Dual-encoder contrastive learning accelerates enzyme discovery

Jason W. Rocks[1], Dat P. Truong[1], Dmitrij Rappoport[1], Sam Maddrell-Mander[1], Daniel A. Martin-Alarcon[1], Toni Lee[1], Steve Crossan[1] and Joshua E. Goldford[1#]

[1]Dayhoff Labs, Inc. Cambridge, MA, USA

[#]To whom correspondence should be addressed: josh@dayhofflabs.com

Keywords: enzymology, deep learning, artificial intelligence

Abstract

The ability to engineer enzymes for desired reactions is a cornerstone of modern biotechnology, yet identifying suitable starting proteins remains a critical bottleneck. Dual-encoder contrastive learning models have emerged as a promising approach for enzyme discovery, learning to match chemical reactions with catalyzing enzymes through a shared embedding space. However, their practical performance beyond computational benchmarks remains unproven. Here, we close this gap with *Horizyn-1*, a deep learning framework for direct reaction-to-enzyme recommendation validated through comprehensive experimental testing. Leveraging a computationally efficient combination of reaction fingerprints and protein language models, we trained *Horizyn-1* on 8.9 million reaction-enzyme pairs to achieve state-of-the-art performance, recovering an enzyme with correct activity within the top 100 hits for over 75% of test reactions. We experimentally validate *Horizyn-1* across three enzyme discovery scenarios: identifying enzymes for orphan reactions, predicting enzyme promiscuity for both characterized and uncharacterized enzymes, and discovering enzymes for non-natural biochemical reactions including lysine-driven transaminations that enable efficient synthesis of non-canonical amino acids. On underrepresented reaction classes, we find that fine-tuning with fewer than 10 additional reactions can dramatically improve performance. Furthermore, a logarithmic scaling of model performance with training dataset size suggests continued improvement with larger and more diverse reaction datasets. *Horizyn-1* addresses the critical bottleneck of sourcing initial enzymes for optimization campaigns, enabling efficient and scalable *in silico* screening for enzymes with desired activities and promising to accelerate future efforts in biocatalysis and metabolic engineering.

Introduction

The emerging field of generative models for enzyme design holds immense promise for creating novel biocatalysts (1, 2). However, the performance and scope of these models are critically dependent on the quality and breadth of the training data, which is dominated by known, wild-type enzymes (3). Consequently, future advancements in this area require robust tools for the discovery and annotation of enzymes exhibiting novel reactivity (3). Advances in high-throughput metagenomic sequencing have yielded an explosion in non-redundant protein sequences throughout the last two decades, many of which contain domains of uncharacterized enzymatic function (4). Identifying previously undescribed catalytic activities within these naturally occurring enzymes presents a grand challenge with significant implications for disciplines as diverse as oncology (5) and microbial ecology (6).

Historically, protein function annotation tools have relied on conventional bioinformatics and sequence-similarity based approaches (7–11). More recently, advances in deep learning have enabled models to learn complex sequence-function relationships from large datasets, leading to accurate predictions for functional descriptors such as E.C. numbers or GO terms (12–14). However, both sets of approaches are fundamentally limited by their reliance on indirect descriptions of catalytic activity (via sequence homology or functional labels) instead of direct representations of the biochemical reactions themselves. The BridgIT algorithm, a notable exception, uses similarity between custom reaction fingerprints to identify plausible enzymes for a query biochemical reaction, but lacks any machine learning capabilities (15).

Contrastive learning is an efficient machine learning technique for linking elements from different domains (16) and has been recently used to directly associate reactions and enzymes (17–20). In particular, it can be readily applied to novel or hypothetical biochemical reactions to efficiently screen large libraries of protein sequences for activity. Despite promising results on machine learning benchmarks, this approach has not undergone experimental validation. In this work, we provide the first experimental validation of contrastive learning models for the enzyme discovery.

Results

Scalable enzyme annotation and discovery with joint reaction and enzyme embedding

To explore the utility of contrastive models to guide experimental enzyme discovery and annotation, we developed *Horizyn-1* (Fig. 1, see Methods). This model adapts the contrastive learning framework popularized by the CLIP image-to-text model to learn associations between reactions and protein sequences (Fig. 1A**)** (16). Input reactions and proteins are converted to vector representations and then projected into a common high-dimensional embedding space using separate Multi-Layer Perceptron (MLP) encoders. Within this space, the model is trained to minimize the distance between known reaction-enzyme pairs (positives) while maximizing distances for unrelated pairs (negatives). This process yields a joint embedding space where proximity correlates with the likelihood a protein catalyzes a given reaction, enabling enzymes or reactions to be ranked in order of relevance to a particular query reaction or enzyme (Fig. 1B-C). *Horizyn-1* computes scores via the dot-product (cosine similarity, ranging between -1 and 1) between the embedding vectors of a given reaction-enzyme pair.

For model development, we constructed a training dataset of 294,166 reaction-enzyme pairs derived from 11,961 biochemical reactions in the Rhea database (21), with 216,132 corresponding enzyme sequences from Swiss-Prot (22). We constructed a representative test set of 1,012 reactions by splitting the data based on reaction similarity, calculated using classical fingerprinting techniques (see Methods). To evaluate model performance, we measured how highly the model ranked enzymes known to catalyze each test reaction relative to all other enzymes in the dataset. We report a variety of standard retrieval metrics, but focus primarily on top-*k* hit rates, measuring the percentage of reactions in the test set where at least one known enzymatic catalyst was successfully recovered from the top *k* ranked sequences. Top-*k* metrics are especially informative because they map directly to experimental design: the top-*k* predicted sequences correspond to the number of variants synthesized and screened in a typical enzyme discovery campaign (see Methods).

While we considered a wide variety of different model architectures and training strategies, we found that the choice of strategy for creating reaction and protein representations were most impactful. Our most successful model architecture combined a hybrid reaction representation—using extended-connectivity fingerprints (ECFP6, also known as Morgan fingerprints) that combine the structures of the reactants and products (structural fingerprints) (23, 24) and Differential Reaction Fingerprints (DRFP) that encode the bonding changes (differential fingerprints) (25) —with pre-computed protein sequence embeddings from the 3-billion-parameter ProtT5 protein language model (26). This combination achieved strong performance, identifying at least one known enzymatic catalyst within the top 100 recommendations for 76.7% of test reactions (Fig. 1D). In contrast, a baseline approach using E.C. numbers up to the third digit to rank sequences—essentially using detailed knowledge of the reaction class—achieved a top-100 hit rate of only 27.5%. Notably, we found that combining our model with partial E.C. class knowledge could boost top-*k* retrieval metrics by approximately 5-10% (see Methods and Table S1). Substituting the reaction representation with approaches based on transformers (rxnfp) (27) or graph neural networks (Chemprop) (28) generally led to lower scores, with significantly reduced top-1 and top-10 hit rates (Fig. 1D). We also tested protein embeddings generated with ESM-2 (650M and 3B) (29), but observed slightly reduced performance, especially for the smaller model (Fig. 1D). In addition to pre-computed protein embeddings, we also explored fine-tuning ESM-2 (650M) to allow its weights to update during training, as well as modifying the architecture to incorporate 3D protein structures. However, neither approach offered substantial performance improvements over static embeddings despite significantly increased computational costs (see Methods and Fig. S1). Finally, we directly compared the model architecture and training procedure of our best model with those of CLIPZyme (17), an influential contrastive model for reaction-enzyme association. We found that our model moderately outperformed CLIPZyme, likely due to the use of a larger protein language model (ProtT5 at 3 billion vs. ESM-2 at 650 million parameters) and fingerprint-based reaction representations rather than a learned graph neural network implementation based on pseudo-transition states (see Table S2, Discussion).

To investigate the relationship between training dataset size and model performance, we performed data-scaling experiments by varying the number of training reactions (clustered on

reaction fingerprint similarity, see Methods). We found that training with only 10% of the data yielded a model with a top-100 hit rate exceeding 40%. We observed logarithmic scaling of model performance with the number of training reactions, indicating that an approximate tenfold increase in reactions could yield a model with a top-100 hit rate approaching 100% (Fig. 1E). These results suggest that further scaling of the model with a substantially larger—and diverse—set of reactions and protein-reaction pairs could yield a model with enough predictive power to construct high-quality panels of enzymes to guide enzyme discovery campaigns with high confidence.

As model performance improves significantly with both the number and diversity of reactions, we assembled a large dataset of reaction-enzyme pairs by combining multiple databases (totalling 8,897,870 pairs, 31,101 reactions, and 7,063,237 protein sequences) and trained a model to be used for experimental inference, which we call *Horizyn-1* (see Methods). Analysis of *Horizyn-1*'s score distribution revealed a strong divergence between scores for positive and negative reaction-enzyme pairs in the training set, allowing us to use thresholds based on these score distributions to identify high-confidence predictions (see Methods, Fig. S2). This model was applied to predict activity for three experimental use cases: (i) identifying enzymes for orphan reactions, (ii) predicting promiscuous enzyme activity, and (iii) discovering enzymes for non-natural reactions.

Experimentally de-orphaning reactions and enzymes with *Horizyn-1*

As a first experimental test case for *Horizyn-1*, we sought to identify enzymes for orphaned biochemical reactions in the Rhea database. Orphaned reactions are those that have been experimentally observed or inferred to exist, but for which a catalyzing protein has not yet been annotated (30), thus providing a strong test case for enzyme discovery. We initially focused on the 6-aminohexanoate aminotransferase activity (RHEA:58200, E.C. 2.6.1.116) observed in *Arthrobacter sp. KI72*, a nylon-degrading bacteria (Fig. 2A) (21), (31). We queried *Horizyn-1* with the 6-aminohexanoate aminotransferase reaction against all protein coding sequences (n = 4,173) from this genome (NCBI accession GCF_002049485.1) (Fig. 2B, Fig. S3). We identified two sequences (WP_021472307.1 and WP_079941468.1) as high confidence candidates. These

proteins were previously annotated as a 4-aminobutyrate--2-oxoglutarate transaminase (GabT, E.C. 2.6.1.19) and an aminotransferase class III-fold pyridoxal phosphate-dependent enzyme (E.C. 2.6.1.-), respectively, via the NCBI prokaryotic genome annotation pipeline (32) (Fig. 2B). Subsequent expression, purification, and biochemical assay of these two proteins confirmed significant 6-aminohexanoate aminotransferase activity relative to the uncatalyzed reaction (Fig. 2C).

The two enzymes identified in *Arthrobacter sp.* KI72 shared significant sequence similarity (43.3% and 43.5%, respectively) with GabT from *E. coli* (P22256) (Fig. 2B). *E. coli* GabT is known to catalyze ω-transamination reactions with various substrates, including 4-aminobutyrate and 4-aminopentanoate. Notably, *Horizyn-1* also ranked GabT as the top hit for 6-aminohexanoate aminotransferase in our training set (Fig. S4), suggesting a potential for promiscuous activity. Based on these observations, we purified and tested *E. coli* GabT for 6-aminohexanoate aminotransferase activity, observing higher activity for this reaction than the two *Arthrobacter sp. KI72* enzymes, and comparable activity to transamination with its annotated substrate, 4-aminobutyrate (RHEA:23352, E.C. 2.6.1.19) (Fig. 2C**,** Fig. S4). Furthermore, *Horizyn-1* predicted strong scores for *E. coli* GabT with two additional orphan ω-transaminase reactions with alternative substrates, β-alanine (RHEA:30699, E.C. 2.6.1.120) and putrescine (RHEA:23816, E.C. 2.6.1.82). Experimental testing of GabT with these additional substrates revealed weak promiscuous activity (Fig. S4), suggesting that *Horizyn-1* may be capable of predicting promiscuous activities for well-characterized enzymes.

We next sought to explore whether *Horizyn-1* could be used to predict the promiscuity of less well-characterized enzymes. To this end, we identified protein sequences from the model thermophilic bacteria *Thermus aquaticus* (NCBI Taxonomy ID: 271) that were identified as enzymes (i.e., had an assigned E.C. number), but were not included in our training data (Fig. 2D). We searched for sequences that had no more than 50% sequence similarity to any enzymes in our training dataset (n=382), and queried *Horizyn-1* for each sequence against all Rhea reactions (n=15,814). After filtering predictions for reactions that corresponded to previously annotated E.C. classes, we identified 818 high scoring (>0.80) hits for promiscuous or alternative activity. Upon inspection, we found several high scoring reactions for a hypothetical D-lactate

dehydrogenase (A0A0N0U829, E.C. 1.1.2.4) (Fig. 2E). The prediction of *Horizyn-1* for the top-scoring reaction for this enzyme (RHEA:13521) was consistent with the result from homology-based models. However, we found that this enzyme also scored highly for (R)-2-hydroxyglutarate--pyruvate transhydrogenase (RHEA:51608, E.C. 1.1.99.40) activity, suggesting this enzyme was promiscuous for alternative electron/donor pairs when oxidizing (R)-lactate (Fig. S5). To test for transhydrogenase activity for this enzyme, we expressed, purified, and developed a coupled assay, measuring pyruvate production via a coupled lactate dehydrogenase assay (see Methods). We observed significant depletion of NADH in the presence of this enzyme, suggesting that A0A0N0U829 can act as a (R)-2-hydroxyglutarate--pyruvate transhydrogenase (Fig. 2G).

As a more challenging test case, we sought to test whether *Horizyn-1* could be used to predict the activity of completely uncharacterized enzymes. To this end, we identified protein sequences from *T. aquaticus* that were both labeled as an "uncharacterized protein" and contained no homologous sequences in our training data (n = 274, see Methods). We then queried *Horizyn-1* for these sequences against all Rhea reactions (n=15,814), and found 61 high scoring (>0.80) pairs of reactions and proteins. Upon inspection, we found that the uncharacterized protein A0A0N0BLM2 scored highly for cytidine monophosphate (CMP) and deoxycytidine monophosphate (dCMP) kinase activities (RHEA:11600, RHEA:25094) (Fig. 2F, Fig. S5). We recombinantly expressed and purified this protein, and measured (d)CMP production using LC/MS-multiple reaction monitoring (MRM) (Methods). We were able to confirm production of dCMP, but not CMP, suggesting that *Horizyn-1* may experience difficulty discriminating between very similar substrates (Fig. 2H**,** Fig. S6). Importantly, traditional homology methods (9, 33), structural comparisons (34), and modern deep learning E.C. classification (12) tools did not yield high confidence predictions. Interestingly, the most similar cytidylate kinase in our training data shared only 28% sequence similarity to this enzyme, suggesting that *Horizyn-1* may be useful for predicting promiscuous activities even for sequences with low homology to previously characterized enzymes.

Enzyme discovery for non-natural biochemical reactions

The experimental confirmation above has been applied to cases where we have screened large libraries of reactions and searched for enzyme-reaction pairs with high model scores. To assess the utility of *Horizyn-1* for an enzyme discovery campaign, we chose to query reactions recently reported to be of interest in the synthesis of non-canonical amino acids (ncAA) used in pharmaceutical manufacturing (35). These reactions use so-called "smart" amine donors, which undergo cyclization after amine transfer to the corresponding keto acid acceptor. This cyclization provides a greater thermodynamic driving force than native amine donors such as glutamate (Fig. 3A). Although prior literature reported the feasibility of such reaction types, the specific transaminases capable of performing these reactions were not reported (35). We used *Horizyn-1* to screen all enzymes in our screening set to generate a panel for experimental validation.

We chose three non-canonical keto acids – phenylglyoxylic acid, 2-thiopheneglyoxylic acid, and 2-furylglyoxylic acid – which can be used to synthesize phenylglycine, 2-thienylglycine, and 2-furylglycine, respectively. Two pathways are possible wherein the amine in either the α or ε position can act as the transamination donor (Fig. 3A). We queried *Horizyn-1* for all six reaction variants against all ~7 million enzymes in our screening set. We then constructed a panel of six transaminases with high scores for all three substrates and at least one of the α- or ε-donating reaction variants.

Of the six enzymes in our panel, we were able to express and purify four. We screened these four enzymes for activity on all substrates using a colorimetric assay, in which cyclized lysine-derived products react with *o*-aminobenzaldehyde to form compounds with absorbance between 450 and 465 nm (Fig. S7, Methods). We observed significant activity for one enzyme, TA2 (P09053), and minor activity for two others on all three substrates (Fig. 3B). We confirmed the production of the expected non-canonical amino acids via LC/MS-MRM and purchased standards; for example, we observed a characteristic transition from 152.2 *m/z* to 107.2 *m/z* for PG (Fig. 3C**)** as well as characteristic transitions for the other ncAAs **(**Fig. S8) (36–38).

