

Applications of Artificial Intelligence in Microbial Bioremediation: Current Advances, Challenges and Future Prospects

Ikponmwen Ivie^{1*}, Ordu-Pac Obuchukwu Precious², Kikshak Rotdung³, Ogbeba Jeremiah⁴

¹ Department of Biological Sciences, Faculty of Sciences and Computing Wellspring University, Nigeria.

² Federal College of Education (Technical), Omoku, Nigeria.

³ Arrit Projects Ltd, Plateau State, Nigeria.

⁴ Federal Polytechnic Bauchi, Science Laboratory Technology (Biology & Microbiology) Department, Nigeria.

* Corresponding author. Email: ivie.ikponmwen@gmail.com (I.I.); obuchukwuordupac@gmail.com (O.-P. O.P.); kikrotz7@gmail.com (K.R.); jerrielife2012@gmail.com (O.J.)

Manuscript submitted September 5, 2026; accepted September 23, 2026; published October 20, 2026.

doi: 10.18178/JAAI.2026.4.4.200-222

Abstract: Artificial Intelligence (AI) is increasingly applied to microbial bioremediation to manage complex relationships among microorganisms, pollutants and environmental conditions. This critical narrative review examines peer-reviewed literature published from 1 January 2020 to the final search date of 21 September 2026 and evaluates AI across a pathway extending from microorganism and enzyme selection to pollutant-removal prediction, process optimization, experimental validation and real-world translation. Recent studies show that machine learning can identify microbial indicators, predict attenuation and removal of heavy metals and organic contaminants, optimize petroleum-, phenol- and PCB-degrading systems, and mine genomic or metagenomic data for candidate enzymes. However, the evidence base remains uneven because datasets are often heterogeneous, models may overfit, external validation is limited, and computational associations do not necessarily establish biological causation. The review therefore distinguishes computational prediction from biochemical, whole-cell, consortium, realistic-matrix and field validation. It argues that the most useful future pathway is an iterative prediction-validation-optimization framework in which AI prioritizes biological candidates and operating conditions, experiments confirm or reject predictions, and monitoring or field observations refine models. AI should consequently be viewed as a decision-support layer that strengthens, rather than replaces, microbiological understanding and experimental validation.

Keywords: artificial intelligence, microbial bioremediation, machine learning, deep learning, metagenomics, enzyme prediction, process optimization, explainable AI, environmental remediation

1. Introduction

1.1. Environmental Pollution and Microbial Bioremediation

Environmental pollution continues to place pressure on soils, surface waters, groundwater and engineered treatment systems because many contaminant classes are persistent, toxic or continuously reintroduced. For the present review, four groups are especially relevant to microbial bioremediation: heavy metals, petroleum hydrocarbons, pharmaceuticals and industrial organic pollutants such as dyes and phenolic compounds. Heavy metals differ from degradable organic contaminants because they cannot be mineralized; microbial treatment therefore relies on processes such as biosorption, bioaccumulation, precipitation, redox

transformation and immobilization. Recent reviews describe bacteria, fungi and other microorganisms as important biological agents for reducing metal mobility or aqueous concentration, while also emphasizing that metal toxicity, mixed contamination and site chemistry can constrain performance [1, 2]. Petroleum hydrocarbons create a different challenge. They are chemically diverse mixtures whose biodegradation depends on compound structure, bioavailability, oxygen status, nutrient supply and the metabolic capacity of the resident or introduced microbial community. Recent work therefore treats petroleum remediation as an ecological and process-control problem rather than as a single-organism reaction [3].

Pharmaceuticals add another layer of complexity because many active compounds occur at low concentrations, enter treatment systems continuously and can coexist with other chemicals and microbial stressors. A recent review of pharmaceutical waste management concluded that microbial consortia can provide a cost-effective treatment route, but also noted important limitations related to treatment time, microbial control, high pollutant concentrations and the handling of engineered organisms [4]. Industrial dyes and related organic pollutants likewise resist many conventional processes. Microbial removal may occur through enzymatic degradation, biosorption, bioaccumulation or combinations of these mechanisms; bacterial, fungal and algal systems have all been investigated. For dye-contaminated matrices, recent reviews highlight the importance of enzymes such as laccases, peroxidases and azoreductases, as well as the need to evaluate detoxification rather than colour removal alone [4, 5]. Across these contaminant groups, the common point is that microbial remediation depends on biological activity interacting with pollutant chemistry and environmental conditions.

Conventional physical and chemical treatments remain essential in environmental management, but they can be expensive, energy-intensive, chemically demanding or incomplete for complex mixtures. Microbial bioremediation is attractive because it uses the metabolic and surface properties of microorganisms to transform, immobilize or remove contaminants under comparatively mild conditions. Its sustainability advantage, however, should not be overstated. A biological process can fail when the target compound is poorly bioavailable, when pollutant concentration is toxic, when electron donors or acceptors are limiting, when indigenous competitors suppress an introduced strain, or when pH and temperature move outside the viable operating window. Reviews covering metals, petroleum hydrocarbons, pharmaceuticals and dyes consistently show that treatment outcome is controlled by multiple simultaneous variables rather than by the identity of a microorganism alone [1–4, 6].

This multivariable character is central to the present review. A microbial strain that performs well in a defined flask experiment may not remain dominant in a complex community. An enzyme predicted to attack a pollutant may be poorly expressed in situ. A microbial consortium may perform differently as nutrient availability, pollutant loading or redox conditions change. For petroleum-contaminated soil, for example, recent synthesis emphasizes that environmental parameters and remediation strategy strongly influence biodegradation [3]. For heavy metals, microbial activity can alter speciation and mobility, but the direction and magnitude of the effect depend on metal identity and local chemistry [1]. In pharmaceutical and dye treatment, co-contaminants and transformation products introduce additional uncertainty [3–5].

The implication is that microbial bioremediation is well suited to data-driven support. The problem is not a shortage of possible microorganisms or treatment variables; it is the difficulty of selecting, combining and optimizing them efficiently. This creates a natural entry point for Artificial Intelligence (AI) and Machine Learning (ML), which can examine nonlinear relationships among microbial features, pollutants and operating conditions. Recent reviews of AI in environmental remediation increasingly frame these methods as decision-support tools for microorganism selection, pollutant screening, process prediction, optimization and monitoring [7–11]. The relevance of AI is therefore not that it replaces microbiology. Its value is that it

may help organize and interpret biological complexity so that microbial remediation can be designed and tested more efficiently.

1.2. Why AI is Becoming Important

Microbial remediation is governed by interacting biological, chemical and operational variables. Microorganism identity, community composition, enzyme repertoire, pollutant concentration, pH, temperature, nutrient availability, oxygen or redox state, treatment time and matrix characteristics can all influence the observed outcome. Traditional experimentation remains indispensable, but exhaustive testing of all possible combinations is costly and often impractical. AI and ML are increasingly used to reduce this search space by learning relationships from existing experimental, monitoring, genomic and process datasets [7–11].

AI terminology is used precisely in this review. Artificial intelligence is the broad field; machine learning is a subset of AI; deep learning is a subset of machine learning; AutoML denotes an automated workflow for model selection and hyperparameter tuning; and Response Surface Methodology (RSM) is a statistical design and optimization approach rather than an AI algorithm. These distinctions are important when comparing studies that combine neural-network models with statistical optimization.

Several recent reviews describe two broad roles for ML in microbial remediation. The first is predictive mapping: a model learns a relationship between input variables and an outcome such as removal efficiency, attenuation rate, microbial growth or treatment stability. The second is importance-based identification: the model is interrogated to determine which microorganisms, environmental variables or molecular descriptors contribute most strongly to the prediction [11]. These roles are already visible in primary studies. In petroleum-contaminated soil, automated ML has been used to relate remediation efficiency to multiple soil, pollutant and treatment variables [12]. In groundwater, random forest and neural-network approaches have linked microbial-community information to natural attenuation of BTEX and chlorinated aliphatic hydrocarbons [13]. In wastewater systems, ML has been applied both to predict nutrient removal and to screen for microbial taxa associated with treatment performance [14].

AI is also becoming important because biological information is moving beyond culture-based observations. Metagenomic and genome-resolved datasets can contain millions of sequences and large numbers of potential functional genes. Such data are too large and complex for manual interpretation alone. Recent reviews of ML in metagenomics emphasize that high-dimensionality, sparsity and noise create analytical challenges, but also show how supervised and unsupervised methods can extract taxonomic and functional patterns [15]. In bioremediation specifically, XenoBug demonstrates the possibility of linking molecular descriptors of xenobiotics to bacterial enzyme information from genome and environmental metagenome datasets [16]. ToxZyme similarly shows that enzyme sequences can be classified according to detoxification potential using ensemble ML methods [17]. These examples move AI from generic process prediction toward biologically interpretable discovery.

