Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Memorization in reaction-conditioned generative enzyme design: folding confidence, database identity, and docking all fail to detect it

Kinsley Torres

kinsley.torres@terramtech.com

Terram Technologies

Article

Keywords: Enzyme design, Sequence-to-sequence, Large language model, Bioremediation, Memorization, Validation, Molecular docking, Protein language model

Posted Date: September 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11039691/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Memorization in reaction-conditioned generative enzyme design: folding confidence, database identity, and docking all fail to detect it

Kinsley Torres*

Department of Research and Development, Terram Tech,

3200 E Camelback Rd, 85018, Arizona, USA

Abstract

Many xenobiotics—halogenated flame retardants, phthalates, organochlorine pesticides—have no documented enzyme capable of efficient degradation. Existing generative tools centre on backbone generation with expert-specified active sites rather than reasoning from pollutant chemistry to a catalytic enzyme. We asked which validation metrics can support claims about generated enzymes. XenoFold couples a frozen ChemBERTa reaction encoder to a trainable ProGen2-small decoder via cross-attention, trained separately for 15 enzyme classes. A comparison against an unconditioned baseline is reported but not interpreted: a post-hoc code audit found the two arms differ in nine respects. BLAST against each model's own training set placed generated sequences at 92.7–99.8% mean identity, with 40% exact copies of training entries, and swapping the conditioning reaction for an unrelated one changed folding confidence non-significantly (82.0 ± 12.2 vs. 87.0 ± 4.5, $p = 0.060$). Generating for substrates with no documented degrading enzyme (PFOA, p,p'-DDE) produced the same pattern as an in-training control (DDT): 21 of 40 sequences were exact copies, regardless of substrate novelty. AutoDock Vina scores were uninformative: no comparison survived multiple-testing correction, and docking-derived K_M estimates deviated from measured values by up to 15,606-fold with an inverted ranking, prompting their removal. We present XenoFold as a case of memorization that pLDDT, *nr*-BLAST, and docking each independently failed to detect.

Scientific Contribution. We show that a reaction-conditioned generative enzyme design pipeline reproduces its training set almost verbatim while three standard in-silico validation

*Corresponding author: kinsley.torres@terramtech.com

metrics—predicted folding confidence, identity against a general sequence database, and molecular docking—each independently fail to detect this. We further show that docking-derived kinetic estimates deviate from measured values by up to 15,606-fold and invert the known ranking of two characterized peroxidases, with no docking comparison surviving multiple-testing correction across twelve tests. We identify three inexpensive controls—training-set identity with a coverage floor, mismatched-conditioning generation, and novel-substrate generation—that detect this failure mode and are not part of current standard practice.

Keywords: Enzyme design; Sequence-to-sequence; Large language model; Bioremediation; Memorization; Validation; Molecular docking; Protein language model

1 Introduction

Xenobiotics, synthetic pollutants such as microplastics, pesticides, industrial solvents, and flame retardants, persist in the environment due to their chemical stability and resistance to natural degradation. Human exposure is substantial and well documented: the World Health Organization estimates approximately 385 million cases of unintentional acute pesticide poisoning annually (Ádam et al., 2024), and untreated sewage discharge contributes to over fifty distinct disease outcomes and roughly half of global child mortality attributable to poor water quality (Lin et al., 2022). Chronic environmental exposure is further associated with cardiovascular and metabolic disease (Münzel et al., 2021), alongside a broader range of documented health effects across xenobiotic classes (Parkinson et al., 2013).

Enzymatic xenobiotic degradation has emerged as one of the most effective and sustainable remediation strategies (Rathore et al., 2022), but naturally occurring enzymes frequently exhibit narrow substrate specificity against synthetic xenobiotics with no evolutionary precedent, leaving many industrially significant pollutants, halogenated flame retardants, phthalate plasticizers, and organochlorine pesticide metabolites, without any documented enzyme capable of their efficient degradation.

RFdiffusion (Watson et al., 2023) performs strongly at de novo backbone generation but centres on expert-specified active-site design rather than modelling degradation as a function of a defined reaction. Conditional protein language models are the closer precedent: ProGen (Madani et al., 2023) demonstrated tag-conditioned generation of experimentally validated enzymes, and ZymCTRL (Munsamy et al., 2022) conditions on EC number. Reaction- and substrate-conditioned approaches include CLIPZyme (Mikhael et al., 2024) and substrate-guided enzyme design (Hua et al., 2024); CLEAN (Yu et al., 2023) is the reference method for EC assignment; and retrieval tools such as Selenzyme (Carbonell et al., 2018) answer the same practical question by database search. XenoFold's position is narrow: a frozen chemical encoder cross-attending into a trainable protein decoder, one model per class. We did not benchmark against any of these systems and make no claim of superiority over them.

This study presents XenoFold, a chemistry-conditioned generative framework for proposing candidate enzymes for xenobiotic degradation. XenoFold conditions an autoregressive protein decoder on a chemical reaction representation via cross-attention, trained separately for 15 enzyme classes relevant to xenobiotic breakdown. Our contribution is twofold and we state both

plainly: a working reaction-conditioned generation architecture, and a quantitative account of which in-silico validation metrics can and cannot support claims about generated enzymes. The second of these produced the more durable result.

A note on scope. An earlier framing described the design target as "extracellular, cofactor-free." That is inconsistent with the four reactions validated here: the two peroxidases are heme-dependent and require H_2O_2, and the 4-chlorobenzoyl-CoA reaction operates on a coenzyme A thioester. Extracellular localization was not tested either, as no signal-peptide analysis was performed. Both remain design goals of the pipeline, not demonstrated properties of its outputs.

2 Methods

2.1 System methods

2.1.1 Sequence-to-sequence encoder-decoder model

Relational target reaction to amino acid sequence data collection Training data pairing enzyme sequences with their catalytic reactions were generated with a custom `DataCollector` framework per enzyme class: UniProt sequence retrieval (EC-number/keyword queries, minimum 50 aa, deduplicated), staged SMILES resolution (local dictionary → CIRpy → PubChem PUG REST → NCI CIR), and RDKit SMARTS-based reaction product prediction. Each predicted reaction was paired with every retrieved sequence for its class to form the final per-class training set (Supplementary Table S1; full parameters in Supplementary Methods S1.1).