Finally, we tested the ability of *Horizyn-1* to distinguish between the α- or ε-donating reaction variants. For TA2, scores were higher for α-transamination compared to ε-transamination for all three substrates, suggesting that the mechanism of TA2 transamination is predominantly from the α-position (Fig. 3D). We devised an isotope-labeling study to confirm this prediction by supplying $^{15}N_{\varepsilon}$-labeled L-lysine and measuring the relative intensities of the monoisotopic (M+0) versus heavy isotope (M+1) of L-phenylglycine **(**Fig. 3E**)**. Consistent with model predictions, we observed comparable heavy versus monoisotopic intensities when supplying either $^{15}N_{\varepsilon}$-labeled L-lysine or $^{14}N_{\varepsilon}$-labeled L-lysine, confirming that transamination predominantly stems from the α-amine.

Efficient inclusion of new experimental data rapidly yields predictive models for new reaction classes

Enzyme discovery often requires venturing into new and unexplored areas of chemistry. To assess the model's ability to generalize to such cases, we tested the model on more challenging splits of our training and evaluation data designed to decrease the attributes shared between the two sets (see Methods). As expected, performance degraded when reactions were stratified to ensure low maximum similarity between training and test reaction sets via fingerprint similarity thresholds or reaction templates (Fig. S9). To more directly assess this relationship between model performance and reaction novelty, we used our best model trained on the training split of the development dataset (Fig. 1B). For each reaction in the development test split, we then correlated the rank of the top-scoring positive enzyme with its distance to the closest reaction in the training split. We found a modest correlation for distances computed using chemical fingerprints (r = 0.29), but a substantially stronger one when using the distance within the model's shared embedding space (r = 0.60) (Fig. S10).

This degradation in performance for out-of-sample reactions underscores the critical need to incorporate data from diverse reaction classes to achieve robust generalizability. However, the quantity of experimental data from a new class necessary for accurate prediction remains unclear. To investigate this, we fine-tuned *Horizyn-1* using data from a novel reaction class,

ene-reductases (EREDs) (Fig. 4A**)**, and evaluated the model's resulting performance specifically for that class.

For this study, we manually curated 406 novel ERED reactions and 903 enzyme-reaction pairs from the literature (see Methods). Consistent with prior results, we found that the baseline inference model's performance on reactions from this new class strongly correlated with distance from the training set (Fig. 4B**,** Fig. S10). Next, we fine-tuned the *Horizyn-1* inference model using small, random subsets of these ERED reactions and assessed performance on corresponding held-out validation sets (see Methods). We found that fine-tuning with as few as 10 new ERED reactions substantially improved performance; the median top-100 hit rate increased from 38% to 71% after fine-tuning (Fig. 4C).

To explore the lower limits of data requirement, we also measured the impact of fine-tuning *Horizyn-1* on single ERED reactions. Remarkably, fine-tuning with just one reaction could, in some cases, markedly increase model performance. For example, using only the reduction of (2-nitro-vinyl)-benzene to (2-nitroethyl)-benzene for fine-tuning resulted in a top-100 hit rate of 84% on the ERED test set. However, this effect was highly variable, with post-fine-tuning hit rates ranging from 26% to 84%. We found that the magnitude of this performance increase correlated with the novelty of the single reaction used for fine-tuning relative to the training set ($r = 0.54$) (Fig. 4D, Fig. S11).

Altogether, these results suggest that acquiring even a small number of enzyme-reaction pairs is an effective strategy for rapidly enhancing predictive skill within a new reaction class. Crucially, *Horizyn-1*'s ability to identify reactions that are most distant from its training data provides a mechanism to proactively design experiments that will be most informative for model improvement. This points toward an efficient, model-guided active learning strategy to systematically expand the model's predictive domain.

Discussion

The full extent of enzyme diversity and promiscuity for both natural and non-natural biochemical reactions likely far exceeds that reported in literature. With increasing volumes of genomic data, machine learning models like *Horizyn-1* are essential for exploring this diversity, yet their true utility can only be confirmed through rigorous, prospective experimental validation. Such forward-looking assessment, which moves beyond the retrospective benchmarks common in machine learning, is vital for establishing the strengths and limitations of these tools in guiding discovery.

Our findings demonstrate the utility of the dual-encoder contrastive approach employed by *Horizyn-1* for critical biocatalysis tasks, including de-orphaning reactions (Fig. 2A-C), characterizing enzyme promiscuity (Fig. 2D-F), and enabling broader biocatalyst discovery for non-natural reactions (Fig. 3).

Our model's strong performance, particularly when compared to alternatives like CLIPZyme (Table S2), is designed to maximize the predictive power of limited available training data. While we sought to maximize our training dataset size, it remains modest compared to those in other domains (16). By leveraging ProtT5, a 3-billion-parameter protein language model, we harness the knowledge learned from the 10^9 protein sequences in its pre-training corpus (26). However, no model of comparable scale exists for chemical reactions, and existing pre-trained models like rxnfp lack sufficient coverage of biochemical transformations (27). Consequently, we found that a simple fingerprint-based approach exhibits superior performance compared to a reaction model trained from scratch (Fig. 1, Table S2). A likely explanation is that deep learning models struggle to learn informative features from a dataset that is both small and chemically diverse.

Although the model performs well, its capabilities are currently constrained by data availability, with opportunities for improvement through data scaling (Fig. 1E) and fine-tuning (Fig. 4). Notably, model performance degrades for reactions that are increasingly dissimilar to the training data (Fig. 4B, Fig. S9, Fig. S10), while training on even small amounts of novel, sufficiently

out-of-sample data can significantly enhance performance (Fig. 4C-D). This result underscores the importance of prioritizing reaction diversity over sheer data volume when collecting new data to improve model performance. Furthermore, distances within *Horizyn-1*'s shared embedding effectively quantify biochemical novelty (Fig. 4B,D, Fig. S10, Fig. S11). This capability could be leveraged to guide the generation of maximally informative data, aiming to build a broadly proficient model (39). This points toward a lab-in-the-loop active learning strategy that could guide enzyme discovery, dynamically balancing the exploitation of high-confidence predictions with the exploration of novel targets to efficiently expand the model's knowledge of chemical space.

Beyond data availability, several other limitations constrain predictive accuracy. For instance, the model associates entire enzyme sequences with reactions, without distinguishing the specific domains that catalyze reaction steps or accounting for regulatory elements. Furthermore, current chemical representations are a challenge; reaction SMILES are insufficient for complex molecular contexts like protein-based metabolites (e.g., ferredoxins) or oligonucleotides. Another key constraint is the lack of explicit negative training examples, which may cause the model to assign high rankings to similar but catalytically inactive proteins. Finally, the omission of reaction directionality and kinetic parameters such as K_{cat} and K_m values prevents the model from explicitly predicting reaction feasibility or catalytic efficiency. These limitations highlight clear opportunities for future development as larger and more diverse datasets become available.

Despite recent advances in generative enzyme design (2, 40), enzyme discovery combined with directed evolution remains the gold standard for generating enzymes with novel functionality (41). *Horizyn-1* promises to accelerate this established discovery pipeline and potentialy serve as a foundational model for generative enzyme design tasks, effectively bridging the gap between discovery and *de novo* creation.

Acknowledgements

We thank Margaux Pinney, Adrian Jinich and Lynn Kamerlin for helpful discussions. We are also grateful to Avi Flamholz for providing valuable comments on the manuscript. We acknowledge support by the Department of Energy Small Business and Innovation Research (SBIR) program

through grant number DE-SC0025760. Finally, we thank Rita Clare and Scivetica for assistance with illustrations.

Competing Financial Interests

All authors are employees and/or equity shareholders in Dayhoff Labs, Inc.

Corresponding Authors

Correspondence to: josh@dayhofflabs.com

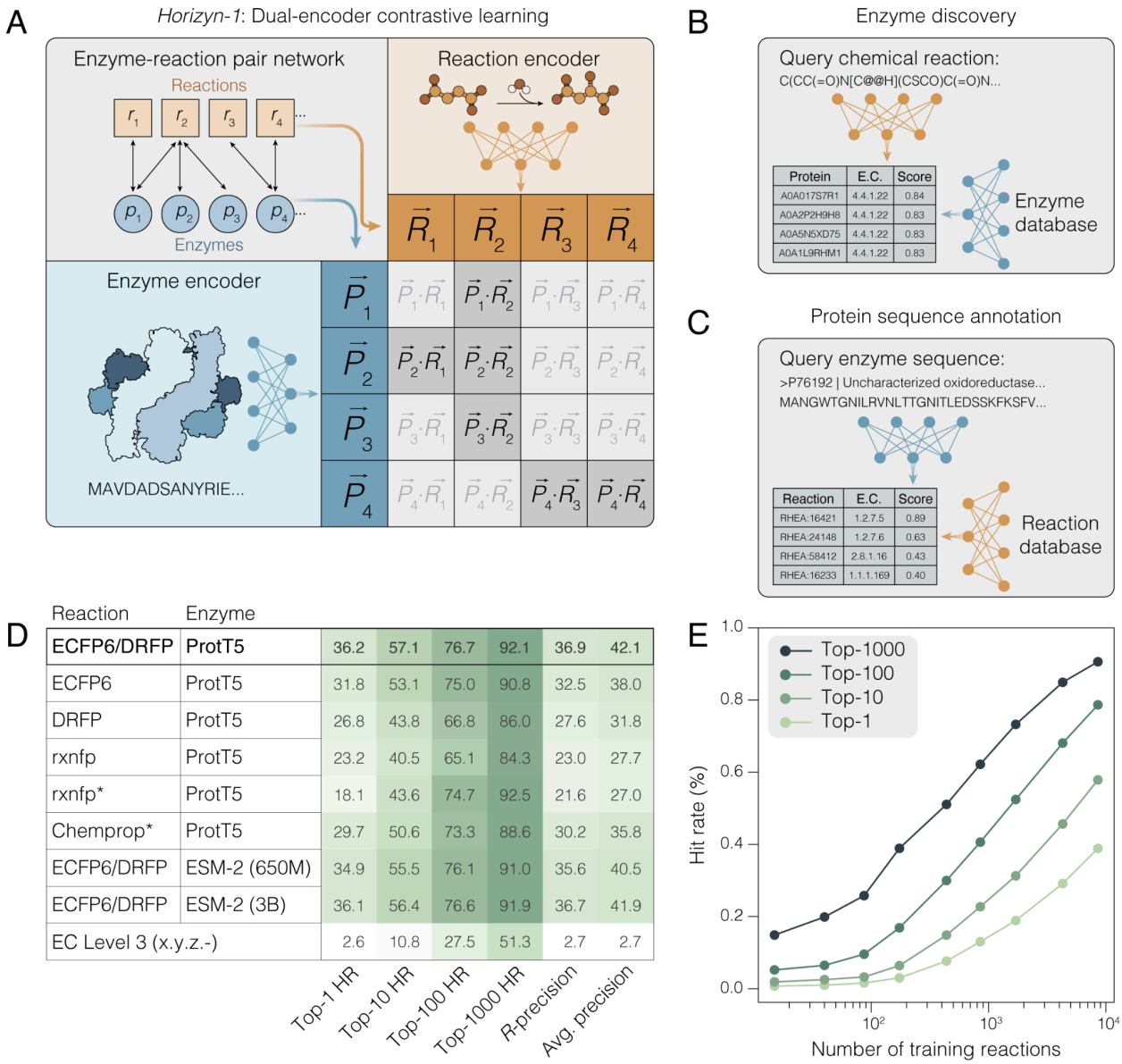


Figure 1. *Horizyn-1* model architecture, applications, and performance. (A) Schematic of the contrastive learning framework. Reactions and protein sequences are independently encoded into vector representations and projected into a shared embedding space. The model is trained to minimize the distance between known reaction-enzyme pairs (positives) and maximize the distance between unrelated pairs (negatives). **(B)** Enzyme discovery workflow: Using a query reaction, candidate enzymes are ranked by their proximity in the embedding space, where closer proximity correlates to a higher predicted likelihood of catalysis. The output is a table of enzymes ranked by their score. Enzyme Commission (E.C.) numbers are also shown for reference. **(C)** Protein sequence annotation workflow: The reverse application, where a query

enzyme is used to identify and rank potential reactions it may catalyze. **(D)** Heatmap comparing retrieval metrics, including top-*k* hit rates (HR, *k*=1, 10, 100, 1000), *R*-precision, and average precision, for different combinations of reaction and protein representations. The highest-performing model (top row) uses a combination of structural and differential fingerprints (ECFP6/DRFP) with static embeddings from the ProtT5 protein language model. *Denotes cases where reaction embedding models were fine-tuned (rxnfp) (27) or learned (Chemprop) (28) during training. **(E)** Data-scaling analysis showing a logarithmic relationship between model performance (top-100 hit rate) and the number of training reactions. This suggests that performance can be further improved with larger datasets.

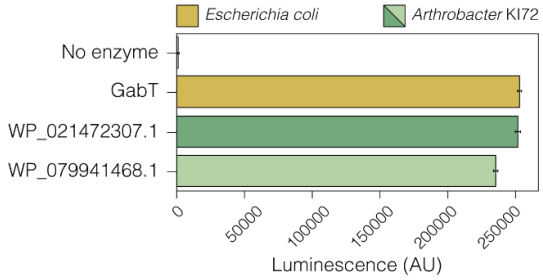
A

Nylon-eating bacteria *Arthrobacter* KI72 orphan reaction
RHEA:58200 (E.C. 2.6.1.116)

B

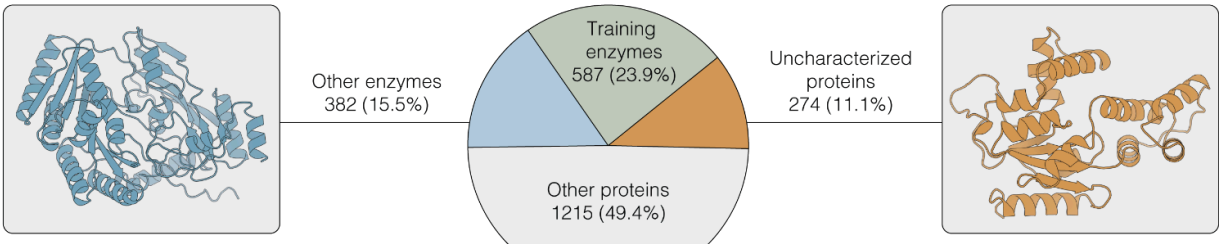
Rank	Protein ID	Predicted Annotation	Score
1	WP_021472307.1	GabT (E.C .2.6.1.19)	0.81
2	WP_079941468.1	TA (E.C. 2.6.1.-)	0.73
3	WP_200769286.1	TA (E.C. 2.6.1.-)	0.52
4	WP_021470822.1	DapC (E.C. 2.6.1.17)	0.47
5	WP_031216719.1	TA (E.C. 2.6.1.-)	0.42

C



D

T. aquaticus Proteome (NCBI Taxonomy ID: 271)



Putative D-lactate dehydrogenase (cytochrome)
A0A0N0U829

Protein of unknown function
A0A0N0BLM2

E

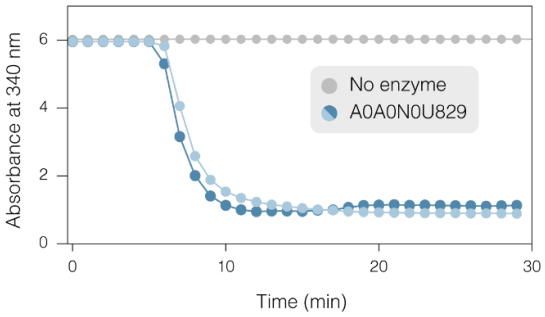
Rank	Reaction definition	Score
1	D-lactate + cyt c(ox) → pyruvate + cyt c(red)	0.88*
2	D-lactate + cyt c_{553}(ox) → pyruvate + cyt c_{553}(red)	0.88
3	D-lactate + UQ_8 → pyruvate + UQH_2	0.84
4	D-2HG + pyruvate → D-lactate + α-KG	0.83
5	D-2HG + A → α-KG + AH_2	0.77

*Pre-existing imputed annotation: E.C. 1.1.2.4

F

Rank	Reaction definition	Score
1	dCMP + ATP → dCDP + ADP	0.88
2	CMP + ATP → CDP + ADP	0.88
3	UMP + ATP → UDP + ADP	0.50
4	AMP + ITP → ADP + IDP	0.46
5	dAMP + ATP → dADP + ADP	0.43

