OXFORD (GIGA)^n SCIENCE

GigaScience, 2025, **14**, 1–12

DOI: 10.1093/gigascience/giaf057

Review

Artificial intelligence: the human response to approach the complexity of big data in biology

Giovanni Melandri[1,2,‡], Georges R-Radohery[1,‡], Chloé Beaumont[1], Sara M. de Cripan[3], Coralie Muller[1,4], Luca Piras[5], Maria Alcina Pereira[6,7], Andreia Ferreira Salvador[6,7], Xavier Domingo-Almenara[3,8], Marie Bolger[9], Sophie Colombié[1,10], Sylvain Prigent[1,10], Biotza Gutierrez Arechederra[5], Nuria Canela Canela[3], and Pierre Pétriacq[1,10,*]

[1]University of Bordeaux, INRAE, UMR 1332 BFP, Villenave d'Ornon 33140, France
[2]School of Plant Sciences, University of Arizona, Tucson, AZ 85721, USA
[3]Centre for Omics Sciences (COS), Eurecat—Technology Centre of Catalonia & Rovira i Virgili University Joint Unit, Unique Scientific and Technical Infrastructures (ICTS), Reus, Catalonia 43204, Spain
[4]Inria, Univ. Bordeaux, INRAE, F-33400, Talence, France
[5]EURECAT—Technology Centre of Catalonia, Barcelona, Catalonia 08005, Spain
[6]Centre of Biological Engineering, University of Minho, 4704-553 Braga, Portugal
[7]LABBELS—Associate Laboratory, Braga/Guimarães, Portugal
[8]Department of Electrical, Electronic and Control Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia 43007, Spain
[9]Institute of Bio- and Geosciences, IBG-4: Bioinformatics, Forschungszentrum Jülich, CEPLAS, BioSC, Jülich 52429, Germany
[10]Bordeaux Metabolome, MetaboHUB, PHENOME-EMPHASIS, Villenave d'Ornon 33140, France
*Correspondence address. Pierre Pétriacq, University of Bordeaux, INRAE, UMR1332 BFP, 33140 Villenave d'Ornon, France. E-mail: pierre.petriacq@inrae.fr
‡Equal contribution.

Abstract

Since the late 2010s, artificial intelligence (AI), encompassing machine learning and propelled by deep learning, has transformed life science research. It has become a crucial tool for advancing the computational analysis of biological processes, the discovery of natural products, and the study of ecosystem dynamics. This review explores how the rapid increase in high-throughput omics data acquisition has driven the need for AI-based analysis in life sciences, with a particular focus on plant sciences, animal sciences, and microbiology. We highlight the role of omics-based predictive analytics in systems biology and innovative AI-based analytical approaches for gaining deeper insights into complex biological systems. Finally, we discuss the importance of FAIR (findable, accessible, interoperable, reusable) principles for omics data, as well as the future challenges and opportunities presented by the increasing use of AI in life sciences.

Keywords: artificial intelligence, machine learning, deep learning, omics, life science, biology

Background

The explosion of omics requires artificial intelligence in the study of life sciences

In the past 2 decades, research and society have entered the "big data" era of life sciences. Technological advances have enhanced our ability to measure qualitative and quantitative variations of internal biological molecules (e.g., DNA, RNA, proteins, metabolites) and phenotypes, making the acquisition of large and complex omics datasets within a single experiment increasingly common.

The explosion of omics data in life sciences began with genomics, which was driven by the emergence of DNA next-generation sequencing (NGS) platforms nearly 20 years ago. While the groundbreaking discovery of the Sanger DNA sequencing method dates back to the 1970s, it took 3 decades for the advent of second-generation short-read sequencing-based NGS to further revolutionize DNA sequencing, dramatically increasing its affordability and throughput. This has led to the *de novo* assembly of thousands of animal and plant genomes [1, 2] and to the discovery of millions of genome-wide single nucleotide polymorphic (SNP) variants. High-throughput analysis of multiple gene transcripts (i.e., transcriptomics) began in the mid-1990s with the introduction of hybridization-based microarray technologies. However, it was not until the 2000s that NGS enabled a more accurate assessment of the qualitative and quantitative diversity (e.g., large dynamic range of expression levels and alternative splicing variants) of messenger RNAs. This technique, known as RNA sequencing (RNA-seq), uses NGS to sequence complementary DNAs (cDNAs) derived from RNA transcripts [3, 4]. The current third-generation single-molecule sequencing technologies (e.g., PacBio and Oxford Nanopore Technologies) have further improved the read length, throughput, and accuracy of data collection in the field of genomics and transcriptomics [5, 6]. The field of proteomics and metabolomics relies on the use of mass spectrometry (MS) techniques to explore the diversity of proteins and metabolites in both a qualitative and quantitative manner. Although mass spectrometers have been available since the late 1940s, it was their integration with gas chromatography (GC) or liquid chromatography (LC) and the development of ionization techniques such as electrospray ionization (ESI) and matrix-assisted laser desorption ionisation (MALDI) in the late 1980s that truly expanded their application to biological research [7, 8]. There are various ionization techniques in mass spectrometry and electronic impact ionization that, while historically important for profiling

Received: October 31, 2024. **Revised:** April 3, 2025. **Accepted:** April 9, 2025

Search query: ((artificial intelligence) AND (omics)) AND (life sciences)

Year	Count
2004	2
2005	0
2006	2
2008	5
2009	3
2010	1
2011	8
2012	9
2013	14
2014	24
2015	26
2016	38
2017	42
2018	78
2019	114
2020	165
2021	183
2022	182
2023	215
2024	251

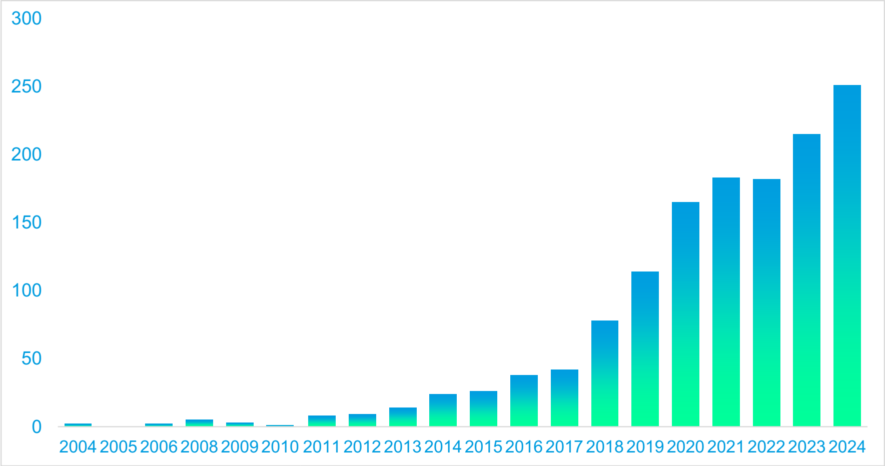

Figure 1: Number of publications found in PubMed including [artificial intelligence] AND [omics] AND [life sciences] from 2004 to 2024. In total, 1,362 publications were found (19 September 2024). Considering the past 20 years, a literature search using the queries [omics] AND [artificial intelligence] AND [life sciences] confirms that AI in life sciences is a rapidly expanding field of research.

primary compounds of biological samples, have been largely superseded by softer ionization methods such as ESI and MALDI. These newer techniques are more suitable for analyzing biomolecules as they cause less fragmentation and tend to preserve the integrity of molecules during ionization. Over the past 20 years, the advancement of high-resolution (HR) MS has been crucial in significantly enhancing the identification of proteins and metabolites. This has driven the widespread application of proteomics and metabolomics in the analysis of complex biological samples [9, 10].

Recent advancements in imaging technologies have improved life science research, benefiting not only the medical field [11] but also plant sciences. The field of plant/crop phenomics has rapidly evolved due to breakthroughs in sensor technology, machine vision, and automation technology [12]. Today, automated, noninvasive, high-throughput imaging and sensor technologies generate vast amounts of image and sensor data, presenting both opportunities and challenges for analysis.

The ability to generate high-throughput large-scale omics data through advanced technologies offers an unprecedented opportunity for exploring the complexity of biological systems in depth. Furthermore, integrating multiple omics datasets from a single experiment facilitates a "holistic" approach, revealing the potential of how the "molecular endophenome" (at the cellular/tissue level) is regulated and connected with the "external phenome" of biological organisms. However, disentangling and deciphering the intricate relationships among tens of thousands (sometimes millions) of molecular variables (i.e., SNPs, transcripts, proteins, and metabolites), which are interconnected among themselves and with the final phenotype, has been a major challenge in biological research over the past 2 decades [13, 14]. The use of high-dimensional solutions on complex omics datasets to address fundamental biological questions exceeds the capacity of the human brain. This requires a computer-based analytical approach, which can benefit from the constant improvements in machine processing power at all levels (single machine or physical/cloud-based clusters). For these reasons, "artificial intelligence" (AI) has emerged as a key tool in life science research (Fig. 1), with the expectation that AI will lead or assist in most of the future biological discoveries.

