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

Fluorine is more abundant in Earth's crust than carbon and nitrogen, yet natural fluorinated compounds are rare in biological systems. In contrast, 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. To advance our understanding of the structural and sequence differences between defluorinating and non-defluorinating dehalogenases, we developed a 96-well colorimetric fluoride assay amenable to bacterial culture-based screening. Additionally, we trained machine learning models to predict defluorination activity and experimentally validated our findings to identify key features of defluorinating enzymes. This opens new research directions in microbiome research for the metabolism of fluorinated compounds, with implications for therapeutic development and enzyme engineering to break carbon-fluorine bonds.

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 confers defluorination activity to a non-defluorinating dehalogenase. Machine learning 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) or activated sludge (17), none were previously associated with the human gut microbiota.

Previously, the discovery of 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- (15) or structure-guided metagenome mining (21) methods 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 found in soil, catalyze carbon-

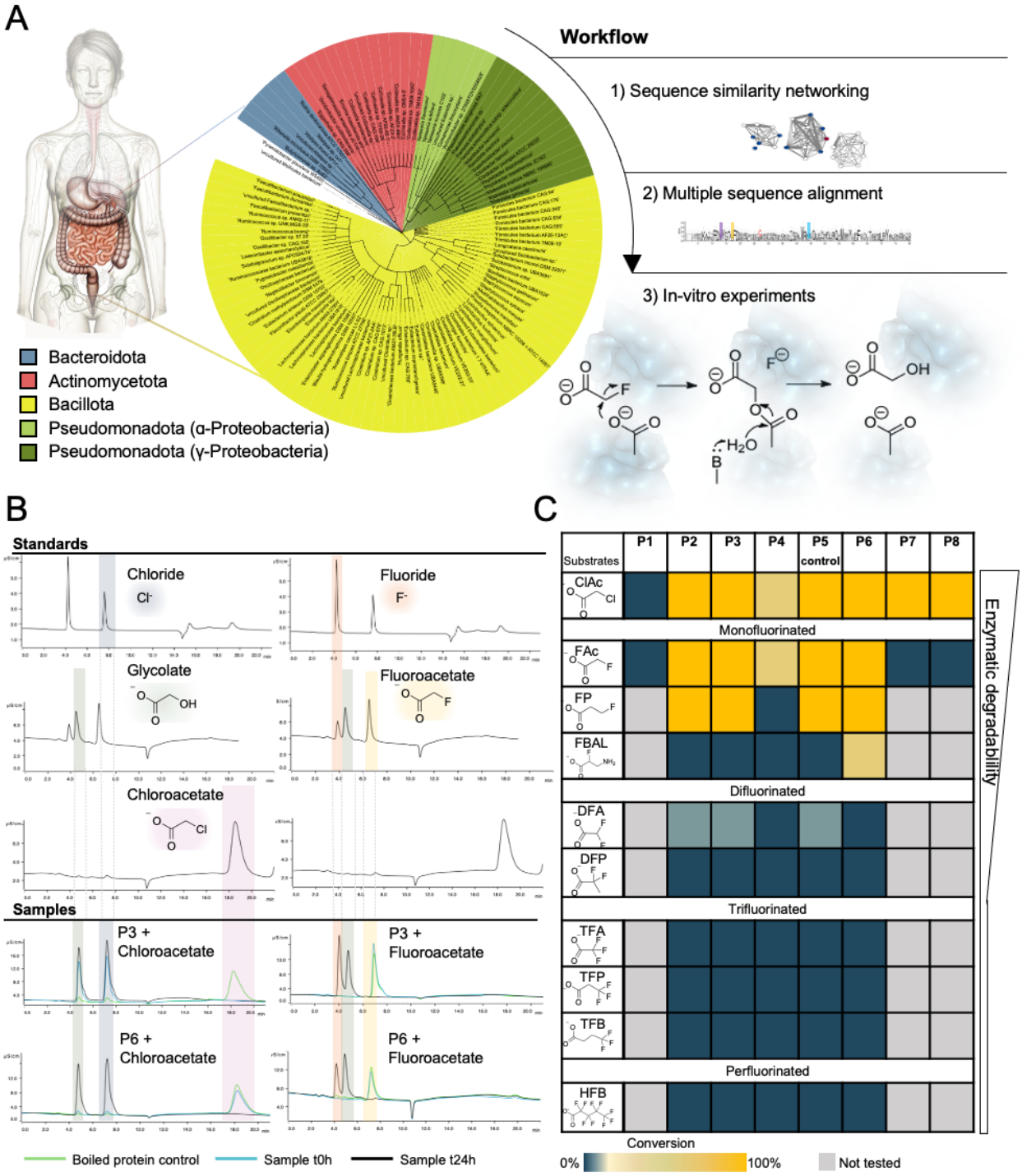
84 halide bond cleavage via a SN_2-like mechanism through nucleophilic attack by an aspartate to
85 displace the halide (25, 27). The active sites of defluorinating and dechlorinating HADs are highly
86 similar (25), thus it remains challenging to mine metagenomic datasets to accurately identify
87 HADs 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). We tested this MD-based classification
90 method and found it did not generalize to a larger set of defluorinating and non-defluorinating
91 HADs, which motivated our search for alternative accessible features to predict metabolic
92 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-terminal region of HADs in defluorination.

