A unified pipeline for discovering previously unknown enzyme activities

Ariane Mora[1,2]*, Julia C. Reisenbauer[1,3*]^, Helen Schmid[1,4], Ikumi Miyazaki[1,5], Yueming Long[1], Jason Yang[1], Ryen O'Meara[1], Frances H. Arnold[1]^

^Corresponding authors. frances@cheme.caltech.edu; Julia.Reisenbauer@ist.ac.at

[1]Division of Chemistry and Chemical Engineering, California Institute of Technology, California Blvd., 91125, California, USA.

[2]Current address: AITHYRA GmbH Research Institute for Biomedical Artificial Intelligence of the Austrian Academy of Sciences Helmut-Qualtinger-Gasse 2, Stg. 2, 1030 Vienna, Austria.

[3]Current address: Institute of Science and Technology Austria (ISTA), Am Campus 1, 3400 Klosterneuburg, Austria.

[4]Department of Biosystems Science and Engineering, ETH Zürich, Basel, Switzerland.

[5]Department of Chemistry, Graduate School of Science, The University of Tokyo, Bunkyo-ku.

*These authors contributed equally.

Abstract

Enzymes catalyze diverse chemical transformations and offer a sustainable approach to both breaking and making chemical bonds. However, finding an enzyme capable of performing a specific chemical reaction remains a challenge. We developed a new framework, Enzyme-toolkit (Enzyme-tk), that integrates 23 open-source tools to enable the discovery of enzymes that have activity toward a specific target reaction. Additionally, we introduce two new methods to facilitate enzyme discovery: (1) Func-e, an ML tool that searches large databases for enzymes that potentially catalyze a specific chemical transformation and (2) Oligopoolio, a gene assembly approach that reduces the cost of accessing protein sequences and thus the barrier to their experimental validation. We applied Enzyme-tk to find enzymes for chemical degradation of two man-made pollutants, di-(2-ethylhexyl) phthalate (DEHP) and triphenyl phosphate (TPP). We demonstrate that new, previously unannotated enzymes with favorable characteristics, such as high thermostability, can be identified using Enzyme-tk for reactions that are dissimilar to the training set.

Introduction

Enzymes can catalyze a plethora of chemical transformations and are used across many industries, from pharmaceutical manufacturing to degradation of environmental pollutants. However, finding an enzyme that performs a specific biocatalytic transformation remains a challenge, especially for unnatural substrates.[1] These enzymatic starting points are usually discovered by screening arrays of diverse enzymes for a desired function, a resource- and time-intensive process. Popular webtools such as MetaCyc,[2] KEGG,[3] and others[4–6] enable enzymes to be queried based on similarity. Results are then based on correlations between ligands or between native enzymatic reactions and the desired chemical transformation given by the user. Alternatively, protein databases such as BRENDA[7] and UniProt[8] can be mined to find enzymes that catalyze a specific known reaction or class

of reactions. However, this discovery process remains *ad hoc* and non-standardized, requiring users to navigate across many resources, either manually filtering or writing bespoke scripts to select for specific characteristics.

Recently, machine learning (ML) methods have been presented to complement sequence- and substrate-based approaches to discover enzymes with no annotated homologs or similar (natural) reactions.[9] Several tools have been developed recently to assign a reaction or an enzyme to an enzyme commission (EC) number or the enzyme's ability to perform a given reaction.[10–12] As with similarity searching tools, there is no standard format for files or the environment in which to run these tools. Finally, most ML methods do not experimentally validate predictions, especially for novel enzymatic functions on non-natural substrates.[10,12,13]

For new or unannotated chemical transformations, both similarity and ML search methods have shown limited generalizability to "out-of-distribution" reactions for novel transformations or substrates that are currently not captured in databases.[13] Experimental success rates for finding active enzymes for even annotated functions remain low and significantly drop for novel transformations.[14] In most cases, prediction methods are only tested *in silico,*[15] in part owing to the high cost of experimental validation and lack of standardized methods for closing the ML-experimental loop. The cost of a full-length gene resulting in a 200 amino acid (AA) sequence is between $70 and $170 (Twist, IDT, September 2025). Several recent approaches have proposed new methods leveraging pools of oligomers that consist of multiple gene fragments, assembling full-length genes using Golden Gate assembly[16–18] or using Polymerase Cycling Assembly (PCA).[19–21] There currently exists no standard pipeline to discover biocatalysts, with tools fragmented between the chemistry and the protein domains, and no guidelines for the integration of experimental data.

To fill this gap, we introduce a framework for enzyme discovery and engineering from diverse data sources, beginning with ML discovery and including both the experimental workflow and feedback loop. Enzyme-tk includes 23 open-source tools, in addition to a new ML classifier, Func-e, to discover enzymes across large databases that may perform specific reactions and an experimental protocol to enable the evaluation of enzymes assembled via a cost-effective gene assembly method, Oligopoolio. We applied our approach to two candidate reactions with unnatural substrates, finding enzymes that catalyze the first step of the degradation pathways of commonly found persistent organic pollutants. We envision that this work will serve as a core framework for ML engineers who may aim to improve and add new open-source ML models and experimentalists who wish to use the tools to facilitate experimental workflows for discovering new enzymes for further engineering.

Results

A generalizable pipeline for enzyme discovery

We developed a new pipeline to bridge the computational and experimental gap in ML-driven discovery of enzymes. The Enzyme-tk pipeline has three core modules: 'predict', 'synthesize', and 'validate', Figure 1.A. Enzyme-tk is a Python package consisting of 23 open-source tools that are composable. Enzyme-tk enables researchers to turn a large set of unannotated sequence reads and a chemical transformation into filtered designs of gene constructs for testing enzymatic activities, Figure 1.B. Standardized data formats are passed between tools to bypass conversion between FASTA, CSV, and other output file formats, see Table 1 for a list of tools. Each module ingests and outputs CSV data, making visualization and interaction simple for researchers. Furthermore, Enzyme-tk handles the different computational environments for each module by using conda, a python environment manager[22], automatically activating and deactivating respective conda

environments. Enzyme-tk is designed with modularity as a principle, which facilitates different enzyme discovery objectives and is designed to be easily extensible by the broader enzyme community. Figure 1.B provides an overview of all modules of the pipeline which can be interchangeably run using a single line of code.

To validate the utility of this pipeline, we sought to identify enzymes capable of performing two specific degradation reactions, 1) hydrolysis of di-(2-ethylhexyl) phthalate (DEHP) to mono(2-ethylhexyl) phthalate (MEHP) and 2) hydrolysis of triphenyl phosphate (TPP) to diphenyl phosphate (DPP), Figure 1.C. DEHP and TPP are representatives of two classes of common pollutants, phthalates and organophosphates. Phthalates are commonly used as plasticizers and, owing to their high use and bioaccumulation, are found abundantly in surface waters and wastewater biosolids.[23] The higher molecular weight species, such as endocrine-disrupting DEHP,[24] pose challenges for natural degradation owing to their sterically demanding side chains.[25] Similar degradation challenges have been encountered for some organophosphate esters, such as TPP, which are used as flame retardants, industrial additives, and plasticizers.[26,27] This phosphate ester has also been linked to reproductive and aquatic toxicity.[28,29,30]

The enzymes already reported to perform these target transformations are not ideal candidates for real-world applications, such as in wastewater plants. Their large size (QHH21706.1 is an esterase from *Bacillus subtilis*[31] that comprises 489 amino acids (AA) and *Sb*-PTE is a phosphotriesterase from Sphingobium sp. TCM1 comprising 583 AA)[26] and structural features (*Sb*-PTE contains disulfide bonds) limit their utility. Given the limitations of these existing enzymes, herein referred to as QHH (ID: 5387) and *Sb*-PTE (ID: 5388), we sought to find alternative enzymes with high activity and more favorable characteristics (small, thermostable, and high expression levels in *Escherichia coli*).