Artificial intelligence, machine learning, and deep learning

Despite its widespread use, the term AI remains an elusive "buzzword." From a scientific perspective, the difficulty in defining AI is associated with the complexity of the concept of intelligence *per se* and with the fact that, despite a resurgence of interest in AI started in the 1990s, fast progresses in AI research rapidly developed only from the 2010s, and, thus, this field of research is far from reaching a level of maturity that can be translated into a clear definition [15].

Oversimplifying, AI can be considered a branch of computer science focused on programming machines (typically 1 or more computers) to perform a specific tasks by learning from the information present in specific dataset(s) [16] (Fig. 2). This definition is appropriate only for "artificial narrow intelligence" or "weak AI," which is currently used for many routine and nearly ubiquitous applications such as spam filtering, speech recognition, language translation, online advertising, image tagging, and so on. However, this definition is not accurate for "artificial general intelligence" or "artificial super intelligence," which are both still far from being achieved. These forms of AI aim to develop machines capable of learning and understand from data in ways that are comparable to or surpass human intelligence [17].

Considering "artificial narrow intelligence" (hereafter AI will refer to this term) and, particularly, its most popular subfield "machine learning" (ML), the "learning" feature defines the process of using an algorithm that finds complex patterns in the training data and translates them into an object-level algorithm (such as a model of a domain problem), which, in turn, is able to make predictions about unobserved data. It is in the context of ML that biological research has benefited the most from the use of large and complex omics data [18, 19]. Biological data-based ML models have the double target of (i) accurately predicting experimental data and (ii) using this predicting ability to inform and direct the efforts of future research. When developing ML models, the characteristics of the training data determine the learning approach. Training data refer to the dataset used to teach an ML model, and a key distinction is whether these training data in-

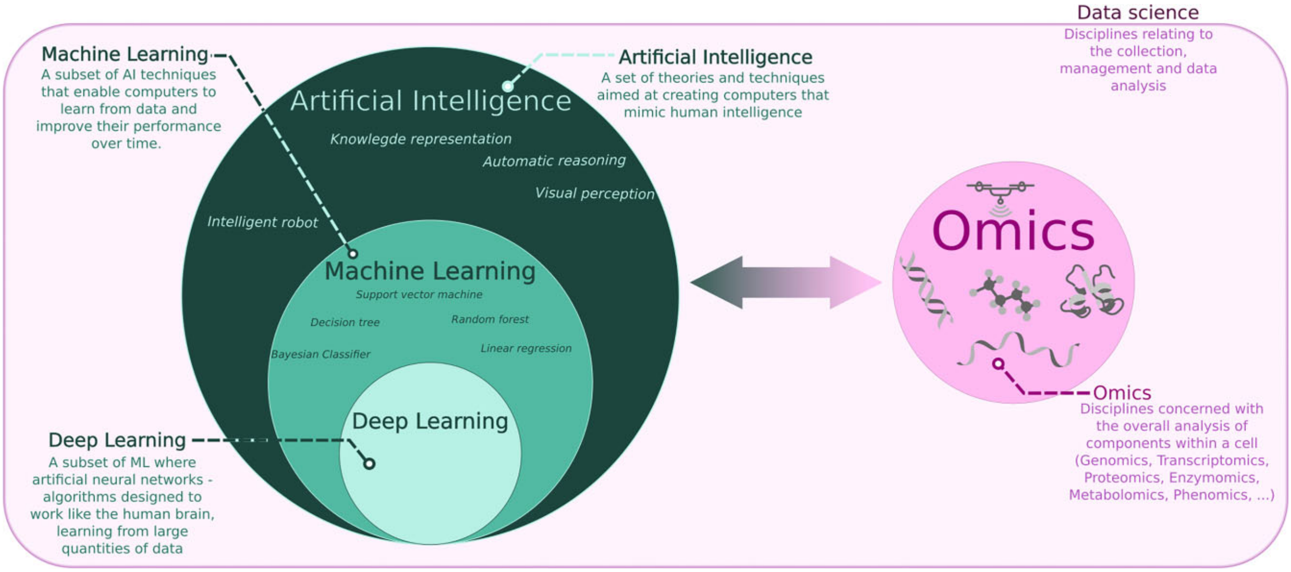
Data science
Disciplines relating to the collection, management and data analysis
Machine Learning
A subset of AI techniques that enable computers to learn from data and improve their performance over time.
Artificial Intelligence
Artificial Intelligence
A set of theories and techniques aimed at creating computers that mimic human intelligence
Knowlegde representation
Automatic reasoning
Visual perception
Intelligent robot
Machine Learning
Support vector machine
Decision tree
Random forest
Bayesian Classifier
Linear regression
Deep Learning
Deep Learning
A subset of ML where artificial neural networks - algorithms designed to work like the human brain, learning from large quantities of data
Omics
Omics
Disciplines concerned with the overall analysis of components within a cell (Genomics, Transcriptomics, Proteomics, Enzymomics, Metabolomics, Phenomics, ...)

Figure 2: Data science in the era of artificial intelligence, machine learning and deep learning: a dynamic schematic breakdown.

clude annotations, which determine the learning method applied. Training data can be labeled or unlabeled. Labeled data contain explicit tags, such as categories or numerical values, allowing the model to learn from predefined outcomes. When the data are labeled, the model follows a supervised learning approach. In contrast, unlabeled data lack predefined tags, requiring the model to extract patterns and relationships independently—a process known as unsupervised learning (Fig. 3) [20]. On the contrary, if the same data are labeled (with qualitative or quantitative tags), the ML model is defined as based on “supervised” learning. Unsupervised ML models are mainly used to deal with clustering problems where the algorithms (e.g., K-means clustering or DBSCAN clustering) find relationships in the overall structure of the training data [20]. In supervised learning, the algorithm uses the provided labels as a guide to map data points to specific outcomes or classifications.

Machine learning is built on a few fundamental algorithms that serve as the foundation for more advanced techniques [21]. Here, we focus on a nonexhaustive list of these algorithms, particularly those that are interpretable and can clarify the importance of each variable in making predictions. This interpretability is especially valuable in life sciences, where it allows for a thorough utilization of information found in omics data—such as genomics, proteomics, and metabolomics—to uncover biological insights [22].

First, linear regression, used in supervised learning, predicts a continuous target variable by establishing a linear relationship between inputs and outputs and adjusting parameters to minimize the difference between expected and actual values. Linear regression is highly interpretable, as it establishes a clear linear relationship between input features and the target variable, allowing a straightforward understanding of how changes in each input affect the predicted outcome. The training process involves iterative adjustments to reduce prediction errors, often guided by optimization techniques [23]. Linear regression forms the basis for methods like Ridge and Lasso regression, which incorporate penalties to mitigate overfitting, where the model performs well on the training data but poorly on new data, by constraining model complexity [24]. These extensions enhance robustness and inform the weight adjustment mechanisms central to neural networks, demonstrating its role as a building block in machine learning [25].

Next, support vector machines (SVMs) address classification by identifying an optimal boundary that maximizes the distance to the nearest data points, which are known as support vectors. The support vectors provide insight into which data points are most crucial for the classification boundary. By examining these vectors and their corresponding features, one can infer which aspects of the data are influential in decision-making. For datasets where linear separation is infeasible, SVMs employ kernel functions—such as polynomial or radial basis functions—to transform the data into a higher-dimensional space, enabling complex separations [26]. This emphasis on margin maximization and spatial transformation influences modern deep learning architectures, notably in convolutional neural networks, where kernel-based operations are prevalent [27].

Decision trees, another supervised learning approach, partition the feature space into distinct regions based on threshold values applied to input variables. Criteria that maximize class separation, such as reducing impurity (e.g., Gini index) or minimizing prediction variance for regression tasks, determine these splits. Their interpretability—from clear, rule-based decisions—makes them particularly appealing for applications requiring transparency, such as omics-driven research [28]. Moreover, integrating them into ensemble methods like random forests, where multiple trees vote to enhance accuracy, or gradient-boosted trees, which iteratively refine predictions, amplifies their utility. These ensembles illustrate how decision trees evolve into robust predictive tools [29].

