An enzyme-level benchmark based on environmental bacterial laccases for predicting contaminant fate in water

Yaochun Yu[1,2,✉], Kunyang Zhang[1,3,4], Vincenz-Maria Steiner[3], Victoria Poltorak[5], Silke I. Probst[5], Serina L. Robinson[5], Jürg Hutter[3], Hiroko Satoh[3,6], Kathrin Fenner[1,3,✉]

[1] Department of Environmental Chemistry, Swiss Federal Institute of Aquatic Science and Technology (Eawag), 8600 Dübendorf, CH.

[2] Department of Civil and Environmental Engineering, Stanford University, 94305 Stanford, CA, USA.

[3] Department of Chemistry, University of Zürich, 8057 Zürich, CH.

[4] Bayer AG, Crop Science Division, 40789 Mohnheim, DE.

[5] Department of Environmental Microbiology, Swiss Federal Institute of Aquatic Science and Technology (Eawag), 8600 Dübendorf, CH.

[6] Joint Support-Center for Data Science Research, Research Organization of Information and Systems (ROIS), 10-3 Midoricho, Tachikawa, Tokyo 190-0014, JP.

✉Corresponding Author: Yaochun Yu (yaochun.yu@stanford.edu); Kathrin Fenner (kathrin.fenner@eawag.ch)

Abstract

Bacterial laccases are widespread multicopper oxidases whose roles in the fate of anthropogenic chemicals in aquatic environments remain poorly understood. Here, we integrate metagenomic analysis, a miniaturized high-throughput assay and machine learning to establish an enzyme-level benchmark for predicting biotransformation of wastewater-relevant trace organic contaminants by laccase-mediator systems. Using a laccase from an ammonia-oxidizing bacterium as a model enzyme, we screened 183 compounds and identified 38 that underwent significant removal. Following phylogenetic analysis of environmental homologs, we expressed and purified two additional laccases from the bacterial methanotrophic phylum *Methylmirabilota* and an archaeal phylum *Thermoproteota*, demonstrating the activity of this enzyme family across domains. Graph convolutional network models trained on the dataset achieved up to 78% accuracy in classifying degradable versus persistent chemicals, while quantum-chemical descriptors highlighted key electronic properties governing oxidation. This bottom-up approach to enzyme-chemical interactions establishes a trajectory towards predicting contaminant persistence in engineered and natural waters.

Introduction

The European Commission's Safe and Sustainable by Design (SSbD) Framework aims to integrate safety and sustainability considerations at the earliest stages of chemical innovation, rather than addressing them afterwards[1]. A central component of this effort is the ability to anticipate how new chemicals will behave once they enter the environment, for example, estimating their environmental half-lives[2], identifying potential biotransformation pathways[3], and assessing whether their transformation products pose additional risks[4, 5]. Yet, despite substantial research on contaminant persistence, the mechanistic understanding of many biotransformation reactions remains limited[6].

The knowledge gap persists for several reasons. Many studies still emphasize system-level parent-compound removal rather than the underlying biochemical mechanisms, leaving the enzymes responsible for individual reactions largely unresolved. Moreover, because many environmentally relevant microorganisms remain uncultivated, direct validation of their metabolic capabilities is limited[7, 8]. Although meta-omics sequencing and high-resolution mass spectrometry can track community shifts and detect transformation products, they rarely provide direct evidence needed to attribute specific biotransformation steps to particular microbial taxa or enzymes[9]. Biochemical approaches, including metabolomics-guided bioinformatics and *in vitro* functional assays, can validate the enzyme activities[6], but available data remain insufficient to elucidate reactivity across diverse contaminant classes. Enzymes can be highly promiscuous[6, 10], and without understanding this promiscuity across environmental homologs, linking chemical structure to community-encoded function, and ultimately predicting chemical fate from genomic information, remains difficult.

Nowadays, artificial intelligence (AI) is beginning to show promise for predicting chemical properties and ecological impacts[11-13]. For example, recent work has used wastewater treatment data to predict contaminant removal efficiency[14-18], and physicochemical parameters such as lipophilicity data to predict toxicity toward nitrifying microorganisms[19-22]. However, the use of AI in this context is still strongly limited by the type and quality of data that are available[23, 24]. From a chemical perspective, predictive models require reactivity data generated under well-defined and comparable conditions, whereas existing datasets are often sparse, inconsistent or difficult to compare across studies. From a microbial perspective, metagenomic and metatranscriptomic datasets contain rich community-level genetic information, yet the specific functional

determinants of individual biotransformation steps are often difficult to extract for predictive modeling. Enzyme activity provides a complementary layer that helps link community-scale sequence information to chemical biotransformation outcomes. On this basis, we hypothesize that a bottom-up strategy focused on enzyme–chemical reactivity provides a practical starting point for reducing system complexity while retaining the resolution required for chemical fate prediction.

In this study, we focused on environmental bacterial laccases, a widespread family of multicopper oxidases, as a model system to test this hypothesis. Using a previously identified bacterial laccase from activated sludge as a reference, we performed metagenomic analyses to map its global distribution, identify more than one thousand environmental homologs, and confirm oxidation activity in two additional phylogenetically distinct representatives. We then applied a laccase–mediator system to create a chemically well-defined oxidative environment and used a miniaturized high-throughput assay to measure the reactivity of 183 wastewater-relevant trace organic contaminants under standardized conditions[25, 26]. To support downstream prediction, we developed a data-cleaning strategy that produces a consistent and machine-learning-ready reactivity dataset. We further trained graph convolutional network models to classify reactive versus persistent chemicals and used quantum chemical descriptors to interpret the key electronic features governing oxidation. Collectively, this study establishes an enzyme-level reactivity benchmark that links biochemical data with predictive modeling, offering a foundational step toward SSbD-aligned tools for assessing the environmental fate of emerging contaminants.

Main text

Prevalence of microorganisms encoding homologs of a novel bacterial multicopper oxidase (MCO1) in global environments

In our previous study, we characterized a novel bacterial multicopper oxidase (MCO1) from *Nitrosomonas* sp., identified through metatranscriptomic sequencing, and demonstrated its capability to degrade eight trace organic contaminants (TrOCs) commonly found in wastewater[26]. However, the environmental prevalence and phylogenetic breadth of this enzyme family remained unclear. Here, we assessed the global distribution and taxonomic diversity of MCO1 homologs using four databases (the National Center for Biotechnology Information (NCBI) non-redundant protein sequences database, the Global Microbial Gene Catalog[27], the BV-BRC database[28], and the Joint Genome Institute Integrated Microbial Genomes (JGI-IMG) database[29]. Our

101 biogeographic analysis revealed that MCO1 homologs were detected in environmental samples
102 spanning at least five continents (Figure 1, Table S1). These homologs were found predominantly
103 in the genomes of bacteria from soil, marine and freshwater systems, lakes, sediments, and the
104 hypersaline Ekho Lake in Antarctica[30]. Genomes of bacterial strains encoding MCO1 homologs
105 have also been isolated from contaminated environments including a crude-oil contaminated
106 seashore[31] and soil polluted with chromate[32] (Figure 1), consistent with previous reports of
107 multicopper oxidases involved in hydrocarbon oxidation[33] and chromate detoxification[34]. While
108 the presence of MCO1 homologs does not by itself establish a direct role in contaminant
109 biotransformation, it is consistent with the broader observation that multicopper oxidases are
110 detected across diverse environments and can participate in redox processes relevant to
111 contaminant transformation[35-37].

112 We identified MCO1 homologs from seventeen different bacterial phyla and one archaeal
113 phylum (Figure 1). Among these, MCO1 homologs were detected in the major genera of ammonia-
114 oxidizing microorganisms including *Nitrosomonas, Nitrosospira,* and the ammonia-oxidizing
115 archaeon Ca. *Nitrosopolaris* sp. (HYV56308.1, phylum *Thermoproteota*, classified in NCBI as the
116 *Nitrososphaerota* and formerly known as the *Thaumarchaeota*[38]). MCO1 homologs were also
117 detected in major methane-oxidizing genera including *Methylobacter, Methylocystis,*
118 *Methyloglobulus, Methylosarcina*, and numerous taxa from the phylum *Methylomirabilota*
119 (formerly NC10). Apart from MCO1, however, none of these bacterial or archaeal-derived
120 laccases have activity which has been determined experimentally. Among the 1103 curated MCO1
121 homologs identified in our analysis (Table S1), we therefore selected two additional candidates to
122 evaluate their oxidation potential (enzyme sequences are noted in Table S1). These candidates
123 were selected on the basis of being divergent MCO1 homologs belonging to different phyla than
124 the original MCO1 (*Nitrosomonas*, belonging to the phylum *Pseudomonadota*). Namely, the
125 candidates belonged to phyla known for their methane oxidation potential (*Methylomirabilota*,
126 here termed mMCO) and an archaeal phylum with ammonia-oxidizing members (*Thermoproteota*,
127 here aMCO). The genes encoding these enzymes were synthesized, cloned, heterologously
128 expressed, and purified (Figure S1). We tested for oxidation activity with the model laccase
129 substrate ABTS (2,2′-azino-bis(3-ethylbenzothiazoline-6-sulfonic acid), a commonly used redox
130 mediator. Both mMCO and aMCO exhibited oxidation activity with ABTS (Figure S1). The ability

of these MCO1 homologs to oxidize ABTS indicates that oxidative capability is conserved, even within a divergent archaeal phylum.

