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

*Correspondence: serina.robinson@eawag.ch

Summary

The human gut microbiota is known for its pharmaceutical biotransformation capabilities. Many pharmaceuticals are fluorinated; however, it was previously unknown whether gut microbial enzymes cleave carbon-fluorine bonds. Here we show that dehalogenases encoded by human gut microbes, including *Clostridia, Bacilli,* and *Coriobacteriia*, can hydrolyze fluorinated amino and organic acids. Through the development of a rapid, miniaturized fluoride detection assay, we characterized five new microbial gut defluorinases. Using alanine scanning and chimeric protein design, we identified a disordered 41-amino acid carboxyl-terminal region which is necessary and sufficient to enable weak defluorination when fused to a non-defluorinating dehalogenase. Based on features of the carboxyl-terminus alone, we were able to predict the defluorination activity of microbial dehalogenases with >83% classification accuracy (compared to 95% based on the entire protein). By uncovering the hidden potential of human gut microbes to catalyze defluorination reactions, our findings lay the foundation for therapeutic interventions relevant to human and gut microbiome health.

Introduction

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

The gut microbiota is known to modify and thereby inactivate or increase the toxicity of various xenobiotics[7,8] including fluorinated compounds. *Escherichia coli* hydrolyzes the antineoplastic prodrug fludarabine, enhancing its cytotoxic activity[9]. The chemotherapeutic gemcitabine is modified through deamination[10] or acetylation[9]. Anti-cancer fluoropyrimidine drugs including capecitabine and 5-FU are inactivated by gut bacteria through a pathway that mimics host metabolism[8,11]. Microbiome-dependent pharmacokinetics[8] have been linked to interindividual differences in the composition and metabolic capabilities of the human gut microbiota[12] 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.

To date, few enzymes are reported to be capable of defluorination[13]. This stems not only from the high strength of the carbon-fluorine bond[14], but also to the scarcity of fluorinated natural products[15] and the toxic effects of fluoride on microorganisms[16]. To enable microbes - including human gut-associated bacteria - to sense and mitigate fluoride toxicity, fluoride-specific riboswitches[17] were discovered to activate the production of fluoride ion channels[18–20], thereby protecting the cell from the toxic fluoride anion which are the product of enzymatic defluorination reactions. Interestingly, fluoride-specific riboswitches and ion channels are frequently co-localized with genes encoding haloacid dehalogenase-like enzymes (HADs)[17]. Although the majority of enzymes in this family actually perform phosphoryl- rather than halide-transfer chemistry[21], a subset of HADs are capable of regio- and stereospecific hydrolysis of halogenated carboxylic acids[22].

HADs are small enzymes which cleave carbon-halide bonds under both aerobic and anaerobic conditions without cofactors or other required partners. Thus, HADs are attractive for medical, environmental, and biotechnological applications, motivating efforts for HAD discovery and characterization from environments such as the human gut microbiome[23]. Although the few HADs capable of carbon-fluorine bond cleavage primarily found in soil have been studied in detail[22,24], it remains challenging to mine metagenomic datasets to accurately identify HADs with specificity for fluorinated compounds[25]. Previously, molecular dynamics (MD) showed that HADs capable of defluorination sampled more compact conformations in the active site compared to non-defluorinating HADs[22]. We discovered that this MD-based classification method did not generalize to the larger set of defluorinating and non-defluorinating HADs tested here, which motivated our search for alternative and accessible features to predict defluorination in the human gut microbiome. This study demonstrates, for the first time, that the human microbiome encodes for enzymes capable of catalyzing carbon-fluorine bond cleavage. Additionally, through alanine scanning and chimeric protein design, we pinpointed an intrinsically disordered region in a gut microbial HAD which can confer defluorination activity when substituted into a non-defluorinating enzyme.

Results

The human gut microbiota encodes HAD-like enzymes

To investigate the potential for defluorination by human gut microbial enzymes, we first tested defluorination activity fluoroacetate as a simple fluorinated substrate and known 5-FU degradation product[26]. Notably, prior studies identified the soil bacterium, *Rhodococcus jostii* RHA1, as a degrader of fluoroacetate, along with the discovery of its responsible HAD family enzyme[27]. These enzymes dehalogenate small chlorinated alkanoic acids with only a few demonstrating the capability to cleave the more recalcitrant carbon-fluorine bond[22]. We therefore used the *Rhodococcus jostii* RHA1 HAD (Rha0230) for a protein homology search in the HumGut database[28]. To our surprise, we detected over 500 HAD homologs encoded in human gut metagenomes (Fig. S1, Table S1).

The identified HAD-encoding genes were predominantly found in bacteria belonging to the phylum Bacillota, formerly named the Firmicutes[29]. However, homologs were also prevalent in the Actinomycetota, Pseudomonadota, and Bacteroidota (Fig. 1A, Fig. S1). To further elucidate the relationships among these proteins, we constructed a protein sequence similarity network (Fig. S2), which revealed several distinct HAD clusters[28,30]. Two of these clusters contained characterized members of the 2-haloacid dehalogenase subfamily[22] within the larger HAD superfamily[21]. Other smaller clusters could be assigned to the HAD-like family phosphatases capable of phosphoryl transfer chemistry and family members with other functions (e.g., azetidine hydrolysis)[21,31]. For subsequent analysis, we focused on clusters containing the 2-haloacid dehalogenases and curated the dataset through a multiple sequence alignment to eliminate redundant sequences and confirmed the presence of the catalytic residues (Fig. S3). Importantly, the selected sequences from the human gut microbiota showed low amino acid sequence identity to previously characterized HADs (18% - 46%; Table S1). Among these, we selected seven new representatives for in vitro experiments (Fig. S2, Table S2).

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

To screen for in vitro defluorination activity, we heterologously expressed, purified, and tested seven selected gut microbial dehalogenases with ten different halogenated substrates (nine fluorinated organic acids and one chlorinated substrate, chloroacetate, Fig. 1, Table S3). The HAD from *Rhodococcus jostii* (P5, Protein Data Bank ID: 3UMG) was included as a positive defluorination control as described previously. All proteins tested except for one (P1,WP_118709078, *E. aldenensis,* Fig. 1C) showed dehalogenase activity with at least one substrate (Fig. S4-S11).

Two enzymes hydrolyzed chloroacetate but were not active with any fluorinated substrates. The remaining four enzymes hydrolyzed fluoroacetate, producing fluoride ion and glycolate stoichiometrically, as measured by ion chromatography (Fig. 1B). Compounds with perfluorinated methyl ($-CF_3$) or perfluorinated methylene ($-CF_2-$) groups including trifluoroacetate, difluoropropionate, trifluoropropionate, trifluoropentonate, and heptafluorobutanoate were not accepted by any of the eight HADs tested (Fig. 1C). Four of the enzymes, including the control were active with known transformation products of fluorinated xenobiotics, namely, 2-fluoropropionate, α-fluoro-β-alanine, and the polyfluorinated compound, difluoroacetate.

To further investigate the activity of our gut microbial HADs in oxygen-depleted environments, such as the gut, we further examined P3 (WP_087189991.1, *Gordonibacter urolithinfaciens*). Under both anaerobic and aerobic conditions, P3 cleaved fluoroacetate to produce glycolate and fluoride ions as would be expected for hydrolytic enzymes (Fig. S12).