G



H

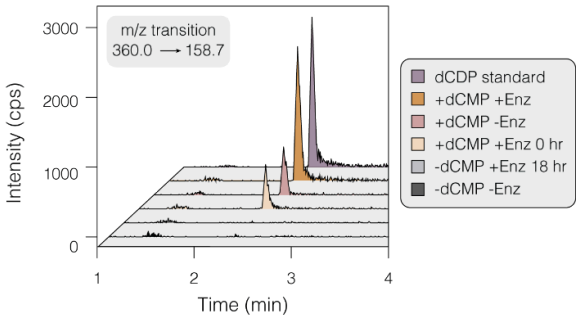


Figure 2. De-orphaning reactions and identification of promiscuous activity in uncharacterized enzymes. (A) The orphan 6-aminohexanoate aminotransferase reaction (RHEA:58200) from the nylon degradation pathway in *Arthrobacter sp.* KI72, in which 6-aminohexanoate undergoes transamination with α-ketoglutarate as the keto acid acceptor. **(B)** Top 5 candidate proteins from the *Arthrobacter sp.* KI72 proteome predicted by *Horizyn-1* to catalyze the reaction in (A), with putative functional annotations from the NCBI prokaryotic genome annotation pipeline (32). The top two hits, WP_021472307.1 and WP_079941468.1, were selected for validation. **(C)** Experimental validation of 6-aminohexanoate aminotransferase activity. Luminescence, which is proportional to L-glutamate production, is shown for the two candidate *Arthrobacter* enzymes and a known ω-transaminase, *E. coli* GabT. **(D)** Pie chart showing the distribution of enzymes, uncharacterized proteins, and other proteins in the *Thermus aquaticus* proteome (NCBI Taxonomy ID: 271). Proteins A0A0N0U829, a putative D-lactate dehydrogenase (cytochrome) and A0A0N0BLM2, a protein of unknown function, were selected for further screening. (**E**) Top 5 Rhea reactions predicted by *Horizyn-1* for the enzyme A0A0N0U829. The model confirms its annotated D-lactate dehydrogenase activity (rank 1) and predicts promiscuous activity for alternative hydride transfer cofactors including (R)-2-hydroxyglutarate–pyruvate transhydrogenase activity (D-2HG:α-KG) (rank 4). **(F)** Top 5 Rhea reactions predicted by *Horizyn-1* for the uncharacterized protein A0A0N0BLM2, suggesting dCMP and CMP kinase activity. **(G)** Experimental validation of the predicted promiscuous activity for A0A0N0U829. Pyruvate production (for the reverse reaction) is measured via a coupled lactate dehydrogenase assay. NADH concentration (absorbance at 340 nm) is depleted over time only in the presence of the enzyme (replicates, blue lines), confirming transhydrogenase activity. **(H)** Experimental validation of A0A0N0BLM2 activity using LC/MS-multiple reaction monitoring (MRM). A peak corresponding to the dCDP product is observed only when dCMP is incubated with the enzyme for 18 hours (+dCMP +Enz 18 hr), confirming dCMP kinase activity.

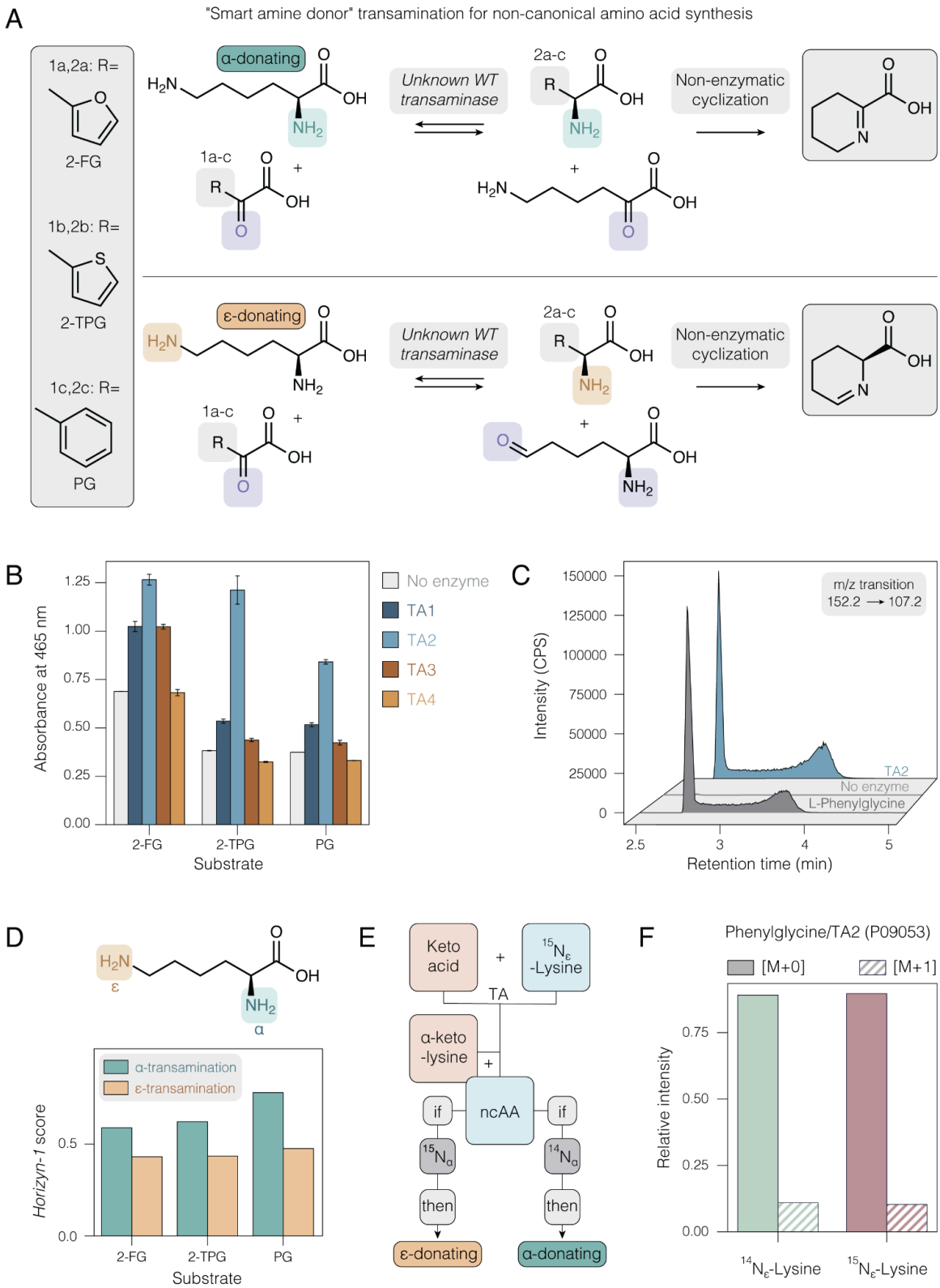

A
"Smart amine donor" transamination for non-canonical amino acid synthesis
1a,2a: R=
2-FG
1b,2b: R=
2-TPG
1c,2c: R=
PG
α-donating
ε-donating
Unknown WT transaminase
Non-enzymatic cyclization
1a-c
2a-c
B
Absorbance at 465 nm
Substrate
2-FG
2-TPG
PG
No enzyme
TA1
TA2
TA3
TA4
C
Intensity (CPS)
Retention time (min)
m/z transition
152.2 → 107.2
TA2
No enzyme
L-Phenylglycine
D
Horizyn-1 score
α-transamination
ε-transamination
E
Keto acid
15Nε -Lysine
TA
α-keto -lysine
ncAA
if
then
ε-donating
α-donating
F
Phenylglycine/TA2 (P09053)
[M+0]
[M+1]
Relative intensity
14Nε-Lysine
15Nε-Lysine

Figure 3. **Discovery and mechanistic validation of transaminases for non-canonical amino acid synthesis. (A)** Schematic of "smart amine donor" transamination for synthesizing non-canonical amino acids (ncAAs). L-lysine can act as an amine donor from either the α- or ε-position. The resulting keto-acid byproduct undergoes non-enzymatic cyclization, providing a thermodynamic driving force for the reaction. **(B)** *In vitro* activity screening of four candidate transaminases (TA1: Q9HUI9, TA2: P09053, TA3: Q9I6J2, TA4: O50131) prioritized by *Horizyn-1*. Activity was measured using a colorimetric assay with three different keto acid substrates: 2-furylglyoxylic acid (2-FG), 2-thiopheneglyoxylic acid (2-TPG), and phenylglyoxylic acid (PG). The enzyme TA2 (P09053) demonstrated the highest activity. **(C)** Confirmation of L-phenylglycine production by TA2 via multiple reaction monitoring (MRM). The extracted ion chromatogram for the reaction product (blue trace, *m/z* = 107.2) matches the retention time of an authentic L-phenylglycine standard (gray trace). CPS = counts per second. (**D**) *Horizyn-1* scores for α- vs. ε-amino-driven transamination for TA2. For all three keto acid substrates, the model predicts that the α-donation pathway is strongly favored. (**E**) Experimental design to distinguish between the two potential mechanisms using lysine labeled with a heavy isotope at the epsilon amine ($^{15}N_{\varepsilon}$-labeled L-Lysine). Label incorporation into the final ncAA product would indicate that the reaction proceeds via ε-transamination. (**F**) Relative intensity of unlabeled (M+0) and labeled (M+1) L-phenylglycine produced by TA2 using either standard $^{14}N_{\varepsilon}$-labeled L-Lysine or $^{15}N_{\varepsilon}$-labeled L-Lysine. The relative intensities are computed from summed ion intensities from multiple reaction monitoring (MRM) fragments. The result, showing negligible label incorporation from $^{15}N_{\varepsilon}$-labeled L-Lysine, confirms that transamination occurs predominantly from the unlabeled α-amine, validating the model's prediction in **(D)**.

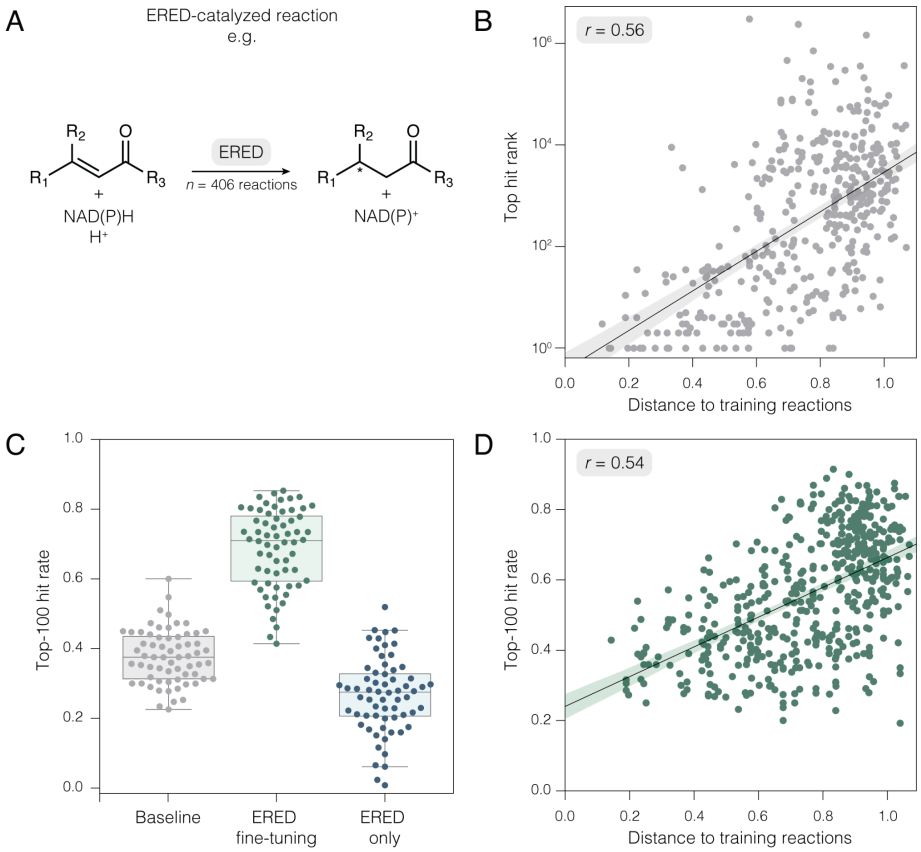


Figure 4. **Fine-tuning boosts performance on novel reaction classes. (A)** Representative ERED-catalyzed reaction involving NAD(P)H-dependent reduction of a C=C bond. This reaction class was not present in the original training data and was used to simulate discovery in a new area of chemistry. **(B)** Baseline model performance on the novel ERED reactions. The rank of the top positive hit (y-axis) for each reaction is plotted against its distance from the training set (x-axis), measured via the minimum distance to any reaction in the training set within the model's shared embedding space. Model performance degrades as reactions become more novel. **(C)** Fine-tuning on just 10 new ERED reactions provides a substantial performance boost, increasing the median top-100 hit rate from 38% (baseline) to 71%. Both of these results are a significant improvement over a model trained only on ERED data, which had a median hit rate of 28%. **(D)** Fine-tuning performance (top-100 hit rate) after training on a single ERED reaction plotted against its distance to the training set. The strong correlation demonstrates that fine-tuning on more novel reactions yields greater performance improvements, suggesting an active learning approach to prioritize data collection. In panels (B) and (D), *r* corresponds to the Pearson correlation coefficient with 95% confidence intervals represented as shaded regions.

Materials and Methods

Software availability

Horizyn-1 will be made available for non-commercial use via a hosted interface following peer review. Due to proprietary restrictions, the source code is not publicly released. However, a detailed description of the model architecture, training setup, and inference workflow is provided in the Methods section. Key algorithmic steps are included in the Supplementary Information.

Datasets and Preparation

Development Dataset

Our dataset used for model development was curated by combining Rhea (release 131, (21), a collection of expert-curated biological reactions linked to enzymes, and Swiss-Prot (release 2023_05), (42), the manually-reviewed, high-quality subset of UniProt. We selected reaction-enzyme pairs from Rhea where the associated enzyme sequence was present in Swiss-Prot. We further excluded any Rhea reactions lacking valid SMILES representations for all participants. The final dataset used for development contains 11,932 unique reactions, 216,132 unique protein sequences, and 294,166 reaction-enzyme pairs.

Inference Dataset

To construct the dataset for training our final inference model, we expanded upon the development dataset by incorporating two additional sources. First, we added EnzymeMap v2 (43) (44), which contributed 19,140 reactions, 15,308 protein sequences, and 55,163 reaction-enzyme pairs primarily curated from BRENDA. In contrast to Rhea, EnzymeMap contains many reactions that are not naturally occurring. Second, we augmented the protein sequences using the TrEMBL subset of UniProt (release 2023_05), (42), adding 26,883,845 computer-annotated sequences. While generally less reliable than Swiss-Prot, TrEMBL introduces significant sequence variation and provides additional indirect links between Rhea reactions lacking overlapping Swiss-Prot enzymes. Finally, we clustered the combined set of all protein sequences at 80% similarity using MMseqs2 (45) (https://github.com/soedinglab/MMseqs2) via the command

```
mmseqs easy-cluster <input_file> <output_prefix> <temp_directory> -c 0.8
--cov-mode 0 --min-seq-id 0.8
```

This step substantially reduced sequence redundancy and created further indirect reaction links via enzymes with highly similar sequences. The resulting inference dataset comprises 31,101 unique reactions, 7,063,237 clustered protein sequences, and 8,897,870 reaction-enzyme pairs.

Enzyme Representation Dataset

To compare different enzyme embedding schemes, we used a smaller version of our development dataset where we clustered sequences at 80% similarity using MMseqs2 (45) . For models that required protein structures, we downloaded 3D coordinates from the AlphaFold Database (46, 47).

CLIPZyme Benchmark Dataset

For comparison with CLIPZyme (17), we used the specific version of the EnzymeMap dataset provided by the authors (http://doi.org/10.5281/zenodo.11187747). This dataset contains reaction-enzyme pairs and a separate screening set of unpaired enzymes. As the original data splits were not included, we used the authors' provided code (https://github.com/pgmikhael/clipzyme) to replicate their training, development (dev), and test splits as closely as possible.

During this replication, we encountered several minor discrepancies compared to the statistics reported in the original publication (reported below). Despite these differences, we believe the comparison to be compatible, and the leakage not to significantly affect the published results.

- **Reaction Splits:** The reported train/dev/test reaction counts (12,629/2,669/1,554) differed from our observed counts (12,603/2,661/1,551).
- **Screening Set Size:** The reported size (260,197) differed from the observed size (261,907).
- **Empty Sequences:** We identified 72 empty sequences within the screening set, likely due to deprecated UniProt entries. Removing these reduced the effective screening set size to

261,835 unique sequences. Some of these missing/empty sequences also affected the paired reaction data, resulting in reduced train/dev/test pair counts (from 34,427/7,287/4,642 to 34,265/7,285/4,638) and reaction counts (12,569/2,661/1,551).

- **Data Contamination:** Crucially, we note that the dataset was split based on reaction-enzyme *pairs*, not solely on reactions. This resulted in data leakage, with 9 training reactions appearing in the dev set and 30 training reactions appearing in the test set.

