Sheppard *et al. Genome Biology* (2025) 26:313
https://doi.org/10.1186/s13059-025-03775-4

Genome Biology

REVIEW

Open Access

Machine learning and statistical inference in microbial population genomics

Samuel K. Sheppard[1], Nicolas Arning[2], David W. Eyre[2,3,4] and Daniel J. Wilson[2,5*]

*Correspondence:
daniel.wilson@bdi.ox.ac.uk

[1] Ineos Oxford Institute for Antimicrobial Research, Department of Biology, University of Oxford, Oxford, United Kingdom
[2] Big Data Institute, Oxford Population Health, University of Oxford, Oxford, United Kingdom
[3] NIHR Oxford Biomedical Research Centre, Oxford, United Kingdom
[4] NIHR Health Protection Research Unit in Healthcare Associated Infections and Antimicrobial Resistance, University of Oxford, Oxford, United Kingdom
[5] Oxford University Department for Continuing Education, Oxford, United Kingdom

Abstract

The availability of large genome datasets has changed the microbiology research landscape. Analyzing such data requires computationally demanding analyses, and new approaches have come from different data analysis philosophies. Machine learning and statistical inference have overlapping knowledge discovery aims and approaches. However, machine learning focuses on optimizing prediction, whereas statistical inference focuses on understanding the processes relating variables. In this review, we outline the different aspirations, precepts, and resulting methodologies, with examples from microbial genomics. Emphasizing complementarity, we argue that the combination and synthesis of machine learning and statistics has potential for pathogen research in the big data era.

Background

Advances in technology and data generation have driven a big data revolution in microbiology, with studies routinely analyzing thousands of whole genome sequences. Datasets generated with ever-increasing volume, variety, and velocity bring tremendous opportunities as well as unique analysis challenges. Inspired by the promise of deeper understanding and driven by high-throughput low-cost DNA sequencing, there are now vast genome libraries of bacterial species approaching one million genomes [1]. Achieving the potential of these resources has required the scaling of conventional statistical methods which face challenges with high-dimensional data, necessitating simplifications and approximations [2]. This is paradoxical, because the vast information content of modern resources should make it easier to glean biological insights about evolutionary origins, transmission dynamics and the genetic basis of phenotypic diversity. Machine learning (ML) approaches offer a potential solution as they can handle very large and heterogeneous datasets [3]. ML is a multidisciplinary pursuit that draws heavily on statistics and computer science. Quantitative approaches to exploiting data underpin both endeavors, but for the purposes of this review we work with the following distinction: statistical inference is a tool for furthering our scientific understanding of the world,

while ML is a tool for engineering automatable solutions to prediction, simulation, and pattern recognition.

ML has driven breakthroughs in generative artificial intelligence (AI) including natural language, image, and audio creation [4]. In the biological sciences, ML has surpassed human-engineered solutions to the prediction of 3D protein structures [5], translating nanopore potentials into DNA base calls [6], and discovering antimicrobial peptides from large protein and metagenomic databases [7, 8]. In microbial population genomics, contemporary big data whole-genome approaches often combine statistical inference and ML to answer diverse questions relating to the evolution and epidemiology of infectious diseases [9]. These include predicting future events (e.g., outbreaks), understanding the impact of variables (e.g., virulence and resistance genes), and discovering data patterns (e.g., commonalities in infection risk). Often the best tool for the job is unclear or ambiguous. Here, we provide some perspective on the suitability of statistical inference versus ML for different problems in microbial population genomics by summarizing the approaches and discussing examples.

Principles of machine learning and statistical inference

ML and statistical inference are tools for modelling often large and complex data that have been numerically encoded into one or more variables, say input features x and outcomes y. For a comprehensive introduction to the subject, see Murphy [10, 11]. A unifying concept is the *data generating process*, which represents the underlying scientific and sampling processes that led to the data at hand. Both ML and statistics attempt to approximate the data generating process as a mathematical function. Broadly speaking, statistics tends to employ models motivated by a desire to understand underlying *processes*, while ML employs flexible models that can faithfully reproduce observed *patterns*, agnostic to the underlying process.

A distinction has been drawn between competing approaches to modelling data generating processes: *data modelling* and *algorithmic modelling* [12] (Fig. 1). Traditionally, data modelling has been the dominant paradigm, particularly in statistics, in which the data generating process is approximated by deriving a model based on assumptions about relationships between variables, both deterministic and stochastic. Data modelling emphasizes interpretability of the model and transparency of modelling assumptions. Model complexity is usually chosen by trading off realism against tractability, considering parsimony and computational burden. Special emphasis is often placed on *domain-specific knowledge* and *probabilistic models* in the data modelling approach. However, models need not be complicated, with simple additive linear assumptions underpinning workhorses like linear regression, logistic regression, and ANOVA.

In contrast, algorithmic modelling aims to provide general purpose approximations to unknown data generating processes without detailed prior knowledge [13]. Recent advances in ML have focused attention on algorithmic modelling, which also encompasses non-parametric statistical techniques. It leans on flexible algorithms capable of accurately reproducing the structure of complex data in very general settings. This flexibility often entails parameter-rich models that require large training datasets. Therefore algorithm development in ML prioritizes *computational efficiency*. Deep neural networks have proved particularly adept for algorithmic modelling [14]. In addition, the ML