We take as the last example naive Bayes, which offers a probabilistic framework for classification. It assumes that features are independent within each class and uses probabilities to determine the most likely class for a given set of data. By applying Bayes’s theorem, it calculates how likely something belongs to a specific category based on past data. Hence, a naive Bayes classifier provides probabilities for each class rather than hard classifications. Furthermore, each feature’s contribution to the final decision can be calculated based on its likelihood of occurrence on each class, thus giving a strong interpretability to the model. However, assumptions like data independence could be

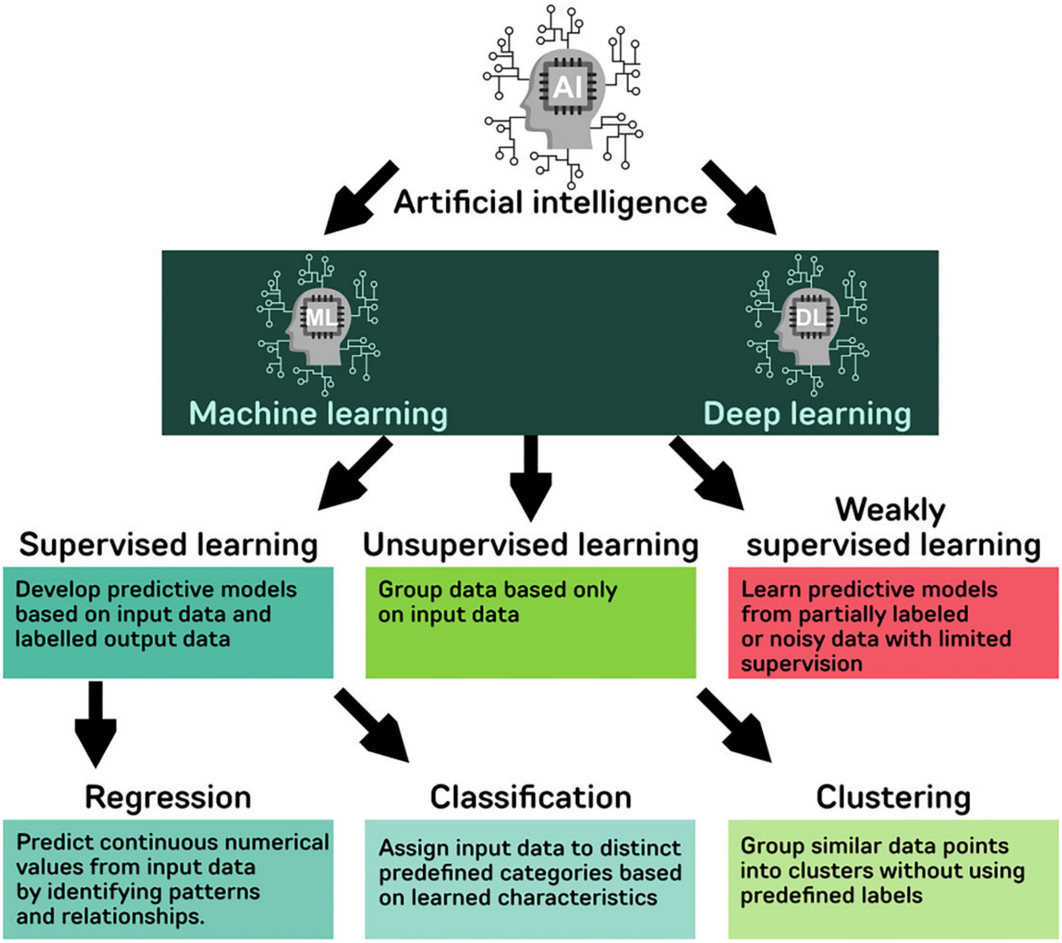
AI
Artificial intelligence
ML
Machine learning
DL
Deep learning
Supervised learning
Develop predictive models based on input data and labelled output data
Unsupervised learning
Group data based only on input data
Weakly supervised learning
Learn predictive models from partially labeled or noisy data with limited supervision
Regression
Predict continuous numerical values from input data by identifying patterns and relationships.
Classification
Assign input data to distinct predefined categories based on learned characteristics
Clustering
Group similar data points into clusters without using predefined labels

Figure 3: Major approaches in machine learning and deep learning.

unrealistic for biological data. Naive Bayes forms the basis for more advanced probabilistic models, like Bayesian networks. A Bayesian network extends the naive Bayes classifier by allowing dependencies between variables, unlike naive Bayes, which assumes all features are conditionally independent given the class label. It represents a probabilistic graphical model where nodes (variables) have directed edges (dependencies) between them [30].

Since its formal introduction in 2006, deep learning (DL) [31], based on diverse artificial neural network (ANN) algorithms, has further boosted the use of ML in many fields of research, particularly in speech recognition and image analysis [32] but also in the biological field, such as in regulatory genomics and protein classification [33, 34] (Fig. 3). Advanced DL-based models represent the state-of-the-art of prediction accuracy in biological sciences [35, 36]. Nevertheless, they require the availability of very large-scale training data (with an associated high computational demand), and their interpretation remains elusive (they are often referred to as "black-box models"), with this elusiveness representing a limitation in biological experiments involving omics data for which identifying the most important predicting features and feature combinations is of primary importance [37]. Thus, when research is aimed at better understanding the functioning of biological systems, DL-based models are still difficult to be commonly applied [38, 39]. It is also for these reasons that in a society where AI algorithms are becoming more central than ever before in all aspects of our daily life, the concepts of "interpretable ML" and "explainable AI" are gaining an always increasing attention and importance [34, 40].

Multiomics integration for ML analysis

As mentioned above, innovations in high-throughput acquisition of different omics data from single experiments are now enabling capturing different layers of biological complexity. In fact, application of omics approaches, such as transcriptomic, proteomics, and metabolomics [41], to large diversity panels and/or samples holds significant promise for unraveling the complexity of living systems. Despite their overall potential for discovery, the diverse nature of omics data acquired by different technological platforms requires the use of integration strategies to effectively harness their complementary information. Recent advances in multiomics analysis have been made possible by the development of various tools and methods that can resolve the heterogeneous nature of biological datasets, enabling their effective integration. Notably, consensus orthogonal partial least squares discriminant analysis (OPLS-DA) has emerged as an effective strategy for fusing multiomics data, combining multiple kernel learning with OPLS-DA [42]. The *mixOmics* R package provides a variety of multivariate methods for integrating omics datasets, including extensions of "Projection to Latent Structure" models for discriminant analysis and molecular signature identification [43]. Additionally, ML techniques, such as network-based diffusion and DL, are increasingly used to capture complex nonlinear associations in multiomics data [44]. Among the available R resources, packages such *moiraine* [45] provide a range of integrative methods for multiomics analyses, including sPLS and DIABLO from the *mixOmics* package [43], sO2PLS from the *OmicsPLS* package [46], and MOFA and MEFISTO from the *MOFA2* package [47].

AI-based analysis of omics data in the fields of plant sciences, animal sciences, and microbial sciences

International initiatives are thriving in the field of AI-based analysis of omics data, aiming to advance the discovery of genotype–phenotype relationships. One such example is the *GLOMICAVE* project (GLobal OMIC data integration on Animal, Vegetal and Environment sectors), an international project that involves all the authors of this review paper. *GLOMICAVE* has created an innovative digital platform that connects genotype to phenotype through Big Data analytics and AI, using extensive public and experimental omic datasets [https://glomicave.eu/]. Likewise, cloud-based platforms like HiOmics offer a comprehensive analysis of biomedical large-scale omics data [48]. Such projects aim to facilitate the analysis of primary data and support large-scale omics experiments, thereby enhancing the utility of omics data on a massive scale and deepening our understanding of entire biological systems. In line with *GLOMICAVE*, and considering that the medical field has been extensively examined from an AI perspective, this review focuses on relevant applications from plant, animal, and microbial sciences.