102 **Results**
103
104 ***The human gut microbiota encodes HAD-like enzymes***
105
106 To investigate the potential for defluorination by human gut microbial enzymes, we first tested
107 defluorination activity for fluoroacetate as a simple fluorinated substrate and a known 5-FU
108 degradation product (28). Notably, prior studies identified the soil bacterium, *Rhodococcus jostii*
109 RHA1, as a degrader of fluoroacetate along with the discovery of its responsible defluorinating
110 HAD superfamily enzyme classified as a L-2-haloacid dehalogenase (15). These enzymes
111 primarily dehalogenate small chlorinated alkanoic acids with only a few demonstrating the
112 capability to cleave the more recalcitrant carbon-fluorine bond (25). We therefore used the
113 *Rhodococcus jostii* RHA1 HAD (Rha0230) for a protein homology search in the HumGut database
114 (29). To our surprise, we detected over 500 HAD homologs encoded in human gut metagenomes
115 (Fig. S1, Table S1).
116
117 The identified HAD-encoding genes were predominantly found in bacteria belonging to the phylum
118 Bacillota, formerly named the Firmicutes (30). However, homologs were also prevalent in the
119 Actinomycetota, Pseudomonadota, and Bacteroidota (Fig. 1A, Fig. S1). To further elucidate the
120 relationships among these proteins, we constructed a protein sequence similarity network (Fig.
121 S2), which revealed several distinct HAD clusters (29, 31). Two of these clusters contained
122 characterized members of the L-2-haloacid dehalogenase subfamily (25) within the larger HAD
123 superfamily (26). Other clusters could be assigned to other HAD superfamily family functions such
124 as phosphonate and azetidine hydrolysis (26, 32). For subsequent analysis, we focused on
125 clusters containing the 2-haloacid dehalogenases and curated the dataset through a multiple
126 sequence alignment to eliminate redundant sequences and confirmed the presence of the

127 catalytic residues (Fig. S3). We selected seven new representatives for in vitro experiments (Fig.
128 S2, Table S2). The selected sequences from the human gut microbiota showed low amino acid
129 sequence identity to previously characterized HADs (18% - 46%; mean: 29%; Table S1). We also
130 investigated the genomic context by extracting ± six genes on either side of the query HADs
131 (Table S3). No CLC^F or Fluc (*crcB*-encoded) fluoride ion transporters or other relevant protein
132 families were detected. Therefore, the role of these HADs in defluorination could not be inferred
133 based on genomic co-localization. The lack of functional prediction based on sequence similarity
134 or genomic context therefore motivated experimental validation of these unknown HADs.
135
136 ***Human gut microbial dehalogenases are active in carbon-fluorine bond cleavage***
137
138 To screen for in vitro defluorination activity, we heterologously expressed, purified, and tested
139 seven selected gut microbial dehalogenases with ten different halogenated substrates (nine
140 fluorinated organic acids and one chlorinated substrate, chloroacetate, Fig. 1, Fig. S4 Table S4).
141 The HAD from the soil bacterium *Rhodococcus jostii* RHA1 (P5, Protein Data Bank (PDB) ID:
142 3UMG) was included as a positive defluorination control as described previously. All proteins
143 tested except for one (*Enterocloster aldenensis*, P1, Fig. 1C) showed dehalogenase activity with
144 at least one substrate (Fig. S5-S12).
145
146 Two enzymes hydrolyzed chloroacetate but were not active with any fluorinated substrates (P7
147 and P8, Fig. 1C). The remaining four enzymes hydrolyzed fluoroacetate, producing fluoride ion
148 and glycolate stoichiometrically, as measured by ion chromatography (Fig. 1B). Compounds with
149 perfluorinated methyl ($-CF_3$) or perfluorinated methylene ($-CF_2-$) groups including trifluoroacetate,
150 difluoropropionate, trifluoropropionate, trifluoropentanoate, and heptafluorobutanoate were not
151 accepted by any of the eight HADs tested (Fig. 1C). Four of the enzymes, including the control
152 were active with known transformation products of fluorinated xenobiotics, namely, 2-
153 fluoropropionate, α-fluoro-β-alanine, and the polyfluorinated compound, difluoroacetate. These
154 active defluorinating enzymes originate from diverse gut bacterial families including
155 Eggerthellaceae (*Gordonibacter*, P2 and P3), Moraxellaceae (*Acinetobacter*, P4), and
156 Christensellaceae (*Guopingia*, P6).
157
158 We further examined the impact of oxygen on one defluorinating enzyme P3 as a representative
159 for our gut microbial HADs. As expected based on the established L-2 haloacid dehalogenase
160 mechanism (24–26), P3 cleaved fluoroacetate to produce glycolate and fluoride ions under both
161 aerobic and anaerobic conditions (Fig. S13). To test whether we could detect the defluorination
162 activity of the gut bacterium *Gordonibacter pamelaeae* anaerobically in gut microbial medium
163 (GMM) with fluoroacetate and measured fluoride concentrations after 24 hours using a fluoride
164 ion electrode (Table S5, Methods). Defluorination activity was not detectable, perhaps due to
165 challenges in expression or transport. Due to the limited genetic tractability of the native gut strains
166 encoding these enzymes, we instead focused efforts on in vitro enzyme characterization. We
167 measured the steady-state enzyme kinetics for P3 with fluoroacetate and compared our results
168 to other defluorinating enzymes including a L-2 haloacid dehalogenase (DeHa2) and
169 fluoroacetate dehalogenases (DeHa4) from *Delftia acidovorans* D4B (33) and the fluoroacetate
170 dehalogenase from *Rhodopseudomonas palustris* (RPA1163, (34)). Briefly, P3 has a K_m of 59.8

171 ± 7.7 mM approximately an order of magnitude higher than the K_m values of DeHa2 and DeHa4
172 (Table S6), although lower than the K_m of RPA1163 (86.8 ± 13.7)(34). The k_{cat} of 0.25 ± 0.02 s^{-1}
173 is about four times less than the k_{cat} of RPA1163 (Table S6). The weak substrate binding and
174 correspondingly high off rate suggested by the K_M may be a contributing factor to the overall low
175 catalytic efficiency.
176



177
178

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 by sequence similarity networking, multiple sequence alignment to 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 degradablity 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 previously, 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). Since the original MD publication (25), 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 the 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 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 elimination to produce HF (40).