A further attraction is that ML can support iterative decision-making. A model can be trained on existing data, used to identify promising conditions or biological candidates, experimentally tested, and then updated with new results. In principle, this creates a prediction-validation-optimization cycle. The critical issue, however, is whether that cycle is actually completed. Reviews published in 2025 repeatedly identify limited data access, interpretability, external validation and real-world translation as unresolved problems [7, 8, 10, 11]. Therefore, the presence of an accurate model should not be equated with successful remediation. The scientific value of AI depends on whether predictions remain valid outside the training set and whether microbiological experiments confirm that the proposed organism, enzyme or operating condition produces the expected remediation outcome.

The decision boundary must remain explicit. AI may rank candidates, estimate likely performance, identify

influential variables and recommend experiments or operating adjustments, but it should not independently authorize field deployment, biosafety-sensitive release, regulatory compliance decisions or claims of successful remediation. These decisions require human scientific judgment, site-specific evidence, experimental confirmation, risk assessment and regulatory oversight. Recommendations should also be withheld or escalated when inputs fall outside the model's applicability domain, uncertainty is high or monitoring data conflict with the prediction.

1.3. Aim and Review Questions

This review examines applications of AI in microbial bioremediation published from 1 January 2020 to 21 September 2026, with emphasis on the pathway from biological selection to practical remediation. It focuses on AI-assisted selection of microorganisms, enzymes and genes; prediction of pollutant removal and biodegradation; optimization of treatment conditions; model reliability and validation; and the laboratory-to-field translation gap. The central review question is: to what extent can AI move microbial bioremediation from computational prediction toward experimentally validated, reproducible and scalable remediation? The overall AI-assisted microbial bioremediation pathway considered in this review is summarized in Fig. 1.

Conceptual workflow of AI-assisted microbial bioremediation

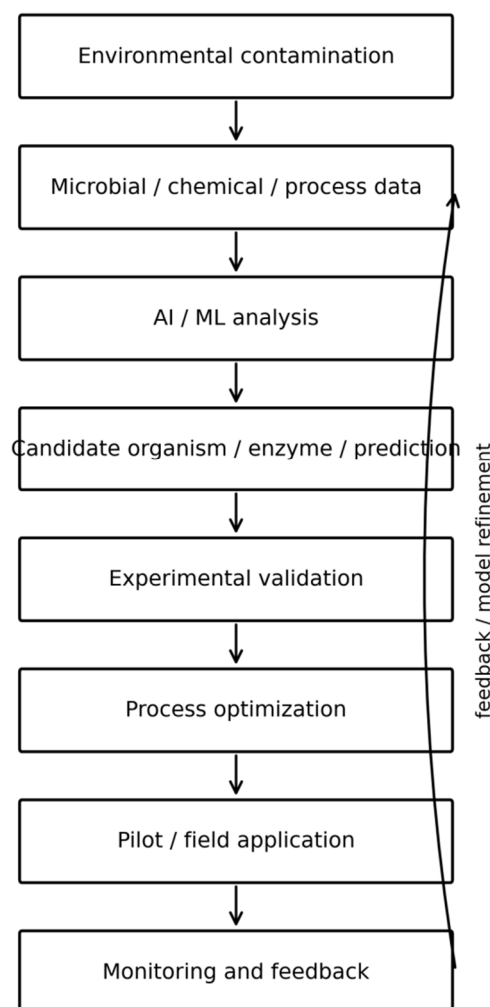


Fig. 1. High-level conceptual workflow of AI-assisted microbial bioremediation, from environmental characterization to monitoring and feedback.

1.4. Review Approach and Literature-Search Transparency

This article is a critical narrative review rather than a systematic review. Targeted literature searches were conducted in PubMed and Google Scholar, and bibliographic details were cross-checked against publisher pages and DOI records. The final search was completed on 21 September 2026. Searches covered peer-reviewed journal articles published from 1 January 2020 to 21 September 2026 and used combinations of the following terms: “artificial intelligence” AND “microbial bioremediation”; “machine learning” AND bioremediation; “machine learning” AND microorganism selection; “enzyme prediction” AND pollutant degradation; “metagenomics” AND bioremediation; “machine learning” AND pollutant removal; “machine learning” AND bioremediation optimization; “artificial neural network” AND biodegradation; “random forest” AND microbial remediation; “support vector machine” AND biodegradation; and “AI” AND bioremediation validation OR scale-up. Eligible studies were peer-reviewed articles in English that directly addressed microbial bioremediation, pollutant biodegradation or removal, microbial or enzyme selection, AI/ML prediction, optimization, validation, or translation. Articles published before 1 January 2020 or after 21 September 2026, preprints, conference abstracts, non-peer-reviewed sources and papers that lacked sufficient relevance to the review questions were excluded. Titles and abstracts were screened for relevance, followed by full-text or publisher-record verification where accessible. Thirty-two journal articles were retained for detailed synthesis. Because the review was narrative and purposive, the literature selection is not presented as exhaustive or PRISMA-based.

2. AI for Microorganism and Enzyme Selection

2.1. Microorganism Selection

Selecting an effective microorganism is one of the most important decisions in bioremediation. Conventional selection often begins with isolation from contaminated environments, enrichment under target-pollutant conditions, physiological testing and measurement of degradation or tolerance. These approaches remain scientifically necessary, but they can be slow when large numbers of strains, communities or environmental samples must be compared. AI can contribute by ranking candidates, identifying microbial indicators, relating community structure to treatment performance and predicting which taxa are most associated with pollutant removal [10, 14–17].

Community-level prediction is especially relevant because real remediation systems frequently depend on consortia rather than single isolates. In activated-sludge systems, Zhang *et al.* [14] applied six ML models to predict total phosphorus, total nitrogen and ammonium removal and combined model interpretation with microbial-community information to identify taxa associated with treatment performance. The important contribution was not merely the prediction of nutrient removal; the model also provided a route for screening potentially influential microorganisms. Jurvansuu *et al.* [18] similarly used 16S rRNA sequencing with ML to identify wastewater-treatment-plant-specific microbial indicators. Although that study focused on source attribution and monitoring rather than direct pollutant degradation, it illustrates how microbial-community signatures can be converted into predictive features.

Groundwater remediation provides a stronger connection between microbial selection and contaminant attenuation. Zhang *et al.* [13] developed a model using microbial-community information and contaminant attenuation rates for BTEX and chlorinated aliphatic hydrocarbons at a historically contaminated site. Random forest classification was combined with regression approaches to predict attenuation potential. The study demonstrates a practical concept: rather than selecting an organism purely from taxonomic presence, ML can relate community patterns to the observed environmental function. Xia *et al.* [19] extended this type of reasoning in a multi-polluted pesticide-site groundwater system, using ML on microbiome data to identify key bioindicators associated with CAH and BTEX biodegradation. The authors linked selected indicators to

co-metabolic and anaerobic degradation processes, providing biologically interpretable information that could inform remediation strategy.

These studies also show why the phrase “microorganism selection” should be interpreted broadly. AI may not always output the name of one strain to inoculate. It can identify a group of taxa, a functional guild, a community pattern or a set of microbial features that predicts successful treatment. This is important because bioremediation performance emerges from ecological interactions. Qattan’s [20] recent review of bacterial consortia emphasizes that mixed communities can combine complementary metabolic capabilities, improve resilience to environmental fluctuations and support removal of contaminants that are difficult for a single strain to manage. Machine learning is therefore potentially useful not only for choosing a microorganism but also for deciding which organisms should be considered together.

The 2026 GENIA study provides the clearest recent example of AI moving from community selection to experimental testing. Vega-Camarillo *et al.* [21] isolated 2,155 bacterial strains from xenobiotic-enriched material, prioritized strains through screening and whole-genome analysis, and then used a machine-learning-guided framework to design a synthetic microbial community. The framework incorporated genomic traits, pathway complementarity and functional redundancy. The resulting nine-member community was experimentally tested and showed simultaneous degradation of lignin, atrazine and PFOS in minimal medium, followed by soil-microcosm validation. This study is important because it closes more of the prediction-to-validation loop than many computational studies: AI-assisted community design was followed by laboratory testing and a more environmentally realistic microcosm step.

Even in this stronger example, field translation remains distinct from laboratory success. Synthetic communities can change when introduced into complex soils or wastewaters because native organisms, nutrient competition, predation, spatial heterogeneity and fluctuating pollutant loads alter ecological relationships. For this reason, AI-assisted selection should be treated as a hypothesis-generation and prioritization tool rather than a substitute for ecological validation. The most defensible workflow is sequential: generate candidate organisms or communities from data, verify their function experimentally, test stability under realistic matrices and then assess performance at pilot or field scale. The highest evidence level demonstrated by GENIA is computational community design followed by laboratory testing and soil-microcosm validation; the study did not report a field trial [21].

Primary studies also reveal that different data types support different forms of selection. Sequence-based models can prioritize functional potential, as in XenoBug and ToxZyme [13, 22]. Community-abundance data can identify taxa correlated with process outcomes [14–17]. Experimental treatment datasets can indicate which remediation strategies or microbial interventions are associated with higher removal efficiency [20, 23]. The integration of these data types is a future opportunity because taxonomic presence alone does not guarantee expression of a degradation pathway, while process performance alone does not reveal the biological mechanism.

