International Journal of Engineering & Tech Development

Volume 2, Issue 9 – September 2026 • E-ISSN: 3067-672X P-ISSN: 3067-669X

AI-Enabled Discovery of Novel Enzymes: Applications in Biotechnology, Food Systems and Green Manufacturing

Timothy Patrick Jenkins

Associate Professor, Technical University of Denmark (DTU), Denmark

Abstract

Enzymes provide selective and comparatively mild routes for transforming biological and chemical materials. Their industrial use, however, is limited by the difficulty of identifying catalysts that remain active, selective, stable, and economically viable under realistic processing conditions. Conventional enzyme discovery depends heavily on cultivable microorganisms, sequence similarity, manual screening, and iterative experimentation. These approaches have generated valuable biocatalysts, but they examine only a small proportion of natural sequence diversity and can become expensive when candidate libraries are large. Artificial intelligence offers a complementary strategy by learning relationships among protein sequence, structure, catalytic function, substrate preference, and process conditions.

This paper examines an AI-enabled enzyme-discovery framework integrating metagenomic exploration, sequence representation, function prediction, structural modeling, candidate ranking, automated screening, and experimental validation. A structured integrative review is combined with a transparent conceptual simulation. The analysis considers applications in industrial biotechnology, food processing, biomass conversion, polymer degradation, low-temperature manufacturing, and environmentally preferable chemical synthesis. It also examines challenges involving incomplete annotations, biased databases, model interpretability, enzyme–substrate interactions, scale-up, biosafety, intellectual property, and responsible use of environmental genetic resources.

A simulated screening portfolio of 20 enzyme candidates is used to illustrate how predicted thermal stability and catalytic activity could support candidate prioritization. The modeled relationship is illustrative and does not represent laboratory measurements. The study finds that AI can reduce an initially vast search space, but computational predictions cannot establish enzyme novelty, activity, safety, or industrial performance without biochemical verification. Reliable discovery therefore requires a closed-loop system in which experimental results continually refine computational models. The paper concludes that the strongest contribution of AI lies not in replacing enzymology but in directing limited experimental resources toward scientifically plausible and industrially relevant candidates while maintaining transparent validation, sustainability assessment, and human oversight.

Keywords: artificial intelligence; biocatalysis; enzyme discovery; enzyme engineering; food biotechnology; green manufacturing; machine learning; metagenomics

1. Introduction

Enzymes are biological catalysts that accelerate chemical reactions with high levels of substrate specificity and reaction selectivity. They are widely used in food processing, pharmaceutical manufacturing, diagnostics, agriculture, textiles, detergents, biofuels, pulp and paper processing, waste treatment, and fine-chemical synthesis. Unlike many conventional catalysts, enzymes frequently operate in water, at moderate temperatures, and under relatively mild pressure and pH conditions. These characteristics can reduce energy consumption, unwanted by-products, and dependence on hazardous reagents.

Industrial enzymes must nevertheless satisfy requirements that differ substantially from those encountered in their natural environments. A naturally occurring enzyme may lose activity at the temperature required by a manufacturing process. It may become unstable in organic solvents, respond poorly to changes in pH, exhibit insufficient substrate specificity, or produce an undesirable stereochemical outcome. Other enzymes may have valuable functions but remain difficult to express, purify, immobilize, or reuse.

The discovery of suitable enzymes has traditionally relied on cultivated microorganisms, biochemical assays, sequence-homology searches, and directed evolution. These methods remain scientifically essential, but each has limitations. Only a fraction of environmental microorganisms can be cultivated using standard laboratory procedures. Homology-based searches favor proteins resembling already characterized enzymes and may overlook distant families. Experimental screening can evaluate functional activity directly, but testing extremely large sequence libraries requires substantial time, instrumentation, reagents, and labor.

Metagenomic sequencing has expanded access to genetic material from microbial communities without requiring each organism to be cultured. Environmental samples collected from soils, oceans, compost, animal digestive systems, industrial residues, hot springs, saline environments, and contaminated sites can contain millions of previously uncharacterized genes. This diversity creates significant opportunity, but it also produces a computational bottleneck. A sequencing project may generate far more candidate proteins than a laboratory can express and test.