This Cartesian pairing matters for interpretation: since each class contributes only a few distinct reaction strings and a separate model is trained per class, the reaction input carries little information beyond the class identity already fixed by model selection. The system is therefore better described as *enzyme-class-conditioned* generation with a reaction-derived input than as generation conditioned on arbitrary substrate chemistry. Demonstrating genuine within-class substrate sensitivity was not performed (Section 4).

Hybrid chemical-conditioned protein decoder architecture Enzyme sequence generation used a custom hybrid architecture, `ImprovedHybridChemProteinModel`, conditioning autoregressive protein generation on a chemical reaction representation via cross-attention rather than a conventional encoder-decoder transformer. One model was trained independently per enzyme class on its corresponding dataset (Section 2.1.1).

Chemical encoder: pretrained ChemBERTa (Chithrananda et al., 2020) (`seyonec/ChemBERTa-zinc-base`) as a fixed feature extractor, all parameters frozen and run in evaluation mode.

Protein decoder: ProGen2-small (`hugohrban/progen2-small`; (Nijkamp et al., 2023)), a causal language model, trainable throughout fine-tuning.

Cross-attention conditioning: reaction context was injected via two sequential multi-head cross-attention blocks. ChemBERTa's and ProGen2's final hidden states were projected into a shared 1024-dimensional space (linear–GELU–LayerNorm); protein representations attended to the projected chemical representations (protein as query, chemical as key/value; 8 heads, dropout 0.1), added residually and layer-normalized, then passed through a second identically-structured block. The result was projected back to ProGen2's native hidden dimension and decoded using ProGen2's own pretrained language-modeling head. Full layer dimensions: Supplementary Methods S1.2.

Reaction input: substrate(s)/product(s) format (Section 2.1.1), with source-format tokens replaced by ChemBERTa's native separator prior to tokenization, tokenized to a fixed maximum length of 256.

Sequence filtering and tokenization: training sequences filtered to 100–507 amino acids, with an explicit `<eos>` appended so the decoder could learn termination rather than rely on a length cutoff.

Training objective: standard next-token autoregressive cross-entropy over the ProGen2 vocabulary, with end-of-sequence tokens upweighted 2× to encourage reliable termination.

Generation: all conditioned batches were decoded by a single routine with identical arguments — temperature 1.1, no nucleus or top-k truncation, minimum length 150 and maximum length 800 tokens, EOS-terminated, one sequence per call and 20 calls per batch. Decoding also applied per-step logit steering toward target hydrophobic and charged composition, an EOS boost increasing with generation progress, and a post-generation quality filter that discarded sequences below a threshold and rank-ordered the remainder; these are specified in full in Supplementary Methods S1.3. No random seed was set. The unconditioned baseline was generated by separate code with different parameters and no steering, filtering or ranking (Supplementary Methods S1.4).

Training procedure: PyTorch/AdamW restricted to trainable parameters (cross-attention, projection layers, ProGen2 weights; ChemBERTa excluded), learning rate 5e-5, batch size 4, gradient-norm clipping at 1.0. Each of the 15 per-class models was trained for up to 15 epochs

subject to a 23-hour budget on a single Nvidia A100 GPU via Google Colab, with checkpoints saved each epoch.

2.2 Validation methods

2.2.1 Enzyme validation methods

Structure prediction and folding confidence via ESMFold Each candidate was structurally validated using ESMFold (Lin et al., 2023) via the BioLM hosted inference API, returning per-residue coordinates and a pLDDT score (0–100; ∼90+ well-defined, below 50 disordered). Average pLDDT was used as a summary folding-confidence metric. We are deliberately careful about what this means: pLDDT reflects the structure predictor's *own confidence*, not thermodynamic stability, expressibility, solubility, or catalytic activity, and is systematically higher for sequences resembling well-represented families. High pLDDT is read here as confidently foldable in silico, not laboratory-stable or active — a distinction that matters throughout, since pLDDT carries this study's principal quantitative result.

Active site prediction via P2Rank The putative catalytic site on each structure was identified using P2Rank (Krivák and Hoksza, 2018), which scores surface points by predicted ligandability and clusters them into ranked pockets. Only the top-scoring pocket per structure was retained, with no validation against known catalytic residues (a limitation addressed in Section 4); its coordinates defined the Vina search space.

Molecular docking via AutoDock Vina Binding affinity was estimated using AutoDock Vina (Trott and Olson, 2010; Eberhardt et al., 2021) via MGLTools. Receptors were converted to PDBQT via `prepare_receptor4.py`; ligands via RDKit (ETKDG, UFF, Gasteiger charges) and Meeko. Docking used Vina 1.2.7 in a fixed 20^3 Å box centred on the top P2Rank pocket, exhaustiveness 8, 9 output modes, one run per sequence.

Estimation of ligand efficiency Ligand efficiency (LE) was calculated as:

$$LE = \frac{-\Delta G}{N}$$

where ΔG is the Gibbs free energy of binding (kcal/mol) predicted by Vina, and N is the number of heavy (non-hydrogen) atoms in the substrate ligand. LE normalizes binding affinity

by ligand size.

A property of this definition governs how LE is used below. Within a single reaction every sequence is docked against the same substrate, so N is constant and LE is a strictly monotone linear transform of the Vina score. Rank-based tests on LE and on Vina must therefore return identical p-values within a reaction, and reporting both as separate evidence double-counts a single measurement. Accordingly, we report LE only where it is informative — comparisons *across* substrates of differing size (Supplementary Tables S10–S13) — and report Vina score alone for all within-reaction comparisons (Tables 2 and 3).

Auditing this relationship against the per-sequence data found two isolated errors, now corrected, affecting no Vina-based result. One halide hydrolase sequence's tabulated LE (0.283) was inconsistent with its own Vina score at the reaction's fixed $N = 57$ (expected 0.125); the other 19 of 20 satisfied $LE = -\text{Vina}/N$ exactly, giving a corrected batch mean of 0.117 ± 0.025 ($n = 20$), not the previously reported 0.125 ± 0.045. For dye-decolorizing peroxidase, LE was recorded for only 19 of 20 sequences; recovering the missing value from its Vina score ($N = 38$) gives 0.237 ± 0.038 ($n = 20$), close to the previous 0.236 ± 0.039 ($n = 19$). Both are corrected in Supplementary Tables S2, S5, and S9.