To search for potential enzymes, we curated two discovery sets, an "extremophile dataset" and a "metagenomics dataset." The first dataset consists of 8,793 enzymes from SwissProt with favorable characteristics: short sequences (100–350 AAs) from bacterial or archaeal thermostable organisms (>50 °C).[32] For the second dataset, we sought to investigate mining unannotated microbial sequences and selected a metagenomics dataset originating from the influent and activated sewage sludge from five wastewater treatment plants across India, the United States, Switzerland, Sweden, and Hong Kong (see Methods for processing details). We filtered the dataset to 2,708 putative esterases by predicting the EC numbers using two ML tools, proteInfer[33] and CLEAN[11], see Methods for details.

Experimental validation of baseline computational methods

To test the discovery loop using conventional methods, we applied two existing approaches to identify potential enzymes. The first is a baseline approach in which sequences are selected based on substrate similarity, and the second is our previously published multi-modal ML approach, CREEP, that uses contrastive learning to embed sequences into a joint space, enabling sequences to be extracted based on reactions that are projected "close" in the latent space.[13] Reviewed bacterial or archaeal sequences from UniProt were filtered for the following EC classes: DEHP: 3.1.1.74, 3.1.1.1, 3.1.1.3; TPP: 3.1.3.48, 3.1.4.4, 3.1.3.2, 1.7.1.6; CREEP ML embedding approach based on DEHP: 3.1.1.3, 3.1.1.101, 3.1.1.74, 3.5.2.6, 3.1.2.2, 3.3.2.10, 3.1.1.76.), resulting in 40 putative enzymes to test for activity, see Methods for details.

The high cost of gene synthesis motivated us to design a new approach to build genes from fragments using pooled Polymerase Cycling Assembly (PCA), Figure 1.D. Ordering gene fragments in a pooled fashion reduces the cost of gene synthesis for a 96-well plate with IDT by 45%. Using the

Oligopoolio submodule of the pipeline to select the optimal fragment size and overlaps, 40 genes were ordered with fragments ranging from 414 to 894 base pairs to test the assembly method. A two-step PCR assembly protocol was developed, see Methods. Full-length genes were identified and sequence-verified using LevSeq[34] in 96-well plate format. We found that 34/40 genes were correctly assembled with 100% sequence identity (> 3.5 pmol/oligo/µL, IDT oligomer pool). Shorter genes assembled more efficiently. Hence, to ensure representation of longer sequences, at least five clones per variant needed to be picked, see SI Section 3 for a comparison to existing methods, S. Table 13 and assembly statistics, S. Figure 2-8. S. Table, 14-16.

The assembled genes were overexpressed in *E. coli* and screened in 96-well plate format for the degradation of DEHP and TPP using clarified cell lysate (see Methods and SI for details). To efficiently detect product formation in our experimental pipeline, we employed relatively high substrate loadings (DEHP: 3.9 g/L; TPP: 326 mg/L), which greatly exceed concentrations typically found in wastewater streams (e.g. DEHP: 0–8840 ng/L[25]; TPP: 0–491 ng/L[35]). Although no enzymes had activity comparable to the enzymes known to have these activities (positive controls), a low level of activity was observed for the carboxylesterase Q06174 (ID: 5395) on TPP, S. Figure 22. To demonstrate the ability to select candidates with desirable traits such as engineerability, we performed one round of random mutagenesis (*ep*PCR) on Q06174, see SI Section 2 and 6 for details. We identified five variants that exhibited increased activity compared to Q06174 under validation conditions. These mutants exhibited activity on TPP equivalent to the positive control, with the best Q06174-3 variant harboring the K138E, D178G, and D236G mutations (ID: 5396), Figure 2. No enzymes were identified with activity for DEHP hydrolysis.

Improving performance with a new ML method for searching sequence space

Given the limited success of the EC-based approaches for identifying enzymes capable of the hydrolysis of DEHP and TPP, we developed a new sequence-only based ML model, Func-e, to predict the activity of an enzyme for a desired reaction. The model uses embeddings of the enzyme sequence and chemical reaction space and combines these with two cross attention heads. These are then fed into a neural network with a prediction head and 13 regression heads that predict features such as molecular weight and sequence length (see Figure 3.A for model architecture). The intersection of the reaction and protein subsets from CARE[13], a recently published benchmark to measure the performance of enzyme function prediction tools, was used for training and *in silico* evaluations (see Methods for details, S. Table 1 for dataset sizes, S. Table 2 for similarity to target reactions and positive controls). After training for 20 epochs, the model shows some *in-silico* generalization capacity, with an expected drop in performance as the test set becomes more dissimilar to the training set, S. Figure 1. The model that was used to inform the final architecture had an accuracy of 0.75, 0.61, and 0.56 on a balanced dataset with sequences which were less than 50% similar to the training set, on the easy (held out reactions from EC 4), medium (entire held out classes from EC 4), and hard test sets (entire held out classes from EC 3 classes), see SI Section1 for details.

In addition to applying the model to the metagenomics set described above, we increased the search space of the extremophile dataset to include unreviewed enzymes from UniProt, resulting in a dataset size of 8,793 putative enzymes. Enzymes with a mean prediction >0.5 and mean distance between the predicted length and the true sequence length of < 50 amino acids were selected. The predicted enzymes were passed through our previously developed active site annotation filter[36] and structural filtering pipeline (https://github.com/moragroup/StructureZyme, see Methods for details). Based on manual inspection of the structures, 14 enzymes from the extremophiles and 17

from the metagenomics set were selected for experimental validation. The 31 predicted genes were assembled using Oligopoolio and screened in 96-well plate format, S.Table 17 for plate format and SI Section 5 for experimental results. Five enzymes showed potential activity and were selected for validation, four from the extremophile dataset and one metagenomics sequence, S.Figure 12 and 16. Of these, two showed significantly higher activity toward the desired reaction compared to the best enzymes from the literature, Q7SIG1 (ID: 5389) for DEHP and I3NWL3 (ID: 5391) for TPP hydrolysis, Figure 3B-C.

Discussion

A significant problem in enzyme engineering is how to best discover an enzyme with activity towards a specific reaction. Recent reports have sought to enhance the identification of new enzymes by using ML to annotate uncharacterized sequences to EC numbers in low homology regimes[11] and new reactions to EC numbers.[13] Building on this work, we developed a ML model, Func-e, to predict whether an enzyme can perform a specific reaction. In combination with this core ML module, we consolidated over twenty publicly available tools and unified them in a single pipeline, Enzyme-tk, to streamline the workflow to identify suitable enzymatic starting points with desirable traits for a specific catalytic function from a user-specified "sequence search space". In addition to the commonly used sequences in SwissProt, datasets were curated to increase the sequence search space for potentially effective biocatalysts. Finally, to reduce the cost of the experimental pipeline, a gene assembly method based on oligos, Oligopoolio, was developed as a sub-module to bridge the gap between *in silico* and experimental testing.

Oliogopoolio is a new joint computational and experimental method for assembling full-length genes using low-cost oligo pools. While various approaches have been reported for achieving similar tasks,[16,17,20,37–39] efficient command line tools for the generation of constructs from multiple distinct sequences using PCA in a single test tube remain scarce. Oligopoolio was used to assemble multiple pools with varying success, often reaching 100% recovery of all desired genes in a single oligo pool sample. However, we noticed that a higher concentration of DNA in the oligo pool drastically impacted the probability of producing sequence-correct variants. Hence, we suggest using 50 pmol pools from IDT (> 3.5 pmol/oligo/µL, IDT opool). With further methodological optimization, we believe that the protocol could also be amenable to lower concentrations of oligo pools. This module of the pipeline could also be swapped for an alternative method such as Golden Gate assembly[16]. Not designed for highly similar genes or for large gene assemblies, Oligopoolio instead bridges the gap between ordering single genes and a plate of diverse genes in a cheap and accessible manner. Assembly of highly similar genes would likely result in chimeric products, and the success rate of assembly in larger pools could be limited by the DNA concentration. We find the oligo pool approach to be an effective method for not only reducing the cost of synthesis but also minimizing the use of expensive enzyme stocks and reagents, as it requires only two pooled PCRs to construct all desired genes in a single tube within a single day.