226
227 ***Phylogenetic analysis does not differentiate defluorinating and non-defluorinating***
228 ***enzymes***
229
230 To further investigate methods to differentiate defluorinating and non-defluorinating enzymes, we
231 generated a phylogenetic tree of the HAD-like family protein sequences (Fig. S20, Table S9). The
232 tree topology revealed that the defluorinating and non-defluorinating enzymes did not form
233 separate or easily-distinguishable clades. The dehalogenating enzymes originated from eight
234 diverse bacterial and archaeal phyla (Fig. S21) which did not group based on the taxonomy of the
235 source organisms.
236
237 ***Rational active site engineering converts an inactive enzyme to a dechlorinating enzyme***
238
239 As neither MD simulations nor phylogenetic analysis yielded conclusive results to distinguish
240 defluorinating from non-defluorinating HADs, we next sought to experimentally test the role of
241 specific residues in the active site. To identify potentially significant amino acids involved in the
242 defluorination process, we used two of seven in vitro tested proteins sharing the highest sequence
243 and structural similarity but distinct activities from each other (Fig. S21) (54% sequence identity,
244 68% sequence similarity). P1, a non-dehalogenating enzyme, and P6 a defluorinating enzyme.
245 Based on comparison between the active sites of the two enzymes, there were only four residues
246 which were significantly different in the active site vicinity including the halide-binding pocket and
247 substrate tunnel (Fig. S21). We predicted these residues may provide substrate specificity for
248 defluorination, therefore we first aimed for rational engineering of the non-dehalogenating enzyme
249 P1 (*E. aldenensis*) via site-directed mutagenesis to generate single, double, and triple variant
250 combinations as well as the full quadruple variant P1 (F54I, T57M, H74F, F187S) (Fig. S22-S25).
251
252 Among the single variants tested, only one (F54I) showed dehalogenation activity. The F54I
253 variant hydrolyzed chloroacetate but did not hydrolyze fluorinated substrates. Based on structural
254 modeling, F54 is positioned at the entrance channel to the substrate binding pocket. In the non-
255 dehalogenating protein P1 (*E. aldenensis*) the larger phenylalanine may exclude chloroacetate at
256 the binding pocket entrance. In contrast, the dehalogenating protein P6 (*Guopingia tenuis*)
257 features an isoleucine at this position, potentially enabling access for chloroacetate. While we did
258 achieve a functional dehalogenase through the single-point mutation in P1 (*E. aldenensis*), the
259 individual active site changes did not singularly confer defluorination activity. We hypothesized
260 epistatic interactions may promote defluorination. We therefore used the F54I variant as the
261 template for generating double, triple, and quadruple active site variants. In the P1 (F54I, H74F)
262 double variant and all triple variants we achieved improved dechlorination activity (Fig. 2C).
263 Surprisingly, we did not observe defluorination activity in any of the rationally-engineered variants,
264 even though the active site residues were 100% identical to the defluorinating reference enzyme
265 P6 (*G. tenuis*). This result suggests that additional second shell interactions or other
266 conformational changes (41) not localized in the active site may be critical for defluorination
267 activity.
268

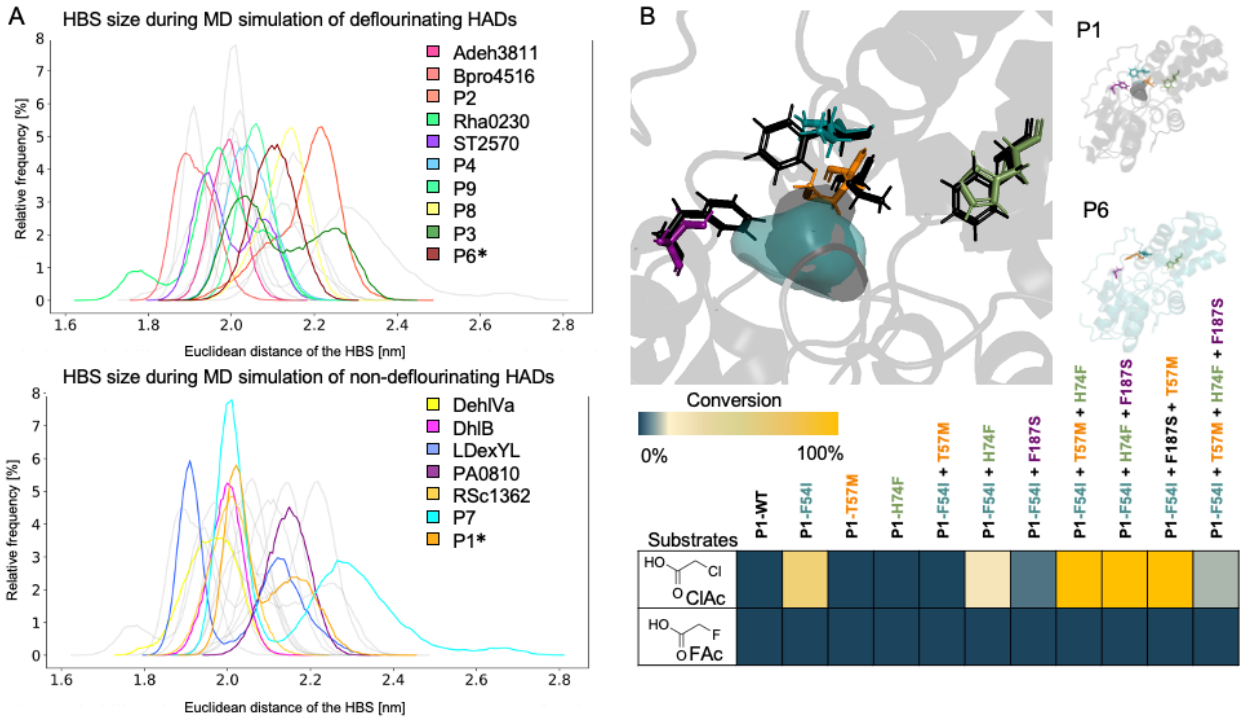


Figure 2. Defluorination capability does not correlate with the active site compactness or amino acid sequence identity. A) Halide binding site (HBS) size in nm during molecular simulations. Distribution of the Euclidean distances of the 3 C_α atoms of the HBS during simulation from 500 ns to 1000 ns. *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

