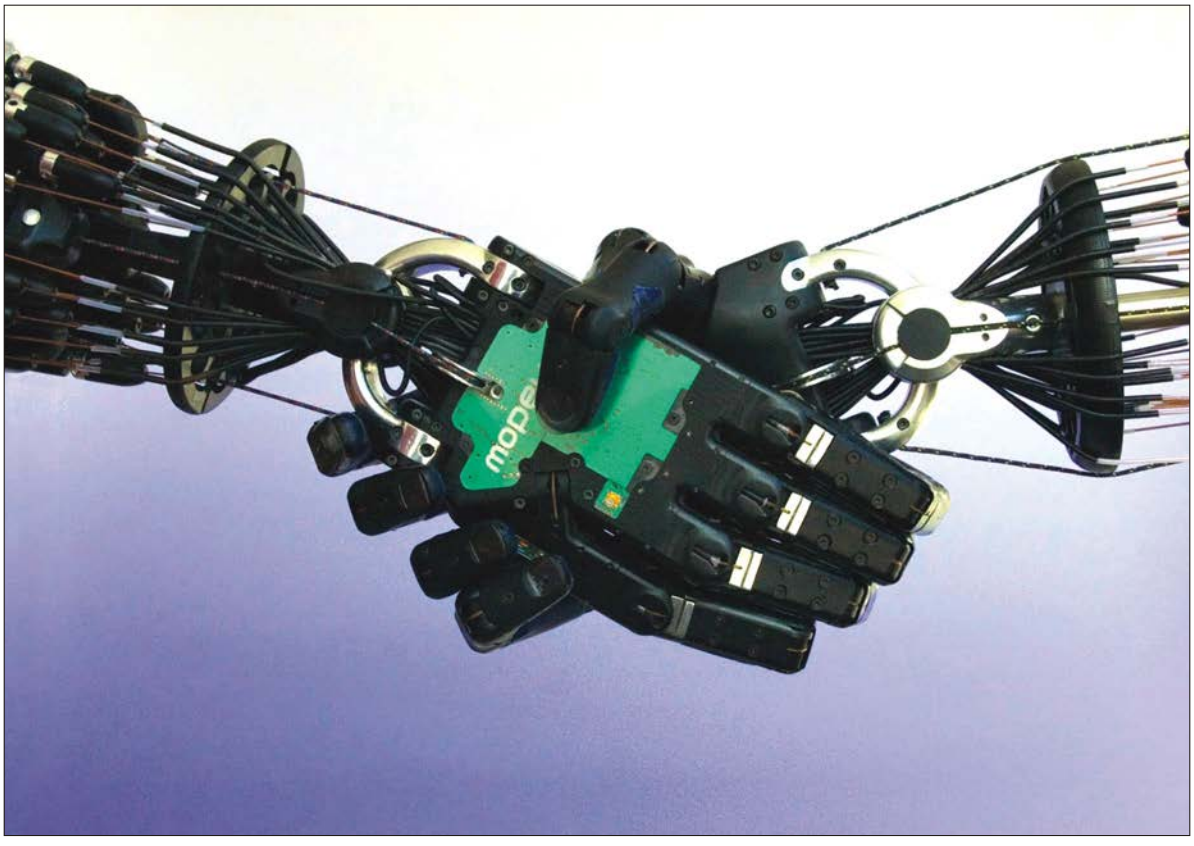

Journal of Artificial Intelligence

ISSN 1994-5450

science alert

ANSInet
an open access publisher
http://ansinet.com

OPEN ACCESS Journal of Artificial Intelligence

ISSN 1994-5450
DOI: 10.3923/jai.2026.11.23

Systematic Review
Artificial Intelligence and Predictive Analytics in Biopesticide Development: A Systematic Review

Shaifali Mathur and Rachana Tiwari

Department of Biotechnology, Faculty of Life Sciences, Shri Shankaracharya Professional University, Durg, Bhilai-490020, Chhattisgarh, India

Abstract
Biopesticides are gaining recognition as a sustainable pest management strategy since they reduce environmental toxicity and human health hazards while maintaining crop protection and agricultural output. Still, their discovery and development remain time-consuming and resource-intensive due to the biological complexity of microbial metabolism, plant secondary metabolites, peptide stability, RNA interference mechanisms and formulation behavior. By facilitating data-driven discovery, optimization and risk assessment throughout the biopesticide development process, artificial intelligence and integrated predictive analytics are emerging as potent tools to address these issues. The ability to find and rank bioactive candidates at scale has greatly increased by genome and sequence-informed techniques like biosynthetic gene cluster mining, multi-omics integration, structural prediction and deep learning-based peptide discovery. Predictive modeling for non-target risk assessment has been enhanced simultaneously by developments in computational toxicology and ecotoxicological evaluations, reinforced by carefully selected datasets, quantitative structure–activity relationship models and graph-based learning techniques. Additionally, model generalizability is limited by heterogeneous datasets and environmental variability in downstream stages like formulation design, production optimization and field translation. The present systematic review examines literature published from 2011 to 2026, employing structured search strategies and critically evaluates the impact of artificial intelligence-driven technological advancements in biopesticide development across microbial, botanical, peptide/protein and RNAi-based biopesticide modalities. The study also delineates significant methodological constraints and proposes future research trajectories to expedite the development of safe, effective and environmentally sustainable biopesticides.

Key words: Biopesticide, artificial intelligence, data-driven discovery, nanoformulation, machine learning, genome mining

Citation: Mathur, S., and R. Tiwari, 2026. Artificial Intelligence and predictive analytics in biopesticide development: A systematic review. J. Artif. Intel., 19: 11-23.

Corresponding Author: Shaifali Mathur, Department of Biotechnology, Faculty of Life Sciences, Shri Shankaracharya Professional University, Durg, Bhilai-490020, Chhattisgarh, India

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Biopesticides are becoming an important part of sustainable agriculture and integrated pest management. Unlike chemical pesticides, they control pests with less harm to the environment and non-target organisms. They are derived from natural sources such as microorganisms, plant compounds, peptides, proteins and RNA interference (RNAi) molecules. Because of their biological origin, they offer targeted pest control and support environmentally friendly farming practices. With increasing concerns about pesticide resistance, environmental pollution and loss of biodiversity, biopesticides are now seen as essential tools for climate-resilient agriculture[1]. However, developing and applying biopesticides at a large scale is challenging. Biopesticide development involves complex biological interactions between microbes, plants, pests and environmental conditions. Variability in microbial metabolism, plant chemical composition, peptide stability and RNAi delivery creates uncertainty during development[2]. Factors such as temperature, humidity, sunlight, soil microbiome and crop type can significantly affect the effectiveness of biopesticides. These complexities lead to longer development times, higher costs and less predictability compared to conventional pesticides[3]. Artificial Intelligence (AI) and predictive analytics are emerging as powerful tools to address these challenges. Machine learning, bioinformatics and computational modeling allow researchers to analyze large biological datasets and identify patterns that are difficult to detect using traditional methods[4]. AI techniques can integrate genomic, chemical and ecological data to identify promising biopesticide candidates more efficiently. Tools such as genome-mining platforms help detect biosynthetic gene clusters responsible for producing bioactive compounds[5]. Similarly, advanced models like deep learning and graph-based methods can predict molecular activity, toxicity and functional properties.