Sequence homology validation via BLAST Generated sequences were queried with blastp (Altschul et al., 1990) against two databases: the NCBI non-redundant protein database (nr), recording percent identity and the annotated function of the nearest hit; and, separately, each per-class *training set* directly, recording percent identity, query coverage, and whether the hit was an exact match, at a 50% coverage floor. The training-set search is decisive where nr is not: because training sequences were drawn from UniProt and UniProt entries are contained in nr, a generated sequence's nearest nr hit is frequently one of its own training examples, so nr identity alone cannot establish novelty. We explicitly do not adopt the convention, used in an earlier version of this work, of treating any identity below 100% as evidence of novel design (Section 3.1).

2.2.2 Benchmarking and baseline comparisons

Comparison to natural/literature enzyme validation scores For each reaction, a literature search identified natural enzymes with documented catalytic activity on the same or closely analogous substrate, processed through the identical fold-dock-validate pipeline (Section 2.2.1) so differences reflect the sequences rather than methodology. Reference sets ranged from one to five enzymes; given negligible power at these sizes, comparisons are reported via percentile positioning, with Mann-Whitney U added where $n \geq 3$, using Vina score and pLDDT (LE and docking-derived kinetics excluded as above).

Ablation against unconditioned sequence generation A baseline was generated using base ProGen2-small with no reaction or class conditioning, in place of the trained architecture (Section 2.1.1), evaluated through the identical pipeline for each of the four reactions.

What this comparison can and cannot establish. The two arms were generated by different code paths, and an audit of those paths performed after the analysis found that they differ in nine respects rather than the two (fine-tuning and conditioning) originally intended. The differences are enumerated in Supplementary Methods S1.4. Four of them act directly on the metric carrying this study's principal quantitative result: the baseline was capped at 250 residues against the conditioned arm's 800, applied no composition steering, used a different EOS schedule, and was taken as raw output whereas conditioned batches were filtered to a quality threshold and rank-ordered. We therefore do not treat the base-versus-trained pLDDT comparison as interpretable evidence about fine-tuning or conditioning, and report it in Table 2 for completeness only. The comparison that does hold generation constant is the mismatched-reaction control (Section 3.3.1), in which both batches were produced by the identical conditioned path.

Unconditioned batches ($n = 6$–8) were smaller than conditioned batches ($n = 19$–20). This was not a design choice: the baseline loop attempted 20 generations per reaction and discarded any attempt whose sequence ESMFold could not fold, a failure the folding code attributes to highly repetitive or low-complexity input. The surviving $n = 6$–8 are therefore a filtered subset, and the discarded sequences were the most degenerate ones. Baseline pLDDT means reported here are consequently biased upward relative to the unfiltered base-model distribution. Shapiro-Wilk normality testing selected between Welch's t-test (Cohen's d) and Mann-Whitney U (rank-biserial) per metric, though Shapiro-Wilk has little power at $n = 6$–8, so this selection

is weakly grounded. p-values are reported raw and Holm-Bonferroni-corrected across the four reactions per metric; full per-reaction results are in Section 3.3.

3 Results

3.1 Enzyme homology results

Sequence identity to nr, reported in an earlier version of this work as evidence of novelty, is a weaker measurement than it appears: training sequences were drawn from UniProt, which is contained in nr, so a generated sequence's nearest nr hit is frequently its own training example. We therefore repeated the search with blastp against each per-class *training set* directly (50% coverage floor), which is decisive where nr is not.

Reaction / substrate	Model (epoch)	% id. to training (mean ± SD)	Median	Exact copies	≥95% id.
Dye-decolorizing peroxidase (validated)	DyP (2)	$97.3\% \pm 6.7\%$	99.7%	7/20	17/20
Ester hydrolase (validated)	Ester (12)	$92.7\% \pm 20.8\%$	99.9%	9/20	17/20
Lignin peroxidase (validated)	Lignin (2)	$96.7\% \pm 6.8\%$	99.2%	4/20	16/20
Halide hydrolase (validated)	Halide (8)	$99.8\% \pm 0.2\%$	100.0%	12/20	20/20
DDT (in-training control)	Halide (8)	$99.6\% \pm 1.0\%$	100.0%	14/20	20/20
PFOA (novel substrate)	Halide (8)	$94.3\% \pm 22.3\%$	99.8%	10/20	18/20
p,p'-DDE (novel substrate)	Halide (8)	$95.9\% \pm 10.7\%$	100.0%	11/20	17/20
BDE-209 (novel)†	DyP (2)	$61.5\% \pm 40.6\%$	76.0%	0/20	7/20

Table 1: Sequence identity to each model's own *training set* (blastp, 50% coverage floor). Top block: the four validated reactions (Sections 3.1), vs. previously reported nr identity of 88.6–96.6%. Bottom block: PFOA and p,p'-DDE have no documented natural degrading enzyme (Section 3.2); DDT is an in-training control. †BDE-209 used a different, less-trained checkpoint and is not a matched control (Section 4). Uniqueness and per-sequence data: Supplementary Tables S14–S15.

Across all four reactions, 32 of 80 generated sequences (40%) are exact copies of training entries and 70 of 80 (87.5%) are at least 95% identical. Training identity exceeds the previously reported nr identity by 4–9 points for three of four reactions (halide +6.5, DyP +8.7, lignin +4.2), confirming nr was not returning the true nearest neighbor. Ester hydrolase's training *mean* (92.7%) is pulled below its nr mean by two low-coverage outliers, but its *median* (99.9%) and 9/20 exact copies tell the same story as the other three.

We no longer describe these results as evidence of designed-enzyme novelty; the straightforward reading is memorization, reproduced with occasional minor perturbation, regardless of whether the mechanism is best described as retrieval or as emitting a stored pattern. This matches the scale of training exposure: 1,656 unique halide sequences against a Cartesian product of 55 reaction strings for up to 8 epochs means individual sequences were seen hundreds of times during fine-tuning.

Functional-category coherence to nr, unaffected by this correction, ranged from 35% (ester hydrolase, direct match) to 85% (DyP peroxidase); the lignin peroxidase model was an outlier, with 80% of sequences matching lipoyl synthase/LipB, a family mechanistically unrelated to lignin peroxidase, consistent with training-data contamination (Supplementary Figs. S1–S8).

3.2 Generation for substrates without known natural enzymes

The scenario XenoFold is designed for is a pollutant with no documented degrading enzyme, so the four validated reactions above — with characterized natural enzymes in the literature and, per Section 3.1, in the training data itself — cannot test the pipeline's stated purpose. We therefore generated additional batches ($n = 20$ each) conditioned on PFOA and p,p'-DDE, both absent from the halide model's 55 training reactions and with no known degrading enzyme, plus DDT as an in-training control, and BDE-209 using the DyP model. The bottom block of Table 1 reports the same training-set BLAST search applied to these batches.