Plant sciences and AI

The explosion of omics has radically transformed research in plant sciences, simultaneously driving the need for ML to handle datasets characterized by high complexity and dimensionality. A paradigmatic example is plant phenomics, which has rapidly shifted from a promising research sector with the potential of bridging the gap with genomic advances to becoming a widespread tool in plant and crop sciences [12]. This rapid progress was enabled by integrating advanced sensors and imaging technologies (e.g., RGB, multispectral, hyperspectral, thermal, and fluorescence cameras and sensors) with unmanned aerial vehicles (or drones) and ground robots, which are able to collect high-throughput phenotyping data. Approaches based on ML algorithms are now a practical and effective strategy for extracting traits and features from massive amounts of imaging- and sensor-based data. DL algorithms (e.g., convolutional neural networks [CNNs]) show the highest versatility and success in image-based plant phenotyping. These algorithms are particularly effective in predicting the effects of biotic and abiotic stresses [49, 50] and enabling rapid and accurate diagnostics of plant diseases [51]. Additionally, AI applications in root system architecture image analysis are emerging as crucial tools for improving this understudied field of research, which holds significant potential to boost a "Second Green Revolution" in agriculture [52]. Plant breeding is another branch of plant science that has been radically transformed by genomic advances, with breeders increasingly relying on genome-wide SNP marker-based genomic prediction (GP) to accelerate genetic gains for target traits in crops. Classic GP models are based on best linear unbiased prediction, but efforts to develop new ML-based and improved GP algorithms are ongoing [53]. Furthermore, different sources of nongenetic variability and nonadditive modes of gene action have made the choice and implementation of GP models challenging for improving complex plant traits, such as biomass and crop yield [54]. One possible solution to this problem is to incorporate other genome-to-phenome intermediate omics data (e.g., transcriptomics, proteomics, metabolomics) into the GP models to enhance their accuracy and predictive power [55]. The potential of ML models based on single intermediate omics, particularly metabolomics, the omics layer closer to the phenotype, has been demonstrated for the accurate prediction of crop yield, notably in maize [56] and rice [57, 58]. However, for plant breeding applications, the integration of large, highly dimensional, and "noisy" omics datasets for complex trait prediction remains a challenging field of study. This challenge will require the use of ML/DL techniques, leveraging their superior capability for Big Data analytics to effectively handle the complexity and scale of these datasets [59]. Interestingly, recent studies have highlighted innovative approaches in metabolomics-based ML prediction of plant complex traits showing innovative routes to identify breeding targets for plant improvement. For example, Colantonio et al. [60] identified candidate metabolites acting as fruit flavor enhancers and suppressors by metabolomics-based ML prediction of tomato and blueberry fruit flavor profiles. In efforts to improve plant tolerance to abiotic stress, Dussarrat et al. [61] applied a holistic ML prediction approach on environmental adaptations based on the multispecies metabolome of plants collected in the Atacama desert. This revealed a core set of metabolite targets for extreme climate resilience (sugars, stress-related amino acids, hormones, and antioxidants, including phenolics and major redox buffers).

Animal sciences and AI

Modern biotechnologies, bio-sensing hardware, and IT infrastructure have led to a high-throughput data collection era in livestock management, driving the need for faster and more efficient computational methods. While traditional information sources in animal breeding included phenotype and pedigree data, the field is increasingly incorporating genomic data such as SNPs, gene annotations, metabolic pathways, protein interaction networks, gene expression, and protein structure information. These data can enhance trait predictions and improve our understanding of the underlying biological phenotypes [62]. Despite these advancements in animal genetics, many challenges still persist. The widespread adoption of omics technologies is hindered by high cost and the need for expertise across diverse fields. Accurate recording of phenotypic data and population/sample size are other constraints that need to be addressed. However, omics technologies have shown their potential to identify superior and disease-resistant animals at an early stage [63]. For example, the metabolomes of healthy and unhealthy chickens were characterized and compared using untargeted mass spectrometry metabolomics [64]. Researchers were able to accurately distinguish chicken health status in multiple countries using a random forest (RF)–based ML model. This approach used raw mass spectrometry signals (unannotated m/z values) as input features, effectively overcoming one of the primary limitations of untargeted metabolomics: the need for metabolite annotation and identification. The use of ML models in animal breeding has recently attracted interest due to their exceptional flexibility and ability to capture patterns in large, noisy datasets [65]. For instance, gradient tree boosting (GTB) has proven to be an effective ML algorithm for predicting different breeding values. GTB-based models have identified a subset of genes contributing to feed efficiency in growing pigs using muscle transcriptome data [66]. The potential of combining metagenomics, metatranscriptomics, and metabolomics data was evaluated in rumen content, demonstrating their value as predictive markers for feed efficiency and their potential applications for selecting cows with high feed efficiency [67]. By using a RF-based model, they were able to predict feed efficiency using a preselected set of metabolites associated with this trait. Antimicrobial-resistant microorganisms pose significant challenges in livestock farming. A recent study [68] evaluated 10 supervised learning classifiers to predict *Escherichia coli* strains susceptible or resistant to 26 different antimicrobials using whole-genome shotgun sequencing in intensive poultry farming. This findings provided evidence of transmissible drug resistance in food-producing animals, which has contributed to the emergence of drug resistance in zoonotic pathogens.

Microbial sciences and AI

Microorganisms exist naturally in microbial communities and establish multiple interactions between each other and with their hosts. Omics experiments play a crucial role in studying these microorganisms in their natural environments, eliminating the need for their isolation and cultivation. However, interpreting omics information and integrating results from different studies remains challenging due to the complexity of omics data. AI has been increasingly applied to help interpret the variations found in microbial communities, particularly in the human microbiome and its relationship to health and disease [69–71]. In the field of environmental microbiology, recent reviews have highlighted major developments in the application of ML to microbial ecology omics [72]. This approach has been primarily applied to omics experiments using 16S rRNA gene sequencing data, which provide taxonomic information on microbial communities. RF-based ML architecture has been widely used due to its ease of implementation, interpretation, low cost, and the requirement of less data compared with DL [72]. Nevertheless, other ML algorithms, such as naive Bayes (NB), SVM, and KNN, have also been applied in the microbiology field [73]. In microbial ecology, the main objective of ML has been to predict the presence of certain microbes (e.g., microbial bioindicators, predicting environmental pollution, and key microbes affecting the performance of biotechnological processes), as well as to predict microbe-microbe and microbe–host interactions and facilitate data mining [72, 73]. For example, in the particular case of anaerobic digestion microbiology (a biotechnological process in which organic waste is converted to methane by microbial communities), there are several studies on AI applied to omics data. Three different algorithms (i.e., linear regression, SVM, and RF regression) were used to predict the production of medium-chain carboxylates, based on microbial community dynamics (16S rRNA) and the bioreactor's productivity data. This study concluded that RF regression was the most effective algorithm for this task [74]. Similarly, another study compared 6 different ML algorithms—namely, GLMNET, RF, NNET, KNN, SVM, and extreme gradient boosting (XGBOOST)—to predict the performance of the anaerobic digestion process, using 16S rRNA genomics as the basis for the analysis [75]. Interactions between microorganisms are highly important and influence the activity of microbial communities. Syntrophic interactions among different species are key examples of microbial interactions, where microbes exchange electrons either via soluble molecules or directly from cell to cell in an interdependent way. ML was recently used to predict the type of syntrophic interaction that prevails in microbial communities by using a Bayesian network approach [76]. This analysis incorporated not only 16S rRNA sequencing but also metagenomics and metatranscriptomics data.

Navigating the frontier: challenges and future horizons in AI innovation

AI in biology research faces several major obstacles that must be addressed through close collaboration between biologists and computer scientists. Such interdisciplinary collaborations are essential to exploit the full potential of AI in life sciences [77].

Tackling technical challenges in AI-based research

A summary of topics that represent challenges in AI-based research is provided in Table 1. For each topic, the description and its connection to ML and/or DL are highlighted. Additionally, the topics are characterized based on 7 main technical challenges: (i) noisy datasets, (ii) high dimensionality, (iii) omics data integration, (iv) interpretability, (v) computational requirements, (vi) FAIR (Findable, Accessible, Interoperable, Reusable) principles, and (vii) data size and diversity.

Importantly, data curation and integration across biological subdisciplines continue to pose significant challenges, requiring the development of new theories and predictive models tailored to biology [77]. A significant problem is the lack of standardized formats across different biological disciplines, which not only complicates the handling of file formats [78] but also makes it difficult to interpret data generated by specialists of each omics data type. Ethical concerns, particularly for animal sciences, and privacy issues surrounding data usage need to be addressed, along with ensuring the reliability and safety of AI models through robust validation and transparency. The explainability of AI methods in biological data science remains a significant challenge, as many current approaches lack interpretability. This can lead to decreased trustworthiness and reliability in decision-making processes. Moreover, improving the interpretability of ML-based models in life science is crucial, as it allows a better understanding of the biological mechanisms behind the models (e.g., by helping to identify important biomarkers, biological pathways, or features that contribute to a specific process) [79].