Artificial intelligence can help manage this imbalance between sequence availability and experimental capacity. Machine-learning models can analyze amino-acid sequences, predicted protein structures, catalytic motifs, evolutionary patterns, physicochemical properties, and substrate information. The resulting predictions can support enzyme classification, active-site recognition, stability estimation, substrate matching, and candidate ranking. Generative models and machine-learning-guided directed evolution can also propose variants expected to improve selected characteristics.

AI-enabled discovery does not eliminate the need for experimentation. Protein function emerges from complex interactions involving folding, dynamics, cofactors, substrates, solvents, temperature, pH, expression systems, and process design. A computationally promising candidate may fail to express or may be inactive under laboratory conditions. Conversely, a sequence with low similarity to known enzymes may exhibit valuable and unexpected catalytic activity. Computational prediction should therefore be positioned as a prioritization layer within an iterative discovery and validation cycle.

2. Review of Literature

2.1 From Environmental Enzyme Mining to Metagenomic Discovery

Microorganisms represent a major source of industrial enzymes because they occupy chemically and physically diverse environments. Their catalytic systems support nutrient acquisition, cellular metabolism, stress tolerance, defense, and environmental adaptation. Microbial enzymes can consequently possess characteristics such as heat tolerance, cold activity, salt tolerance, pressure resistance, solvent stability, or activity under acidic and alkaline conditions.

Traditional discovery begins with the isolation and cultivation of microorganisms, followed by screening for a desired biochemical activity. This strategy has produced commercially important amylases, proteases, lipases, cellulases, pectinases, and other catalysts. Its main limitation is that laboratory cultivation accesses only part of microbial diversity. Environmental conditions and microbial interactions may be difficult to reproduce, while slow-growing or symbiotic organisms may remain inaccessible.

Metagenomics addresses this limitation by extracting and sequencing DNA directly from environmental samples. Sequence-based metagenomics searches for genes with motifs or similarities associated with known functions. Functional metagenomics places environmental DNA fragments in expression hosts and screens the resulting clones for observable activity. The sequence-based route is scalable but dependent on existing annotations. Functional screening can reveal unexpected activities but may fail when environmental genes are poorly expressed by the selected host.

Zarafeta et al. demonstrated how metagenomic investigation of a thermophilic environment could identify esterolytic enzymes with characteristics relevant to industrial conditions. Uria et al. reviewed the mining of commercially useful enzymes through metagenomics and emphasized the importance of linking sequence exploration with functional characterization. Such studies establish environmental DNA as a major discovery resource, while also showing that sequencing alone cannot establish catalytic usefulness.

Environmental sampling should be guided by biological reasoning. Compost and decaying plant material may be productive sources of lignocellulose-degrading enzymes. Cold oceans can contain cold-active enzymes suitable for low-temperature processing. Saline and alkaline habitats may yield catalysts that tolerate conditions relevant to detergents, food processing, and chemical manufacturing. Ecological metadata therefore provide informative features rather than merely describing the origin of a sequence.

2.2 Computational Prediction and Machine Learning

Early computational enzyme annotation relied strongly on sequence alignment, conserved motifs, protein domains, and phylogenetic relationships. Tools such as BLAST and DIAMOND remain valuable because homologous proteins frequently share functional characteristics. However, sequence similarity does not always indicate identical catalytic activity. Small changes in an active site can alter substrate preference, while structurally related proteins may perform different reactions.

Machine learning expands the range of computable relationships. Models can combine amino-acid composition, sequence patterns, evolutionary conservation, predicted secondary structure, three-dimensional structure, physicochemical descriptors, and functional annotations. Supervised models learn from labeled enzymes, whereas unsupervised and self-supervised models extract representations from large collections of unlabeled protein sequences.

Protein language models treat amino-acid sequences as structured symbolic data. By learning which residues tend to occur in particular contexts, these models develop representations that can support

downstream tasks such as function prediction, localization, stability assessment, and mutation-effect estimation. Work by Alley et al., Rao et al., Rives et al., and Yang et al. demonstrated that learned protein representations can capture biologically meaningful properties and improve prediction even when labeled experimental data are limited.