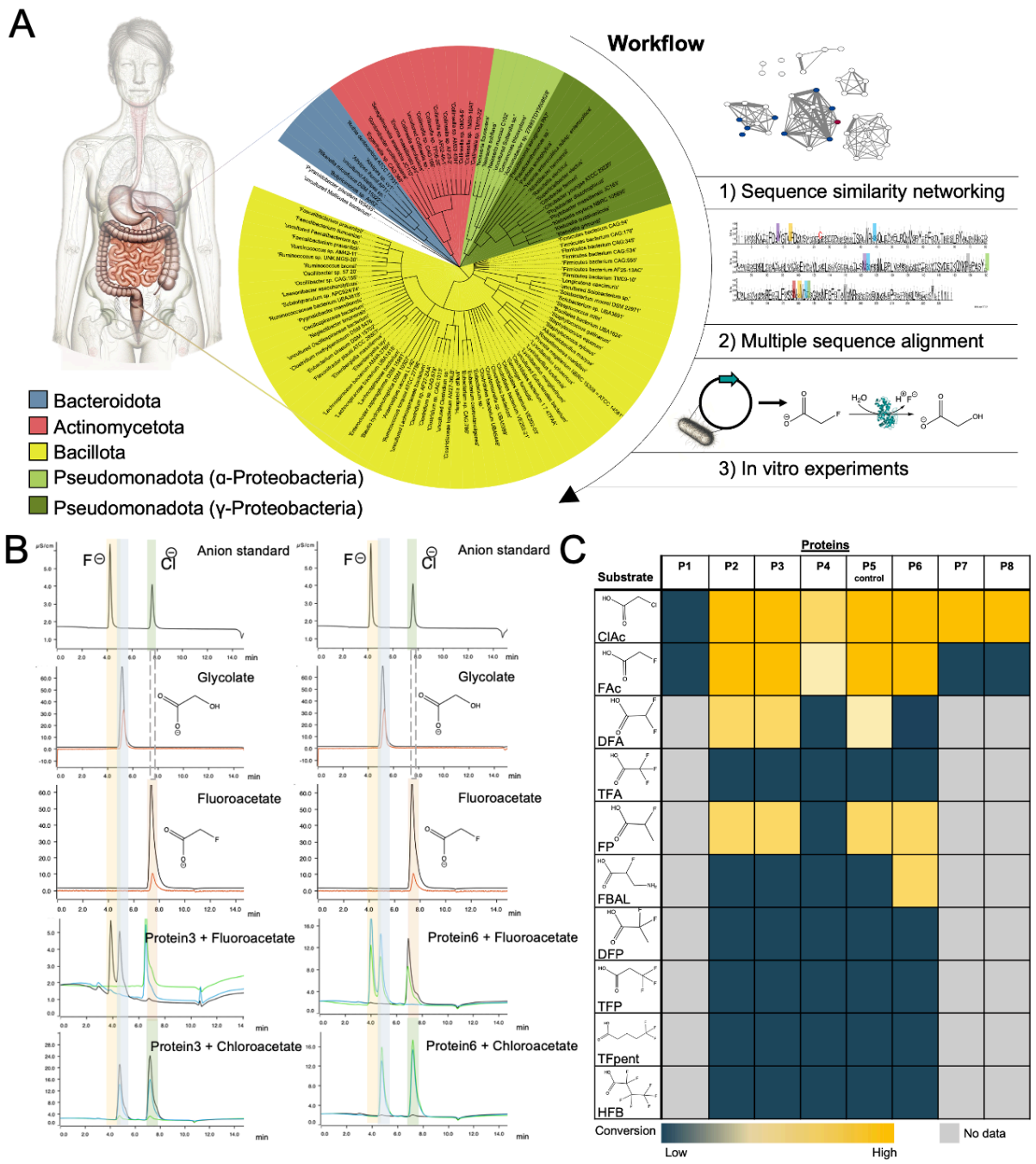

Figure 1. Haloacid dehalogenases are widespread across human gut bacterial phyla and exhibit broad substrate specificity. A) Overview of the identification pipeline for the selected haloacid dehalogenases. We reduced initial homology search hits by sequence similarity networking, multiple sequence alignment to eight candidates. B) Ion chromatography traces of standards and the halogenated organic acids tested with protein 3 and protein 6. C) Substrates specificity of ten different substrates with the eight selected proteins. Yellow: active; light yellow: medium active; blue: non-active; gray: not tested. ClAc: Chloroacetate; FAc: Fluoroacetate, DFA: Difluoroacetate; TFA: Trifluoroacetate; FP: Fluoropropionate; FBAL: α-fluoro-β-alanine; DFP: Difluoropropionate; TFP: Trifluoropropionate; TFpent: Trifluoropentonate; HFB: Heptafluorobutanoate

149 ***Molecular dynamics of defluorinating and non-defluorinating enzymes reveals range of***
150 ***conformational compactness***
151
152 Previously, MD simulations showed that HADs with defluorination activity sampled more
153 compact conformations in the halide binding pocket relative to non-defluorinating enzymes[22].
154 Since the HADs identified here from the human gut microbiota were not tested previously, we
155 first replicated the molecular dynamics (MD) simulations as described[22]. In a subsequent
156 round of simulations, we then increased the simulation time from 300 ns to 1 µs using the
157 Amber99SB-ILDN force field instead of OPLS/A because of better performance for proteins
158 (Fig. S13)[32,33]. Since the original MD publication[22], major advancements have been made in
159 protein structure prediction using AlphaFold (AF)[34,35], even enabling accurate MD simulations
160 using relaxed AF models[36]. Thus, for proteins where crystal structures were not available, we
161 used AF to generate models (Fig. S14). In a benchmark comparison with Chan et al., we
162 obtained generally comparable results in MD simulations of the available X-ray crystal
163 structures and their respective AF models (Fig. S15). Simulating all 11 available defluorinating
164 dehalogenases and 7 non-defluorinating dehalogenases in water for 1 µs, while tracking the
165 size of the halide binding site, did not yield any significant differences between the two groups
166 (Fig. 2A). Further analysis on all conserved amino acids of HADs and their relative distances
167 between each other also showed no significant differences (Fig. S16; Table S4).
168
169 ***Phylogenetics and alignment-independent motifs do not differentiate defluorinating***
170 ***and non-defluorinating enzymes***
171
172 To further investigate methods to differentiate defluorinating and non-defluorinating enzymes,
173 we generated a phylogenetic tree of the HAD-like family protein sequences (Fig. S17). The
174 tree topology revealed that the defluorinating and non-defluorinating enzymes did not form
175 separate or easily-distinguishable clades. The dehalogenating enzymes originated from eight
176 diverse bacterial and archaeal phyla (Fig. S18) which do not group based on the taxonomy of
177 the source organisms.
178
179 ***Active site variants convert inactive enzyme to dechlorinating enzyme***
180
181 As neither MD simulations nor phylogenetic analysis yielded conclusive results to distinguish
182 defluorinating from non-defluorinating HADs, we next sought to experimentally test the role of
183 specific residues in the active site. These residues were identified by comparing the non-
184 defluorinating enzyme P1 (*E. aldenensis)* with the active defluorinase from P6
185 (WP_178618037.1, *G. tenuis)*. Based on comparison between the active sites of the two
186 enzymes (68% sequence similarity, bitscore 268), there were only four residues which were
187 significantly different in the general active site and binding pocket (Fig. S19). We predicted
188 these residues may provide substrate specificity for defluorination, therefore we first aimed for
189 rational engineering of the non-defluorinating enzyme P1 (*E. aldenensis*) via site-directed
190 mutagenesis to generate single, double, and triple variant combinations as well as the full
191 quadruple variant P1 (F541, T57M, H74F, F187S) (Fig. S20-S23, Table S5).
192
193 Among the single variants tested, only one (F54I) showed dehalogenation activity. The F54I
194 variant hydrolyzed chloroacetate but did not hydrolyze fluorinated substrates. Based on
195 structural modeling, F54 is positioned at the entrance channel to the substrate binding pocket.
196 In the non-dehalogenating protein P1 (*E. aldenensis*) the larger phenylalanine likely blocks

the binding pocket entrance. In contrast, the dehalogenating protein P6 (*G. tenuis*) features a smaller isoleucine at this position, potentially enabling access for chloroacetate. While we did achieve a functional dehalogenase through the single-point mutation in P1 (*E. aldenensis*), the individual active site changes did not singularly confer defluorination activity. We hypothesized epistatic interactions may promote defluorination. We therefore used the F54I variant as the template for generating double, triple, and quadruple active site variants. In the P1 (F54I, H74F) double variant and all triple variants we achieved improved dechlorination activity (Fig. 2C). Surprisingly, we did not observe defluorination activity in any of the rationally-engineered variants, even though the active site residues were 100% identical to the defluorinating reference enzyme P6 (*G. tenuis*). This result suggests that additional second shell interactions or other conformational changes[37] not localized in the active site may be critical for defluorination activity.