AI also supports iterative "design-test-learn" cycles, where computational predictions guide experiments and experimental results improve the models[6]. However, the use of AI is still uneven. It is highly effective in early discovery stages but less developed in downstream processes like formulation, large-scale production and field application due to data limitations and environmental variability[7,8]. Additionally, predictive models are increasingly used in environmental risk assessment, helping evaluate toxicity to non-target organisms and supporting regulatory decision-making. Toxicity assessment and quantitative structure–activity relationship (QSAR) models have therefore become important tools for predicting potential ecological risks before large-scale testing[9,10]. The availability of curated datasets for honey-bee toxicity and other ecotoxicological endpoints has enabled the development of reproducible machine-learning models aligned with regulatory expectations[11]. Despite progress, the integration of AI across the full biopesticide development lifecycle remains limited. Research is often fragmented across disciplines such as microbiology, computational biology and agricultural sciences. There is a lack of comprehensive understanding of how predictive analytics can be applied across all stages from discovery to field use and regulatory evaluation. In addition, while some stages of development (like discovery) have sufficient data, later stages such as formulation design, production optimization and field-level performance suffer from limited and inconsistent datasets[3,10]. This makes it difficult to build reliable predictive models across the entire development pipeline.

This systematic review aims to provide a clear and comprehensive overview of the role of AI and predictive analytics in biopesticide development. It examines applications across key stages, including discovery, formulation design, production optimization, field efficacy prediction and environmental risk assessment. The study also identifies current trends, evaluates the strengths and limitations of existing approaches and highlights areas that need further research. Ultimately, this review seeks to demonstrate how integrating AI with experimental and regulatory frameworks can support the development of safer, more effective and sustainable biopesticides for future agriculture.

MATERIALS AND METHODS

Study design: A systematic literature review was conducted to synthesize current research on the application of AI and predictive analytics in biopesticide development. The review followed established systematic review principles to ensure transparency, reproducibility and methodological rigor.

Literature search strategy: A structured and comprehensive literature search was performed across major academic databases, including Scopus, Web of Science, PubMed and Google Scholar. The search focused on peer-reviewed articles published between 2011 and 2026. Controlled vocabulary and Boolean operators were employed to capture relevant studies at the intersection of biopesticide research and computational modeling.

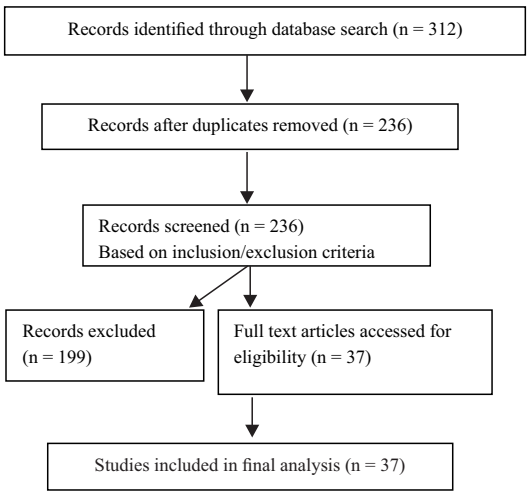

Fig. 1: PRISMA flow diagram

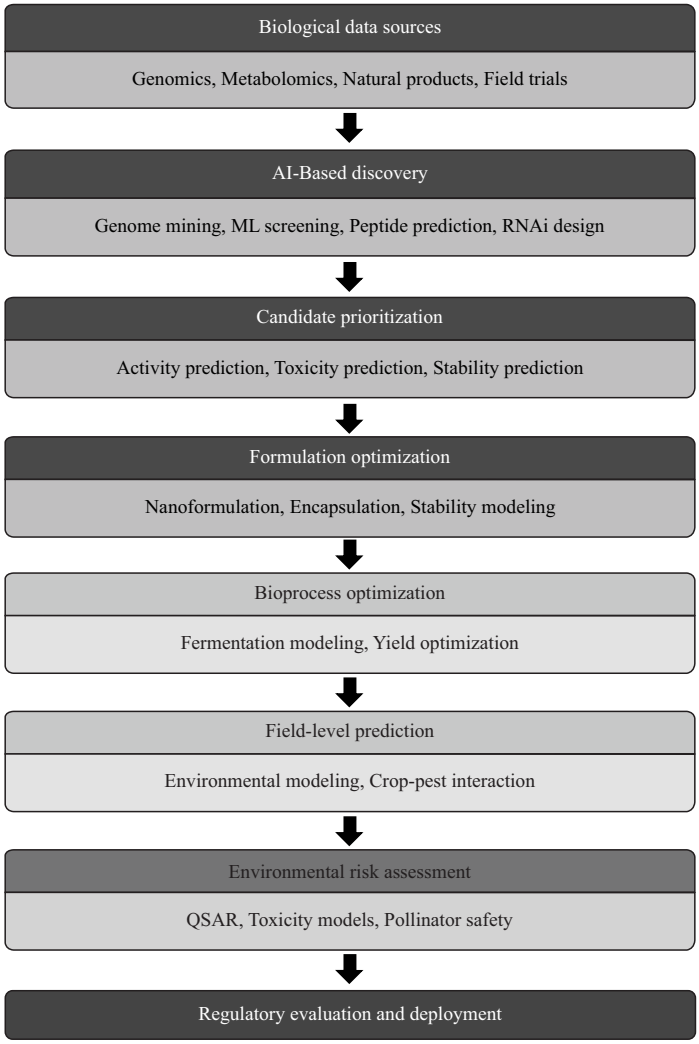

Fig. 2: AI-driven biopesticide development stages

Schematic representation of an integrated, data-driven pipeline for biopesticide discovery and deployment. The process begins with biological data and AI-based discovery to candidate prioritization, bioprocess optimization, environment risk assessment and regulatory deployment

Table 1: Applications of artificial intelligence across major biopesticide modalities.