2.3 Industrial Enzyme Engineering and Application Requirements

The discovery of a natural enzyme is often followed by engineering. Directed evolution generates protein variants and selects those with improved properties. Rational design modifies residues based on structural and mechanistic knowledge. Semi-rational design combines focused variation with computational prediction. Bornscheuer et al. described enzyme engineering as a major stage in the development of practical biocatalysis, while Arnold emphasized the ability of directed evolution to produce functions not readily achieved through purely rational reasoning.

Machine learning can support engineering by learning sequence–activity relationships from experimental variants. Rather than testing every possible mutation, a model can prioritize combinations expected to improve stability, activity, specificity, or expression. This process can be iterative: selected variants are synthesized and tested, experimental results are added to the training set, and the updated model recommends another group.

Application requirements vary considerably. Food-processing enzymes must satisfy safety, regulatory, sensory, and allergenicity considerations. A catalyst suitable for biomass conversion may need thermostability and resistance to inhibitors. An enzyme used in pharmaceutical synthesis may require exceptional stereoselectivity. A detergent enzyme must tolerate surfactants, oxidants, and repeated temperature changes. Therefore, a universally “best” enzyme does not exist; performance is conditional on the intended process.

Table 1: Evidence domains supporting AI-enabled enzyme discovery

Evidence domain	Primary contribution	Typical output	Important limitation
Metagenomic sequencing	Expands access to uncultured microbial diversity	Candidate protein sequences	Function remains unverified
Functional screening	Detects catalytic activity directly	Active clones or purified enzymes	Expression-host and throughput constraints
Sequence modeling	Learns relationships among residues and protein families	Functional or property predictions	Training-data bias and annotation errors
Structural modeling	Estimates folds, pockets, and residue interactions	Structural hypotheses	Does not fully capture catalytic dynamics

Evidence domain	Primary contribution	Typical output	Important limitation
Automated experimentation	Tests prioritized candidates at scale	Activity and stability measurements	Requires suitable assays and quality control
Process engineering	Evaluates performance under operational conditions	Yield, productivity, and robustness	Laboratory success may not survive scale-up

The literature supports a hybrid discovery model. Computational systems can expand and organize the search space, while experiments establish whether predicted activities are real and reproducible. Industrial translation then requires process-level evaluation, regulatory review, sustainability assessment, and economic analysis.

3. Objectives

The main objective of this study is to examine how artificial intelligence can improve the identification, evaluation, and engineering of novel enzymes for biotechnology, food systems, and environmentally preferable manufacturing. The research treats enzyme discovery as a connected sequence of biological sampling, computational prediction, experimental testing, and industrial translation rather than as an isolated classification task.

A further objective is to distinguish scientifically defensible AI support from exaggerated claims of autonomous discovery. The framework recognizes that an enzyme becomes a credible discovery only after its sequence, expression, biochemical function, novelty, reproducibility, and application relevance have been established through appropriate evidence.

The specific objectives are:

- To examine how metagenomic and environmental sequence resources can expand the searchable space of potential industrial enzymes.
- To assess the contributions of sequence analysis, protein language models, structural prediction, and supervised machine learning to enzyme-function prediction.
- To explain how AI-based ranking can allocate limited experimental resources to candidates with stronger predicted relevance.
- To evaluate enzyme requirements across biotechnology, food processing, biomass conversion, and green chemical manufacturing.
- To identify the experimental evidence needed to verify computationally predicted enzyme activity and novelty.
- To examine the risks created by incomplete annotations, biased databases, model uncertainty, intellectual-property conflicts, and inappropriate sustainability claims.
- To develop a closed-loop discovery model connecting computational screening with laboratory and process-level feedback.

- To illustrate candidate prioritization through a clearly labeled simulation without presenting the values as experimental observations.

Five research questions guide the paper:

- How can AI extract useful functional evidence from large and incompletely annotated protein-sequence repositories?
- Which computational and experimental stages are required to validate a putatively novel enzyme?
- How do the performance requirements of biotechnology, food systems, and green manufacturing differ?
- Which technical and governance risks may undermine AI-enabled enzyme discovery?
- How can computational prioritization and experimental learning be integrated into a reproducible discovery pipeline?