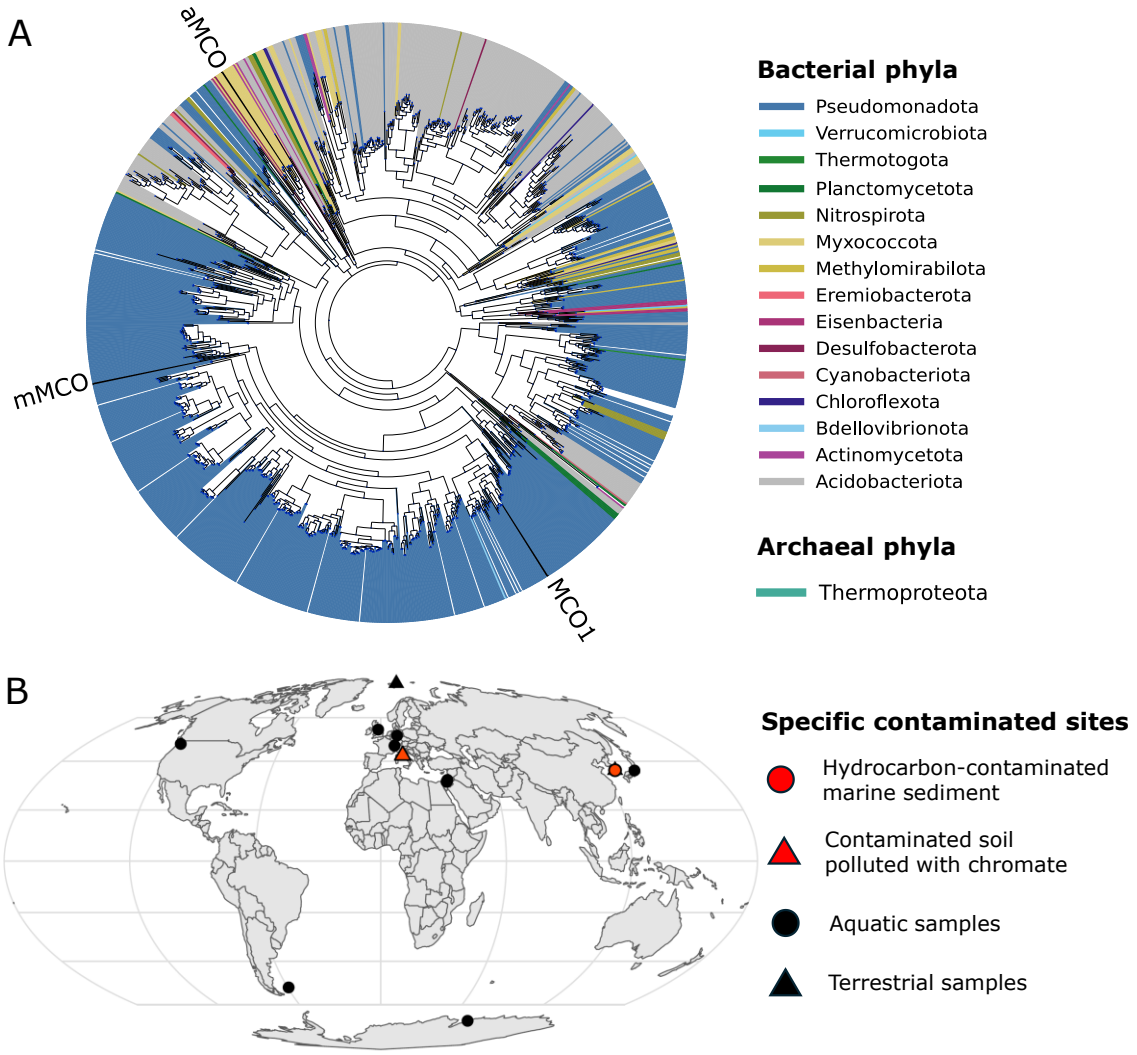

Figure 1. (A) Approximate maximum-likelihood phylogenetic tree of the multicopper oxidase MCO1 from *Nitrosomonas* sp. and two new laccases experimentally validated in this study from a methane-oxidizing bacterium (mMCO, NCBI ID: HEY7867587.1) and a soil ammonia-oxidizing archaeon (aMCO, HYV56308.1) displayed relative to other uncharacterized homologs in global metagenomic datasets colored by metagenomically-assigned phylum of origin. (B) Biogeographic distribution of MCO1 homologs from samples where location information was recorded in environmental sample metadata. Two samples were from notably contaminated sites: Red triangle = soil site contaminated with 1000 mg Cr (VI) kg^{-1}. Red circle = crude-oil contaminated marine sediment site.

143 **High-throughput screening of 183 TrOCs biotransformation in a miniaturized laccase–**
144 **mediator system.**

145 Having established that MCO1 is widely distributed in the environment and its function is
146 conserved across diverse microbial lineages, we next used this enzyme to explore its reactivity
147 against a broader chemical space. To create an enzyme-level benchmark suitable for predictive
148 modeling, we assembled a library of 183 wastewater-relevant trace organic contaminants and
149 quantified their biotransformation by MCO1 under standardized conditions.

150 To enable higher throughput screening while maintaining a chemically controlled oxidative
151 environment, we implemented a miniaturized laccase–mediator system (LMS) using a 96-well
152 plate setup and ABTS as the model mediator (Figure 2A)[6, 26]. Across all compounds, 38 exhibited
153 significant biotransformation (criteria detailed in Methods) (Figure 2B, Table S2). Several
154 compounds that showed extensive biotransformation in previous larger-volume, higher-enzyme-
155 load assays displayed lower reactivity here (Table S2). For instance, amisulpride, sulfadiazine, and
156 isoproturon, which exhibited more than 70% removal in the previous investigation, did not meet
157 the significant biotransformation criteria in this scaled-down high-throughput assay (Table S2)[26].
158 Ranitidine, which was fully degraded in the high-enzyme-load system, demonstrated only partial
159 removal under the current conditions (Figure 2C). Nonetheless, the significant biotransformation
160 observed for 38 TrOCs (Figure 2B) demonstrates that this miniaturized LMS is well suited for
161 surveying enzyme-substrate reactivity across large and diverse contaminant libraries.

162 Compared to 24-well plate setups that are better suited for deriving robust kinetic rate
163 constants under environmentally relevant conditions[39], we used a 96-well format to enable parallel
164 screening of 183 contaminants distributed across 14 sub-mixtures (including all controls) under a
165 unified reaction framework. To balance this focus on throughput with the need for downstream
166 model input, we propose a refined data-interpretation framework (Figure S2) beyond a traditional
167 binary classification of "reactive" versus "non-reactive." Specifically, we define a third
168 "inconclusive" category to capture compounds whose apparent reactivity is sensitive to assay
169 configuration or falls within low-signal fluctuation ranges (Figure S2). Compounds are thus
170 assigned to "reactive", "non-reactive", or "inconclusive", where the "inconclusive" group
171 represents borderline behaviors that cannot be distinguished reliably under the current conditions.
172 We exclude these "inconclusive" cases from the initial predictive modeling to reduce bias.

Collectively, the miniaturized LMS provides a scalable platform for profiling enzyme–substrate interactions.

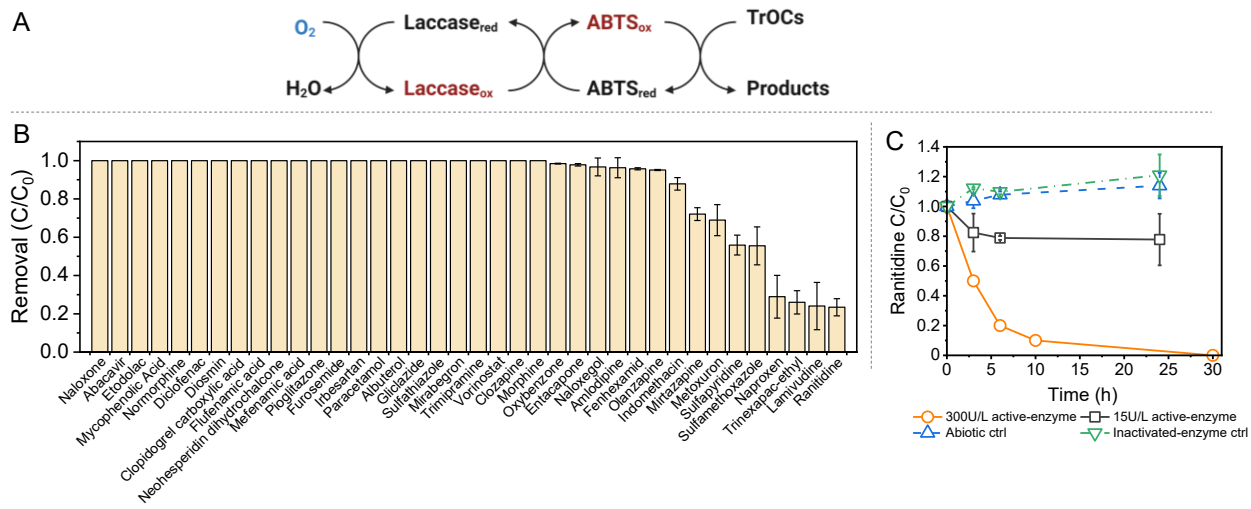

Figure 2. Biotransformation of TrOCs in the LMS. (A) Schematic of the laccase-mediator system; (B) TrOCs removal in the 96-well high-throughput screening assay (15U/L active enzyme in a 250 μL reaction volume); (C) Comparison of reactivity for ranitidine in different reaction systems (yellow, circle: 300U/L active enzyme in a 1000 μL reaction volume; black, box: 15U/L active enzyme in a 250 μL reaction volume; blue, up-triangle: abiotic buffer control with 250 μL reaction volume; green down-triangle: inactivated enzyme control with 250 μL reaction volume).

Graph Convolutional Network (GCN) model for TrOCs biotransformation prediction.