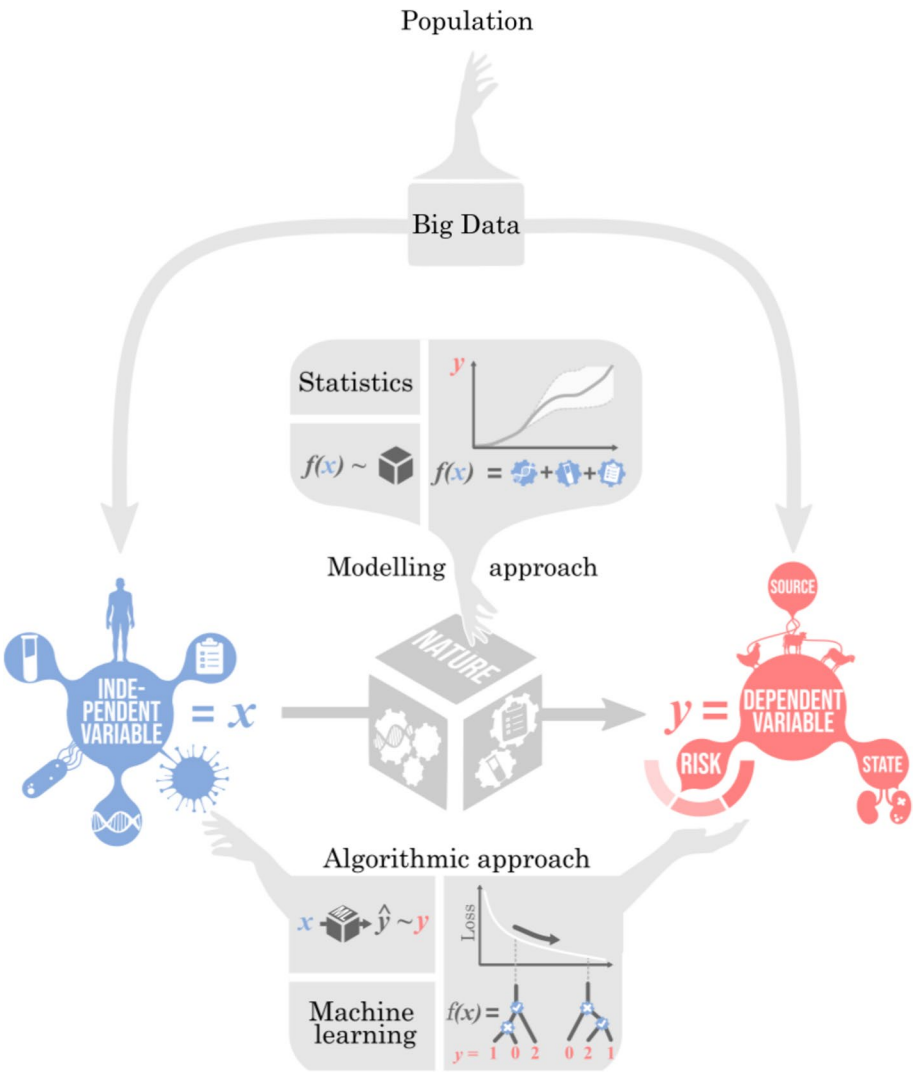

Fig. 1 Modelling and algorithmic approaches. Big data drawn from an example population that describes the objects from which data are randomly sampled (hand). This contains features, otherwise known as independent variables, predictors or regressors, and outcomes, otherwise known as dependent variables, labels, classes, or targets, whereby changes in the features lead to changes in the outcomes. Relating the two is the data generating process, or "nature". Statistics (or, more precisely, data modelling in Breiman's dichotomy [12]) aims to understand the underlying processes while ML (or, more precisely, algorithmic modelling in Breiman's dichotomy) aims to faithfully reproduce the observed patterns to achieve optimal prediction, for instance

toolkit comprises a wide variety of techniques, many of which are available in *Python* software libraries such as Scikit-Learn, PyTorch, and TensorFlow [15–17].

Supervised vs unsupervised learning

In *supervised learning*, a mathematical function models the relationship between variables representing the *features* x and the *outcomes* y, often with the aim of explaining or predicting y in terms of x. Very often y is low-dimensional; it can be binary, e.g., describing if an event does or does not occur, categorical, describing one of several possible outcomes, or continuous. In contrast, x is often high-dimensional, consisting of

many inputs that potentially influence or predict the outcome of interest. In microbial population genomics, y would often be a phenotype, such as drug susceptibility, and x could represent the genome sequences. Genomes are typically encoded numerically for such analyses as described below. Supervised learning includes familiar approaches such as *classification* and *regression* for modelling genotype-to-phenotype relationships (e.g., [18, 19]). In *unsupervised learning*, a mathematical function models the relationship within the data x, often to reveal hidden structure or simulate new data. In recent approaches, such as the large language models (LLMs) that power ChatGPT, x represents digitally encoded text numbering several hundred billion words [20]. In microbial genomics, an important application of unsupervised learning is the detection of genetic clustering (e.g., [21–24]).

Feature engineering for genome sequence data

Before analyzing genome data, the molecular sequences must be encoded as *features*, or vectors of numbers. Typically, features are defined in terms of genetic variation—the parts of the sequence that differ between two or more genomes. Features are often defined relative to a reference genome. For example, single-nucleotide polymorphisms (*SNPs*), which may be encoded as elements of a binary vector representing the reference allele (e.g., by the number 0) or non-reference allele (1) in each genome. If more than one non-reference allele exists, the second and third non-reference alleles are represented by additional binary vectors, so a single SNP generates multiple features, known as *dummy variables* or *one-hot encoding*. Likewise, *alleles* at a specific locus can be represented with binary vectors recording the presence (1) or absence (0) of each non-reference sequence in each genome. If there are K alleles at a locus, this produces $K-1$ features. For accessory genes, *presence or absence* of the entire locus can be encoded as a binary vector.

Reference-free approaches are also popular. Genome assemblies or protein sequences can be chopped into short, overlapping windows of oligonucleotides or oligopeptides known as k-mers, where k represents the sequence length. The presence or absence of each k-mer in each genome can be encoded as a binary vector. Very short k-mers ($k<5$) are informative about nucleotide composition, whereas k-mers in the range 10–50 can capture locus-specific variation in SNPs, indels, and gene presence or absence. If k is much longer, the k-mers become rare or unique to individual genomes, and thereby less useful for inference or prediction. More advanced uses of k-mers reduce the number of features while retaining their biological meaning, for example by merging k-mers that always (or sometimes) share the same (or similar) pattern of presence versus absence across genomes into a smaller number of *unitigs* (or *embeddings*) which are encoded as binary (or continuous) vectors (e.g., [24, 25]).

Biological questions and analysis goals

Framing the goal of an analysis firmly in terms of the biological question helps narrow down the appropriate ML or statistical inference tools. Biological questions map on to analysis goals including (i) data exploration, (ii) prediction, (iii) parameter estimation, and (iv) hypothesis testing. In *data exploration*, the aim is often familiarization, visualization, or hypothesis generation. These aims are open-ended, but they share a common

theme of identifying or communicating important characteristics of the data or—conversely—of not missing important aspects of the data. Often, an analysis goal can be formalized quantitatively through a *loss function* that is constrained or minimized. Considering loss functions helps when comparing ML and statistical approaches.