With the modular Enzyme-tk pipeline in hand, we first evaluated the performance of a similarity-based module, instead of the more complex Func-e ML tool, to filter for potential enzyme candidates applied to SwissProt sequences. This subset of the pipeline enabled us to identify an enzyme, Q06174, from an extremophilic thermostable organism catalyzing the hydrolysis of TPP to DPP. One additional round of directed evolution provided Q06174-3 (with K138E_D178G_D236G substitutions; ID: 5396), which is less than half the length of literature-known *Sb*-PTE and exhibits modest activity toward TPP hydrolysis. This demonstrates the modularity and compact workflow from a sequence dataset to filtering and validation of capable enzyme variants.

To further extend the search space to our curated metagenomics and extremophile datasets, the developed ML model, Func-e, was implemented to search for enzymes for the partial degradation of DEHP into MEHP and TPP into DPP. Neither of the tested reaction smiles are similar to the training set, for DEHP, the closest reaction similarity is 0.46 and for TPP it is 0.48, using Tanimoto similarity with structural fingerprints, S. Table 1. *In-silico* predicted enzymes were then assembled via Oligopoolio and experimentally evaluated. For the DEHP hydrolysis, an unreviewed enzyme from Uniprot, Q7SIG1 (ID: 5389), originating from thermophilic *Bacillus acidocaldarius*, demonstrated catalytic activity. Q7SIG1 is a hydrolase (no EC annotation in UniProt[31][40] This is in contrast to the literature enzyme QHH, which is able to fully degrade DEHP into phthalic acid (PA). Though this is unsurprising, as the first degradation step of this transformation was the sole input into the ML model. For the partial TPP degradation, another unreviewed carboxylesterase originating from *Geobacillus sp*. ZH1, I3NWL3 (ID: 5391, EC:3.1.1.1, 246 AA length), was identified, which is 337 AAs shorter than previously reported *Sb*-PTE. Translating to a less than 30% sequence identity between the newly discovered enzymes and the positive control sequences.[41] demonstrating the predictive and filtering capabilities of our computational framework. Neither of these enzymes would have been discovered using a simple EC search on annotated data: I3NWL3 was active on TPP, yet EC3.1.1.1 was not predicted as a top EC for this reaction, while for Q7SIG1 no EC number is assigned in UniProt. Both enzymes were identified by screening fewer than 30 enzymes (out of 8,793 thermophilic enzymes considered), reducing experimental cost and discovery time. Of note, given the high lipophilicity of these substrates and the associated risk of phase separation at such concentrations under aqueous conditions, a qualitative comparison is most appropriate under the developed analytical conditions and sufficient for finding initial activities that can be improved by protein engineering/directed evolution.

In contrast to other models,[10] we opted to design a protein sequence-only method to enable rapid *in-silico* filtering of predictions, before passing the predicted sequences to more computationally intensive structural filtering steps. In future work, ensembling of multiple ML models could potentially improve the activity prediction, further increasing the success rate of finding active enzymes within a small pool of experimentally tested variants. The pipeline has been designed modularly to enable exactly such integrations.

Conclusion

We designed Enzyme-tk to be a platform for the prediction, synthesis, and experimental validation of enzymes with a desired catalytic activity. The code is fully open source and build as an airflow-style interface to enable the composition of sub-modules. This modularity will allow its adoption across disciplines, where enzymes with specific traits for desired chemical transformations need to be identified or improved. In showcasing the utility of the pipeline for two bioremediation tasks, we also provide the community multiple ready-to-use pipelines that will be of broad utility, including 1) metagenomics reads to EC classification and reaction prediction, 2) sequence filtering for enzyme traits, prediction, active site annotation, structural filtering, and oligo design. These pipelines along with documentation are available on our GitHub (https://github.com/ArianeMora/enzyme-tk). We hope to build a thriving open-source community that intersects the enzyme annotation and enzyme design domains.

Acknowledgments

The authors thank Charlie Trimble and the Resnick Institute through the Ideation Grant for their generous contributions to the project. A.M. is supported by the Schmidt Science Fellows, in

partnership with the Rhodes Trust. J.C.R. acknowledges partial support from the Swiss National Science Foundation (SNSF) Postdoc. Mobility (P500PN_214290). I.M. thanks the JST SPRING (JPMJSP2108) and the Forefront Physics and Mathematics Program to Drive Transformation (FoPM) within the World-leading Innovative Graduate Study (WINGS) Program of the University of Tokyo. J.Y. is supported by the NSF Graduate Research Fellowship Program and a Google PhD Fellowship. R.L.O. is supported by the National Science Foundation Graduate Research Fellowship Program (NSF grant DGE-1745301). The authors would like to thank Sabine Brinkmann-Chen for careful proofreading and valuable discussions, and Jarrid Rector-Brooks for careful proofreading.

Declaration of interests

The authors declare no competing interests.

Author contributions

A.M., J.C.R., and F.H.A. conceived the idea. A.M. developed the ML model. A.M., H.S., and I.M. developed the computational pipeline, with input from R.L.O. on the dataset conceptualization and J.C.R., J.Y., and Y.L. on the pipeline conceptualization. H.S. developed the structural pipeline. J.Y. ran the CREEP model. J.C.R. developed screening assays and the experimental testing pipeline. J.C.R., H.S., I.M., A.M., R.L.O., and Y.L. performed experimental validations. A.M., J.C.R., H.S., and I.M. developed the oligopool assembly strategy. A.M. and Y.L. developed the oligopool sequencing analysis pipeline. A.M., J.C.R., and F.H.A. wrote the manuscript, and all authors edited and revised the manuscript.