297 assay which can be reduced to below the limit-of-detection with washing steps (Figure S28,
298 Methods).
299
300 To test our miniaturized screening method, we selected eleven candidate dehalogenases
301 spanning eleven additional gut microbial genera (Table S12). Based on the colorimetric screen,
302 we detected an active gut microbial defluorinase from *Priestia megaterium* sharing less than 35%
303 amino acid identity to our reference defluorinase (PDB ID: 3UMG). We purified the *P. megaterium*
304 enzyme to homogeneity and validated its defluorination activity using ion chromatography (Fig.
305 S29). This highlights the potential of microtiter-well assays to rapidly identify defluorinating
306 enzymes.
307
308 With our new microtiter defluorination assay in hand, we next screened the alanine scanning
309 library with a dual-assay approach for both defluorination and dechlorination activity (Fig. 3A).
310 The screening was conducted in two independent experimental campaigns each with biological
311 replicates as further described in the Methods section, with and without dilution, and normalization
312 for protein concentration (Experiment 1 and Experiment 2, respectively). Across independent
313 experiments and biological replicates of the alanine scanning library (Methods), 39.8-41.7% of
314 alanine variants lost defluorination activity. A slightly lower proportion of variants (33.0-33.5%)
315 lost dechlorination activity in experiment one. In experiment two 31%-37% of alanine variants lost
316 defluorination activity and 39-41% lost dechlorination activity. We identified the top 19 alanine
317 variants which showed dechlorination activity but lost defluorination activity in both experiments
318 (Fig. S32-S34, Table S14-15). Overall, although Experiment 1 and 2 varied due to natural
319 biological variability in growth and expression, despite protein normalization correction (Fig. S30-
320 31), both experiments identified the same key regions and residues to be important for
321 defluorination (Fig. S35, Table S13). Subsequent machine learning results are reported here for
322 Experiment 1 for clarity (Exp. 2 results are reported in Table S16 and Fig. S36).
323
324 ***Machine learning highlights important protein regions for defluorination***
325
326 To analyze our alanine scanning data, we used the random forest algorithm to train a machine
327 learning model to predict defluorination activity. Random forest is suitable for both regression and
328 classification tasks even under low-data regimes (n=474 data points from the alanine scanning
329 experiments). Our regression model trained on 80% of the alanine scanning data was capable of
330 predicting quantitative changes in defluorination activity as the result of single alanine point
331 mutations with the remaining 20% of the data with a root-mean square error (RMSE) of 0.2 ± 0.02
332 enzyme activity units (on a min-max normalized activity scale between 0-1 defluorination activity).
333 Based on the relative feature importance, the most predictive feature was the position of each
334 amino acid within the protein sequence rather than the size, charge, identity, codon diversity, or
335 polarity of each amino acid converted to alanine (Fig. 3B). These results showing position carried
336 the highest importance for defluorination activity corresponds with the visible 'defluorination'
337 hotspots on the protein structure (Fig. 3C).
338
339 To train a more generalizable model with applications beyond our selected alanine scanning
340 libraries P1 (F54I, T57M, H74F, F187S) and P6, we used random forest to classify 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 (F1-score = 0.928 ± 0.028, Table S16). To test the generalizability of our classification model, we used the model to predict defluorination activity for a new set of five uncharacterized HAD homologs (Methods, Fig. S35). We obtained a defluorination probability score of 0.9 for a HAD (WP_139652913.1) encoded in the genome of *Raoultibacter phocaeensis* Marseille-P8396^{T} isolated from human stool samples (44). We heterologously expressed and confirmed positive defluorination activity in the *R. phocaeensis* HAD. The other four HADs in the dataset expressed well in *E. coli* but were not active in defluorination. Our models predicted a low defluorination probability of 0.5 or less (Fig. S37, Table S17), supporting the ability of our models to discern high-probability defluorinating HADs from others.

Although recognizing the limited scope of this model to HAD-like family enzymes, this classification analysis pinpointed residues that were important for defluorination activity (Fig. S35-36). Strikingly, the top five most important residues (Experiment 1) for defluorination activity prediction accuracy were localized in the C-terminus (final 41 amino acids). 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 = 0.728 ± 0.129, Experiment 1, Table S16).

Additionally, we computed the root mean square fluctuations (RMSF) from our molecular dynamics (MD) simulations of P6 (*G. tenuis*), allowing us to compare the spatial fluctuations of different regions within the protein structural model. Our analysis identified two regions exhibiting high RMSF values. Both regions are situated at the channel entry leading to the active site (Fig. S38), 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.