In *prediction*, the aim is to predict, impute, classify or simulate new, unobserved, or deliberately masked data by exploiting patterns in observed data, while simultaneously minimizing prediction error: the difference between the truth and the prediction. Common loss functions for prediction include squared error for continuous outcomes or 0–1 mis-classification error for discrete outcomes, where 1 indicates misclassification and 0 indicates correct classification [26, 27].

In *parameter estimation*, the aim is to precisely quantify the parameters of a mathematical model assumed to describe the data generating process. Common loss functions for estimation include the error, absolute error, and squared error. Finally, in *hypothesis testing*, the intention is to draw qualitative conclusions, for example that a variable affects the outcome. Here, false positives are commonly encoded as a 0–1 loss function indicating whether a null hypothesis has been rejected erroneously (1) or not (0).

Comparing performance

The performance of ML and statistical methods can be compared by averaging the loss function across observed datapoints (*empirical risk*), or across a prior distribution (*Bayes risk*), or across theoretical re-enactments of the data generating process (*frequentist risk*). Empirical risk is convenient, but requires a *ground truth*, and therefore is most applicable to prediction, where predictions might be compared directly to observed data that were deliberately masked or set-aside to measure prediction accuracy. ML offers a rich toolbox of flexible algorithms for prediction, which often helps the analyst achieve a smaller empirical risk than using traditional statistical approaches alone.

Statistical methods help when a ground truth is unavailable, for example when estimating parameters and testing hypotheses about unobserved processes. *Maximum likelihood estimation* and *likelihood ratio tests* are widely used classical approaches that minimize or constrain frequentist risks such as the mean squared error (for estimation) and family-wise error rate (for hypothesis tests). These guarantees are subject to technical assumptions like large sample size and (for hypothesis tests) nesting of models. When we are willing to make prior assumptions about the likely values of unknown parameters, Bayesian inference is useful for parameter estimation and hypothesis testing, because it minimizes or constrains Bayesian loss functions like the mean squared error (for prediction or estimation) and false discovery rate (for hypothesis tests). It does not rely on assumptions like large sample size, but Bayesian approaches can be computationally intensive.

Fitting models

Data are typically split into *training and testing* sets when a ground truth is available, allowing empirical risk to be minimized (during training) and measured (during testing). Parameters are optimized using training data, then final performance is evaluated using testing data. The idea is to obtain an independent, unbiased estimate of performance, but this may be undermined by dependence between training and testing data.

Sometimes ML models entail hyper-parameters that are difficult to fit during training, so an intermediate *validation* set is employed to optimize them using grid search (Fig. 2). *Cross-validation* is a popular technique used to average over different ways of splitting the data [28]. In classical and Bayesian statistics, for estimation and hypothesis testing, often the whole data is used to fit the model, since Bayes risk or frequentist risk can be optimized theoretically. This makes more efficient use of the data.

In both ML and statistical inference, particularly in parameter-rich or data-limited settings, *over-fitting* risks noisy parameter estimates and poor generalizability to other data [29, 30]. To mitigate over-fitting, it is common to practice *regularization*, in which parameter values are constrained in some way. Examples of regularization include penalized likelihoods and Bayesian priors. *Ensemble methods*, like bootstrap aggregating in

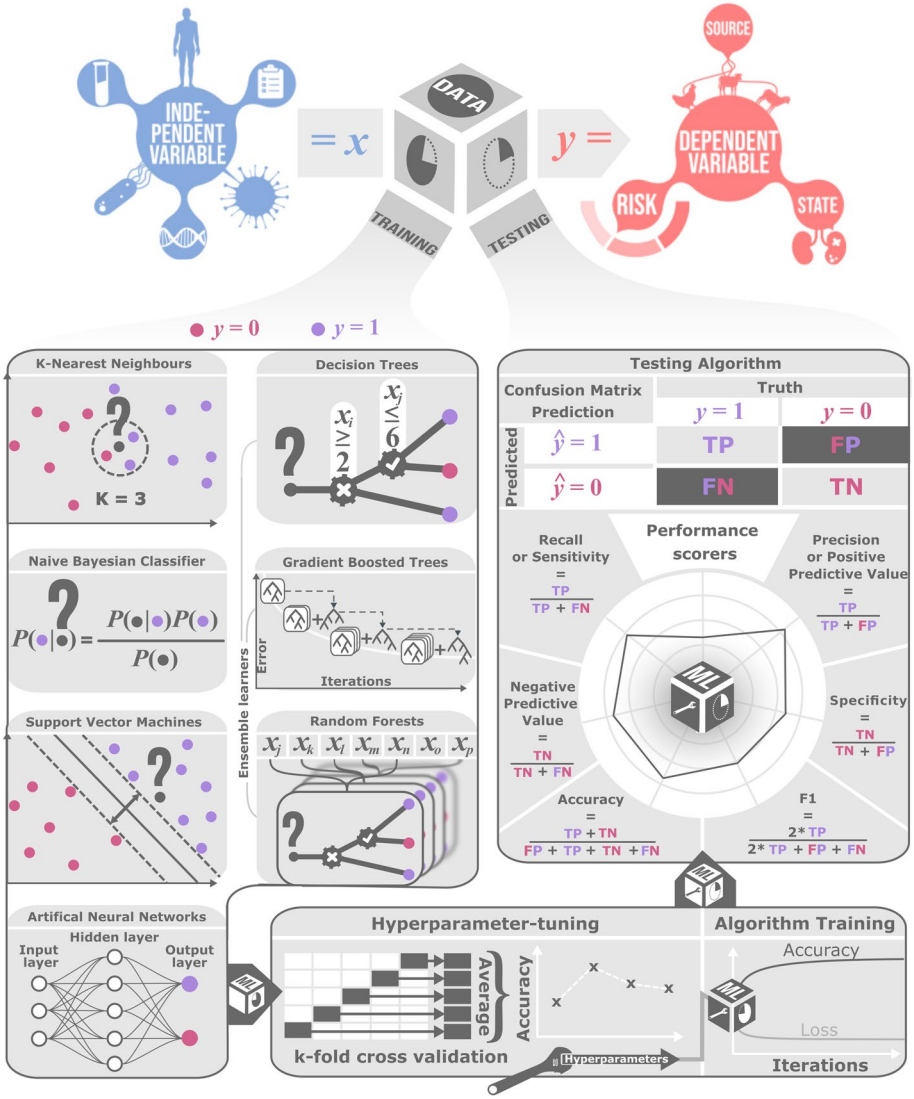

Fig. 2 Machine learning workflow in classification tasks. The data is split into training and testing, after which a suitable general-purpose algorithm is chosen, its hyper-parameters tuned and fitted to the training data. The performance of the fitted classifier is subsequently measured using a metric of choice