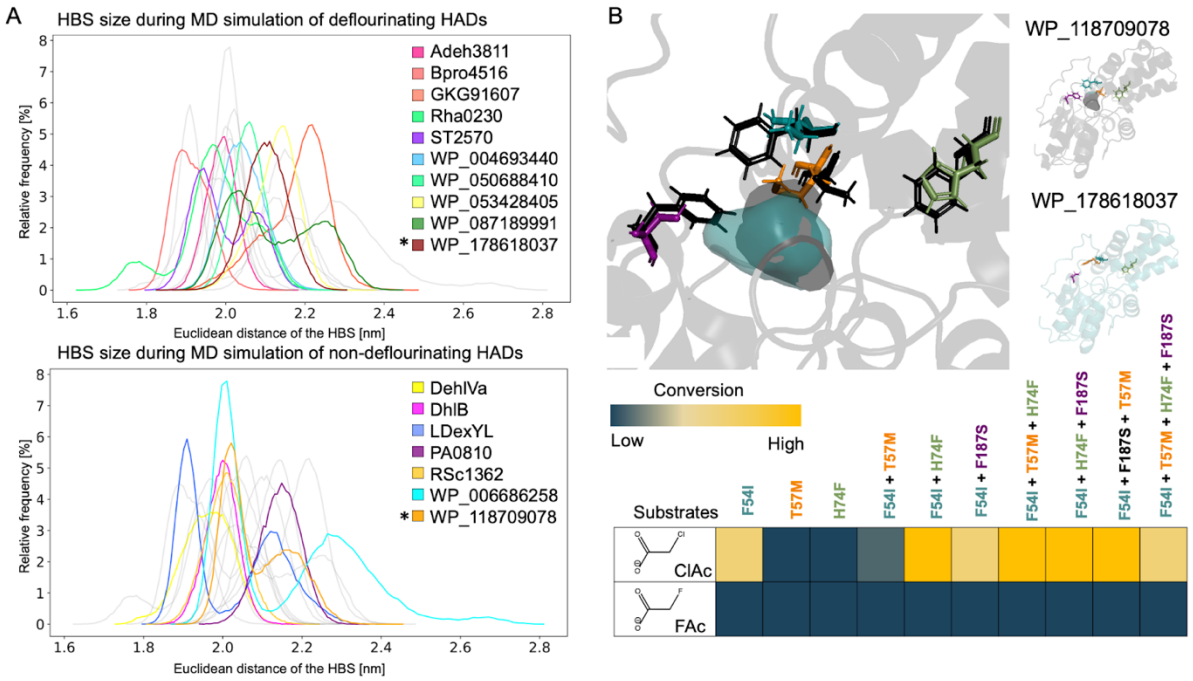

Figure 2. Defluorination capability does not correlate with the active site compactness or amino acid sequence identity. A) Halide binding site (HBS) size during molecular simulation. Distribution of the Euclidean distances of the 3 C_α atoms of the HBS during simulation from 500 ns to 1000 ns. *Asterisks 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. aldensis*) black and the defluorinating protein P6 (WP_178618037, *G. tenuis*) teal. The variants introduced in P1 (WP_, *E. aldensis*) are F54I (blue) T57M (orange), H74F (green) and F187S (purple). Substrate specificity of the different protein variants with the two substrates chloroacetate (ClAc) and fluoroacetate (FAc).

Alanine scanning and a new microtiter well plate liquid assay

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 (F541, T57M, H74F, F187S) (*E. aldensis*). In total, the library consisted of 474 single alanine variants. To screen the alanine

228 library, we devised a microtiter well plate colorimetric screening method for defluorination to
229 avoid lengthy ion chromatography steps for analysis. Previously, a lanthanum-alizarin assay
230 for microtiter well plates[38] and a xylenol orange assay for an agar plate-based method[39] for
231 colony screening and a nanoparticle based fluoride sensor[40] were reported. We tested the
232 lanthanum-alizarin assay[38] and additionally developed a miniaturized liquid adaptation of the
233 xylenol orange screening method. Both assays yielded a clear change in absorbance in the
234 presence of free fluoride ions as validated by ion chromatography results. Due to its enhanced
235 color intensity, we report here the results of screening using the miniaturized xylenol orange
236 assay. As the xylenol orange assay has not been described previously for use in liquid,
237 microtiter well plate format, a detailed methods section and list of interfering substances is
238 included (Fig. S24-S25, Table S6).
239
240 To test our miniaturized screening method, we selected eleven dehalogenases spanning
241 eleven additional gut microbial genera (Table S7). Based on the colorimetric screen, we
242 detected an active gut microbial defluorinase from *Priestia megaterium* sharing less than 35%
243 amino acid identity to the reference enzyme (PDB ID: 3UM9). We purified the *P. megaterium*
244 enzyme to homogeneity and validated its defluorination activity using ion chromatography
245 (Fig. S26). This highlights the potential of microtiter-well assays to rapidly identify
246 defluorinating enzymes.
247
248 With our new defluorination assay in hand, we next screened the alanine scanning library with
249 a dual-assay approach for both defluorination and dechlorination activity (Fig. 3A). Across
250 independent biological replicates of the alanine scanning library (Methods), 39.8-41.7% of
251 alanine variants lost defluorination activity. A slightly lower proportion of variants (33.0-33.5%)
252 lost dechlorination activity. We identified 64 alanine variants which showed dechlorination
253 activity but lost defluorination activity (Fig. S27-S29).
254
255 ***Statistical machine learning highlights the importance of the C-terminal region***
256
257 To analyze our alanine scanning data, we used the random forest algorithm to train a machine
258 learning model to predict defluorination activity. Random forest is suitable for low-data regimes
259 (n=474 data points from alanine scanning data) and was employed for both regression and
260 classification tasks. Our regression model trained on 80% of the alanine scanning data was
261 capable of predicting quantitative changes in defluorination activity as the result of single
262 alanine point mutations with the remaining 20% of the data with a root-mean square error
263 (RMSE) of 0.2 ± 0.02 enzyme activity units (on a normalized activity scale between 0-1
264 defluorination activity). Based on the relative feature importance, the most predictive feature
265 was the position of each amino acid within the protein sequence rather than the size, charge,
266 identity, codon diversity, or polarity of each amino acid converted to alanine (Fig. 3B). These
267 results showing position carried the highest importance for defluorination activity corresponds
268 with the visible 'defluorination' hotspots on the protein structure (Fig. 3C).
269
270 To train a more generalizable model with applications beyond our selected alanine scanning
271 libraries P1 (F541, T57M, H74F, F187S) and P6, we used a random forest classifier to predict
272 defluorinating vs. non-defluorinating enzymes based on general amino acid features
273 extendable to untested HAD sequences (Fig. 3D). We obtained 95% training and test set
274 classification accuracy (average F1-score = 0.93, Table S8). Although recognizing the limited
275 scope of this model to HAD-like family enzymes, this classification analysis further enabled us

to pinpoint residues that were important for defluorination activity (Fig. S30). Strikingly, the top five most important residues for defluorination activity prediction accuracy were localized in the C-terminus (final 41 amino acids). Moreover, a simplified model trained on only the 41 C-terminal residues as features still achieved 83% testing and training set prediction accuracy (average F1-score = 0.73, Table S8).