Figures

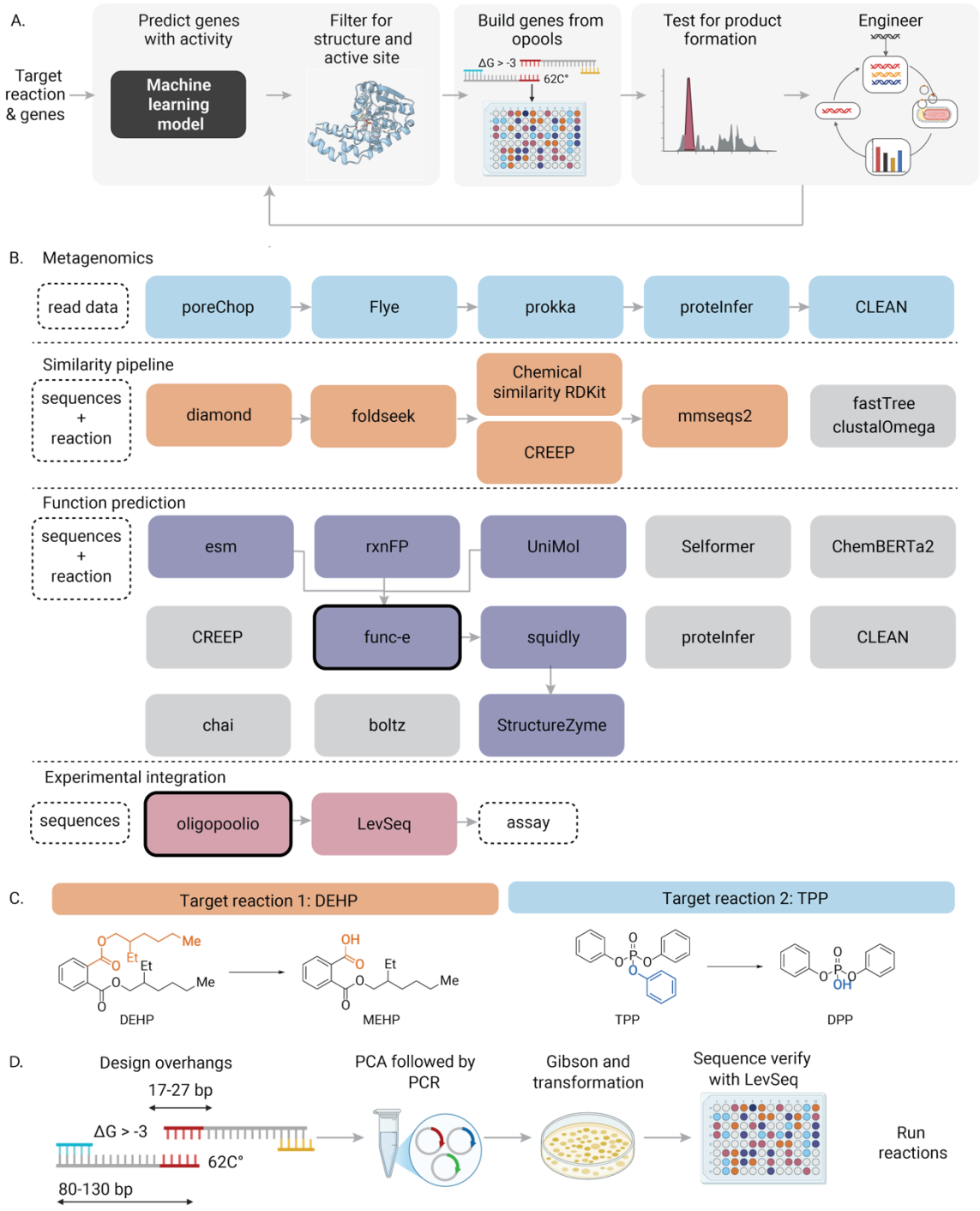

Figure 1. *A. High-level overview of the Enzyme-tk pipeline. Users input a desired target reaction and a sequence database on which they wish to predict activity for, the program then predicts which sequences within this set are most likely to be able to catalyze the reaction. Predictions are filtered using structural and active site predictions. Genes are then synthesized using an oligomer assembly*

method, Oligopoolio, designed to reduce the cost of gene synthesis from oligopools (opools). Experimental validation is performed by testing for a specific target transformation. Successful variants can be further explored and engineered for a desired process or re-included in the ML loop, depending on the scientific or engineering objective. B. 23 computational open-source tools are currently included in Enzyme-tk, here organized around three pipelines created for this study, which are shown highlighted with the colored boxes. Each module can be linked together interchangeably or substituted by alternative tools (highlighted in grey boxes; e.g. in the function prediction module, CLEAN could be used instead of Func-e). Novel tools introduced in this study (black outlined boxes). C. Target reactions of the two partial degradation pathways, 1) hydrolysis of di-(2-ethylhexyl) phthalate (DEHP) to mono(2-ethylhexyl) phthalate (MEHP) and 2) hydrolysis of triphenyl phosphate (TPP) to diphenyl phosphate (DPP). D. Overview of the Oligopoolio pipeline: overhangs are automatically designed, ordered from a DNA provider, and assembled with a polymerase cycling assembly step, followed by PCR to amplify the full assemblies. Circular vectors are created using Gibson assembly and transformed into E. coli. Inserts are sequence-verified for accuracy with LevSeq[34].

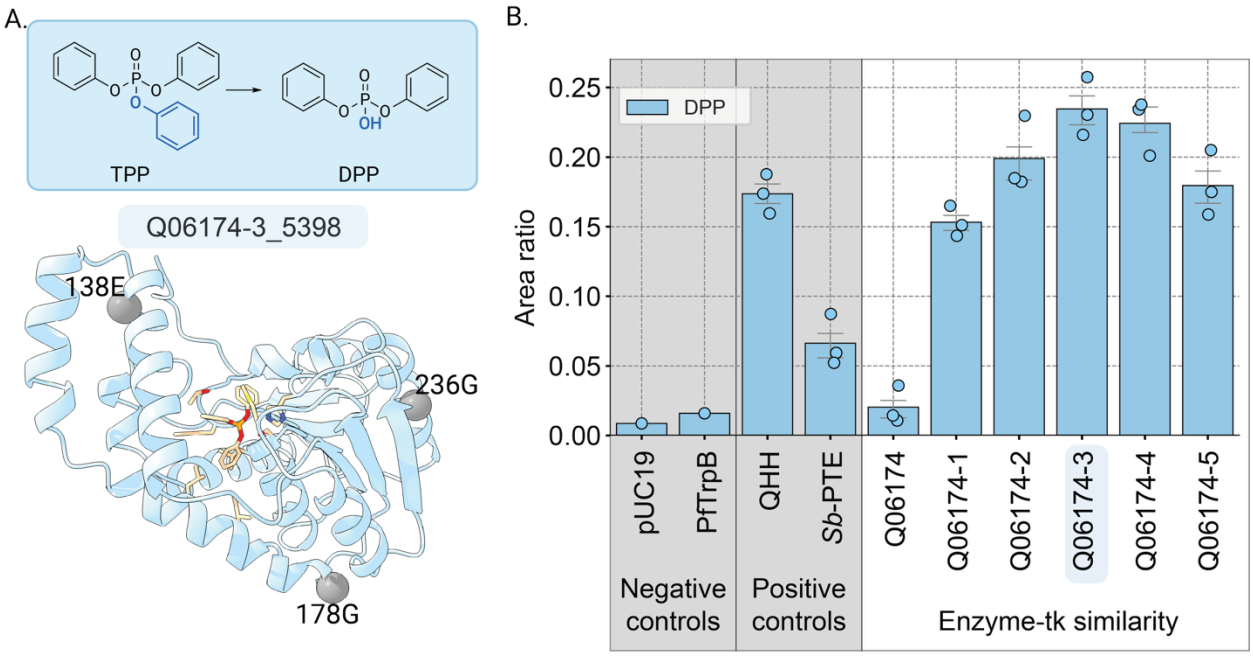

Figure 2. *A. Structure of Q06174 co-folded with TPP using Chai[42] showing the substitutions of the best performing variant Q06174-3 (with K138E_D178G_D236G substitutions). B. Validation results for the epPCR library showing area ratio (integrated product area/integrated internal standard area) for the formation of DPP from TPP. Reactions were performed in technical triplicate under validation conditions, and papaverine hydrochloride was used as an internal standard when analyzing product formation via HPLC-(MS)-analysis.*