To evaluate whether enzyme-level reactivity data can support predictive modeling, we trained machine-learning classifiers on the curated LMS dataset. After excluding compounds categorized as "inconclusive", 138 chemicals with confident "reactive" (n=38) or "non-reactive" (n=100) assignments were retained for model development. We first applied GCNs to construct a model for chemical biotransformation prediction. GCNs are particularly suitable for predicting biotransformation potential due to their ability to naturally incorporate the graph-like structure of chemical compounds into their learning process. By leveraging the graph representation of molecules, where atoms are nodes and bonds are edges, GCNs can effectively learn complex patterns and interactions crucial for distinguishing between reactive and non-reactive compounds. Unlike traditional molecular descriptors, which rely on pre-defined features and can miss subtle structural variations, GCNs automatically learn molecular features during training. To prevent overfitting, we first pre-trained our GCN using a contrastive learning approach on a dataset of 1

195 million PubChem molecules. Contrastive learning is a technique that helps the model learn
196 meaningful representations by contrasting similar pairs and dissimilar pairs, encouraging the
197 model to map similar molecules closer in the feature space and dissimilar ones further apart[40]. By
198 doing so, the generalization capabilities of models can be improved when applied to downstream
199 tasks, such as molecule classification or property prediction. After pre-training, an average
200 classification accuracy of 0.78 ± 0.08 was achieved for the GCN model fine-tuned on 138
201 compounds (Table S2) with a 10-fold train-test split. By contrast, a benchmark multilayer
202 perceptron (MLP) model trained on the same dataset using Molecular ACCess System (MACCS)
203 descriptors as input achieved an average test accuracy of 0.73 ± 0.12 (Table 1).
204 Furthermore, the 64-dimensional descriptors extracted from the trained GCN model
205 effectively differentiated compounds that underwent biotransformation from those that did not
206 (Figure 3A). The Principal Component Analysis (PCA) plot of these 64-dimensional descriptors
207 for all compounds clearly showed a separation between the two classes (Figure 3A), whereas the
208 PCA of the 167-dimensional MACCS descriptors (Figure 3B) failed to show such a distinction.
209 These results highlight the potential of GCN models in capturing complex molecular patterns
210 relevant to biotransformation prediction. By learning high-quality molecular representations
211 directly from the graph structure of compounds, GCNs provide an excellent tool for predicting the
212 biodegradability of new chemicals.
213 However, the descriptors generated by GCNs remain latent features, making it challenging
214 to interpret the molecular factors driving individual predictions. For applications aimed at the
215 development of SSbD chemicals, understanding why a molecule is predicted to be degradable or
216 persistent, and which structural or chemical properties govern this outcome, is critical. Given that
217 laccase-mediated oxidation is fundamentally governed by electron-transfer processes, we therefore
218 explored whether quantum-chemical descriptors could complement the GCN framework by
219 providing mechanistically-interpretable features. As a proof-of-concept, we analyzed a set of
220 oxidation-relevant electronic parameters[41] to evaluate whether they capture structure–reactivity
221 relationships.
222
223
224

Table 1: GCN prediction performance on train and test data. F1 is defined as the harmonic mean of precision and recall, providing a balanced measure of classification performance by jointly considering false positives and false negatives. AUROC represents the area under the receiver operating characteristic curve and reflects the model's ability to discriminate between positive and negative classes across different classification thresholds.

	GCN		MLP with MACCS	
	Train	Test	Train	Test
Accuracy	0.84±0.02	0.78±0.08	0.76±0.01	0.73±0.12
F1 Score	0.65±0.06	0.63±0.10	0.58±0.02	0.51±0.08
AUROC Score	0.91±0.01	0.82±0.07	0.86±0.02	0.80±0.08

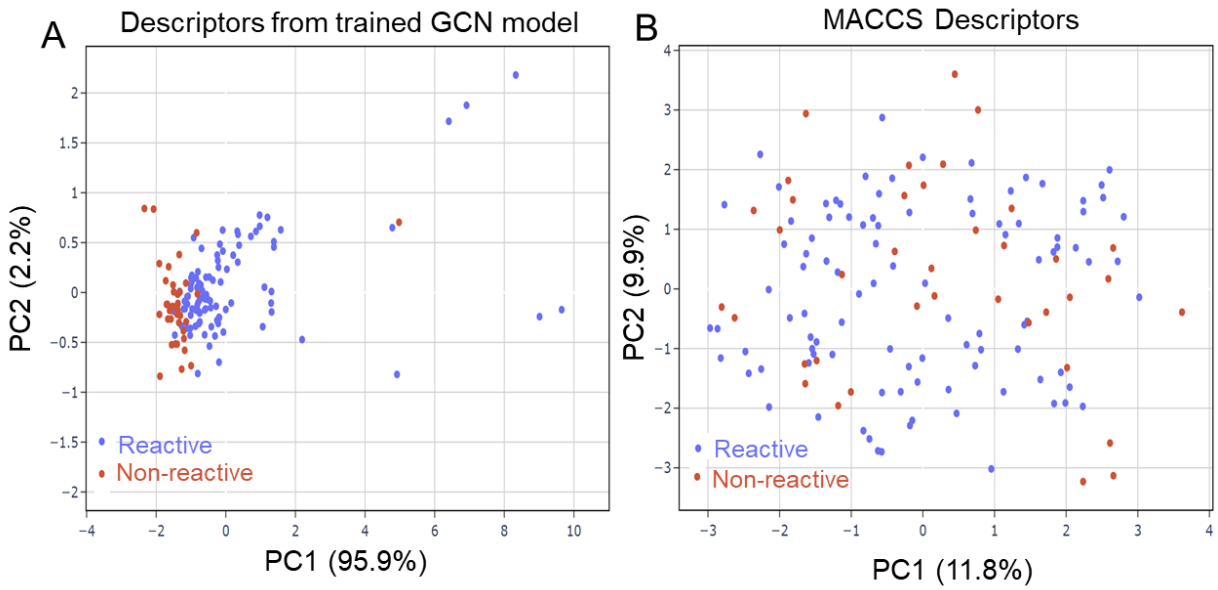

Figure. 3. (A) Principal Component Analysis (PCA) of the descriptors generated by the 64-dimensional GCN model and (B) the MACCS fingerprints. Blue dots represent compounds with positive biotransformation, while red dots represent compounds with no biotransformation.

Exploration of Descriptors from Quantum Chemical (QC) Calculations

Thanks to improved computational resources, it is now feasible to use quantities derived from advanced quantum chemical (QC) calculations as descriptors[41, 42]. The QC descriptors extend classic chemical descriptors based on electronic structure theory. In the case of the LMS,

241 contaminant removal involves the interactions between TrOCs and the mediator-derived radicals,
242 which ultimately initiate oxidation reactions. In other words, the key degradation process in the
243 LMS is the chemical oxidation of the TrOCs by the mediator-derived radical. In the case of ABTS
244 as mediator, the rate-limiting step is likely to be a one-electron transfer from the respective TrOC
245 to the ABTS radical. Thus, we tried to find relevant QC parameters that can effectively capture the
246 potential for TrOCs to be oxidized.

247 As an initial test, we examined the 20 molecules characterized in our previous large-
248 volume experiment (LVE)[26], including 16 molecules also present in the high-throughput LMS
249 dataset, to directly compare between the two experimental setups. In order to search for indicators
250 that can distinguish between the experimentally observed reacted and not-reacted molecular
251 groups, we analyzed the correlations between the reactivity (degradability) and nine QC
252 parameters: one-electron oxidation potential (E_{ox})[43], electronegativity[44], ionization energy[44],
253 electron affinity[44], chemical hardness[44], highest occupied molecular orbital (HOMO) energy,
254 lowest unoccupied molecular orbital (LUMO) energy, HOMO-LUMO energy gap ($\Delta E_{HOMO\text{-}LUMO}$)
255 and polarizability, calculated at the M062X/6-311+G(2df,2p)//M06L/6-311+G(2df,2p) level with
256 the SMD solvent model[45] (1st QC). The results show that the combination of E_{ox} and $\Delta E_{HOMO\text{-}LUMO}$
257 clearly distinguishes between reacted and not-reacted molecules in LVE (Figure 4A). The
258 classification was further validated by employing a support vector machine (SVM) classifier,
259 achieving a cross-validation accuracy score of 0.95, a near-perfect accuracy. Importantly, these
260 two parameters are fundamentally related to oxidation potential, providing a solid theoretical basis
261 for separating the two classes of molecule.

262 Encouraged by these promising results, we extended this approach to the 138 molecules
263 with confident reacted or not-reacted assignments in the curated LMS dataset. In order to
264 effectively handle the larger set of compounds, we opted for a slightly lower level in the QC
265 calculations (2nd-QC) allowing for accelerated calculations while keeping the necessary accuracy
266 for the relevant parameters (see Materials and Methods). The 2nd-QC level yields the same
267 classification accuracy of 0.95 for the reactivity in LVE (Figure 4B), which demonstrated its
268 suitability.

269 We used the 16 molecules used in both experimental systems for a first test. Under the
270 miniaturized LMS conditions in this study, two compounds (1: amisulpride and 2: isoproturon)

were newly determined to be non-reactive due to the lower activity of the miniatured LMS as discussed before. These discrepancies reduced the classification accuracy to 0.77 (1st-QC) and 0.83 (2nd-QC) (Figure 4C–D). When the analysis was extended to the 138 molecules, the boundary between the two classes became less distinct with a decreased accuracy of 0.71 (Figure 5A). The reacted molecules are distributed in the bottom right region, while the top left region contains mostly unreacted molecules. Although the separation is not perfect, this model can still predict rather confidently that an entry in the upper left region is a non-reactive molecule. The original QC calculations treated molecules in their neutral forms without consideration of ionization states. However, several compounds are expected to exist predominantly in ionic rather than neutral forms at the experimental pH of approximately 5–5.5. We then selected 58 from 138 molecules, which were predicted to be predominantly neutral at pH 5 by the Chemaxon method[46]. However, the results did not improve, showing a slightly lower SVM accuracy of 0.67 (Figure 5B). The results indicate that while these two QC descriptors remain important, a more general predictive model will likely require additional molecular descriptors to capture other confounding factors not represented by the current electronic features.

