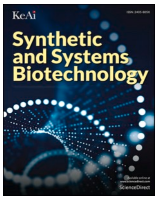
Contents lists available at ScienceDirect

Synthetic and Systems Biotechnology

journal homepage: www.keaipublishing.com/en/journals/synthetic-and-systems-biotechnology



Original Research Article

Sequence and taxonomic feature evaluation facilitated the discovery of alcohol oxidases

Yilei Han [a,1], Xuwei Ding [b,1], Junjian Tan [a], Yajuan Sun [b], Yunjiang Duan [a], Zheng Liu [a], Gaowei Zheng [b,*], Diannan Lu [a,**]

[a] *Department of Chemical Engineering, Tsinghua University, Beijing, 100084, China*
[b] *State Key Laboratory of Bioreactor Engineering, Shanghai Collaborative Innovation Center for Biomanufacturing, East China University of Science and Technology, Shanghai, 200237, China*

ARTICLE INFO

Keywords:
Enzyme discovery
alcohol oxidase
sequence similiarty network
Machine learning

ABSTRACT

Recent advancements in data technology offer immense opportunities for the discovery and development of new enzymes for the green synthesis of chemicals. Current protein databases predominantly prioritize overall sequence matches. The multi-scale features underpinning catalytic mechanisms and processes, which are scattered across various data sources, have not been sufficiently integrated to be effectively utilized in enzyme mining. In this study, we developed a sequence- and taxonomic-feature evaluation driven workflow to discover enzymes that can be expressed in *E. coli* and catalyze chemical reactions *in vitro*, using alcohol oxidase (AOX) for demonstration, which catalyzes the conversion of methanol to formaldehyde. A dataset of 21 reported AOXs was used to construct sequence scoring rules based on features, including sequence length, structural motifs, catalytic-related residues, binding residues, and overall structure. These scoring rules were applied to filter the results from HMM-based searches, yielding 357 candidate sequences of eukaryotic origin, which were categorized into six classes at 85 % sequence similarity. Experimental validation was conducted in two rounds on 31 selected sequences representing all classes. Among these selected sequences, 19 were expressed as soluble proteins in *E. coli*, and 18 of these soluble proteins exhibited AOX activity, as predicted. Notably, the most active recombinant AOX exhibited an activity of 8.65 ± 0.29 U/mg, approaching the highest activity of native eukaryotic enzymes. Compared to the established UniProt-annotation-based workflow, this feature-evaluation-based approach yielded a higher probability of highly active recombinant AOX (from 8.3 % to 19.4 %), demonstrating the efficiency and potential of this multi-dimensional feature evaluation method in accelerating the discovery of active enzymes.

1. Introduction

Recent years have witnessed the rapid development of data technology for enzyme discovery and design for unprecedented applications [1–3]. Growing efforts have been directed to establish an accurate function-to-protein mapping [4–8], which is essential for the success of data-driven enzyme discovery [9]. The predictions based solely on sequence data are prone to error propagation, where inaccuracies in raw annotations compound over iterative algorithmic cycles, leading to significant downstream errors [10]. Thus, novel approaches to improve annotation accuracy in large-scale sequence databases has been developed [11–13]. For enzyme annotation, numerous databases provide a wealth of resources for function inference from various structural scales including sequence, domains, motifs, and residues [14,15]. The Enzyme Function Initiative (EFI) has advanced the field of Genomic Enzymology [16–19], leveraging genomic environments of protein families to develop comprehensive strategies for discovering novel enzyme mechanisms, reactions, and metabolic pathways. EFI tools such as Sequence Similarity Networks [20] (SSN) and Genome Neighborhood Network [21] (GNN) are integral to these efforts. The InterPro [22] database

Peer review under the responsibility of Editorial Board of Synthetic and Systems Biotechnology.
* Corresponding author.
** Corresponding author.
E-mail addresses: gaoweizheng@ecust.edu.cn (G. Zheng), ludiannan@tsinghua.edu.cn (D. Lu).
[1] First authors.

https://doi.org/10.1016/j.synbio.2025.04.014
Received 20 January 2025; Received in revised form 18 April 2025; Accepted 21 April 2025
Available online 22 April 2025

classifies proteins by structural domains, using domain arrangements as "protein signatures" to infer functions [23]. This approach has been validated by studies such as the discovery of oxazolidinone synthases [24]. The M-CSA [25,26] database captures enzyme reaction mechanisms and active site information, providing a chemical perspective for functional inference. Considering the diverse types of information associated with enzyme catalysis, we directed our efforts towards a data integration workflow that correlates enzyme function with various features indicative of functional relevance. This flexible, one-stop information analysis approach enables us to identify the most functionally relevant features from vast amounts of data.

In this study, we proposed a workflow facilitated by sequence- and taxonomic-feature evaluation for enzyme discovery, using alcohol oxidase (AOX) as a demonstration. The physiological roles of AOX include methanol catabolism in methylotrophic yeast, providing energy and carbon sources, and the generation of hydrogen peroxide in filamentous fungi during wood degradation [27,28]. Recently, the potential of AOX in CO_2 fixation and utilization has been explored, owing to its ability to convert methanol—a central product of CO_2 fixation—into formaldehyde, a versatile precursor for various synthetic applications [29–34]. The discovery of AOX dates back over 50 years, yet its structure-based characterization [35,36] and engineering [37,38] have only recently attracted significant attention, making it an ideal example for validating the proposed workflow. In our previous work [30], we successfully expressed recombinant AOX in *E. coli*, enabling rapid validation of newly discovered enzymes.

The computational workflow is outlined in Fig. 1, consisting of (i) extracting sequence and taxonomy features for referenced sequences, (ii) filtering and diversifying potentially active sequences, and (iii) selecting sequences for experimental validation. The BRENDA [15] and UniProt [39] databases were used as primary data sources for referenced sequences. The candidate sequences were identified through an evaluation of sequence length, domain composition, and key catalytic and binding residues. A sequence similarity network aligned with the taxonomy of the source organism was employed to guide enzyme selection. We then expressed the representative sequences and obtained novel enzymes with high activity and protein yield in previously uncharacterized classes. Statistical analyses of both *in silico* and *in vitro* screening results demonstrated the effectiveness of this feature evaluation-guided enzyme discovery workflow.

2. Methods

2.1. Data and software

The amount of data from BRENDA [15] and UniProt [39] databases (accessed on 2023.10.30) for sequences, references, and source organisms was summarized in Supplementary Table S1. The bioinformatics tools were employed for sequence processing and analysis. CD-HIT [40] (v4.8.1) was used for sequence clustering, and T-coffee's online server [41] (v11.00) was used for structure-guided multiple sequence alignment (MSA). Hmmer [42] software (v3.3) was used to construct the Hidden Markov Model (HMM) and perform HMM-based MSA. HMM-based sequence searches were conducted using Hmmer's online server [43] (v2.41.2). InterProScan [22] (v5.57–90.0) was used for structural domain assignment. USEARCH [44] (v11.0.667) was used to calculate the pairwise sequence similarity. Homology modeling was performed with the Swiss-Model online server [45] (accessed on 2023.11.04) in automated mode, and pairwise structural similarity was calculated using TM-align [46] (v20190822). Jalview [47] (v2.11.2.7) was used to visualize and analyze MSA results, and CytoScape [48] (v3.9.1) was employed for sequence similarity network visualization. Python scripts were used to construct and parse input and output files for these tools. PyMOL [49] (v2.5.2) was used for protein structure visualization and key residue analysis, with the DockingPie [50] plugin (v1.2) integrated with AutoDock Vina [51] for methanol docking.

General data processing was conducted using python (3.8.20) and Jupyter notebook. The main packages included: biopython (1.81) for processing biological sequences and MSA data, scipy (1.10.1) and scikit-learn (1.3.2) for clustering algorithms and statistical analysis, numpy (1.24.4), pandas (2.0.3) and matplotlib (3.7.2) for data manipulation, analysis, and visualization. The statistical analysis was performed using Fisher's exact test on 2 × 2 contingency tables. All source code and datasets are available on GitHub at: https://www.github.com/ZOOEEER/enzyme-mining-aox.