Ene-Reductase (ERED) Dataset

The ene-reductase dataset was compiled from reactions available in the BRENDA database (E.C. classes 1.3.1.-, 1.6.99.-) (48) and the public version of the RetroBioCat ERED dataset (49). This collection was supplemented with additional manual curation from primary literature sources. This dataset totaled 543 reactions, 299 proteins, and 1136 pairs. To ensure reactions were distinct from the inference dataset, they were filtered to remove those with identical fingerprints (see Reaction Clustering and Data Splitting). This process resulted in a final dataset of 406 reactions, 221 proteins, and 903 pairs.

Genome annotation

Protein coding sequences form *Arthobacter* sp. KI72 (*Paenarthrobacter ureafaciens* KI72), a nylon-eating bacteria, were obtained from the NCBI ftp server using the accession GCF_002049485.1, with annotations derived from the NCBI Prokaryotic Genome Annotation Pipeline (32). Protein coding sequences (and corresponding E.C. annotations) for *Thermus aquaticus* were obtained from the Uniprot reference proteome for Taxonomy ID: 271. For enzyme promiscuity analysis, we screened all enzymes with no annotation in Rhea (release 131), and had at most 50% sequence similarity to enzymes in our training data. Sequence similarity was computed using MMseqs2 (45) using default settings.

Reaction SMILES Standardization and Augmentation

To ensure representation consistency across data sources and compatibility with the reaction fingerprinting algorithms we investigated, we used RDKit (50) to standardize all reaction SMILES strings in the development, inference, and ene-reductase datasets with the following steps:

1) Convert hypervalent N, P, Cl, Br, I to zwitterionic structures (rdmolops.SanitizeMol with SANITIZE_CLEANUP flag)
2) Neutralize acidic and basic sites by adjusting ionizable protons (rdMolStandardize.Uncharger)
3) Detach metals by converting covalent metal–heteroatom bonds to ionic bonds (rdMolStandardize.MetalDisconnector)

We found these steps to be sufficient for reactions sourced from Rhea and EnzymeMap. However, many of the ene-reductase-catalyzed reactions were manually curated directly from literature, resulting in many cases where the same reaction was represented with different choices of SMILES. Therefore, we applied two extra standardization steps for this dataset:

4) Generate canonical tautomer (rdMolStandardize.TautomerParentInPlace)
5) Kekulize aromatic structures (rdmolops.Kekulize)

Because many of the reaction representations we investigated were not invariant to reaction direction, we augmented reactions by including separate copies of each reaction proceeding in both the forward and backward directions. This resulted in each reaction-enzyme pair being duplicated with both directions during training and for computing performance metrics. We always included both copies of each reaction in the same split during model training and evaluation, to avoid contamination.

Reaction Similarity Clustering and Data Partitioning

The strategy for splitting data into training, validation, and test sets is crucial for obtaining realistic performance estimates. We advocate for splitting datasets at the reaction level, rather than, for example, splitting on reaction-enzymes pairs or enzymes. This approach directly reflects the real-world use case where models must generalize to entirely new, unseen classes of reactions.

A key challenge in reaction splitting is avoiding data contamination, where reactions highly similar to those in the training set appear in the validation or test sets. Such leakage can lead to inflated performance metrics that do not translate to real-world utility. To mitigate this, we cluster reactions based on their similarity and ensure that all reactions within a cluster are assigned to the same data split.

Several methods have been employed historically for assessing reaction similarity for clustering. Template-based approaches, such as those used in methods like CLIPZyme (17), attempt to identify conserved reaction patterns. However, templates may not be available for all reaction types, and a single reaction can sometimes be assigned multiple templates, leading to ambiguity. Another approach, employed by methods like Reactzyme, uses fuzzy string matching on reaction SMILES strings. While more general, this method can be susceptible to inconsistencies arising from the canonicalization process of SMILES representations.

Given these limitations, we used a historically standard and robust method: representing reactions with structural reaction fingerprints (24) that concatenate extended-connectivity fingerprints (ECFP6, also known as Morgan fingerprints) (23) of the reactants and products and quantifying their similarity using the Tanimoto coefficient. This approach is broadly applicable to diverse reaction types and exhibits stable behavior. To ensure the robustness of our splits and understand the model's generalization capabilities, we systematically analyzed the validation error as a function of the Tanimoto similarity threshold used for clustering during the data splitting process (see Fig. S9 and Reaction Similarity Clustering). We also compared our fingerprint-based clustering to a separate template-based approach.

Structural Fingerprint Clustering Protocol

Our default clustering protocol quantifies similarities between reactions by computing Tanimoto (Jaccard) coefficients between their structural fingerprints. First, we applied our standardization protocol to all reaction SMILES (see Reaction SMILES Standardization and Augmentation). Next, we used the RDKit library (50) to generate Morgan 2048-bit extended-connectivity fingerprints with radius 3 (ECFP6) for each molecule via the *rdFingerprintGenerator.GetMorganGenerator* class. We enabled options to incorporate chirality, bond orders, and ring membership information. For each reaction, we consolidated the reactant

and product fingerprints into separate binary vectors using a bitwise OR operation and then concatenated these vectors to create a single 4096-bit reaction fingerprint (24). The concatenation order reflects the reaction direction, with the reactant fingerprint preceding the product fingerprint. To ensure directional invariance, we computed Tanimoto coefficients between two reactions for both the forward and reverse directions of each, taking the maximum value from the four possible comparisons. We grouped reactions with a Tanimoto coefficient of 1.0 into the same cluster. Finally, we randomly assigned these clusters to the training, validation, or test sets based on the desired split ratio.

Template Clustering Protocol

In order to create a split of the Rhea dataset by reaction templates, we generated atom mappings between reactants and products for all reactions using the RXNMapper transformer model (51). Subsequently, we constructed the condensed graph of reaction (also referred to as "imaginary transition state") (52, 53), of which we retained only the atoms and bonds involved in bonding changes. The retained substructures of the reactants and products were saved as reaction templates in SMARTS notation. Due to failures of atom mapping, reaction templates were constructed for 10,473 of the 11,932 Rhea reactions.

Model Architecture and Training Protocol

Reaction Representations

Extended-connectivity Fingerprints (ECFP6)

To capture molecular structure, we generated extended-connectivity level 3 fingerprints (ECFP6, Morgan) (23) as specified in Structural Fingerprint Clustering Protocol. For use in *Horizyn-1*, we used fingerprints composed of 1024 bits, with 512 bits allocated for each of the reactant and product subcomponents.

DRFP Fingerprints

To incorporate the approximate reaction mechanism and bond changes, we employed a different fingerprint known as DRFP (Difference Reaction Fingerprint) (25). We used the default DRFP

parameters (min_radius=0, radius=3, rings=True, root_central_atom=True, include_hydrogens=False), with the exception of setting the output fingerprint dimension to 1024 bits.

ECFP6/DRFP Composite Fingerprints

We constructed composite reaction fingerprints by concatenating the ECFP6 and DRFP fingerprints described above. This allowed us to capture both molecular structure and reaction transformations. The final reaction fingerprint used in our study is the direct concatenation of the 1024-bit ECFP6 structure fingerprint and the 1024-bit DRFP transformation fingerprint, yielding a total size of 2048.

rxnfp

The rxnfp model (27) is a pre-trained BERT-style transformer encoder. This model was trained on 2.6 million reactions from the Pistachio dataset using a Masked Language Modeling (MLM) objective, where it learned to predict masked tokens within reaction SMILES strings.

For our dual contrastive model, we used the pre-trained rxnfp model to process reaction SMILES strings. Following the fine-tuning methodology proposed by the original authors, we extract embeddings from the final classification token ([CLS]) of the transformer output. This process yields a single vector representation for each reaction, which serves as the input embedding for the subsequent Multi-Layer Perceptron (MLP) head in our model. We investigated two strategies: using the frozen pre-trained weights and fine-tuning (unfreezing) the rxnfp weights during the training of our contrastive model.

Chemprop

Chemprop (28) is a graph neural network model based on the Directed-Message Passing Neural Network (D-MPNN) architecture, originally developed for predicting molecular properties. While often applied to single molecules, Chemprop's framework can be adapted for reaction modeling, particularly as it can handle multi-component systems.

In our implementation, we use a Chemprop style architecture consisting of three layers of directed message passing. To obtain a fixed-size representation for each reaction system (reactants and products), we aggregated the node features learned by the message-passing layers using a global pooling operation (e.g., summing node features) followed by an MLP projection. This results in a single vector representation in the shared embedding space. This Chemprop model was trained from scratch using our reaction dataset as part of the end-to-end training process for the contrastive model.

Enzyme Representations

To generate enzyme vector embeddings, we employed two pre-trained protein language models (pLMs): ProtT5 and ESM-2. Inspired by transformer models in the natural language processing domain, pLMs treat protein sequences analogously to text with amino acids acting as the vocabulary. Both models were trained on a Masked Language Modelling (MLM) objective, where the goal is to predict the identity of masked amino-acid tokens from their context. This relatively simple self-supervised training objective, when applied to a diverse collection of sequences, enables the models to learn rich embeddings of both structure and function.

ProtT5

ProtT5 is a 3-billion-parameter encoder-decoder transformer pre-trained on large sequence datasets totaling over 2.1 billion sequences (26). To create a fixed-size (1024-dimension) embedding for each protein, we extracted the final decoder output embeddings for each token (residue) and applied mean pooling. Sequences longer than the model's limit (5,000 residues) were processed in chunks, and their embeddings were averaged. We precomputed the embeddings before training for efficiency.

ESM-2

The ESM-2 family of models uses encoder-only (BERT-style) transformers pre-trained on over 65 million sequences (29) . For this model, we explored multiple configurations, varying the model size, the embedding extraction strategy, and the training approach (using pre-computed embeddings vs. end-to-end fine-tuning).

Pre-computed [CLS] Token Embeddings

This strategy used static, pre-computed, embeddings generated from the final hidden state of the [CLS] token, which aggregates sequence-level information. We generated these embeddings for both the 650M and 3B parameter versions of ESM-2. Sequences longer than the 1,500-residue input limit were truncated.

Live Integration with Mean-Pooling

This approach used the ESM-2 (650M) model. Sequence embeddings were generated by mean-pooling the final hidden states of all tokens, which were then passed to a final MLP layer. We investigated two training schemes: one with the base ESM-2 weights frozen and another where the entire model was fine-tuned.

Live Integration with Structure-Biased Attention Pooling

To incorporate 3D structural information, we augmented the ESM-2 (650M) model with a structure-aware attention layer inserted before the final pooling step. This added layer used a hidden dimension of 1280 (matching the base model), a single attention head, and a 4x expanded feed-forward dimension of 5120. The self-attention mechanism is biased by the inverse pairwise Euclidean distances between residues, calculated from AlphaFold 3-predicted structures. This design encourages attention between spatially proximal residues while maintaining SE(3) invariance. This architecture was also tested with both frozen and fine-tuned base model weights.

Dual Encoder Architecture

Our model maps reaction and enzyme representations into a shared embedding space via two multilayer perceptrons (MLPs) that act in parallel. Each MLP accepts representation-specific input dimensions and uses two 4096-dimensional hidden layers with ReLU activation. The resulting embedding vectors are L2-normalized, and cosine similarity (dot product) is used to compute their distances.

Contrastive Learning Loss

See SI note (16, 54, 55)

General Training Protocol

All models were trained using PyTorch v. 2.0.0 (56) and Pytorch Lightning v. 2.3.3 (57). We employed the AdamW optimizer with parameters $(\beta_1, \beta_2) = (0.9, 0.999)$, a weight decay of 10^{-2}, and a fixed learning rate of 5×10^{-5}. A default batch size of 16,384 was used unless otherwise specified. With the exception of the ESM-2 protein language model, all models were trained on a single NVIDIA T4 GPU.

To ensure consistent batch sizes, crucial for the MLNCE loss (see Contrastive Learning Loss), the final batch of each epoch was padded to the required size. Padding involved randomly duplicating reaction-enzyme pairs from the training data.

We used a fixed inverse temperature parameter β=10 to set the distance scale of our loss function (see Contrastive Learning Loss). While the original CLIP implementation treats this as a learnable parameter, we observed slightly improved results and more stable training dynamics by fixing its value.

Training duration was determined based on validation set performance, approximately targeting epochs after the point where the top-100 hit rate on a held-out validation set began to plateau, but before the onset of overfitting.

Model Evaluation

Model performance was assessed using an evaluation set containing reactions, associated enzymes, and a broader screening set. For each reaction, the primary evaluation measured how effectively the model ranked known catalyzing enzymes (true positives) higher than other enzymes within a screening set. For a particular reaction, positive enzymes were those derived from reaction-enzyme pairs involving that reaction.

Performance was quantified using several standard information retrieval metrics. Each metric was averaged over each reaction, proceeding in both the forwards and backwards directions.

Top-k hit rate, which measures whether at least one of the true positives appears within the top k ranked results returned by the model, was chosen to reflect the practical, real-world use case of identifying relevant enzymes from a large pool with k values of 1, 10, 100, and 1000.

R-precision and average precision offer more stringent assessments of retrieval quality. R-precision is the strictest metric, measuring the fraction of a reaction's true positive enzymes found within the top R ranked results, where R equals the total number of true positives for that reaction. Average precision provides a less strict, but more comprehensive measure of how highly positive enzymes are ranked relative to others across the entire list, offering greater sensitivity to the ranking of all relevant items and being less susceptible to class imbalance issues compared to metrics like AUROC.

Additionally, Boltzmann-Enhanced Discrimination of Receiver Operating Characteristic (BEDROC) and Enrichment Factor (EF) were included, primarily for comparison purposes with the CLIPZyme results. BEDROC is a metric that heavily weights the ranking of true positives found early in the list, assessing initial retrieval performance. Enrichment Factor (EF) measures the concentration of true positives within a specific top fraction of the ranked list compared to what would be expected from a random ranking. For consistency with the CLIPZyme results, we report BEDROC scores with α equal to 20 and 85, and EF scores for the top 5% and 10% of predictions.

Protein Language Model Compute Cost

A key consideration in model development is the trade-off between performance and computational efficiency. Large Transformer-based models, like protein language models (pLMs), can demand orders of magnitude more compute resources for training or fine-tuning compared to smaller architectures such as Multi-Layer Perceptrons (MLPs). We therefore analyze the relative performance of our different protein embedding strategies against their estimated computational cost, measured in Floating Point Operations (FLOPs).

We compare two main paradigms for using pLMs to generate protein embeddings (see Enzyme Representations):

1. **Pre-computed Embeddings:** This approach involves a one-time process where a pre-trained pLM generates fixed embeddings for the entire protein dataset. These static embeddings are then reused across all training epochs for downstream tasks. While the initial embedding generation incurs a significant one-time cost, this is amortized over the training duration. Numerically, this is equivalent to using a fully frozen pLM during training but is often more efficient in practice.
2. **Live pLM Integration:** This method keeps the pLM weights loaded in memory during training, allowing gradients to propagate through the model. This enables fine-tuning of the pLM weights for a specific task, as well as providing access to per-residue embeddings where applicable, but substantially increases the computational cost per training step due to the large number of active parameters in every forward and backward pass.

Larger models inherently require more FLOPs per training step and often necessitate longer training durations and more powerful hardware, impacting both financial costs and the speed of experimental iteration. Furthermore, the considerable memory footprint of models like ESM-2 (650M) often compels the use of smaller batch sizes compared to models trained on pre-computed embeddings. For instance, our ESM-based models were limited to a batch size of 32. In contrast, models using pre-computed embeddings could potentially leverage much larger batch sizes (e.g., 16k, as used for a reference model comparison), which can also influence training dynamics and final performance.

Furthermore, the contrastive loss function has a unique complication with respect to batch size. Typically where the data points can be considered drawn from an i.i.d distribution, a desired global batch size can easily be achieved through combining batches via gradient accumulation and data parallel training despite small local mini-batch sizes. However, this averaging cannot be employed so trivially with standard contrastive losses that employ random negative sampling, as all pairs in a batch contribute to the denominator of each loss term. An online-MLNCE loss function, drawing from the online-softmax implementations is an area for further work.

Given these trade-offs, a central question emerges: Does the potential performance gain from unfreezing the pLM, or from adding architectural modifications like a structure-aware layer, justify the dramatically increased computational expense?

To address this, we compare the performance-per-FLOP across several model configurations:

- **Pre-computed Embeddings:** ESM-2 (650M), ProtT5-XL-U50 (referred to as ProtT5 3B), and ESM-2 (3B). Embeddings were generated once; downstream models were trained using these fixed embeddings. We tested batch sizes of 32 (matching the live models) and 16k (for reference) where applicable. These models were trained for 100 epochs.
- **Live ESM-2 (650M) Models:** These models were trained with a batch size of 32:
    - Frozen base model (trained for 45 epochs).
    - Unfrozen base model (fine-tuned for 10 epochs).
    - Frozen base model + structure-aware layer (trained for 60 epochs).
    - Unfrozen base model + structure-aware layer (fine-tuned for 60 epochs).