The scarcity of labeled data for training AI models

Labeling large amounts of data has become one of the main bottleneck in the development of AI systems [80]. Over the past 15 years, advanced ML models, particularly those based on deep neural networks (DNNs), have enabled unprecedented results in a variety of fields, including omics research in life sciences [81]. However, these models require vast amounts of labeled training data, which in many practical scenarios are either unavailable or very arduous to obtain [82, 83]. Creating hand-labeled training datasets is expensive and time-consuming, often taking months or years to develop, particularly when domain expertise is required. In response to this technical challenge, a subfield of ML know as *weakly supervised learning*, a concept developed back in the 1960s, has evolved into an approach capable of generating large training datasets more rapidly. These datasets, though noisier and of lower quality, are constructed via strategies such as using cheaper annotators, programmatic scripts, or more creative and high-level input from domain experts. In principle, these techniques offer higher-level or less precise forms of supervision, which, while less accurate, are faster and easier to obtain than manual annotation [84]. Another approach motivated by the same goal is *semi-supervised learning*, which strives to create large training datasets by combining a small amount of labeled data with a much larger amount of unlabeled data [85]. Omics-based research in life sciences has quickly adopted solutions derived from these approaches across various applications, such as molecular path-

Table 1: Major technical challenges in AI-based research

Technical challenge	Description	Connection to ML and DL
1. Noisy datasets		
Impact on model performance	Noisy or erroneous data can degrade AI model performance, leading to inaccurate predictions, especially in high-precision fields like life sciences.	**ML**: Often struggles with noisy data unless advanced preprocessing is applied. **DL**: Sensitive to noise, impacting performance.
Data cleaning	Effective noise reduction and robust data cleaning are essential but challenging, particularly at large scales.	**ML**: Requires preprocessing techniques to handle noisy data. **DL**: Needs data cleaning to improve model accuracy.
2. High dimensionality		
Curse of dimensionality	High-dimensional data can lead to overfitting, making models perform well on training data but poorly on unseen data.	**ML**: Can overfit if dimensionality is not managed; requires feature selection. **DL**: Needs strategies to handle high dimensions.
Feature selection	Identifying relevant features from a large number of variables is complex and requires advanced techniques to prevent redundancy and enhance model efficiency.	**ML**: Involves sophisticated techniques for effective feature selection. **DL**: Uses embedded feature selection or reduction techniques.
3. Omics data integration		
Heterogeneity	Omics data from various sources (e.g., genomics, proteomics) are often heterogeneous, differing in scale, format, and noise, complicating integration.	**ML**: Requires methods to handle heterogeneous data. **DL**: Needs effective data fusion strategies for multiomics.
Data fusion	Developing methods for effective multiomics data fusion that preserves biological context and relationships is an ongoing challenge.	**ML**: Must integrate diverse data types. **DL**: Benefits from advanced fusion techniques for comprehensive analysis.
4. Interpretability of results		
Complex models	Deep learning models, especially those with complex architectures, can act as "black boxes," making it hard to interpret how conclusions are reached.	**ML**: Generally more interpretable than DL but still faces challenges. **DL**: Requires explainability techniques for transparency.
Explainability techniques	Emerging techniques like sHapley additive exPlanations (SHAP) or local intrepretable model-agnostic explanations (LIME) offer ways to explain AI decisions but may not always provide comprehensive or intuitive insights.	**ML**: May utilize various explainability methods. **DL**: Needs specific techniques for understanding model behavior.
5. Computational requirements		
Resource intensity	Training state-of-the-art AI models, particularly deep learning models, requires significant computational resources, including high-performance GPUs and extensive memory.	**ML**: Generally less resource-intensive but can still require significant computational power. **DL**: Highly resource-demanding.
Scalability	Ensuring algorithms scale efficiently with increasing data sizes and complexity without excessive computational costs is a critical challenge.	**ML**: Needs to manage scalability efficiently. **DL**: Must handle large-scale data and complex models effectively.
6. Importance of FAIR principles		
Findable, Accessible, Interoperable, Reusable (FAIR)	Adhering to FAIR principles for data and scripts is essential for reproducibility and collaboration but challenging, particularly in standardizing metadata and documentation.	**ML**: Requires well-documented datasets for reproducibility. **DL**: Benefits from FAIR practices for consistent data use.
Data sharing	Facilitating access to well-documented, standardized datasets while maintaining privacy and security can be complex.	**ML**: Needs secure and standardized data-sharing practices. **DL**: Requires access to high-quality, FAIR-compliant datasets.
7. Data size and diversity		
Scalability of models	Handling and processing large-scale datasets requires models that can manage and learn from vast amounts of data without compromising performance.	**ML**: Must be scalable to handle large data. **DL**: Efficiently manages large datasets but with high computational costs.
Bias and generalization	Ensuring data diversity to avoid biases and ensure models generalize well across different populations or conditions is crucial. Imbalanced datasets can lead to skewed results.	**ML**: Needs diverse data to prevent bias. **DL**: Requires careful data handling to ensure generalization across conditions.

ways status prediction in cancer [86] or protein-DNA binding prediction [87], and in the field of plant sciences (applications specific to plant and field phenomics) [88–91]. These examples provide evidence of the effectiveness of *weakly* and *semi-supervised learning* when applied to omics science and indicate a promising future direction.

AI for the prediction and annotation of metabolites

Recent developments in AI-based metabolite annotation reflects significant advancements in the application of ML and DL techniques to improve the accuracy and efficiency of metabolite identification and characterization in mass spectrometry–based studies [92]. As an example, the chemical language model "DeepMet" utilizes CNNs to learn features from raw MS/MS spectral data and predict human metabolite identities [93]. Similarly, the "MetFID" model uses ANNs to predict molecular fingerprints from MS/MS data, enhancing annotation accuracy compared to existing tools [94]. Computational annotation strategies, including peak grouping, ion adduction analysis, and incorporation of biological knowledge, help overcome the limitations of accurate mass searching alone [95]. ML-based approaches and molecular networking have shown promise in large-scale metabolite annotation, particularly in natural product discovery [96]. Another compelling ML-based tool includes the "PeakDecoder" algorithm, which enables metabolite annotation and accurate profiling in multidimensional mass spectrometry measurements [97]. Despite the availability of ML-based tools for metabolite annotation, inconsistencies in their benchmarking hinder users from selecting the most appropriate method for their research, highlighting the need for standardized evaluation practices [96].

In the context of ecosystem metabolomics, computational methods can now predict previously unobserved metabolites in new microbial communities by leveraging paired metabolome and metagenome data, achieving over 50% accuracy for related metabolites [98]. Additionally, knowledge-based and ML-driven approaches are being developed to refine metabolite identification and analyze primary microbial metabolism in mixed samples [99]. This demonstrates that predictive metabolomics can aid experimental design and reveal valuable insights into numerous community profiles where only metagenomic data are available.

AI-based gene annotation

Advances in genomics have been largely driven by the increasing throughput and lower cost of DNA sequencing. This has made it possible to sequence thousands of individual genomes within a species and a large number of new species. While generating sequencing data has become a relatively straightforward task, the subsequent processing steps to produce a genome assembly with structural annotations of genomic elements (e.g., genes, promoters, and regulatory elements) and gene functional annotations still represent a challenge. Long-read sequencing technologies have alleviated some of these issues, particularly for genome assembly, but the structural annotation of genes, especially in novel genomes, remains problematic in the absence of other extrinsic data sources. Well-known structural annotation tools, such as AUGUSTUS [100], use hidden Markov models (HMMs) for intrinsic *ab initio* gene finding. A recent *ab initio* gene calling tool, Helixer [101], uses DNNs combined with HMMs to identify genes in genomes without the need for extrinsic data and has shown promising results. Gene functional annotation has traditionally relied on homology to characterize proteins for ascribing a function to newly identified genes. The bottleneck of this methodology is mainly due to knowledge gaps that are producing annotation of genes of "unknown function." DeepGO [102] is a tool that employs DL methods and interactive networks to annotate protein sequences with Gene Ontology (GO) terms. A later improvement, DeepGOPlus [103] removed many of the restrictions of the earlier version and no longer needs the interaction networks. DeepGOPlus has the additional advantage of being species agnostic and gives equally good results from protein sequences derived from genomes of newly sequenced species and clades.

FAIR practices for omics data and AI

Despite the advances outlined above, challenges in standardizing methods and interpreting results persist, highlighting the need for FAIR practices and proper benchmarking to ensure reproducibility and reliability in multiomics and AI research. In this context, ontologies play a crucial role by tagging datasets with metadata, thereby improving data understanding and interoperability [104]. They define domain-specific concepts and relationships, making data both human- and machine-readable for easier reuse. However, identifying relevant ontologies can be difficult due to the large amount available. For example, as of September 2024, 1,147 different ontologies are available in BioPortal [105], including 24 specific for plants and 37 for animal science. As ML becomes increasingly indispensable, ensuring data privacy, algorithmic fairness, and transparency will be paramount for maintaining public trust and ensuring equitable access to the benefits of ML-driven advancements [106]. Additionally, many open data sources in the life sciences are not yet fully FAIR-compliant, with issues related to the absence of proper metadata, inadequate data documentation, and the lack of crosslinking between datasets. This requires significant effort to upgrade their FAIRness for integration into semantic web platforms [107]. While the FAIR principles aim to enhance machine readability and processing of scientific data, concerns have been raised about potential epistemic losses, such as a reduction in semantic freedom and the displacement of human expertise, which could discourage trust in AI [108]. To address skepticism and foster trust among stakeholders, a more balanced discussion of both the benefits and epistemic costs of implementing FAIR is needed. Remarkably, a systematic review of 124 LC/MS metabolomics software that subsequently retained 61 for detailed analysis based on FAIR Principles for Research Software (FAIR4RS) criteria reported that software fulfillment of these criteria ranged from 21.6% to 71.8%, with no significant improvement in FAIRness over time [109]. Key issues identified included the lack of semantic annotation (0%, i.e., no software had semantic annotation of key information), low registration on Zenodo with DOIs (6.3%), limited containerization of code or use of virtual machines (14.5%), and insufficiently documented functions in code (16.7%). This recent work highlights clear caveats that need to be addressed in further Big Data–based life science research. To further advance the FAIR principles, collaboration between researchers, data scientists, and data managers is more than ever needed.