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

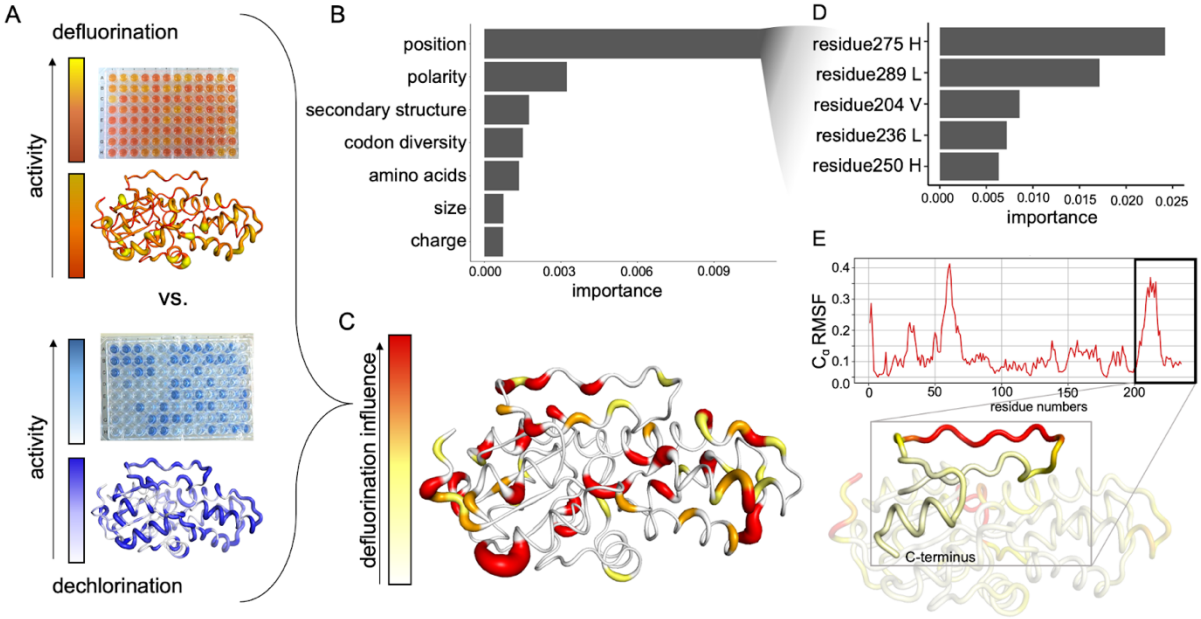

Figure 3. The intrinsically disordered C-terminal region is important for defluorination activity. A) Results of the colorimetric assays for defluorination and dechlorination activity, respectively, mapped onto the P6 protein model. The color gradient from white (low activity) to yellow (medium activity) to red (high activity) and line thickness is used to visualize activity differences based on location in the wild-type protein structure. B) Mean relative feature importance plot extracted from random forest regression model using alanine scanning data to predict defluorination activity. C) Protein 'hotspots' important for defluorination and dechlorination based on the alanine scanning data, where line thickness and coloring corresponds to a higher importance for defluorination alone. D) Top five residue positions with the highest mean Gini importance in random forest classification for defluorinating vs. non-defluorinating proteins E) Root mean square fluctuation (RMSF, reported in nm) of the C_α atoms indicates high movement in the C-terminal region of the protein.

Chimeric defluorinases indicate importance of the C-terminal region

To test the relevance of the C-terminal region experimentally, we generated chimeric proteins by combining different regions from the quad variant non-defluorinating protein P1 (*E. aldenensis* F54I-T57M-F74H-F187S*)* with the defluorinating enzyme P6 (*G. tenuis)* (Fig. S32). Strikingly, the single P1 quad variant (*E. aldensis* 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. S32). Notably, the C-terminal region has not been previously reported for its importance in dehalogenation or fluorinated compound substrate specificity. Analyzing our alanine scanning data based on our chimeric segments (Fig. 4B), we observed a pattern in the activity distribution of defluorination activity (Fig. 4B). Based on our MD simulations, we observed that the C-terminal chimeric segment also had the highest overall movement relative to other protein chimeric segments (Fig. 4C). Overall, the combination of alanine scanning and chimeric mutagenesis indicated that the C-terminal 41 amino acid segment was both necessary and sufficient to install defluorination activity when fused to a non-defluorinating homologous enzyme.

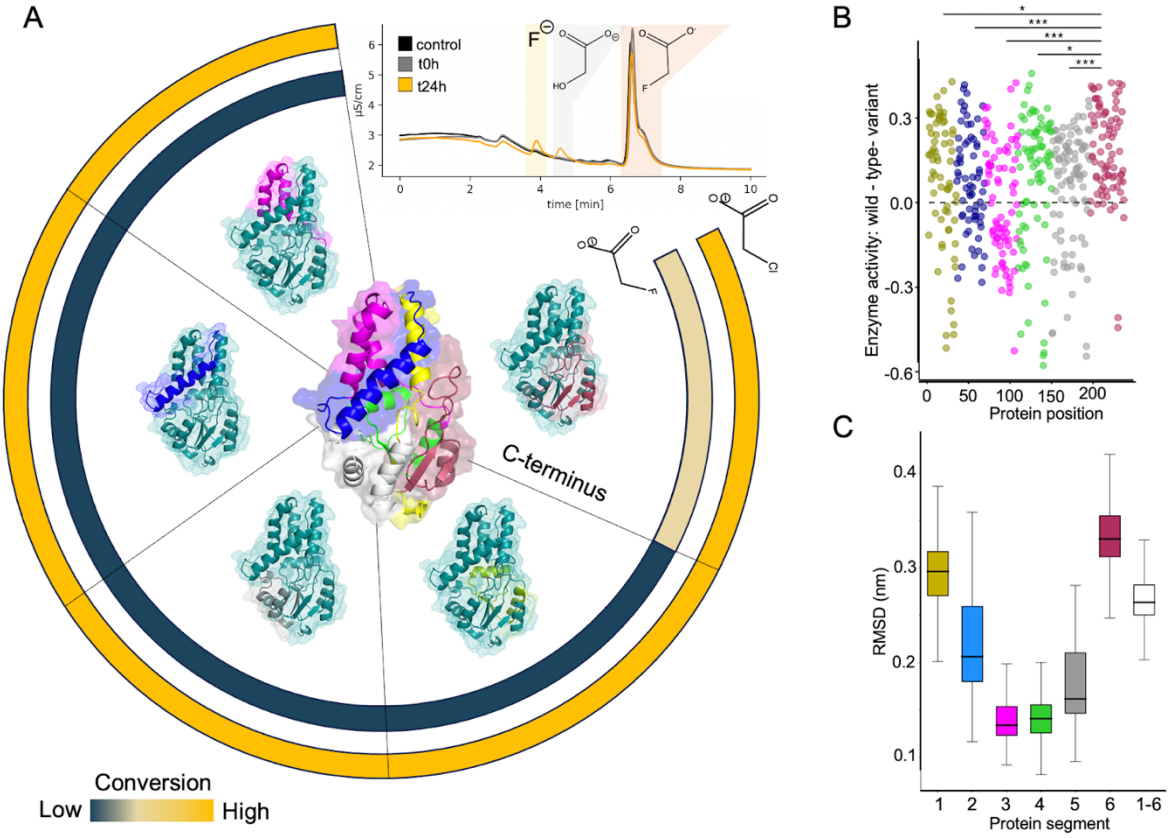

Figure 4. Swapping the C-terminus of a non-defluorinating HAD with the C-terminus of a defluorinating homologous enzyme enables defluorination activity. A) Activity of the chimeras generated from low conversion (blue) to high conversion (gold) of two substrates chloroacetate (outer ring) and fluoroacetate (inner ring). Inset of the ion chromatogram shows an engineered chimeric protein with a swapped C-terminal segment enables weak conversion of fluoroacetate to glycolate and fluoride ion. B) Protein defluorination activity of the alanine scanning library with all data points shown for two independent biological replicates. The x-axis indicates the position in the protein and the y-axis indicates the change in normalized activity for each alanine scanning variant relative to wild-type. Points are colored by chimeric segments as depicted in other figure panels. The C-terminal chimeric segment exhibits a significantly different distribution of active to inactive alanine variants from the other chimeric segments of the protein ($p<0.05$, Kruskal-Wallis test). Only one variant (L231A) in the C-terminal segment displays higher activity than wild-type. C) Boxplot of root mean squared distance (RMSD) in nm for different protein segments derived from MD simulations with the C-terminal chimeric segment exhibiting the highest overall movement.