By translating the training epochs for each model into an estimated total FLOP count, we establish a fairer comparison based on a fixed computational budget. All models were trained on a special version of our development dataset (see Enzyme Representation Dataset and ESM-2 Live Integration and Fine-tuning). The results, presented in Fig. S1, illustrate the top-1, top-10, and top-100 hit rates on the validation set plotted against cumulative FLOPs, enabling a direct comparison of efficiency across strategies.

Unsurprisingly, models using pre-computed embeddings achieve the best final performance for the lowest computational cost. When comparing the live ESM-2 models, the frozen configurations generally yield better performance per FLOP than their unfrozen counterparts for the more challenging top-1 and top-10 hit rate tasks, although both significantly lag behind the pre-computed embedding approach in overall efficiency. However, the unfrozen models demonstrate competitive performance for the top-100 hit rate metric relative to their FLOP cost. Ultimately, the performance of all live ESM-2 models (frozen or unfrozen) appears constrained by the small batch size (32) necessitated by memory limitations, compared to the much larger batch size (16k) feasible with pre-computed embeddings.

Model Confidence Heuristic

To provide a practical framework for interpreting model outputs in experimental settings, we established a qualitative confidence heuristic based on the *Horizyn-1* score distributions from the inference model training set (Fig. S2C). This approach categorizes predictions into three confidence tiers based on the score distributions of positive and negative enzyme-reaction pairs. The tiers are defined as follows:

- **Low Confidence (Score < 0.10):** Scores falling below the 99th percentile of the negative pair distribution.
- **Moderate Confidence (Score 0.10 to 0.72):** Scores between the 99th percentile of negative pairs and the 1st percentile of positive pairs.
- **High Confidence (Score ≥ 0.72):** Scores at or above the 1st percentile of the positive pair distribution.

This heuristic is intended to guide the prioritization of candidates for experimental screening.

Additional Modelling Study Details

Development Model Training and Evaluation

The development model was trained on the development dataset (see Development Dataset), derived from Rhea and Swiss-Prot. The training and test sets were created by splitting on reactions and subsequently filtering the test set to exclude reactions that had fingerprint similarities of 1.0 with any reactions in the training set (see Reaction Similarity Clustering). The final training and test splits contained 10,785 and 1,012 reactions, respectively. The model was trained for 100 epochs with batch sizes of 16,384 (default), 256 (Chemprop), and 16 (unfrozen rxnfp). For evaluation, the model identified the correct enzyme for each test reaction from a screening pool of all proteins in the development dataset. Results are shown in Fig. 1D.

E.C. Conditioning Analysis

For the E.C. number conditioning analysis (Fig. 1D, Table S1), we first assigned E.C. numbers to each protein based on the Rhea annotations of its associated reactions. We included reactions and

enzymes in our retrieval metrics even if they did not have any assigned E.C. numbers. When scoring proteins for a given query reaction, we assigned a score from 0 to 4, equal to the number of matching digits in their respective E.C. numbers. For the E.C.-only baseline, ties were broken randomly, and final retrieval metrics were averaged over 10 such permutations. To condition the performance of our model, we combined the scores using the formula model score + 2 * E.C. score, essentially using our model to break ties for the E.C. scores.

Training Set Size Scaling Analysis

To analyze the impact of training set size (Fig. 1E), we first subdivided the training subset of the development dataset (see Development Model Training and Evaluation) into a fixed random 80%/20% training/validation split based on reaction fingerprint clusters (see Reaction Similarity Clustering). To create datasets for each run, the 80% training portion was then systematically subsetted by reaction clusters to achieve the desired training set sizes. All models were trained for 100 epochs and evaluated on the same fixed validation set, using the full development dataset's protein set as the screening pool.

Reaction Similarity Clustering Analysis

To assess how the difficulty of reaction train/test splits impacts model performance (Fig. S9), we used the training portion of the development dataset (see Development Model Training and Evaluation). This dataset was subdivided into approximate 80%/20% training/validation splits using different random seeds for each replicate. These splits were generated based on either fingerprint clusters at a defined threshold or reaction templates (see Reaction Similarity Clustering and Data Partitioning). For the reaction template-based splits, reactions that could not be mapped to a valid template were assigned to the training set; the remaining reactions were then randomly divided to meet the 80%/20% ratio. All models were trained for 100 epochs and evaluated using the corresponding validation set and all proteins from the full development dataset as the screening set for reaction queries.

ESM-2 Live Integration and Fine-tuning

For models incorporating ESM-2 (650M) live integration (see ESM-2, Protein Language Model Compute Cost, Fig. S1), we used the enzyme representation dataset. This dataset, which is a

version of the development dataset with sequences clustered at 80% similarity (see Enzyme Representation Dataset), was subdivided into a fixed 80%/20% training/validation split based directly on reactions, rather than reaction fingerprint clusters.

Training durations varied by model type:

- The mean-pooling variant with a frozen ESM-2 base was trained for 45 epochs.
- The mean-pooling variant with fine-tuning was trained for 10 epochs.
- Models with structure-biased attention pooling were trained for 60 epochs.

All 650M parameter model training used 4 NVIDIA A100 GPUs with DeepSpeed ZeRO-2 for distributed training and memory management. This approach balanced memory savings with communication volume, which was especially important during full fine-tuning. Finally, models were evaluated on the validation set, using all proteins from the development dataset (clustered at 80% sequence similarity) as the screening set.

CLIPZyme Benchmarking

For CLIPZyme benchmarking (Table S2), the CLIPZyme benchmark dataset was used (see CLIPZyme Benchmark Dataset). This dataset came with a predefined 80%/10%/10% train/dev/test split. Models were trained for 30 epochs and evaluated on the test split and the screening set provided within the CLIPZyme Dataset.

Inference Model

The inference model was trained using the inference dataset, derived from Rhea, EnzymeMap, Swiss-Prot, and TrEMBL (see Inference Dataset). For this model, the entire dataset was used as the training set, with all proteins used for the screening set. The model was trained for 200 epochs.

Ene-Reductase Fine-tuning

The fine-tuning studies (Fig. 4) used a manually curated ene-reductase dataset (see Ene-Reductase Dataset). The data splits for fine-tuning were prepared using the following procedure:

1) Randomly split the ene-reductase dataset into a 90%/10% train/val split based on reaction fingerprint clusters (see Structural Fingerprint Clustering Protocol).
2) Randomly select the desired number of reaction clusters from the training split. Results in Fig. 4C and Fig. 4D used 10 and 1 reaction clusters, respectively.
3) Add an additional 10 random reaction-enzyme pairs from the inference dataset to the training split, with each reaction belonging to a different fingerprint cluster. These additional pairs served to regularize the model during fine-tuning and prevent total collapse of the shared embedding space.

Each training replicate used a different random seed to randomize each of these steps. Model evaluation was performed on reaction-enzyme pairs from the validation split, using a comprehensive screening set that combined all proteins from both the inference and ene-reductase datasets, totaling 7,063,403 proteins. Training for fine-tuning was conducted for 10 epochs. Models were initialized either from scratch with random weights or with the weights of the previously established inference model for fine-tuning. The batch size was set to twice the number of pairs in the final training split; the factor of two accounted for reaction direction augmentation.

Experimental procedures

Chemical Reagents

The compounds α-ketoglutaric acid sodium salt, lysine, $MnSO_4$, $Co_3(PO_4)_2$, $MgCl_2$, $ZnCl_2$ (R)-lactate, L-Lactate Dehydrogenase, NADH, UMP, and CMP were purchased from Millipore Sigma. Additional chemicals including 6-aminohexanoic acid, monosodium L-glutamate, ATP, PLP, o-aminobenzaldehyde, and 2-thiopheneglyoxylic acid were also sourced commercially (VWR).

Cloning and transformation

The native E. coli GabT gene (Genbank M88334) encoding the Uniprot P22256 sequence was trimmed to remove the start and stop codons and ordered from Twist Biosciences as a clonal

gene inserted between the NdeI and HindIII restriction sites in pET29b(+). The insertion resulted in a full length transcript containing an additional seven amino acid linker (KLAAALE), C-terminal his tag, and stop codon encoded in the plasmid sequence.

DNA sequences for GabT transaminase homologs from *Arthrobacter* sp. KI72 (NCBI sequences WP_021472307.1 and WP_079941468.1), Lysine TAs (UniProt:Q9HUI9, P09053, Q9I6J2, O50131), Transhydrogenase from *Thermus aquaticus* (UniProt:A0A0N0U829), and dCMP kinase from *Thermus aquaticus* (UniProt:A0A0N0BLM2) were synthesized without the start and original stop codon from Twist Biosciences and modified to contain a C-terminal GGSGSG linker, 6-his tag, and stop codon. Full ORFs were codon optimized for expression in *E. coli*. These sequences were delivered as clonal constructs in pET29b(+) with gene insertion at the NdeI and XhoI sites.

Protein overexpression and purification

Plasmids harboring GabT, GabT transaminase homologs, Lysine TAs, and dCMP kinase were transformed into chemically competent BL21(DE3) cells (Agilent Technologies). Plasmid harboring transhydrogenase was transformed into chemically competent OverExpress C41(DE3) cells (Sigma-Aldrich). Cells were plated on LB agar containing 50 ug/ml kanamycin for overnight growth. Single colony transformants were grown aerobically overnight at 37 °C in 5 mL of LB medium supplemented with 100 µg/mL kanamycin. Cells from the starter cultures were added to 100 mL LB with 100 mg/L kanamycin and cultivated at 37 °C until OD600 reached 0.6-1.0. Expression was induced with 0.2 mM IPTG for 16 hr at 25 °C for GabT. For other enzymes, expression was induced with 0.2 mM IPTG for 16 hr at 16 °C. Cells were then harvested by centrifugation (e.g., 8000 x g, 15 min, 4 ∘C) and stored at -20 °C until use.

Enzymes were purified through Ni-NTA gravity column purification. Cell pellets were thawed and resuspended in one-fifth the volume of the culture with Lysis Buffer (20 mM Tris, pH 7.5, 0.5 M NaCl, 100 µM PLP, 10 mM imidazole, protease inhibitor cocktail (Roche), DNase, and lysozyme) and then lysed by sonication on ice using a microtip on a Branson sonifier at 70% amplitude with five 1 s on 1 s off cycles, pausing for 30 s on ice between cycles. The resulting lysate was centrifuged for 30 minutes at 20,000 x g. Clarified lysate was collected and incubated

with 500 µL of washed Ni-NTA beads (ThermoFisher) per manufacturer's instructions with rocking on a nutator for 30 minutes.

Lysate was loaded onto a gravity column and the resin was washed with 10 CV of Wash Buffer (20 mM Tris pH 7.4, 0.5 M NaCl, and 20 mM imidazole). The protein was eluted with 10 CV of Elution Buffer (20 mM Tris pH 7.4, 0.5 M NaCl, and 250 mM imidazole). Fractions were collected and analyzed by SDS-PAGE (GenScript). Purified enzymes were dialyzed with 50 mM potassium phosphate pH 8.0 with a 30 kDa MWCO spin column (Cytiva) and stored in 12.5% glycerol at -80 °C until use. The concentration of purified enzyme was determined by the Pierce BCA assay (Thermo Fisher) per manufacturer's protocol.

Enzymatic reaction and assays

Transaminase Assay: *E. coli* GabT and GabT transaminase homologs from *Arthrobacter* sp. KI72 were tested for aminotransferase activity using the amino donor 6-aminohexanoic acid and the amino acceptor 2-oxoglutarate. Reactions were assembled containing 50 mM sodium phosphate, pH 8.0, 1 mM PLP, 100 mM 2-oxoglutarate, and 50 mM of amino donor (6-aminohexanoate, β-alanine, or putrescine). To initiate the reaction, purified enzyme was added at a final concentration of 10 µM. For negative control, purified enzyme was replaced with the same volume of 50 mM sodium phosphate buffer. Reactions were performed at room temperature for approximately 16 hours. The formation of glutamate in transaminase enzymatic assays was detected by using the Glutamate-Glo™ Assay (Promega) following the manufacturer's protocol. A luminescence signal indicating glutamate formation was measured by a POLARStar Omega plate reader (BMG Labtech). All data points were measured in duplicates.

Lysine Transaminase Assay: Ability to transfer an amino group from lysine to an oxoacid to form a non-canonical amino acid was determined for Q9HUI9, P09053, Q9I6J2, and O50131. Reaction mixtures contained 50 mM sodium phosphate, pH 8.0, 0.4 mM PLP, 5 mM *o*-aminobenzaldehyde, 100 mM ketoacid, and 100 mM L-lysine. Reactions were initiated through addition of 10 µM of purified enzymes and allowed to react at 37°C for approximately 18 hours. For negative controls, purified enzyme was replaced with the same volume of 50 mM sodium phosphate buffer. Reactions containing lysine as the amino donor were tested

colorimetrically, taking advantage of the spontaneous cyclization of products and reactivity with o-aminobenzaldehyde to form colored compounds. If the α amine is used as a donor, 6-amino-2-oxohexanoic acid is formed which spontaneously cyclizes to 2-piperideine-2-carboxylate. The piperideine reacts with o-aminobenzaldehyde to form a three-membered ring product with an absorbance maximum of 450 nm (37, 38). Use of the ε amine as a donor results in formation of L-allysine. L-allysine similarly cyclizes to (S)-1-piperideine-6-carboxylate and reacts with o-aminobenzaldehyde to form a three-membered ring product with an absorbance maximum at 465 nm (36–38). Absorbance was measured by a POLARStar Omega plate reader (BMG Labtech). Quantification was performed using an extinction coefficient of 2.8 mM^{-1} cm^{-1}. One unit of enzyme activity is defined as catalyzing formation of 1 nmol of product per min. All data points were measured in duplicates.

Transhydrogenase Assay: Transhydrogenase A0A0N0U829 from *Thermus aquaticus* was tested for transhydrogenase activity. Transhydrogenase reactions contained 50 mM Tris, pH 7.5, 0.1 mM $ZnCl_2$ or 1 mM of other divalent cations ($MnSO_4$, $Co_3(PO_4)_2$), $MgCl_2$), 50 mM 2-ketoglutarate, 200 mM (R)-lactate, and 10 µM purified enzyme. For negative controls, purified enzyme was replaced with the same volume of 50 mM Tris, pH 7.5 buffer. Reactions were run at 55 °C for approximately 18 hours and stopped by heating at 95 °C for 3 minutes. Samples were centrifuged for 10 minutes at 16,000 x g at 4 °C to remove protein precipitation and the pyruvate formation was determined via L-LDH (Roche) by monitoring loss of absorbance of NADH at 340 nm. The L-LDH assay mixture contained 100 mM Tris-HCl, pH 7.5, 3 mM NADH, 0.1 unit L-LDH, 50 uL reaction mix in a total reaction volume of 200 µL. Absorbance was measured on a POLARStar Omega plate reader (BMG Labtech). All data points were measured in duplicates.

Kinase Assay: A0A0N0BLM2 from *Thermus aquaticus* was tested for kinase activity. Reactions contained 50 mM Tris, pH 7.5, 5 mM $MgCl_2$, 1 mM ATP, 1 mM CMP or dCMP, 10 µM purified enzyme. For negative controls, purified enzyme was replaced with the same volume of 50 mM Tris, pH 7.5 buffer. Reactions were run at room temperature for ~18 hours. The formation of ADP was detected by ADP-Glo™ Assay (Promega) following the manufacturer's protocol. CDP and dCDP were detected by LCMS as described in the LCMS section. All data points were measured in duplicates.