For the halide hydrolase model, substrate novelty made no detectable difference. DDT produced 14/20 exact training copies; PFOA and p,p'-DDE, absent from the training reaction set entirely, produced 10/20 and 11/20 respectively — 35 of 60 sequences (58%) across the three conditions. The model reproduces its training distribution regardless of whether the requested substrate has a documented natural counterpart, the scenario this framework is intended to address.

BDE-209, generated with the DyP peroxidase model, departed further: zero exact copies, median identity 76.0%, and 5/20 sequences below the coverage floor entirely, folding confidently (pLDDT 83.8$\pm$3.2) with no relationship between identity and fold quality among the rest (Supplementary Table S16) — consistent with genuine departure rather than degradation. However, this condition used a different, less-trained checkpoint (2 epochs vs. halide's 8), and the DyP ablation baseline was already lower than halide's. A matched, in-training DyP control is needed before this can be read as evidence of extrapolation, and we did not run it.

3.3 Ablation study

The ablation study compared unconditioned generation from the base ProGen2-small language model against the full XenoFold pipeline, jointly conditioned on reaction SMILES and predicted enzyme class, across all four reactions examined. Both configurations were evaluated through the identical fold-dock-BLAST validation pipeline described in Section 2.2. Unconditioned batch

sizes (n = 6–8 per reaction) were smaller than conditioned batches (n = 19–20 per reaction) due to the computational cost of large-scale folding and docking for this baseline; this disparity is addressed in Section 4. For each metric, Shapiro-Wilk normality testing determined whether Welch's t-test (Cohen's d) or the Mann-Whitney U test (rank-biserial r) was used.

Reaction	pLDDT (base vs. trained)	pLDDT p	Holm p	Effect size	Vina p	Vina Holm p
Dye-decolorizing peroxidase	41.4 ± 6.9 vs. 85.3 ± 9.0	< 0.0001	0.0002^{***}	$r = 1.00$	0.089	0.267 (n.s.)
Ester hydrolase	41.7 ± 10.2 vs. 80.2 ± 14.7	0.0009	0.0009^{***}	$r = 0.92$	0.764	1.000 (n.s.)
Lignin peroxidase†	49.4 ± 11.7 vs. 85.8 ± 5.9	0.0001	0.0003^{***}	$d = 4.73$	0.0495	0.198 (n.s.)
Halide hydrolase	45.2 ± 13.8 vs. 87.0 ± 4.5	0.0002	0.0004^{***}	$d = 5.33$	0.956	1.000 (n.s.)

Table 2: Base ProGen2-small vs. the trained, reaction/EC-conditioned pipeline. **These measurements are reported for completeness and are not interpretable as evidence about fine-tuning or conditioning.** The two arms were produced by different generation code differing in nine respects, including a 250- vs. 800-residue length cap and a quality filter applied to the conditioned arm only (Supplementary Methods S1.4) (Section 4.1). Base n = 6–8 per reaction (of 20 attempted; the remainder failed to fold and were discarded, biasing the baseline mean upward); trained n = 19–20 per reaction. Holm-Bonferroni correction applied within each metric family across the four reactions. Ligand efficiency is omitted: within a reaction the substrate is fixed, so LE is a linear transform of the Vina score and carries no independent information. †The lignin peroxidase model is affected by diagnosed training-data contamination (Section 3.1); its results are reported for completeness but should not be treated as support for any general claim. n.s. = not significant at $\alpha = 0.05$. Supplementary Figs. S9–S15.

Folding confidence was higher for the trained pipeline than for base ProGen2-small in all four reactions, with all four results surviving Holm correction ($p \leq 0.0009$) and effect sizes from large to very large (d = 4.73–5.33 parametric; r = 0.92–1.00 non-parametric). Docking behaved entirely differently: no Vina comparison survived correction. The lignin peroxidase result, the only one significant at the raw threshold, moves from $p = 0.0495$ to Holm $p = 0.198$ — and it comes from the one model this study independently diagnoses as trained on contaminated data. With that model set aside, the docking result across the remaining three reactions is zero of three.

Two things follow. The effect is large and unambiguous: base ProGen2-small produces sequences at mean pLDDT 41–49, below the conventional disorder threshold of 50, while the trained pipeline reaches 80–87. That separation cannot, however, be attributed to fine-tuning or to conditioning: the two arms were generated by different code paths differing in nine respects (Section 2.2.2; Supplementary Methods S1.4), including a 250- versus 800-residue length cap and a quality filter applied to the conditioned arm only. We report the comparison for completeness and draw no inference from it. The evidence that bears on conditioning is the mismatched-reaction control (Section 3.3.1), where generation was held constant, together with the training-identity results of Section 3.1 (40% exact copies). What the docking comparison supports is