2.2. Sequence mining

The Hidden Markov Model (HMM) was employed for sequence mining. A dataset of non-redundant active AOX sequences was manually curated and clustered at a 90 % similarity threshold, resulting in representative sequences for HMM construction. The HMM model was applied to search the UniProt reference proteome (accessed on 2023.10.30) using default parameters. This search yielded 30,240 sequences, with a maximum E-score was 0.98, and a minimum HMM score was 19.5.

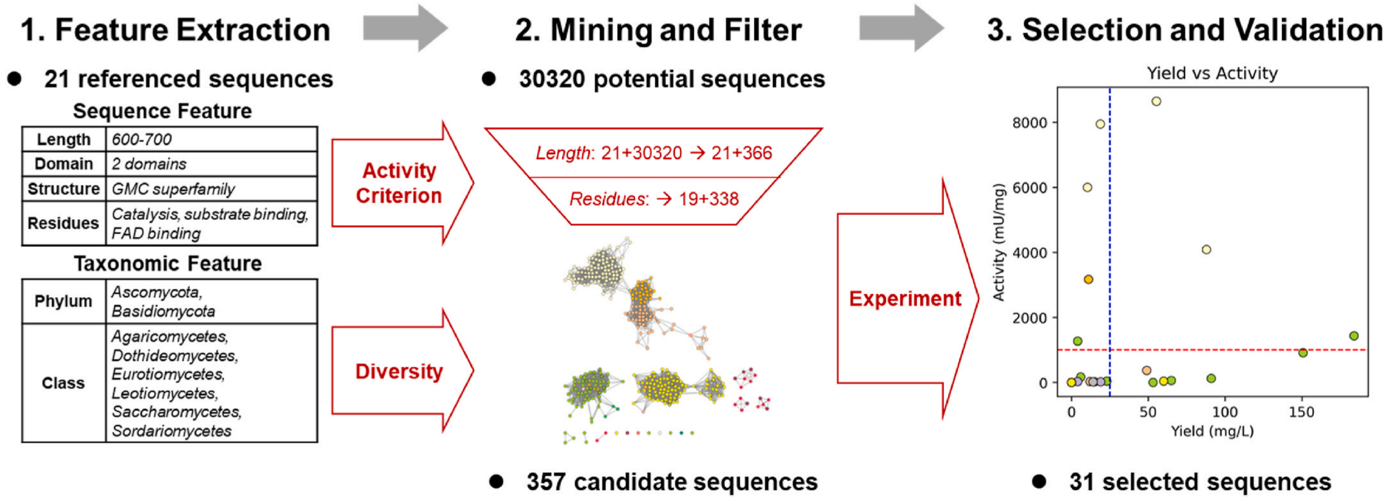


Fig. 1. The workflow for enzyme discovery of AOX facilitated by sequence- and taxonomic-feature evaluation.

2.3. Sequence evaluation

2.3.1. Sequence scoring

Scoring rules were made to evaluate post-MSA sequences. Each scoring rule assigns a specific score corresponding to the presence of a defined residue type set (A) at a given position (k). The residue type set (A) could be a subset of the 21 residue types, including 20 standard amino acids and the gap ("-"). The sequence-based score, **seq_score**(s), is the sum of the scores obtained by applying all the scoring rules to a given sequence s. The formula is as follows:

$$\mathbf{seq\_score}(s) = \sum_{i=1}^{m} score_i \cdot \mathbf{IF}\left(a_{k_i} \in A_i\right)$$

where m is that total number of scoring rules, $score_i$ is an arbitrary given score for the i-th scoring rule, a_k denotes the amino acid at the k-th position of the sequence, A_i is the specified residue type set for the i-th scoring rule, and the **IF** function returns 1 if the condition $a_{k_i} \in A_i$ is true, and 0 otherwise.

The sequence score quantifies the extent to which a sequence matching predefined sequence patterns. A high score indicates strong alignment with expected functional patterns, suggesting a greater likelihood of the desired function. These predefined scoring rules can be derived either heuristically through structure-function relationship modeling or statistically via sequence or structural alignment analyses.

2.3.2. Taxonomy scoring

The BRENDA database provides a list of active organisms associated with each EC number, along with relevant reference information. These active organisms were initially presented as strings, which were subsequently mapped to taxonomy units (taxid) from the NCBI Taxonomy [52] database (downloaded on 2023.11.04) through string matching to eliminate redundancy and renaming issues. For AOX, the active species was mapped to the taxonomic units at the genus level.

Each reference was treated as independent evidence of the active organism. For each reference p_i, all taxonomic units (super-kingdom, kingdom, phylum, order, family, species, genus) associated with it were counted once. For a specific taxonomic unit t_j, the total number of references associated with it was counted. To normalize the value, the occurrences of t_j in the NCBI Taxonomy database (**Total_count**(t_j)) was used as a denominator. The taxonomic score (**Tax_score**(t_j)) for the unit t_j was then calculated as follows:

$$\mathbf{Tax\_score}\left(t_j\right) = \frac{\sum_{i=1}^{n} \mathbf{Related}\left(p_i, t_j\right)}{\mathbf{Total\_count}\left(t_j\right)}$$

where n is the total number of references. The **Related** function returns 1 if the reference p_i reported the assay of activity on the taxonomic unit t_j, and 0 otherwise. Taxonomic units that were not reported in the references were assigned a score of 0.

For a given sequence s, which is mapped to various taxonomic levels, the set of corresponding taxonomic units is denoted as **TAXID**(s). The taxonomic score (**tax_score**(s)) for sequence s is the sum of the taxonomic scores (Tax_score(t)) for each taxonomic unit in the set **TAXID**(s). The formula is as follows:

$$\mathbf{tax\_score}(s) = \sum_{t \in \mathbf{TAXID}(s)} \mathbf{Tax\_score}(t)$$

The taxonomy score summarizes evidence from the literature on the source organisms of known enzymes, while accounting for phylogenetic relationships. A high taxonomy score indicates that the source organism of the sequence is more closely related phylogenetically to species with known enzymatic activity, suggesting a higher likelihood that the sequence possesses the corresponding function.

2.4. Sequence selection

A total of 357 candidate sequences were identified through sequence-based scoring. To construct sequence similarity networks (SSN), pairwise sequence similarity was calculated using USEARCH. Homology-based Structural models were used to calculate pairwise structural similarity. Since the TM-align score was found to be high (>85 %), no further classification based on structure was performed.

As a control group, a pre-experiment based on UniProt annotation was conducted, selecting 12 sequences from the 104 sequences labeled as "alcohol oxidase [EC 1.1.3.13]" for experiment. For the newly constructed candidate sequences, two rounds of selection were performed based on sampling at the class level (the taxonomic rank in biological classification). The first round focused on selecting sequences from previously reported classes, while the second round targeted unreported classes and those showing high activity in the first round of experiment. For each class, the sequences were randomly selected, with the number of sequences chosen proportional to the size of the class, and with a bias towards sequences that had already been reported. The two rounds of selection resulted in 19 and 12 sequences, respectively. In total, 43 sequences were experimentally validated.

2.5. Experimental validation

2.5.1. Reagents

4-aminoantipyrine (4-AAP) and 2,4,6-tribromo-3-hydroxybenzoic acid (TBHBA) were purchased from D&B Biological Science and Technology (Shanghai), and horseradish peroxidase (HRP) was purchased from Macklin Biochemical (Shanghai).