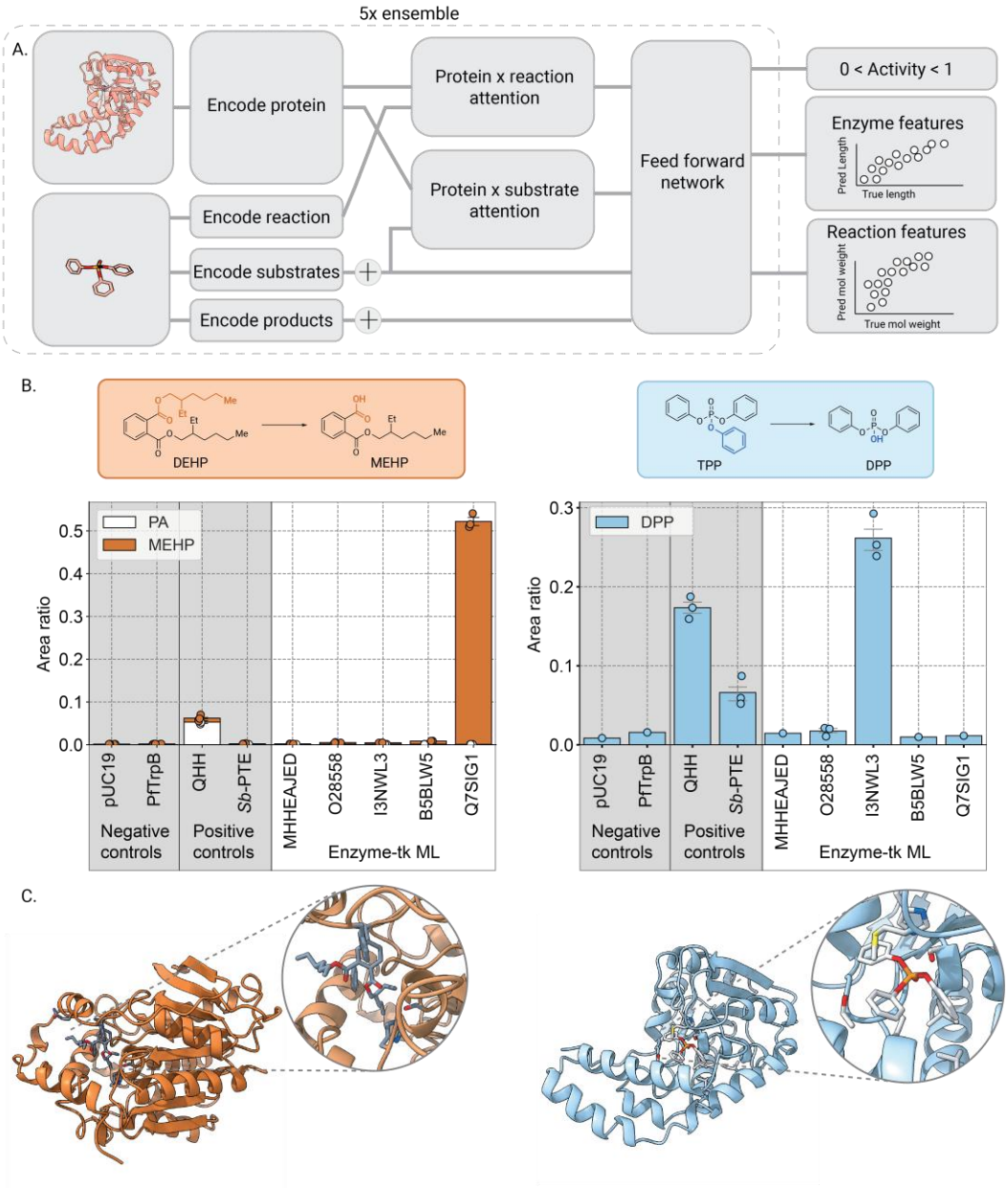

Figure 3. *A. Overview of the architecture of our predictive ML model, Func-e. B. Validation to test enzyme activity towards MEHP and DPP formation. Clarified cell lysate reactions were performed under validation conditions in technical triplicate, and papaverine hydrochloride was used as an internal standard when analyzed by HPLC-(MS)-analysis. Results show integrated product area normalized by the internal standard area for each enzyme and target reaction. PUC19 and PfTrpB are negative controls, QHH is a positive control for DEHP to PA hydrolysis and additionally has promiscuous activity towards TPP degradation. Sb-PTE is a positive control for the TPP to DPP hydrolysis. C. Co-folded structure using Chai for Q7SIG1 with the DEHP substrate and I3NWL3 with the TPP substrate.*

Table 1: Description of tools currently available in Enzyme-tk. Some tools contain multiple functions, however, the main function used in Enzyme-tk has been listed in the table.

Module	Name	Description	Reference
Metagenomics	PoreChop	Used to filter adapters for nanopore sequences in metagenomics pipeline.	[43]
Metagenomics	Flye	Used to assemble the metagenomes.	[44]
Metagenomics	Prokka	Annotation of genes within the genome.	[45]
Function prediction	Proteinfer	Annotation of genes to function (GO or EC class) using ML.	[33]
Function prediction	CLEAN	Annotation of genes to EC class using ML.	[11]
Function prediction	CREEP	Annotation of genes to EC class using ML.	[13]
Function prediction	Func-e	Annotation of genes to reaction using ML.	This study.
Function prediction	Squidly	Annotation of catalytic residues using ML.	[36]
Embedding generation	ESM2 & 3	Conversion of amino acid sequence to a numerical embedding using a PLM.	[46,47]
Embedding generation	RxnFP	Conversion of reaction smiles to a numerical embedding using a language model.	[48]
Embedding generation	Selformer	Conversion of reaction selfies to a numerical embedding using a language model.	[49]
Embedding generation	Uni-mol	Conversion of molecule smiles to a numerical embedding using a language model.	[50]
Embedding generation	ChemBERTa2	Conversion of reaction smiles to a numerical embedding using a language model.	[51]
Docking	Chai	Diffusion based folding of a protein and ligand.	[42]
Docking	Boltz	Diffusion based folding of a protein and ligand.	[52]
Similarity	Diamond	Sequence similarity calculation using basic local alignment search.	[53]
Similarity	Foldseek	Fast structure similarity search.	[54]
Similarity	MMseqs	Fast sequence clustering.	[55]
Docking	StructureZyme	Alignment and calculation of structure metrics.	[56]
Oligo design	Oligopoolio	Calculation of oligo fragments for gene assembly.	This study.
Sequencing	LevSeq	Sequence verification of protein variants.	[34]
MSA generation	ClustalOmega	Creation of multiple sequence alignments (MSA).	[57]
Phylogenetic tree generation	FastTree	Creation of multiple phylogenetic trees.	[58]

Materials and Methods

Data selection and processing

Data were downloaded from UniProt on the 17th of February 2025 by selecting reviewed sequences and the following additional information (Fragment, Length, Mass, Organism ID, Taxonomic Lineage (IDs), Active Site, Cofactor, EC number, Reviewed), totaling 572,970 protein sequences. These were filtered for size (minimum length 100, maximum 1024), to be non-fragments, and proteins with an EC annotation, resulting in 261,254 potential enzymes. For each sequence, the mean summarized esm3 embeddings (esm3-open, mean of the per-token-embedding)[46,47] were computed.

Reaction data were downloaded from EnzymeMap,[59] containing 62,896 deduplicated reactions, or 54,090 products and 37,263 substrates. Properties were calculated for each substrate and product by using RdKit version 2024.3.3, calculated properties were molecular weight (MolWt), total polar surface area (TPSA), calculated solubility proxy (MolLogP), and the maximum and minimum Gasteiger partial charges (MaxPartialCharge and MinPartialCharge). For reactions with multiple substrates and products, these were summed to create a single embedding. For each reaction, chemberta2[51] and rxnFP[48], additionally for each substrate and product the uniMol[50] representation was calculated.

Metagenomics data processing pipeline

Long-read metagenomics data were downloaded from a study[60] that investigated antibiotic resistance genes both from influent and activated sewerage sludge from five wastewater treatment plants from India, the United States, Switzerland, Sweden, and Hong Kong. These were available via the SRA under the following accessions, SRX8190902, SRX8190901, SRX8190900, SRX8190899, SRX8190898, SRX8190897. SRA toolkit was used to download and unzip the data. Porechop[43] was used to trim the nanopore adaptors, before Flye was used to assemble the metagenomes.[44] Genes were annotated to the genomes using Prokka,[45] before enzyme commission numbers were predicted for each gene using ProteInfer[33] and CLEAN.[11] Default parameters were used for all tools. Genes were retained if they fell within the EC numbers, EC:3.1 or 3.- (i.e. retained if unable to be assigned at the second level), and finally, filtered for length (100 < length < 900 amino acids). These genes were used as the "metagenomics set" referred to herein.

Extremophile dataset processing

The temperature was annotated to the enzymes using data from[32] a resource containing the optimal growth temperature data for 21,498 organisms. Of the 261,254 UniProt enzymes, 218,840 were able to be mapped to an organism temperature. Growth temperature is used as a proxy for thermostability. Temperature annotations were mapped at the species level if it existed, otherwise genus, then family, and order. If no order was identified, temperature was omitted from this sample. After the first pass on the EC classification dataset, we extended this to unreviewed sequences in UniProt (filtering for bacteria or archaea and protein existance-1-2, or annotation level of 4-5), resulting in a dataset of 8,793 thermophilic enzymes.