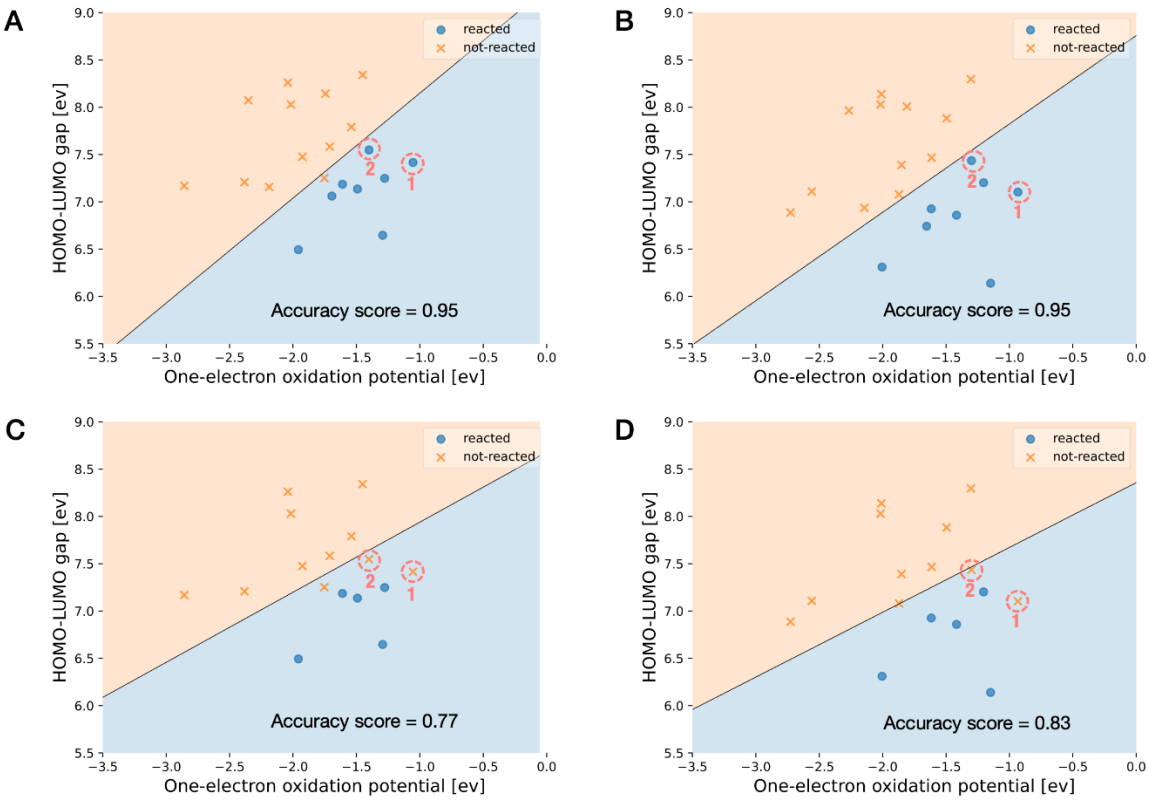

Figure 4. SVM classification of molecules used in the large-volume experiment (LVE) (20 molecules)[26] and in this study (16 molecules). Amisulpride (1) and isoproturon (2) were reactive in LVE but not-reacted in this study. (A) LVE data for 20 molecules calculated at the 1st-QC level. SVM weight (the number of reacted compounds: the number of not-reacted compounds) = 1:1. (B) LVE, 20 molecules, 2nd-QC, SVM weight = 1:1. (C) This study, 16 molecules, 1st-QC, SVM weight = 0.45:1. (D) This study, 16 molecules, 2nd-QC, SVM weight = 0.45:1.

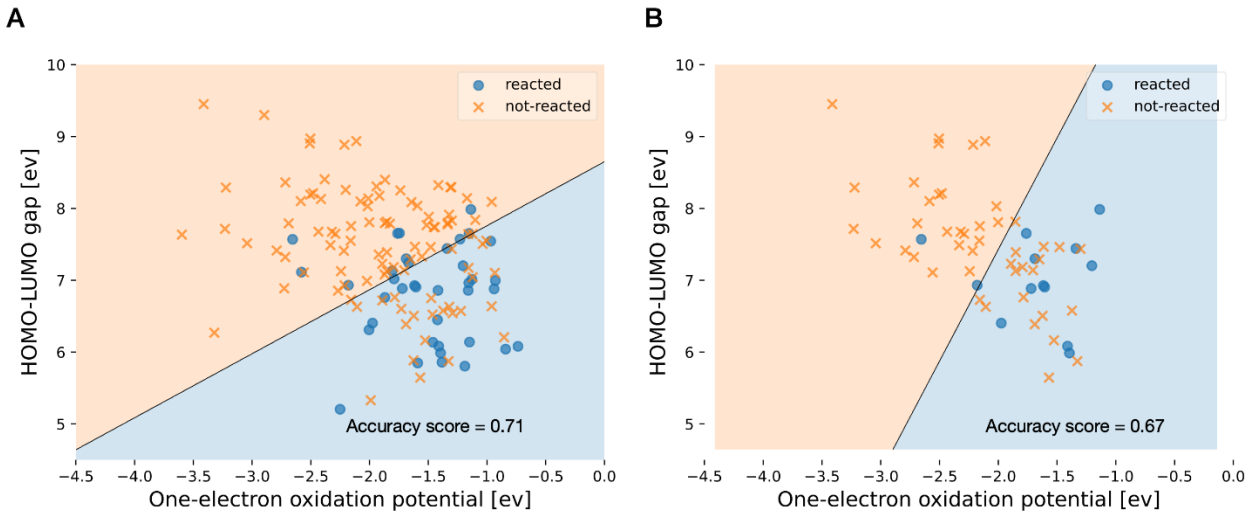

Figure 5. SVM classification of compounds in this study. (A) 138 compounds at the 2nd-QC level, SVM weight = 0.4:1. (B) 58 neutral molecules, 2nd-QC, SVM weight = 0.3:1.

Environmental Implications

Predicting differences in biotransformation of TrOCs across environmental systems remains a fundamental challenge. Existing approaches rely on parameters such as pH, oxygen demand or organic carbon content, which reflect system conditions but do not capture the microbial processes that ultimately determine whether a chemical transforms or persists. Although the need to incorporate microbial signatures into predictive frameworks has been well recognized, progress has been limited because metagenomic and transcriptomic datasets are complex, redundant and only loosely connected to actual chemical reactivity. These limitations highlight the importance of furthering our understanding of chemical-enzyme interactions and developing prediction models that originate from chemical–enzyme interactions rather than from indirect community-level descriptors.

309 This study demonstrates an enzyme-level benchmark for bacterial laccases belonging to
310 the MCO1-family. It additionally broadened the known sequence space of MCO1-family
311 homologs by purifying and testing the activity of laccases from a newly characterized methane-
312 oxidizing bacteria phylum and an ammonia-oxidizing archaeal phylum. Historically, the
313 contribution of ammonia-oxidizing microorganisms to TrOC biotransformation has been
314 attributed to the catalytic promiscuity of ammonia monooxygenase (AMO)[47, 48], a copper-
315 dependent membrane-bound enzyme known for its broad substrate specificity. However,
316 validating the direct role of AMO in TrOC biotransformation has been challenging because such
317 membrane-bound enzymes cannot be readily tested *in vitro*[49]. Evidence from previous studies
318 already hinted that AMO is unlikely to be the only contributor[50]. For example, in nitrifying
319 activated sludge, inhibition with the copper chelator allylthiourea caused a stronger reduction in
320 TrOC biotransformation than treatments with more AMO-specific inhibitors such as octyne[50]. This
321 pattern suggests that additional copper-dependent enzymes within nitrifying communities
322 contribute to TrOC removal. Our findings support this broader view, suggesting a previously
323 overlooked role of bacterial and archaeal laccases in contaminants removal.

324 We screened the activity of the ammonia-oxidizing bacterial laccase MCO1 across a set
325 of >180 wastewater-relevant TrOCs. We combine this high-throughput enzymatic dataset with
326 machine learning and quantum-chemical descriptors to derive mechanistically grounded
327 predictions of biotransformation potential. This work represents an initial step toward building
328 predictive tools that link chemical structure, enzyme specificity and reaction thermodynamics.
329 Although this is a pilot study, the workflow is inherently scalable, and the integration of
330 experimental enzymatic data with mechanistic descriptors offers a versatile foundation for future
331 model expansion and multi-enzyme extensions. One potential direction is to integrate the QC
332 descriptors with conventional QSAR features, which may better capture both local reactivity and
333 global molecular behavior, thereby improving robustness across diverse chemical libraries. In
334 addition, when more experimental records become available in large and chemically diverse
335 datasets, more flexible modeling approaches, including deep learning architectures, could be
336 explored to learn complex interactions between QC descriptors and chemical endpoints.

337 To evaluate how far this bottom-up framework can extend toward describing behavior in a
338 relevant microbial community, we compared our results with an activated-sludge study that
339 examined the same set of chemicals[51]. Ten of the forty-two compounds classified as undergoing

340 oxidative catabolic degradation in that study were also transformed in our enzyme assay (Table
341 S2), representing approximately one quarter of the catabolic group. At least another six compounds
342 active in our assay showed behavior that made a clear delineation between catabolic and co-
343 metabolic transformation difficult in the activated sludge study. This overlap is notable given that
344 our system contains only one oxidative enzyme yet still seems to capture a measurable fraction of
345 community-level oxidative biotransformation. As additional enzyme families are incorporated,
346 this approach is expected to expand in predictive coverage and provide a pathway toward
347 mechanism-informed tools that support environmental fate assessment and the design of more
348 degradable chemicals.
349

350 **Materials and Methods**

351 **Bioinformatic analysis**

352 MCO1 homologs in different publicly available sequencing databases were filtered for protein
353 sequences meeting the following criteria: at least 35% amino acid sequence identity, at least 85%
354 alignment, and a bit score greater than 270. The National Center for Biotechnology Information
355 (NCBI) non-redundant protein sequences database (nr) was queried using BLASTP[52] with the
356 above-listed stringent homology parameters, resulting in 960 unique protein sequences. Since the
357 nr database is not comprehensive for environmental metagenomic sequences, we additionally
358 queried three other databases with the same criteria: the Global Microbial Gene Catalog[27], the BV-
359 BRC database[28], and the Joint Genome Institute Integrated Microbial Genomes (JGI-IMG)
360 database[29]. All databases were accessed on 14 June, 2024. All hits were collated and duplicates
361 were removed, resulting in 1,103 non-redundant MCO1 homologs.