2.5.2. Cell culture and protein expression

All alcohol oxidase (AOX) genes were codon optimized and synthesized by GenScript Biotech and cloned into the pET-28a protein expression vector between the *Nde*I and *Xho*I restriction sites, facilitating N-terminal His-tag fusion. The constructs were transformed into *E. coli* BL21(DE3) chemically competent cells and incubated at 37 °C for 12 h. A single colony was then inoculated into 4 mL of Luria-Bertani (LB) medium containing 50 μg/mL kanamycin and cultured at 37 °C for 8 h. For protein expression, the cultures were transferred to Terrific Broth (TB) medium with 50 μg/mL kanamycin and incubated at 37 °C until the optical density at 600 nm (OD_{600}) reached 0.8. Protein expression was induced by adding 0.2 mM IPTG, followed by incubation at 16 °C for 24 h. Cells were harvested by centrifugation at 4 °C and 12,000 rpm. For protein purification, the cell pellet was resuspended in Buffer A (100 mM Tris, 300 mM NaCl, 10 mM imidazole, pH 7.5) and lysed by ultrasonication. The crude lysate was clarified by centrifugation, and the supernatant was loaded onto a nickel-affinity chromatography column. Non-specific proteins were eluted with a wash buffer composed of 10 % Buffer B (100 mM Tris, 300 mM NaCl, 500 mM imidazole, pH 7.5) and 90 % Buffer A, while AOX proteins were eluted using 60 % Buffer B and 40 % Buffer A. The eluted protein solution was concentrated using a 10 kDa molecular weight cutoff filter and subjected to buffer exchange with 20 mL of Buffer C (100 mM Tris, 100 mM NaCl, 1 mM dithiothreitol, pH 7.5). Protein concentration was measured using a NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific), quantifying the characteristic absorbance of protein at 280 nm (A_{280}). The absorbance at 450 nm (A_{450}) of the FAD cofactor extracted from purified proteins were also characterized. The results showed a good correlation between the two measurements (see **Supplementary Methods**). Soluble expression of the AOX proteins was confirmed via SDS-PAGE analysis of the supernatant and pellet fractions obtained after centrifugation.

2.5.3. AOX activity assay and enzyme kinetics

The AOX activity was determined through three parallel experiments using a cascade reaction involving AOX and HRP, based on the previously reported method [30] with minor modifications. The reaction produced a magenta-colored product, which was quantified spectrophotometrically at 510 nm.

One unit of AOX activity was defined as the amount of enzyme required to consume 1 μmol of H_2O_2 per minute under the assay conditions. The 1 mL reaction mixture comprised 250 μL of a mixed reagent (containing 6 mM 4-AAP, 20 mM TBABA, 50 U/mL HRP, and 2 % v/v DMSO), 2.5–120 mM methanol (50 mM for other alcohols except methanol), an appropriate amount of purified AOX, and 100 mM Tris buffer (pH 7.5) supplemented with 5 mM $MgCl_2$. In the activity assay, the AOX solution was gradually diluted to the desired concentration, ensuring that the absorbance at 510 nm remained below 1.5 (a.u.) so that all measurements fell within the linear range. Reactions were conducted at 30 °C for 1–3 min prior to analysis. The specific activity was calculated using the following formula:

$$\text{Specific activity } (\text{mU} / \text{mg}) = \frac{\Delta A \times V \times 1000}{t \times 29.4 \times l \times m_{protein}}$$

Where: ΔA: change of absorbance value (a.u.); V: total volume (mL); t: time (min); l: path length (cm); m_{protein}: purified enzyme (mg).

3. Results

3.1. Analysis of sequence features in referenced AOXs related to activity and expression

A dataset of referenced AOX sequences comprising 21 sequences (Supplementary Table S2) was constructed from selected literatures recorded in the BRENDA database [27,53–65]. The lengths of these sequences were tightly clustered, ranging from 638 to 687, distinguishing them from other alcohol oxidases with different substrate specificities [27].

The structure characteristics of AOX were analyzed to construct features associated with activity, as shown in Fig. 2. According to the InterPro-Scan results, all referenced AOXs consist of two domains: *Glucose-methanol-choline* (*GMC*) *oxidoreductase*, *N-terminal* and *GMC oxidoreductase*, *C-terminal*, which correspond to the FAD cofactor binding domain and the substrate binding domain, respectively [35]. In the cofactor binding domain, a conserved nucleotide binding site (GXGXXG) was identified at positions 13–18 (using *Kpa*AOX from *Ascomycota* (PDB: 5HSA, NCBI Accession: AAB57849.1) as a reference). The presumed catalytic residues, His567 and Asn616, are likely involved in capturing proton from the substrate methanol. In the substrate binding pocket, three large aromatic amino acids—Phe98, Trp566 and Phe417—were identified, which may form the active pocket and create a hydrophobic environment to selectively accommodate small alcohols as substrates. This structural feature is also conserved in another alcohol oxidase [37] from *Basidiomycota*, *Pch*AOX (PDB: 6H3G, NCBI Accession: CDG66232.1), where the corresponding residues are Phe101, Trp560, and Tyr422.

Multiple sequence alignment was performed across the 21 sequences to access the conservation of these residues, as shown in Fig. 3. The residues were highly conserved across the referenced AOXs, except in two sequences, CAM84031.1 and AFO55203.1, that lacked one aromatic amino acid in their substrate pockets respectively, which may explain the reported higher *K*m values for small substrates like methanol [56, 59]. Therefore, we used the co-existence of all these activity-related residues as a criterion for filtering active enzymes (Supplementary Table S3).

Another feature of sequence is associated with the cellular localization of AOX from different phyla. The sequence of methylotrophic yeast (*Ascomycota*) contains a C-terminal peroxisomal targeting signal (PTS), a four-residue motif (660–663), which has been linked to the formation of the active form of AOX as octamers [66]. In contrast, AOXs from *Basidiomycota* were localized in the hyphal periplasmic space, the cell wall, and on extracellular tripartite membranes and slime [57], and typically lacked the PTS motif. Since the PTS motif is not directly related to enzyme activity, the PTS-related score was used as a categorization feature for sequences, rather than as a criterion for activity.

3.2. Construction of the candidate AOX library by filtering HMM-results using activity criterion

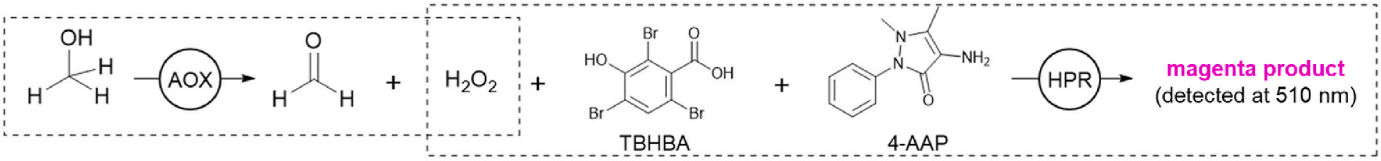
The genome database was first mined to expand the sequence spaces for potential AOXs. The referenced sequences were initially clustered at 90 % similarity, yielding 12 representative sequences. An HMM model was constructed using these sequences and searched against the UniProt reference proteomics database, resulting in a total of 30,240 sequences. Sequences shorter than 600 amino acids and longer than 700 were excluded, leaving 369 sequences. The distribution of sequence length versus HMM score, as shown in Supplementary Fig. S1, revealed that the sequence length criterion effectively removed a significant number of irrelevant fragment sequences.

The structural domains and structural similarities of these sequences were then assessed. The InterPro-Scan assignments showed that all the sequences contained the two typical structural domains of AOXs described earlier. The overall structural similarity between each pair of structures (Supplementary Fig. S2), evaluated by the TM-align score, was greater than 0.88. These results suggest that all these potential sequences share the typical structural features of AOXs.

To further confirm the presence of activity, sequence scoring rules (Supplementary Table S3) were applied to the sequences after MSA along with the referenced sequences. Only sequences satisfying all the activity criteria were retained, resulting in a final candidate library of 357 sequences, including 19 referenced sequences and 338 inferred sequences. The correlation between the sequence scoring results and HMM scores was analyzed (Supplementary Fig. S3), and while higher HMM scores helped filter out some irrelevant sequences, a few sequences with high HMM scores still lacked certain essential residues. This indicated that the activity criterion used here was more rigorous than the sequence match score alone in potential enzyme filtering.

3.3. Selection of experimental sequences according to sequence similarity network aligned with source organism taxonomy

The sequence similarity network (SSN) was employed to select sequences for experimental assays. A 60 % similarity threshold clustered the sequences into two groups (Supplementary Fig. S4), which aligned with the taxonomy at the phylum level. To refine the clustering further, the similarity threshold was increased to 85 %, resulting in class-level clusters, as shown in Fig. 4. Most of the reported active sequences are from *Saccharomycetes* [54,58–62,64,65], with a smaller number coming



AOX: alcohol oxidase, TBHBA: 2,4,6-tribromo-3-hydroxybenzoic acid, 4-AAP: 4-aminoantipyrine, HRP: horseradish peroxidase.