A final caution concerns causal interpretation. Feature importance in a random forest or another predictive model does not prove that a selected taxon directly causes pollutant removal. A microorganism may be a reliable indicator because it co-occurs with the actual degrader, responds to the same environmental condition, or occupies a related ecological niche. Therefore, AI-assisted microorganism selection is strongest when model outputs are combined with genomic evidence, known metabolic pathways, targeted cultivation, isotope or metabolite evidence, and perturbation experiments. The recent literature is moving in this direction, but the evidence base remains uneven [10, 16, 17, 24]. The practical contribution of AI is thus to narrow the experimental search space and reveal candidate relationships that can be tested microbiologically.

The selection problem also exposes a distinction between indicator organisms and intervention organisms. An indicator organism is valuable because its abundance or activity predicts the state of a contaminated

system; an intervention organism is valuable because adding, stimulating or preserving it changes the remediation outcome. ML studies based on 16S rRNA profiles are often stronger at identifying indicators than proving intervention value. This is scientifically useful for monitoring, but it should be reported accurately. For example, a taxon that helps classify a wastewater plant or predict attenuation may be an excellent diagnostic marker without being the best organism for bioaugmentation [14–17]. Moving from indicator discovery to intervention design requires perturbation experiments in which the candidate is enriched, removed, introduced or otherwise manipulated and the treatment response is measured.

Data leakage is another practical concern in microorganism-selection studies. If samples from the same reactor, site or time series appear in both training and testing sets, models may learn location-specific signatures instead of transferable microbial relationships. Independent-site or temporally separated validation is therefore particularly important when the intended use is to select organisms for new environments. Future studies should report how biological replicates, time points and sites are partitioned during model development. This will make claims of generalizable microbial selection more convincing and help distinguish models that recognize a known system from models that identify transferable ecological rules.

2.2. Enzyme and Gene Prediction

Enzymes and functional genes provide a more mechanistic target for AI than taxonomic identity alone. A microorganism may carry many genes unrelated to remediation, and closely related strains can differ substantially in catabolic potential. Predicting enzyme-pollutant relationships therefore offers a way to move from “which microbe is present?” to “which biochemical reaction may occur?” Recent developments in sequence analysis, molecular descriptors and metagenomic machine learning have made this an increasingly active area [11–13, 22].

XenoBug is a computational prioritization tool developed by Malwe *et al.* [16] to predict bacterial enzymes, associated enzymatic reactions, candidate bacterial species and the metagenomic source of predicted enzymes for xenobiotic degradation. The model used molecular information from 6,814 substrates involved in approximately 141,200 biochemical reactions together with enzyme information from bacterial genomes and environmental metagenomes. Validation was computational and included held-out xenobiotics from multiple chemical classes; the study did not demonstrate biochemical enzyme activity or whole-cell pollutant degradation for every prediction. Accordingly, XenoBug should be interpreted as a candidate-prioritization platform whose outputs require experimental confirmation.

The significance of this approach is not simply computational scale. Traditional enzyme discovery often depends on known degraders, enrichment culture or homology to previously characterized proteins. ML can search a broader functional space because it can combine sequence-derived features, reaction information and molecular descriptors. This may help identify previously unreported candidates that are not obvious from simple similarity searches. At the same time, XenoBug's validation illustrates an important limitation: computational prediction of a metabolic enzyme is not equivalent to proof that the enzyme will transform a pollutant under environmental conditions. Expression level, cofactor availability, transport into the cell, competing substrates, toxicity and pathway completeness can all determine whether the predicted reaction contributes meaningfully to remediation.

ToxZyme represents a related but distinct strategy. Sahibzada *et al.* [17] developed an ensemble ML model to classify enzymes according to toxin-degradation potential. The work used positive and negative enzyme datasets and combined multiple machine-learning approaches. This type of model is useful for prioritization because it can reduce a large sequence space to a smaller set of candidates for experimental characterization. It also demonstrates the broader trend toward function prediction from protein-level information. In bioremediation, such tools could support enzyme discovery for pollutants that are difficult to study using

whole-cell systems.

Metagenomics expands the opportunity because many environmentally important microorganisms cannot be readily cultured. Sharma *et al.* [15] reviews how ML is being used to handle the high-dimensional, sparse and noisy nature of metagenomic data, including classification, time-series analysis, transfer learning and newer causal or generative approaches. For microbial bioremediation, metagenomic ML can help identify candidate genes and pathways in contaminated environments without first isolating every organism. Siddique *et al.* similarly emphasize the integration of AI with bioinformatics and metagenomic platforms for identifying microbial communities, genes and degradation pathways [22]. These approaches are particularly valuable for complex communities in soil, sediment and wastewater where functional potential is distributed across many taxa. Representative AI applications across pollutant classes, biological targets, data types and validation levels are summarized in Table 1.

Table 1. AI Applications in Microbial Bioremediation

Pollutant	Microorganism / enzyme	AI method	Data used	Main application	Validation evidence	Ref.
Heavy metals (Cd, Cu, Pb, Zn)	Sulfate-reducing bacteria	CatBoost and other ML models	Treatment variables	Predict removal and influential conditions	Experimental microbial-removal dataset	[23]
Cadmium	Wild/recombinant E. coli	AdaBoost, ExtraTrees; deep-learning image analysis	Removal experiments and microscopy	Predict bioremoval	Experimental validation sets	[25]
Petroleum hydrocarbons	Mixed remediation strategies	H ₂ O AutoML	545 compiled data points	Predict optimal conditions	Compiled-data validation	[12]
Phenol	Bacterial consortium	BP-ANN	Experimental consortium conditions	Predict biodegradation	Measured degradation + metabolic/community analysis	[26]
PCBs	Bacillus sp. + Stutzerimonas sp.	ANN + RSM (hybrid ML/statistical optimization)	Experimental conditions	PCB Optimize congeners	six Kinetic metabolomic validation and	[27]
Diesel	Soil microbial strategies	ANN	NA, bioaugmentation, biostimulation experiments	Predict remediation time/performance	Experimental comparison	[28]
Petroleum hydrocarbons	Microbial degradation system	ANN + RSM	Experimental variables	operating Optimize degradation rate	Experimental comparison	[29]
BTEX and CAHs	Indigenous groundwater communities	RF + ANN	Long-term microbial/contaminant data	field Predict natural attenuation	Field monitoring / cross-validation	[13]
CAHs and BTEX	Groundwater microbiome	ML bioindicator models	Microbiome contaminant data	and Identify bioindicators	Field-derived data	[19]
Multiple xenobiotics	Predicted bacterial enzymes	XenoBug	Molecular descriptors + genomes/metagenomes	Predict enzymes/species	Computational validation	[16]
Lignin, atrazine, PFOS	Nine-member synthetic bacterial community	Genome-informed ML	2,155 isolates genome/pathway features	+ Design synthetic consortium	Lab degradation + soil microcosm	[21]

However, gene prediction and pathway reconstruction introduce several layers of uncertainty. A predicted gene may be incorrectly annotated; a correctly annotated gene may be inactive; an active enzyme may produce a transformation product rather than complete mineralization; and pathway steps may be split across different community members. Therefore, enzyme and gene prediction should ideally be followed by transcriptomic or proteomic confirmation, enzyme assays, metabolite analysis and, where relevant, community-level experiments. The 2026 GENIA study provides an example of integrating genome-informed prediction with experimental community construction and kinetic testing [21], although its central objective was community design rather than individual enzyme validation.

AI can also help bridge chemical diversity. Jiang *et al.* [30] developed ML-based quantitative structure-biodegradability relationship models for predicting primary and ultimate biodegradation rates across organic chemicals. Although this approach is not itself an enzyme-discovery platform, it complements enzyme prediction by indicating which molecular features are associated with faster or slower biodegradation. Such information can prioritize pollutants for mechanistic study and help identify cases where a compound may require specialized enzymatic activity.

A key future direction is to connect these layers directly: pollutant structure, predicted enzyme, host microorganism, metagenomic context, pathway completeness, expression evidence and experimental degradation. Current tools often address only part of this chain. XenoBug links pollutants to predicted bacterial enzymes and metagenomic sources [16]; ToxZyme classifies detoxification enzymes [17]; metagenomic ML provides broader community and functional analysis [15]. The next step is not merely a larger model, but a biologically integrated workflow in which predictions are traceable to testable mechanisms.

Table 2. Major challenges and priority responses

Challenge	Why it matters	Priority response
Small datasets	Complex models may fit noise instead of transferable biological relationships.	Multi-site datasets, simpler models where appropriate, prospective validation.
Poor data quality / heterogeneity	Different analytical methods, endpoints and matrices reduce comparability.	Harmonized metadata, standardized endpoints and reusable raw datasets.
Overfitting / data leakage	Internal performance may be inflated when related samples occur in training and testing sets.	Biologically independent splits and external validation.
Weak interpretability	Feature importance may be mistaken for biological causation.	Explainable AI combined with mechanistic and perturbation evidence.
Limited experimental validation	Computational predictions do not prove enzyme activity or pollutant degradation.	Biochemical, whole-cell, consortium and prospective validation.
Scale-up / domain shift	Laboratory optima may fail in different soils, waters, climates or communities.	Staged microcosm-to-pilot-to-field validation and applicability-domain checks.
Cost / biosafety / regulation	A computationally optimal treatment may be impractical or unsafe.	Multi-objective optimization incorporating cost, risk, monitoring and regulatory constraints.

