymidylate synthase as a 5-fluorouracil resistance
933 mechanism. *Biochim. Biophys. Acta.* **1587**, 194–205 (2002).

934 47. A.A. Adhikari, et al. Development of a covalent inhibitor of gut bacterial bile salt hydrolases.
935 Nature Chemical Biology 16(3), 318-326 (2020).

936 48. E. Wang, C. Walsh, Suicide substrates for the alanine racemase of *Escherichia coli* B.

937 *Biochemistry* **17**, 1313–1321 (1978).

938 49. C. Tysoe, S.G. Withers Fluorinated mechanism-based inhibitors: common themes and
939 recent developments. Current Topics in Medicinal Chemistry 14(7), 865-874 (2014).

940 50. V.M. Rekdal, E.N. Bess, J.E. Bisanz, P.J. Turnbaugh, E.P. Balskus, Discovery and
941 inhibition of an interspecies gut bacterial pathway for Levodopa metabolism. Science 364:
942 eaau6323 (2019).

943 51. U. Dorgerloh, R. Becker, M. Kaiser, Evidence for the Formation of Difluoroacetic Acid in
944 Chlorofluorocarbon-Contaminated Ground Water. *Molecules* **24** (2019).

945 52. J. Herbst, U. Köster, R. Kerssebaum, W. Dekant, Role of P4502E1 in the metabolism of
946 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-ethane. *Xenobiotica* **24**, 507–516 (1994).

947 53. A. P. Bhat, W. C. K. Pomerantz, W. A. Arnold, Finding Fluorine: Photoproduct Formation
948 during the Photolysis of Fluorinated Pesticides. *Environ. Sci. Technol.* **56**, 12336–12346
949 (2022).

950 54. V. N. Uversky, P. Kulkarni, Intrinsically disordered proteins: Chronology of a discovery.
951 *Biophys. Chem.* **279**, 106694 (2021).

952 55. L. C. James, D. S. Tawfik, Conformational diversity and protein evolution--a 60-year-old
953 hypothesis revisited. *Trends Biochem. Sci.* **28**, 361–368 (2003).

954 56. S. Sharma, M. R. Schiller, The carboxy-terminus, a key regulator of protein function. *Crit.*
955 *Rev. Biochem. Mol. Biol.* **54**, 85–102 (2019).

956 57. M. Kumar, *et al.*, ELM-the Eukaryotic Linear Motif resource-2024 update. *Nucleic Acids*
957 *Res.* **52**, D442–D455 (2024).

958 58. T. Zarin, *et al.*, Identifying molecular features that are associated with biological function of
959 intrinsically disordered protein regions. *Elife* **10** (2021).

960 59. D. P. Sargeant, *et al.*, Secondary structure, a missing component of sequence-based
961 minimotif definitions. *PLoS One* **7**, e49957 (2012).

962 60. I. S. Hsu, *et al.*, A functionally divergent intrinsically disordered region underlying the
963 conservation of stochastic signaling. *PLoS Genet.* **17**, e1009629 (2021).

964 61. I. Pritišanac, *et al.*, A functional map of the human intrinsically disordered proteome.
965 *bioRxiv* (2024).

966 62. M. Y. Lobanov, E. I. Furletova, N. S. Bogatyreva, M. A. Roytberg, O. V. Galzitskaya, Library
967 of disordered patterns in 3D protein structures. *PLoS Comput. Biol.* **6**, e1000958 (2010).

968 63. N. L. Ogihara, *et al.*, Design of three-dimensional domain-swapped dimers and fibrous
969 oligomers. *Proceedings of the National Academy of Sciences* **98**, 1404–1409 (2001).

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

972 65. J. I. Tietz, *et al.*, A new genome-mining tool redefines the lasso peptide biosynthetic

973 landscape. *Nat Chem Biol* **13**, 470–478 (2017).

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

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

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

981 69. M. M. Bradford, A rapid and sensitive method for the quantitation of microgram quantities of
982 protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).

983 70. M. Mirdita, *et al.*, ColabFold: making protein folding accessible to all. *Nat. Methods* **19**,
984 679–682 (2022).

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

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

989 73. P. Bauer, B. Hess, E. Lindahl, GROMACS 2022.5 Manual. (2023). Available at:
990 https://zenodo.org/record/7586765.

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

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

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

997 77. U. Essmann, *et al.*, A smooth particle mesh Ewald method. *J. Chem. Phys.* **103**, 8577–
998 8593 (1995).

999 78. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J.*
1000 *Chem. Phys.* **126**, 014101 (2007).

1001 79. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, J. R. Haak,
1002 Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684–3690
1003 (1984).

1004 80. UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids
1005 Res. 51, D523–D531 (2023).

1006 81. F. Sievers, et al., Fast, scalable generation of high-quality protein multiple sequence
1007 alignments using Clustal Omega. Mol. Syst. Biol. 7, 539 (2011).

1008 82. P. J. A. Cock, et al., Biopython: freely available Python tools for computational molecular

1009 biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009).

1010 83. S. Kalyaanamoorthy, B. Q. Minh, T. K. F. Wong, A. von Haeseler, L. S. Jermiin,
1011 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**,
1012 587–589 (2017).

1013 84. B. Q. Minh, *et al.*, Corrigendum to: IQ-TREE 2: New Models and Efficient Methods for
1014 Phylogenetic Inference in the Genomic Era. *Mol Biol Evol* **37**, 2461 (2020).

1015 85. J. Huerta-Cepas, F. Serra, and P. Bork. ETE 3: reconstruction, analysis, and visualization
1016 of phylogenomic data. Mol Biol Evol 33, 1635-1638 (2016).

1017 86. S. L. Robinson, M. D. Smith, J. E. Richman, K. G. Aukema, L. P. Wackett, Machine
1018 learning-based prediction of activity and substrate specificity for OleA enzymes in the
1019 thiolase superfamily. *Synth. Biol.* **5** (2020).

1020 87. S. L. Robinson, *et al.*, Global analysis of adenylate-forming enzymes reveals β-lactone
1021 biosynthesis pathway in pathogenic Nocardia. *J. Biol. Chem.* **295**, 14826–14839 (2020).

1022 88. M. N. Wright, A. Ziegler, Ranger: A fast implementation of random forests for high
1023 dimensional data in C++ and R. *J. Stat. Softw.* **77** (2017).

1024 89. W. R. Atchley, J. Zhao, A. D. Fernandes, T. Drüke, Solving the protein sequence metric
1025 problem. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 6395–6400 (2005).

1026 90. E. Wright, Accurately clustering biological sequences in linear time by relatedness sorting.
1027 *Nat. Commun.* **15**, 3047 (2024).

1028 91. F. Baty, *et al.*, A Toolbox for Nonlinear Regression in R: The Package nlstools. *J. Stat. Soft.*
1029 **66**, 1–21 (2015).

1030