Biopesticide modality	Data type	AI methods	Applications	References
Microbial biocontrol	Genomes, Biosynthetic Gene Clusters (BGCs)	Random Forest, DeepBGC, Neural Networks	Metabolite discovery	Medema *et al.*[6]
Peptide/Protein biopesticides	Protein sequences	Protein language models, CNNs	Peptide activity prediction	Jumper *et al.*[7] and Su *et al.*[8]
Botanical biopesticides	Chemical descriptors	QSAR, Graph Neural Networks (GNNs)	Bioactivity prediction	Rios-Martinez *et al.*[12]
RNAi-based pesticides	Gene sequences	Sequence-based ML models	Off-target prediction	Li *et al.*[13]
Formulation systems	Physicochemical and formulation data	Artificial Neural Networks (ANN), Support) Vector Machines (SVM), Genetic Algorithms (GA)	Nanoformulation optimization	Kah *et al.*[14]

Representative search query: ("Biopesticide" OR "microbial pesticide" OR "botanical pesticide" OR "RNAi pesticide" OR "dsRNA pesticide") AND ("artificial intelligence" OR "machine learning" OR "predictive model" OR "deep learning" OR "computational modeling").

Study selection process: The initial search identified 312 records. After removing duplicates, 236 unique articles remained and were subjected to title and abstract screening. A detailed eligibility assessment based on predefined criteria resulted in the inclusion of 37 studies for final analysis. Additional relevant articles were identified through manual screening of the reference lists of selected papers. To enhance transparency, the study selection process is illustrated using a PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) flow diagram Fig. 1, detailing the stages of identification, screening, eligibility and final inclusion.

Inclusion and exclusion criteria: Inclusion criteria:

- Studies investigating AI or machine-learning applications in biopesticide discovery, formulation, production, optimization, or environmental risk assessment
- Studies reporting experimental, computational, or predictive modeling outcomes

Exclusion criteria:

- Studies focused solely on agronomic pest management without computational or modeling components
- Non-English publications
- Studies lacking clear methodological or analytical contributions
- Data extraction and analysis

Relevant data were systematically extracted from the selected studies, including:

- Biopesticide modality
- AI methodology
- Dataset characteristics
- Validation strategies
- Performance metrics

The extracted data were analysed to identify methodological trends, key research gaps and emerging opportunities in AI-driven biopesticide development.

Biopesticide modalities and data types: Biopesticide development encompasses several distinct biological modalities, including microbial biocontrol agents, microbial natural products derived from biosynthetic gene clusters (BGCs), pesticidal peptides and proteins, complex botanical extracts and formulations and nucleic-acid based actives[2,3]. Because these modalities differ in their molecular structure, biological mechanisms and development pipelines, the role of artificial intelligence and predictive analytics varies across them. AI-driven biopesticide development pipeline is depicted in Fig. 2. Applications of artificial intelligence across major biopesticide modalities are summarized in Table 1.

Microbial biopesticides: In microbial and metabolite discovery, predictive analytics is commonly applied to genome mining, BGCs prediction and metabolite prioritization using machine-learning models trained on genomic and metabolomic datasets. Tools such as antiSMASH and DeepBGC have enabled systematic identification of candidate secondary metabolites from microbial genomes, accelerating natural-product discovery[6,9].

Peptide and protein-based biopesticides: For peptide and protein based biopesticides, sequence-based deep-learning models and structural prediction platforms help identify candidates with pesticidal activity and improved stability. Advances in protein structure prediction, particularly AlphaFold2, have significantly improved the ability to model pesticidal proteins and peptide toxins and guide rational screening of candidate molecules[7,8,12].

Botanical biopesticides: In botanical and small-molecule-based biopesticides, graph-based machine-learning models and QSAR approaches are often used to predict bioactivity, toxicity and physicochemical behavior of plant-derived compounds such as terpenoids and essential-oil components[9].

Predictive analytics also contributes to formulation science by modeling nano-formulation stability, delivery efficiency and environmental persistence, which are critical parameters for the development of nano-enabled pesticide delivery systems[14,15].

Nucleic-acid-based biopesticides: In nucleic-acid–based biopesticides such as dsRNA, sequence-design algorithms and off-target prediction models are used to optimize gene-silencing efficiency while minimizing risks to non-target organisms. Bioinformatic pipelines for RNAi design have been applied in pest management for species such as *Helicoverpa armigera* and *Diabrotica virgifera*, improving target specificity and ecological safety[16].

AI IN BIOPESTICIDE DISCOVERY AND CANDIDATE PRIORITIZATION

As genomic datasets continue to expand and machine-learning architectures become more sophisticated, genome-informed predictive analytics is likely to play a central role in unlocking the vast microbial and peptide diversity that remains largely untapped for sustainable pest management. Genome-informed machine learning has emerged as a powerful strategy for accelerating the discovery of microbial biocontrol agents by identifying links between genomic features, particularly BGCs and antifungal or antimicrobial phenotypes[17]. Microbial genomes encode a large diversity of specialized metabolites with pesticidal potential, but traditional discovery approaches based on culture screening and chemical isolation are slow and resource-intensive. Genome mining has therefore become a foundational approach for exploring microbial metabolic potential. Early computational platforms such as antiSMASH enabled systematic detection and annotation of BGCs in bacterial and fungal genomes, revealing that many microorganisms possess far greater biosynthetic capacity than previously recognized[6]. Global surveys of microbial genomes further demonstrated that thousands of previously uncharacterized BGCs exist across bacterial taxa, suggesting a vast reservoir of unexplored bioactive compounds relevant for agriculture and biotechnology[18].

At an earlier stage of the discovery pipeline, machine learning approaches are increasingly used for BGC detection and characterization. Deep-learning frameworks can represent BGCs as sequences of Pfam protein domains, enabling models to capture long-range dependencies in biosynthetic pathways. The DeepBGC algorithm developed by Hannigan *et al.*[19] employs recurrent neural networks to detect BGC boundaries and classify predicted clusters according to product type and potential biological activity. Subsequent developments have incorporated self-supervised representation learning, which enables models to learn latent features of biosynthetic pathways without requiring extensive labeled datasets. Rios-Martinez *et al.*[12] demonstrated that the BiGCARP model applied masked-language modeling to approximately 127,000 BGC Pfam sequences, showing that self-supervised training can improve both BGC detection and product-class prediction. Other computational frameworks have similarly combined genomics with metabolomics to map biosynthetic diversity and prioritize novel natural products for experimental validation[19,20].