For review purposes, this distinction matters because papers may report impressive classification performance without demonstrating pollutant removal. Enzyme and gene prediction are most valuable for microbial bioremediation when they generate experimentally testable candidates. The evidence should therefore be described in tiers: computational identification, biochemical validation, whole-cell validation,

community validation and realistic environmental testing. The recent literature shows strong progress in the first two tiers and emerging progress in the later ones, but routine field-scale validation remains uncommon [10–13, 22, 24]. The major methodological challenges and priority responses identified across the reviewed literature are summarized in Table 2, while Table 3 provides an evidence matrix focused on AI-assisted microorganism and enzyme selection.

Table 3. Evidence Matrix for AI-Assisted Microorganism and Enzyme Selection

Pollutant / target	Candidate organism / enzyme / gene	AI model	Input data	Predicted output	Experimental validation	Environmental matrix	Key limitation	Ref.
WWTP nutrient removal	Treatment-associated microbial taxa	Six ML models	Activated-sludge community + process data	Removal performance and crucial microorganisms	Retrospective model validation	Wastewater treatment plants	Association does not establish intervention causality	[14]
BTEX / CAHs	Indigenous groundwater community / key genera	RF + ANN	Long-term microbial-community + contaminant data	Natural attenuation feasibility/rate	Cross-validation on field monitoring data	Contaminated groundwater	Observational; no manipulation of candidate taxa	[13]
CAHs / BTEX	Microbial bioindicators	ML bioindicator models	Microbiome + contaminant data	Key indicators associated with biodegradation	Field-derived observational validation	Pesticide-site groundwater	Indicator status does not prove causal degradation	[19]
Multiple xenobiotics	Predicted bacterial enzymes / species	XenoBug	Molecular descriptors + bacterial genomes + environmental metagenomes	Enzymes, reactions, species and metagenomic source	Computational held-out xenobiotic validation	Genome/metagenome data	No universal biochemical or whole-cell validation	[16]
Environmental toxins	Detoxification enzymes	ToxZyme ensemble ML	Positive and negative enzyme sequence datasets	Classify toxin-degrading enzymes	Computational classification validation	Sequence database	Predicted function requires biochemical confirmation	[17]
Lignin, atrazine, PFAS/PFOS	Nine-member synthetic microbial community	Genome-informed GNN / network modeling	2,155 isolates; 45 sequenced strains; pathway features	Optimal community assembly	Laboratory degradation + soil microcosm	Minimal medium and soil microcosm	No field trial reported	[21]

3. AI for Predicting and Optimizing Bioremediation

3.1. Prediction of Pollutant Removal

Prediction of pollutant removal is among the more mature AI applications in the bioremediation literature because treatment datasets naturally contain input variables and measurable outcomes. These outcomes include percentage removal, residual concentration, degradation rate, attenuation class and time required to reach a target. Machine-learning models can capture nonlinear relationships that are difficult to represent with simple empirical equations, especially when microbial and environmental variables interact [10, 27, 29].

Predicted removal is a model output, not proof of successful remediation. The endpoint must be defined

precisely: removal can reflect biodegradation, mineralization, natural attenuation, adsorption, precipitation, immobilization or partitioning, depending on the pollutant and system. For heavy metals, which cannot be biodegraded, concentration loss generally reflects immobilization, adsorption, precipitation or redox transformation rather than mineralization.

Heavy-metal removal by sulfate-reducing bacteria is a clear example. Xiong *et al.* [23] compared four ML models for predicting removal of Cd, Cu, Pb and Zn by sulfate-reducing bacteria. CatBoost showed the strongest overall predictive performance, and feature analysis identified temperature, pH, sulfate concentration and the carbon-to-sulfur ratio as important determinants. The study is valuable because the prediction target is directly tied to a microbial treatment process. It also illustrates how model interpretation can inform process understanding by highlighting variables that deserve experimental attention.

Solmaz *et al.* [25] examined cadmium bioremoval using wild and recombinant *Escherichia coli* strains and compared a large set of ML models. AdaBoost and ExtraTrees performed strongly for the two strain types, while image processing with a deep-learning architecture was also explored to estimate cadmium-related bacterial responses. This work shows how prediction can extend beyond tabular operating data to image-derived features. Importantly, the authors still relied on experimental bioremediation data for model development and validation, reinforcing that AI performance depends on the quality and representativeness of the biological dataset.

Organic pollutants provide similarly diverse examples. Bing *et al.* isolated a phenol-degrading consortium capable of treating high phenol concentrations and used a back-propagation artificial neural network to predict biodegradation under different conditions [26]. In this study, the model was embedded within a microbiological investigation that also examined community composition and metabolism. Thathola and Haldar [27] integrated ANN-based optimization with microbial consortium treatment of six PCBs and complemented model-based optimization with degradation kinetics and metabolomic analysis. These studies are stronger than purely computational exercises because pollutant removal was experimentally measured and the model was used to improve or interpret an observed biological process.

Petroleum contamination has produced several prediction studies. Maluleka *et al.* [28] compared bioremediation strategies for diesel-contaminated soil and developed both kinetic and machine-learning models, reporting that a bi-layered ANN outperformed linear regression for prediction. Wang *et al.* used automated ML on a compiled petroleum-hydrocarbon remediation dataset and evaluated multiple variables affecting degradation efficiency [12]. Wang *et al.* [29] later combined response-surface methodology with ANN modeling to optimize petroleum hydrocarbon degradation conditions. Together these studies show two common data pathways: models built from newly generated experiments and models built from aggregated literature or historical datasets. Each has advantages. Experimental datasets provide stronger control over measurement consistency, while aggregated datasets increase coverage but may introduce heterogeneity.

Groundwater attenuation studies illustrate prediction at a more realistic environmental scale. Zhang *et al.* [13] used long-term monitoring data for microbial communities and BTEX/CAH contaminants at a historic pesticide-manufacturing site to predict natural attenuation. Xia *et al.* [19] similarly used ML to identify microbial indicators and interpret biodegradation processes in multi-polluted groundwater. These studies are especially important for translation because they use field-derived environmental data rather than only controlled reactors or flasks. However, field-derived data also contain confounding factors and spatial heterogeneity, so model interpretation requires caution.

Wastewater studies broaden the concept from a single pollutant to treatment-system performance. Zhang *et al.* [14] predicted nitrogen and phosphorus removal from WWTP data and linked prediction to microbial-community screening. Reviews of ANN applications in wastewater treatment likewise show extensive use of neural networks to estimate pollutant removal and process outputs across diverse treatment

systems [24]. Ganthavee and Trzcinski describe AI and ML applications in pharmaceutical wastewater, including contaminant monitoring, prediction of process parameters and optimization [10]. These bodies of work support the view that AI is particularly useful when the response of interest is continuously measurable.

Prediction quality should nevertheless be judged in context. A high coefficient of determination on a test split does not establish that the model will work at another site, with another microbial community or under a new pollutant range. Internal cross-validation primarily measures performance within the available data distribution. For practical bioremediation, stronger evidence comes from temporally separated testing, external-site validation, prospective prediction followed by experiment, or transfer to pilot and field conditions. The recent literature contains examples of these stronger designs, but they remain less common than retrospective modeling [10, 16–19, 24].

A critical distinction must be maintained between removal and degradation. A model may accurately predict a decrease in aqueous concentration without identifying whether the mechanism is biodegradation, biosorption, precipitation, dilution or partitioning. For metals, “removal” generally refers to immobilization, adsorption, precipitation or transformation rather than biodegradation. For organic pollutants, concentration loss should be interpreted alongside mass balance, transformation products, mineralization indicators or metabolomic evidence where degradation is claimed. The phenol and PCB studies are stronger in this respect because model outputs were interpreted together with measured biodegradation and metabolic or metabolomic evidence [21, 28].

The strongest use of prediction is therefore not to replace experimentation but to make it more targeted. A model can indicate which parameter combinations, microorganisms or pollutant classes are likely to produce high removal, allowing researchers to allocate experimental effort more efficiently. The scientific endpoint remains an experimentally demonstrated remediation outcome.

3.2. Optimization of Remediation Conditions

Optimization extends prediction by asking which controllable conditions are most likely to maximize treatment performance. In microbial bioremediation, the variables most frequently considered include pH, temperature, nutrient supply, microbial biomass or inoculum, pollutant concentration and treatment time. These factors are biologically meaningful: pH and temperature affect enzyme activity and growth; nutrients influence biomass and co-metabolism; pollutant concentration may provide substrate at low levels but become toxic at high levels; and treatment time determines whether slow pathways can proceed to completion [1–4, 6].

The petroleum AutoML study by Wang *et al.* illustrates optimization using a broad compiled dataset [12]. Eight variables describing soil properties, petroleum characteristics and remediation conditions were used to predict remediation efficiency. The model indicated that petroleum concentration and cultivation period were particularly influential, and the analysis suggested a concentration range and treatment-duration window associated with improved degradation. The value of this approach is that it can detect patterns across many previous experiments that would be difficult to reproduce in one laboratory.