Discussion

Here, we demonstrated that HAD family enzymes capable of carbon-fluorine bond cleavage are widely distributed across bacterial phyla common in the human gut. To our knowledge, this is the first report of enzymes sourced from the human gut microbiota capable of catalyzing this reaction. Despite sharing sequence and structural similarity, the proteins in this study exhibited variable substrate specificity including fluoroacetate and α-fluoro-β-alanine. These are known transformation products of 5-FU, a widely-used cancer medication on the World Health Organization's list of Essential Medicines[41]. Fluoroacetate, its transformation product fluorocitrate, and α-fluoro-β-alanine resulting from the metabolism of fluoropyrimidine drugs were previously shown to drive the variable and severe side effects, including neurotoxicity[6,42] and cardiotoxicity[43,44].

In this study, we also identified four new HADs active on difluoroacetate, doubling the number of hydrolytic dehalogenases experimentally shown to defluorinate $-CF_2$ groups[45]. Difluoroacetate is a transformation product of various xenobiotics including the refrigerant chlorotrifluoroethylene (R1113)[46] and the anesthetic 1,1,2,2-tetrafluoro-1-(2,2,2-trifluoroethoxy)-ethane[47]. It is a byproduct of the photodegradation of penoxsulam, a widely used pesticide[48]. Difluoroacetate may also come in contact with gut microbes through ingestion via pesticide residues on food[48]. In general, fluorinated substances enter our environment and can seep into groundwater, where they are subsequently consumed by humans, thus coming in contact with the gut microbiome. It has been shown that long-term ingestion of per- and polyfluoroalkyl substances (PFAS) including pharmaceuticals containing $-CF_3$ groups cause shifts in the gut microbiota[24]. Taken together with our findings, this invites further investigation into whether long-term consumption of fluorinated compounds may contribute to defluorinating HADs identified here in the human gut microbiome.

Specificity-determining residues or structural features that differentiate defluorinating enzymes from their non-defluorinating counterparts in the HAD family has remained elusive[22]. Alanine scanning proved useful to identify protein variants that retained dechlorination activity but lost their defluorination capability and identified several amino acids critical for defluorination. Surprisingly, these amino acids were primarily located in the peripheral regions of the protein with a higher percentage of significant residues located in the C-terminal segment of the protein. These observations, along with results from more recent studies[49], support the growing complex understanding of protein function than the simplistic lock-and-key model[50]. Consequently, our protein engineering efforts shifted towards engineering chimeric proteins by replacing segments of non-defluorinating protein variants with segments from a defluorinating enzyme. We observed chimeric proteins with weak, but robust and reproducible defluorination activity only when the C-terminal segment was exchanged.

The C-terminal regions of proteins have been extensively studied in eukaryotic cells[51,52], where 'short linear motifs' within intrinsically disordered regions[53] have been associated with a broad spectrum of functions, including substrate recognition[51,54], signaling[55], and disease risk[51,56]. This functional diversity is attributed to the solvent exposure and higher degree of disorder of C-termini, which facilitate protein-protein interactions[57]. In a prokaryotic HAD family phosphatase, the flexible C-terminal tail was also previously shown to interact with the active site of an adjacent monomer to facilitate product release[37]. Our MD simulations of gut microbial HADs here suggest C-termini exhibit greater flexibility compared to other protein regions.

Combined with our 'gain-of-function' result exclusive to our C-terminal chimeric design, 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 also adds a prokaryotic dimension for deciphering the 'C-terminome' to more accurately predict and modulate protein functions across the domains of life.

Data and code availability

Scripts associated with analysis presented in this manuscript are available at: https://github.com/MSM-group/biodefluorination-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. We thank the entire Joint Genome Institute team including Miranda Harmon-Smith, Yasuo Yoshikuni, and many more for their exceptional support.

Funding

The work (proposal: 10.46936/10.25585/60008420) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231. S.P. was supported by funding from the Helmut Horten Foundation. V.P. and S.L.R. acknowledge funding from the Swiss National Science Foundation (grant no. 501100001711-209124, PZPGP2_209124).

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., 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[22] from NCBI (WP_011593529.1 and WP_011481495.1). The sequences were used as queries for BLAST searches within the HumGut human gut genome collection (Version: February 2023)[28]. A BLASTdb was created using the 30,691 HumGut genomes and a tblastn search was performed to identify homologs of the WP_011593529.1 and WP_011481495.1 proteins. The 500 sequence hits with bitscores greater than 40 were subjected for a reverse blast search against the NCBI database. Hits that could be mapped to NCBI database entries with 100% sequence identity and 90-100% query coverage were further curated for redundancies. The resulting list of 114 protein sequences (Table S1) were aligned using Clustal Omega[58] and the presence of the catalytic active residues were verified. Code and query sequences are available at https://github.com/MSM-group/biodefluorination-paper.
Subsequently, we used the EFI-EST webtool to generate a sequence similarity network (SSN) with default parameters[59]. We visualized the 95% amino acid identity representative network using Cytoscape[60] (Fig. S2) and seven SSN cluster representatives were selected for experimental validation. Genes were synthesized as described below for in vitro work. InBr a second 'validation' round, 11 new proteins were selected based on high sequence similarity alignment scores for additional experimental validation (Table S2, S7).

2. Cloning and heterologous expression of candidate enzymes

The first batch of full-length dehalogenases (n=7) were codon-optimized for *E. coli* codon usage using the Integrated DNA Technologies Codon Optimization Tool. The sequences were synthesized as gBlocks by Integrated DNA Technologies. The gBlocks were cloned individually by Gibson Assembly into multiple cloning site 1 of pCFDuet-1 vectors (spectinomycin/streptomycin resistance marker) with tobacco etch virus (TEV) cleavage sites and C-terminal hexahistidine tags (Table S5). Constructs were transformed into *E. coli* DH5α cells and verified by Sanger sequencing (Microsynth, Balgach, Switzerland). Sequence-verified constructs were transformed into *E. coli* T7 Express cells for protein production (New England Biolabs Frankfurt, Germany, NEB C2566I). The second batch of synthesized genes (n=11) were cloned by the U.S. Department of Energy Joint Genome Institute. These were constructed identically to the first batch with the exception of using the build-optimization software, BOOST[61] for codon optimization and expression in T7 Express *E. coli*. Codon-optimized sequences are available at:
https://github.com/MSM-group/biodefluorination-paper/tree/main/protein_sequences_info.

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

For scale up and purification, induced cultures were poured into 50 mL Falcon tubes and centrifuged at 4000 RCF for 30 min. The supernatant was decanted, and cell pellets were resuspended in 4 mL buffer A (20 mM Tris-HCl, 500 mM NaCl, 20 mM imidazole) per gram