Concluding remarks

In conclusion, AI has already transformed biomedical research by accelerating drug discovery, enhancing clinical trials, and providing powerful tools for analyzing complex biological data [110]. Its ability to optimize processes, reduce costs, and increase precision is revolutionizing how researchers approach biological challenges. The 2020s is the decade of AI applied to biology: as AI continues to advance, its impact on animal, plant, and environmental research will be paramount. AI is reshaping animal research by enhancing data analysis, improving animal welfare, and reducing reliance on traditional testing methods. Through predictive modeling, AI helps refine experimental designs, minimizing the number of animals used while increasing the accuracy of results. It also supports the monitoring of animal behavior and health, contributing to better care and more ethical practices. The growing role of AI in animal research will likely lead to more humane, efficient, and scientifically robust studies. Additionally, the evolution of ML in plant biology, ranging from its early explorations to its current prominence as a transformative tool, demonstrates its remarkable potential. As ML continues to advance, its integration with other AI techniques, real-time data processing, and ethical considerations, including agroecological transitions, will shape the future of plant biology research and agricultural practices. In a wider context, AI is making significant strides in environmental research by providing sophisticated tools for monitoring ecosystems, predicting climate patterns, and analyzing environmental data. Its ability to process vast amounts of information and identify complex patterns helps in understanding and mitigating the impacts of climate change, pollution, and habitat loss. AI promises to enhance our capacity for environmental stewardship, driving more effective and data-driven strategies to protect and sustain our planet.

Abbreviations

AI: artificial intelligence; ANN: artificial neural network; cDNA: complementary DNA; CNN: convolutional neural network; DL: deep learning; DNN: deep neural network; ESI: electrospray ionization; GC: gas chromatography; GO: Gene Ontology; GP: genomic prediction; GTB: gradient tree boosting; HMM: hidden Markov model; HR: high resolution; LC: liquid chromatography; LIME: local intrepretable model-agnostic explanations; MALDI: matrix-assisted laser desorption ionization; ML: machine learning; MS: mass spectrometry; NGS: next-generation sequencing; OPLS-DA: orthogonal partial least squares discriminant analysis; RF: random forest; RNA-seq: RNA sequencing; SHAP: sHapley additive exPlanations; SNP: single nucleotide polymorphic; SVM: support vector machine.

Funding

The authors are grateful for financial support from the European Commission's Horizon 2020 Research and Innovation program via the GLOMICAVE (grant agreement no. 952908), MetaboHUB infrastructure funded by the Agence Nationale de la Recherche under the France 2030 program (MetaboHUB ANR-11-INBS-0010; MetEx+ ANR-21-ESRE-0035; MetaboHUB (JVCE) ANR-24-INBS-0012), and PHENOME infrastructure (ANR-11-INBS-0012) projects. X.D. was supported by "La Caixa" Foundation (ID 100010434) via the Junior Leader Fellowship LCF/BQ/PR21/11840001.

Data Availability

No data are associated with this article.

Competing Interests

The authors declare that they have no competing interests.

References

1. Stephens ZD, Lee SY, Faghri F, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13:e1002195. https://doi.org/10.1371/journal.pbio.1002195.
2. Giani AM, Gallo GR, Gianfranceschi L, et al. Long walk to genomics: history and current approaches to genome sequencing and assembly. Comput Struct Biotechnol J. 2020;18:9–19. https://doi.org/10.1016/j.csbj.2019.11.002.
3. Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63. https://doi.org/10.1038/nrg2484.
4. Lowe R, Shirley N, Bleackley M, et al. Transcriptomics technologies. PLoS Comput Biol. 2017;13:e1005457. https://doi.org/10.1371/journal.pcbi.1005457.
5. Amarasinghe SL, Su S, Dong X, et al. Opportunities and challenges in long-read sequencing data analysis—genome biology—full text. Genome Biol. 2020;21:1–16. https://doi.org/10.1186/s13059-020-1935-5.
6. Marx V. Method of the year: long-read sequencing. Nat Methods. 2023;20:6–11. https://doi.org/10.1038/s41592-022-01730-w.
7. Griffiths J. A brief history of mass spectrometry. Anal Chem. 2008;80:5678–83. https://doi.org/10.1021/ac8013065.
8. McLafferty FW. A century of progress in molecular mass spectrometry. Annu Rev Anal Chem. 2011;4:1–22. https://doi.org/10.1146/annurev-anchem-061010-114018.
9. Mann M, Kelleher NL. Precision proteomics: the case for high resolution and high mass accuracy. Proc Natl Acad Sci USA. 2008;105:18132–38. https://doi.org/10.1073/pnas.0800788105.
10. Alseekh S, Fernie AR. Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J. 2018;94:933–42. https://doi.org/10.1111/tpj.13950.
11. Hussain S, Mubeen I, Ullah N, et al. Modern diagnostic imaging technique applications and risk factors in the medical field: a review. Biomed Res Int. 2022;2022:Na. https://doi.org/10.1155/2022/5164970.
12. Yang W, Feng H, Zhang X, et al. Crop phenomics and high-throughput phenotyping: past decades, current challenges, and future perspectives. Mol Plant. 2020;13:187–214. https://doi.org/10.1016/j.molp.2020.01.008.
13. Joyce AR, Palsson B. The model organism as a system: integrating "omics" data sets. Nat Rev Mol Cell Biol. 2006;7:198–210. https://doi.org/10.1038/nrm1857.
14. Picard M, Scott-Boyer MP, Bodein A, et al. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46. https://doi.org/10.1016/j.csbj.2021.06.030.
15. Wang P. On defining artificial intelligence. J Artif Gen Intell. 2019;10:1–37. https://doi.org/10.2478/jagi-2019-0002.
16. Samoili S, López Cobo M, Gómez E, et al. AI watch: defining artificial intelligence: towards an operational definition and taxonomy of artificial intelligence. Luxembourg: Publications Office of the European Union, 2020. https://dx.doi.org/10.2760/382730.
17. Kaplan A, Haenlein M. Siri, Siri, in my hand: who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Bus Horiz. 2019;62:15–25. https://doi.org/10.1016/j.bushor.2018.08.004.
18. Murdoch WJ, Singh C, Kumbier K, et al. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci USA. 2019;116:22071–80. https://doi.org/10.1073/pnas.1900654116.
19. Li R, Li L, Xu Y, et al. Machine learning meets omics: applications and perspectives. Briefings Bioinf. 2022;23:1–22. https://doi.org/10.1093/bib/bbab460.
20. Sohail A, Arif F. Supervised and unsupervised algorithms for bioinformatics and data science. Prog Biophys Mol Biol. 2020;151:14–22. https://doi.org/10.1016/j.pbiomolbio.2019.11.012.