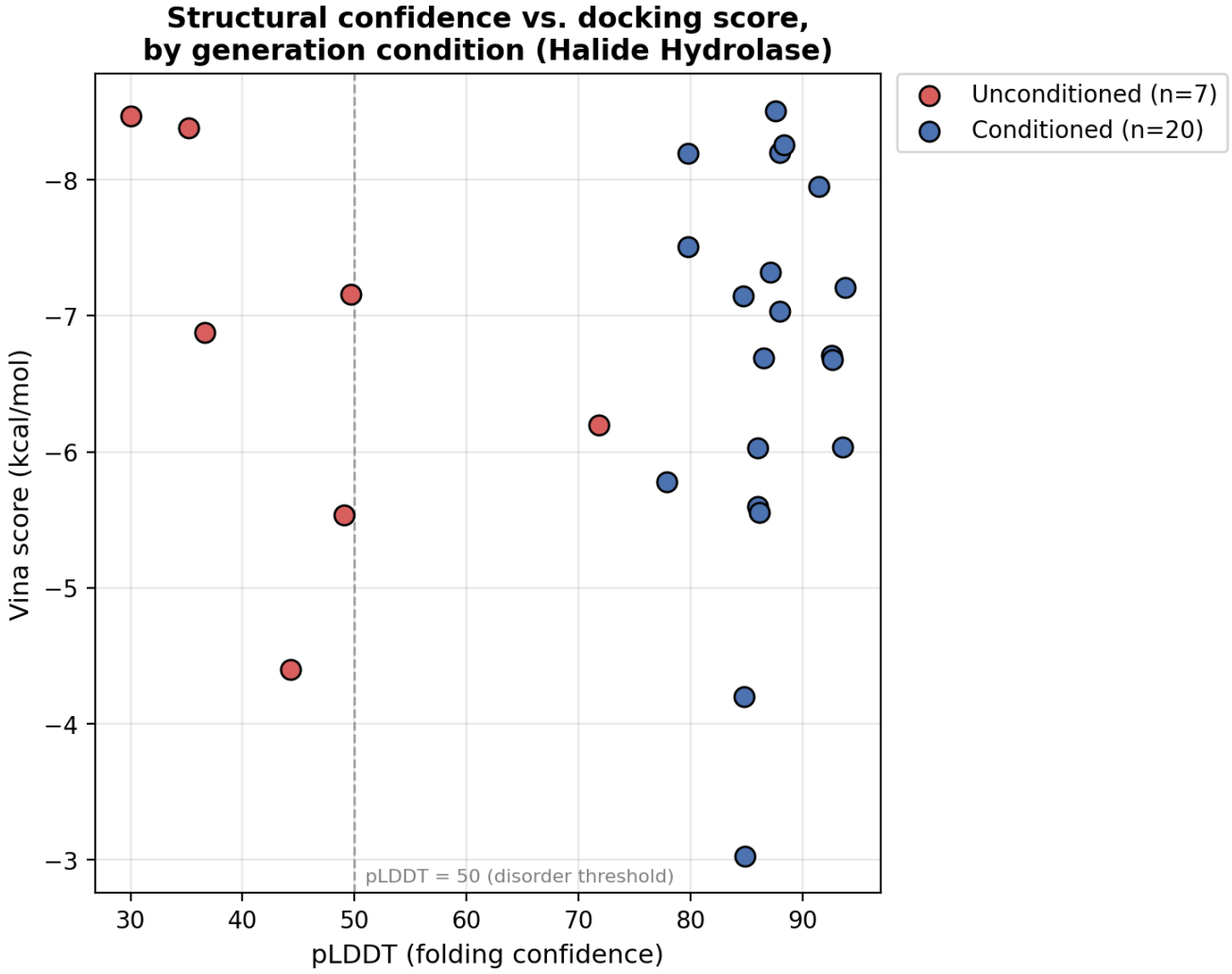

Figure 1: Relationship between folding confidence (pLDDT) and docking score (Vina) for unconditioned (red) and conditioned (blue) halide hydrolase generations, the reaction showing the clearest dissociation between structural plausibility and raw docking metrics. The dashed line marks the pLDDT = 50 disorder threshold. Corresponding plots for the remaining three reactions are provided in Supplementary Figs. S9–S14.

narrower but firmer and does not depend on the confound: whatever improves pLDDT so dramatically leaves Vina scores untouched (Fig. 1), a finding about docking-based validation rather than about XenoFold specifically.

3.3.1 Reaction specificity: a mismatched-reaction control

The confound above leaves open whether reaction conditioning contributes anything on top of fine-tuning. We tested this directly using one of the three controls proposed in Section 4.1: the same trained halide hydrolase model, fed a chemically unrelated reaction (pyrethroid ester hydrolysis) instead of its target 4-chlorobenzoyl-CoA reaction, decoding otherwise identical

($n = 20$).

Folding confidence for this mismatched-reaction batch (82.0±12.2) did not differ significantly from the matched-reaction batch (87.0±4.5; Mann-Whitney $p = 0.060$), while both vastly exceed the unconditioned baseline (45.2 ± 13.8). Docking succeeded for only 6/20 mismatched-reaction structures, too few to interpret (Supplementary Table S17).

Supplying the wrong reaction to an already-trained model therefore changes output only marginally and without significance at this sample size. Taken with the exact-copy rates in Sections 3.1–3.2, the most parsimonious account is that per-class fine-tuning, not the chemistry supplied at generation time, determines what the model produces.

3.3.2 Benchmarking against external baselines

Generated sequences for each of the four reactions were compared against real, literature-characterized natural enzymes with documented catalytic activity on the same or a closely analogous substrate, processed through the identical fold-dock-validate pipeline (Section 2.2). The number of natural reference enzymes varied by reaction (1–5), so comparisons are reported primarily via percentile positioning, with a Mann-Whitney U test additionally reported where $n \geq 3$. Table 3 summarizes results across all four reactions; full per-enzyme percentile tables and overlay plots are provided in Supplementary Tables S6–S9 and Supplementary Figs. S16–S19.

Reaction (n_{natural})	**pLDDT**	**Vina score**
Dye-decolorizing peroxidase (2)	15th–25th pct.	55th–100th pct.
Ester hydrolase (5)	$p = 0.164$ (n.s.)	$p = 0.243$ (n.s.)
Lignin peroxidase (1)	30th pct.	100th pct.
Halide hydrolase (3)	$p = 0.0071^{**}$, $r = 1.00$	$p = 0.355$ (n.s.)

Table 3: Percentile position (or Mann-Whitney p-value, where $n \geq 3$) of real natural enzymes within the corresponding generated-sequence distribution. Ligand efficiency is omitted for the reason given in Table 2. Reference-set sizes of $n = 1$–5 give these tests very little power, so a null result here is weak evidence of equivalence rather than evidence of no difference. n.s. = not significant at $\alpha = 0.05$.

The relationship between generated and natural folding confidence was reaction-dependent: generated sequences scored above the natural references in two reactions (DyP peroxidase, halide hydrolase) and at or below them in the other two, while Vina scores showed no significant difference in any reaction. The halide result (87.0 vs. 72.2 pLDDT, $p = 0.0071$) should not be read as the generated enzymes being "better" than functional natural enzymes: pLDDT reflects the structure predictor's confidence, which is systematically higher for sequences resembling well-

represented families — consistent with the identity distributions in Section 3.1 — rather than a measure of superior design, and no functional evidence in this study supports the alternative reading. The null docking results likewise should not be over-read: with $n = 1$–5 natural references per reaction, these tests have almost no power, and failing to detect a difference is not evidence that none exists.

3.4 Designed enzyme candidates

One candidate per reaction — highest average pLDDT after residue-validity filtering — is given with full sequence, Vina score, and ligand efficiency in Supplementary Tables S10–S13 (LE informative here since substrate sizes differ, 21–57 heavy atoms). Given Section 3.1, none are novel designs: the halide candidate is 99.7% identical to a specific training entry (one residue difference over ∼300), and the other three are comparably close. We report them as representative high-scoring outputs, not evidence of de novo design.