4. Research Methodology

4.1 Evidence Identification and Selection

The paper employed a structured integrative-review design. Relevant literature was identified from PubMed, Web of Science-oriented scholarly sources, Crossref-indexed publications, Google Scholar, and publisher databases covering biotechnology, bioinformatics, enzymology, food science, biochemical engineering, and sustainable manufacturing. The final search was conducted on August 25, 2026, while the eligibility boundary restricted cited research to publications listed in the References.

Search combinations included artificial intelligence enzyme discovery, machine learning enzyme engineering, protein language model function prediction, metagenomic enzyme mining, novel biocatalyst discovery, enzyme food processing, green biocatalysis, enzyme stability prediction, and high-throughput enzyme screening. Reference lists of relevant reviews and foundational methodological studies were also examined.

Eligible evidence included peer-reviewed research and review articles that addressed enzyme mining, protein representation, functional annotation, structural prediction, directed evolution, biocatalysis, or industrial enzyme applications. Studies were excluded from the analytical core when they lacked a clear relationship to enzyme discovery or provided only promotional claims without sufficient methodological information.

4.2 Analytical Framework

Evidence was coded across seven discovery stages: environmental or database sourcing, sequence processing, functional prediction, structure and property prediction, candidate prioritization, experimental validation, and process translation. Application evidence was classified under biotechnology, food systems, and green manufacturing.

The conceptual candidate score was defined as:

$$P_i = w_a A_i + w_s S_i + w_e E_i + w_n N_i + w_p R_i - w_u U_i$$

where:

- P_i represents the overall priority of candidate i;
- A_i represents predicted catalytic activity;
- S_i represents predicted operational stability;

- Ei represents expected expression feasibility;
- Ni represents evidence of functional novelty;
- Ri represents application relevance;
- Ui represents predictive uncertainty; and
- w represents the application-specific weight assigned to each criterion.

This expression is conceptual. It does not imply that all variables can be measured perfectly or combined without validation. Weighting should be established through expert input, experimental objectives, process constraints, and sensitivity analysis.

4.3 Simulation Procedure and Data Integrity

A synthetic dataset of 20 hypothetical enzyme candidates was created to demonstrate two-variable screening. Candidates were assigned predicted thermal-stability and catalytic-activity scores on a standardized 0–100 scale. Seven candidates were associated with general biotechnology, six with food systems, and seven with green manufacturing.

The simulated scores were not produced by a trained biological model and do not correspond to real enzymes. No statistical significance, biological activity, industrial performance, or causal relationship should be inferred from them. The purpose is to demonstrate how a research team could visualize candidate portfolios before experimental screening.

Table 2: Supporting values used in the simulated enzyme-candidate graph

Application group	Number of candidates	Stability-score range	Activity-score range	Mean activity score
Biotechnology	7	42–54	38–56	47.0
Food systems	6	56–64	58–70	64.2
Green manufacturing	7	65–78	72–91	80.9
Complete portfolio	20	42–78	38–91	64.4

The visible association between the two scores is a property of the constructed scenario. An empirical study would require independent test data, uncertainty intervals, error analysis, biological replicates, benchmark comparisons, and prospective laboratory validation.

5. Analysis

An AI-enabled discovery program begins with a clearly defined target-product profile. The research team must specify the reaction, substrate, desired product, process temperature, pH, solvent, cofactors, selectivity, required turnover, expression system, safety conditions, and estimated production cost.

Without these constraints, a model may rank sequences that are scientifically interesting but unsuitable for the intended application.

Environmental selection should follow the target profile. If the required enzyme must operate at high temperature, thermophilic ecosystems offer a biologically plausible search space. Cold-active enzymes may be investigated in polar, alpine, and deep-marine environments. Saline or alkaline habitats may yield proteins adapted to chemically challenging conditions. Agricultural residues, compost, and herbivore-associated microbiomes are relevant to carbohydrate-degrading enzymes.