The cascade reaction for AOX activity assay.

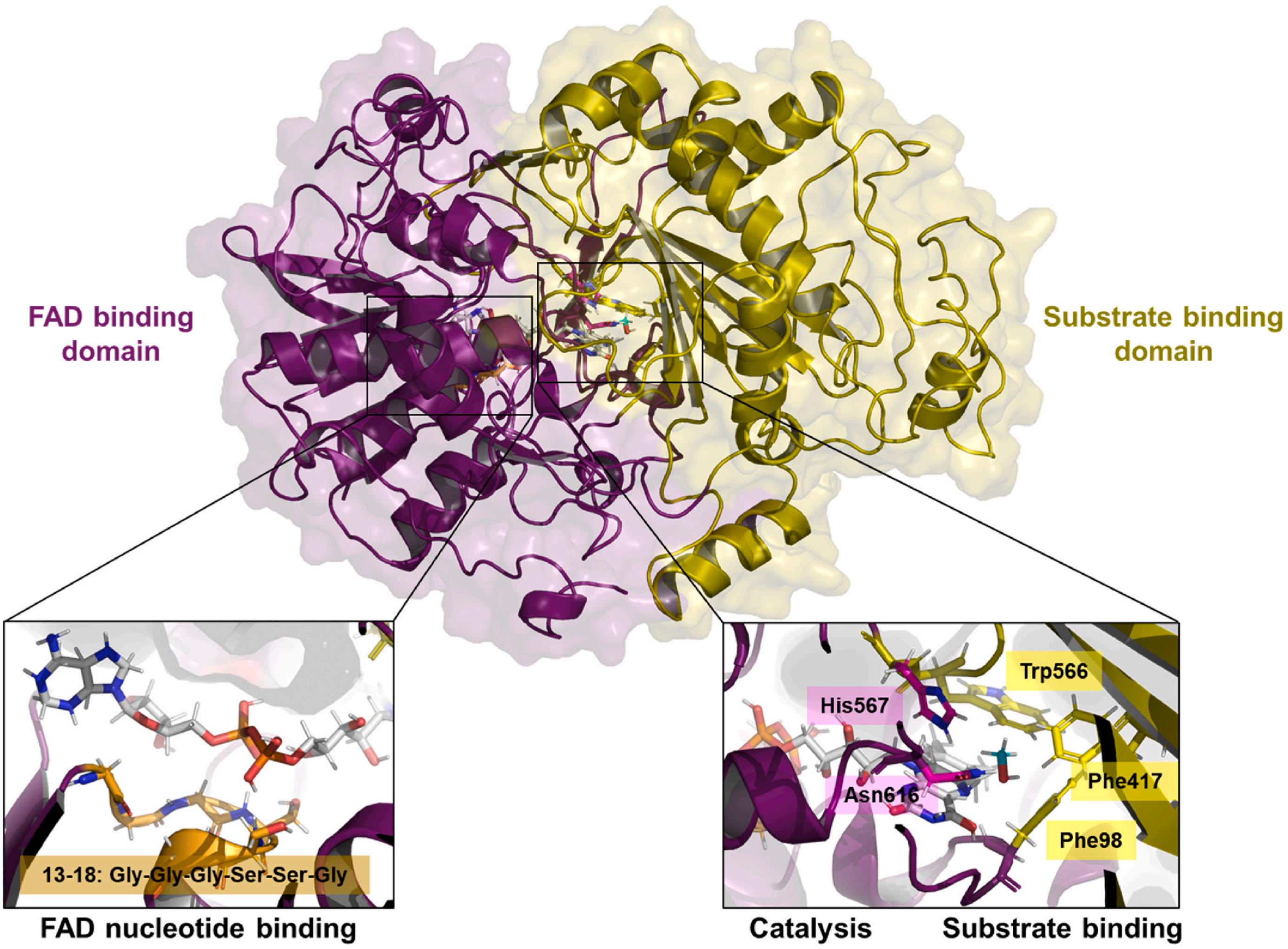


Fig. 2. Sequence and structural features of AOX. The structure of *Kpa*AOX (PDB: 5HSA.A, NCBI Accession: AAB57849.1) was used to illustrate the domain and residue features. The structural domains are distinguished by color: purple for the FAD-binding domains (residues 2–155, 192–306, 568–664), and yellow for the substrate-binding domains (residues 156–191, 307–567). Enlarged views of the FAD-binding and substrate-binding sites show key residues, labeled and colored as follows: purple for catalytic residues (His567, Asn616), yellow for substrate-binding residues (Phe98, Phe417, Trp566), and orange for FAD nucleotide-binding residues (13–18: GXGXXG). The FAD cofactor (white) is positioned as determined in the initial crystal structure. Methanol (cyan), the substrate, was docked into the substrate pocket using the PyMOL (v2.5.2) Plugin, DockingPie (v1.2) integrated with AutoDock Vina.

from *Agaricomycetes* [56,57], *Dothideomycetes* [53,55], and *Eurotiomycetes* [27,63].

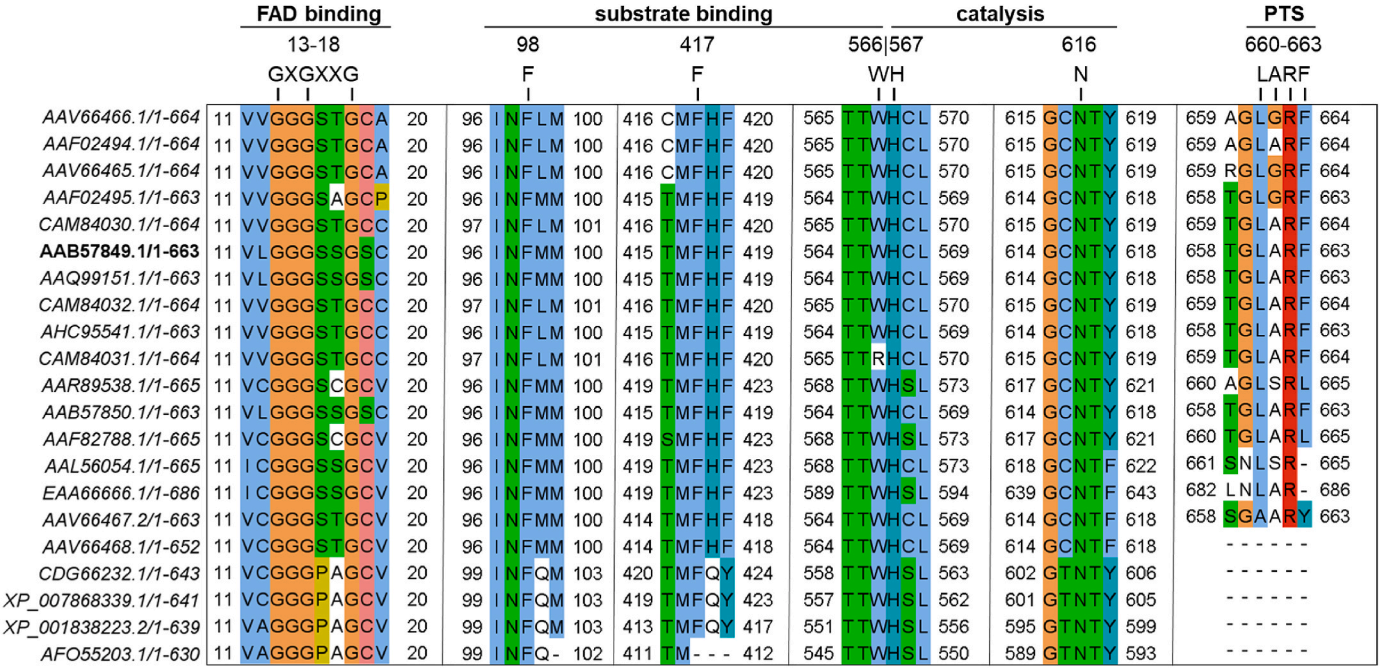
To quantify the source organisms and leverage the extensive data available in the BRENDA database, which primarily reported activity on organisms rather than the sequence, we developed scoring rules based on taxonomy. These rules favored source microorganisms with reported activity. Once the target source organisms were identified, a randomization strategy was employed for sequence selection. For each target class, 2–6 sequences were randomly selected for *in vitro* testing, as shown in Supplementary Fig. S5.