Random Forest and boosting in Gradient Boosted Trees, reduce over-fitting by optimizing performance across pseudo-replicated data. In contrast, *dropouts* in artificial neural networks avoid over-fitting by optimizing performance across randomly pruned networks to build resilience and avoid unstable over-specialization of neurons during training. Training algorithms in ML can often be tuned to reduce over-fitting by modifying a tuning parameter known as the *learning rate*, and developing strategies known as *early stopping rules*. Concerns about over-fitting must be weighed against over-correcting by poorly or under-fitting the model, a balance known as the *bias-variance trade-off*.

Machine learning classifiers common in microbial genomics

In *classification*, the challenge is to predict or explain the outcome variable, y, a categorical variable (or "class") that takes one of a fixed number of values (or "labels"), using information in the features, x. Typically, algorithms have parameters that are calibrated by optimizing accuracy in a training dataset. There are several common ML classifiers used in microbial genomic analyses with varying levels of complexity. Among the earliest classification algorithms is *k-Nearest Neighbors*. Here, the inferred class is the one most frequently observed in the k datapoints from the training data that are closest to x, in some sense. This requires a distance metric [31, 32]. Applications include predicting gene function and phenotypes from DNA sequences [33, 34]. Another relatively simple method is the highly scalable *Naïve Bayes* approach, where the class is assigned using Bayes' theorem assuming independence among the features. Here, the inferred class is the one with the highest posterior probability [35–37]. Statistical distributions (e.g., Gaussian, Bernoulli) are assumed for conditional likelihoods, whose parameters must be learned. Applications include disease diagnosis [38] and sequence-based taxonomy for genomes, meta-genomes [39], and horizontally transferred genes [40].

There are several more sophisticated approaches including Support Vector Machines, Decision Trees and Artificial Neural Networks. *Support Vector Machines* offer a flexible approach to classification based on kernels, which measure the similarity of features between datapoints. Non-linear kernels facilitate classification in difficult problems like image analysis. Results can be sensitive to tuning the parameters [41–43]. Applications include detection of horizontal gene transfer [44], predicting molecular phenotypes from genome sequences [45, 46] and classifying host specificity [47].

Decision Trees can be compared to keys in biology field guides used to identify species. Here, decision trees represent a hierarchical sequence of rules that use features to assign a label or class. Rules are trained using heuristic "greedy" algorithms and pruned to mitigate over-fitting. Easy-to-interpret, individual decision trees are usually used ensemble to improve accuracy and reduce noise [48, 49]. Well-known *Random Forests* are an ensemble method where features and datapoints are repeatedly subsampled ("bootstrapped") when training to build many decision trees. The most frequent classification across trees is used ("aggregated"), which improves accuracy [50–52]. Applications include predicting pathogenicity, disease status, antimicrobial resistance, genome content, and host specificity [53–62]. *Gradient Tree Boosting* is another ensemble method in which a forest of decision trees is grown stepwise, with the last tree trained to improve the previous step's prediction, assessed via a loss function [63–65]. Applications include

predicting pH preference and antimicrobial resistance from relevant gene sequences [66, 67].

Finally, inspired by neuroscience, *Artificial Neural Networks (ANNs)* have become a popular ML approach in microbial genomics. ANNs comprise directed graphs (networks) of simple functions (artificial neurons). ANNs vary in architecture but typically organize neurons into observed (input and output) layers and one or more hidden layers [68]. Communication occurs between layers of the ANN [69]. *Deep learning* employs ANNs with multiple hidden layers, which produces complex and flexible models with large information processing capacity [70, 71]. Advances in big data availability, GPUs (graphics processing units) and theoretical innovations have allowed parameter-rich ANNs to be fitted efficiently. Applications include identifying species, strains, and gene function from DNA sequences [72–75]. ANNs perform well partly because they act as universal function approximators through approximating arbitrary continuous relationships, given enough hidden neurons [76], and partly because the fitting techniques are thought to impose regularization (e.g., [77]). *Attention mechanisms* enable some ANNs, notably *transformers*, to dynamically weight the influence of input elements based on context, rather than relying on fixed connection patterns [78]. This allows the network to selectively focus on the most relevant parts of the input, regardless of their position. Attention is useful for dependencies in molecular sequences or three-dimensional protein structures, where traditional architectures struggle to propagate long-range information. Attention mechanisms allow every input element to directly consider all others in parallel, avoiding the dilution of important but distant signals. Attention has driven breakthroughs in generative AI [79], antibiotic prediction [7, 8], and protein structure prediction [5].

Strengths and weaknesses of machine learning and statistics

A clear statement of the biological questions informs the analysis goals by determining what type of loss to minimize. Minimizing estimation error versus prediction error versus false positives guides the choice of method. Data analysis aimed at understanding underlying processes causally is better served by statistical inference, because it will minimize the (Bayesian or frequentist) risk associated with estimation and hypothesis testing. Data analysis aimed at optimizing model-agnostic problem-solving performance is better served by ML, because it can minimize the (empirical) risk of prediction whenever a ground truth is available [29, 80]. The dominant statistical paradigm emphasizes principles like *parsimony* and *explainability*, whereas sophisticated ML algorithms can produce demonstrably superior performance over simpler models common in statistics. This is exemplified by classic supervised learning examples like the XOR problem, where the output is not a linear function of the input data.