Two caveats attach to them. The tabulated stoichiometries omit required co-substrates — H_2O_2 for both peroxidase reactions — and represent the tetrabromobisphenol A transformation as a hydrolytic debromination, whereas peroxidases debrominate this substrate oxidatively via radical chemistry. These are artifacts of the SMARTS-based product rules used to build training labels, and they mean the conditioning reactions do not always match the true mechanism of the target class. Second, docking for the two peroxidases used apo ESMFold models with no heme, placing the ligand in a cavity from which the catalytic centre is absent; those Vina scores are not mechanistically interpretable.

4 Discussion

We report limitations in order of how much they constrain the conclusions, beginning with the one that most directly bounds the central result.

4.1 The ablation is confounded and does not isolate reaction conditioning

This is the principal limitation of the present work, and it is larger than we stated in an earlier version. That version described the two arms as differing in two respects — fine-tuning and conditioning. An audit of the generation code, performed after the analysis was complete, established that they differ in nine. The conditioned arm used a hand-written sampling

loop at temperature 1.1 with a maximum length of 800 residues, per-step logit steering toward target hydrophobic and charged composition, an EOS boost linear in generation progress, no restriction of the output alphabet, and a post-generation quality filter that discarded sequences below a threshold and rank-ordered the remainder. The unconditioned baseline used HuggingFace `generate()` at temperature 1.0 with nucleus sampling ($p = 0.9$), a maximum length of 250 residues, no composition steering, an EOS boost quadratic in progress, logit masking to the 20 canonical residues, and no quality filter or ranking. The full comparison is given in Supplementary Methods S1.4.

Four of these differences act directly on pLDDT. The length cap is the most consequential: it was set to 250 for the baseline to reduce peak GPU memory during folding, a resource constraint rather than a modelling decision, and folding confidence is sensitive to whether a sequence contains a complete structural unit. The quality filter is the second: it selects against repeats, low-complexity windows and off-target composition — the same properties that depress pLDDT — and was applied to one arm only. We therefore withdraw the claim that the base-versus-trained pLDDT difference is informative about fine-tuning, about conditioning, or about anything else. Table 2 is retained so the measurements are on record, not as support for any inference. The very large effect sizes ($d > 4.7$, $r = 1.00$) should be read as reflecting the accumulated procedural differences between two pipelines, not a graded improvement in generation quality.

Of three inexpensive controls that would isolate the conditioning signal, we ran one (Section 3.3.1): the identically trained decoder fed a mismatched reaction. That result — no significant change in folding confidence when the supplied reaction is wrong — argues against a reaction-specific conditioning effect, but it does not decompose the base-versus-trained gap itself into fine-tuning and conditioning components, which requires either a cross-attention-ablated retrain or a plain per-class fine-tune with no chemical encoder. Neither was run. Baseline batches should also be generated at $n = 20$ to match the conditioned arm.

4.2 Sequence similarity to natural proteins

This measurement was completed and is reported in Sections 3.1–3.2: 40% of generated sequences across four reactions are exact training-set copies, and this persists for substrates with no natural degrading enzyme. Two things remain unresolved. First, only the single top-scoring training alignment was examined per sequence; a small number of additional hits per query

would confirm whether near-copies (95–99% identity) match one training entry consistently or recombine features of several. Second, decoding was not seeded, so the specific sequences evaluated here cannot be regenerated byte-for-byte; they are archived in full, and all decoding parameters — constant across every condition — are reported in Supplementary Methods S1.3. Because those parameters never varied, the range in exact-copy rate reported in Sections 3.1–3.2 cannot be a decoding artefact.

4.3 Validation gaps: docking, catalytic machinery, and wet-lab confirmation

Three related gaps bound how far the in-silico results can be trusted. First, docking was uncontrolled: no negative controls (substrates in unrelated receptors; decoys in the designed enzymes) were run, receptor preparation removed all hetero-atoms so the two peroxidase systems were docked without heme, and a single Vina run per sequence (exhaustiveness 8, no replicate seeds) leaves run-to-run variance unquantified for ligands flexible enough that it may rival the differences being tested; P2Rank's top pocket was used without checking it against expected catalytic residues, so binding was assessed for occupancy, never catalytic competence. Second, no sequence-level check was performed for the catalytic machinery each class requires (the Ser-His-Asp triad, class II peroxidase histidines, the HAD-family aspartate nucleophile); their presence would be far stronger evidence than pLDDT, and fold-level agreement, signal peptides, and aggregation propensity were likewise not measured. Third, all results are computational: no generated sequence was expressed, purified, or assayed, so structural and binding predictions remain predictions of plausibility rather than demonstrations of function.

4.4 Exclusion of docking-derived kinetic estimates

An early version of this pipeline estimated Michaelis constants (K_M) deterministically from AutoDock Vina binding free energy via the relationship $\Delta G = RT \ln(K_d)$, under the assumption $K_M \approx K_d$. This approach was validated against two real, literature-characterized natural enzymes (Pleos-DyP1 and Pleos-DyP4; (Fernández-Fueyo et al., 2015)) processed through the identical pipeline, which revealed substantial deviation from experimentally measured K_M values (738-fold and 15,606-fold, respectively) and, critically, an inversion of the two enzymes' correct relative binding ranking: Pleos-DyP1 exhibits the tighter experimentally measured K_M but a weaker predicted Vina score than Pleos-DyP4. Given this demonstrated unreliability, docking-derived K_M/K_d estimates are not reported anywhere in this study. AutoDock Vina

binding score and ligand efficiency were used as the primary computational binding metrics instead.

The substitute metrics fared little better. After Holm correction, *no* Vina comparison distinguished trained from base generation, or generated from natural enzymes. The single raw-threshold result (lignin peroxidase, $p = 0.0495 \rightarrow$ Holm $p = 0.198$) came from the model with diagnosed contamination. Ligand efficiency added nothing: within a reaction it is a linear transform of the Vina score, so the two were never independent tests.

Our conclusion is stronger than a recommendation for caution. Across every comparison in this study, rigid-receptor docking on predicted apo structures produced no result surviving correction for multiple testing; the same pipeline also inverted the known ranking of two characterized enzymes. We do not think these scores carry usable signal about designed-enzyme quality as implemented here, and report them so the negative result is on record.