In addition to microbial metabolite discovery, AI-driven methods are increasingly applied to protein- and peptide-based biopesticide candidates, particularly antimicrobial peptides (AMPs) targeting plant pathogens. Advances in sequence modeling have enabled large-scale exploration of peptide space using deep learning architectures. Su *et al.*[8] screened over 6,700 *Bacillus* genomes to identify short peptides (10-100 amino acids) using deep-learning models including BERT, Mamba, CNN-LSTM and CNN-Attention. The analysis predicted 4,993,389 potential antimicrobial peptides. Two high-confidence candidates, cAMP1 and cAMP2, were validated through molecular dynamics and experimental assays, showing activity against *E. coli*, *Staphylococcus aureus* and several agricultural pathogens. Importantly, the study released datasets and code, highlighting the increasing emphasis on reproducibility and open computational workflows in natural-product discovery. Candidate peptides were subsequently filtered using cross-validation and structural simulations and selected peptides were experimentally validated for antimicrobial activity. Earlier studies had previously demonstrated that machine-learning models trained on peptide databases can accurately classify antimicrobial peptides and predict functional properties such as toxicity, stability and membrane activity[21,22].

Deep learning has also been used to design new peptides with improved bioactivity through generative modeling approaches. For example, variational autoencoders and generative adversarial networks have been applied to generate novel antimicrobial peptide sequences that maintain structural and physicochemical features associated with antimicrobial activity[22]. Such generative approaches allow exploration of vast sequence spaces that would be inaccessible through traditional mutagenesis or rational design strategies.

Table 2: Representative AI tools, their modeling approaches, datasets, application stages and their key functions

AI tool/Framework	Model type	Data type/Dataset size	Application stage	Key function	References
antiSMASH	Rule-based+ML-assisted genome mining	Thousands of microbial genomes	Discovery	Detection and annotation of biosynthetic gene clusters (BGCs) for natural product discovery	Medema *et al.*[6]
DeepBGC	Recurrent neural network (RNN)+Random forest	Bacterial genome datasets with Pfam domain sequences	Discovery	Predicts BGC boundaries and classifies product types and potential activity	Raies and Bajic[10]
BiGCARP	Self-supervised masked language model	~127,000 BGC Pfam domain sequences	Discovery/ Characterization	Improves BGC detection and product-class prediction through representation learning	Rios-Martinez *et al.*[12]
PRISM	Genome mining+ cheminformatics modeling	Microbial genome datasets	Discovery	Predicts chemical structures of natural products encoded by BGCs	Skinnider *et al.*[24]
MIBiG database Models	Database-driven ML annotation	>2,000 experimentally characterized BGCs	Discovery/ Validation	Benchmark dataset for training ML models to predict metabolite classes	Kautsar *et al.*[20]
AMP prediction Networks (AMPScanner, etc.)	CNN/Deep Neural Networks	Peptide sequence databases (10,000+ peptides)	Candidate identification	Predict antimicrobial and pesticidal peptide activity	Lee *et al.*[22]
DeepAMP/AMP-BERT	Transformer-based language models	Large peptide sequence datasets	Candidate discovery	Identifies antimicrobial peptides and predicts stability and toxicity	Das *et al.*[23]
Generative AMP	Variational Autoencoder/ GAN Models	Peptide sequence datasets	Candidate design	Generates novel antimicrobial peptides with improved bioactivity	Dean *et al.*[21]
dsRNA design pipelines (siRNA / RNAi tools)	Sequence alignment+ ML off-target prediction	Insect genome and transcriptome datasets	Target identification	Designs dsRNA molecules targeting pest genes with minimal off-target effects	Li *et al.*[13] and Taning *et al.*[16]
RNAi design bioinformatics platforms	Sequence-based predictive models	Pest genomic datasets	Target validation	Predict gene-silencing efficiency and ecological safety	Li *et al.*[13] and Taning *et al.*[16]

Similarly, computational approaches support the design of RNA interference (RNAi)–based biopesticides, where sequence-design algorithms and off-target prediction models help optimize dsRNA molecules for species-specific gene silencing while minimizing ecological risks[6,7]. Several computational frameworks have therefore been developed to support AI-driven discovery of biopesticides across microbial metabolite discovery, antimicrobial peptide prediction and RNA interference design. Representative tools, their modeling approaches, datasets and application stages are summarized in Table 2.

AI IN FORMULATION AND DELIVERY SYSTEMS

Formulation represents a major translational bottleneck in biopesticide development, particularly for modalities such as botanical extracts, microbial agents and RNAi-based products. Machine-learning approaches can assist formulation optimization by modeling relationships between formulation parameters and physicochemical performance. Botanical biopesticides, including essential oils and plant-derived secondary metabolites, often suffer from high volatility, rapid photodegradation and poor water solubility, which can significantly reduce their field persistence and efficacy[24,25]. Similarly, microbial biopesticides must maintain cell viability, spore stability and metabolic activity during storage and application, while RNAi-based biopesticides face challenges related to dsRNA degradation, environmental instability and delivery efficiency[16]. Nano-formulation and encapsulation technologies have therefore been explored to enhance stability and controlled release of biopesticide active ingredients. Nanoemulsion-based delivery systems have been developed to improve the stability and bioavailability of essential oils such as thymol, eugenol and citronella oil, which exhibit pesticidal activity but degrade rapidly under environmental conditions[15]. Earlier studies of essential-oil nanoformulations describe the use of algorithms such as artificial neural networks (ANN), genetic algorithms, support vector machines and random forest to optimize nanoemulsion droplet size, encapsulation efficiency and formulation stability. However, these studies also highlight the lack of standardized datasets and reporting frameworks for pesticide-grade formulations as a key limitation, which constrains the reproducibility and generalizability of machine-learning models across different formulation systems[12,13].

AI IN BIOPROCESS OPTIMIZATION AND SCALABILITY

AI plays a transformative role in bioprocessing by enabling more efficient optimization, reducing trial-and-error experimentation and supporting scalable, cost-effective production of microbial biopesticides. Machine learning in bioprocess optimization focuses less on individual algorithms and more on its ability to model complex, multivariable fermentation systems and improve production outcomes. By learning relationships between key parameters including temperature, pH, nutrient levels and aeration, AI-driven models can predict optimal conditions that maximize microbial performance. In microbial biopesticide production, this translates into tangible improvements in critical outputs, including higher spore yields, enhanced biomass production and increased metabolite titers[2,4,5]. Machine learning also helps identify process bottlenecks, reduce variability and improve production stability over time. Evidently, AI does not function as a standalone solution. Its effectiveness depends on data quality, experimental design and system observability. As a result, the most successful strategies integrate data-driven models with mechanistic understanding and controlled experimentation[26].