21. Domingos P. The master algorithm: how the quest for the ultimate learning machine will remake our world. New York: Basic Books, 2015. ISBN: 0465065708
22. van Dijk ADJ, Kootstra G, Kruijer W, et al. Machine learning in plant science and plant breeding. iScience. 2021;24:101890. https://doi.org/10.1016/j.isci.2020.101890.
23. Schneider A, Hommel G, Blettner M. Linear regression analysis. Dtsch Arztebl Int. 2010;107:776–82. https://doi.org/10.3238/arztebl.2010.0776.
24. Swindel BF. Geometry of ridge regression illustrated. Am Stat. 1981;35:12–15. https://doi.org/10.1080/00031305.1981.10479296.
25. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge, MA: The MIT Press, 2016. https://www.deeplearningbook.org/.
26. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97. https://doi.org/10.1007/BF00994018.
27. Mairal J, Koniusz P, Harchaoui Z, et al. Convolutional kernel networks. arXiv, 2014. https://doi.org/10.48550/arXiv.1406.3332. Accessed 2 April 2025.
28. Breiman L, Friedman JH, Olshen RA, et al. Classification and regression trees. Cole Statistics/Probability Series. Monterey, CA: Wadsworth and Brooks; 1984.
29. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;29:1189–232. https://doi.org/10.1214/aos/1013203451.
30. Pearl J. Probabilistic Reasoning in intelligent systems. Elsevier, 1988. https://doi.org/10.1016/C2009-0-27609-4.
31. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18:1527–54. https://doi.org/10.1162/neco.2006.18.7.1527.
32. Lecun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44. https://doi.org/10.1038/nature14539.
33. Senior AW, Evans R, Jumper J, et al. Improved protein structure prediction using potentials from deep learning. Nature. 2020;577:706–10. https://doi.org/10.1038/s41586-019-1923-7.
34. Novakovsky G, Dexter N, Libbrecht MW, et al. Obtaining genetics insights from deep learning via explainable artificial intelligence. Nat Rev Genet. 2023;24:125–37. https://doi.org/10.1038/s41576-022-00532-2.
35. Mahmud M, Kaiser MS, McGinnity TM, et al. Deep learning in mining biological data. Cogn Comput. 2021;13:1–33. https://doi.org/10.1007/s12559-020-09773-x.
36. Sapoval N, Aghazadeh A, Nute MG, et al. Current progress and open challenges for applying deep learning across the biosciences. Nat Commun. 2022;13:1728. https://doi.org/10.1038/s41467-022-29268-7.
37. Ching T, Himmelstein DS, Beaulieu-Jones BK, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141):Na. https://doi.org/10.1098/rsif.2017.0387.
38. Greener JG, Kandathil SM, Moffat L, et al. A guide to machine learning for biologists. Nat Rev Mol Cell Biol. 2022;23:40–55. https://doi.org/10.1038/s41580-021-00407-0.
39. Xu C, Jackson SA. Machine learning and complex biological data the revolution of biological techniques and demands for new data mining methods. Genome Biol. 2019;20:1–42019. https://doi.org/10.1186/s13059-019-1689-0.
40. Adadi A, Berrada M. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access. 2018;6:52138–60. https://doi.org/10.1109/ACCESS.2018.2870052.
41. Hajjar G, Barros Santos MC, Bertrand-Michel J, et al. Scaling-up metabolomics: current state and perspectives. Trends Anal Chem. 2023;167:117225. https://doi.org/10.1016/j.trac.2023.117225.
42. Boccard J, Rutledge DN. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock omics data fusion. Anal Chim Acta. 2013;769:30–39. https://doi.org/10.1016/j.aca.2013.01.022.
43. Rohart F, Gautier B, Singh A, et al. mixOmics: an R package for ‘omics feature selection and multiple data integration. PLoS Comput Biol. 2017;13:e1005752. https://doi.org/10.1371/journal.pcbi.1005752.
44. Cominetti O, Agarwal S, Oller-Moreno S. Editorial: advances in methods and tools for multi-omics data analysis. Front Mol Biosci. 2023;10:1–2. https://doi.org/10.3389/fmolb.2023.1186822.
45. Angelin-Bonnet O. moiraine: Construction of reproducible pipelines for testing and comparing multi-omics integration tools. R package version 1.0.0.9000, https://plant-food-research-open.github.io/moiraine/. GitHub, 2025. https://github.com/Plant-Food-Research-Open/moiraine. Accessed 2 April 2025.
46. el Bouhaddani S, Uh HW, Jongbloed G, et al. Integrating omics datasets with the OmicsPLS package. BMC Bioinf. 2018;19:1–9. https://doi.org/10.1186/s12859-018-2371-3.
47. Argelaguet R, Velten B, Arnol D, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. Mol Syst Biol. 2018;14:1–13. https://doi.org/10.15252/msb.20178124.
48. Li W, Zhang Z, Xie B, et al. HiOmics: a cloud-based one-stop platform for the comprehensive analysis of large-scale omics data. Comput Struct Biotechnol J. 2024;23:659–68. https://doi.org/10.1016/j.csbj.2024.01.002.
49. Singh AK, Ganapathysubramanian B, Sarkar S, et al. Deep learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci. 2018;23:883–98. https://doi.org/10.1016/j.tplants.2018.07.004.
50. Islam S, Reza MN, Samsuzzaman S, et al. Machine vision and artificial intelligence for plant growth stress detection and monitoring: a review. Precis Agric Sci Technol. 2024;6:33–57. https://doi.org/10.12972/pastj.20240003.
51. Natarajan S, Chakrabarti P, Margala M. Robust diagnosis and meta visualizations of plant diseases through deep neural architecture with explainable AI. Sci Rep. 2024;14:1–14. https://doi.org/10.1038/s41598-024-64601-8.
52. Weihs BJ, Heuschele DJ, Tang Z, et al. The state of the art in root system architecture image analysis using artificial intelligence: a review. Plant Phenomics. 2024;6:0178. https://doi.org/10.34133/plantphenomics.0178.
53. Azodi CB, Bolger E, McCarren A, et al. Benchmarking parametric and machine learning models for genomic prediction of complex traits. G3 (Bethesda). 2019;9:3691–702. https://doi.org/10.1534/g3.119.400498.
54. Rice BR, Lipka AE. Diversifying maize genomic selection models. Mol Breeding. 2021;41:Na. https://doi.org/10.1007/s11032-021-01221-4.
55. Tong H, Nikoloski Z. Machine learning approaches for crop improvement: leveraging phenotypic and genotypic big data. J Plant Physiol. 2021;257:153354. https://doi.org/10.1016/j.jplph.2020.153354.
56. Riedelsheimer C, Czedik-Eysenberg A, Grieder C, et al. Genomic and metabolic prediction of complex heterotic traits in hybrid maize. Nat Genet. 2012;44:217–20. https://doi.org/10.1038/ng.1033.