In the first round, 19 sequences were selected, primarily from *Saccharomycetes*, *Agaricomycetes*, *Eurotiomycetes* and *Dothideomycetes*—-classes that have already been characterized for active sequences. Although a high proportion of sequences exhibited measurable activity, the absolute activity values were low. To obtain highly active sequences, a second round of sampling was conducted, focusing on *Agaricomycetes*, which showed higher activity in the first round, as well as the unreported classes *Leotiomycetes* and *Sordariomycetes*. Surprisingly, enzymes with significantly higher activity were obtained from these unreported source organisms. The activities and protein yields of all characterized sequences are depicted in Fig. 5.

For the six newly identified highly active sequences, the substrate specificity characterization and enzyme kinetics analysis were conducted (Supplementary Fig. S6; Table 1 and Supplementary Fig. S7). These enzymes exhibited high specificity towards methanol, with activity rapidly decreasing as the carbon chain length increased, confirming their identity as methanol oxidases. The most active enzyme, *Cta*AOX (A0A4U6X6L6), demonstrated a K_M value of 3.6 mM for methanol, suggesting a highly efficient binding affinity. This enzyme represents a promising starting point for further engineering aimed to enhance enzymatic activity.

3.4. Retrospective analysis of experimental validation results reveals the utility of features

Compared to the control group, the proportion of new identified enzymes with measurable activity increased from 16.7 % to 58.1 % with statistical significance (Odds Ratio = 0.144, *p*-value = 0.0193). A retrospective analysis of the control group based on UniProt-annotations revealed that 7 of the 12 sequences could have been excluded in advance due to the absence of essential residues (Supplementary Table S4). Among the 104 sequences provided by UniProt, only 35 sequences (33.65 %) retained the required active sites. Even if all these sequences were soluble and active, the current workflow still demonstrated a



	FAD binding 13-18 GXGXXG	substrate binding 98 F	substrate binding 417 F	catalysis 566|567 WH	catalysis 616 N	PTS 660-663 LARF
AAV66466.1/1-664	11 VVGGGSTGCA 20	96 INFLM 100	416 CMFHF 420	565 TTWHCL 570	615 GCNTY 619	659 AGLGRF 664
AAF02494.1/1-664	11 VVGGGSTGCA 20	96 INFLM 100	416 CMFHF 420	565 TTWHCL 570	615 GCNTY 619	659 AGLARF 664
AAV66465.1/1-664	11 VVGGGSTGCA 20	96 INFLM 100	416 CMFHF 420	565 TTWHCL 570	615 GCNTY 619	659 RGLGRF 664
AAF02495.1/1-663	11 VVGGGSAGCP 20	96 INFMM 100	415 TMFHF 419	564 TTWHCL 569	614 GCNTY 618	658 TGLGRF 663
CAM84030.1/1-664	11 VVGGGSTGCC 20	97 INFLM 101	416 TMFHF 420	565 TTWHCL 570	615 GCNTY 619	659 TGLARF 664
AAB57849.1/1-663	11 VLGGGSSGSC 20	96 INFMM 100	415 TMFHF 419	564 TTWHCL 569	614 GCNTY 618	658 TGLARF 663
AAQ99151.1/1-663	11 VLGGGSSGSC 20	96 INFMM 100	415 TMFHF 419	564 TTWHCL 569	614 GCNTY 618	658 TGLARF 663
CAM84032.1/1-664	11 VVGGGSTGCC 20	97 INFLM 101	416 TMFHF 420	565 TTWHCL 570	615 GCNTY 619	659 TGLARF 664
AHC95541.1/1-663	11 VVGGGSTGCC 20	96 INFLM 100	415 TMFHF 419	564 TTWHCL 569	614 GCNTY 618	658 TGLARF 663
CAM84031.1/1-664	11 VVGGGSTGCC 20	97 INFLM 101	416 TMFHF 420	565 TTRHCL 570	615 GCNTY 619	659 TGLARF 664
AAR89538.1/1-665	11 VCGGGSCGCV 20	96 INFMM 100	419 TMFHF 423	568 TTWHSL 573	617 GCNTY 621	660 AGLSRL 665
AAB57850.1/1-663	11 VLGGGSSGSC 20	96 INFMM 100	415 TMFHF 419	564 TTWHCL 569	614 GCNTY 618	658 TGLARF 663
AAF82788.1/1-665	11 VCGGGSCGCV 20	96 INFMM 100	419 SMFHF 423	568 TTWHSL 573	617 GCNTY 621	660 TGLARL 665
AAL56054.1/1-665	11 ICGGGSSGCV 20	96 INFMM 100	419 TMFHF 423	568 TTWHCL 573	618 GCNTF 622	661 SNLSR- 665
EAA66666.1/1-686	11 ICGGGSSGCV 20	96 INFMM 100	419 TMFHF 423	589 TTWHSL 594	639 GCNTF 643	682 LNLAR- 686
AAV66467.2/1-663	11 VCGGGSTGCV 20	96 INFMM 100	414 TMFHF 418	564 TTWHCL 569	614 GCNTF 618	658 SGAARY 663
AAV66468.1/1-652	11 VCGGGSTGCV 20	96 INFMM 100	414 TMFHF 418	564 TTWHCL 569	614 GCNTF 618	------
CDG66232.1/1-643	11 VCGGGPAGCV 20	99 INFQM 103	420 TMFQY 424	558 TTWHSL 563	602 GTNTY 606	------
XP_007868339.1/1-641	11 VCGGGPAGCV 20	99 INFQM 103	419 TMFQY 423	557 TTWHSL 562	601 GTNTY 605	------
XP_001838223.2/1-639	11 VAGGGPAGCV 20	99 INFQM 103	413 TMFQY 417	551 TTWHSL 556	595 GTNTY 599	------
AFO55203.1/1-630	11 VAGGGPAGCV 20	99 INFQ- 102	411 TM--- 412	545 TTWHSL 550	589 GTNTY 593	------

Fig. 3. Multiple sequence alignment of 21 referenced AOXs. The residues are labeled using the AAB57849.1 sequence from *Komagataella pastoris* as the reference. The alignment is colored according to the Clustal style.