4.5 Sample sizes and statistical power

Baseline batches ($n = 6$–8) and natural-enzyme reference sets ($n = 1$–5) are small; normality-based test selection at these sizes is weakly grounded, and null results in the natural-enzyme comparison should be read as uninformative rather than as evidence of equivalence.

4.6 Training data quality and class-specific generative failures

- **Reduced generative diversity for the bacterial copper laccase model.** Repeated stochastic sampling produced substantial output redundancy for this reaction (this exploratory check predates the evaluated batches and used temperatures between 0.75 and 1.1; all batches reported in this study used 1.1), with only 7 of 19 sampled sequences unique despite non-deterministic decoding confirmed to be functioning correctly, suggesting the underlying generative model for this enzyme class may be constrained by limited training data diversity.

- **Training data contamination for the lignin peroxidase model.** BLAST homology analysis revealed that 80% of generated lignin peroxidase sequences matched lipoyl synthase or lipoyl(octanoyl) transferase (LipB), enzymes mechanistically and structurally unrelated to lignin peroxidase (radical SAM/[4Fe-4S] chemistry versus heme-dependent oxidoreductase chemistry), with full-length query coverage confirming genuine, not coincidental, matches. This pattern is most consistent with contamination or mislabeling in

the training data collection process for this specific model. This issue was isolated to the lignin peroxidase model; the remaining three reactions examined showed 75–85% functional coherence with their intended enzyme class (75% and 85% direct matches for halide hydrolase and DyP peroxidase, respectively, and 80% when direct and mechanistically-adjacent same-chemistry matches are combined for ester hydrolase; see Section 3.1 for the basis of this categorization).

These findings collectively indicate that data curation and per-class training set diversity are the primary drivers of generative reliability in this pipeline, more so than the underlying model architecture, and should be prioritized in future extensions of this work.

4.7 Modeling simplifications

Two further simplifications bear on interpretation. Structure predictions used ESMFold in single-sequence mode, efficient but less accurate than MSA-based methods for sequences with few close homologs, and docking used a single static structure per sequence with no accounting for conformational flexibility or induced-fit effects. Separately, reaction product prediction for training-data generation relied on RDKit SMARTS-based functional group detection with simplified atom-addition rules (e.g. generic hydroxyl addition for hydrolysis) that do not account for regiochemistry or stereochemistry — a reasonable approximation for generating training labels at scale, but a simplification that should be considered when interpreting the chemical specificity of generated candidates.

4.8 Scope of validation

This study's ablation, natural-enzyme comparison, and BLAST homology analyses were performed for four of the fifteen enzyme classes supported by this pipeline (dye-decolorizing peroxidase, ester hydrolase, lignin peroxidase, and halide hydrolase), each evaluated on a single representative reaction. Sample sizes for unconditioned-baseline and natural-enzyme comparisons were necessarily limited by the computational and API-access constraints of large-scale sequence folding and docking (unconditioned baselines: $n = 6$–8; natural-enzyme references: $n = 1$–5 per reaction, depending on availability of characterized enzymes in the literature), which limits statistical power, particularly for the Vina score and ligand efficiency comparisons discussed in Section 4.4. A comparison against structure-generation-based tools such as

RFdiffusion combined with inverse-folding via ProteinMPNN was not performed, given the substantial difference in design objective (unconditioned or motif-scaffolded backbone generation, versus direct reaction- and enzyme-class-conditioned sequence generation in this work) and the considerable additional infrastructure required for a fair comparison. Nor was any comparison made against the conditional protein language models closest to this work — ZymCTRL, which is already EC-conditioned, or ProGen — or against retrieval-based reaction-to-enzyme tools such as Selenzyme, all of which would be substantially cheaper points of reference than RFdiffusion and more directly informative about whether this pipeline offers anything those methods do not. We regard these comparisons, rather than broader class coverage, as the most important missing benchmarks.

5 Conclusions

We presented XenoFold, a chemistry-conditioned generative framework for candidate enzyme design, and a validation study that changed our own reading of what it does.

The trained pipeline folds with substantially higher confidence than an untrained base model, but we no longer present that comparison as a result: the two arms were generated by different code differing in nine respects, four of which act directly on folding confidence (Section 4.1). The findings below do not depend on it. BLAST against each model's own training set showed 40% of sequences to be exact copies and 87.5% at least 95% identical, and swapping the conditioning reaction for an unrelated one changed folding confidence only marginally and non-significantly ($p = 0.060$). The parsimonious account is that fine-tuning teaches the decoder to reproduce its training distribution, with little detectable contribution from the chemical encoder. This held for substrates with no documented degrading enzyme: PFOA and p,p'-DDE produced exact copies at rates comparable to an in-training control (DDT). A fourth novel substrate on a less-trained checkpoint departed further from training while remaining confidently folded; we report this as an open, confounded observation, not a result.

The results we consider most durable are negative. Docking-derived K_M estimates deviated from measured values by up to 15,606-fold for two characterized DyP peroxidases and inverted their ranking. After correction, no Vina comparison distinguished trained from untrained generation, or generated from natural enzymes, in any reaction — rigid-receptor docking on predicted apo structures carried no usable signal in twelve comparisons across four reactions.

This study's central finding is therefore about validation rather than about design capability. pLDDT, nr-BLAST, and docking each independently indicated plausible, apparently novel enzymes; only identity against the model's own training set revealed memorization, and it was not part of the original protocol. We regard this as a documented failure mode for reaction-conditioned generative protein design under Cartesian sequence-reaction pairing and small per-class datasets: the checks that would have caught it are individually inexpensive and, in our view, prerequisites rather than optional validation.

XenoFold in its current form does not produce enzymes ready for experimental deployment, and we cannot claim it narrows sequence space more effectively than family-based database mining. Whether a matched, adequately trained model behaves differently for chemistry genuinely absent from its training data remains unanswered here, and is, in our view, the necessary next experiment before any design claim can be revisited.

List of abbreviations

BDE-209: decabromodiphenyl ether; BLAST: Basic Local Alignment Search Tool; CoA: coenzyme A; DDE: dichlorodiphenyldichloroethylene; DDT: dichlorodiphenyltrichloroethane; DyP: dye-decolorizing peroxidase; EC: Enzyme Commission; HAD: haloacid dehalogenase; K_d: dissociation constant; K_M: Michaelis constant; LE: ligand efficiency; LiP: lignin peroxidase; LipB: lipoyl(octanoyl) transferase; MSA: multiple sequence alignment; PFOA: perfluorooctanoic acid; pLDDT: predicted local distance difference test; SMARTS: SMILES arbitrary target specification; SMILES: simplified molecular-input line-entry system; TBBPA: tetrabromobisphenol A.