57. Xu S, Xu Y, Gong L, et al. Metabolomic prediction of yield in hybrid rice. Plant J. 2016;88:219–27. https://doi.org/10.1111/tpj.13242.
58. Melandri G, Monteverde E, Riewe D, et al. Can biochemical traits bridge the gap between genomics and plant performance? A study in rice under drought. Plant Physiol. 2022;189:1139–52. https://doi.org/10.1093/plphys/kiac053.
59. Yan J, Wang X. Machine learning bridges omics sciences and plant breeding. Trends Plant Sci. 2023;28:199–210. https://doi.org/10.1016/j.tplants.2022.08.018.
60. Colantonio V, Ferrão LFV, Tieman DM, et al. Metabolomic selection for enhanced fruit flavor. Proc Natl Acad Sci USA. 2022;119:1614–28. https://doi.org/10.1073/pnas.2115865119.
61. Dussarrat T, Prigent S, Latorre C, et al. Predictive metabolomics of multiple Atacama plant species unveils a core set of generic metabolites for extreme climate resilience. New Phytol. 2022;234:1614–28. https://doi.org/10.1111/nph.18095.
62. Nayeri S, Sargolzaei M, Tulpan D. A review of traditional and machine learning methods applied to animal breeding. Anim Health Res Rev. 2019;20:31–46. https://doi.org/10.1017/S1466252319000148.
63. Chakraborty D, Sharma N, Kour S, et al. Applications of omics technology for livestock selection and improvement. Front Genet. 2022;13:1–16. https://doi.org/10.3389/fgene.2022.774113.
64. Wolthuis JC, Magnúsdóttir S, Stigter E, et al. Multi-country metabolic signature discovery for chicken health classification. Metabolomics. 2023;19:1–14. https://doi.org/10.1007/s11306-023-01973-4.
65. Chafai N, Hayah I, Houaga I, et al. A review of machine learning models applied to genomic prediction in animal breeding. Front Genet. 2023;14:1–18. https://doi.org/10.3389/fgene.2023.1150596.
66. Messad F, Louveau I, Koffi B, et al. Investigation of muscle transcriptomes using gradient boosting machine learning identifies molecular predictors of feed efficiency in growing pigs. BMC Genomics. 2019;20:1–14. https://doi.org/10.1186/s12864-019-6010-9.
67. Xue MY, Xie YY, Zhong Y, et al. Integrated meta-omics reveals new ruminal microbial features associated with feed efficiency in dairy cattle. Microbiome. 2022;10:1–14. https://doi.org/10.1186/s40168-022-01228-9.
68. Peng Z, Maciel-Guerra A, Baker M, et al. Whole-genome sequencing and gene sharing network analysis powered by machine learning identifies antibiotic resistance sharing between animals, humans and environment in livestock farming. PLoS Comput Biol. 2022;18:e1010018. https://doi.org/10.1371/journal.pcbi.1010018.
69. Pasolli E, Truong DT, Malik F, et al. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. PLoS Comput Biol. 2016;12:e1004977. https://doi.org/10.1371/journal.pcbi.1004977.
70. Topçuoğlu BD, Lesniak NA, Ruffin MT, et al. A framework for effective application of machine learning to microbiome-based classification problems. mBio. 2020;11. https://doi.org/10.1128/mBio.00434-20.
71. Krause T, Wassan JT, Mc Kevitt P, et al. Analyzing large microbiome datasets using machine learning and big data. BioMedInformatics. 2021;1:138–65. https://doi.org/10.3390/biomedinformatics1030010.
72. McElhinney JMWR, Catacutan MK, Mawart A, et al. Interfacing machine learning and microbial omics: a promising means to address environmental challenges. Front Microbiol. 2022;13:Na. https://doi.org/10.3389/fmicb.2022.851450.
73. Qu K, Guo F, Liu X, et al. Application of machine learning in microbiology. Front Microbiol. 2019;10:827. https://doi.org/10.3389/fmicb.2019.00827.
74. Liu B, Sträuber H, Saraiva J, et al. Machine learning-assisted identification of bioindicators predicts medium-chain carboxylate production performance of an anaerobic mixed culture. Microbiome. 2022;10:48. https://doi.org/10.1186/s40168-021-01219-2.
75. Long F, Wang L, Cai W, et al. Predicting the performance of anaerobic digestion using machine learning algorithms and genomic data. Water Res. 2021;199:117182. https://doi.org/10.1016/j.watres.2021.117182.
76. Yuan H, Wang X, Lin TY, et al. Disentangling the syntrophic electron transfer mechanisms of Candidatus geobacter eutrophica through electrochemical stimulation and machine learning. Sci Rep. 2021;11:15140. https://doi.org/10.1038/s41598-021-94628-0.
77. Hassoun S, Jefferson F, Shi X, et al. Artificial intelligence for biology. Integr Comp Biol. 2022;61:2267–75. https://doi.org/10.1093/icb/icab188.
78. Thessen AE, Patterson DJ. Data issues in the life sciences. ZooKeys. 2011;150:15–51. https://doi.org/10.3897/zookeys.150.1766.
79. Sidak D, Schwarzerová J, Weckwerth W, et al. Interpretable machine learning methods for predictions in systems biology from omics data. Front Mol Biosci. 2022;9:1–28. https://doi.org/10.3389/fmolb.2022.926623.
80. Zhou ZH. A brief introduction to weakly supervised learning. Natl Sci Rev. 2018;5:44–53. https://doi.org/10.1093/nsr/nwx106.
81. Zhang Z, Zhao Y, Liao X, et al. Deep learning in omics: a survey and guideline. Briefings Functional Genomics. 2019;18:41–57. https://doi.org/10.1093/bfgp/ely030.
82. Camargo G, Bugatti PH, Saito PTM. Active semi-supervised learning for biological data classification. PLoS One. 2020;15:e0237428. https://doi.org/10.1371/journal.pone.0237428.
83. Huang D, Song B, Wei J, et al. Weakly supervised learning of RNA modifications from low-resolution epitranscriptome data. Bioinformatics. 2021;37:i222–i30. https://doi.org/10.1093/bioinformatics/btab278.
84. Ratner A, De Sa C, Wu S, et al. Data programming: creating large training sets, quickly. In: Lee DD, von Luxburg U, Garnett R, Sugiyama M, Guyon I, eds. Advances in Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc.; 2016: 3574–82.
85. van Engelen JE, Hoos HH. A survey on semi-supervised learning. Mach Learn. 2020;109:373–440. https://doi.org/10.1007/s10994-019-05855-6.
86. Bilal M, Raza SEA, Azam A, et al. Development and validation of a weakly supervised deep learning framework to predict the status of molecular pathways and key mutations in colorectal cancer from routine histology images: a retrospective study. Lancet Digital Health. 2021;3:e763–e72. https://doi.org/10.1016/S2589-7500(21)00180-1.
87. Zhang Q, Zhu L, Bao W, et al. Weakly-supervised convolutional neural network architecture for predicting protein-DNA binding. IEEE/ACM Trans Comput Biol and Bioinf. 2020;17:679–89. https://doi.org/10.1109/TCBB.2018.2864203.
88. Ghosal S, Zheng B, Chapman SC, et al. A weakly supervised deep learning framework for sorghum head detection and

counting. Plant Phenomics. 2019;2019:1525874. https://doi.org/10.34133/2019/1525874.

89. Petti D, Li C. Weakly-supervised learning to automatically count cotton flowers from aerial imagery. Comput Electron Agric. 2022;194:106734. https://doi.org/10.1016/j.compag.2022.106734.
90. Chen J, Deng X, Wen Y, et al. Weakly-supervised learning method for the recognition of potato leaf diseases. Artif Intell Rev. 2023;56:7985–8002. https://doi.org/10.1007/s10462-022-10374-3.
91. Yan J, Wang X. Unsupervised and semi-supervised learning: the next frontier in machine learning for plant systems biology. Plant J. 2022;111:1527–38. https://doi.org/10.1111/tpj.15905.
92. Sen P, Lamichhane S, Mathema VB, et al. Deep learning meets metabolomics: a methodological perspective. Briefings Bioinf. 2021;22:1531–42. https://doi.org/10.1093/bib/bbaa204.
93. Wang F, Liigand J, Tian S, et al. CFM-ID 4.0: more accurate ESI-MS/MS spectral prediction and compound identification. Anal Chem. 2021;93:11692–700. https://doi.org/10.1021/acs.analchem.1c01465.
94. Fan Z, Alley A, Ghaffari K, et al. MetFID: artificial neural network-based compound fingerprint prediction for metabolite annotation. Metabolomics. 2020;16:1–11. https://doi.org/10.1007/s11306-020-01726-7.
95. Domingo-Almenara X, Montenegro-Burke JR, Benton HP, et al. Annotation: a computational solution for streamlining metabolomics analysis. Anal Chem. 2018;90:480–89. https://doi.org/10.1021/acs.analchem.7b03929.
96. de Jonge NF, Mildau K, Meijer D, et al. Good practices and recommendations for using and benchmarking computational metabolomics metabolite annotation tools. Metabolomics. 2022;18:1–22. https://doi.org/10.1007/s11306-022-01963-y.
97. Bilbao A, Munoz N, Kim J, et al. PeakDecoder enables machine learning-based metabolite annotation and accurate profiling in multidimensional mass spectrometry measurements. Nat Commun. 2023;14:2461. https://doi.org/10.1038/s41467-023-37031-9.
98. Mallick H, Franzosa EA, Mclver LJ, et al. Predictive metabolomic profiling of microbial communities using amplicon or metagenomic sequences. Nat Commun. 2019;10:3136. https://doi.org/10.1038/s41467-019-10927-1.
99. Bartmanski BJ, Rocha M, Zimmermann-Kogadeeva M. Recent advances in data- and knowledge-driven approaches to explore primary microbial metabolism. Curr Opin Chem Biol. 2023;75:102324. https://doi.org/10.1016/j.cbpa.2023.102324.
100. Stanke M, Diekhans M, Baertsch R, et al. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24:637–44. https://doi.org/10.1093/bioinformatics/btn013.
101. Holst F, Bolger A, Günther C, et al. Helixer–de novo prediction of primary eukaryotic gene models combining deep learning and a hidden Markov model. Biorxiv. 2023. https://doi.org/10.1101/2023.02.06.527280. Accessed 2 April 2024.
102. Kulmanov M, Khan MA, Hoehndorf R. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34:660–68. https://doi.org/10.1093/bioinformatics/btx624.
103. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020;36:422–29. https://doi.org/10.1093/bioinformatics/btz595.
104. Dumschott K, Dörpholz H, Laporte MA, et al. Ontologies for increasing the FAIRness of plant research data. Front Plant Sci. 2023;14:1279694. https://doi.org/10.3389/fpls.2023.1279694.
105. Whetzel PL, Noy NF, Shah NH, et al. BioPortal: enhanced functionality via new web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res. 2011;39:W541–W45. https://doi.org/10.1093/nar/gkr469.
106. Gardezi M, Joshi B, Rizzo DM, et al. Artificial intelligence in farming: challenges and opportunities for building trust. Agron J. 2024;116:1217–28. https://doi.org/10.1002/agj2.21353.
107. Kamdar MR, Musen MA. An empirical meta-analysis of the life sciences linked open data on the web. Sci Data. 2021;8:1–21. https://doi.org/10.1038/s41597-021-00797-y.
108. Chatterjee A, Swierstra T. Making FAIR trustworthy. SocArXiv. 2021. https://doi.org/10.31235/osf.io/x4csm. Accessed 2 April 2025.
109. Du X, Dastmalchi F, Ye H, et al. Evaluating LC-HRMS metabolomics data processing software using FAIR principles for research software. Metabolomics. 2023;19:11. https://doi.org/10.1007/s11306-023-01974-3.
110. Leite ML, de L, Costa LS, et al. Artificial intelligence and the future of life sciences. Drug Discov Today. 2021;26:2515–26. https://doi.org/10.1016/j.drudis.2021.07.002.

Received: October 31, 2024. **Revised:** April 3, 2025. **Accepted:** April 9, 2025