362 Protein sequences were aligned using the DECIPHER sequence- and structure-based multi-
363 sequence alignment algorithm[53] The protein alignment was assessed and alignment edges (first
364 255 alignment positions, last 2078 positions) were trimmed. The trimmed alignment was used for
365 approximate maximum likelihood phylogenetic tree estimation using FastTree v2.0[54]. Tree clades
366 were colored based on phylum information from Genome Taxonomy Database release version 220
367 (GTDB)[55], using the ETE 3 toolkit v 3.1.3[56]. For the BV-BRC sequences with linked metadata,
368 geographic coordinates were extracted and mapped using rnaturalearth[57].
369

370 **Heterologous gene expression.**

371 The investigated multicopper oxidase encoding gene (i.e., MCO1) was identified from the
372 metatranscriptomics sequencing dataset of activated sludge samples, as previously described[26].
373 For gene expression, the MCO1-transformed *Escherichia coli* BL21(DE3) cryo stocks (stored at
374 –80°C) were streaked onto LB agar plates supplemented with 100 μg/mL ampicillin and incubated
375 at room temperature for 48 h. A single MCO1-expressing colony was picked and inoculated into
376 5 mL of terrific broth containing 100 μg/mL ampicillin and cultured overnight at 37°C with
377 shaking. A 500 μL aliquot of this pre-grown culture was then transferred into 500 mL of terrific
378 broth supplemented with 100 μg/mL ampicillin and incubated at 37°C with shaking at 200 rpm
379 until the optical density at 600 nm (OD_{600}) reached approximately 0.5. Protein expression was
380 induced by adding 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). To promote the formation
381 of the functional multicopper oxidase, $CuCl_2$ was added together with IPTG at a final concentration
382 of 0.25 mM. Cultures were transferred to 18°C and shaken for 4 h, followed by overnight
383 incubation at 18°C without shaking to create microaerobic conditions favorable for active enzyme
384 production. Cells were harvested by centrifugation at 5,000 × *g* for 10 min, and the cell pellets
385 were stored at −20°C until use. For the two selected MCO1 homologs, protein expression was
386 carried out using *Escherichia coli* T7. All subsequent induction and cultivation steps followed the
387 same procedures described above for MCO1.
388
389 **Enzyme purification and SDS page.**
390 The cell pellets were resuspended in lysis buffer on ice (50 mM Tris-HCl, pH 7.5, 300 mM NaCl,
391 50% glycerol, 5 mM imidazole, 1 mM PMSF, 1 mg/mL lysozyme) and lysed using a French press
392 (Avestin Emulsiflex C3). Cell debris and other impurities was removed by centrifugation, and the
393 resulting supernatants were collected for enzyme purification. For benchtop purification,
394 supernatants were applied to a Ni-NTA agarose column (Qiagen) pre-equilibrated with buffer A
395 (20 mM Tris-HCl, 500 mM NaCl, 10% v/v glycerol, and 20 mM imidazole). The column was then
396 washed with buffer A followed by two additional washes with buffer B (40 mM Tris-HCl, 500
397 mM NaCl, and 10% v/v glycerol). Final proteins were eluted using buffer C (500 mM imidazole,
398 500 mM NaCl, and 10% v/v glycerol). Protein fractions were collected and analyzed by SDS-
399 PAGE (Figure S1A and Figure S1B). Fractions containing enzyme of interest, consistent with their
400 respective molecular weights, were pooled and subjected to buffer exchange into TNG buffer (50
401 mM Tris-HCl, pH 7.4, 0.1 M NaCl, and 10% v/v glycerol) to remove imidazole. Purified enzymes

402 were aliquoted, flash-frozen in liquid nitrogen, and stored at −80°C until further use. For the
403 purification of MCO1 homologs, Fast Protein Liquid Chromatography (FPLC) was performed
404 using an ÄKTA Pure system (Cytiva) following the procedures established in our previous work.
405 After buffer exchange into TNG buffer, the purified homologs were aliquoted and stored at −80°C
406 as described above for MCO1.
407
408 **Screening of laccase activity on a 96-well plate using a microplate reader.**
409 Laccase activity was quantified colorimetrically using 2,2'-azino-bis(3-ethylbenzothiazoline-6-
410 sulfonic acid) (ABTS) as a chromogenic substrate (Figure S1C and Figure S1D). The assay
411 followed a modified protocol adapted from previously described methods[26]. For MCO1, enzyme
412 concentrations were first determined using the Bradford colorimetric assay. For activity
413 measurements, enzyme stock solutions were diluted in ammonium acetate buffer (pH 5.6).
414 Reactions were carried out in clear, flat-bottom 96-well plates by mixing 20 µL of diluted enzyme
415 samples with 180 µL of buffer containing 500 µM ABTS. Laccase activity was quantified based
416 on the initial rate of absorbance increase at 420 nm within 5 seconds of shaking at 30°C in a
417 microplate reader. Kinetic measurements were conducted for one hour at one-minute intervals.
418 The laccase activity toward ABTS is determined based on Equation 1[58]. For the two additional
419 MCO1 homologs, enzyme activity was assessed qualitatively after purification by directly adding
420 the enzyme to ABTS solution and visually confirming oxidation based on the characteristic color
421 change (Figure S1D).
422
423 $$U/L = \frac{\Delta A \times Df \times 10^6 (\frac{\mu mole}{mole})}{t\ (min) \times \varepsilon\ (M^{-1} * cm^{-1}) \times L\ (cm)} \times \frac{V_{total}(\mu L)}{V_{samplel}(\mu L)} \qquad \text{(Equation 1)}$$
424
425 Where, **U**: enzyme activity (µmole/(min*L)); **ΔA**: the change in absorbance and is calculated
426 as the final absorbance minus the initial absorbance; **Df**: dilution factor; **t**: reaction time (min); **ε**:
427 molar extinction coefficient (36000 M^{-1} cm^{-1}, M=mole/L); **L**: path length, which is 0.56 cm for a
428 96-well plate containing 200µL of solution; $\mathbf{V_{total}}$: total volume of the reaction (µL); $\mathbf{V_{sample}}$:
429 Sample volume (µL).
430
431 **Compound selection and clustering**

432 A total of 183 wastewater-relevant trace organic contaminants (TrOCs), including commonly used
433 pesticides, pharmaceuticals, and artificial sweeteners, were selected for this study. Individual stock
434 solutions were prepared in dimethyl sulfoxide (DMSO) at 10 mM and stored at −20 °C until use.
435 To reduce potential competitive inhibition in the high-throughput screening assay, we partitioned
436 the 183 compounds into 14 sub-mixtures (final concentration at 200 μM for each TrOC in ethanol)
437 using hierarchical clustering. The clustering was based on the compound structural information
438 together with predicted biotransformation reaction rules from enviPath[3, 59]. This design distributes
439 structurally and mechanistically similar compounds across different sub-mixtures, thereby limiting
440 the number of closely related substrates co-incubated in the same well and reducing the likelihood
441 of strong mixture effects (*e.g.*, shared-enzyme competition or combined toxicity). All sub-mixtures
442 were stored at −20 °C for subsequent biotransformation experiments. Detailed sub-mixture
443 information is included in Table S2, same clustering strategy has been reported previously[6].
444
445 **High-throughput biotransformation assay of the investigated enzyme and selected TrOCs.**
446 TrOCs biotransformation by MCO1 was assessed using a clear-bottomed 96-well plate (Greiner).
447 Individual TrOC sub-mixtures were first added to an empty 96-well plate. After complete
448 evaporation of the organic solvent (i.e., ethanol), 240 μL of ammonium acetate buffer (pH ~5.5)
449 was added to each well. The plate was then subjected to shaking at 50 rpm for 15 minutes to
450 redissolve the TrOCs. Subsequently, 5 μL of ABTS stock solution (final concentration 200 μM)
451 was added to the well, followed by the addition of 5 μL of purified enzyme solution (with a final
452 enzyme activity of ~15 U/L). The plate was incubated in a self-designed humidified chamber at
453 30°C[6], with continuous shaking at 150 rpm. To monitor potential evaporation, the total weight of
454 the plate was measured before and after each sampling event. No significant evaporation was
455 detected during the experimental period. Additionally, to ensure sufficient oxygen exposure for
456 the enzyme, in addition to shaking, each well in the 96-well plate was aerated thoroughly using a
457 multichannel pipette every three hours throughout the duration of the experiment. Samples for LC-
458 HRMS/MS analysis were collected at 0, 3, 6, and 24 hours. At each sampling event, a 20 μL sample
459 was taken and mixed with the same amount of pure methanol to completely denature the enzyme.
460 A 160 μL of ammonium acetate buffer was then added to the sample mixture to achieve an
461 approximate 10-fold dilution of the original sample. Following centrifugation at 13,000 × *g* for 15
462 minutes at 4°C, the supernatant was collected and stored at 4°C for LC-HRMS/MS analysis.

463 Additionally, enzyme-free abiotic fresh buffer control and trichloroacetic acid (TCA) treated
464 inactivated enzyme control were prepared in parallel with the biological experiment. All
465 experiments were conducted in duplicate, and samples were measured by LC-HRMS/MS within
466 two days of collection. Matrix-matched calibration standards were used for quantification, with
467 concentrations determined from peak area integration using TraceFinder 5.1 (Thermo Fisher
468 Scientific).
469
470 **Ultrahigh-performance liquid chromatography coupled to high-resolution tandem mass**
471 **spectrometry (UHPLC-HRMS/MS) analysis.**
472 The TrOCs were analyzed using UHPLC–HRMS/MS on a Q Exactive system (Thermo Fisher
473 Scientific) following a method adapted from our previous work[6]. Briefly, for the LC separation,
474 25 µL of each sample was loaded onto an ACQUITY Premier BEH C_{18} column (1.7 µm, 100 ×
475 2.1 mm, Waters) and eluted with nano-pure water (A) and methanol (B), each containing 0.1%
476 formic acid. The flow rate was set at 300 µL/min. A linear elution gradient was as follows: 95%
477 A: 0 – 1.5 min, 95% – 5% A:1 – 7.5 min, 5% A: 7.5 – 9.5 min, and 95% A: .5 – 11.5 min. For
478 HRMS, mass spectra were acquired in full scan mode with a resolution of 70,000 at m/z 200, using
479 positive/negative switching electrospray ionization mode over a scan range of m/z 100 – 1000.
480 Xcalibur 4.0 (Thermo Fisher Scientific) and TraceFinder 5.1 (Thermo Fisher Scientific) were used
481 for data acquisition and analysis. ChemDraw Professional 20.0 and MarvinSketch (v19.20.0,
482 ChemAxon, http://www.chemaxon.com) were used for drawing, displaying, and characterizing
483 chemical structures
484
485 **Data Cleaning Framework for Predictive Modeling in Chemical Fate and Biotransformation**
486 **Studies**
487 Current predictive models for TrOC biotransformation or chemical fate assessment are often built
488 from datasets curated from the literature, where compound classifications typically rely on strict
489 but somewhat arbitrary criteria. A commonly used approach defines "Significant
490 Biotransformation" as >20% biological removal relative to abiotic and dead-cell controls (Figure
491 S2)[47]. Although such thresholds help minimize false positives, they may also misclassify weakly
492 reactive or slowly transforming compounds as non-reactive, thereby biasing predictive models
493 toward reduced sensitivity for these compounds. To better align experimental measurements with

computational model requirements, here, we applied a refined classification scheme that introduces an intermediate "inconclusive" category (Figure S2).