483 cell pellet, 10% glycerol, pH = 8) per gram of pellet weight. Cells were lysed by 3 cycles of
484 sonication using a 6 mm needle, 30% amplitude and 10 x 10 second pulses with a rest time
485 of 10 seconds in between (total sonication time per cycle = 3 minutes). Lysate was centrifuged
486 at 20,000 RCF at 4°C for 45 min. For benchtop purification, supernatants were applied to a
487 column packed with Ni-NTA agarose beads (Qiagen) pre-equilibrated in buffer A. The column
488 was subsequently washed with buffer A and then subjected to two washes with buffer B (40
489 mM Tris-HCl, 500 mM NaCl, 10% glycerol and 40 mM imidazole). Enzymes were eluted from
490 the column using buffer C (40 mM Tris-HCl, 500 mM NaCl, 10% glycerol, and 500 mM
491 imidazole) The fractions with the highest protein concentration was determined qualitatively
492 by Bradford assay[62]. Imidazole and NaCl were removed by running the sample over a PD-10
493 desalting column (Cytiva) and eluting in buffer (100 mM Tris-SO_4, pH = 8.5). The concentration
494 of the purified proteins were determined by Qubit™ protein assay and the proteins were flash
495 frozen using liquid nitrogen in 500 µL aliquots and subsequently stored at -80°C.
496
497
498 **3. Assessment of enzyme-substrate specificity**
499 Stock solutions containing 100 mM of the test substrates listed in Table S3 dissolved in water
500 were prepared. The concentration of the Ni-NTA purified proteins were determined by using
501 the Qubit™ protein assay and the proteins were diluted to a final concentration of 0.5 mg/mL
502 in 100 mM Tris-SO_4 buffer, pH 8.5. As controls the proteins were boiled for 10 min at 100°C.
503 The test substrates were added to a final concentration of 1 mM and the samples were
504 incubated at 37°C. 50 µL samples were taken for the different timepoints and mixed with 50
505 µL acetonitrile to quench the reaction. The samples were centrifuged at 13 000 RPM for 5 min
506 and the supernatant transferred into plastic vials for ion chromatography (IC).
507
508 An IC system (Metrohm, Switzerland) was used consisting of an IC sampling center 889
509 coupled to an ion chromatograph (930 Compact IC Flex) with a conductivity detector. Samples
510 were incubated at 10°C in the sample center with 10 µL injection volumes first onto a
511 precolumn (Metrosep A Supp 16, Guard, 4 mm) and subsequently onto a Metrosep A Supp
512 16 column (250 x 4.0 mm, 5 µm). The temperature in the column was kept constant at 30°C.
513 The eluent consisted of 2.0 mM Na_2CO_3 and 5.0 mM $NaHCO_3$. The flow rate was set to 0.8
514 mL/min. IC data processing and quantification was performed using MagIC Net Professional
515 (version 3.3).
516
517
518 **4. AlphaFold and structural comparison**
519 Structural models for the seven initial dehalogenase sequences selected from the HumGut
520 databases, were generated using AlphaFold2 (AF) ColabFold[63]. Models with the highest rank
521 for each dehalogenase were prepared with relaxation using the Amber99SB force field[64].
522 Additionally, we used AF to generate models for the HADs PA0810, RSC1362, ST2570 and
523 RHA0230. The accuracy of AF models was compared to known X-ray crystal structures, where
524 available[22].
525
526
527 **5. Molecular dynamics simulations**
528 Molecular dynamics simulations were performed using the GROMACS package version
529 2022.5[65,66] compiled with CUDA support and run on NVIDIA GeForce RTX 2080 Ti graphic
530 cards. The crystal structures for HADs PA0810, RSC1362, ST2570 and RHA0230 were

downloaded from RCSB PDB (PDB IDs: 3UMC, 3UMB, 2W43 and 3UMG, respectively. For PDB files which contained selenomethionine, selenomethionine was exchanged with methionine. Relaxed AF models were also subjected to simulations. The system was prepared with GROMACS tools, parameterized with the force field Amber99SB-ILDN[67], and solvated in explicit water using the default water model TIP3P[68] 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[69]. 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[70] and a Fourier spacing of 0.14 nm. The temperature was controlled using a modified Berendsen thermostat[71] 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[72] 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 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[73]. To measure the Root Mean Square Fluctuation 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. S17-S18) were performed to gain a better perspective on the sequence features of the HAD-like enzymes. We performed these two analyses with a data set of the entire HAD-like family available in UniProt[74] 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[75] and the phylogenetic tree (Fig. S17-S18) was generated using the DistanceTreeConstructor method from the biopython library[76].

7. Rational active site engineering

To identify potentially significant amino acids involved in the defluorination process, we used two of seven in vitro tested proteins sharing high structural and sequential similarity but distinct activities from each other (Fig. S19) (54% sequence identity, 68% sequence similarity). WP_1187090781.1, a non-defluorinating enzyme, and WP_1786180371.1, a defluorinating enzyme. We targeted amino acids within WP_1187090781.1, particularly those situated in or

near the active site, for substitution with their counterparts from WP_1786180371.1. Four single variants were constructed: F54I, T57M, F74H, and F187S and their combinations resulting in 14 variants encoded by point mutations in codons for up to four different active site residues. The mutant gene constructs were generated by using mismatch primers (Table S5). A 3 µL Kinase, Ligase & DpnI (KLD) Treatment Reaction containing 0.5 µL of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes at 37°C.

8. Colorimetric assay for fluoride detection

A colorimetric fluoride screening method was developed here for 96-well plate format based on the agar-plate based colony screening method initially described by Davis et al.[39] Xylenol orange stock solution was prepared with a concentration of 20 mM. To prepare the zirconium oxychloride octahydrate stock solution, 0.3022 g of zirconium(IV)-oxychloride octahydrate was dissolved in 925 µL water and 1.85 mL 12.1 M HCl added and then boiled for 5 min to achieve final concentrations of 500 mM fully depolarized zirconium and 8 M HCl. After cooling, the stock solution was diluted to a concentration of 50 mM zirconium. Shortly before the experiment, the working solution was prepared with 7.8 M HCl, 400 µM xylenol orange and the complex was initiated by the addition of 250 µM zirconium. A list of potential interfering substances and buffers were tested and results are listed in Table S6.

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

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

9. Chloride detection assay

A chloride detection assay kit (QuantiChrom™ Chloride Assay Kit) was obtained from Huber lab and assays were run according to the manufacturer's instructions with the following changes: Briefly, *E. coli* T7 expression strains carrying the corresponding plasmids were

grown in TB medium in a 2 mL deep well plate (Corning, Axygen Inc P-DW-20-C-S) with 100 µg/mL spectinomycin until an optical density of 0.8 was reached. The cells were induced with 0.1mM IPTG and incubated for 24 h. The plate was centrifuged at 4000 RCF for 10 min, the supernatant was discarded and the cells were washed twice with 2 mL 20 mM HEPES (pH 8). After centrifugation at 4000 RCF for 10 min the cells were resuspended in 1 mL 20 mM HEPES (pH 8) and chloroacetate was added to a final concentration of 30 mM prior to incubation at 37°C for 24 h. The cells were centrifuged at 4000 RCF for 15 min and 2.5 µL of the supernatant was transferred to a new 96 well plate and 100 µL of the colorimetric working solution was added. The chloride concentration is measured at 610 nm with an Eon spectrophotometer (Biotek). The limit-of-quantification was 0.2 mM.

To increase the comparability between the fluoride and chloride assay results, the values of both assays were 0 to 1 linearized and thresholds used to decide whether they are positive for their respective activity. The normalization formula used is as follows: (raw value - lowest value)/(highest value - lowest value). For the chloride assay values below 0.15 were counted as negative for dechlorination and for the defluorination assay values below 0.4 were counted as negative for defluorination. Differences in the chloride and fluoride values greater than 0.275 were counted as exhibiting different dechlorination vs. defluorination activity.

10. Whole-gene alanine-scanning mutagenesis and chimeragenesis

High-throughput alanine-scanning mutagenesis was performed using the Q5® Site-Directed Mutagenesis Protocol (New England Biolabs Inc.) with the following amendments. Briefly, back-to-back forward and reverse primers were designed using NEBaseChanger (https://nebasechanger.neb.com/), where the alanine codon substitution was contained in the center of the forward mutagenic primer. Note that a histidine codon was substituted in place of any native alanine codons. Primers were ordered from Integrated DNA Technologies. The complete set of alanine (n=475) are available on https://github.com/MSM-group/biodefluorination-paper/tree/main/Alanine_scanning_primers. A 6 µL PCR reaction, containing 1x Q5 Hot Start High-Fidelity 2X Master Mix (New England Biolabs), 0.5 µM of the forward and reverse primer, and 10 ng template DNA, was carried out for each variant. P6, WP_178618037.1 (*Guopingia tenuis*) and P1 quadruple variant, WP_118709078.1 F54I-T57M-F74H-F187S (*Enterocloster aldenensis*) cloned into the pCDFDuet-1 vector was used as template DNA. PCR cycles were 98°C for 30 seconds, followed by 30 cycles of 98°C for 10 seconds, 65°C for 20 seconds and 72°C for 2 minutes, followed by a 2-minute final extension at 72°C. A 3 µl KLD Treatment Reaction containing 0.5 µl of the unpurified PCR product, 1x KLD Reaction Buffer and 1x KLD Enzyme Mix (New England Biolabs) was incubated at room temperature for 10 minutes followed by 10 minutes at 37°C. The entire reaction was transformed into NEB® DH5α competent *E. coli* using the manufacturer's recommended transformation protocol and sequence verified on the Pacific Biosciences Revio platform (Pacific Biosciences, CA) and analyzed using custom pipelines. Plasmid DNA of sequence-verified mutant variants were isolated and transformed into T7 Express cells (New England Biolabs), which was once again sequence verified (Pacific Biosciences). We engineered five chimeric proteins by combining longer sequences from P6, WP_1786180371.1 and P1 quad variant, WP_1187090781.1 F54I-T57M-F74H-F187S (Table S5). Gibson assembly (primers in Table S5) was used for the construction of the plasmids.