Literature search

To find a positive control for the degradation of DEHP and TPP, scientific literature was searched by using the terms "DEHP/TPP enzymatic degradation", finding several papers that met our criteria. The final selection was based on reports describing the culturing, isolation, and testing of specific enzymes in detail, selecting QHH21706.1, a *para*-nitrobenzyl esterase from *Bacillus subtilis,*[31] and *Sb*-PTE, a phosphotriesterase from *Sphingobium sp*. TCM1, as our positive controls.[26]

Training and test sets setup

Training and test sets were created using the splits as defined in CARE.[13] The reaction and enzyme datasets were combined by taking the intersection of the training and test sets for protein similarity 30-50 (EC's and enzymes). These sets were filtered to ensure no contamination between the test and training sets, by running diamondBlast[53](v 2.1.11) against the training set for each test set (sequence similarity), and FoldSeek[54](v 10.941cd33) (structural similarity). The PDB was used as the database in FoldSeek with the pdb structures mapped to their respective UniProt identifier. Any training data entries that mapped to the test sets with an identity > 0.5 (both structural and sequence) were omitted from the training datasets. The reactions for each EC were also tested for similarity to the training set, here, the reaction distance (Tanimoto similarity) from RDkit was used (v 2024.3.3). Note, some reactions had a high identity to a reaction in another EC class; in these instances, the reaction was considered ambiguous and removed from both EC classes.

To create the training and test sets, pairs of positive and negative data were created by selecting one reaction as a positive from the EC annotated to the enzyme and one as a negative from an EC number that was not annotated to the enzyme, and these were annotated as 1 and 0, respectively. This was done for each of the 184,529 enzymes in the training set for the protein space. Note only ECs from each of the reaction sets were used, making an "easy", "medium", and "hard" training set for reactions and enzymes. The test sets were created in a similar manner, where a sequence was annotated as positive for a single reaction for the annotated EC number and also assigned to another random EC as a negative data point, for which a complementary positive and negative data point was also created. Different EC number levels were used to make easier and more challenging training points.

ML Model, func-e

We designed a new ML model, func-e, with a binary class classifier with 13 regression heads, with five enzyme features: enzyme length, mass, polarity, temperature, and the total polar surface area, and four features each from the substrates and the products: molecular weight (MolWt), total polar surface area (TPSA), calculated solubility proxy (MolLogP), and the maximum and minimum Gasteiger partial charges (MaxPartialCharge and MinPartialCharge). The model had a self-attention mechanism on the substrate (uniMol), product (uniMol), reaction (rxnFP), and enzyme (esm) heads. This was concatenated with a bidirectional cross attention head (eight heads, 128 embedding size) between the substrate and the enzyme, and the reaction and the enzyme. The attention was added to the unaltered inputs for substrate (N=768) and enzyme (N=1280), before being fed through a feedforward neural network of three layers, 1024, 512, and 256. The loss was calculated as the mean of activity (binary cross entropy loss), and the masked mean-squared error (MSE) of the regression features. Dropout layers were used (0.2), with a 1,000 batch size, learning rate of 0.01 and with early stopping for a maximum of 30 epochs. An ensemble was created across five models with the mean prediction and variance recorded across each model. The ensemble was selected across four EC classes with different inferred negatives selected for each, and the final models were trained using the "easy" training set, for further details see SI Section 1. The reaction used as input was the smiles string:
CCCCC(COC(C1=CC=CC=C1C(OCC(CCCC)CC)=O)=O)CC>>CCCCC(COC(C2=CC=CC=C2C(O)=O)=O)CC for DEHP and O=P(OC1=CC=CC=C1)(OC2=CC=CC=C2)OC3=CC=CC=C3>>O=P(O)(OC4=CC=CC=C4)OC5=CC=CC=C5 for TPP.

Structural pipeline

Final candidate selections for experimental validations were performed after manual evaluation of all filtered enzymes using a structural pipeline (https://github.com/moragroup/StructureZyme). In short, enzymes were first assigned catalytic residues using our recently developed catalytic residue prediction model, Squidly.[36] Structures were then co-folded with the predicted substrate (TPP or DEHP) using Chai[42] and Boltz2[52]. If Alpha Fold 2 (AF2)[61] structures were in the AF2 database, then these were also collected. The physics-based tool autodock Vina[62] was used to then re-dock the substrate in either the AF2 or Chai structure. Metrics from each of the structural tools were extracted. Additionally, for each protein, pairwise superimposition over the predicted structures was performed, and features were extracted using PLIP[63]. The distance between the ligand and the nucleophile was calculated using the docked structures in Python. Enzymes were also manually visualized and selected for further validation.

Annotation filtering steps based on UniProt

Initially, we aimed to keep our filtering to a minimum, filtering enzymes for origin (bacteria and archaea), length (<300 AA), simple structures (monomers/homodimers), and for the hydrolysis enzymes, no cofactors (i.e. ester cleavage by a catalytic triad). In the second step, we reduced the constraints of no cofactors and instead added an extra constraint of expression in the cytoplasm.

CREEP Model baseline

We trained a new CREEP model, using all available (no held-out) reaction-protein pairwise data, as given by the reaction2EC.csv and protein2EC.csv files in the CREEP repository. Afterward, we performed retrieval across all the SwissProt enzymes compiled in CREEP (protein2EC.csv) by searching based on the first step in the DEHP degradation. The top 1000 retrieved enzyme sequences (not grouped by EC number) were then filtered to include only short (<350 AA) archaeal or bacterial sequences, that were not annotated to require a cofactor and formed either a homodimer or monomer leaving 25 sequences. Of these, we removed EC numbers outside the class of hydrolases, selecting the following EC classes, 3.1.1.3, 3.1.1.101, 3.1.1.74, 3.5.2.6, 3.1.1.-, 3.1.1.3, 3.1.2.2, 3.3.2.10, and 3.1.1.76.

Similarity selection baseline

In addition to enzymes selected based on machine learning, we sought to benchmark these compared with a similarity-based approach, searching the reviewed protein space in UniProt. We tested the baseline substrate similarity approach by searching for proteins that use the most similar substrates to a target reaction, using the Tanimoto similarity using Morgan Fingerprints with parameters from Rdkit (radius=2, fpSize=2048). For TPP and DEHP, we filter based on sequence length (<300 AA), origin (bacteria or archaea), cofactors (none), and oligomeric state filter (homodimer or monomer). For TPP, this reduced the number of enzymes to 50. To remove redundancy, mmseqs2 was used to cluster sequences at 90% identity. From the 18 clusters, the centroid sequence from each cluster was selected. For DEHP, the initial filter left 12 enzymes to test. For the second set of filters, we also selected for cytosolic expression as some of the secreted enzymes may have disulfide bonds.

Computational oligo assembly method (Oligopoolio)

Oligo-sequences are short sequences (<300 nt) that are generally less expensive per base pair compared to ordering full-length genes. We designed a computational approach using polymerase cycling assembly (PCA) to assemble full gene constructs from short oligos. The tool includes two computational choices, first, a swarm-based optimization algorithm that aims to optimize on