By contrast, Thathola and Haldar [27] used ANN and response-surface methodology within an experimental PCB biodegradation study. Here, optimization was directly connected to a defined bacterial consortium and specific congeners. The combination of ML/statistical optimization, kinetic analysis and metabolomics provides stronger biological grounding because predicted optimal conditions can be checked against measured degradation and transformation pathways. Bing *et al.* [26] similarly used a BP-ANN to predict phenol biodegradation by a bacterial consortium across changing conditions. These examples support a workflow in which experimental design generates training data and the model then helps locate promising regions of the parameter space.

Heavy-metal systems show the same principle with different mechanisms. Xiong *et al.* [23] found that pH,

temperature, sulfate concentration and carbon-to-sulfur ratio strongly influenced ML predictions of Cd, Cu, Pb and Zn removal by sulfate-reducing bacteria. Because the biological process involves sulfate reduction and metal precipitation, these variables have clear mechanistic relevance. Solmaz *et al.* identified initial cadmium concentration as a particularly influential factor in bacterial removal and demonstrated that ML could distinguish performance across wild and recombinant *E. coli* strains [25]. These studies show why optimization should be tied to microbial physiology rather than treated as a purely mathematical exercise.

Petroleum degradation studies also illustrate the use of neural networks as optimization companions. Maluleka *et al.* [28] compared natural attenuation, bioaugmentation, biostimulation and combined treatment in diesel-contaminated soil and used ANN modeling to predict remediation behavior. Wang *et al.* [29] combined response-surface methodology and ANN modeling to estimate optimal petroleum degradation conditions. Taken together with the AutoML study [12], these papers show a progression from strategy-level selection to parameter-level optimization.

AI-based optimization can reduce the number of laboratory experiments required when the model is adequately validated, but it cannot safely extrapolate far outside the training domain. A model trained on moderate pollutant concentrations may recommend unrealistic conditions when applied to extreme contamination. Likewise, an optimum derived from one microbial consortium may not apply to another strain or field site. Biological systems can also change over time as communities adapt, inhibitors accumulate or substrate availability declines. Static optimization should therefore be complemented by iterative monitoring and model updating where possible.

A further issue is multi-objective optimization. Maximizing removal percentage alone may favor long treatment times, high nutrient inputs or expensive operating conditions. Practical remediation may instead require balancing removal efficiency, treatment duration, biomass stability, energy demand, chemical addition and risk of toxic by-products. The current microbial-remediation literature is beginning to discuss AI for broader process optimization and real-time decision support, but many primary studies still focus on a single response variable [8–11, 27, 29]. Future work should define optimization objectives that reflect real treatment constraints.

For field translation, the most useful AI system would operate as an adaptive loop. Sensors and microbiological measurements would feed a model; the model would recommend changes to pH, nutrient dosing, aeration, retention time or inoculation; treatment performance would then provide new data for model updating. Current studies provide the predictive components of such a system, while reviews of AI in wastewater and microbial remediation identify real-time monitoring as an important future direction [7–11, 27]. The remaining challenge is to connect prediction, control and biological validation in stable operational platforms.

3.3. Common AI Models

Although the wider AI literature contains many algorithms, three model families are particularly useful for framing microbial bioremediation: Artificial Neural Networks (ANNs), Random Forests (RFs) and Support Vector Machines (SVMs). Their importance is not that they are universally superior, but that they represent different ways of learning nonlinear biological relationships and they recur across recent environmental applications [10, 11, 27, 29].

ANNs are flexible function approximators that can learn complex interactions among multiple variables. They have been used to predict phenol degradation by microbial consortia [26], diesel bioremediation [28], petroleum-hydrocarbon degradation [29] and PCB biodegradation optimization [27]. Their strength is the ability to model nonlinear relationships without an explicit mechanistic equation. Their weakness is interpretability: unless additional tools are used, it may be difficult to determine why a network generated a particular prediction. Performance can also be unstable when datasets are small relative to model complexity.

Random forests use ensembles of decision trees and are often attractive for environmental data because they can handle nonlinear relationships, mixed variable types and interactions while also providing feature-importance measures. They have been used to predict natural attenuation from microbial-community data [13], identify microbial bioindicators in contaminated groundwater [19], and model process or community variables in wastewater systems [14]. Random forest approaches are also central to discussions of ML in microbial-remediation reviews [11]. Their feature-importance outputs are useful, but should not be interpreted automatically as causal biological mechanisms.

SVMs are well suited to moderate-sized datasets and can model nonlinear boundaries through kernel functions. Jiang et al. used SVM among the methods applied to primary and ultimate biodegradation-rate modeling of organic chemicals [30]. Reviews of AI-driven bioremediation and bioinformatics also identify SVM as a common method for environmental prediction and classification [22]. SVM performance depends strongly on feature representation, kernel choice and hyperparameter tuning, and model explanations can be less intuitive than simple regression.

The choice among ANN, RF and SVM should therefore be driven by data structure and scientific objective rather than popularity. Community-abundance datasets, molecular descriptors, image features and reactor operating data have different dimensionality and noise characteristics. Comparisons within the same dataset are more informative than claims that one algorithm is generally “best”. Recent primary studies frequently compare several models and then select the strongest performer for that particular task [18, 19, 26]. This is preferable to assuming superiority a priori. These considerations are integrated in the multi-objective optimization and validation loop shown in Fig. 2.

Multi-objective AI optimization and experimental validation loop

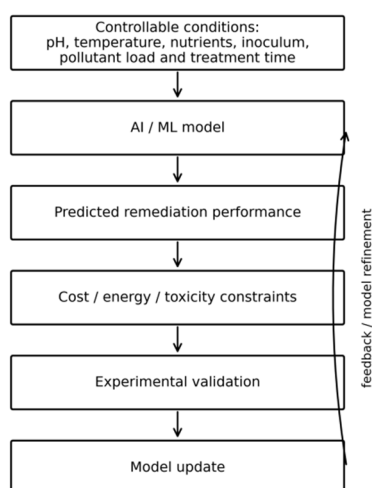


Fig. 2. Multi-objective optimization loop showing how performance, cost, energy and toxicity constraints inform experimental validation and model refinement.

4. Challenges and the Translation GAP

4.1. Data Limitations

Data limitations are a major recurring barrier in the reviewed AI-assisted bioremediation literature rather

than a universal property of every dataset. Microbiological studies can contain relatively few independent observations while including many predictors such as microbial taxa, genes, environmental variables, pollutant descriptors and operational parameters [10, 12]. This high-dimensional, low-sample setting increases the risk that models learn experiment-specific patterns rather than transferable biological relationships.

Heterogeneity creates a second problem. Literature-derived datasets may combine studies using different analytical methods, inoculum definitions, pollutant mixtures, temperatures, endpoints and reporting conventions. Wang *et al.*'s [12] petroleum AutoML study demonstrates the power of aggregation, but also illustrates why harmonization is essential when data are compiled from multiple experiments. Reviews of ML in microbial remediation call for improved data accessibility and standardization because inconsistent datasets reduce model portability [10, 29].

Metagenomic data introduce different limitations. Sharma *et al.* [15] describe high dimensionality, sparsity, compositionality and noise as core challenges for ML analysis of microbial communities. In bioremediation, these issues are compounded by functional uncertainty: a sequence may indicate potential, but not necessarily expression or activity. Preprocessing choices can therefore influence which microbial features appear important.

The retained literature more consistently documents data heterogeneity, limited metadata and high-dimensional biological features than it quantifies publication bias. Publication bias is therefore treated here as a potential concern rather than an established property of all bioremediation datasets. Regardless of cause, negative results, low-removal conditions and unsuccessful inoculations are valuable for defining the limits of a model's response surface and should be reported when available.

A practical response is to design future bioremediation studies with machine learning in mind. This means standardized metadata, clear definitions of endpoints, replicate measurements, reporting of unsuccessful conditions and deposition of reusable datasets. Where sample size remains small, model complexity should be constrained and uncertainty should be reported explicitly. Synthetic data generation may eventually help, but generated data cannot substitute for missing biological reality unless grounded in validated mechanisms [10, 12].

4.2. Model Reliability and Validation

Model reliability depends on more than fitting the available dataset. Overfitting occurs when a model learns noise, study-specific features or accidental correlations and therefore performs well during development but poorly on genuinely new data. This risk is especially high in microbial bioremediation because datasets are often small, heterogeneous and high-dimensional [10, 12]. Cross-validation reduces but does not eliminate overfitting and data-leakage risks; folds created from the same experiment can still be more similar to one another than to a new site.