In the dataset used for modeling (Table S2), the maximum removal percentage of individual compounds for different control groups was calculated using Equation 2. Compounds were classified as Positive if they exhibited more than 20% removal and significantly higher removal in the biological control compared to both abiotic and inactive-enzyme controls ($p < 0.05$), consistent with criteria used in previous biodegradation studies. Rather than grouping all remaining compounds as Negative, we assigned compounds showing 10–20% removal and/or lacking statistically significant differences from controls ($p \geq 0.05$) to the "inconclusive" category (Figure S2). Compounds with less than 10% removal were classified as Negative, indicating their persistence under specific conditions. Compounds assigned to the "inconclusive" category were excluded from model training. This selection strategy ensures that the predictive model is calibrated with only reliable data, resulting in more accurate and robust predictions of chemical biotransformation. By applying this criterion, we identified 38 positive and 100 negative results suitable for downstream computational modeling. The detailed compounds classification is included in Table S2.

$$Removal\ Percentage\ (\%) = \frac{(C_0 - C_{final})}{C_0} \times 100\% \quad \text{(Equation 2)}$$

Where, $\mathbf{C_0}$ represents the initial concentration measured by LC-HRMS/MS, and $\mathbf{C_{final}}$ denotes the final concentration after biotransformation experiments.

Graph Convolutional Network (GCN) model construction.

Our Graph Convolutional Network (GCN) model, an extension of the neural network class (nn.Module) in PyTorch, was designed for graph-structured data analysis. The model was initialized with three graph convolutional layers to process the features of nodes, each followed by a ReLU activation function to introduce non-linearity. Convolution layers progressively transformed the node features from the input space to a hidden space comprising 64 channels. To aggregate features of nodes into a graph-level representation, we applied global mean pooling and achieved a 64-dimensional descriptor for each compound. Our model contained a dropout layer for regularization and a linear classifier that mapped the 64-dimensional descriptor to the class space of the dataset. In addition, we designed a dataset class for the systematic conversion of molecular data into graph representations, suitable as input for our model. This class processed

525 molecules encoded as SMILES strings, translating them into graphs and extracting the adjacency
526 information of atoms. Within these graphs, nodes represented atoms and edges denoted chemical
527 bonds. The node features encompassed atomic number, degree, formal charge, hybridization,
528 aromaticity, total hydrogen count, radical electron count, inclusion in rings, and chirality. The edge
529 features included bond type and ring membership. Ten neutral compounds were randomly selected
530 as the test set, while the remaining 128 compounds were used for training. The train-test split was
531 repeated 10 times to ensure a comprehensive evaluation. We also developed a benchmark model
532 by removing the convolutional components from our GCN model and retaining only the final
533 linear classifier, using MACCS descriptors as the input. It allows us to compare the 64-dimensional
534 descriptors learned by the GCN model with the predefined MACCS descriptors[60].
535

536 **Exploration of Quantum Chemical Descriptors**

537 All QC calculations were performed using the Gaussian 16 program package[61]. PubChemPy (ver.
538 1.0.4) was used to obtain geometries that were used as the initial structures for QC calculations.
539 We used Scikit-learn (ver. 1.1.3) for the SVM modeling and Mlxtend (ver. 0.23.1) and Matplotlib
540 (ver. 3.6.2) for the plotting. The ionization states of each compound at pH 4-6 were determined
541 using the Chemaxon pKa calculator (https://chemaxon.com/calculators-and-predictors).

542 The initial structures were obtained from PubChem (https://pubchem.ncbi.nlm.nih.gov)
543 automatically from the compound names. If there is no three-dimensional geometry available, the
544 initial structure was manually built by using GaussView6[61]. The first step is geometry optimization
545 for the initial structure at the M06L/6-311+G(2df,2p) level with the SMD solvent model to give
546 result 1 (R1) followed by a single point calculation for the structure of R1 at the M062X/6-
547 311+G(2df,2p) level to give result 2 (R2). The geometry of R1 is used for geometry optimization
548 of the cation as an open shell doublet system at the M06L/6-311+G(2df,2p) level with the SMD
549 solvent model to give result 3 (R3), followed by a single point calculation to give result 4 (R4).
550 Using these results, one-electron oxidation potential, E_{ox}, was calculated by Equation 3[43].

551 $$E_{ox} = E(P^+) - E(P) + \Delta G(P^+) - \Delta G(P) - 4.28 \text{ eV}, \qquad \text{(Equation 3)}$$

552 where, $E(P^+)$ and $E(P)$ are the potential energies of R4 and R2, respectively, $dG(P^+)$ and $dG(P)$
553 are thermal corrections to the Gibbs free energy of R3 and R1, respectively. The unit is electron
554 volt (eV). The energies of anionic and cationic species for other parameters including

555 electronegativity[44] were obtained by single point calculations of open shell systems. HOMO and
556 LUMO energies were obtained from R2.
557 To obtain the lower QC level for acceleration, we searched for a relevant level by
558 comparing the accuracy of E_{ox} and $\Delta E_{HOMO-LUMO}$ with respect to the 1st QC level with the
559 following four combinations of QC level and methods: (i) PM7//PM7, (ii) M062X/6-31+G(d,p)
560 with SMD//PM7, (iii) M062X/6-31+G(d,p)//PM7 both with SMD and (iv) M062X/6-
561 31+G(d,p)//BLYP/3-21G(d) both with SMD, where PM7 is a semi-empirical method, and SMD is
562 a solvent model. The method before and after the notation "//" is for obtaining energy and geometry,
563 respectively, i.e., the former was used for a single point calculation, and the latter was used for
564 geometry optimization. We employed a dataset containing 167 molecules calculated at the 1st QC
565 level. The R^2 value for E_{ox} and $\Delta E_{HOMO-LUMO}$ calculated with these methods against those by the
566 1st QC level are shown in Table S4. As shown in Table S4, most of the methods show a good
567 correlation for $\Delta E_{HOMO-LUMO}$ but mostly fail for E_{ox}. Only method (iv) resulted in an acceptable
568 R^2 value (0.907). We inspected also the similarity of geometries as well as the relevancy of the
569 parameter values as indicators to separate reactive and non-reactive molecules predicted by the
570 SVM model built using the 20 molecules examined in the previous experimen[26]. Thus, we decided
571 to use method (iv) as the 2nd QC level. E_{ox} is calculated by Equation 3 but without thermal
572 correlation to Gibbs free energy. Using this level, computation became about four times faster
573 compared to that at the 1st QC level.
574
575 **Training Support Vector Machine (SVM) Model**
576 We use the linear kernel function for the SVM modeling. The cross-validation score was
577 performed by five times splitting of data and fitting the model for the 20 and 16 molecules (Figure
578 4) and eight times for the 138 and 58 molecules (Figure 5). For plots in which "reacted" and "not-
579 reacted" compounds are partly mixed, i.e., all plots with the current experiment (Figure 4C, Figure
580 4D, Figure 5A and Figure 5B), weights corresponding to the ratio of the number of "reacted" and
581 "not-reacted" compounds was considered in the SVM modeling. The weight values are listed in
582 the respective figure captions.
583
584 **Supplementary information**

The Supplementary Information includes four Supplementary Tables (Tables S1–S4), provided in an accompanying Excel file. Figures S1 and S2 present the SDS-PAGE characterization of purified MCO1 and its homologs, , as well as the proposed data-cleaning framework used for predictive modeling.

CRediT Authorship Contribution Statement

Yaochun Yu: Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Kunyang Zhang:** Formal analysis, Investigation, Methodology, Writing – original draft, Writing – review & editing. **Vincenz-Maria Steiner:** Formal analysis, Methodology, Writing – review & editing. **Victoria Poltorak:** Formal analysis, Methodology, Visualization, Writing – review & editing. **Silke I. Probst**: Methodology, Writing – review & editing. **Serina L. Robinson:** Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. **Jürg Hutter:** Conceptualization, Supervision, Writing – review & editing. **Hiroko Satoh:** Conceptualization, Formal analysis, Investigation, Methodology, Visualization, Supervision, Writing – original draft, Writing – review & editing. **Kathrin Fenner:** Conceptualization, Supervision, Funding acquisition, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare no competing financial interest.

Data Availability

Scripts used for the bioinformatics analysis in this manuscript are available at : https://github.com/MSM-group/laccase-paper. Geometries of all compounds as well as all computational descriptors are available at https://github.com/shiroco/Data/tree/main/wasser-degradation/138. Scripts related to Graph Convolutional Network (GCN) model in this manuscript are available at https://github.com/zhangky12/mco1

Acknowledgments

This work was supported by the Swiss National Science Foundation (project no. 200021L_201006, for Y.Y. and K.F.; PZPGP2_209124, no. 501100001711-209124 for V.P. and S.L.R.), the Eawag Academic Transition Grant (for Y.Y.), the European Union's H2020 research and innovation program under the Marie Sklodowska Curie grant agreement (award no. MSCA-ITN-H2020[956496], for K.Z.), and the Peter und Traudl Engelhorn Stiftung (for S.I.P.). We thank Philipp Longree from the Department of Environmental Chemistry (Uchem) at Eawag for the method development on the UHPLC-HRMS/MS. We thank Elia Ceppi (Uchem, Eawag) for assisting with the sample measurement on UHPLC-HRMS/MS. We thank Cleo Soldini (Uchem, Eawag), Till Epprecht (Department of Environmental Microbiology, Umik, Eawag), Niklas Ferenc Trottmann (Umik, Eawag), and René Gall (Umik, Eawag) for their technical support on the enzyme purification. We thank Martina Kalt (Uchem, Eawag) and Miguel Heussi (Uchem, Eawag) for preparing chemical stock solutions. We thank Dr. Sema Karakurt-Fischer (Process Engineering Department, Eawag) for designing and kindly sharing the humidified chamber. We thank Dr. Anastasia Athanasakoglou (formerly a postdoctoral researcher at Uchem, Eawag) for the thorough discussions about the enzyme purification protocol.