LC-MS

Non-canonical amino acid products of lysine transaminases and dinucleotide products of A0A0N0BLM2 from *Thermus aquaticus* were confirmed via LC-MS consisting of a LC-20AD liquid chromatography system (Shimadzu) and 3200 QTRAP MS/MS (SciEx). Separations were performed on an ACQUITY UPLC HSS T3 column, 100Å, 1.8 µm, 2.1 mm X 150 mm (Waters). Mobile phase A was composed of 0.1% formic acid in water, and mobile phase B was acetonitrile with 0.1% formic acid. The gradient was as follows: 0 min, 0% B; 0 to 2.5 min, 4.8% B; 2.5 to 3.5 min, 4.8 to 30% B; 3.5 to 4.10 min, 30 to 100% B; 4.1 to 5.1 min, 100% B; 5.1 to 5.2 min, 100% to 0% B; 5.2 to 10 min, 0% B. The flow rate was 0.2 mL/min. The injection volume was 10 µL and the column was maintained at 40 °C. Experiments were performed on a triple quadrupole electrospray ionization (ESI) 3200 QTRAP MS/MS (SciEx) operating in positive ion mode. Analytes were fragmented using nitrogen gas as the collision gas and optimized multiple reaction monitoring transitions were recorded for each analyte. Mass spectrometry parameters were as follows: A) L-phenylglycine: ion source gas 1 (GS1) 50.00 arbitrary units, ion source gas 2 (GS2) 50.00 arbitrary units, curtain gas (CUR) 20.00 arbitrary units, ionization spray 5000.00, CAD Medium, declustering potential (DP) 20.00 V, entrance potential (EP) 10.00 V, collision cell entrance potential (CE) 15.30 V, collision cell exit potential (CXP) 4.00 V, source temperature 500 °C, B) 2-thienylglycine: GS1 50.00, GS2 50.00, CUR 20.00, IS 5000.00, CAD Medium, DP 17.00 V, EP 4.00 V, CE 22.00, CXP 2.50, source temperature 500 °C, C) 2-Furylglycine: GS1 50.00, GS2 50.00, CUR 20.00, IS 5000.00, CAD Medium, DP 20.00 V, EP 10.00 V, CE 30.00, CXP 5.00, source temperature 500 °C, D) CDP: GS1 50.00, GS2 50.00, CUR 20.00, IS 5000.00, CAD Medium, DP -47.00 V, EP -4.00 V, CE -31.00, CXP -1.00, source temperature 500 °C, E) dCDP: GS1 50.00, GS2 50.00, CUR 20.00, IS 5000.00, CAD Medium, DP -45.00 V, EP -5.00 V, CE -34.00, CXP -1.36, source temperature 500 °C. MRM transitions and retention times are listed in the Table below. For phenylglycine, peak areas were integrated between 2.97 and 4.20 min. *m/z* transitions and retention times used in this study are in Table S3. Note, in Fig. 3, we cannot distinguish between L- and D-enantiomers of the amino acid products because the chromatography was performed on an achiral column, which does not resolve stereoisomers.

Supplementary Information

Supplementary Tables

Model/ E.C. Level	**Top-1 Hit Rate**	**Top-10 Hit Rate**	**Top-100 Hit Rate**	**Top-1000 Hit Rate**	***R*-Precision**	**Average Precision**
E.C. 1 (a.-.-.-)	0.1	0.9	5.6	19.6	0.1	0.1
E.C. 2 (a.b.-.-)	0.9	5.0	16.5	40.0	1.0	1.0
E.C. 3 (a.b.c.-)	2.6	10.8	27.5	51.3	2.7	2.7
E.C. 4 (a.b.c.d)	60.1	62.6	63.5	64.6	60.0	60.5
Horizyn-1	36.2	57.1	76.7	92.1	36.9	42.1
Horizyn-1 + E.C. 1 (a.-.-.-)	38.3	60.5	81.4	94.9	39.5	44.9
Horizyn-1 + E.C. 2 (a.b.-.-)	40.1	63.4	84.9	96.7	41.7	47.3
Horizyn-1 + E.C. 3 (a.b.c.-)	41.9	65.6	86.7	97.6	43.9	49.5
Horizyn-1 + E.C. 4 (a.b.c.d)	81.9	90.8	96.0	98.7	82.2	84.5

Table S1. Performance conditioned on E.C. numbers. This table presents retrieval performance metrics on the test set of the development dataset, comparing rankings that use

partial E.C. numbers, *Horizyn-1* scores (trained on the training split of the development dataset), and a composite score. For the composite method, enzymes are prioritized by E.C. number matches, with *Horizyn-1* scores used to rank enzymes within those groups. Full analysis details can be found in the E.C. Conditioning Analysis section of the Methods.

Reaction Embeddings	Enzyme Embeddings	BEDROC 20	BEDROC 85	EF0.05	EF0.1	Model/ Source
ECFP6/DRFP	**ProtT5**	**66.3**	**48.5**	**14.7**	**8.3**	***Horizyn-1***
ECFP6/DRFP	ESM-2 (3B)	59.8	41.6	13.5	7.6	
ECFP6/DRFP	ESM-2 (650M)	57.8	42.4	12.7	7.3	
PTS-GNN	ESM-2 (650M)	29.4	17.8	6.6	4.2	(17)
PTS-GNN	ESM-2 (650M) + EGNN	63.0	44.7	14.1	8.1	CLIPZyme (17)

Table S2. Comparison of *Horizyn-1* and Clipzyme on the EnzymeMap benchmark. Performance of *Horizyn-1* (trained on the EnzymeMap training split) and CLIPZyme variants was assessed using early recognition metrics. Specifically, we report the Boltzmann-Enhanced Discrimination of ROC (BEDROC) with α parameters of 20 and 85, and the Enrichment Factor (EF) at the top 5% and 10% of the ranked list. Full definitions for these metrics are located in the Model Evaluation section. *Horizyn-1* variants combine ECFP6/DRFP reaction fingerprints with different protein language models [ProtT5, ESM-2 (3B), and ESM-2 (650M)]. Data for the CLIPZyme variants are reproduced from (17); these models use a pseudo-transition state graph neural network (PTS-GNN) for reaction embeddings and ESM-2 (650M) for protein embeddings, with and without a structure-aware EGNN module.

Analyte	**Ion mode**	**Multiple Reaction Monitoring (MRM) transitions m/z**	**Retention time (RT) min**
2-furylglycine	Positive	*142.2 → 97.1* 142.2 → 69.2	2.2
L-phenylglycine	Positive	152.2 → 135.0 *152.2 → 107.2* 152.2 → 79.2	3.0
2-thienylglycine	Positive	158.1 → 141.2 *158.1 → 113.1* 158.1 → 85.1	2.7
CDP	Negative	*402.1 → 158.7* 402.1 → 110.1 402.1 → 78.9	2.4
dCDP	Negative	*386.0 → 158.7* 386.0 → 78.7	2.4

Table S3. Multiple reaction monitoring (MRM) m/z transitions Ion mode, *m/z* transitions, and retention times for analytes measured in this study. Transitions in italics represent data used for analysis.

Supplementary Figures

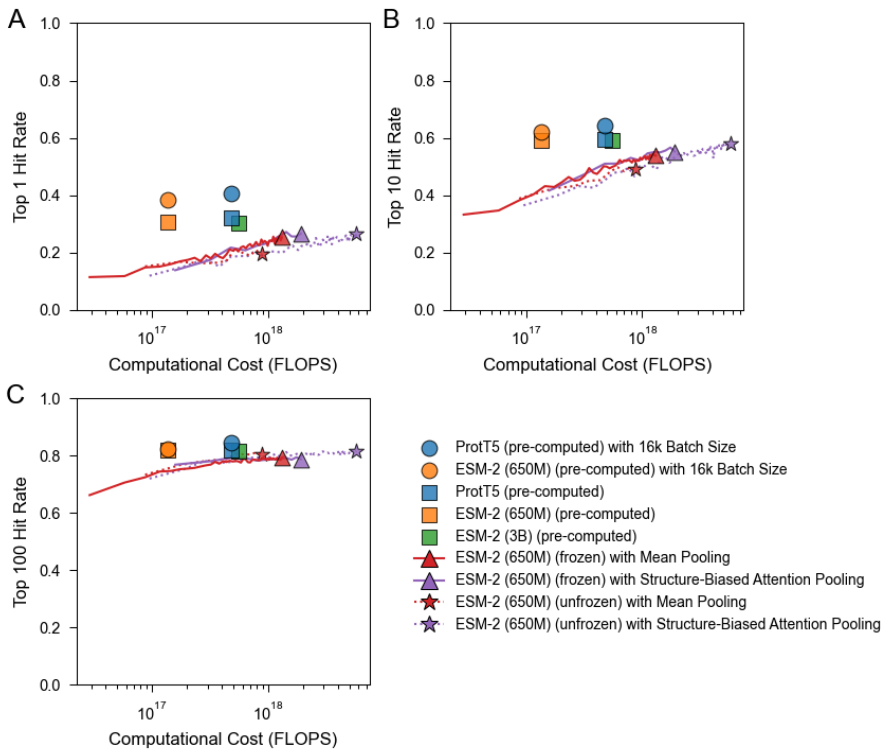


Figure S1. Performance vs. computational cost for different protein embedding strategies. The trade-off between model performance and computational efficiency is shown for various protein language model (pLM) embedding strategies. The **(A)** top-1, **(B)** top-10, and **(C)** top-100 hit rates on the validation set are plotted against the cumulative computational cost, measured in Floating Point Operations (FLOPs). Lines represent the performance trajectory during training, while the solid markers indicate the final performance of each fully trained model. Two main paradigms are compared: models trained on pre-computed pLM embeddings (circles and squares) and models that integrate a live ESM-2 (650M) model during training (triangles and stars). The live models were trained with either a frozen or unfrozen pLM, and with or without a structure-aware attention pooling layer. See Protein Language Model Compute Cost for analysis details.

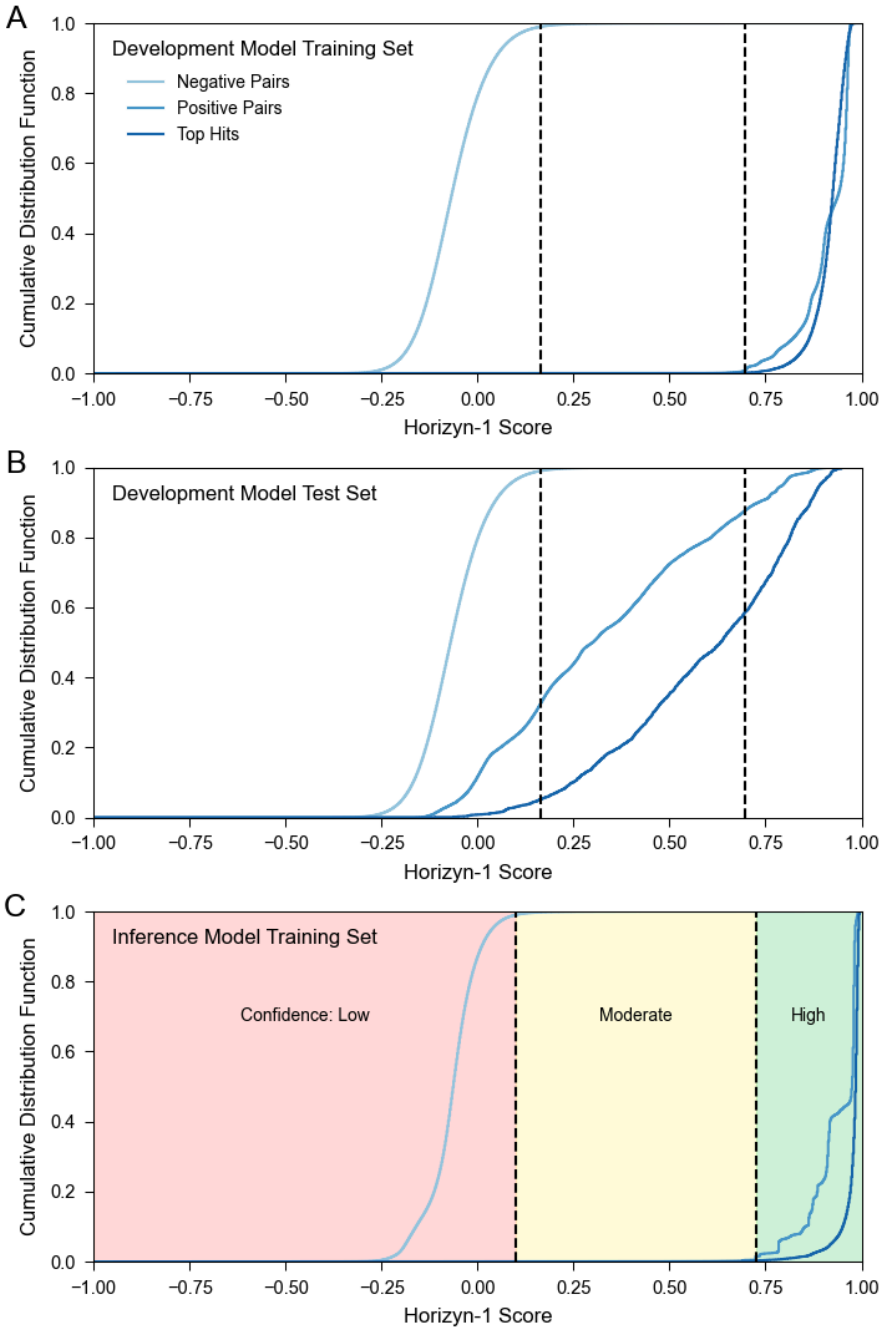


Figure S2. Score distributions and confidence heuristic. Cumulative Distribution Functions (CDFs) of model scores are shown for three categories: negative pairs (light blue), positive pairs (medium blue), and top hits (dark blue). Score distributions generated by the model trained on the training split of the development dataset are plotted for **(A)** the development training set and **(B)** the development model test set, demonstrating clear separation between the scores of positive and negative pairs. **(C)** The score distributions generated by *Horizyn-1* for the inference

training set are used to define a three-tiered confidence heuristic, visually represented by the shaded regions. The dashed lines denote boundaries between 'Low,' 'Moderate,' and 'High' confidence predictions, which are based on percentiles of the negative and positive score distributions (see Model Confidence Heuristic).

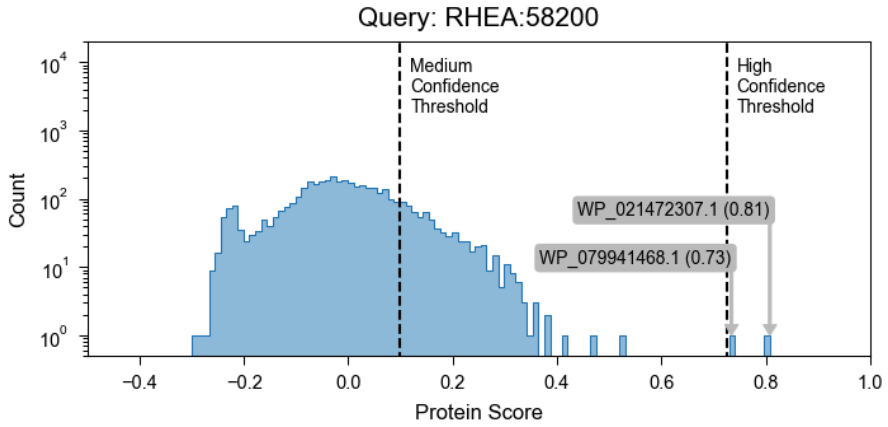


Figure S3. *Horizyn-1* predictions for 6-aminohexanoate transaminase orphan reaction. (RHEA:58200) among *Arthrobacter sp.* KI72 genome. Distribution of Horizyn-1 protein scores for the 6-aminohexanoate aminotransferase orphan reaction (RHEA:58200) queried against all 4,173 protein-coding sequences from the *Arthrobacter sp.* KI72 genome. The scores for the two highest-ranking proteins, WP_021472307.1 (a GabT homolog) and WP_079941468.1, are explicitly annotated (0.81 and 0.73, respectively). Both candidates surpass the high-confidence threshold (right dashed line), supporting their selection for subsequent experimental characterization.