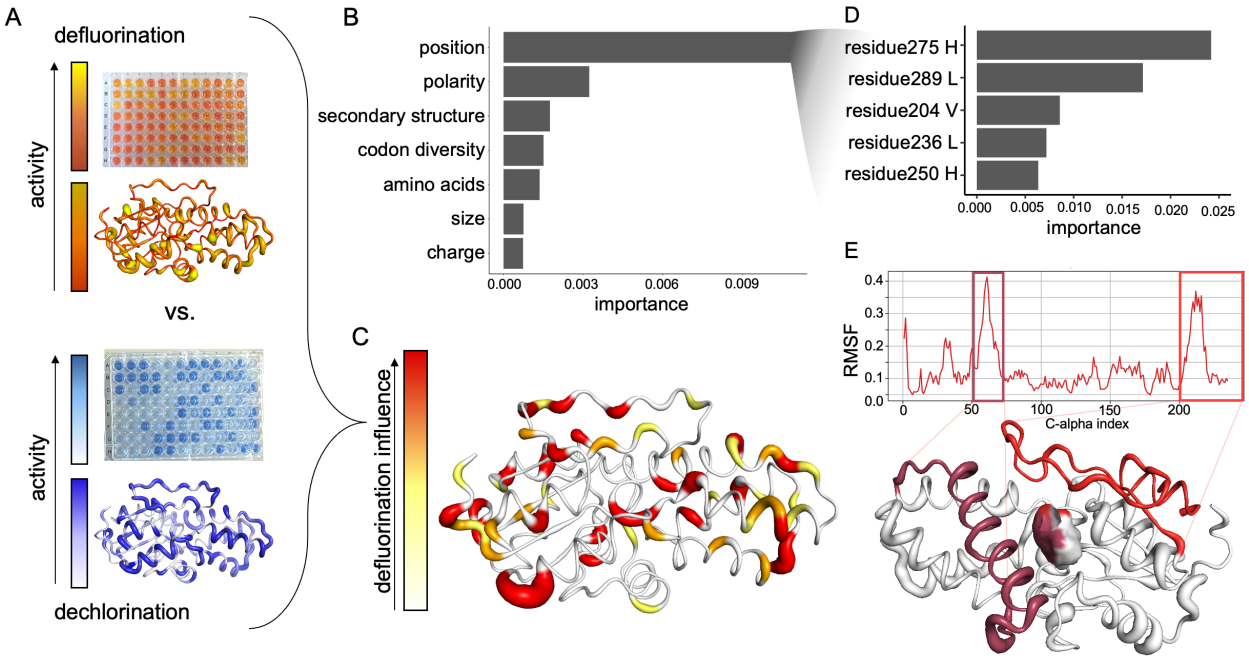


Figure 3. The intrinsically disordered C-terminal region is important for defluorination activity. A) Results of Experiment 1 colorimetric assays for defluorination and dechlorination activity, respectively, mapped onto the P6 protein model. The color gradient from white (low activity) to yellow (medium 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 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 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.

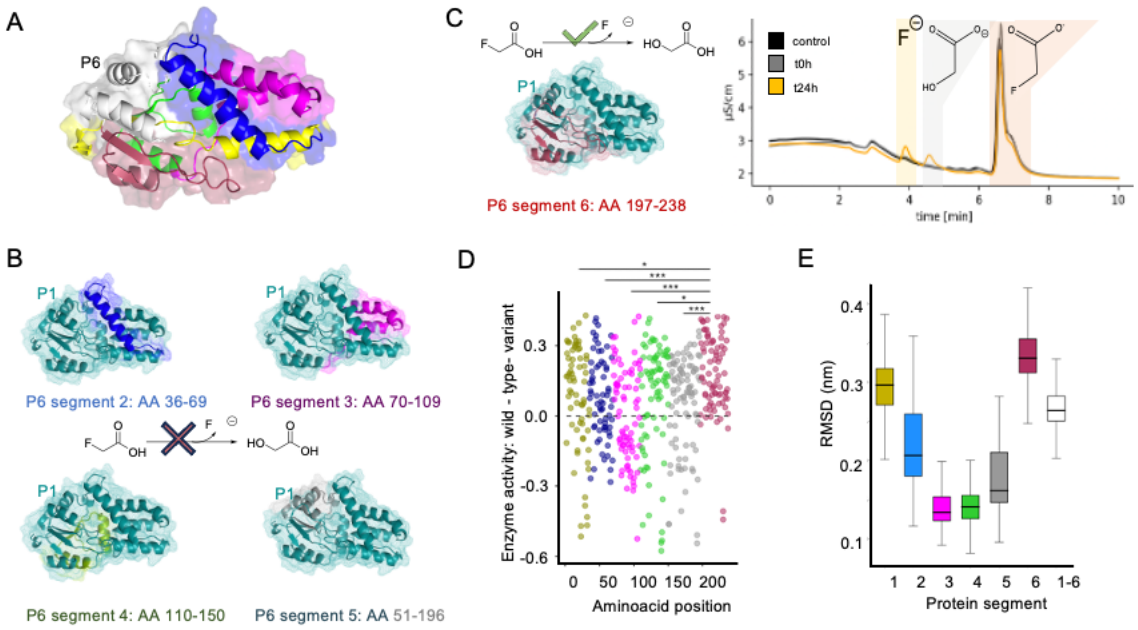


Figure 4. Swapping the C-terminus of a non-defluorinating HAD with the C-terminus of a defluorinating homologous enzyme enables defluorination activity. A) Representation of the defluorinating protein P6. The colored parts represent the exchanged segments used for generating the chimeric protein variants 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 P3 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 (dUMP) by forming a covalent ternary complex with TS and methylenetetrahydrofolate (CH_2-THF)(46). This mechanism, known as suicide inhibition, has been a significant driving force

429 behind research into the interactions and defluorination activities of human enzymes. Studies on
430 fluorinated amino acid analogs that inhibit alanine racemases (47) other pyridoxal phosphate
431 (PLP)-dependent enzymes (48), fluorinated phosphate esters that target acetylcholinesterases
432 (49) and serine hydrolases (50), have not only enhanced our understanding of the previously
433 unknown ability of certain proteins to interact with fluorinated substrates but also contributed to
434 advancements in human health. Although some of these enzymes were found in *Escherichia coli,*
435 none of these studies have directly linked them to human gut strains or utilized enzymes
436 specifically encoded by gut microbes.