The distinction between internal and external validation is therefore critical. Internal validation includes train-test splits, k-fold cross-validation or repeated resampling within one dataset. External validation uses data collected independently, ideally from another experiment, time period or site. Many recent studies report strong internal metrics. Consistent with the scope described in Section 2.2, XenoBug was evaluated computationally using held-out xenobiotics from multiple chemical classes, including compounds outside the training and test data [16]. This evaluation did not constitute biochemical, whole-cell or field validation of every predicted enzyme-pollutant relationship. ToxZyme similarly evaluates the classification of detoxification-related enzymes using separated datasets [17]. These computational validations are important, but biochemical or whole-cell testing remains necessary before a predicted enzyme can be treated as a proven remediation agent.

Experimental validation adds another level. The phenol study combined ANN prediction with measured

consortium biodegradation and metabolic analysis [26]. PCB work integrated ANN optimization with actual consortium degradation, kinetics and metabolite analysis [27]. Heavy-metal and cadmium studies likewise developed models from experimental microbial-removal data [18, 19]. These designs strengthen confidence because the model is anchored to biological measurements. However, they do not automatically establish transferability beyond the specific experimental system.

The GENIA study provides an unusually strong example of prospective validation. A genome-informed, ML-guided synthetic community was designed computationally, assembled experimentally and tested for multi-pollutant degradation, with additional soil-microcosm validation [21]. This is closer to the ideal prediction-to-validation pathway because the model influenced biological design rather than simply fitting historical outcomes. Even here, field deployment remains a further step.

Interpretability is another aspect of reliability. Feature-importance outputs and SHAP-type explanations are useful for hypothesis generation because they identify predictors or associations used by the model, but they do not by themselves establish biological causation [12, 31]. Mechanistic measurements such as metabolites, enzyme activity, redox products, mass balance or perturbation experiments should accompany biological claims. Recent environmental-XAI synthesis likewise cautions that explainability does not automatically create trustworthiness and should be paired with transparent model evaluation and uncertainty reporting [31].

Reliability also requires correct definition of the prediction target. In heavy-metal treatment, “removal” may mean biosorption, precipitation or immobilization rather than biodegradation. In organic-pollutant treatment, concentration loss may reflect adsorption rather than transformation. A model can be statistically reliable while the biological interpretation is wrong. Mechanistic measurements such as metabolites, enzyme activity, redox products or mass balance should therefore accompany AI predictions when claims concern degradation mechanisms.

Finally, uncertainty should be visible. Point predictions create a false sense of precision if the model is being applied near or beyond the boundaries of its training data. Prediction intervals, applicability-domain checks and explicit reporting of extrapolation risk would improve environmental decision-making. In high-stakes field remediation, a model should help prioritize actions, not conceal uncertainty.

Prospective validation should become a more common standard. In retrospective modeling, researchers already know the treatment outcomes before the model is built. A stronger test is to freeze the model, generate predictions for new organisms or new operating conditions, and then perform experiments without revising the model after seeing the results. This design measures whether AI can genuinely guide discovery. XenoBug illustrates computational candidate prioritization, whereas GENIA extends AI-guided prioritization to experimental community assembly and soil-microcosm testing [22, 24]. The broader literature nevertheless contains many studies in which model evaluation remains retrospective. Prospective testing would also reduce selective reporting because failed predictions would become part of the evidence base.

Reproducibility is related. For a bioremediation model to be useful beyond its original group, the training variables, preprocessing steps, hyperparameters and validation procedure should be described sufficiently for independent implementation. Open code and data are especially valuable when conclusions depend on feature importance or a claimed performance advantage among algorithms. In environmental microbiology, reproducibility must also include biological metadata: sampling depth, matrix, incubation conditions, sequencing platform, inoculum history and analytical method can influence the data used to train the model. A computationally reproducible workflow with biologically incomplete metadata is still difficult to transfer.

4.3. Laboratory-to-Field Translation

Laboratory-to-field translation is a central bottleneck for AI-assisted microbial bioremediation. Flask studies offer controlled pH, temperature, nutrient levels and pollutant concentrations. Field sites do not. Soil

structure, hydrology, pollutant aging, sorption, competing electron acceptors, native microbial communities and seasonal variation can all alter treatment performance. A model trained on laboratory data may therefore be accurate within a narrow experimental domain but unreliable in a real contaminated environment [7, 8, 10].

Evidence should progress explicitly from controlled laboratory testing to realistic matrices or microcosms, then to pilot systems and finally to field validation. Each step introduces new sources of domain shift, and independent validation is needed before claims of field transferability are justified.

Field-derived groundwater studies provide useful intermediate evidence. Zhang *et al.* [13] modeled natural attenuation using long-term monitoring of microbial communities and BTEX/CAH contaminants at a historic site, while Xia *et al.* [19] identified microbial indicators in multi-polluted groundwater. These studies avoid some of the artificiality of laboratory experiments because their data come from real contaminated systems. However, they are observational. Correlations between microbial indicators and attenuation do not automatically demonstrate that manipulating those organisms will improve remediation.

Pilot-scale and microcosm validation can bridge this gap. The GENIA study moved from high-throughput strain isolation and genome-informed ML design to experimental community testing and soil microcosms [21]. This progression is valuable because each step adds ecological complexity. Similar staged validation is needed for enzyme predictions, process-optimization models and microbial-selection tools. A proposed organism should first demonstrate the predicted function, then survive and perform in realistic matrices, then maintain activity under changing operational conditions.

Scale-up also changes process economics and monitoring requirements. A laboratory optimum may require nutrient concentrations, mixing intensity or retention times that are impractical in the field. For petroleum hydrocarbon treatment, compiled-data and experimental studies show that concentration, cultivation period and treatment strategy strongly affect predicted outcomes [20, 23, 25]. At field scale, these variables interact with cost, land availability, oxygen transfer and regulatory constraints. AI models should therefore incorporate operational feasibility rather than optimize removal percentage alone.

Real-time monitoring is frequently proposed as part of the solution. Reviews of AI in microbial remediation, pharmaceutical wastewater and AI-driven bioinformatics identify continuous sensing and adaptive control as future opportunities [7–11, 27]. The concept is attractive: sensor data could detect process drift, microbial measurements could track community changes, and an AI model could recommend operational adjustments. Yet real-time microbial information remains difficult to obtain compared with conventional variables such as pH, dissolved oxygen or conductivity. Bridging this gap will require robust biosensors, rapid molecular methods or surrogate variables that are demonstrably linked to microbial function.

Ecological safety must also be considered when AI recommends engineered organisms or synthetic communities. A computationally optimal consortium may not be ecologically appropriate for release. Persistence, horizontal gene transfer, competition with native communities and containment are relevant concerns. Reviews of microbial and genetically enhanced bioremediation repeatedly note ecological and regulatory barriers [1, 2, 30]. AI can improve candidate selection, but it does not remove the need for risk assessment.

The practical standard for “successful translation” should therefore be explicit and demanding. A field-ready AI-assisted bioremediation system should have: a clearly defined biological mechanism; validated predictions; evidence across realistic matrices; a known applicability domain; operationally feasible control variables; monitoring that can detect loss of performance; and a risk-management plan. Few current studies satisfy all these criteria. The literature instead shows a pipeline in development, with strong progress in prediction and optimization, growing experimental validation, and much thinner evidence at pilot and field scale [7, 10, 24].

This translation gap is not a failure of AI; it reflects the complexity of microbial ecology. The appropriate response is to design models and experiments together. Field constraints should shape the training data, and model uncertainty should guide where new experiments are most informative. In this sense, AI is most powerful when it becomes part of an iterative scientific process rather than a final prediction layer.

Another scale-up problem is domain shift. Training data may come from one geography, wastewater composition, soil type or climatic regime, while deployment occurs in another. The statistical distribution of the inputs then changes, and model performance can deteriorate even when the algorithm is unchanged. Microbial communities are especially sensitive to domain shift because they are shaped by local history and environmental selection. A robust field model should therefore include checks that determine whether new samples fall within the range represented in the training data. When they do not, the system should flag uncertainty and request new measurements rather than produce an apparently precise recommendation.

Economic and regulatory factors further separate laboratory feasibility from practical application. Bioaugmentation may require cultivation and transport of biomass; nutrient amendment may increase downstream oxygen demand; genetically modified organisms may face regulatory restrictions; and long treatment periods may be unacceptable at active industrial sites. AI optimization that ignores these constraints can produce scientifically interesting but operationally unusable solutions. Multi-objective models should therefore incorporate cost, treatment duration, resource use and risk alongside pollutant removal. This would align computational optimization with the decisions that remediation practitioners actually have to make.

4.4. Responsible AI, Biosafety and Data Governance

Responsible deployment requires attention to data provenance, uncertainty, reproducibility and the ecological consequences of biological interventions. Environmental AI systems can inherit bias from incomplete or spatially unrepresentative datasets, while opaque preprocessing or poorly documented model updates make recommendations difficult to audit [31]. When engineered microorganisms or synthetic communities are considered, biosafety concerns include persistence, containment, horizontal gene transfer, ecological interactions and regulatory oversight. Recent guidance on environmental release of engineered bacteria emphasizes performance control, persistence, biocontainment and regulatory review as central design considerations [32]. AI therefore does not reduce biosafety obligations; it should be embedded in a governance framework that makes data origin, model uncertainty, decision logic and biological risk explicit. The resulting hierarchy of evidence, from computational prediction through field validation, is summarized in Table 4.