PREDICTIVE MODELLING FOR FIELD EFFICACY AND TRANSLATION

A persistent challenge in biopesticide development is the domain shift between laboratory efficacy assays and real-world field performance. Laboratory bioassays typically evaluate pesticidal activity under controlled conditions, whereas field outcomes depend on complex environmental variables including temperature, humidity, ultraviolet radiation, host plant physiology, soil characteristics, microbiome interactions and application practices[4]. These factors can substantially influence the survival, persistence and activity of microbial or botanical biopesticides, often leading to discrepancies between laboratory predictions and field-level pest suppression[9]. Consequently, predictive modeling approaches that incorporate environmental and biological variables are increasingly explored as tools for improving efficacy prediction and translation from laboratory studies to field applications[27]. Machine-learning models integrating environmental sensor data and crop characteristics have been used to forecast pest outbreaks and plant disease risk in agricultural systems, demonstrating the feasibility of data-driven forecasting approaches for crop-protection strategies. Despite these promising developments, predictive modeling for field translation remains limited by data scarcity and heterogeneity across experimental systems. Field trials often differ in crop species, pest populations, environmental conditions and application protocols, making it difficult to develop generalized predictive models. As a result, many current ML applications focus on intermediate mechanistic traits, such as microbial colonization, metabolite production, or environmental persistence, rather than direct yield outcomes. These surrogate endpoints can nevertheless provide valuable insights into the factors that determine field efficacy and guide the design of targeted field experiments[27,28].

ENVIRONMENTAL RISK ASSESSMENT AND ECOTOXICOLOGY

Ecotoxicology is one area where machine-learning applications have achieved relatively mature benchmarking practices, particularly in the context of pesticide risk assessment. Although biopesticides are derived from biological sources, regulatory agencies such as the U.S. Environmental Protection Agency (EPA) and the European Food Safety Authority (EFSA) still require comprehensive ecological safety evaluation, particularly for non-target organisms, including pollinators such as the honey bee (*Apis mellifera*)[11]. As honey bees play a crucial role in global agricultural productivity and ecosystem functioning, toxicity testing and predictive modeling for bee safety have become a central component of regulatory ecotoxicology. As a result, computational toxicology approaches originally developed for synthetic pesticides are increasingly being adapted to evaluate potential ecological risks associated with emerging biopesticide products[29].

Predictive toxicology models rely heavily on high-quality datasets and well-defined toxicity endpoints. However, many studies highlight that inconsistent reporting, heterogeneous experimental protocols and limited dataset availability often complicate the evaluation and comparison of machine-learning models. To address these challenges, several recent initiatives have focused on developing curated benchmarking datasets specifically designed for reproducible ecotoxicology modeling. One prominent example is the ApisTox dataset, a curated resource for honey-bee toxicity prediction. ApisTox integrates toxicity information from multiple sources, including the ECOTOX database, the Pesticide Properties Database (PPDB) and other publicly available chemical toxicity repositories. Following extensive curation and deduplication, the dataset contains 1,035 chemical compounds, representing a diverse set of molecular structures with 424 Bemis-Murcko scaffolds, many of which appear only once in the dataset.

Importantly, ApisTox introduces clearly defined training-testing splits, including a time-based split derived from PubChem literature publication dates. This temporal validation strategy mimics real-world discovery scenarios in which models trained on previously known compounds are evaluated on newly identified molecules. Such approaches allow researchers to assess the out-of-distribution generalization ability of predictive models, which is a critical requirement for regulatory applications[11]. Complementary resources such as BeeToxAI further provides a fully documented modeling workflow that incorporates external validation procedures, applicability-domain estimation and model interpretability tools, aligning its methodology with Organisation for Economic Co-operation and Development (OECD) principles for quantitative structure-activity relationship (QSAR) modeling[30].

Beyond dataset development, researchers have increasingly explored advanced machine-learning architectures to improve predictive accuracy in ecotoxicology modeling. Graph-based neural networks are particularly suitable for this task because they represent chemical molecules as graph structures composed of atoms and bonds. For instance, a graph attention convolutional neural network (GACNN) trained on 720 pesticide compounds and evaluated on an external validation set of 90 pesticides demonstrated improved predictive performance compared with traditional machine-learning models such as random forests and support vector machines when predicting honey-bee toxicity. Importantly, the attention mechanism within the model highlights molecular substructures associated with toxicological outcomes, enabling interpretable insights into chemical features linked to bee toxicity[29]. Similar graph neural network approaches have been applied to broader environmental toxicology datasets such as ToxCast and ECOTOX, where deep-learning models have demonstrated improved predictive performance for chemical hazard classification compared with conventional QSAR approaches[7,29]. Traditional QSAR models were widely used to estimate pesticide toxicity for aquatic organisms, birds and pollinators, providing foundational methodologies for modern machine-learning-based ecotoxicology[11,31].

REGULATORY ALIGNMENT IN BIOPESTICIDE DEVELOPMENT

Regulatory assessment plays a pivotal role in biopesticide development because it defines the data requirements, acceptable endpoints and documentation standards necessary for product approval. Unlike pharmaceutical pipelines, where regulatory review often occurs after development, pesticide regulation directly shapes the research design, experimental endpoints and evidence generation throughout the development process. Consequently, regulatory frameworks strongly influence how discovery, efficacy testing and environmental risk assessments are conducted.

In the United States, pesticide registration is governed by the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) and the associated regulations under 40 CFR Part 158, which outline the minimum data requirements for pesticide registration. These guidelines distinguish between conventional chemical pesticides, biochemical pesticides and microbial pesticides, recognizing the unique biological characteristics of biopesticides and allowing flexibility for case-specific data requirements depending on the active ingredient. For microbial pesticides, regulatory dossiers typically require information on strain identity, pathogenicity, infectivity, secondary metabolite production, environmental persistence and potential effects on non-target organisms, whereas biochemical pesticides may emphasize environmental fate and ecological exposure[32].

In the European Union, microbial biopesticides are regulated under the plant protection product framework established by Regulation (EC) No 1107/2009, which governs the approval of active substances and plant protection products. Supporting guidance documents emphasize that regulatory requirements must accommodate the diversity of microorganisms used as biological control agents. As a result, not all data requirements apply uniformly; applicants must justify the relevance of specific endpoints for a given strain. Key regulatory considerations include strain-level characterization, potential production of toxins or secondary metabolites, antimicrobial resistance risks and impacts on non-target organisms such as pollinators, soil microorganisms and aquatic species. Regulatory challenges are particularly evident for RNAi-based biopesticides, such as double-stranded RNA (dsRNA) products that silence pest genes. Because dsRNA operates through sequence-specific gene silencing, regulatory agencies emphasize bioinformatic screening to detect potential off-target effects in non-target organisms[14,32,33].