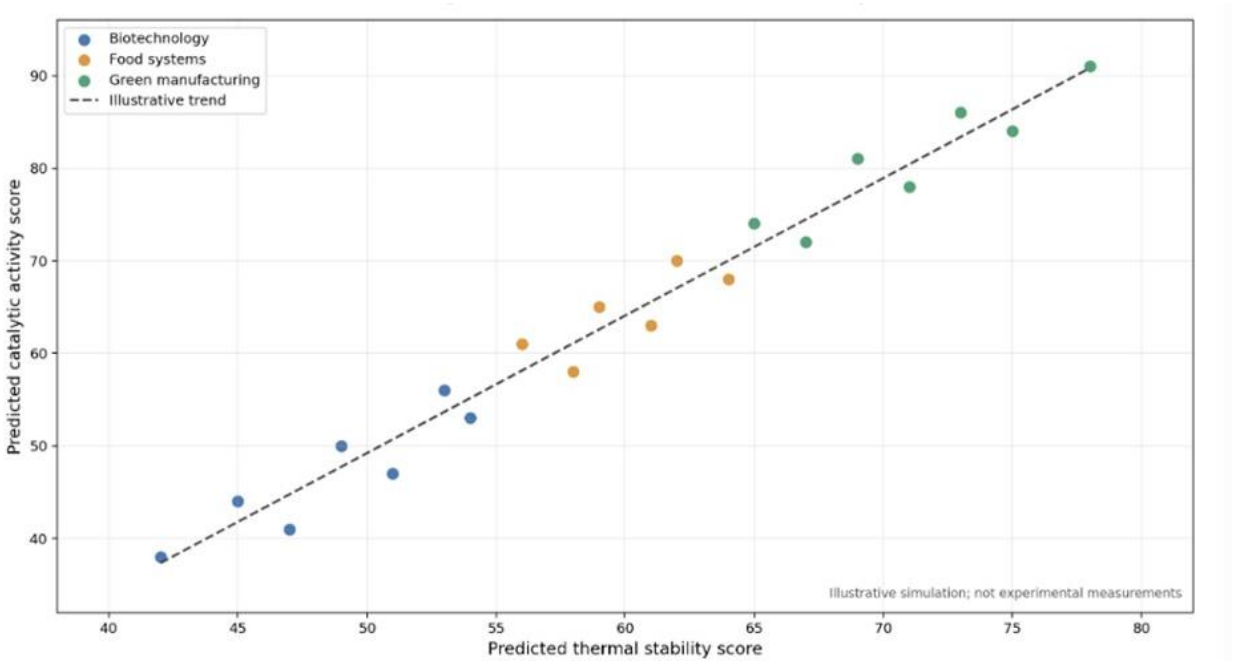
Sequence-processing quality affects every downstream stage. Raw reads must be checked, assembled where appropriate, and translated into candidate open reading frames. Redundant sequences can be clustered, while fragments and low-quality predictions require careful treatment. Taxonomic and ecological metadata should remain connected to sequences because environmental origin may help explain catalytic properties.

Initial annotation can combine sequence alignment, domain detection, catalytic-motif recognition, evolutionary analysis, and machine-learning prediction. Candidates strongly matching known enzymes may support reliable functional inference, whereas distant sequences require more cautious interpretation. Novelty should not be defined merely as low sequence identity. It may refer to a new sequence, fold, substrate, reaction, mechanism, operational property, or industrial application. Each form of novelty requires different evidence.

Protein language models can generate numerical representations of sequences that preserve patterns learned from large protein databases. These representations can support classifiers for enzyme families, functional classes, localization, solubility, or stability. However, a model may learn biases associated with heavily represented organisms and enzyme families. Performance should therefore be evaluated separately for common, rare, and evolutionarily distant proteins.

Structural prediction can help identify active-site geometry, substrate-access channels, oligomerization interfaces, and stabilizing interactions. Molecular docking and simulation may add hypotheses regarding binding, but predicted affinity is not equivalent to catalytic turnover. Catalysis involves transition-state stabilization, conformational dynamics, solvent organization, proton transfer, and cofactor behavior that may not be represented adequately by a static model.