Table 4. Evidence Levels for AI-Assisted Microbial Bioremediation

Evidence level	What it demonstrates	Representative retained study / status	Highest level achieved	Interpretive caution
Computational prediction	Model predicts a candidate enzyme, organism or treatment outcome	XenoBug [16]; ToxZyme [17]	Computational validation	Does not prove enzyme activity or pollutant degradation
Biochemical validation	Purified enzyme or biochemical assay confirms predicted reaction	No retained AI enzyme-prediction study provided a complete enzyme-assay validation chain	Evidence gap	Needed before enzyme prediction can be treated as demonstrated catalytic function
Whole-cell validation	A microorganism experimentally	Cadmium-bacteria study [25]	Laboratory whole-cell	May remain matrix- and condition-specific

	removes/transforms the target			
Consortium validation	A defined microbial community demonstrates treatment function	Phenol consortium [26]; PCB consortium [27]	Laboratory consortium	Community stability outside the experimental system remains uncertain
Realistic matrix / microcosm	Validated system performs in a more environmentally realistic matrix	GENIA [21]	Soil microcosm	Not equivalent to field validation
Pilot scale	AI-guided intervention tested in pilot equipment or pilot field setting	No clear prospective AI-guided pilot intervention among retained primary studies	Evidence gap	Pilot evidence should precede broad field claims
Field validation	AI-guided approach validated under field conditions	Groundwater studies [16, 17] provide field-derived observational data, not prospective AI-guided intervention	Field-derived observational evidence	Association/prediction is not the same as field intervention efficacy

5. Future Prospects

Future progress is likely to come from tighter integration of AI with biological measurement, validation and mechanistic knowledge rather than from algorithmic complexity alone. Four opportunities are particularly important: AI with metagenomics, AI-assisted enzyme discovery, AI-guided microbial consortia and AI-supported real-time process monitoring.

AI plus metagenomics can improve access to the functional potential of complex microbial communities. Metagenomic datasets contain taxonomic, gene and pathway information from organisms that may be difficult or impossible to culture. Sharma *et al.* [15] describe how ML can address classification, dimensionality reduction, time-series analysis and other tasks in metagenomic data while also emphasizing interpretability and data-quality challenges. In bioremediation, the most valuable development would be to connect metagenomic predictions directly to pollutant fate. Instead of only identifying taxa enriched at a contaminated site, future models should prioritize genes and pathways, predict community interactions and then test whether the predicted functions are expressed during treatment.

AI-assisted enzyme discovery is already moving in this direction. XenoBug links xenobiotic molecular descriptors with bacterial enzyme information from genomes and environmental metagenomes [16], while ToxZyme classifies enzymes according to detoxification potential [17]. Future platforms could combine sequence embeddings, structural prediction, reaction chemistry and environmental context to rank candidate enzymes. The critical next step is high-throughput biochemical validation. Without enzyme assays or whole-cell confirmation, improved computational accuracy will not necessarily translate into better remediation.

AI-guided microbial consortia offer perhaps the strongest route to functional robustness. Natural communities often distribute a degradation pathway across multiple organisms, and a single strain may be unable to perform every required reaction. Reviews of bacterial consortia emphasize complementary metabolism and resilience [20]. The GENIA study demonstrates that genome-informed, ML-guided community design can be linked to experimental multi-pollutant degradation and soil microcosm testing [21]. Future work should extend this concept to community stability, invasion resistance, spatial organization and dynamic adjustment of community composition under fluctuating pollutant loads.

Real-time monitoring is the fourth opportunity. Current bioremediation control often depends on periodic

sampling, which can miss rapid changes in microbial activity. AI could integrate continuous process variables with intermittent molecular or biochemical measurements to provide early warning of treatment failure. Reviews of AI in environmental management, pharmaceutical wastewater and wastewater ANN applications repeatedly identify monitoring and predictive control as important future directions [7–11, 27]. For this vision to become practical, sensor outputs must be linked to biologically meaningful states rather than only to bulk water-quality parameters.

A broader future framework should also incorporate uncertainty and mechanistic knowledge. Purely data-driven models can perform well but extrapolate poorly. Hybrid models that combine microbial growth, mass-transfer or reaction knowledge with ML may be more reliable when data are sparse. Explainable AI can help identify influential variables, while causal and knowledge-graph approaches may help distinguish correlation from mechanism [10, 12]. The objective should not be to replace mechanistic microbiology but to make models biologically constrained and experimentally informative.

Standardized, reusable datasets will be essential. The current literature contains many small studies with different endpoints and reporting formats. Community repositories should include raw process data, negative results, microbial metadata, pollutant chemistry and validation outcomes. This would allow models to be compared across sites and reduce the risk of repeatedly training on narrow local datasets.

Finally, future research should measure success at the level of translation. A study should not end when a model predicts a high-performing organism or operating condition. Stronger evidence would proceed through laboratory validation, realistic matrix testing, pilot-scale assessment and, where appropriate, field demonstration. This staged framework makes AI accountable to the biological outcome it is intended to improve.

Prediction-to-validation framework for AI-assisted microbial bioremediation

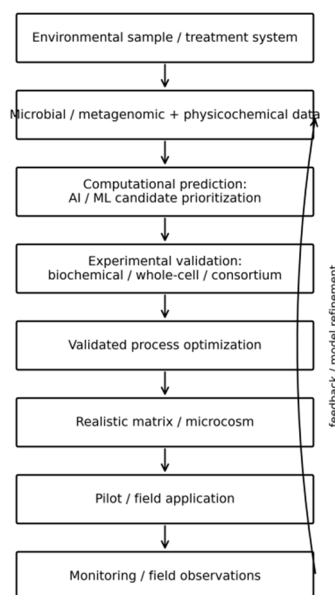


Fig. 3. Detailed prediction-to-validation framework showing the evidence ladder from computational prioritization through experimental and realistic-matrix validation to pilot or field application and feedback-driven model refinement.

5.1. Proposed AI-Bioremediation Framework

Fig. 3 consolidates the preceding evidence into an operational framework. Environmental and biological data inform AI-assisted candidate prioritization; experiments then test those candidates before process optimization and staged transfer to realistic matrices, pilot systems or field settings. Monitoring data are

returned to the model to update predictions and guide subsequent experiments [10, 16, 20, 22, 24].

Unlike Fig. 1, which provides a high-level overview, Fig. 3 emphasizes the evidence ladder from computational prioritization to experimental, realistic-matrix and field validation. Its feedback loop allows laboratory results to reject model predictions, field observations to revise laboratory optima and uncertainty to identify where new data are most valuable. AI therefore remains subordinate to experimentally observed biological function.

6. Conclusion

Artificial intelligence is becoming a useful extension of microbial bioremediation because it can help researchers interpret complex datasets, prioritize microorganisms and enzymes, predict pollutant removal, and identify operating conditions associated with improved treatment. Recent studies demonstrate meaningful progress across heavy metals, petroleum hydrocarbons, phenol, PCBs, groundwater contaminants and wastewater processes. Sequence- and metagenome-based tools further expand the search for pollutant-degrading enzymes and functional genes.

The evidence also shows a clear limitation: many studies remain strongest at prediction, retrospective modeling or controlled laboratory validation. Fewer demonstrate independent external validation, realistic-matrix performance, economic feasibility and field-scale operation. The central research need is therefore not simply higher model accuracy, but stronger integration of computational prediction with experimental microbiology, uncertainty assessment and staged scale-up. AI-assisted bioremediation will become more credible as studies adopt standardized datasets, transparent models, prospective validation, mechanistic evidence, multi-objective optimization and continuous monitoring. The future of the field lies in an iterative process in which AI identifies and prioritizes possibilities while experiments and field performance determine which possibilities are biologically real, safe, experimentally validated and practically scalable.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

I.I. conceptualized and coordinated the review, conducted the literature search and synthesis, and drafted the manuscript. O.-P. O.P., K.R., and O.J. contributed to the literature search, critical evaluation and synthesis of the reviewed literature, and critical revision and editing of the manuscript. All authors contributed to the revision of the manuscript and approved the final version.