11. Machine learning

Random forest regression and classification was performed based on scripts and featurization methods published previously[77,78] with the random forest algorithm as implemented in the ranger package[79]. Briefly, for activity prediction using regression based on the alanine scanning data, each point variant (n=474) was featurized using the method of Atchley et al.[80] to a vector representing the position index, size, charge, identity, codon diversity, and polarity of each amino acid which was converted to alanine. These features were subsequently used to predict the change or delta in defluorination enzyme activity from wild-type as a result of the alanine substitution. For the random forest classification task, the sequences of all HAD family enzymes which have been characterized in this study and previously in the literature were aligned using a protein structure-aware alignment method[81]. Invariant residues in the alignment were trimmed and the remaining residues were one-hot encoded and used as features to classify HAD sequences as 1: defluorinating or 0: non-defluorinating. For both classification and regression tasks, the data were split by stratified random sampling into 80% training and 20% testing sets. Grid search was used to tune model hyperparameters by 10-fold cross validation repeated in triplicate. All forests were grown to a size of 1000 trees. Hyperparameters that were tuned during cross-validation included the number of variables randomly sampled as candidates at each split, the minimum size of terminal nodes and the splitting rule methods. The variability in training and testing set prediction performances were examined with 1000 different random seeds for training-test splits. To assess the specific importance of the C-terminal region for classification prediction, only the final 41 residues in the HAD alignment were used as features and model performance was compared to models trained on the full-length alignment. Scripts for the machine learning analysis are available at: https://github.com/MSM-group/biodefluorination-paper/tree/main/machine_learning.

References

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

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

3. Kotthoff, M., Müller, J., Jürling, H., Schlummer, M., and Fiedler, D. (2015). Perfluoroalkyl and polyfluoroalkyl substances in consumer products. Environ. Sci. Pollut. Res. *22*, 14546–14559.

4. Phelps, D.W., Parkinson, L.V., Boucher, J.M., Muncke, J., and Geueke, B. (2024). Per- and Polyfluoroalkyl Substances in Food Packaging: Migration, Toxicity, and Management Strategies. Environ. Sci. Technol. *58*, 5670–5684.

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

6. Okeda, R., Shibutani, M., Matsuo, T., Kuroiwa, T., Shimokawa, R., and Tajima, T. (1990). 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.

715 7. Koppel, N., Maini Rekdal, V., and Balskus, E.P. (2017). Chemical transformation of
716 xenobiotics by the human gut microbiota. Science *356*.
717 https://doi.org/10.1126/science.aag2770.

718 8. Javdan, B., Lopez, J.G., Chankhamjon, P., Lee, Y.-C.J., Hull, R., Wu, Q., Wang, X.,
719 Chatterjee, S., and Donia, M.S. (2020). Personalized Mapping of Drug Metabolism by the
720 Human Gut Microbiome Cell *181*, 1661–1679.e22.

721 9. Lehouritis, P., Cummins, J., Stanton, M., Murphy, C.T., McCarthy, F.O., Reid, G., Urbaniak,
722 C., Byrne, W.L., and Tangney, M. (2015). Local bacteria affect the efficacy of
723 chemotherapeutic drugs. Sci. Rep. *5*, 1–12.

724 10. Geller, L.T., Barzily-Rokni, M., Danino, T., Jonas, O.H., Shental, N., Nejman, D., Gavert, N.,
725 Zwang, Y., Cooper, Z.A., Shee, K., et al. (2017). Potential role of intratumor bacteria in
726 mediating tumor resistance to the chemotherapeutic drug gemcitabine. Science *357*, 1156–
727 1160.

728 11. Spanogiannopoulos, P., Kyaw, T.S., Guthrie, B.G.H., Bradley, P.H., Lee, J.V., Melamed, J.,
729 Malig, Y.N.A., Lam, K.N., Gempis, D., Sandy, M., et al. (2022). Host and gut bacteria share
730 metabolic pathways for anti-cancer drug metabolism. Nat Microbiol *7*, 1605–1620.

731 12. Jennings, B.A., Loke, Y.K., Skinner, J., Keane, M., Chu, G.S., Turner, R., Epurescu, D.,
732 Barrett, A., and Willis, G. (2013). Evaluating predictive pharmacogenetic signatures of
733 adverse events in colorectal cancer patients treated with fluoropyrimidines. PLoS One *8*,
734 e78053.

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

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

740 15. Walker, M.C., and Chang, M.C.Y. (2014). Natural and engineered biosynthesis of fluorinated
741 natural products. Chem. Soc. Rev. *43*, 6527–6536.

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

744 17. Baker, J.L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R.B., and Breaker, R.R.
745 (2012). Widespread genetic switches and toxicity resistance proteins for fluoride. Science
746 *335*, 233–235.

747 18. McIlwain, B.C., Gundepudi, R., Koff, B.B., and Stockbridge, R.B. (2021). The fluoride
748 permeation pathway and anion recognition in Fluc family fluoride channels. Elife *10*.
749 https://doi.org/10.7554/eLife.69482.

750 19. Stockbridge, R.B., Robertson, J.L., Kolmakova-Partensky, L., and Miller, C. (2013). A family
751 of fluoride-specific ion channels with dual-topology architecture. Elife *2*, e01084.

752 20. Brammer, A.E., Stockbridge, R.B., and Miller, C. (2014). F-/Cl- selectivity in CLCF-type F-
753 /H+ antiporters. J. Gen. Physiol. *144*, 129–136.

754 21. Burroughs, A.M., Allen, K.N., Dunaway-Mariano, D., and Aravind, L. (2006). Evolutionary
755 genomics of the HAD superfamily: understanding the structural adaptations and catalytic
756 diversity in a superfamily of phosphoesterases and allied enzymes. J. Mol. Biol. *361*, 1003–
757 1034.

758 22. Chan, P.W.Y., Chakrabarti, N., Ing, C., Halgas, O., To, T.K.W., Wälti, M., Petit, A.-P., Tran,
759 C., Savchenko, A., Yakunin, A.F., et al. (2022). Defluorination Capability of l-2-Haloacid
760 Dehalogenases in the HAD-Like Hydrolase Superfamily Correlates with Active Site
761 Compactness. Chembiochem *23*, e202100414.

762 23. Jia, B., Han, X., Kim, K.H., and Jeon, C.O. (2022). Discovery and mining of enzymes from
763 the human gut microbiome. Trends Biotechnol. *40*, 240–254.

764 24. Rashid, F., Dubinkina, V., Ahmad, S., Maslov, S., and Irudayaraj, J.M.K. (2023). Gut
765 Microbiome-Host Metabolome Homeostasis upon Exposure to PFOS and GenX in Male
766 Mice. Toxics *11*. https://doi.org/10.3390/toxics11030281.

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

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

772 27. Chan, W.Y., Wong, M., Guthrie, J., Savchenko, A.V., Yakunin, A.F., Pai, E.F., and Edwards,
773 E.A. (2010). Sequence- and activity-based screening of microbial genomes for novel
774 dehalogenases. Microb. Biotechnol. *3*, 107–120.

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

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

780 30. Gerlt, J.A., Bouvier, J.T., Davidson, D.B., Imker, H.J., Sadkhin, B., Slater, D.R., and Whalen,
781 K.L. (2015). Enzyme Function Initiative-Enzyme Similarity Tool (EFI-EST): A web tool for
782 generating protein sequence similarity networks. Biochim. Biophys. Acta *1854*, 1019–1037.