Out-of-the-box, many ML approaches deal with *collinearity*, *non-linearity*, and *interactions* better than traditional statistical approaches like regression. An experienced statistician might employ regularization to counter unreliable parameter estimates and high uncertainty caused by strongly correlated or collinear features, but regularization is built in to many ML algorithms as standard. Non-linear relationships between features and outcomes, and interactions between features, can also be modelled statistically, but this requires some sophistication and manual intervention on the part of the data

analyst, whereas many ML algorithms are designed to model these phenomena automatically. ML algorithms can often prioritize among thousands of features, allowing the user to take an agnostic approach to feature selection. However, the cost of sophisticated ML is models whose workings and parameters are less transparent to interpretation [81], often termed *black boxes* [19].

The strong performance of machine automation and the advantages of model agnosticism have de-emphasized the perceived importance of human accountability for *data quality* issues; this is known as *automation bias*. *Biased sampling* and *batch effects* create problems for both ML and statistical inference, by generating conclusions that may be misleading or poorly generalizable (see "Data quality and interrogating results"). Moreover, concerns relating to *interpretability*, *equality*, and *accountability* are important in many settings, particularly healthcare [82]. Therefore, trade-offs that exist between the performance of a model in a specific, loss function sense, versus its wider utility for society, may shift the balance between preferring ML versus statistical inference. The ML vs statistics and data modelling vs algorithmic modelling dichotomies recall a more fundamental distinction between *deductive* (logic-based) vs *inductive* (observation-based) scientific inference. The fundamentally *empirical* approach to modelling of ML is data-driven and data-hungry, explaining its reliance on big data and susceptibility to biased datasets, but also its superior flexibility to fit the data more closely.

Data quality and interrogating results

The adage "garbage in, garbage out" is a truism in ML and statistics: proper data preparation and quality checking (*QC*) are indispensable to any analysis. Researchers must adopt strategies to diagnose *data quality* issues before and after analysis.

As a first step, it is essential to understand the *provenance* of the data, its limitations, and whether it is adequate for the analysis goals. Next, data must be quality-checked using approaches including summary statistics and visuals to diagnose issues such as data entry errors, outliers, missing values, and special values. Data must be properly encoded, especially missing or special values, to ensure ML or statistical algorithms handle them appropriately. An *imputation* step, to predict missing values, may be required. Alongside QC, data exploration is valuable for hypothesis generation and selecting an appropriate model that makes reasonable assumptions.

Before merging datasets that may have been collected in different places, at different times, by different processes, or for different purposes, it is important to consider how the analysis might be influenced by *heterogeneity*—systematic differences between the datasets. For example, there might be unmeasured *confounders* that differ between them. Systematic differences in outcomes across datasets make analysis particularly vulnerable to so-called *batch effects*. Sometimes heterogeneities are "controlled" by including the batch label as a feature. A more robust but less efficient approach is *meta-analysis*, in which datasets are analyzed separately and results are compared post-analysis, merging if appropriate. Often this fits neatly into training, testing, and validation, especially since *out-of-sample prediction* is a more robust indicator of generalizability than splitting a single dataset.

After analysis, it is essential to interrogate the data again to understand where informative signals come from and diagnose unresolved QC issues or implementation *bugs*. A

healthy *skepticism*, especially for surprising results, is important and consideration of questions such as: (i) How do results compare to the *literature*? (ii) Are results robust to the analysis assumptions? Benchmarking against simpler methods can help here. Unless the signal in the data can be explained—for example using visualization or explainable AI—it may be difficult to convince peers. Experimental *validation* and *replication* in independent datasets are often required to build credibility and, to repeat another truism, "extraordinary claims require extraordinary evidence."

Applications of ML and statistics in microbial genomics

In this section, we consider applications of ML and statistics in microbial genomics and discuss the relative advantages of the competing approaches in the context of three examples: source attribution in zoonotic bacteria, genome-wide association studies of antimicrobial resistance, and predicting antimicrobial resistance from genome sequences.

Example 1: Source attribution in campylobacter

#prediction #classification #supervised_learning #machine_learning.

Features ($\boldsymbol{x}$): genome sequences. Outcomes ($\boldsymbol{y}$): host species of origin.

Identifying the population-of-origin of bacterial infections has practical applications for a range of pathogens, particularly multi-host organisms that cause zoonotic infections in humans like *Salmonella*, *Escherichia coli*, and *Campylobacter*. Person-to-person transmission of *Campylobacter*, a common cause of gastro-enteritis in humans, is rare, with most cases caused by consumption of contaminated food. *Campylobacter* commonly colonizes the guts of birds and mammals, including animals farmed for meat and poultry, and is found in environmental water. Therefore, each human case is thought to originate from one of the source reservoirs, and it is useful to predict, or "attribute," the source. Source attribution helps prevent future human cases by informing efforts to disrupt the transmission chain.

DNA sequencing has been exploited for source attribution in *Campylobacter* using a variety of tools. Data typically comprise DNA sequences of *Campylobacter* isolated from cases of human infection, and, for comparison, from animal and environmental reservoirs. Early approaches pursued statistical epidemiological models, using strain-level designations to rule out transmission (e.g., [83, 84]). Later, statistical models grounded in population genetics, like *Structure* and *iSource*, were applied, which exploited more of the information in the DNA [85, 86]. However, source attribution can be formulated as a straightforward ML problem, where the analysis goal is to minimize mis-classification error. DNA sequences of *Campylobacter* sampled directly from source populations can be used to train a classifier with known labels (e.g., cattle, sheep, pigs, chicken, environmental water). Classifier accuracy can be tested using cross-validation. The population-of-origin of each human case can then be predicted from the DNA sequence. ML classifiers proved to be faster and around 11% more accurate than the established statistical approaches applied to multi-locus sequence typing (71% vs 64%), and readily generalized to the analysis of whole genome sequencing (WGS) data, permitting 33% gains in accuracy (85% vs 64%) [58, 59]. Random Forest [51] and XGBoost [87] produced the greatest improvements. Key to the success of ML in this context has been the availability

of big data comprising thousands of whole genomes with a high degree of replication: 5799 genomes sampled from the source populations of interest, together with 15,988 genomes from human infection.