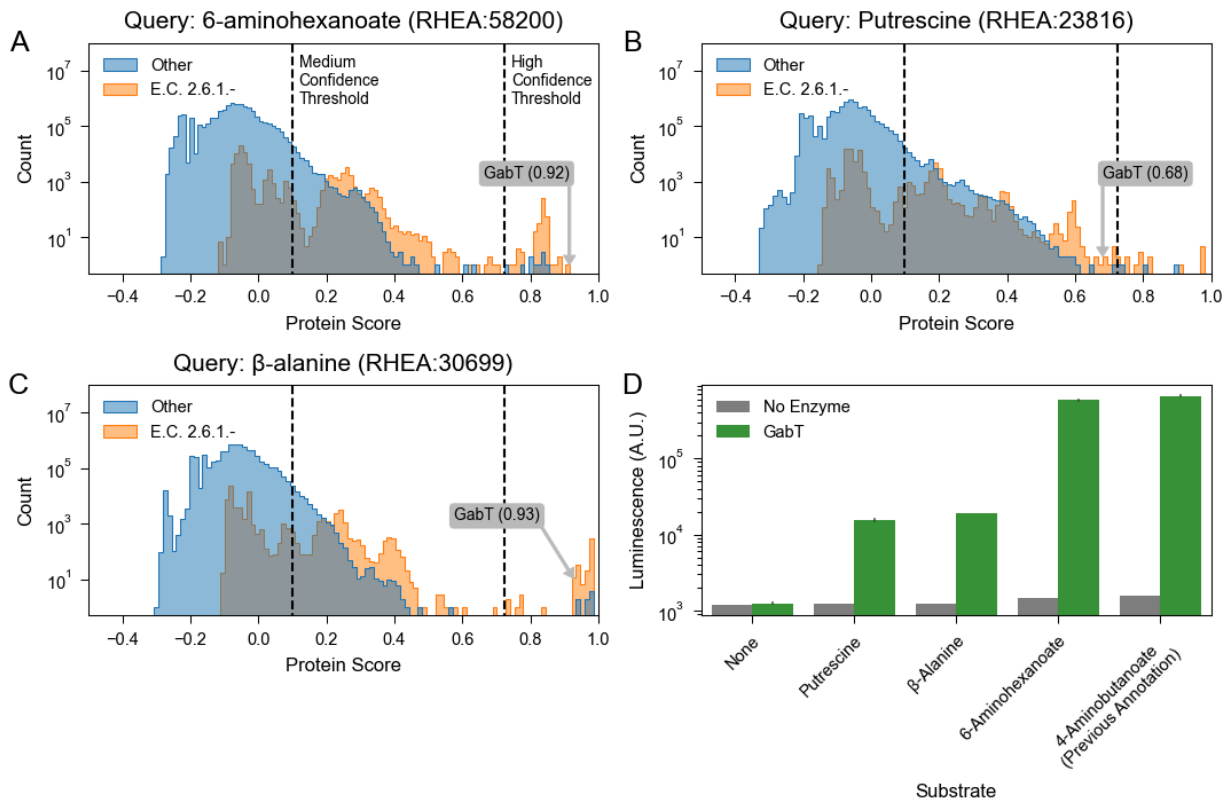


Figure S4. *Horizyn-1* prediction and experimental confirmation of promiscuous ω-transaminase activity for *E. coli* GabT. (A-C) Distributions of *Horizyn-1* protein scores for three orphan ω-transaminase reactions queried against our inference screening set of ~7 million enzymes. The reactions shown are for (A) 6-aminohexanoate (RHEA:58200), (B) putrescine (RHEA:23816), and (C) β-alanine (RHEA:30699). In each plot, the score for *E. coli* GabT (P22256) is highlighted with its score in parentheses, ranking it as a high-confidence candidate for these promiscuous activities. The histograms differentiate between known transaminases (E.C. 2.6.1.-; orange) and all other enzymes(blue). **(D)** Experimental validation of GabT activity. L-glutamate production (measured as luminescence on a logarithmic scale) was assayed for the predicted substrates and the previously annotated substrate, 4-aminobutanoate (RHEA:23352). Green bars indicate activity with purified GabT, while gray bars represent the background signal from the uncatalyzed reaction. The results confirm that GabT has strong activity with 6-aminohexanoate and displays weaker, promiscuous activity with putrescine and β-alanine. Error bars denote the standard deviation from duplicate measurements.

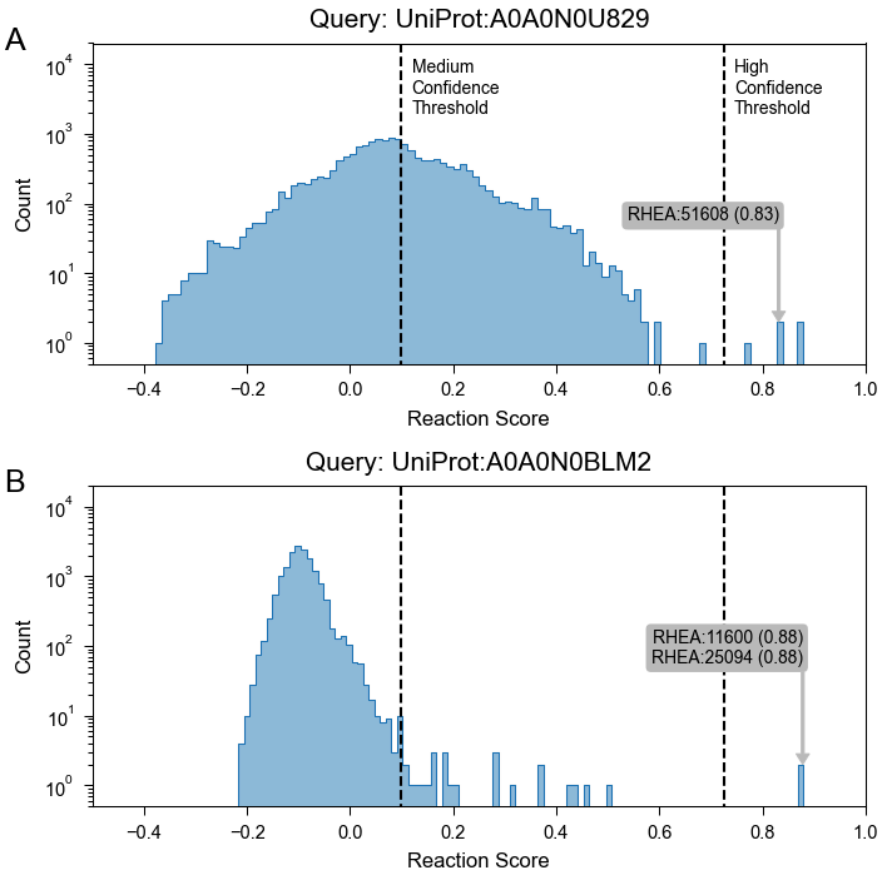


Figure S5. Horizyn-1 predictions for promiscuous activity for enzymes from *Thermus aquaticus*. (A) Distribution of Horizyn-1 reaction scores for the hypothetical D-lactate dehydrogenase A0A0N0U829, queried against all 15,814 Rhea reactions. While the protein's annotated function is lactate oxidation, the analysis predicted a high-scoring promiscuous activity: (R)-2-hydroxyglutarate–pyruvate transhydrogenase (RHEA:51608, score = 0.83). This prediction, which surpassed the high-confidence threshold, suggested the enzyme could utilize alternative electron donor/acceptor pairs and was subsequently confirmed experimentally. **(B)** Distribution of Horizyn-1 reaction scores for the uncharacterized protein A0A0N0BLM2, queried against all Rhea reactions. The model predicted two high-scoring kinase activities: cytidine monophosphate (CMP) kinase (RHEA:11600, score = 0.88) and deoxycytidine monophosphate (dCMP) kinase (RHEA:25094, score = 0.88). Subsequent experimental testing confirmed the dCMP kinase activity, demonstrating Horizyn-1's ability to successfully assign function to a previously unannotated protein.

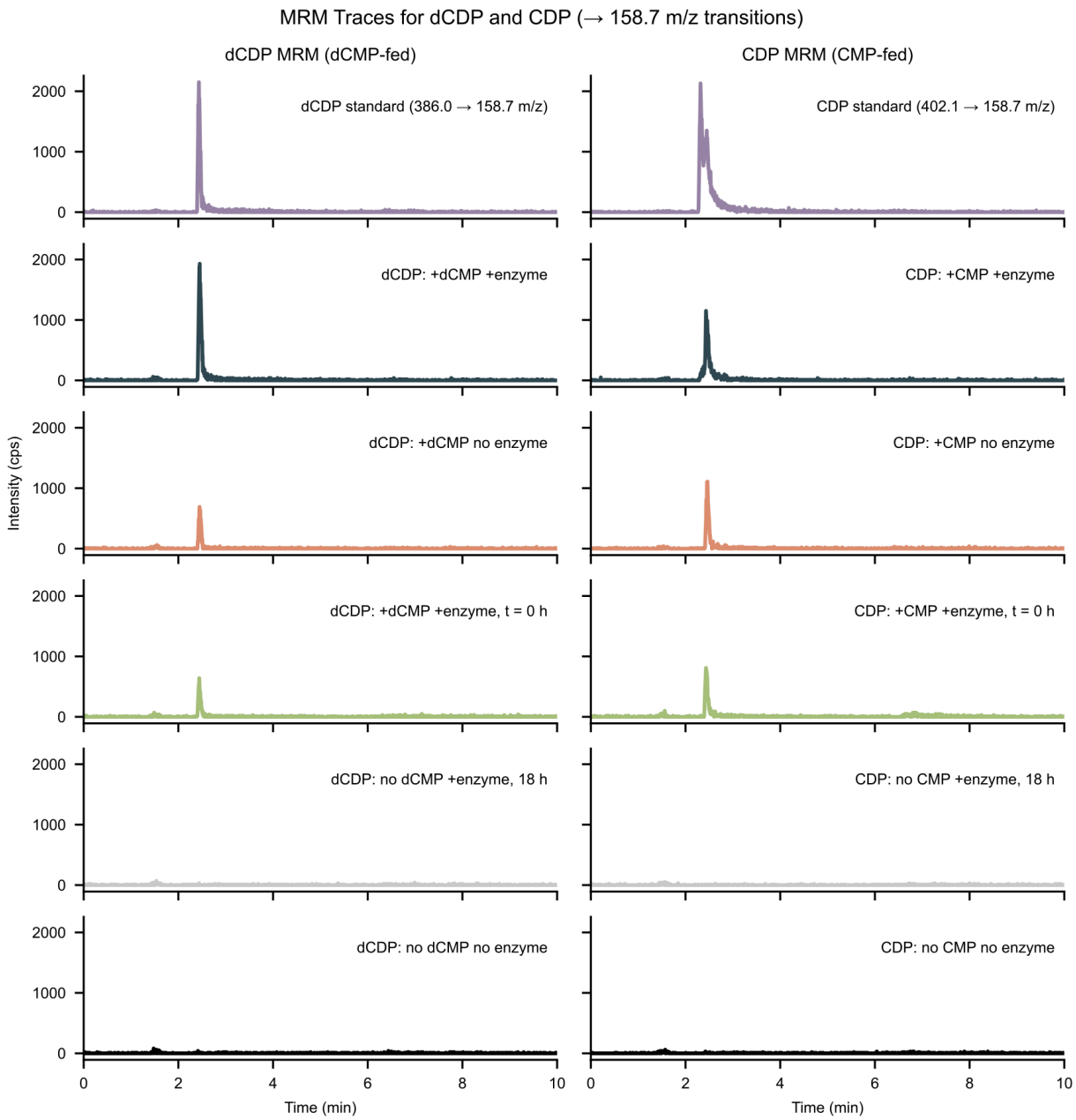


Figure S6. MRM analysis of (d)CMP kinase activity on dCMP and CMP substrates. Multiple reaction monitoring (MRM) chromatograms show extracted ion transitions for dCDP (386.0 → 158.7 m/z, left) and CDP (402.1 → 158.7 m/z, right) under various reaction conditions. Each row corresponds to a unique condition: reactions with or without enzyme (A0A0N0BLM2), with or without nucleotide substrate (dCMP or CMP), and an early timepoint control (t = 0 h). A strong dCDP signal is observed only in the presence of both dCMP and enzyme, indicating conversion of dCMP to dCDP after 18h of incubation. In contrast, no

appreciable CDP signal is observed relative to controls in reactions containing CMP and enzyme. These results demonstrate that the enzyme selectively phosphorylates dCMP but not CMP under the conditions tested.

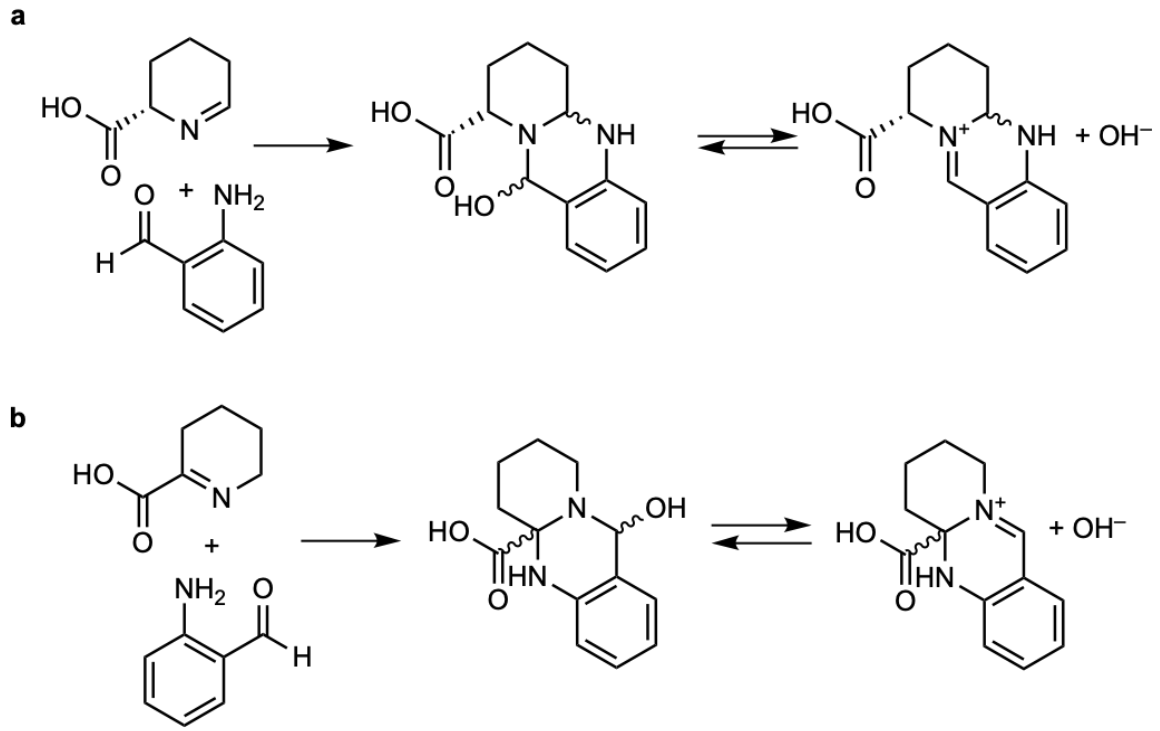


Figure S7. Reaction schemes for L-lysine-driven transamination assay. Reaction schemes for the detection of lysine **(a)** ε-amine and **(b)** α-amine transamination products via o-aminobenzaldehyde reactivity. Lysine transaminase screening may be performed colorimetrically by taking advantage of the spontaneous cyclization of products that react with o-aminobenzaldehyde. Transamination utilizing the ε-amine of lysine produces *L*-allysine which spontaneously cyclizes to form (*S*)-1-piperideine-6-carboxylate. This molecule reacts with o-aminobenzaldehyde to form a three-membered ring product with an absorbance maximum at 465 nm (36–38). Quantification can proceed using the extinction coefficient of 2.8 mM^{-1} cm^{-1}. One unit of enzyme activity is defined as catalyzing formation of 1 nmol of product per min. Similarly, transamination utilizing the α-amine of lysine produces 6-amino-2-oxohexanoic acid which spontaneously cyclizes to form 1-piperideine-2-carboxylate. Reaction with *o*-aminobenzaldehyde produces a three-membered ring product with an absorbance maximum of 450 nm (36–38) .

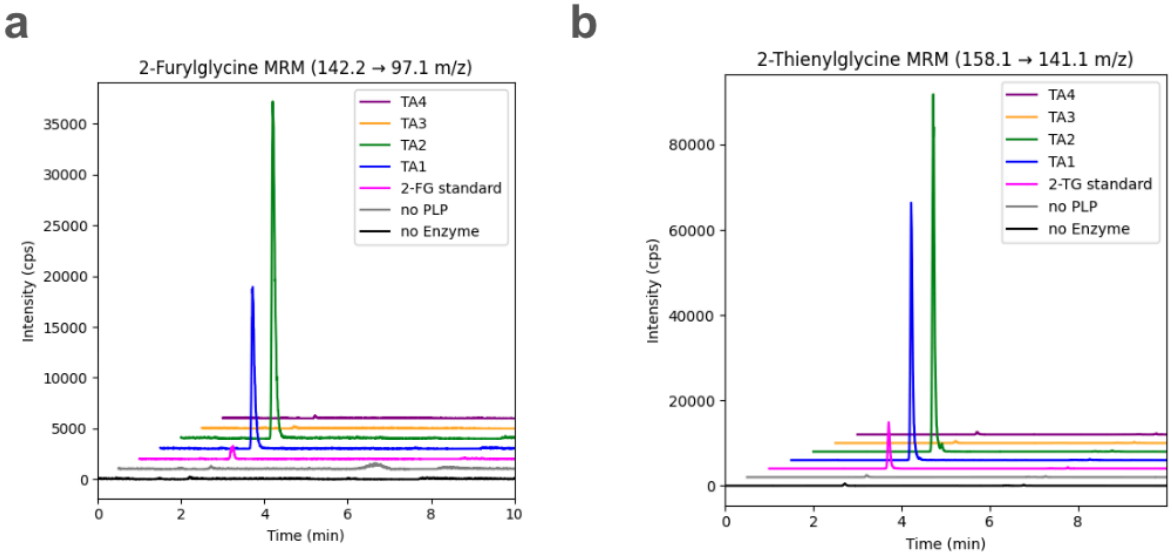


Figure S8. MRM for non-canonical amino acid detection. Positive mode MRM of characteristic transitions for non-canonical amino acids synthesized via smart lysine transaminases. Traces are shown for **(a)** 2-furylglycine and **(b)** 2-thienylglycine. Negative controls in which the PLP cofactor or the enzyme were omitted from the reaction are shown in gray and black traces, respectively.