Graph 1: Simulated Screening Profile of AI-Prioritized Enzyme Candidates



The graph shows a positive modeled relationship between predicted thermal stability and predicted catalytic activity. Candidates associated with green manufacturing occupy the upper-right region because the simulation assumes that industrial chemical applications prioritize both high-temperature resilience and strong catalytic performance. Food-system candidates occupy the middle range, while the biotechnology candidates illustrate a broader early-stage portfolio.

The upper-right region would be a logical priority area if stability and activity were the only decision criteria. Real candidate selection would also consider selectivity, substrate concentration, toxicity, solubility, expression yield, cofactors, immobilization, food safety, and manufacturing cost. A candidate with moderate activity but excellent specificity may be more useful than one with high apparent activity and undesirable side reactions.

Key finding: The scatter plot demonstrates how AI predictions can organize an experimental portfolio, but it cannot establish actual activity or superiority. Every candidate remains hypothetical until its protein is expressed, purified, characterized, and tested under application-relevant conditions.

Experimental validation should proceed in defined stages. The sequence is first synthesized or amplified and introduced into an appropriate expression host. Successful expression does not prove that the protein is soluble or correctly folded. Purification and activity assays must then determine whether the predicted reaction occurs. Appropriate positive controls, negative controls, blanks, calibration procedures, and biological or technical replicates are required.

6. Challenges

Training data are a central limitation. Protein databases contain large numbers of sequences, but only a smaller subset has been characterized experimentally. Many annotations are transferred computationally from homologous proteins. If an incorrect annotation is propagated across databases, a machine-learning system may learn and amplify the error.Data imbalance further complicates prediction. Common enzyme families may contain thousands of examples, while rare functions have only a few. Aggregate evaluation

can obscure poor performance in underrepresented classes. Models should therefore report per-class performance, uncertainty, calibration, and behavior on sequences with low similarity to training data.

Novel enzymes create an inherent out-of-distribution problem. A model trained on known protein families is most reliable when predicting proteins similar to its training examples. Yet discovery is often concerned with unusual sequences and functions. Confidence scores generated by a model may be misleading if the candidate lies outside the learned sequence distribution.Enzyme function cannot always be represented by a single label. Some proteins are multifunctional, display catalytic promiscuity, require interaction partners, or change behavior under different conditions. Assigning one enzyme commission number may simplify a complex biochemical reality. Models should allow hierarchical and multilabel predictions where justified.

The integration of substrate information remains difficult. Two enzymes with similar structures may act on different substrates, while the same enzyme can transform several molecules. Reliable prediction requires representations of proteins, substrates, reactions, cofactors, and environmental conditions. Incomplete reaction databases restrict this integration.Model interpretability is scientifically important. A prediction should ideally identify the residues, motifs, structural regions, or comparative evidence contributing to its conclusion. Interpretability cannot prove that the model's reasoning is biologically correct, but it can help experts design mutations, detect artifacts, and select validation experiments.

Experimental bottlenecks remain significant. Gene synthesis, cloning, expression, purification, and biochemical assays are resource-intensive. Some proteins require specialized hosts, post-translational modifications, membrane environments, or cofactors. A highly ranked candidate may repeatedly fail because the experimental system cannot reproduce its native biological context.Scale-up creates additional uncertainty. Enzyme behavior in a small assay may not predict performance in a reactor. Mass transfer, mixing, substrate solubility, temperature gradients, contamination risk, and enzyme recovery influence productivity. Immobilization can improve reusability but may reduce apparent activity or alter substrate access.

7. Discussion

7.1 AI as an Experimental Prioritization System

The principal value of AI lies in increasing the informational quality of candidate selection. A laboratory may be unable to test hundreds of thousands of environmental sequences, but it can evaluate a carefully ranked subset. Computational models can combine evidence that would be difficult to review manually, including remote sequence relationships, predicted structural features, environmental metadata, and application constraints.

This prioritization should not be mistaken for discovery completion. A model output is a hypothesis concerning function or performance. Enzyme discovery becomes credible through controlled biochemical evidence. The appropriate language is therefore "AI-prioritized candidate" or "computationally predicted function" until experimental confirmation is obtained.

Model performance should be evaluated prospectively. Randomly dividing a highly similar sequence database into training and testing sets can create optimistic results because close homologs occur in both subsets. More demanding evaluations use sequence-identity thresholds, family-level separation, temporal validation, or prospective testing of newly selected candidates.