783 31. Gross, C., Felsheim, R., and Wackett, L.P. (2008). Genes and enzymes of azetidine-2-
784 carboxylate metabolism: detoxification and assimilation of an antibiotic. J. Bacteriol. *190*,
785 4859–4864.

786 32. Lindorff-Larsen, K., Maragakis, P., Piana, S., Eastwood, M.P., Dror, R.O., and Shaw, D.E.
787 (2012). Systematic validation of protein force fields against experimental data. PLoS One *7*,
788 e32131.

789 33. Piana, S., Lindorff-Larsen, K., and Shaw, D.E. (2011). How robust are protein folding
790 simulations with respect to force field parameterization? Biophys. J. *100*, L47–L49.

791 34. Abramson, J., Adler, J., Dunger, J., Evans, R., Green, T., Pritzel, A., Ronneberger, O.,
792 Willmore, L., Ballard, A.J., Bambrick, J., et al. (2024). Accurate structure prediction of
793 biomolecular interactions with AlphaFold 3. Nature. https://doi.org/10.1038/s41586-024-
794 07487-w.

795 35. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O.,
796 Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate
797 protein structure prediction with AlphaFold. Nature *596*, 583–589.

798 36. Guo, H.-B., Huntington, B., Perminov, A., Smith, K., Hastings, N., Dennis, P., Kelley-
799 Loughnane, N., and Berry, R. (2024). AlphaFold2 modeling and molecular dynamics
800 simulations of an intrinsically disordered protein. PLoS One *19*, e0301866.

801 37. Biswas, T., Yi, L., Aggarwal, P., Wu, J., Rubin, J.R., Stuckey, J.A., Woodard, R.W., and
802 Tsodikov, O.V. (2009). The tail of KdsC: conformational changes control the activity of a
803 haloacid dehalogenase superfamily phosphatase. J. Biol. Chem. *284*, 30594–30603.

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

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

810 40. Mukherjee, S., Shah, M., Chaudhari, K., Jana, A., Sudhakar, C., Srikrishnarka, P., Islam,
811 M.R., Philip, L., and Pradeep, T. (2020). Smartphone-based Fluoride-specific Sensor for
812 Rapid and Affordable Colorimetric Detection and Precise Quantification at Sub-ppm Levels
813 for Field Applications. ACS Omega *5*, 25253–25263.

814 41. Miura, K., Kinouchi, M., Ishida, K., Fujibuchi, W., Naitoh, T., Ogawa, H., Ando, T., Yazaki,
815 N., Watanabe, K., Haneda, S., et al. (2010). 5-fu metabolism in cancer and orally-
816 administrable 5-fu drugs. Cancers *2*, 1717–1730.

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

819 43. Muneoka, K., Shirai, Y., Yokoyama, N., Wakai, T., and Hatakeyama, K. (2005). 5-
820 Fluorouracil cardiotoxicity induced by alpha-fluoro-beta-alanine. Int. J. Clin. Oncol. *10*, 441–
821 443.

822 44. Franck, C., Malfertheiner, P., and Venerito, M. (2017). Safe administration of S-1 after 5-
823 fluorouracil-induced cardiotoxicity in a patient with colorectal cancer. BMJ Case Rep. *2017*.
824 https://doi.org/10.1136/bcr-2016-219162.

825 45. Khusnutdinova, A.N., Batyrova, K.A., Brown, G., Fedorchuk, T., Chai, Y.S., Skarina, T., Flick,
826 R., Petit, A.-P., Savchenko, A., Stogios, P., et al. (2023). Structural insights into hydrolytic

defluorination of difluoroacetate by microbial fluoroacetate dehalogenases. FEBS J. https://doi.org/10.1111/febs.16903.

46. Dorgerloh, U., Becker, R., and Kaiser, M. (2019). Evidence for the Formation of Difluoroacetic Acid in Chlorofluorocarbon-Contaminated Ground Water. Molecules *24*. https://doi.org/10.3390/molecules24061039.

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

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

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

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

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

52. Kumar, M., Michael, S., Alvarado-Valverde, J., Zeke, A., Lazar, T., Glavina, J., Nagy-Kanta, E., Donagh, J.M., Kalman, Z.E., Pascarelli, S., et al. (2024). ELM-the Eukaryotic Linear Motif resource-2024 update. Nucleic Acids Res. *52*, D442–D455.

53. Zarin, T., Strome, B., Peng, G., Pritišanac, I., Forman-Kay, J.D., and Moses, A.M. (2021). Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. Elife *10*. https://doi.org/10.7554/eLife.60220.

54. Sargeant, D.P., Gryk, M.R., Maciejewski, M.W., Thapar, V., Kundeti, V., Rajasekaran, S., Romero, P., Dunker, K., Li, S.-C., Kaneko, T., et al. (2012). Secondary structure, a missing component of sequence-based minimotif definitions. PLoS One *7*, e49957.

55. Hsu, I.S., Strome, B., Lash, E., Robbins, N., Cowen, L.E., and Moses, A.M. (2021). A functionally divergent intrinsically disordered region underlying the conservation of stochastic signaling. PLoS Genet. *17*, e1009629.

56. Pritišanac, I., Alderson, T.R., Kolarić, Đ., Zarin, T., Xie, S., Lu, A., Alam, A., Maqsood, A., Youn, J.-Y., Forman-Kay, J.D., et al. (2024). A functional map of the human intrinsically disordered proteome. bioRxiv. https://doi.org/10.1101/2024.03.15.585291.

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

58. Madeira, F., Pearce, M., Tivey, A.R.N., Basutkar, P., Lee, J., Edbali, O., Madhusoodanan, N., Kolesnikov, A., and Lopez, R. (2022). Search and sequence analysis tools services from EMBL-EBI in 2022. Nucleic Acids Res. *50*, W276–W279.

59. Zallot, R., Oberg, N., and Gerlt, J.A. (2019). 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.

60. Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

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

62. Bradford, M.M. (1976). 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.

63. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682.

64. Eastman, P., Swails, J., Chodera, J.D., McGibbon, R.T., Zhao, Y., Beauchamp, K.A., Wang, L.-P., Simmonett, A.C., Harrigan, M.P., Stern, C.D., et al. (2017). OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLoS Comput. Biol. *13*, e1005659.

65. Abraham, M.J., Murtola, T., Schulz, R., Páll, S., Smith, J.C., Hess, B., and Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. SoftwareX *1-2*, 19–25.

66. Bauer, P., Hess, B., and Lindahl, E. (2023). GROMACS 2022.5 Manual. Preprint at Zenodo, https://doi.org/10.5281/ZENODO.7586765 https://doi.org/10.5281/ZENODO.7586765.

67. Lindorff-Larsen, K., Piana, S., Palmo, K., Maragakis, P., Klepeis, J.L., Dror, R.O., and Shaw, D.E. (2010). Improved side-chain torsion potentials for the Amber ff99SB protein force field. Proteins *78*, 1950–1958.

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

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

70. Essmann, U., Perera, L., Berkowitz, M.L., Darden, T., Lee, H., and Pedersen, L.G. (1995). A smooth particle mesh Ewald method. J. Chem. Phys. *103*, 8577–8593.

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

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

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

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

75. Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. *7*, 539.

76. Cock, P.J.A., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422–1423.

77. Robinson, S.L., Smith, M.D., Richman, J.E., Aukema, K.G., and Wackett, L.P. (2020). Machine learning-based prediction of activity and substrate specificity for OleA enzymes in the thiolase superfamily. Synth. Biol. *5*. https://doi.org/10.1093/synbio/ysaa004.

78. Robinson, S.L., Terlouw, B.R., Smith, M.D., Pidot, S.J., Stinear, T.P., Medema, M.H., and Wackett, L.P. (2020). Global analysis of adenylate-forming enzymes reveals β-lactone biosynthesis pathway in pathogenic. J. Biol. Chem. *295*, 14826–14839.

79. Wright, M.N., and Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. *77*. https://doi.org/10.18637/jss.v077.i01.

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

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