437 In this study, we also identified four new HADs active on difluoroacetate, doubling the number of
438 hydrolytic dehalogenases experimentally shown to defluorinate -CF_2 groups (24). Difluoroacetate
439 is a transformation product of various xenobiotics including the refrigerant R1113,
440 chlorotrifluoroethylene, (51) and the anesthetic 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-
441 ethane (52). It is a byproduct of the photodegradation of penoxsulam, a widely used pesticide
442 (53). Difluoroacetate may also come in contact with gut microbes through ingestion via pesticide
443 residues on food (53). In general, fluorinated substances enter our environment and can seep
444 into groundwater, where they are subsequently consumed by humans, thus coming in contact
445 with the gut microbiota. This study is limited in its scope to overexpressed recombinant HADs
446 sourced from the gut microbiota. In vivo and pure culture experiments remain inconclusive due to
447 unknown conditions for gene expression and interference from complex matrices, requiring further
448 investigation.
449
450 Specificity-determining residues or structural features that differentiate defluorinating enzymes
451 from their non-defluorinating counterparts in the L-2 haloacid dehalogenases has remained
452 elusive (25). Alanine scanning proved useful to identify protein variants that retained
453 dechlorination activity but lost their defluorination capability and identified several amino acids
454 critical for defluorination. Surprisingly, these amino acids were primarily located in the peripheral
455 regions of the protein with a higher percentage of significant residues located in the C-terminal
456 segment of the protein. These observations, along with results from more recent studies (54),
457 support the growing complex understanding of protein function than the simplistic lock-and-key
458 model (55). Consequently, our protein engineering efforts shifted towards engineering chimeric
459 proteins by replacing segments of non-defluorinating protein variants with segments from a
460 defluorinating enzyme. We observed chimeric proteins with weak, but robust and reproducible
461 defluorination activity only when the C-terminal segment was exchanged.
462
463 The C-terminal regions of proteins have been extensively studied in eukaryotic cells (56, 57),
464 where 'short linear motifs' within intrinsically disordered regions (58) have been associated with
465 a broad spectrum of functions, including substrate recognition (56, 59), signaling (60), and
466 disease risk (56, 61). This functional diversity is attributed to the solvent exposure and higher
467 degree of disorder of C-termini, which facilitate protein-protein interactions (62) and
468 oligomerization facilitated through natural C-domain swapping (63). Previously, in a HAD
469 superfamily phosphatase, the flexible C-terminal tail was shown to interface with adjacent
470 monomers to facilitate product release (41). Our MD simulations suggested gut microbial HAD C-
471 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. 501100001711-209124, PZPGP2_209124), the Uniscientia Foundation, and the Helmut Horten Foundation.

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 a tblastn search was performed to identify homologs of the WP_011593529.1 and

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

560 at 4°C for 45 min. For benchtop purification, supernatants were applied to a column packed with
561 Ni-NTA agarose beads (Qiagen) pre-equilibrated in buffer A. The column was subsequently
562 washed with buffer A and then subjected to two washes with buffer B (40 mM Tris-HCl, 500 mM
563 NaCl, 10% glycerol and 40 mM imidazole). Enzymes were eluted from the column using buffer C
564 (40 mM Tris-HCl, 500 mM NaCl, 10% glycerol, and 500 mM imidazole) The fractions with the
565 highest protein concentration was determined qualitatively by Bradford assay(69). Imidazole and
566 NaCl were removed by running the sample over a PD-10 desalting column (Cytiva) and eluting in
567 buffer (100 mM Tris-SO_4, pH = 8.5). The concentration of the purified proteins were determined
568 by Qubit™ protein assay and the proteins were flash frozen using liquid nitrogen in 500 µL aliquots
569 and subsequently stored at -80°C.
570

571 **3. Assessment of enzyme-substrate specificity**

572 Stock solutions containing 100 mM of the test substrates listed in Table S5 dissolved in water
573 were prepared. The concentration of the Ni-NTA purified proteins were determined by using the
574 Qubit™ protein assay and the proteins were diluted to a final concentration of 0.5 mg/mL in 100
575 mM Tris-SO_4 buffer, pH 8.5. As controls the proteins were boiled for 10 min at 100°C. The test
576 substrates were added to a final concentration of 1 mM and the samples were incubated at 37°C.
577 50 µL samples were taken for the different timepoints and mixed with 50 µL acetonitrile to quench
578 the reaction. The samples were centrifuged at 13 000 RPM for 5 min and the supernatant
579 transferred into plastic vials for ion chromatography (IC).
580

581 An IC system (Metrohm, Switzerland) was used consisting of an IC sampling center 889 coupled
582 to an ion chromatograph (930 Compact IC Flex) with a conductivity detector. Samples were
583 incubated at 10°C in the sample center with 10 µL injection volumes first onto a precolumn
584 (Metrosep A Supp 16, Guard, 4 mm) and subsequently onto a Metrosep A Supp 16 column (250
585 x 4.0 mm, 5 µm). The temperature in the column was kept constant at 30°C. The eluent consisted
586 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
587 and quantification was performed using MagIC Net Professional (version 3.3).
588 All experiments were conducted at least in triplicates.
589

590 **4. AlphaFold and structural comparison**

591 Structural models for the seven initial dehalogenase sequences selected from the HumGut
592 databases, were generated using AlphaFold2 (AF) ColabFold(70). Models with the highest rank
593 for each dehalogenase were prepared with relaxation using the Amber99SB force field(71).
594 Additionally, we used AF to generate models for the HADs PA0810, RSC1362, ST2570 and
595 RHA0230. The accuracy of AF models was compared to known X-ray crystal structures, where
596 available (25).
597