7.2 Sector-Specific Translation

The three application areas considered in this paper share a common discovery pipeline but differ in decision criteria. General biotechnology may prioritize reaction novelty and molecular specificity. Food systems place stronger emphasis on safety, sensory neutrality, regulatory acceptance, and performance in complex edible matrices. Green manufacturing often prioritizes productivity, stability, solvent tolerance, catalyst reuse, and measurable lifecycle improvement.

These differences should influence model design. A general enzyme-function predictor may identify likely catalytic families but cannot rank candidates appropriately without application-specific features. Industrial discovery therefore requires collaboration among computational scientists, enzymologists, process engineers, food scientists, toxicologists, sustainability researchers, and regulatory experts.

Commercial success also depends on production feasibility. An enzyme with exceptional catalytic properties may remain unsuitable if it is expressed at extremely low levels or requires expensive purification. Multi-objective models should balance activity with manufacturability rather than optimizing a single biological property.

7.3 Closed-Loop Learning and Responsible Governance

Closed-loop learning connects prediction, experiment, and process feedback. Initial models rank candidates; laboratories test a selected group; and measured outcomes are returned to the computational system. Active-learning methods can select the next experiments based on expected performance and information gain. This approach uses negative and ambiguous results productively.

Reproducibility requires detailed records of sequence databases, preprocessing, model architecture, training partitions, hyperparameters, software versions, candidate-selection rules, assay conditions, and exclusion decisions. When proprietary data prevent full disclosure, researchers should provide enough methodological detail to support independent evaluation of the central claims.

Responsible governance should begin at project design rather than after a candidate becomes commercially attractive. It should address environmental-sample permissions, benefit sharing, biosafety, data access, model security, intellectual property, research integrity, and the communication of uncertainty. Human expert review remains essential when models rank biologically unusual or potentially sensitive functions.

8. Conclusion

AI-enabled enzyme discovery can connect expanding biological-sequence resources with the limited capacity of experimental laboratories. Metagenomics, protein representation learning, functional prediction, structural modeling, and active learning can identify scientifically plausible candidates from search spaces too large for manual analysis. Their principal benefit is therefore improved experimental prioritization rather than the elimination of experimental science.

The simulation used in this study demonstrates how predicted catalytic activity and thermal stability might be visualized across a hypothetical candidate portfolio. The modeled trend is not biological evidence. It illustrates a decision-support method that would require validation through expression,

controlled activity assays, kinetic characterization, benchmark comparison, and application-relevant testing.

Applications in biotechnology, food systems, and green manufacturing require different performance profiles. Food enzymes require rigorous safety and matrix evaluation; manufacturing enzymes must demonstrate process stability, productivity, and genuine environmental advantage; and broader biotechnology applications require reliable specificity, expression, and system compatibility. AI models should reflect these distinctions through multi-objective candidate ranking.

The responsible future of enzyme discovery lies in a closed-loop partnership among computation, experimentation, and process engineering. AI can direct attention toward promising and previously overlooked regions of protein diversity, while human expertise establishes biological meaning, safety, novelty, and industrial relevance. When combined with transparent methodology, lifecycle assessment, biosafety review, and equitable management of genetic resources, this approach can support more efficient biotechnological innovation and more sustainable production systems.