EMERGING COMPUTATIONAL APPROACHES IN AI-DRIVEN BIOPESTICIDE RESEARCH

Recent advances in artificial intelligence for biological sciences have introduced foundation models and protein language models, which are increasingly transforming molecular discovery and bioactive compound design. Unlike

traditional machine-learning models trained for specific prediction tasks, foundation models are trained on extremely large biological datasets and learn generalized representations of biological sequences and molecular structures. Protein language models such as ProtBERT, ESM (Evolutionary Scale Modeling) and ESMFold apply transformer-based architectures originally developed for natural language processing to protein sequences, allowing the extraction of structural and functional information directly from amino-acid sequences[7,34]. These models have demonstrated strong performance in protein structure prediction, functional annotation and peptide activity prediction. The ESMFold integrates large-scale protein language modeling with structural inference to predict protein structures with high accuracy and significantly reduced computational cost compared with traditional methods. Similarly, generative biology frameworks such as ProtGPT2 enable the design of novel protein sequences with desired functional properties by sampling from learned sequence distributions[34]. In the context of biopesticide development, these tools offer promising opportunities for the design of pesticidal peptides, enzyme-based biocontrol agents and microbial metabolites, enabling computational exploration of biological sequence space beyond naturally occurring molecules[5,6].

Another emerging direction in computational agriculture is multimodal learning, which integrates heterogeneous biological and environmental datasets within unified predictive frameworks. Biopesticide development involves multiple layers of information, including genomic sequences, metabolomic profiles, microbial community interactions, plant physiology, formulation chemistry and environmental conditions. Traditional machine-learning models typically analyze these datasets independently, which limits their ability to capture complex cross-domain interactions. Multimodal learning approaches address this limitation by combining diverse data types into joint representation models that simultaneously analyze genomic, chemical, ecological and formulation-related variables. Integrating genomic data from microbial strains, metabolomic signatures of secondary metabolites and environmental field variables can help identify microbial candidates that are both biologically active and environmentally robust. Similarly, combining chemical descriptors of botanical compounds with formulation parameters and environmental persistence data can improve predictive models for formulation stability and field efficacy. As large-scale agricultural and environmental datasets continue to expand, multimodal AI frameworks are expected to play a central role in connecting molecular discovery with real-world agricultural performance[22,26,27].

A complementary methodological approach gaining importance in biological discovery pipelines is active learning, which integrates machine-learning predictions with iterative experimental validation[29]. In conventional machine-learning workflows, models are trained on fixed datasets and subsequently used to generate predictions for new candidate molecules or organisms. However, in biological systems where experimental validation is costly and time-consuming, active learning allows models to dynamically select the most informative experiments for improving predictive performance. In this framework, the model identifies candidate compounds, microbial strains, or peptide sequences with high uncertainty or high predicted utility and prioritizes them for laboratory testing. The resulting experimental data are then incorporated into the training dataset, enabling the model to iteratively refine its predictions. Active-learning strategies are particularly useful in biopesticide discovery pipelines, when millions of potential peptides, metabolites, or RNAi sequences may be generated computationally but only a limited number can be experimentally validated. By guiding experimental design toward the most informative candidates, active learning can significantly reduce experimental costs while improving model reliability and translational relevance[35].

Equally important for advancing AI-driven biopesticide research is the development of standardized and interoperable biological datasets. Machine-learning models depend heavily on the availability of well-structured and reproducible data, yet biopesticide research data are often fragmented across genomic repositories, laboratory experiments, agricultural field trials and regulatory documentation. The adoption of findable, accessible, interoperable and reusable data (FAIR) has therefore been widely advocated to improve data sharing and computational reproducibility in life sciences[36]. FAIR-compliant datasets facilitate the integration of heterogeneous data sources and enable more reliable benchmarking of machine-learning models. In microbial ecology and environmental genomics, the Minimum Information about any (x) Sequence (MIxS) metadata standards provide structured guidelines for describing genomic and environmental sequencing data, including contextual information about sample origin, environmental conditions and experimental protocols. Such metadata standards are critical for ensuring that genomic data used in machine-learning models are properly contextualized and comparable across studies. Additionally, the increasing availability of open metabolomics repositories, such as GNPS (Global Natural Products Social Molecular Networking) and MetaboLights, provides valuable resources for studying microbial secondary metabolites and plant-derived

compounds with pesticidal potential[37]. Integrating these open data infrastructures with machine-learning workflows can significantly improve the reproducibility, scalability and collaborative potential of AI-driven biopesticide discovery.

CHALLENGES AND LIMITATIONS

Biopesticide development faces several structural and methodological challenges that limit the effective translation of artificial intelligence and predictive analytics into deployable agricultural products. One major constraint is data fragmentation and limited standardization. Relevant information is often dispersed across genomic databases, laboratory assay records, site-specific field trials and regulatory submissions, frequently stored in incompatible formats and units. Even in relatively well-studied areas such as ecotoxicology, datasets may contain inconsistencies, duplicates and varying experimental protocols, requiring extensive preprocessing before they can be used for machine-learning workflows[3,26]. A second limitation arises from bias, confounding variables and domain shift. Many biopesticide outcomes depend strongly on environmental context, including host plant genotype, microbiome composition, climatic conditions and formulation stability. Models trained on narrow experimental datasets can therefore capture spurious correlations and may perform poorly when applied to real agricultural environments. Benchmark datasets such as ApisTox attempt to address this issue through time-based or distribution-aware validation strategies[11], but comparable standardized validation practices remain rare in microbial or formulation research. Another challenge relates to model interpretability and regulatory acceptance. While researchers may prioritize predictive accuracy, regulatory agencies often require mechanistic justification, applicability-domain analysis and transparent uncertainty estimates. In addition, experimental validation remains the primary bottleneck, as computational pipelines can generate large numbers of candidate peptides, metabolites, or RNA-based molecules, but confirming efficacy, stability, manufacturability and non-target safety requires time-consuming laboratory and field testing. Finally, regulatory documentation requirements present significant barriers. Microbial biopesticide dossiers must address organism identity, pathogenicity potential, metabolite production and ecological effects under frameworks[22,23], which often require additional organism-specific data beyond minimum requirements. These limitations highlight the need for standardized datasets, interpretable modeling frameworks and stronger integration between computational predictions and experimental validation.

FUTURE PROSPECTS OF AI-DRIVEN BIOPESTICIDE DEVELOPMENT

Biopesticide research and predictive analytics integrate heterogeneous data sources including genomic sequences, metabolomic profiles, chemical descriptors, formulation variables, ecological parameters and experimental bioassay results to identify potential bioactive compounds and predict their efficacy, stability and safety prior to large-scale laboratory or field evaluation. By shifting the discovery process from largely empirical trial-and-error experimentation to a more rational and data-informed approach, AI-based predictive systems have the potential to substantially improve the efficiency, speed and reliability of biopesticide development.