598 **5. Molecular dynamics simulations**

599 Molecular dynamics simulations were performed using the GROMACS package version 2022.5
600 (72, 73) compiled with CUDA support and run on NVIDIA GeForce RTX 2080 Ti graphic cards.
601 The crystal structures for HADs PA0810, RSC1362, ST2570 and RHA0230 were downloaded
602 from RCSB PDB (PDB IDs: 3UMC, 3UMB, 2W43 and 3UMG, respectively. For PDB files which
603 contained selenomethionine, selenomethionine was exchanged with methionine. Relaxed AF

models were also subjected to simulations. The system was prepared with GROMACS tools, parameterized with the force field Amber99SB-ILDN (74), and solvated in explicit water using the default water model TIP3P (75) 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 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 (76). 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 algorithm (77) and a Fourier spacing of 0.14 nm. The temperature was controlled using a modified Berendsen thermostat (78) 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 (79) 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 4.5×10^{-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 root mean square deviation (RMSD) of the simulated proteins relative to the starting structure (C_α 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 C_α atoms was tracked during the simulation after an equilibration period of 500 ns with the built-in tool gmx_distance (80). To measure the Root Mean Square Fluctuation (RMSF) of all C_α 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. S20-S21, Table S9) 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 (81) 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 (82) and the phylogenetic tree (Fig. S20-21) was generated using the IQ-TREE algorithm (83, 84). The model selection was performed by IQ-TREE model finder function (83). The best model was selected (substitution model Q.pfam+R8) and a maximum-likelihood tree tree was constructed with 1000 bootstrap replicates (85).

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. S21) (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 S21). 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. (43) 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 S11. Sodium phosphate monobasic (hereafter phosphate) interferes with the formation of the xylenol orange-fluoride complex at phosphate concentrations above 100 µM. The time dependence of phosphate interference was analyzed over a range of phosphate concentrations between 100 and 1000 µM (Fig. S28).

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 decanted 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.
Two independent experiments were performed.
The first experiment was performed in duplicate for all protein variants and assessed for defluorination activity. The second experiment was conducted in triplicate for all protein variants and tested for defluorination but with normalization to protein concentrations.

The absorbances used to calculate fluoride ion concentration was measured at 540 nm and the isosbestic point was measured at 480 nm using an Eon spectrophotometer (Biotek). The approximate fluoride concentration is calculated using the following formula: x=(abs(480 nm)) / (abs(540 nm)) where a ratio above 1.8 indicates a [F^-] greater than 100 µM. Strongly positive samples (>100 µM [F^-], x>1.8) appear yellow and negative samples appear red (<10 µM [F^-],

692 x<1.2). Fluoride calibration curves were prepared by spiking sodium fluoride amounts between 0
693 μM and 100 μM into cell supernatants taken from negative control samples. The assay shows a
694 2nd degree polynomial relationship for fluoride concentrations between 0 μM and 100 μM. Fluoride
695 concentrations above 100 μM were diluted 1:3 (Fig. S26) and quantified using a 1:3 diluted
696 calibration curve. More information and a spreadsheet to prepare the diluted and undiluted
697 calibration curves is available https://github.com/MSM-
698 group/gut_microbe_defluorination_paper/tree/main/Dehalogenation_colorimetric_assays/Experi
699 ment2. For the quantification of undiluted samples, 180 μl cell supernatant and 20 μL sodium
700 fluoride solution with 10x fluoride concentration was used.
701
702 **9. Chloride detection assay**
703 A chloride detection assay kit (QuantiChrom™ Chloride Assay Kit) was obtained from Huber lab
704 and assays were run according to the manufacturer's instructions with the following changes:
705 Briefly, *E. coli* T7 expression strains carrying the corresponding plasmids were grown in TB
706 medium in a 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 μg/mL
707 spectinomycin until an optical density of 0.8 was reached. The cells were induced with 0.1mM
708 IPTG and incubated for 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant
709 was discarded, and the cells were washed twice with 2 mL 20 mM HEPES (pH 8). After
710 centrifugation at 4000 RCF for 10 min the cells were resuspended in 1 mL 20 mM HEPES (pH 8)
711 and chloroacetate was added to a final concentration of 30 mM prior to incubation at 37°C for 24
712 h. The cells were centrifuged at 4000 RCF for 15 min and 2.5 μL of the supernatant was
713 transferred to a new 96 well plate and 100 μL of the colorimetric working solution was added. The
714 chloride concentration was measured at 610 nm with an Eon spectrophotometer (Biotek). The
715 limit-of-quantification was 0.2 mM.
716 Two independent experiments were performed. The first experiment was performed in duplicate
717 for all protein variants and assessed for dechlorination activity. The second experiment was
718 conducted in triplicate for all protein variants and tested for dechlorination but with normalization
719 to protein concentrations.
720
721 **10. Whole-gene alanine-scanning mutagenesis and chimeragenesis**
722 High-throughput alanine-scanning mutagenesis was performed using the Q5® Site-Directed
723 Mutagenesis Protocol (New England Biolabs Inc.) with the following amendments. Briefly, back-
724 to-back forward and reverse primers were designed using NEBaseChanger
725 (https://nebasechanger.neb.com/), where the alanine codon substitution was contained in the
726 center of the forward mutagenic primer. Note that a histidine codon was substituted in place of
727 any native alanine codons. Primers were ordered from Integrated DNA Technologies. The
728 complete set of alanine (n=474) are available on https://github.com/MSM-group/biodefluorination-
729 paper/tree/main/Alanine_scanning_primers. A 6 μL PCR reaction, containing 1x Q5 Hot Start
730 High-Fidelity 2X Master Mix (New England Biolabs), 0.5 μM of the forward and reverse primer,
731 and 10 ng template DNA, was carried out for each variant. P6, WP_178618037.1 (*Guopingia*
732 *tenuis*) and P1 quadruple variant, WP_118709078.1 F54I-T57M-F74H-F187S (*Enterocloster*
733 *aldenensis*) cloned into the pCDFDuet-1 vector was used as template DNA. PCR cycles were
734 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 65°C for 20 seconds and
735 72°C for 2 minutes, followed by a 2-minute final extension at 72°C. A 3 μl KLD Treatment Reaction