References

(1) Caldeira, C.; Farcal, L.; Garmendia, I.; Mancini, L.; Tosches, D.; Antonio, A.; Rasmussen, K.; Rauscher, H.; Riego Sintes, J.; Sala, S. Safe and Sustainable by Design chemicals and materials- Framework for the definition of criteria and evaluation procedure for chemicals and materials. EUR 31100 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-53264-4, doi:10.2760/487955, JRC128591.

(2) Hafner, J.; Fenner, K.; Scheidegger, A. Systematic Handling of Environmental Fate Data for Model Development─Illustrated for the Case of Biodegradation Half-Life Data. *Environ. Sci. Technol. Lett.* **2023**, *10* (10), 859-864. DOI: 10.1021/acs.estlett.3c00526.

(3) Zhang, K.; Fenner, K. enviRule: an end-to-end system for automatic extraction of reaction patterns from environmental contaminant biotransformation pathways. *Bioinform.* **2023**, *39* (7). DOI: 10.1093/bioinformatics/btad407.

(4) Zahn, D.; Arp, H. P. H.; Fenner, K.; Georgi, A.; Hafner, J.; Hale, S. E.; Hollender, J.; Letzel, T.; Schymanski, E. L.; Sigmund, G.; et al. Should Transformation Products Change the Way We

Manage Chemicals? *Environ. Sci. Technol.* **2024**, *58* (18), 7710-7718. DOI: 10.1021/acs.est.4c00125.

(5) Löffler, P.; Lai, A.; Henschel, H.; Spilsbury, F.; Deviller, G.; Tarazona, J. V.; Lai, F. Y. Enhanced Risk Assessment of Transformation Products through Chemical Similarity Analysis. *ACS ES&T water* **2024**, *4* (5), 1949-1951. DOI: 10.1021/acsestwater.4c00240.

(6) Yu, Y.; Trottmann, N. F.; Schärer, M. R.; Fenner, K.; Robinson, S. L. Substrate promiscuity of xenobiotic-transforming hydrolases from stream biofilms impacted by treated wastewater. *Water Res.* **2024**, *256*, 121593. DOI: https://doi.org/10.1016/j.watres.2024.121593.

(7) Hofer, U. The majority is uncultured. *Nat. Rev. Microbiol.* **2018**, *16* (12), 716-717. DOI: 10.1038/s41579-018-0097-x.

(8) Salcher, M. M.; Layoun, P.; Fernandes, C.; Chiriac, M.-C.; Bulzu, P.-A.; Ghai, R.; Shabarova, T.; Lanta, V.; Callieri, C.; Sonntag, B.; et al. Bringing the uncultivated microbial majority of freshwater ecosystems into culture. *Nat. Commun.* **2025**, *16* (1), 7971. DOI: 10.1038/s41467-025-63266-9.

(9) Fenner, K.; Elsner, M.; Lueders, T.; McLachlan, M. S.; Wackett, L. P.; Zimmermann, M.; Drewes, J. E. Methodological Advances to Study Contaminant Biotransformation: New Prospects for Understanding and Reducing Environmental Persistence? *ACS ES&T water* **2021**, *1* (7), 1541-1554. DOI: 10.1021/acsestwater.1c00025.

(10) Marti, T. D.; Schweizer, D.; Yu, Y.; Schärer, M. R.; Probst, S. I.; Robinson, S. L. Machine Learning Reveals Signatures of Promiscuous Microbial Amidases for Micropollutant Biotransformations. *ACS Environ. Au* **2025**, *5* (1), 114-127. DOI: 10.1021/acsenvironau.4c00066.

(11) Wang, H.; Fu, T.; Du, Y.; Gao, W.; Huang, K.; Liu, Z.; Chandak, P.; Liu, S.; Van Katwyk, P.; Deac, A.; et al. Scientific discovery in the age of artificial intelligence. *Nature* **2023**, *620* (7972), 47-60. DOI: 10.1038/s41586-023-06221-2.

(12) Gallegos, M.; Vassilev-Galindo, V.; Poltavsky, I.; Martín Pendás, Á.; Tkatchenko, A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nat. Commun.* **2024**, *15* (1), 4345. DOI: 10.1038/s41467-024-48567-9.

(13) Maeda, K.; Hirano, M.; Hayashi, T.; Iida, M.; Kurata, H.; Ishibashi, H. Elucidating Key Characteristics of PFAS Binding to Human Peroxisome Proliferator-Activated Receptor Alpha:

An Explainable Machine Learning Approach. *Environ. Sci. Technol.* **2024**, *58* (1), 488-497. DOI: 10.1021/acs.est.3c06561.

(14) Cordero Solano, J. A.; Hafner, J.; McLachlan, M. S.; Singer, H.; Fenner, K. Predicting Micropollutant Removal in Wastewater Treatment Based on Molecular Structure: Benchmark Data and Models. *Environ. Sci. Technol.* **2025**, *59* (41), 22020-22028. DOI: 10.1021/acs.est.5c09314.

(15) Yang, Y.; Shan, C.; Pan, B. Machine learning modeling of fluorescence spectral data for prediction of trace organic contaminant removal during UV/H2O2 treatment of wastewater. *Water Res.* **2024**, *255*, 121484. DOI: https://doi.org/10.1016/j.watres.2024.121484.

(16) Farzanehsa, M. S. Z.; Carvajal, G.; McDonald, J.; Khan, S. J. Predicting the concentration range of trace organic contaminants in recycled water using supervised classification. *J. Water Proc. Eng.* **2024**, *58*, 104709. DOI: https://doi.org/10.1016/j.jwpe.2023.104709.

(17) Zhang, C.; Yu, Q.; He, Y.; Wu, G.; Fan, F.; Xu, K.; Ren, H.; Geng, J. Decoding pharmaceutical removability in full-scale wastewater treatment plants via machine learning model integrating modified electrophilicity index. *Water Res.* **2025**, *286*, 124218. DOI: https://doi.org/10.1016/j.watres.2025.124218.

(18) Rios-Miguel, A. B.; van Bergen, T. J. H. M.; Zillien, C.; Ragas, A. M. J.; van Zelm, R.; Jetten, M. S. M.; Hendriks, A. J.; Welte, C. U. Predicting and improving the microbial removal of organic micropollutants during wastewater treatment: A review. *Chemosphere* **2023**, *333*, 138908. DOI: https://doi.org/10.1016/j.chemosphere.2023.138908.

(19) Zhang, K.; Schwaller, P.; Fenner, K. Predicting Toxicity toward Nitrifiers by Attention-Enhanced Graph Neural Networks and Transfer Learning from Baseline Toxicity. *Environ. Sci. Technol.* **2025**, *59* (9), 4518-4529. DOI: 10.1021/acs.est.4c12247.

(20) Escher, B. I.; Baumer, A.; Bittermann, K.; Henneberger, L.; König, M.; Kühnert, C.; Klüver, N. General baseline toxicity QSAR for nonpolar, polar and ionisable chemicals and their mixtures in the bioluminescence inhibition assay with Aliivibrio fischeri. *Environ. Sci.: Processes Impacts* **2017**, *19* (3), 414-428, 10.1039/C6EM00692B. DOI: 10.1039/C6EM00692B.

(21) Maeder, V.; Escher, B. I.; Scheringer, M.; Hungerbühler, K. Toxic Ratio as an Indicator of the Intrinsic Toxicity in the Assessment of Persistent, Bioaccumulative, and Toxic Chemicals. *Environ. Sci. Technol.* **2004**, *38* (13), 3659-3666. DOI: 10.1021/es0351591.

705 (22) Qin, W.; Henneberger, L.; Glüge, J.; König, M.; Escher, B. I. Baseline Toxicity Model to
706 Identify the Specific and Nonspecific Effects of Per- and Polyfluoroalkyl Substances in Cell-
707 Based Bioassays. *Environ. Sci. Technol.* **2024**, *58* (13), 5727-5738. DOI:
708 10.1021/acs.est.3c09950.
709 (23) Zhu, J.-J.; Yang, M.; Ren, Z. J. Machine Learning in Environmental Research: Common
710 Pitfalls and Best Practices. *Environ. Sci. Technol.* **2023**, *57* (46), 17671-17689. DOI:
711 10.1021/acs.est.3c00026.
712 (24) Zhu, J.-J.; Boehm, A. B.; Ren, Z. J. Environmental Machine Learning, Baseline Reporting,
713 and Comprehensive Evaluation: The EMBRACE Checklist. *Environ. Sci. Technol.* **2024**, *58*
714 (45), 19909-19912. DOI: 10.1021/acs.est.4c09611.
715 (25) Cañas, A. I.; Camarero, S. Laccases and their natural mediators: Biotechnological tools for
716 sustainable eco-friendly processes. *Biotechnol. Adv.* **2010**, *28* (6), 694-705. DOI:
717 https://doi.org/10.1016/j.biotechadv.2010.05.002.
718 (26) Athanasakoglou, A.; Fenner, K. Toward characterizing the genetic basis of trace organic
719 contaminant biotransformation in activated sludge: the role of multicopper oxidases as a case
720 study. *Environ. Sci. Technol.* **2022**, *56* (1), 313-324. DOI: 10.1021/acs.est.1c05803.
721 (27) Coelho, L. P.; Alves, R.; del Río, Á. R.; Myers, P. N.; Cantalapiedra, C. P.; Giner-Lamia, J.;
722 Schmidt, T. S.; Mende, D. R.; Orakov, A.; Letunic, I.; et al. Towards the biogeography of
723 prokaryotic genes. *Nature* **2022**, *601* (7892), 252-256. DOI: 10.1038/s41586-021-04233-4.
724 (28) Olson, R. D.; Assaf, R.; Brettin, T.; Conrad, N.; Cucinell, C.; Davis, James J.; Dempsey,
725 Donald M.; Dickerman, A.; Dietrich, Emily M.; Kenyon, Ronald W.; et al. Introducing the
726 Bacterial and Viral Bioinformatics Resource Center (BV-BRC): a resource combining PATRIC,
727 IRD and ViPR. *Nucleic Acids Res.* **2022**, *51* (D1), D678-D689. DOI: 10.1093/nar/gkac1003.
728 (29) Grigoriev, I. V.; Nordberg, H.; Shabalov, I.; Aerts, A.; Cantor, M.; Goodstein, D.; Kuo, A.;
729 Minovitsky, S.; Nikitin, R.; Ohm, R. A.; et al. The Genome Portal of the Department of Energy
730 Joint Genome Institute. *Nucleic Acids Res.* **2011**, *40* (D1), D26-D32. DOI: 10.1093/nar/gkr947.
731 (30) Labrenz, M.; Collins, M. D.; Lawson, P. A.; Tindall, B. J.; Schumann, P.; Hirsch, P.
732 *Roseovarius tolerans* gen. nov., sp. nov., a budding bacterium with variable bacteriochlorophyll
733 a production from hypersaline Ekho Lake. *Int. J. Syst. Bacteriol.* **1999**, *49 Pt 1*, 137-147. DOI:
734 10.1099/00207713-49-1-137.