Example 2: genome-wide association studies of antimicrobial resistance

#hypothesis_testing #parameter_estimation #regression #statistics.

Features ($\boldsymbol{x}$): genome sequences. Outcomes ($\boldsymbol{y}$): antimicrobial resistance or sensitivity.

A major aim of biology in the twenty-first century is to unravel the genetic architecture of phenotypic diversity within species [88]. In microbiology, there is particular interest in traits that affect the outcome of human colonization and infection, like virulence (the frequency or severity of disease) and antimicrobial resistance (AMR). Early approaches to such questions studied candidate genes, for example using PCR to test for differences in the frequency of genetic markers between cases and controls (e.g., [89]). With the advent of technologies like genotyping arrays and, later, whole-genome sequencing, the accepted approach to such questions has been to scan the genome for evidence of associations between allelic differences and phenotypic differences. So-called genome-wide association studies (GWAS) address concerns that candidate gene approaches are vulnerable to selection and reporting bias, and struggle to control artefactual associations caused by population stratification of phenotypes, for example when phenotypes differ between strains [90–92].

GWAS is motivated by a desire to learn about the causal process underlying the data, and pains are taken to avoid artefactual signals of association, while recognizing that observational studies cannot prove causality (see, e.g., [93, 94]). This is a statistical inference problem in which the parameters of a relatively simple and readily interrogated general linear model are interpreted to identify genetic variants responsible for observable phenotypic diversity. Special emphasis is placed on limiting the expected losses caused by false positive associations. In bacteria, GWAS have been applied to a range of traits and species (e.g., [54, 95–99]). While ML approaches have been applied to this problem, and while informative for data exploration and hypothesis generation, particularly in expert hands [100], ML approaches only return "high-leverage" genes or genetic variants that help predict the outcome. Out-of-the-box they neither test nor quantify the evidence for the hypothesis that these variants directly influence the outcome. Nor do they offer theoretical or empirical tools for easily controlling family-wise error or false discovery rates across loci. Statistical approaches address these foundational issues, and mapping of genes underlying AMR has proved particularly fruitful (e.g., [101, 102]), presumably because mechanisms of genetic resistance are often direct, almost deterministic. GWAS depends on big data to find signals of association, but interpretation of those signals relies on explicit modelling assumptions, and not on training a general-purpose algorithm using datasets of many known genotype-to-phenotype associations, which as yet do not exist.

Example 3: predicting antimicrobial resistance from genome sequences

#prediction #classification #supervised_learning #interpretable_machine_learning.

Features ($\boldsymbol{x}$): genome sequences. Outcomes ($\boldsymbol{y}$): antimicrobial resistance or sensitivity.

Related to the problem of inferring which genes confer antimicrobial resistance is the problem of predicting antimicrobial resistance from an individual bacterial genome. Modernizing microbiological diagnostics in clinical practice has been a major focus of research over the last 15 years, with aspirations to replace a battery of phenotypic tests with a streamlined WGS and phenotype prediction pipeline [103]. WGS has become routine in some healthcare settings, particularly for organisms that are challenging to test in the laboratory, like the slow-growing and high biosafety level pathogen *Mycobacterium tuberculosis* [104, 105].

The statistical models used for GWAS could be turned to prediction, but the superior flexibility of ML algorithms to fit data more closely make them a natural choice for predicting AMR (e.g., [100, 106–110]). In this setting, the analysis goal is to minimize prediction error, which can be quantified empirically because a ground truth is available. Large datasets have been generated comprising WGS and traditional AMR phenotyping assays and based on these, automated predictions with high accuracy have been achieved—in some cases exceeding the standards required of traditional laboratory diagnostics [111, 112]—confirming the excellent performance of ML algorithms for general-purpose prediction.

ML performance in AMR prediction has established it as an important tool for predicting all manner of bacterial phenotypes from WGS data. However, there is a question of accountability: in a medical setting, decision-taking responsibility lies with the clinical microbiologist. Therefore the ML algorithm needs to present the evidence for its prediction transparently for interpretation by the domain-specific expert. Scenarios like this create a need for *explainable AI* that goes significantly beyond outputting coefficients for predictive features, which may be mere confounders, rather than biologically causal genetic variants, particularly in the presence of population stratification [113]. Approaches to explainable AI include *attribution algorithms*, which may impose post hoc linearization of the predictions (e.g., [114, 115]). This leads back to simpler, more transparent data models resembling additive or linear models. Alternatively, *ablation algorithms* systematically drop features-of-interest from the model to assess their impact on performance [116]. Consequently, even when pursuing prediction via complex ML, efforts to interpret prediction may resemble more traditional statistical analysis in which high importance is attached to understanding in a causal way the conclusions and interpretation of the data.

Statistics versus machine learning: right tool for the right job

The boundary between ML and statistics is blurred, with cross-over methods like the elastic net, bootstrap, non-parametric statistics, and Bayesian-inspired approaches. The labels "machine learning" and "statistics" are typically less useful than a clear definition of the analysis goals—prediction, exploratory data analysis, parameter estimation, hypothesis testing—which in turn are framed by the biological questions. Where a project has multiple goals, such as prediction and hypothesis testing, it is reasonable to apply different analysis approaches to the same data. However, as example 3 illustrates, even when a task clearly fits the goal of prediction, the choice of method is influenced by context-specific considerations, notably explainability and accountability. Frequently in scientific applications, there is an emphasis on understanding and interpreting the

data generating process, and this may tip the balance away from ML and towards statistical inference. Interrogating results in real data analysis, detecting data quality issues like batch effects, explaining which signals drive the results, controlling for confounding factors, and understanding the limits to generalizability, are essential to the integrity of scientific outputs. Developing strategies to check scientific results is a key step towards scientific independence that allows a researcher to take responsibility for final conclusions. The risk of automation bias, in which responsibility for final conclusions is delegated to opaque algorithms, and abdication of critical thinking, are rightly of concern.