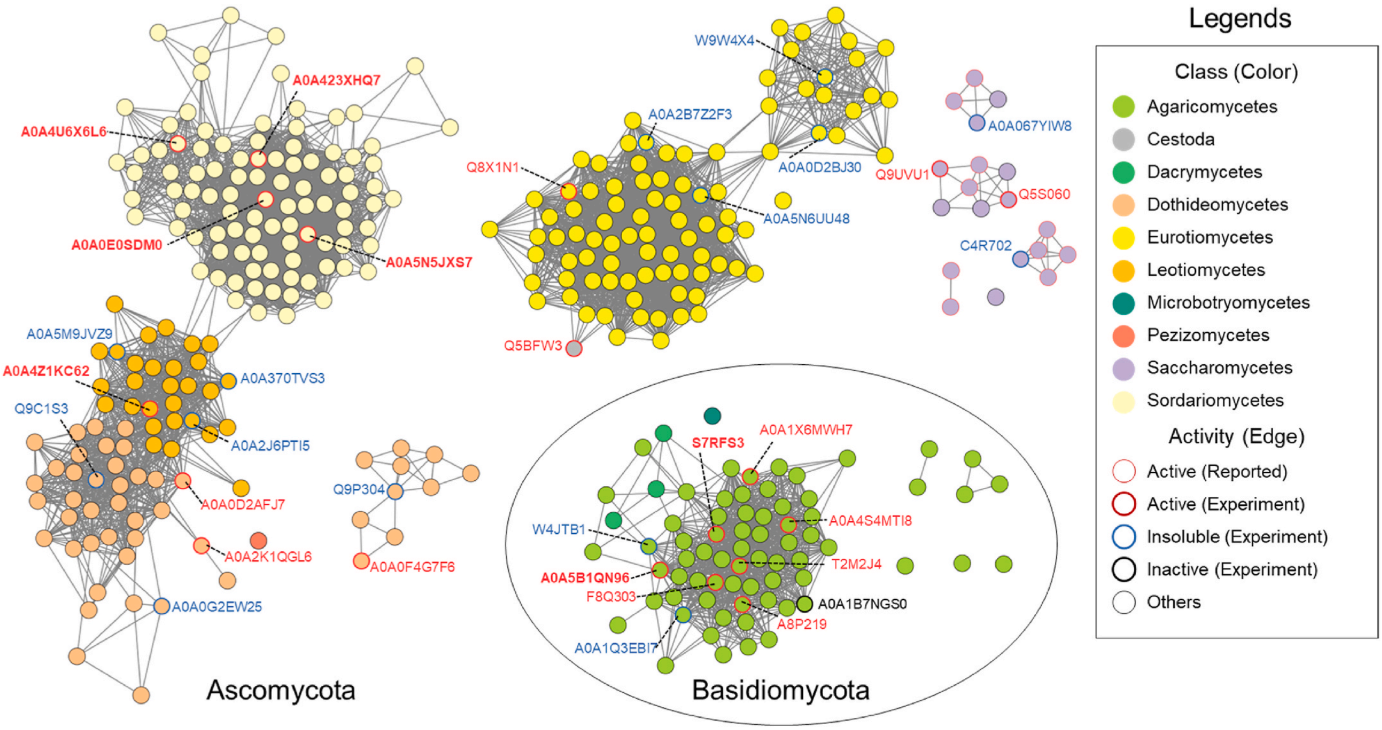


Fig. 4. Sequence similarity network of 357 candidate AOXs. The SSN is constructed at the 85 % similarity threshold. Nodes are colored according to the class of the source organisms, with the borders of experimentally selected nodes highlighted in bold. The border color of the nodes indicates activity: red for measured or reported activity, blue for insoluble enzymes, and black for soluble but inactive enzymes. UniProt Accessions for the experimental sequences are labeled, with active ones colored in red, insoluble sequences in blue, and soluble/inactive sequences in black. Accessions with activity greater than 1 U/mg are bolded.

significantly higher probability of identifying active sequences (Odds Ratio = 0.366, *p*-value = 0.0207). These results validate the effectiveness of incorporating the activity criterion into the virtual screening process for sequence selection.

The proportion of newly discovered enzymes with high activity (more than 1 U/mg) increased from 8.3 % to 19.4 % (Odds Ratio = 0.379, *p*-value = 0.652). Although the overall improvement in the probability of identifying highly active enzymes was not statistically significant, an analysis based on enzyme classes (Supplementary Fig. S8A) revealed that enzymes from *Sordariomycetes* exhibited a statistically significant higher proportion of high activity compared to other reported classes (*p*-value <0.05). Notably, the highest enzyme

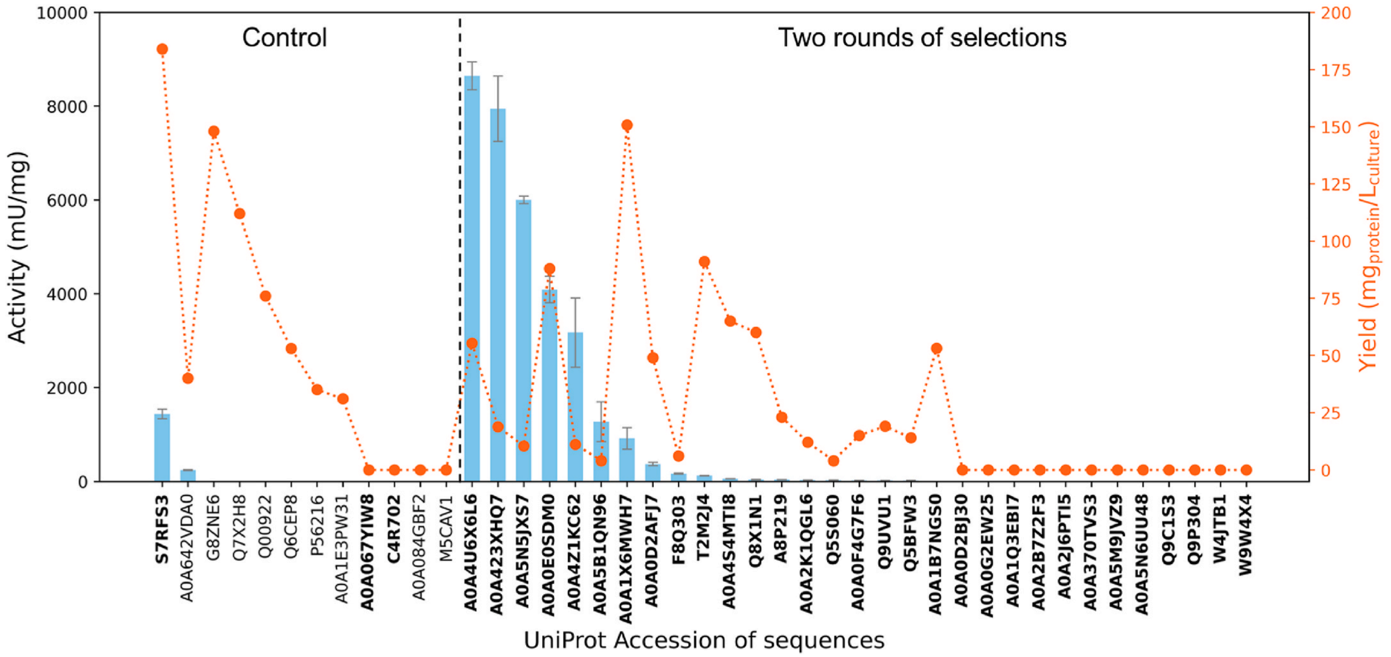


Fig. 5. Activity and protein yield of experimental sequences. The bolded accessions indicate sequences that belong to the candidate sequences. The activity of each AOX was determined through three parallel experiments.

Table 1
Kinetic parameters of the most active AOXs from the two rounds of selections.

UniProt Accession	k_{cat} (s^{-1})	K_M (mM)	k_{cat}/K_M (s^{-1} mM^{-1})	Organism
A0A4U6X6L6	9.9 ± 0.1	3.6 ± 0.2	2.7	*Colletotrichum tanaceti*
A0A423XHQ7	8.6 ± 0.4	22.9 ± 2.6	0.4	*Cytospora leucostoma*
A0A5N5JXS7	7.0 ± 0.6	35.6 ± 6.9	0.2	*Coniochaeta* sp. 2T2.1
A0A0E0SDM0	4.1 ± 0.1	7.2 ± 0.8	0.6	*Gibberella zeae*
A0A4Z1KC62	3.1 ± 0.2	20.1 ± 4.5	0.2	*Botrytis elliptica*
A0A5B1QN96	1.5 ± 0.1	30.8 ± 4.5	0.05	*Dentipellis* sp. KUC8613

activity recorded in this study (*Cta*AOX, A0A4U6X6L6, 8.6 U/mg) surpassed the activity of a previously engineered enzyme (*Gtr*AOX-M2, 4.8 U/mg) obtained through two rounds of directed evolution [30] and was comparable to the commercial AOX from yeast (*Ppa*AOX, 10–40 U/mg). This demonstrates that taxonomy feature-based classification could systematically aid in identifying families of enzymes with high activity.

Unlike the taxonomy-based class labels, taxonomic scoring developed based on published literature was not effective in guiding the identification of active recombinant enzymes, either at the inter-class or intra-class level (Supplementary Fig. S8B). In this study, the most active enzymes were obtained from *Sordariomycetes* and *Leotiomycetes*, two classes with no prior reports of active sequences. Consequently, these enzymes had very low taxonomic scores. Furthermore, for previously reported active sequences, most failed to produce soluble proteins, or their soluble forms exhibited only minimal activity (Supplementary Fig. S8C). Differences in expression chassis [67,68] may be an important factor. SDS-PAGE analysis confirmed protein expression in all samples (Supplementary Fig. S9), regardless of whether the proteins were soluble or formed precipitates. The primary location of the protein correlated well with its expression levels. Across different expression levels, both highly active and low-activity enzymes were observed, suggesting that low activity may result from either misfolding/aggregation or the activity absence. More than half of the proteins exhibited both low soluble expression levels and low activity, which could be attributed to the formation of inclusion bodies during expression.