Declarations

Availability of data and materials

All data, code, and model weights supporting the conclusions of this article are publicly available without registration or login.

- **Project name:** XenoFold
- **Project home page:** `https://github.com/KinsleyTorres/XenoFold`

- **Archived version (data and analysis code):** generated sequences, ESMFold-predicted structures, AutoDock Vina outputs, BLAST results, the four per-class training datasets, and the analysis scripts reproducing every table and figure are archived at `https://doi.org/10.5281/zenodo.22066938`.
- **Archived version (model weights):** the four fine-tuned checkpoints evaluated in this study are archived and citable as Torres (2026c), Torres (2026a), Torres (2026b), and Torres (2026d). Exact repository revisions are pinned in `config/models.yaml` in the project repository and are passed explicitly via the `revision` argument at load time, so that the weights retrieved are those that produced the results reported here.
- **Operating system:** platform independent
- **Programming language:** Python
- **Other requirements:** PyTorch, HuggingFace Transformers, RDKit, AutoDock Vina 1.2.7, MGLTools, NCBI `blastp`
- **License:** MIT (source code)
- **Any restrictions to use by non-academics:** none

Fifteen per-class models were trained; the four listed above are those evaluated in this study and are the only checkpoints on which any reported result depends.

Competing interests

The author declares no competing interests.

Funding

No specific funding was received for this work.

Authors' contributions

KT conceived the study, implemented the pipeline, performed all computational experiments and statistical analysis, and wrote the manuscript. The author read and approved the final manuscript.

Acknowledgements

Not applicable.

Ethics approval and consent to participate

Not applicable. This study is entirely computational and did not involve human participants, animal subjects, or sensitive personal data.

Additional files

Additional file 1 — Supplementary Material (PDF). Supplementary Methods S1, Supplementary Tables S1–S17 and Supplementary Figures S1–S19, referenced throughout the main text.

References

Balázs Ádam, Paola Cocco, and Lode Godderis. Hazardous effects of pesticides on human health. *Toxics*, 12(3):186, 2024.

Stephen F. Altschul, Warren Gish, Webb Miller, Eugene W. Myers, and David J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, 1990.

Pablo Carbonell, Jerry Wong, Neil Swainston, Eriko Takano, Nicholas J. Turner, Nigel S. Scrutton, Douglas B. Kell, Rainer Breitling, and Jean-Loup Faulon. Selenzyme: enzyme selection tool for pathway design. *Bioinformatics*, 34(12):2153–2154, 2018. doi: 10.1093/bioinformatics/bty065.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 2021.

Elena Fernández-Fueyo, Dolores Linde, David Almendral, María F. López-Lucendo, Francisco J. Ruiz-Dueñas, and Angel T. Martínez. Description of the first fungal dye-decolorizing peroxidase oxidizing manganese(ii). *Applied Microbiology and Biotechnology*, 99(21):8927–8942, 2015.

Chenqing Hua, Jiarui Lu, Yong Liu, Odin Zhang, Jian Tang, Rex Ying, Wengong Jin, Guy Wolf, Doina Precup, and Shuangjia Zheng. Reaction-conditioned de novo enzyme design with genzyme. *arXiv preprint arXiv:2411.16694*, 2024.

Radoslav Krivák and David Hoksza. P2rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *Journal of Cheminformatics*, 10(1):39, 2018.

Li Lin, Haoran Yang, and Xiaocang Xu. Effects of water pollution on human health and disease heterogeneity: A review. *Frontiers in Environmental Science*, 2022.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41(8):1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Peter G. Mikhael, Itamar Chinn, and Regina Barzilay. Clipzyme: Reaction-conditioned virtual screening of enzymes. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 35647–35663, 2024.

Geraldene Munsamy, Sebastian Lindner, Philipp Lorenz, and Noelia Ferruz. Zymctrl: A conditional language model for the controllable generation of artificial enzymes. In *Machine Learning for Structural Biology Workshop, NeurIPS*, 2022.

Thomas Münzel et al. Transportation noise pollution and cardiovascular disease. *Nature Reviews Cardiology*, 2021.

Erik Nijkamp, Jeffrey Ruffolo, Eli N. Weinstein, Nikhil Naik, and Ali Madani. Progen2: Exploring the boundaries of protein language models. *Cell Systems*, 2023.

Andrew Parkinson, Brian W. Ogilvie, David B. Buckley, Faraz Kazmi, Maciej Czerwinski, and Oliver Parkinson. Biotransformation of xenobiotics. In Curtis D. Klaassen, editor, *Casarett and Doull's Toxicology: The Basic Science of Poisons*. McGraw-Hill Education, 8th edition, 2013.

S. Rathore, A. Varshney, S. Mohan, and P. Dahiya. An innovative approach of bioremediation in enzymatic degradation of xenobiotics. *Biotechnology & Genetic Engineering Reviews*, 2022.

Kinsley Torres. Xenofold dye-decolorizing peroxidase model (epoch 2). Hugging Face. `https://doi.org/10.57967/hf/10132`, 2026a.

Kinsley Torres. Xenofold ester hydrolase model (epoch 12). Hugging Face. `https://doi.org/10.57967/hf/10135`, 2026b.

Kinsley Torres. Xenofold halide hydrolase model (epoch 8). Hugging Face. `https://doi.org/10.57967/hf/10134`, 2026c.

Kinsley Torres. Xenofold lignin peroxidase model (epoch 2). Hugging Face. https://doi.org/10.57967/hf/10133, 2026d.

Oleg Trott and Arthur J. Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 2010.

Joseph L. Watson, David Juergens, Nathaniel R. Bennett, Brian L. Trippe, Jason Yim, Helen E. Eisenach, Woody Ahern, Andrew J. Borst, Robert J. Ragotte, Lukas F. Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620:1089–1100, 2023.

Tianhao Yu, Haiyang Cui, Jianan Canal Li, Yunan Luo, Guangde Jiang, and Huimin Zhao. Enzyme function prediction using contrastive learning. *Science*, 379(6639):1358–1363, 2023. doi: 10.1126/science.adf2465.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- XenoFoldJCheminfsupplementarysource1.zip