Moreover, several emerging computational paradigms are expected to further enhance the capabilities of AI-driven discovery pipelines. The application of foundation models and generative artificial intelligence in biological sequence design is particularly promising. Protein language models trained on massive sequence databases are capable of learning latent representations of protein structure and function directly from amino-acid sequences. These models can be used to predict enzymatic activity, structural stability and interaction properties of proteins and peptides. In the context of biopesticide development, such approaches may enable the rational design of pesticidal peptides, antimicrobial proteins, or microbial enzymes with improved target specificity and environmental stability. Generative AI systems further expand these possibilities by enabling the computational design of entirely novel protein sequences or biosynthetic pathways, thereby expanding the chemical and biological diversity available for biological pest control.

Additional future development involves multimodal machine-learning frameworks capable of integrating diverse biological and environmental datasets within unified predictive models. Biopesticide performance is influenced by a complex combination of factors, including microbial metabolism, plant-microbe interactions, pest physiology, formulation chemistry and environmental conditions such as temperature, humidity and soil characteristics. Multimodal learning architectures allow these heterogeneous datasets to be analyzed simultaneously, enabling more comprehensive modeling of biological interactions. Integrating microbial genomic data with metabolomic signatures and environmental variables may improve the identification of microbial strains capable of producing bioactive metabolites under real field conditions. Similarly, combining chemical descriptors of botanical compounds with formulation parameters and ecological exposure data could enhance predictions of formulation stability and environmental persistence.

The integration of active learning and iterative experimental validation also represents a promising strategy for improving the efficiency of biopesticide discovery. In active-learning frameworks, machine-learning models dynamically select the most informative experiments to perform based on prediction uncertainty or expected information gain. Experimental outcomes are then incorporated into the training dataset, allowing the model to iteratively refine its predictions through a continuous design–test–learn cycle. Such approaches are particularly valuable in biological research, where experimental validation is expensive and time-consuming. By prioritizing the most informative candidate molecules, microbial strains, or RNAi sequences for testing, active learning can significantly reduce experimental costs while improving model reliability and translational relevance.

Future progress will also depend on improvements in data infrastructure, interoperability and standardization. Machine-learning models require high-quality, well-annotated datasets for reliable training and validation; however, biopesticide research data are often fragmented across genomic repositories, laboratory experiments, agricultural field trials and regulatory documentation. Expanding open biological databases, including genomic, metabolomic and natural-product repositories, will further enhance the availability of training data for predictive modeling.

CONCLUSION

The AI driven predictive analytics refers to the application of computational algorithms, machine-learning models and data-driven frameworks to analyze complex biological datasets in order to identify patterns, predict outcomes and guide scientific decision making. This review highlights the transformative role of AI and predictive analytics in biopesticide development, from discovery and candidate optimization to formulation, field-efficacy prediction and environmental risk assessment. AI-driven approaches, including machine learning and deep neural networks, have expanded the identification of microbial metabolites, pesticidal peptides and botanical compounds, while computational toxicology improves non-target safety predictions. Key challenges remain, including fragmented and heterogeneous datasets, limited model interpretability and experimental validation bottlenecks. Nevertheless, emerging AI strategies such as foundation models, multimodal learning and active-learning frameworks, along with FAIR data practices and standardized metadata, offer promising avenues to accelerate biopesticide development. Integrating these computational innovations with robust experimental workflows can enhance the creation of safe, effective and sustainable biopesticides, supporting global food security and climate-resilient agriculture.

SIGNIFICANCE STATEMENT

This systematic review highlights the transformative potential of artificial intelligence in accelerating sustainable biopesticide development. By integrating genome mining, deep learning and predictive ecotoxicology, AI addresses traditional bottlenecks in discovery, formulation and regulatory safety. These computational innovations offer a data-driven pathway to enhance crop protection, supporting global food security and climate-resilient agricultural practices.

REFERENCES

1. Fenibo, E.O. and T. Matambo, 2025. Biopesticides for sustainable agriculture: Feasible options for adopting cost-effective strategies. Front. Sustainable Food Syst., Vol. 9. 10.3389/fsufs.2025.1657000.
2. Mawcha, K.T., L. Malinga, D. Muir, J. Ge and D. Ndolo, 2025. Recent advances in biopesticide research and development with a focus on microbials. F1000Research, Vol. 10. 10.12688/f1000research.154392.5.
3. Chitranshi, R., Enespa, R. Singh and P. Chandra, 2021. Opportunities and Challenges in the Application of Biopesticides in Organic Farming. In: Biopesticides in Organic Farming: Recent Advances, Awasthi, L.P. (Ed.), CRC Press, Boca Raton, Florida, ISBN: 9781003027690, pp: 23-28.
4. Kashyap, G.S., P. Kamani, M. Kanojia, S. Wazir, K. Malik, V.K. Sehgal and R. Dhakar, 2024. Revolutionizing Agriculture: A Comprehensive Review of Artificial Intelligence Techniques in Farming. In: Research Advances in Intelligent Computing, Verma, A. and P. Verma (Eds.), CRC Press, Boca Raton, Florida, ISBN: 9781003433941, pp: 82-112.
5. Zhu, S., H. Xu, Y. Liu, Y. Hong, H. Yang, C. Zhou and L. Tao, 2025. Computational advances in biosynthetic gene cluster discovery and prediction. Biotechnol. Adv., Vol. 79. 10.1016/j.biotechadv.2025.108532.
6. Medema, M.H., K. Blin, P. Cimermancic, V. de Jager and P. Zakrzewski *et al*., 2011. antiSMASH: Rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Res., 39: W339-W346.