Conclusions and future directions

We are currently in a period of exploration, as ML and AI are increasingly applied to diverse questions like "what is the genetic architecture of virulence," "why do dangerous pathogens emerge," and "how do we fight the spread of antimicrobial resistance"? In allied fields, we have seen transformative innovations ranging from the prediction of 3D molecular structure [5] to antimicrobial peptide discovery [7, 8] and, looking ahead, the design of novel proteins and molecular systems based on free text (e.g., [117–120]). In microbial population genomics, we anticipate that ML will continue to play a leading role, both by improving on previous approaches, and by opening new avenues of research and understanding. If, in the years to come, there were to be a final analysis of the role of ML and AI in microbial genomics, no doubt it would re-emphasize the enduring importance of deductive statistical thinking, currently less fashionable as the new opportunities presented by ML take precedence. Statistics provides a foundation for scientific thought, clarifying concepts like study design, randomization, replication, control, batch effects, mediation and confounding, causation, and correlation. Scientific progress is a continual process, so there will be no final analysis. Instead, we expect a gradual assimilation of recent developments in AI/ML together with well-established statistical approaches into a new and emerging field of Data Science.

Glossary 1: Terms in statistical inference and machine learning, generated by ChatGPT-4o and manually curated

Attribution algorithms

Methods that assign importance scores to input features by estimating their contribution to a model's prediction, often using gradients, perturbations, or local surrogate models (e.g., SHAP, LIME).

Ablation algorithms

Techniques that assess feature importance by systematically removing or masking input features and measuring the resulting impact on model performance or predictions.

Automation bias

The tendency for humans to over-rely on automated systems, such as ML models, even when they may be incorrect.

Batch effects

Non-biological variations introduced into data during different processing times, instruments, or sample batches, which can confound results.

Bias-variance trade-off

A fundamental concept that describes the balance between underfitting and over-fitting. High bias models are too simple and may miss patterns (underfitting), while high variance models are too complex and may capture noise as if it were signal (over-fitting). Optimal performance is achieved by balancing these two sources of error.

Biased sampling

Occurs when the sample used for training or testing a model is not representative of the overall population-of-interest, leading to biased, misleading, or non-generalizable results.

Black box

A term used to describe models (such as deep neural networks) that are complex and difficult to interpret, where the internal workings are not easily understood.

Computational efficiency

Refers to the amount of computational resources (time and memory) required to train and use a model. More efficient models can handle larger datasets or run faster.

Collinearity

Collinearity occurs when two or more features are highly correlated, meaning they share a linear relationship. This makes it difficult to estimate the unique contribution of each predictor, leading to unreliable estimation with high uncertainty.

Cross-validation

A technique for assessing how well a model generalizes to unseen data by partitioning the dataset into multiple subsets and training/testing the model on different combinations of these subsets.

Data generating process

The scientific and sampling mechanisms by which the observed data in a study were produced.

Data quality

Refers to the accuracy, completeness, and reliability of data, which directly affects the performance of statistical inference and ML.

Data vs algorithmic modelling

In Breiman's dichotomy, *data modelling* focuses on building models that capture the essence of the underlying data generating process, whereas *algorithmic modelling* focuses on flexible prediction algorithms that exploit the structure in the observed data, without making assumptions about the underlying data generating process.

Deep learning

A subset of ML that involves neural networks with many layers (deep architectures) used to model complex patterns in data, particularly useful for image, speech, and sequence tasks.

Deductive vs inductive reasoning

Deductive reasoning draws specific conclusions from general principles or theories, whereas *inductive reasoning* infers general patterns or rules from specific observations or data.

Domain-specific knowledge

Expert knowledge about the particular field or domain of application (e.g., genomics) that helps guide model development and interpretation of results.

Dropouts

A regularization technique commonly used in neural networks where random units (artificial neurons) are "dropped" or ignored during training to prevent over-fitting.

Early stopping rules

A technique used to stop training a model once its performance on a validation set starts to degrade, preventing over-fitting.

Empirical

Based on observation or experimentation rather than theory. Empirical data is gathered from real-world experiments or observations.

Ensemble methods

ML methods that combine the predictions of multiple models to improve accuracy and robustness. Common examples include Random Forest and Gradient Boosting.

Explainability

The ability to interpret and understand how a ML model makes decisions, particularly in complex or high-dimensional models.

False discovery rate (FDR)

The expected proportion of false positives among all rejected null hypotheses in multiple hypothesis testing.

Family-wise error rate (FWER)
The probability of making one or more false positive errors when performing multiple hypothesis tests.

Features
Individual measurable properties or characteristics of the data used to train a model; they are analogous to independent variables in traditional statistical terminology, representing the inputs used to predict or explain an outcome.

Hypothesis test
A statistical method used to determine whether there is enough evidence to reject a null hypothesis, usually based on the comparison of a test statistic to a critical value.

Interactions
Interactions occur when the effect of one feature on the outcome depends on, or is modified by, the value of another feature.

Interpretable machine learning
A branch of ML focused on developing models that provide human-understandable explanations for their predictions.

Interpretability, equality, and accountability
Important ethical considerations in ML that refer to the clarity of model outputs (*interpretability*), fairness across different groups (*equality*), and responsibility for decisions made by models (*accountability*).

Learning rate
A hyper-parameter that controls how much a ML model's weights are updated with respect to the gradient of the loss function during training.

Loss function
Quantifies the (lack of) quality of a model's performance relative to the biological aims. It guides the optimization process during training. Examples include the mean squared error, calculated between a prediction or estimate and the truth, and 0–1 loss, where a value of 1 indicates a misclassification error or a false positive. Usually it is *risk*, rather than loss, that is minimized.

Maximum likelihood estimate (MLE)
A method of estimating the parameters of a statistical model by maximizing the likelihood function, which measures how likely it is to observe the given data under different parameter values.

Maximum a posteriori estimate (MAP)
An estimation method that incorporates prior knowledge or beliefs about the parameters in addition to the likelihood of the data, often used in Bayesian statistics.

Non-linearity

Non-linearity refers to relationships between variables that cannot be adequately captured by a straight line. In a non-linear relationship, changes in one variable do not simply lead to proportional changes in another.

Outcomes

The target variables that a model aims to predict or explain; they are analogous to dependent variables in traditional statistical terminology, representing the outputs that depend on the input features.

Over-fitting

Occurs when a model learns not only the underlying pattern in the training data but also the noise, leading to poor performance on new, unseen data.