469 multiple constraints using Pyswarms (v. 1.3.0), second a deterministic method that iteratively
470 selects the optimal overhangs. The constraints are, melting temperature of 62 °C for overhangs, DG
471 of homodimer > –3 kcal mol^{-1}, not starting or terminating on a triplet (e.g. AAA), and having a minimum
472 Hamming distance of eight to other overlaps. These are optimized by altering the cut sites along a
473 gene, with the oligo lengths between 80 and 130 nts and the overlap or primer length between 18 and
474 27 nts. For the stochastic approach, as there are no guarantees as to reaching the global optimum,
475 we perform the swarm five times and retain the oligo cut sites from the best cost. Each of these oligos
476 is then iteratively flipped, starting with a reverse strand and ending either on a forward strand. These
477 sequences are then collectively evaluated for overlaps with other oligos in the pool, and any minimal
478 Hamming distances are reported to the user, see https://github.com/ArianeMora/oligopoolio for
479 details.
480
481 **Experimental oligo assembly method (Oligopoolio)**
482 Oligopools (opools) were resuspended according to the suppliers' recommendations and
483 assembled via a two-step PCR protocol. The initial PCR was aimed at receiving full-length genes with
484 annealing temperatures ($T_{annealing}$ = 61.8 °C) based on the overlap regions, while the second one
485 served as a clean-up and amplification protocol (see SI Section 3 for details). After clean-up, these
486 gene inserts were then cloned into a pET22b(+)-vector between the *Nde*I and *Xho*I restriction sites
487 using Gibson protocol in a single sample.[64] Freshly transformed competent *E. coli* (BL21(DE3)) cells
488 were plated on Luria broth (LB) agar plates containing 100 µg/mL carbenicillin, and stationary
489 incubated at 37 °C overnight. Single colonies were picked into a 96-deep well plate containing LB_{Amp}
490 (500 µL) to prepare starter cultures and incubated at 220 rpm for 12–16 hours at 37 °C. To confirm
491 fully assembled sequences of desired genes, LevSeq[34] was implemented, and cells were stored in
492 glycerol stocks (25% glycerol) at -80 °C until used further. 96-well plates were re-arrayed to only
493 contain desired genes, and starter cultures were prepared. Final variants were sequence-verified
494 using Laragen Inc. (Culver City, CA), the DNA and protein sequences can be found in SI Section 7.
495
496 **Screening in 96-well plate format**
497 LB_{Amp} starting cultures (50 µL) were used to inoculate TB_{Amp} pre-expression cultures (950 µL). These
498 were then incubated at 37 °C for 2.5 hours at 220 rpm. After cooling on ice, protein expression was
499 induced by the addition of β-D-1-thiogalactopyranoside (IPTG, 0.5 mM) and incubated at 22 °C for 16
500 – 21 h at 220 rpm. After harvesting, cells were frozen at -20 °C at least overnight before further use.
501 Chemical lysis using 10 mM KP_i buffer (pH = 8) was performed, and clarified cell lysate was used to
502 perform reactions in 96-well plates (for details see SI). Each well contained 10 mM KP_i buffer (pH =
503 8, 690 µL) and the respective substrate (10 µL of 1 M DEHP stock in DMSO, $[DEHP]_{final}$ = 10 mM; 10 µL
504 of 100 mM TPP stock in DMSO, $[TPP]_{final}$ = 1 mM) and clarified cell lysate (300 µL). Reactions were
505 performed under aerobic conditions, sealed and shaken at 220 rpm for 4 hours at 22 °C. Reactions
506 were quenched by the addition of acetonitrile and analyzed by HPLC-(MS) analysis.
507
508 **Experimental validation**
509 To validate identified hits from 96-well plate screens, glycerol stocks were streaked on LB_{Carb} agar
510 plates and single colonies were picked to inoculate LB_{Amp} starter cultures (6 mL), which were used
511 to inoculate TB_{Amp} pre-expression cultures (addition of 1 mL starter culture into 50 mL TB_{Amp}). These
512 were grown to an OD_{500} = 0.8 – 1.0, before cooling for 30 min on ice, and protein overexpression was
513 induced with IPTG (0.5 mM). Cells were incubated at 22 °C for 16 – 21 h, then harvested by
514 centrifugation, frozen overnight at – 20 °C, and chemical lysis was performed after thawing at room
515 temperature by adding four times the volume of the wet-pellet weight (see SI Section 2 for
516 procedures and SI Section 4 and 5 for results). Lysis buffer contained 10 mM KP_i buffer with 1 mg/mL

517 lysozyme, 0.07 mg mL^{-1} DNAse, and 1μL mL^{-1} 2 M $MgCl_2$. Clarified lysate was used to perform
518 reactions in technical triplicate in 1.5-mL screw-cap glass vials. Each vial contained 10 mM KP_i buffer
519 (pH = 8, 690 μL) and the respective substrate (10 μL of 1 M DEHP stock in DMSO, $[DEHP]_{final}$ = 10 mM;
520 10 μL of 100 mM TPP stock in DMSO, $[TPP]_{final}$ = 1 mM) and clarified lysate (300 μL). Protein
521 concentrations were determined using the Pierce BCA Protein Assay Kit (Thermo Fisher Scientific)
522 according to the manufacturer's recommendations.
523

524 **References**

525 1. Buller, R. *et al.* From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**,
526 eadh8615 (2023).
527 2. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc
528 collection of Pathway/Genome Databases. *Nucleic Acids Research* **42**, D459–D471 (2014).
529 3. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for
530 taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research* **51**, D587–D592
531 (2023).
532 4. Gao, J., Ellis, L. B. M. & Wackett, L. P. The University of Minnesota
533 Biocatalysis/Biodegradation Database: improving public access. *Nucleic Acids Research* **38**,
534 D488–D491 (2010).
535 5. Malwe, A. S., Longwani, U. & Sharma, V. K. XenoBug: machine learning-based tool to
536 predict pollutant-degrading enzymes from environmental metagenomes. *NAR Genomics and*
537 *Bioinformatics* **7**, lqaf037 (2025).
538 6. Shi, Z. *et al.* REME: an integrated platform for reaction enzyme mining and evaluation.
539 *Nucleic Acids Research* **52**, W299–W305 (2024).
540 7. Chang, A. *et al.* BRENDA, the ELIXIR core data resource in 2021: new developments and
541 updates. *Nucleic Acids Research* **49**, D498–D508 (2021).
542 8. Consortium, T. U. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids*
543 *Research* **51**, D523–D531 (2022).
544 9. Moorhoff, F. *et al.* Machine Learning-Driven Enzyme Mining: Opportunities, Challenges,
545 and Future Perspectives. *ACS Catal.* **16**, 12–30 (2026).

10. Liu, Y. *et al.* EnzymeCAGE: A Geometric Foundation Model for Enzyme Retrieval with Evolutionary Insights. 2024.12.15.628585 Preprint at https://doi.org/10.1101/2024.12.15.628585 (2024).
11. Yu, T. *et al.* Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).
12. Mikhael, P. G., Chinn, I. & Barzilay, R. CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes. Preprint at https://doi.org/10.48550/arXiv.2402.06748 (2024).
13. Yang, J. *et al.* CARE: a Benchmark Suite for the Classification and Retrieval of Enzymes. Preprint at https://doi.org/10.48550/arXiv.2406.15669 (2025).
14. Johnson, S. R. *et al.* Computational scoring and experimental evaluation of enzymes generated by neural networks. *Nat Biotechnol* **43**, 396–405 (2025).
15. Merlicek, L. P. *et al.* AI.zymes: A Modular Platform for Evolutionary Enzyme Design. *Angewandte Chemie International Edition* **64**, e202507031.
16. Freschlin, C. R., Yang, K. K. & Romero, P. A. Scalable and cost-efficient custom gene library assembly from oligopools. 2025.03.22.644747 Preprint at https://doi.org/10.1101/2025.03.22.644747 (2025).
17. Qian, J. *et al.* Accelerating protein design by scaling experimental characterization. 2025.08.05.668824 Preprint at https://doi.org/10.1101/2025.08.05.668824 (2025).
18. Pryor, J. M. *et al.* Enabling one-pot Golden Gate assemblies of unprecedented complexity using data-optimized assembly design. *PLOS ONE* **15**, e0238592 (2020).
19. Luna-Cerralbo, D. *et al.* CertPrime: a new oligonucleotide design tool for gene synthesis. 2025.04.29.650686 Preprint at https://doi.org/10.1101/2025.04.29.650686 (2025).