7. Jumper, J., R. Evans, A. Pritzel, T. Green and M. Figurnov *et al*., 2021. Highly accurate protein structure prediction with AlphaFold. Nature, 596: 583-589.
8. Su, H., M. Gu, Z. Qu, Q. Wang and J. Jin *et al*., 2025. Discovery of antimicrobial peptides from *Bacillus* genomes against phytopathogens with deep learning models. Chem. Biol. Technol. Agric., Vol. 12. 10.1186/s40538-025-00751-9.
9. Javed, K., G. Smagghe, Q. Wang, H. Javed and Y. Wang, 2025. Artificial intelligence in crop protection: Revolutionizing agriculture for a sustainable future. Inf. Process. Agric., 10.1016/j.inpa.2025.12.003
10. Raies, A.B. and V.B. Bajic, 2016. *In silico* toxicology: Computational methods for the prediction of chemical toxicity. WIREs Comput. Mol. Sci., 6: 147-172.
11. Adamczyk, J., J. Poziemski and P. Siedlecki, 2025. ApisTox: A new benchmark dataset for the classification of small molecules toxicity on honey bees. Sci. Data, Vol. 12. 10.1038/s41597-024-04232-w
12. Rios-Martinez, C., N. Bhattacharya, A.P. Amini, L. Crawford and K.K. Yang, 2023. Deep self-supervised learning for biosynthetic gene cluster detection and product classification. PLoS Comput. Biol., Vol. 19. 10.1371/journal.pcbi.1011162.
13. Li, Z., Y. Liu, Y. Liang, T. Pan and J. Liu, 2026. RNA interference-based pesticides: Mechanism, application, and commercialization in sustainable pest management. Pestic. Biochem. Physiol., Vol. 219. 10.1016/j.pestbp.2026.107034.
14. Kah, M., S. Beulke, K. Tiede and T. Hofmann, 2013. Nanopesticides: State of knowledge, environmental fate and exposure modeling. Crit. Rev. Environ. Sci. Technol., 43: 1823-1867.
15. Mustafa, I.F. and M.Z. Hussein, 2020. Synthesis and technology of nanoemulsion-based pesticide formulation. Nanomaterials, Vol. 10. 10.3390/NANO10081608.
16. Taning, C.N.T., S. Arpaia, O. Christiaens, A. Dietz-Pfeilstetter and H. Jones *et al*., 2020. RNA-based biocontrol compounds: Current status and perspectives to reach the market. Pest. Manage. Sci., 76: 841-845.
17. Navarro-Muñoz, J.C., N. Selem-Mojica, M.W. Mullowney, S.A. Kautsar and J.H. Tryon *et al*., 2020. A computational framework to explore large-scale biosynthetic diversity. Nat. Chem. Biol., 16: 60-68.
18. Cimermancic, P., M.H. Medema, J. Claesen, K. Kurita and L.C.W. Brown *et al*., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell, 158: 412-421.
19. Hannigan, G.D., D. Prihoda, A. Palicka, J. Soukup and O. Klempir *et al*., 2019. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. Nucleic Acids Res., 47: e110-e110.
20. Kautsar, S.A., K. Blin, S. Shaw, J.C. Navarro-Muñoz and B.R. Terlouw *et al*., 2020. MIBiG 2.0: A repository for biosynthetic gene clusters of known function. Nucleic Acids Res., 48: D454-D458.Dean, S.N. and S.A. Walper, 2020. Variational autoencoder for generation of antimicrobial peptides. ACS Omega, 5: 20746-20754.
21. Lee, H., S. Lee, I. Lee and H. Nam, 2023. AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. Protein Sci., Vol. 32. 10.1002/pro.4529.
22. Das, P., T. Sercu, K. Wadhawan, I. Padhi and S. Gehrmann *et al*., 2021. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nat. Biomed. Eng., 5: 613-623.
23. Skinnider, M.A., N.J. Merwin, C.W. Johnston and N.A. Magarvey, 2017. PRISM 3: Expanded prediction of natural product chemical structures from microbial genomes. Nucleic Acids Res., 45: W49-W54.
24. Pavela, R. and G. Benelli, 2016. Essential oils as ecofriendly biopesticides? Challenges and constraints. Trends Plant Sci., 21: 1000-1007.
25. Cheng, Y., X. Bi, Y. Xu, Y. Liu and J. Li *et al*., 2023. Artificial intelligence technologies in bioprocess: Opportunities and challenges. Bioresour. Technol., Vol. 369. 10.1016/j.biortech.2022.128451.
26. Madhuri, E.V., J.S. Rupali, S.P. Sharan, N.S. Pooja and G.S. Sujatha *et al*., 2025. Transforming pest management with artificial intelligence technologies: The future of crop protection. J. Crop Health, Vol. 77. 10.1007/s10343-025-01109-9.
27. Ajakwe, S.O., N.F. Esomonu, O. Deji-Oloruntoba, I.U. Ajakwe, J.M. Lee and D.S. Kim, 2024. Machine Learning in UAV-Assisted Smart Farming. In: Applications of Machine Learning in UAV Networks, Hassan, J. and S. Alsamhi (Eds.), IGI Global, United States, ISBN: 9798369305782, pp: 217-245.
28. Wang, F., J.F. Yang, M.Y. Wang, C.Y. Jia, X.X. Shi, G.F. Hao and G.F. Yang, 2020. Graph attention convolutional neural network model for chemical poisoning of honey bees' prediction. Sci. Bull., 65: 1184-1191.
29. Moreira-Filho, J.T. , R.C. Braga, J.M. Lemos, V.M. Alves and J.V.V.B. Borba *et al*., 2021. BeeToxAI: An artificial intelligence-based web app to assess acute toxicity of chemicals to honey bees. Artif. Intell. Life Sci., Vol. 1. 10.1016/j.ailsci.2021.100013.
30. Gramatica, P., 2007. Principles of QSAR models validation: Internal and external. QSAR Combinatorial Sci., 26: 694-701.
31. Shekhar, C., R. Khosya, K. Thakur, D. Mahajan, R. Kumar, S. Kumar and A.K. Sharma, 2024. A systematic review of pesticide exposure, associated risks, and long-term human health impacts. Toxicol. Rep., Vol. 13. 10.1016/j.toxrep.2024.101840.

32. Vermelho, A.B., J.V. Moreira, I.T. Akamine, V.S. Cardoso and F.R.P. Mansoldo, 2024. Agricultural pest management: The role of microorganisms in biopesticides and soil bioremediation. Plants, Vol. 13. 10.3390/plants13192762.
33. Ferruz, N., S. Schmidt and B. Höcker, 2022. ProtGPT2 is a deep unsupervised language model for protein design. Nat. Commun., Vol. 13. 10.1038/s41467-022-32007-7.
34. Anandhi, G. and M. Iyapparaja, 2024. Systematic approaches to machine learning models for predicting pesticide toxicity. Heliyon, Vol. 10. 10.1016/j.heliyon.2024.e28752.
35. Wilkinson, M.D., M. Dumontier, I.J. Aalbersberg, G. Appleton and M. Axton *et al.*, 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data, Vol. 3. 10.1038/sdata.2016.18.
36. Wang, Y., Y. Wang, Z. Zhang, K. Xu, Q. Fang, X. Wu and S. Ma, 2025. Molecular networking: An efficient tool for discovering and identifying natural products. J. Pharm. Biomed. Anal., Vol. 259. 10.1016/j.jpba.2025.116741.