AOXs from *Ascomycota* are known to require assembly and translocation into the peroxisome to achieve full functionality [69], which cannot be fulfilled in the *E. coli* chassis. This process is mediated by the PTS motif. Therefore, for naturally PTS-deficient sequences, their assembly may not rely on peroxisome-mediated maturation, making them more likely to achieve high soluble expression in *E. coli*. The correlation of protein yields with the integrity of the PTS motif (Supplementary Fig. S10) showed that all sequences with high protein yields (>25 mg/L) were PTS-deficient. However, the correlation between expression levels and the presence of the PTS motif was not statistically significant (*p*-value = 0.0770), suggesting that other factors may also be associated with expression levels.

4. Discussion

Enzyme discovery serves as the foundation for further molecular engineering of enzymes [70], such as directed evolution [71] and rational design [72]. With advancements in data technology, these molecular engineering approaches are becoming increasingly predictive and programmable [73,74]. In this study, we presented an enzyme discovery workflow characterized by the sequence and taxonomic feature evaluation. These features were constructed based on referenced sequences, considering sequence and source organism information related to activity and expression. Based on the distribution, these features could be divided into two categories. The first category includes features that are consistent across active enzymes, such as sequence length, overall structural family, domain composition, and the presence of key catalytic and binding sites, as used in this study. These features also have clear functional relevance, making them useful as activity criteria to exclude irrelevant sequences. In contrast, the second category is more diverse, such as the phylogenetic classification of the source organisms. Attempts to use taxonomy-based scoring to guide the discovery of enzymes with high activity and good expression have proven unsuccessful. As an alternative approach, systematically sampling based on the taxonomy has led to the identification of new enzyme sources with high activity [75]. Recently, Seo et al. [76] applied sequence similarity networks constructed from natural sequences along with stratified sampling and cluster sampling methods to discover PET depolymerases with high fitness, highlighting the importance of global landscape profiling and systematic sampling.

Enzyme discovery often relies on custom pipelines that integrate various tools. This study demonstrates an example of leveraging function-related information to discover new enzymes, achieving accuracy comparable to approaches that employ more complex integrated evaluation strategies [77]. The current data-mining workflow is compatible with additional evaluation methods, such as soluble expression prediction models [78] and machine learning models focused on activity prediction [79], and thus holds promise for extension to precise quantitative functional assessments.

5. Conclusion

A new enzyme discovery workflow facilitated by sequence and taxonomic feature evaluation was proposed and demonstrated for mining highly active AOXs that are expressible in *E. coli*. Based on the analysis of sequence features of the 21 referenced sequences collected, a candidate enzyme library comprising 357 sequences was constructed. Of those, 34 sequences were experimentally validated, with 19 showing measurable activity and 7 exhibiting activity more than 1 U/mg. The sequence similarity network-guided mining identified two new classes—*Sordariomycetes* and *Leotiomycetes*—as promising sources for highly active AOXs. A comparison with workflows lacking *in silico* activity assessment, which was based on the UniProt annotation, showed that the sequence and taxonomic feature evaluation significantly increased the probability of obtaining highly active enzymes. The workflow has been released as open source, providing a foundation for further development and adaptation to other enzymes and tailored scoring rules.

CRediT authorship contribution statement

Yilei Han: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Formal analysis, Data curation, Conceptualization. **Xuwei Ding:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Conceptualization. **Junjian Tan:** Data curation. **Yajuan Sun:** Methodology, Data curation. **Yunjiang Duan:** Data curation. **Zheng Liu:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Gaowei Zheng:** Writing – review & editing, Validation, Supervision, Funding acquisition, Conceptualization. **Diannan Lu:** Writing – review & editing, Validation, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by National Key Research and Development Program of China (no. 2022YFC2105900).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.synbio.2025.04.014.

References

[1] Dauparas J, et al. Robust deep learning–based protein sequence design using ProteinMPNN. Science 2022;378(6615):49–56.
[2] Yu T, et al. Enzyme function prediction using contrastive learning. Science 2023;379(6639):1358–63.
[3] Madani A, et al. Large language models generate functional protein sequences across diverse families. Nat Biotechnol 2023;41(8):1099–106.
[4] Robinson SL, Piel J, Sunagawa S. A roadmap for metagenomic enzyme discovery. Nat Prod Rep 2021;38:1994–2023.
[5] Vanacek P, et al. Exploration of enzyme diversity by integrating bioinformatics with expression analysis and biochemical characterization. ACS Catal 2018;8(3):2402–12.
[6] Hon J, et al. EnzymeMiner: automated mining of soluble enzymes with diverse structures, catalytic properties and stabilities. Nucleic Acids Res 2020;48(W1):W104–9.
[7] Shi Z, et al. REME: an integrated platform for reaction enzyme mining and evaluation. Nucleic Acids Res 2024;52(W1):W299–305.
[8] Song Y, et al. Accurately predicting enzyme functions through geometric graph learning on ESMFold-predicted structures. Nat Commun 2024;15(1):8180.
[9] Yu T, et al. Machine learning-enabled retrobiosynthesis of molecules. Nat Catal 2023;6(2):137–51.
[10] Schnoes AM, et al. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol 2009;5(12):e1000605.
[11] Radivojac P, et al. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10(3):221–7.
[12] Jiang Y, et al. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 2016;17(1):184.
[13] Zhou N, et al. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol 2019;20(1):244.
[14] Holliday GL, et al. Evaluating functional annotations of enzymes using the gene ontology. In: Dessimoz C, Škunca N, editors. The gene ontology handbook. New York, NY: Springer New York; 2017. p. 111–32.
[15] Chang A, et al. BRENDA, the ELIXIR core data resource in 2021: new developments and updates. Nucleic Acids Res 2021;49(D1):D498–508.
[16] Gerlt JA. Genomic enzymology: web tools for leveraging protein family sequence–function space and genome context to discover novel functions. Biochemistry 2017;56(33):4293–308.
[17] Zallot R, Oberg N, Gerlt JA. The EFI web resource for genomic enzymology tools: leveraging protein, genome, and metagenome databases to discover novel enzymes and metabolic pathways. Biochemistry 2019;58(41):4169–82.
[18] Allen KN, Whitman CP. The birth of genomic enzymology: discovery of the mechanistically diverse enolase superfamily. Biochemistry 2021;60(46):3515–28.
[19] Knox HL, Allen KN. Expanding the viewpoint: leveraging sequence information in enzymology. Curr Opin Chem Biol 2023;72:102246.
[20] Atkinson HJ, et al. Using sequence similarity networks for visualization of relationships across diverse protein superfamilies. PLoS One 2009;4(2):e4345.
[21] Zhao S, et al. Prediction and characterization of enzymatic activities guided by sequence similarity and genome neighborhood networks. Elife 2014;3:e03275.
[22] Paysan-Lafosse T, et al. InterPro in 2022. Nucleic Acids Res 2023;51(D1):D418–27.
[23] Rentzsch R, Orengo CA. Protein function prediction using domain families. BMC Bioinf 2013;14(Suppl 3):S5.
[24] de Rond T, Asay JE, Moore BS. Co-occurrence of enzyme domains guides the discovery of an oxazolone synthetase. Nat Chem Biol 2021;17(7):794–9.
[25] Ribeiro AJM, et al. Mechanism and Catalytic Site Atlas (M-CSA): a database of enzyme reaction mechanisms and active sites. Nucleic Acids Res 2018;46(D1):D618–23.
[26] Ribeiro AJ, et al. A global analysis of function and conservation of catalytic residues in enzymes. J Biol Chem 2020;295(2):314–24.
[27] Goswami P, et al. An overview on alcohol oxidases and their potential applications. Appl Microbiol Biotechnol 2013;97:4259–75.
[28] Pawlik A, Stefanek S, Janusz G. Properties, physiological functions and involvement of basidiomycetous alcohol oxidase in wood degradation. Int J Mol Sci 2022;23(22):13808.
[29] Cai T, et al. Cell-free chemoenzymatic starch synthesis from carbon dioxide. Science 2021;373(6562):1523–7.
[30] Ding X-W, et al. De novo multienzyme synthetic pathways for lactic acid production. ACS Catal 2024;14(7):4665–74.
[31] Zhou J, et al. Three multi-enzyme cascade pathways for conversion of C1 to C2/C4 compounds. Chem Catal 2022;2(10):2675–90.
[32] Zhang J, et al. Hybrid synthesis of polyhydroxybutyrate bioplastics from carbon dioxide. Green Chem 2023;25(8):3247–55.
[33] Liu J, et al. Turn air-captured CO_2 with methanol into amino acid and pyruvate in an ATP/NAD(P)H-free chemoenzymatic system. Nat Commun 2023;14(1):2772.
[34] Lundberg DJ, et al. Concerted methane fixation at ambient temperature and pressure mediated by an alcohol oxidase and Fe-ZSM-5 catalytic couple. Nat Catal 2024;7(12):1359–71.
[35] Koch C, et al. Crystal structure of alcohol oxidase from *Pichia pastoris*. PLoS One 2016;11(2):e0149846.
[36] Vonck J, Parcej DN, Mills DJ. Structure of alcohol oxidase from *Pichia pastoris* by cryo-electron microscopy. PLoS One 2016;11(7):e0159476.
[37] Nguyen Q-T, et al. Structure-based engineering of *Phanerochaete chrysosporium* alcohol oxidase for enhanced oxidative power toward glycerol. Biochemistry 2018;57(43):6209–18.