736 containing 0.5 µl of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix
737 (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes
738 at 37°C. The entire reaction was transformed into NEB® DH5α competent *E. coli* using the
739 manufacturer's recommended transformation protocol and sequence verified on the Pacific
740 Biosciences Revio platform (Pacific Biosciences, CA) and analyzed using custom pipelines.
741 Plasmid DNA of sequence-verified variants were isolated and transformed into T7 Express cells
742 (New England Biolabs), which was once again sequence verified (Pacific Biosciences). We
743 engineered five chimeric proteins by combining longer sequences from P6, WP_1786180371.1
744 and P1 quad variant, WP_1187090781.1 F54I-T57M-F74H-F187S (Table S5). Gibson assembly
745 (primers in Table S5) was used for the construction of the plasmids.

746 **11. Machine learning**

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

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

771 Random forest classification models were used to predict defluorination probability for HAD
772 homologs in NCBI RefSeq and assembled coding sequences from human gut metagenomes in
773 the Short Read Archive (BioProject ID: PRJNA268964). Five diverse HAD sequences (denoted
774 P20-P24) with defluorination prediction probabilities ranging between 0.38 and 0.85 (Fig. S35)
775 were selected for experimental validation. Genes were codon optimized and synthesized as
776 described above by IDT and cloned by Gibson Assembly into multiple cloning site 1 of pCFDuet-
777 1 vectors. Codon optimized sequences are available here: https://github.com/MSM-
778 group/gut_microbe_defluorination_paper/tree/main/protein_sequences_info/. 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). Defluorination activity was then assessed using the colorimetric assay method described above and compared to machine learning predictions for defluorination.

13. In vivo validation of P2 from *Gordonibacter pamelaeae* DSM 110924

Gordonibacter pamelaeae was inoculated from a cryostock into 8 mL gut microbiota medium (GMM: tryptone peptone 2 g/L; yeast extract 1 g/L; raffinose 10 g/L; L-cysteine 0.5 g/L; meat 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 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 mL/L; ATCC vitamin mix 10 mL/L; ATCC trace mineral mix 10 mL/L; resazurin 1 mg/L) in triplicates and incubated anaerobically for 8 days until it reached an OD of 0.1. Fluoroacetate was added to a final concentration of 10 mM and the cultures were incubated overnight. The culture was centrifuged and the supernatant used for fluoride determination before transfer of 2 mL of supernatant into a small beaker with 10 mL TISAB-II-solution (ammonium acetate 77.1 g/L; Titriplex IV 4 g/L; ammonium chloride 26.75 g/L; acetic acid 96% 59 mL/L) and filled up to 40 mL. The fluoride concentration was measured by using an ISE (fluoride-ion sensitive electrode Orion type No. 9-09-00 and a reference electrode Orion type No. 90-02-00) by adding stepwise a 10 mg/L fluoride standard. The values obtained in mV were used for fluoride quantification (Table S4).

References

1. Current Contributions of Organofluorine Compounds to the Agrochemical Industry. *iScience* **23**, 101467 (2020).

2. M. Inoue, Y. Sumii, N. Shibata, Contribution of Organofluorine Compounds to Pharmaceuticals. *ACS Omega* **5**, 10633–10640 (2020).

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

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

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

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

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

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

9. Personalized Mapping of Drug Metabolism by the Human Gut Microbiome. *Cell* **181**, 1661–1679.e22 (2020).

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

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

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

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

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

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

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

17. S. Che, *et al.*, Structure-Specific Aerobic Defluorination of Short-Chain Fluorinated Carboxylic Acids by Activated Sludge Communities. *Environ Sci Technol Lett* **8**, 668–674 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

47. E. Wang, C. Walsh, Suicide substrates for the alanine racemase of *Escherichia coli* B. *Biochemistry* **17**, 1313–1321 (1978).

48. Y. Xu, R. H. Abeles, Inhibition of tryptophan synthase by (1-fluorovinyl)glycine. *Biochemistry* **32**, 806–811 (1993).

49. E. H. Hanneman, Diisopropylfluorophosphate inhibits acetylcholinesterase activity and disrupts somitogenesis in the zebrafish. *J Exp Zool* **263**, 41–53 (1992).

50. G. F. Makhaeva, *et al.*, Fluorinated alpha-aminophosphonates--a new type of irreversible inhibitors of serine hydrolases. *Dokl Biochem Biophys* **400**, 92–95 (2005).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

65. J. I. Tietz, *et al.*, A new genome-mining tool redefines the lasso peptide biosynthetic landscape. *Nat Chem Biol* **13**, 470–478 (2017).

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

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

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

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

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

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

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

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

74. K. Lindorff-Larsen, *et al.*, Improved side-chain torsion potentials for the Amber ff99SB

protein force field. *Proteins* **78**, 1950–1958 (2010).

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

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

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

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

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

80. GROMACS 2024.1 Manual. https://doi.org/10.5281/zenodo.10721192.

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

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

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

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

85. P. J. A. Cock, *et al.*, Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

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

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

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

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

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