735 (31) Lee, D. H.; Moon, S. R.; Park, Y. H.; Kim, J. H.; Kim, H.; Parales, R. E.; Kahng, H. Y.
736 *Pseudomonas taeanensis* sp. nov., isolated from a crude oil-contaminated seashore. *Int. J. Syst.*
737 *Evol. Microbiol.* **2010**, *60* (Pt 12), 2719-2723. DOI: 10.1099/ijs.0.018093-0.
738 (32) Viti, C.; Decorosi, F.; Tatti, E.; Giovannetti, L. Characterization of chromate-resistant and -
739 reducing bacteria by traditional means and by a high-throughput phenomic technique for
740 bioremediation purposes. *Biotechnol. Prog.* **2007**, *23* (3), 553-559. DOI:
741 https://doi.org/10.1021/bp0603098.
742 (33) Diefenbach, T.; Sumetzberger-Hasinger, M.; Braunschmid, V.; Konegger, H.; Heipieper, H.
743 J.; Guebitz, G. M.; Lackner, M.; Ribitsch, D.; Loibner, A. P. Laccase-mediated degradation of
744 petroleum hydrocarbons in historically contaminated soil. *Chemosphere* **2024**, *348*, 140733.
745 DOI: https://doi.org/10.1016/j.chemosphere.2023.140733.
746 (34) Alshiekheid, M. A.; Umar, A.; Ameen, F.; Alyahya, S. A.; Dufossé, L. Biodegradation of
747 chromium by laccase action of Ganoderma multipileum. *J. King Saud Univ. Sci.* **2023**, *35* (10),
748 102948. DOI: https://doi.org/10.1016/j.jksus.2023.102948.
749 (35) Rodríguez Couto, S.; Toca Herrera, J. L. Industrial and biotechnological applications of
750 laccases: A review. *Biotechnol. Adv.* **2006**, *24* (5), 500-513. DOI:
751 https://doi.org/10.1016/j.biotechadv.2006.04.003.
752 (36) Gao, Y.; Xiao, M.; Zou, H.; Nurwono, G.; Zgonc, D.; Birch, Q.; Nadagouda, M. N.; Park, J.
753 O.; Blotevogel, J.; Liu, C.; et al. Laccase Immobilized on Arginine-Functionalized Boron Nitride
754 Nanosheets for Enhanced Atrazine Degradation. *Environ. Sci. Technol.* **2024**, *58* (34), 15111-
755 15119. DOI: 10.1021/acs.est.4c02849.
756 (37) Clark, M.; Tepper, K.; Petroll, K.; Kumar, S.; Sunna, A.; Maselko, M. Bioremediation of
757 Industrial Pollutants by Insects Expressing a Fungal Laccase. *ACS Synth. Biol.* **2022**, *11* (1), 308-
758 316. DOI: 10.1021/acssynbio.1c00427.
759 (38) Pessi, I. S.; Rutanen, A.; Hultman, J. *Candidatus Nitrosopolaris*, a genus of putative
760 ammonia-oxidizing archaea with a polar/alpine distribution. *FEMS Microbes* **2022**, *3*. DOI:
761 10.1093/femsmc/xtac019.
762 (39) Partanen, S. B.; Mueller, N.; Fenner, K. High-Throughput Miniaturized Biotransformation
763 Testing Using Activated Sludge Enables Rapid Chemical Persistence Assessment. *Environ. Sci.*
764 *Technol. Lett.* **2025**, *12* (11), 1561-1566. DOI: 10.1021/acs.estlett.5c00859.

(40) Wang, Y.; Wang, J.; Cao, Z.; Barati Farimani, A. Molecular contrastive learning of representations via graph neural networks. *Nat. Mach. Intell.* **2022**, *4* (3), 279-287. DOI: 10.1038/s42256-022-00447-x.

(41) Satoh, H.; Hafner, J.; Hutter, J.; Fenner, K. Can AI help improve water quality? Towards the prediction of degradation of micropollutants in wastewater. *CHIMIA* **2023**, *77* (1/2), 48-55. DOI: 10.2533/chimia.2023.48.

(42) Satoh, H.; Steiner, V.-M.; Hutter, J. "Quantum-Chemoinformatics" for design and discovery of new molecules and reactions. In *Drug Development Supported by Informatics*, Satoh, H., Funatsu, K., Yamamoto, H. Eds.; Springer Nature Singapore, 2024; pp 225-245.

(43) Arnold, W. A. One electron oxidation potential as a predictor of rate constants of N-containing compounds with carbonate radical and triplet excited state organic matter. *Environ. Sci.: Processes Impacts* **2014**, *16* (4), 832-838, 10.1039/C3EM00479A. DOI: 10.1039/C3EM00479A.

(44) Grimmel, S. A.; Reiher, M. On the predictive power of chemical concepts. *CHIMIA* **2021**, *75* (4), 311. DOI: 10.2533/chimia.2021.311.

(45) Marenich, A. V.; Cramer, C. J.; Truhlar, D. G. Universal solvation model based on solute electron density and on a continuum model of the solvent defined by the bulk dielectric constant and atomic surface tensions. *J. Phys. Chem. B* **2009**, *113* (18), 6378-6396. DOI: 10.1021/jp810292n.

(46) Chemaxon pKa calculator*, https://playground.calculators.cxn.io/ (accessed on 20 January 2026)*

(47) Yu, Y.; Han, P.; Zhou, L.-J.; Li, Z.; Wagner, M.; Men, Y. Ammonia monooxygenase-mediated cometabolic biotransformation and hydroxylamine-mediated abiotic transformation of micropollutants in an AOB/NOB coculture. *Environ. Sci. Technol.* **2018**, *52* (16), 9196-9205. DOI: 10.1021/acs.est.8b02801.

(48) Han, P.; Yu, Y.; Zhou, L.; Tian, Z.; Li, Z.; Hou, L.; Liu, M.; Wu, Q.; Wagner, M.; Men, Y. Specific micropollutant biotransformation pattern by the comammox bacterium *Nitrospira inopinata*. *Environ. Sci. Technol.* **2019**, *53* (15), 8695-8705. DOI: 10.1021/acs.est.9b01037.

(49) Arp, D. J.; Sayavedra-Soto, L. A.; Hommes, N. G. Molecular biology and biochemistry of ammonia oxidation by *Nitrosomonas europaea*. *Arch. Microbiol.* **2002**, *178* (4), 250-255. DOI: 10.1007/s00203-002-0452-0.

(50) Men, Y.; Achermann, S.; Helbling, D. E.; Johnson, D. R.; Fenner, K. Relative contribution of ammonia oxidizing bacteria and other members of nitrifying activated sludge communities to micropollutant biotransformation. *Water Res.* **2017**, *109*, 217-226. DOI: https://doi.org/10.1016/j.watres.2016.11.048.

(51) Kalt, M.; Ceppi, E.; Robinson, S. L.; Fenner, K. Breaking it down: Unveiling the roles of chemical structure, concentrations and technology on micropollutant biotransformation in wastewater treatment plants. *Water Res.* **2026**, *292*, 125283. DOI: https://doi.org/10.1016/j.watres.2025.125283.

(52) Camacho, C.; Coulouris, G.; Avagyan, V.; Ma, N.; Papadopoulos, J.; Bealer, K.; Madden, T. L. BLAST+: architecture and applications. *BMC Bioinform.* **2009**, *10* (1), 421. DOI: 10.1186/1471-2105-10-421.

(53) Wright, E. S. Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R. *R J.* **2016**, *8* (1), 352-359.

(54) Price, M. N.; Dehal, P. S.; Arkin, A. P. FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. *Plos One* **2010**, *5* (3). DOI: 10.1371/journal.pone.0009490.

(55) Parks, D. H.; Chuvochina, M.; Chaumeil, P.-A.; Rinke, C.; Mussig, A. J.; Hugenholtz, P. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **2020**, *38* (9), 1079-1086. DOI: 10.1038/s41587-020-0501-8.

(56) Huerta-Cepas, J.; Serra, F.; Bork, P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.* **2016**, *33* (6), 1635-1638. DOI: 10.1093/molbev/msw046.

(57) Massicotte, P. S., A.; Hufkens, K. rnaturalearth. World Map Data from Natural Earth, Version 1.1.0. **2025**. DOI: DOI: 10.32614/CRAN.package.rnaturalearth.

(58) Baltierra-Trejo, E.; Márquez-Benavides, L.; Sánchez-Yáñez, J. M. Inconsistencies and ambiguities in calculating enzyme activity: The case of laccase. *J. Microbiol. Methods.* **2015**, *119*, 126-131. DOI: https://doi.org/10.1016/j.mimet.2015.10.007.

(59) Wicker, J.; Lorsbach, T.; Gütlein, M.; Schmid, E.; Latino, D.; Kramer, S.; Fenner, K. enviPath – The environmental contaminant biotransformation pathway resource. *Nucleic Acids Res.* **2016**, *44* (D1), D502-D508. DOI: 10.1093/nar/gkv1229 (acccessed 12/5/2025).

(60) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL Keys for Use in Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273-1280. DOI: 10.1021/ci010132r.

(61) Frisch, M. J.; Trucks, G. W.; Schlegel, H. B.; Scuseria, G. E.; Robb, M. A.; Cheeseman, J. R.; Scalmani, G.; Barone, V.; Petersson, G. A.; Nakatsuji, H.; et al. Gaussian 16 Rev. C.01. In Wallingford, CT; 2016.