References

1. Mazurenko, S., Prokop, Z., & Damborsky, J. (2020). Machine learning in enzyme engineering. *ACS Catalysis, 10*(2), 1210–1223. https://doi.org/10.1021/acscatal.9b04321
2. Singh, N., Malik, S., Gupta, A., & Srivastava, K. R. (2021). Revolutionizing enzyme engineering through artificial intelligence and machine learning. *Emerging Topics in Life Sciences, 5*(1), 113–125. https://doi.org/10.1042/ETLS20200257
3. Feehan, R., Montezano, D., & Slusky, J. S. (2021). Machine learning for enzyme engineering, selection and design. *Protein Engineering, Design and Selection, 34*, gzab019. https://doi.org/10.1093/protein/gzab019
4. Jang, W. D., Kim, G. B., Kim, Y., & Lee, S. Y. (2022). Applications of artificial intelligence to enzyme and pathway design for metabolic engineering. *Current Opinion in Biotechnology, 73*, 101–107. https://doi.org/10.1016/j.copbio.2021.07.024
5. Mowbray, M., Savage, T., Wu, C., Song, Z., Cho, B. A., Del Rio-Chanona, E. A., & Zhang, D. (2021). Machine learning for biochemical engineering: A review. *Biochemical Engineering Journal, 172*, 108054. https://doi.org/10.1016/j.bej.2021.108054
6. Khan, M. A., et al. (2021). A hierarchical deep learning based approach for multi-functional enzyme classification. *Protein Science, 30*(9), 1995–2005. https://doi.org/10.1002/pro.4146
7. Ralbovsky, N. M., & Smith, J. P. (2021). Machine learning and chemical imaging to elucidate enzyme immobilization for biocatalysis. *Analytical Chemistry, 93*(35), 11973–11981. https://doi.org/10.1021/acs.analchem.1c01909
8. Robinson, S. L., Piel, J., & Schmidt, E. W. (2021). A roadmap for metagenomic enzyme discovery. *Natural Product Reports, 38*, 2162–2186. https://doi.org/10.1039/D1NP00006C
9. Ferruz, N., & Höcker, B. (2022). Controllable protein design with language models. *Nature Machine Intelligence, 4*, 521–532. https://doi.org/10.1038/s42256-022-00499-z
10. Yu, J., et al. (2022). Data-driven enzyme engineering to identify function-enhancing enzymes. *Protein Engineering, Design and Selection, 36*, gzac009. https://doi.org/10.1093/protein/gzac009
11. Yang, K. K., Wu, Z., & Arnold, F. H. (2019). Machine-learning-guided directed evolution for protein engineering. *Nature Methods, 16*, 687–694. https://doi.org/10.1038/s41592-019-0496-6

12. Wu, Z., Kan, S. J., Lewis, R. D., Wittmann, B. J., & Arnold, F. H. (2019). Machine learning-assisted directed protein evolution with combinatorial libraries. *Proceedings of the National Academy of Sciences, 116*(18), 8852–8858. https://doi.org/10.1073/pnas.1901979116
13. Zhou, J., & Huang, M. (2024). Navigating the landscape of enzyme design: From molecular simulations to machine learning. *Chemical Society Reviews, 53*, 8202–8239. https://doi.org/10.1039/D4CS00196F
14. Notin, P., Rollins, N., Gal, Y., Sander, C., & Marks, D. S. (2024). Machine learning for functional protein design. *Nature Biotechnology, 42*, 216–228. https://doi.org/10.1038/s41587-024-02127-0
15. Hogg, E. J., et al. (2024). The impact of metagenomics on biocatalysis. *Angewandte Chemie International Edition*. https://doi.org/10.1002/anie.202402316
16. Abdelraheem, E. M. M., Busch, H., Hanefeld, U., & Tonin, F. (2019). Biocatalysis explained: From pharmaceutical to bulk chemical production. *Reaction Chemistry & Engineering, 4*, 1878–1894. https://doi.org/10.1039/C9RE00301K
17. Zhang, Y., Geary, T., & Simpson, B. K. (2019). Genetically modified food enzymes: A review. *Current Opinion in Food Science, 25*, 14–18. https://doi.org/10.1016/j.cofs.2019.01.002
18. Ozturk, B., et al. (2019). State-of-the-art strategies and applied perspectives of enzyme biocatalysis in food sector—Current status and future trends. *Critical Reviews in Food Science and Nutrition, 60*(12), 2063–2081. https://doi.org/10.1080/10408398.2019.1627284
19. Liu, J., et al. (2019). Current methods and applications in computational protein design for food industry. *Critical Reviews in Food Science and Nutrition, 60*(19), 3285–3301. https://doi.org/10.1080/10408398.2019.1682513
20. Kuddus, M. (Ed.). (2019). *Enzymes in food biotechnology: Production, applications, and future prospects*. Academic Press. https://doi.org/10.1016/C2016-0-04555-2