568 20. Kosuri, S. *et al.* Scalable gene synthesis by selective amplification of DNA pools from high-
569 fidelity microchips. *Nat Biotechnol* **28**, 1295–1299 (2010).
570 21. Sidore, A. M., Plesa, C., Samson, J. A., Lubock, N. B. & Kosuri, S. DropSynth 2.0: high-
571 fidelity multiplexed gene synthesis in emulsions. *Nucleic Acids Res* **48**, e95 (2020).
572 22. Anaconda Software Distribution. Anaconda Inc. (2020).
573 23. Richman, T., Arnold, E. & Williams, A. J. Curation of a list of chemicals in biosolids from
574 EPA National Sewage Sludge Surveys & Biennial Review Reports. *Sci Data* **9**, 180 (2022).
575 24. Payne-Sturges, D., De Saram, S. & Cory-Slechta, D. A. Cumulative Risk Evaluation of
576 Phthalates Under TSCA. *Environ. Sci. Technol.* **57**, 6403–6414 (2023).
577 25. Sousa, J. C. G., Ribeiro, A. R., Barbosa, M. O., Pereira, M. F. R. & Silva, A. M. T. A review
578 on environmental monitoring of water organic pollutants identified by EU guidelines.
579 *Journal of Hazardous Materials* **344**, 146–162 (2018).
580 26. Mabanglo, M. F., Xiang, D. F., Bigley, A. N. & Raushel, F. M. Structure of a Novel
581 Phosphotriesterase from Sphingobium sp. TCM1: A Familiar Binuclear Metal Center
582 Embedded in a Seven-Bladed β-Propeller Protein Fold. *Biochemistry* **55**, 3963–3974 (2016).
583 27. van der Veen, I. & de Boer, J. Phosphorus flame retardants: Properties, production,
584 environmental occurrence, toxicity and analysis. *Chemosphere* **88**, 1119–1153 (2012).
585 28. Yu, J., Zou, H., Zhang, T. & Zhang, J. Nuclear Receptors-Mediated Endocrine Disrupting
586 Effects of Non-Phthalate Plasticizers: A Review. *J. Agric. Food Chem.* **71**, 16902–16918
587 (2023).
588 29. Li, Y. *et al.* Adverse Effects of Aryl Organophosphate Esters on Reproduction. *Environ. Sci.*
589 *Technol.* **59**, 10122–10135 (2025).

590 30. Xie, Z. *et al.* Organophosphate ester pollution in the oceans. *Nat Rev Earth Environ* **3**, 309–
591 322 (2022).
592 31. Fan, S. *et al.* Biodegradation of phthalic acid esters (PAEs) by *Bacillus* sp. LUNF1 and
593 characterization of a novel hydrolase capable of catalyzing PAEs. *Environmental*
594 *Technology & Innovation* **32**, 103269 (2023).
595 32. Engqvist, M. K. M. Correlating enzyme annotations with a large set of microbial growth
596 temperatures reveals metabolic adaptations to growth at diverse temperatures. *BMC*
597 *Microbiology* **18**, 177 (2018).
598 33. Sanderson, T., Bileschi, M. L., Belanger, D. & Colwell, L. J. ProteInfer, deep neural
599 networks for protein functional inference. *eLife* **12**, e80942 (2023).
600 34. Long, Y. *et al.* LevSeq: Rapid Generation of Sequence-Function Data for Directed Evolution
601 and Machine Learning. *ACS Synth. Biol.* **14**, 230–238 (2025).
602 35. Australian Industrial Chemicals Introduction Scheme (AICIS). Triphenyl phosphate and
603 diphenyl phosphate - Evaluation statement - 26 June 2023.
604 https://www.industrialchemicals.gov.au/sites/default/files/2023-06/EVA00108%20-
605 %20Evaluation%20Statement%20-%2026%20June%202023.pdf (2023).
606 36. Rieger, W. J., Boden, M., Arnold, F. & Mora, A. Squidly: Enzyme Catalytic Residue
607 Prediction Harnessing a Biology-Informed Contrastive Learning Framework. *eLife* **14**,
608 (2025).
609 37. Tian, S., Yesselman, J. D., Cordero, P. & Das, R. Primerize: automated primer assembly for
610 transcribing non-coding RNA domains. *Nucleic Acids Res* **43**, W522–W526 (2015).

611 38. Patsch, D., Eichenberger, M., Voss, M., Bornscheuer, U. T. & Buller, R. M. LibGENiE – A
612 bioinformatic pipeline for the design of information-enriched enzyme libraries.
613 *Computational and Structural Biotechnology Journal* **21**, 4488–4496 (2023).
614 39. Edinburgh-Genome-Foundry/DnaWeaver: A route planner for DNA assembly.
615 https://github.com/Edinburgh-Genome-Foundry/DnaWeaver.
616 40. De Simone, G. *et al.* A snapshot of a transition state analogue of a novel thermophilic
617 esterase belonging to the subfamily of mammalian hormone-sensitive lipase. *J Mol Biol* **303**,
618 761–771 (2000).
619 41. Zhu, Y. B. UniProt Entry I3NWL3. *UniProt*
620 https://www.uniprot.org/uniprotkb/I3NWL3/entry (2025).
621 42. Chai Discovery. Chai-1: Decoding the molecular interactions of life. *bioRxiv*
622 https://doi.org/10.1101/2024.10.10.615955 (2024) doi:10.1101/2024.10.10.615955.
623 43. Wick, R. Porechop, https://github.com/rrwick/Porechop. (2024).
624 44. Kolmogorov, M. *et al.* metaFlye: scalable long-read metagenome assembly using repeat
625 graphs. *Nat Methods* **17**, 1103–1110 (2020).
626 45. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069
627 (2014).
628 46. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language
629 model. *Science* **379**, 1123–1130 (2023).
630 47. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. *Science*
631 **387**, 850–858 (2025).
632 48. Schwaller, P. *et al.* Mapping the space of chemical reactions using attention-based neural
633 networks. *Nat Mach Intell* **3**, 144–152 (2021).

49. Yüksel, A., Ulusoy, E., Ünlü, A. & Doğan, T. SELFormer: Molecular Representation Learning via SELFIES Language Models. Preprint at https://doi.org/10.48550/arXiv.2304.04662 (2023).
50. Zhou, G. *et al.* Uni-Mol: A Universal 3D Molecular Representation Learning Framework. in (The Eleventh International Conference on Learning Representations, 2022).
51. Ahmad, W., Simon, E., Chithrananda, S., Grand, G. & Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. Preprint at https://doi.org/10.48550/arXiv.2209.01712 (2022).
52. Cho, Y., Pacesa, M., Zhang, Z., Correia, B. E. & Ovchinnikov, S. Boltzdesign1: Inverting All-Atom Structure Prediction Model for Generalized Biomolecular Binder Design. 2025.04.06.647261 Preprint at https://doi.org/10.1101/2025.04.06.647261 (2025).
53. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**, 59–60 (2015).
54. van Kempen, M. *et al.* Fast and accurate protein structure search with Foldseek. *Nat Biotechnol* **42**, 243–246 (2024).
55. Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**, 1026–1028 (2017).
56. Schmid, H. StructureZyme, https://github.com/moragroup/StructureZyme.
57. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* **7**, 539 (2011).
58. Price, M. FastTree 2: Approximately-Maximum-Likelihood Trees for Large Alignments, https://github.com/morgannprice/fasttree. (2025).

656 59. Heid, E., Probst, D., Green, W. H. & Madsen, G. K. H. EnzymeMap: Curation, validation
657 and data-driven prediction of enzymatic reactions. Preprint at
658 https://doi.org/10.26434/chemrxiv-2023-jzw9w (2023).
659 60. Dai, D. *et al.* Long-read metagenomic sequencing reveals shifts in associations of antibiotic
660 resistance genes with mobile genetic elements from sewage to activated sludge. *Microbiome*
661 **10**, 20 (2022).
662 61. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**,
663 583–589 (2021).
664 62. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: New
665 Docking Methods, Expanded Force Field, and Python Bindings. *J. Chem. Inf. Model.* **61**,
666 3891–3898 (2021).
667 63. Adasme, M. F. *et al.* PLIP 2021: expanding the scope of the protein–ligand interaction
668 profiler to DNA and RNA. *Nucleic Acids Research* **49**, W530–W534 (2021).
669 64. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases.
670 *Nat Methods* **6**, 343–345 (2009).