References

- [1] Abo-Alkasem, M. I., Hassan, N. H., & Abo Elsoud, M. M. (2023). Microbial bioremediation as a tool for the removal of heavy metals. *Bulletin of the National Research Centre*, 47, 31. doi: 10.1186/s42269-023-01006-z.
- [2] Saravanan, A., Yaashikaa, P. R., Ramesh, B., Shaji, A., & Deivayanai, V. C. (2024). Microorganism-mediated bioremediation of dyes from contaminated soil: Mechanisms, recent advances, and future perspectives. *Food and Chemical Toxicology*, 185, 114491. doi: 10.1016/j.fct.2024.114491
- [3] Elshafei, A. M., & Mansour, R. (2024). Microbial bioremediation of soils contaminated with petroleum hydrocarbons. *Discover Soil*, 1, 9. doi: 10.1007/s44378-024-00004-5
- [4] Srinivasan, K., Hariharapura, R. C., & Mallikarjuna, S. V. (2025). Pharmaceutical waste management through microbial bioremediation. *Environmental Monitoring and Assessment*, 197(4), 488. doi: 10.1007/s10661-025-13912-y

- [5] Zhang, D., Zhao, Z., Zhang, J., Wang, Z., Zhou, H., Li, M., & Hao, L. (2025). Machine learning applications in prediction and optimization of heavy metal bioremediation: A review. *Journal of Environmental Chemical Engineering*, 13(5), 118845. doi: 10.1016/j.jece.2025.118845.
- [6] Swaminaathan, P., Thamarai, P., Yaashikaa, P. R., Saravanan, A., & Vickram, A. S. (2025). Microbial bioremediation of dyes, metals, and microplastics for ecological sustainability. *Energy, Ecology and Environment*, 10(1), 45–65. doi: 10.1007/s40974-024-00335-7
- [7] Upadhyay, R., Przysaś, W., & Dave, B. (2025). Myco-remediation of synthetic dyes: A comprehensive review on contaminant alleviation mechanism, kinetic study and toxicity analysis. *International Journal of Environmental Science and Technology*, 22(1), 521–538. doi: 10.1007/s13762-024-05793-4
- [8] Singh, K., Sarkar, S., & Sharma, S. (2025). Artificial intelligence in bioremediation and environmental management. *Environmental Quality Management*, 35(2), e70192. doi: 10.1002/tqem.70192
- [9] Blessing, A.-A., & Olateru, K. (2025). AI-driven optimization of bioremediation strategies for river pollution: A comprehensive review and future directions. *Frontiers in Microbiology*, 16, 1504254. doi: 10.3389/fmicb.2025.1504254
- [10] Ganthavee, V., & Trzcinski, A. P. (2024). Artificial intelligence and machine learning for the optimization of pharmaceutical wastewater treatment systems: A review. *Environmental Chemistry Letters*, 22(5), 2293–2318. doi: 10.1007/s10311-024-01748-w
- [11] Zhang, Y., Sasaki, J., Li, A., & Chen, J. (2025). Application and challenges of machine learning in microbial remediation: A review of current status and future directions. *Critical Reviews in Environmental Science and Technology*, 55(22), 1657–1682. doi: 10.1080/10643389.2025.2560435
- [12] Wang, J., Peng, C., Man, Q., Guo, R., Yang, Z., & Ma, X. (2024). Prediction of optimal bioremediation conditions for petroleum hydrocarbon contaminated soil by automated machine learning-based analysis. *Journal of Cleaner Production*, 479, 144042. doi: 10.1016/j.jclepro.2024.144042
- [13] Zhang, X., Long, T., Deng, S., Chen, Q., Chen, S., Luo, M., Yu, R., & Zhu, X. (2023). Machine learning modeling based on microbial community for prediction of natural attenuation in groundwater. *Environmental Science & Technology*, 57(50), 21212–21223. doi: 10.1021/acs.est.3c05667
- [14] Zhang, Y., Wu, H., Xu, R., Wang, Y., Chen, L., & Wei, C. (2024). Machine learning modeling for the prediction of phosphorus and nitrogen removal efficiency and screening of crucial microorganisms in wastewater treatment plants. *Science of the Total Environment*, 907, 167730. doi: 10.1016/j.scitotenv.2023.167730
- [15] Sharma, S., Narahari, H. P., & Raman, K. (2025). Harnessing machine learning for metagenomic data analysis: Trends and applications. *mSystems*, 10(11), e01642-24. doi: 10.1128/msystems.01642-24
- [16] Malwe, A. S., Longwani, U., & Sharma, V. K. (2025). XenoBug: Machine learning-based tool to predict pollutant-degrading enzymes from environmental metagenomes. *NAR Genomics and Bioinformatics*, 7(2), Article lqaf037. doi: 10.1093/nargab/lqaf037
- [17] Sahibzada, K. I., Shahid, S., Akhter, M., Faisal, M., Abd El Rahman, R. A., Imran, M., Lv, Y., Wei, D., & Hu, Y. (2025). Advancing enzyme-based detoxification prediction with ToxZyme: An ensemble machine learning approach. *Toxins*, 17(4), 171. doi: 10.3390/toxins17040171
- [18] Jurvansuu, J., Länsivaara, A., Palmroth, M., Kaarela, O., Hyöty, H., Oikarinen, S., & Lehto, K.-M. (2025). Machine learning-based identification of wastewater treatment plant-specific microbial indicators using 16S rRNA gene sequencing. *Scientific Reports*, 15(1), 23771. doi: 10.1038/s41598-025-07952-0
- [19] Xia, F., Fan, T., Wang, M., Yang, L., Ding, D., Wei, J., Zhou, Y., Jiang, D., & Deng, S. (2025). Biodegradation of CAHs and BTEX in groundwater at a multi-polluted pesticide site undergoing natural attenuation: Insights from identifying key bioindicators using machine learning methods based on microbiome data. *Ecotoxicology and Environmental Safety*, 291, 117609. doi: 10.1016/j.ecoenv.2024.117609.
- [20] Qattan, S. Y. A. (2025). Harnessing bacterial consortia for effective bioremediation: Targeted removal of

- heavy metals, hydrocarbons, and persistent pollutants. *Environmental Sciences Europe*, 37, 85. doi: 10.1186/s12302-025-01103-y
- [21] De la Vega-Camarillo, E., Arreola-Vargas, J., Antony-Babu, S., Mathur, S. K., Santos, J. A., & Shim, W. B. (2026). Machine learning-guided synthetic microbial communities enable functional and sustainable degradation of persistent environmental pollutants. *Environmental Science & Technology*, 60(18), 13500–13519. doi: 10.1021/acs.est.6c01112
- [22] Siddique, K. R., Bose, D., Bhattacharya, R., Rodriguez, R. V., & Ray, A. (2025). Artificial intelligence driven bioinformatics for sustainable bioremediation: Integrating computational intelligence with ecological restoration. *Environmental Science: Advances*, 4(12), 1987–1997. doi: 10.1039/D5VA00240K
- [23] Xiong, B., Chen, K., Ke, C., Zhao, S., Dang, Z., & Guo, C. (2024). Prediction of heavy metal removal performance of sulfate-reducing bacteria using machine learning. *Bioresource Technology*, 397, 130501. doi: 10.1016/j.biortech.2024.130501
- [24] Ibrahim, M., Haider, A., Lim, J. W., Mainali, B., Aslam, M., Kumar, M., & Shahid, M. K. (2024). Artificial neural network modeling for the prediction, estimation, and treatment of diverse wastewaters: A comprehensive review and future perspective. *Chemosphere*, 362, 142860. doi: 10.1016/j.chemosphere.2024.142860
- [25] Solmaz, M. K., Akkurt, Ş., Alkan Uçkun, A., & Uçkun, M. (2025). Integrating image processing and machine learning for efficient cadmium removal from water by bacteria. *Journal of Water Process Engineering*, 77, 108475. doi: 10.1016/j.jwpe.2025.108475
- [26] Bing, W., Li, X., Zhao, Y., Wang, Y., Zhang, J., Zhang, J., & Liang, J. (2024). Collaboration of bacterial consortia for biodegradation of high concentration phenol and potential application of machine learning. *Chemico-Biological Interactions*, 399, 111153. doi: 10.1016/j.cbi.2024.111153
- [27] Thathola, P., & Halder, S. (2025). Optimizing biodegradation of polychlorinated biphenyls (PCBs) using bacteria isolated from common effluent treatment plant (CETP) sludge: Integration of machine learning, kinetic studies, and metabolomic analysis. *International Biodeterioration & Biodegradation*, 198, 105992. doi: 10.1016/j.ibiod.2024.105992
- [28] Maluleka, K., Roopchand, R., Seedat, N., & Sillanpää, M. (2024). Developing a predictive machine learning model and a kinetic model for the bioremediation of terrestrial diesel spills. *Results in Engineering*, 23, 102378. doi: 10.1016/j.rineng.2024.102378
- [29] Wang, Z., Sun, X., Song, K., Lu, J., & Wu, M. (2025). Combined optimization of petroleum hydrocarbon degradation rate prediction model using response surface methodology and artificial neural networks. *Journal of the Indian Chemical Society*, 102(9), 101937. doi: 10.1016/j.jics.2025.101937
- [30] Jiang, S., Liang, Y., Shi, S., Wu, C., & Shi, Z. (2023). Improving predictions and understanding of primary and ultimate biodegradation rates with machine learning models. *Science of the Total Environment*, 904, 166623. doi: 10.1016/j.scitotenv.2023.166623.
- [31] Schiller, J., Stiller, S., & Ryo, M. (2025). Artificial intelligence in environmental and Earth system sciences: Explainability and trustworthiness. *Artificial Intelligence Review*, 58, 316. doi: 10.1007/s10462-025-11165-2
- [32] Y Chemla, Y., Sweeney, C. J., Wozniak, C. A., & Voigt, C. A. (2025). Design and regulation of engineered bacteria for environmental release. *Nature Microbiology*, 10(2), 281–300. doi: 10.1038/s41564-024-01918-0