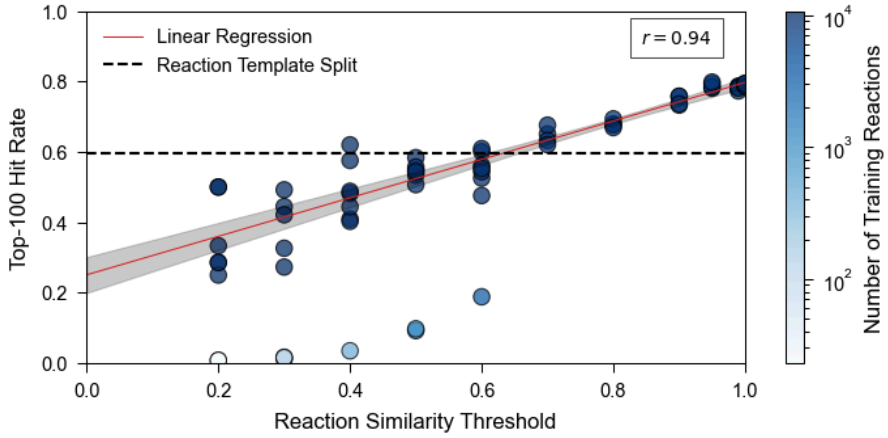


Figure S9. Impact of training and validation data similarity on model performance. The scatter plot shows the top-100 hit rate for models trained on data splits generated using various fingerprint-based reaction similarity thresholds. A lower threshold creates a more challenging split by increasing the dissimilarity between reactions in the training and validation sets. Each point represents an independently trained model, colored by the number of training reactions in its respective split (color bar). A linear regression (red line) with 95% confidence intervals (gray shading) demonstrates a clear positive correlation between the similarity threshold and model performance (Pearson's r = 0.94). For comparison, the dashed line shows the average performance (top-100 hit rate of 60%) when splitting data by reaction templates (see Reaction Similarity Clustering Analysis).

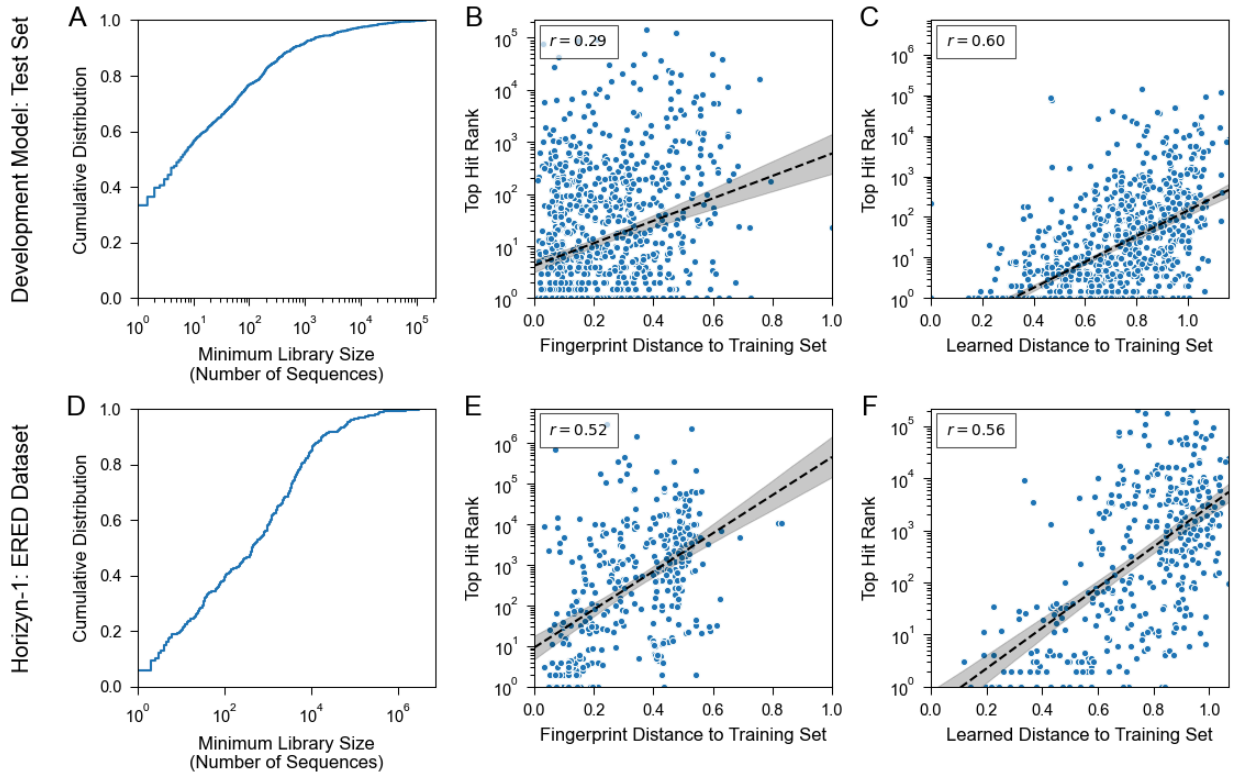


Figure S10. **Model performance and reaction novelty.** Model performance and its relationship to reaction novelty are presented for **(A-C)** a model trained on the development dataset training split and evaluated on the development test split and **(D-F)** *Horizyn-1* (trained on the inference dataset) evaluated on the ERED dataset. **(A, D)** The cumulative distribution of the top hit ranks; the x-axis represents the minimum number of sequences that would need to be experimentally screened to find the first positive hit for a given reaction. For a given position *k* on the x-axis, the position on the y-axis is the top-*k* hit rate. **(B, C, E, F)** Scatter plots correlating the rank of the top-scoring enzyme hit with the reaction's novelty. Novelty is defined as the minimum distance from a test reaction to any reaction in the corresponding training set. This distance is measured using either **(B, E)** Tanimoto distance of chemical fingerprints (see Structural Fingerprint Clustering Protocol) or **(C, F)** Euclidean distance within the learned embedding space. For both datasets, the performance is more strongly correlated with the learned distance than the fingerprint distance (Pearson's r = 0.60 vs. 0.29 for the development test set; r = 0.56 vs. 0.52 for the ERED set), demonstrating that the model's learned embedding space is a more effective measure of biochemical novelty.

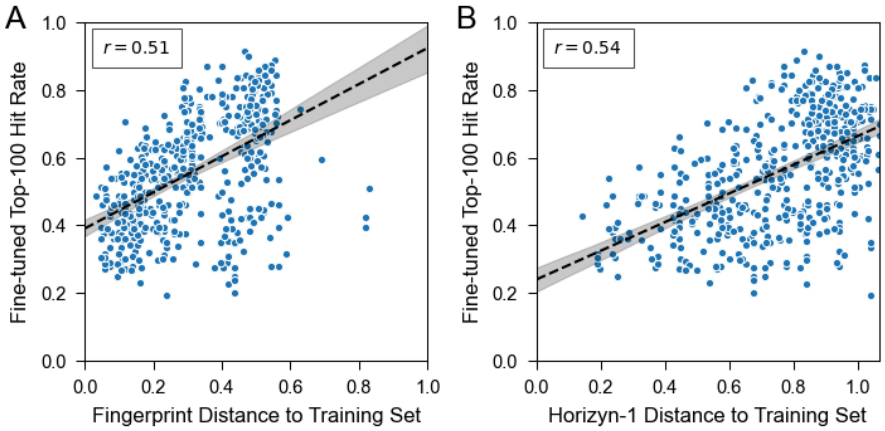


Figure S11: Fine-tuning improvement vs. reaction novelty. The scatter plots correlate the top-100 hit rate on a random ERED validation set after fine-tuning the *Horizyn-1* inference model on a single reaction with the novelty of that reaction (see Ene-Reductase Fine-tuning). Novelty is quantified as the minimum distance of the reaction used for fine-tuning to any reaction in the original training set, measured using either **(A)** chemical fingerprint Tanimoto distance (see Structural Fingerprint Clustering Protocol) or **(B)** Euclidean distance within the *Horizyn-1* embedding space. The performance after fine-tuning is strongly correlated with both distances (r = 0.54 vs. r = 0.51), suggesting that biochemical novelty leads to performance improvements upon fine-tuning.

References

1. R. Buller, *et al.*, From nature to industry: Harnessing enzymes for biocatalysis. *Science* **382**, eadh8615 (2023).

2. A. Lauko, *et al.*, Computational design of serine hydrolases. *Science* **388**, eadu2454 (2025).

3. H. R. Kim, H. Ji, G. B. Kim, S. Y. Lee, Enzyme functional classification using artificial intelligence. *Trends Biotechnol.* **0** (2025).

4. V. de Crécy-Lagard, *et al.*, A roadmap for the functional annotation of protein families: a community perspective. *Database* **2022** (2022).

5. A. Mardinoglu, B. Ø. Palsson, Genome-scale models in human metabologenomics. *Nat. Rev. Genet.* **26**, 123–140 (2025).

6. N. Quinn-Bohmann, A. V. Carr, C. Diener, S. M. Gibbons, Moving from genome-scale to community-scale metabolic models for the human gut microbiome. *Nat. Microbiol.* **10**, 1055–1066 (2025).

7. A. MacDougall, *et al.*, UniRule: a unified rule resource for automatic annotation in the UniProt Knowledgebase. *Bioinformatics* **36**, 4643–4648 (2020).

8. T. Aramaki, *et al.*, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *Bioinformatics* **36**, 2251–2252 (2020).

9. S. R. Eddy, Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).

10. E. M. Zdobnov, R. Apweiler, InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**, 847–848 (2001).

11. S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman, Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

12. T. Yu, *et al.*, Enzyme function prediction using contrastive learning. *Science* **379**, 1358–1363 (2023).

13. J. Y. Ryu, H. U. Kim, S. Y. Lee, Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 13996–14001 (2019).

14. G. B. Kim, *et al.*, Functional annotation of enzyme-encoding genes using deep learning with transformer layers. *Nat. Commun.* **14**, 7370 (2023).

15. N. Hadadi, H. MohammadiPeyhani, L. Miskovic, M. Seijo, V. Hatzimanikatis, Enzyme annotation for orphan and novel reactions using knowledge of substrate reactive sites. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 7298–7307 (2019).

16. A. Radford, *et al.*, Learning transferable visual models from natural language supervision. *arXiv [cs.CV]* (2021).

17. P. Mikhael, I. Chinn, R. Barzilay, CLIPZyme: Reaction-Conditioned Virtual Screening of Enzymes in *International Conference on Machine Learning*, (PMLR, 2024), pp. 35647–35663.

18. C. Hua, *et al.*, ReactZyme: A benchmark for enzyme-reaction prediction. *arXiv [cs.LG]* (2024).

19. Y. Liu, *et al.*, EnzymeCAGE: A geometric foundation model for enzyme retrieval with evolutionary insights. *bioRxiv* 2024.12.15.628585 (2024).

20. J. Yang, *et al.*, CARE: A benchmark suite for the Classification and Retrieval of Enzymes. *arXiv [q-bio.BM]* (2024).

21. P. Bansal, *et al.*, Rhea, the reaction knowledgebase in 2022. *Nucleic Acids Res.* **50**, D693–D700 (2022).

22. E. Boutet, D. Lieberherr, M. Tognolli, M. Schneider, A. Bairoch, UniProtKB/Swiss-Prot. *Methods Mol. Biol.* **406**, 89–112 (2007).

23. D. Rogers, M. Hahn, Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).

24. N. Schneider, D. M. Lowe, R. A. Sayle, G. A. Landrum, Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **55**, 39–53 (2015).

25. D. Probst, P. Schwaller, J.-L. Reymond, Reaction classification and yield prediction using the differential reaction fingerprint DRFP. *Digit Discov* **1**, 91–97 (2022).

26. A. Elnaggar, *et al.*, ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans. Pattern Anal. Mach. Intell.* **44**, 7112–7127 (2022).

27. P. Schwaller, *et al.*, Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **3**, 144–152 (2021).

28. E. Heid, *et al.*, Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **64**, 9–17 (2024).

29. Z. Lin, *et al.*, Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

30. M. Sorokina, M. Stam, C. Médigue, O. Lespinet, D. Vallenet, Profiling the orphan enzymes. *Biol. Direct* **9**, 10 (2014).

31. I. Takehara, *et al.*, Metabolic pathway of 6-aminohexanoate in the nylon oligomer-degrading bacterium Arthrobacter sp. KI72: identification of the enzymes responsible for the conversion of 6-aminohexanoate to adipate. *Appl. Microbiol. Biotechnol.* **102**, 801–814 (2018).

32. T. Tatusova, *et al.*, NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res.* **44**, 6614–6624 (2016).

33. S. F. Altschul, *et al.*, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).

34. M. van Kempen, *et al.*, Fast and accurate protein structure search with Foldseek. *Nat. Biotechnol.* **42**, 243–246 (2024).

35. N. P. Dunham, *et al.*, Transaminase-catalyzed synthesis of β-branched noncanonical amino acids driven by a lysine Amine donor. *J. Am. Chem. Soc.* **146**, 16306–16313 (2024).

36. R. Kawakami, *et al.*, Crystal structure of a novel type of ornithine δ-aminotransferase from the hyperthermophilic archaeon Pyrococcus horikoshii. *Int. J. Biol. Macromol.* **208**, 731–740 (2022).

37. J. C. Fothergill, J. R. Guest, Catabolism of L-lysine by Pseudomonas aeruginosa. *J. Gen. Microbiol.* **99**, 139–155 (1977).

38. B. Holmstedt, R. Tham, A spectrophotometric method for determination of diamine oxidase (DAO) activity. *Acta Physiol. Scand.* **45**, 152–163 (1959).

39. Y.-C. Chang, *et al.*, COMBREX-DB: an experiment centered database of protein function: knowledge, predictions and knowledge gaps. *Nucleic Acids Res.* **44**, D330–D335 (2016).

40. J. L. Watson, *et al.*, De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).

41. F. H. Arnold, Innovation by Evolution: Bringing New Chemistry to Life (Nobel Lecture). *Angew. Chem. Int. Ed Engl.* **58**, 14420–14426 (2019).

42. UniProt Consortium, UniProt: The universal protein knowledgebase in 2025. *Nucleic Acids Res.* **53**, D609–D617 (2025).

43. E. Heid, D. Probst, W. H. Green, G. K. H. Madsen, EnzymeMap: curation, validation and data-driven prediction of enzymatic reactions. *Chem. Sci.* **14**, 14229–14242 (2023).

44. E. Heid, D. Probst, W. H. Green, G. K. H. Madsen, EnzymeMap. Zenodo. https://doi.org/10.5281/ZENODO.8254726. Deposited 2023.

45. M. Steinegger, J. Söding, MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **35**, 1026–1028 (2017).

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

47. M. Varadi, *et al.*, AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res.* **52**, D368–D375 (2024).

48. A. Chang, *et al.*, BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Res.* **49**, D498–D508 (2021).

49. W. Finnigan, *et al.*, RetroBioCat Database: A Platform for Collaborative Curation and Automated Meta-Analysis of Biocatalysis Data. *ACS Catal.* **13**, 11771–11780 (2023).

50. G. Landrum, *et al.*, *rdkit/rdkit: 2025_03_3 (Q1 2025) Release* (Zenodo, 2025).

51. P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Sci Adv* **7** (2021).

52. A. Varnek, D. Fourches, F. Hoonakker, V. P. Solov'ev, Substructural fragments: an universal language to encode reactions, molecular and supramolecular structures. *J. Comput. Aided Mol. Des.* **19**, 693–703 (2005).

53. S. Fujita, Graphic characterization and taxonomy of organic reactions. *J. Chem. Educ.* **67**, 290 (1990).

54. E. Rusak, *et al.*, InfoNCE: Identifying the gap between theory and practice. *arXiv [cs.LG]* (2024).

55. Z. Ma, M. Collins, Noise Contrastive Estimation and negative sampling for conditional models: Consistency and statistical efficiency. *arXiv [cs.CL]* (2018).

56. A. Paszke, *et al.*, PyTorch: An imperative style, high-performance deep learning library. *arXiv [cs.LG]* (2019).

57. W. Falcon, The PyTorch Lightning team, *PyTorch Lightning* (Zenodo, 2024).