[38] Wu B, et al. Structure-based redesign of a methanol oxidase into an "aryl alcohol oxidase" for enzymatic synthesis of aromatic flavor compounds. J Agric Food Chem 2023;71(16):6406–14.

[39] UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. Nucleic Acids Res 2023;51(D1):D523–31.

[40] Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.

[41] Di Tommaso P, et al. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 2011;39(suppl_2):W13–7.

[42] Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 2011;39(suppl_2):W29–37.

[43] Potter SC, et al. HMMER web server: 2018 update. Nucleic Acids Res 2018;46(W1): W200–4.

[44] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 2010;26(19):2460–1.

[45] Waterhouse A, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res 2018;46(W1):W296–303.

[46] Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302–9.

[47] Waterhouse AM, et al. Jalview Version 2—a multiple sequence alignment editor and analysis workbench. Bioinformatics 2009;25(9):1189–91.

[48] Shannon P, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 2003;13(11):2498–504.

[49] DeLano WL. Pymol: an open-source molecular graphics tool. CCP4 Newsl. Protein Crystallogr 2002;40(1):82–92.

[50] Rosignoli S, Paiardini A. DockingPie: a consensus docking plugin for PyMOL. Bioinformatics 2022;38(17):4233–4.

[51] Trott O, Olson AJ. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem 2010;31(2):455–61.

[52] Schoch CL, et al. NCBI Taxonomy: a comprehensive update on curation, resources and tools. Database 2020:baaa062.

[53] Soldevila AI, Ghabrial SA. A novel alcohol oxidase/RNA-binding protein with affinity for mycovirus double-stranded RNA from the filamentous Fungus *Helminthosporium* (*cochliobolus*) *victoriae*: molecular and functional characterization. J Biol Chem 2001;276(7):4652–61.

[54] Szamecz B, et al. Identification of four alcohol oxidases from methylotrophic yeasts. Yeast 2005;22(8):669–76.

[55] Segers G, et al. Alcohol oxidase is a novel pathogenicity factor for *Cladosporium fulvum*, but aldehyde dehydrogenase is dispensable. Mol Plant Microbe Interact 2001;14(3):367–77.

[56] de Oliveira BV, et al. A potential role for an extracellular methanol oxidase secreted by *Moniliophthora perniciosa* in Witches' broom disease in cacao. Fungal Genet Biol 2012;49(11):922–32.

[57] Daniel G, et al. Characteristics of *Gloeophyllum trabeum* alcohol oxidase, an extracellular source of H2O2 in brown rot decay of wood. Appl Environ Microbiol 2007;73(19):6241–53.

[58] Ledeboer A, et al. Molecular cloning and characterization of a gene coding for methanol oxidase in *Hansenula polymorpha*. Nucleic Acids Res 1985;13(9): 3063–82.

[59] Dmytruk KV, et al. Isolation and characterization of mutated alcohol oxidases from the yeast *Hansenula polymorpha* with decreased affinity toward substrates and their use as selective elements of an amperometric biosensor. BMC Biotechnol 2007;7: 33.

[60] Ellis SB, et al. Isolation of alcohol oxidase and two other methanol regulatable genes from the yeast *Pichia pastoris*. Mol Cell Biol 1985;5(5):1111–21.

[61] Koutz P, et al. Structural comparison of the *Pichia pastoris* alcohol oxidase genes. Yeast 1989;5(3):167–77.

[62] Promdonkoy P, et al. Methanol-inducible promoter of thermotolerant methylotrophic yeast *Ogataea thermomethanolica* BCC16875 potential for production of heterologous protein at high temperatures. Curr Microbiol 2014;69: 143–8.

[63] Holzmann K, Schreiner E, Schwab H. A *Penicillium chrysogenum* gene (aox) identified by specific induction upon shifting pH encodes for a protein which shows high homology to fungal alcohol oxidases. Curr Genet 2002;40:339–44.

[64] Raymond CK, et al. Development of the methylotrophic yeast *Pichia methanolica* for the expression of the 65 kilodalton isoform of human glutamate decarboxylase. Yeast 1998;14(1):11–23.

[65] Ozimek P, Veenhuis M, van der Klei IJ. Alcohol oxidase: a complex peroxisomal, oligomeric flavoprotein. FEMS Yeast Res 2005;5(11):975–83.

[66] Waterham HR, et al. Peroxisomal targeting, import, and assembly of alcohol oxidase in *Pichia pastoris*. J Cell Biol 1997;139(6):1419–31.

[67] Jiang R, et al. Strategies to overcome the challenges of low or no expression of heterologous proteins in *Escherichia coli*. Biotechnol Adv 2024;75:108417.

[68] Montgomery SL, et al. Characterization of imine reductases in reductive amination for the exploration of structure-activity relationships. Sci Adv 2020;6(21): eaay9320.

[69] Ozimek P, et al. Pyruvate carboxylase is an essential protein in the assembly of yeast peroxisomal oligomeric alcohol oxidase. Mol Biol Cell 2003;14(2):786–97.

[70] Trudeau DL, Tawfik DS. Protein engineers turned evolutionists—the quest for the optimal starting point. Curr Opin Biotechnol 2019;60:46–52.

[71] Wu Z, et al. Machine learning-assisted directed protein evolution with combinatorial libraries. Proc Natl Acad Sci U S A 2019;116(18):8852–8.

[72] Reetz M. Making enzymes suitable for organic chemistry by rational protein design. Chembiochem 2022;23(14):e202200049.

[73] Lovelock SL, et al. The road to fully programmable protein catalysis. Nature 2022; 606(7912):49–58.

[74] Nestl BM, et al. The development and opportunities of predictive biotechnology. Chembiochem 2024;25(13):e202300863.

[75] Fisher BF, et al. Site-selective C–H halogenation using flavin-dependent halogenases identified via family-wide activity profiling. ACS Cent Sci 2019;5(11): 1844–56.

[76] Seo H, et al. Landscape profiling of PET depolymerases using a natural sequence cluster framework. Science 2025;387(6729):eadp5637.

[77] Johnson SR, et al. Computational scoring and experimental evaluation of enzymes generated by neural networks. Nat Biotechnol 2024;43(3):396–405.

[78] Hon J, et al. SoluProt: prediction of soluble protein expression in *Escherichia coli*. Bioinformatics 2021;37(1):23–8.

[79] Kroll A, et al. Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning. Nat Commun 2023;14(1):4139.