Parsimony

A principle that prefers simpler models over more complex ones when both explain the data equally well, often used interchangeably with Occam's Razor.

Probabilistic models

Models that incorporate uncertainty by assigning probabilities to different outcomes, useful for reasoning about uncertainty in data.

Python

A high-level programming language widely used in data science and ML due to its simplicity, extensive libraries and strong community support.

Regression vs classification

Regression models continuous outcomes, while *classification* models categorical outcomes, in both cases fitting observed or predicting new outcomes based on the features of other input data.

Regularization

A technique used to prevent over-fitting by incorporating a penalty on the parameter values into the loss function (e.g., L1, L2 regularization).

Risk

The expected value of a *loss function*, defined as an arithmetic mean over (i) observed datapoints (*empirical risk*), (ii) a prior distribution (*Bayes risk*), or (iii) hypothetical repetitions of the data generating process (*frequentist risk*). Usually it is risk, not loss, that can be minimized by model fitting.

Supervised vs unsupervised learning

In *supervised learning*, a statistical or ML algorithm is trained on labelled outcome data, whereas in *unsupervised learning*, the algorithm learns from unlabelled data, discovering patterns without explicit outcomes.

Training, testing, and validation

A *training set* comprises data used to fit or train a model. A *validation set* is a separate subset of data used to tune model parameters and assess performance during training, where necessary. The *test set* is another, separate set of data used to evaluate the model's performance after training is complete.

Acknowledgements
Not applicable.

Peer review information
Claudia Feng was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
NA, SKS and DJW conceived the idea and shaped the structure of the manuscript. NA, SKS, DWE and DJW wrote the manuscript. NA conducted the original literature review and assembled the figures. All authors read and approved the final manuscript.

Funding
SKS was supported by an Ineos Oxford Institute grant; Wellcome Trust grant 088786/C/09/Z, and UKRI grants MR/L015080/1, MR/V001213/1, MR/S009264/1, and MR/T030062/1. NA was supported by a BBSRC scholarship BB/M011224/1. DWE was supported by the NIHR Oxford Biomedical Research Centre, the NIHR Health Protection Research Unit in Healthcare Associated Infection and Antimicrobial Resistance and by a Robertson Fellowship. DJW was supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society 101237/Z/13/B and by a Robertson Fellowship. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.
Ineos Oxford Institute, Wellcome (088786/C/09/Z, 101237/Z/13/B), UK Research and Innovation (MR/L015080/1), Biotechnology and Biological Sciences Research Council (BB/M011224/1), NIHR Oxford Biomedical Research Centre, NIHR Health Protection Research Unit in Healthcare Associated Infection and Antimicrobial Resistance, Robertson Foundation, Royal Society (101237/Z/13/B).

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 28 October 2024 Accepted: 3 September 2025
Published online: 27 September 2025

References

1. Blackwell GA, Hunt M, Malone KM, Lima L, Horesh G, Alako BTF, et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. PLoS Biol. 2021;19:e3001421 (Hanage WP, editor.).
2. Wong ZSY, Zhou J, Zhang Q. Artificial intelligence for infectious disease big data analytics. Infect Dis Health. 2019;24:44–8.
3. Ow GS, Tang Z, Kuznetsov VA. Big data and computational biology strategy for personalized prognosis. Oncotarget. 2016;7:40200–20.
4. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the Opportunities and Risks of Foundation Models. arXiv; 2021 Available from: https://arxiv.org/abs/2108.07258. [cited 2025 Sept 2].
5. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630:493–500.
6. Pagès-Gallego M, De Ridder J. Comprehensive benchmark and architectural analysis of deep learning models for nanopore sequencing basecalling. Genome Biol. 2023;24:71.
7. Torres MDT, Brooks EF, Cesaro A, Sberro H, Gill MO, Nicolaou C, et al. Mining human microbiomes reveals an untapped source of peptide antibiotics. Cell. 2024;187:5453-5467.e15.
8. Wan F, Torres MDT, Peng J, De La Fuente-Nunez C. Deep-learning-enabled antibiotic discovery through molecular de-extinction. Nat Biomed Eng. 2024;8:854–71.
9. Iwashyna TJ, Liu V. What's So Different about Big Data?. A Primer for Clinicians Trained to Think Epidemiologically. Annals ATS. 2014;11:1130–5.

10. Murphy KP. Probabilistic machine learning: an introduction. Cambridge, Massachusetts: The MIT Press; 2022.
11. Murphy KP. Probabilistic machine learning: advanced topics. Cambridge, Massachusetts: The MIT Press; 2023.
12. Breiman L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). Statist Sci. 2001;16. Available from: https://projecteuclid.org/journals/statistical-science/volume-16/issue-3/Statistical-Modeling--The-Two-Cultures-with-comments-and-a/10.1214/ss/1009213726.full. [cited 2025 Sept 2].
13. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15:233–4.
14. Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw. 2015;61:85–117.
15. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O. Scikit-learn: Machine learning in Python. J Mach Learn Res. 2011;12:2825–30.
16. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Advances in Neural Information Processing Systems 32. Curran Associates, Inc; 2019;8024–35.
17. TensorFlow Developers. TensorFlow. Zenodo; 2024. Available from: https://zenodo.org/doi/10.5281/zenodo.12726004. [cited 2025 Sept 2].
18. Greene AC, Giffin KA, Greene CS, Moore JH. Adapting bioinformatics curricula for big data. Brief Bioinform. 2016;17:43–50.
19. Wiemken TL, Kelley RR. Machine learning in epidemiology and health outcomes research. Annu Rev Public Health. 2020;41:21–36.
20. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Adv Neural Inf Process Syst. 2020;33:1877–901.
21. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of human migrations in *Helicobacter pylori* populations. Science. 2003;299:1582–5.
22. Corander J, Marttinen P. Bayesian identification of admixture events using multilocus molecular markers. Mol Ecol. 2006;15:2833–43.
23. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. Fast hierarchical Bayesian analysis of population structure. Nucleic Acids Res. 2019;47:5539–49.
24. Lees JA, Tonkin-Hill G, Yang Z, Corander J. Mandrake: visualizing microbial population structure by embedding millions of genomes into a low-dimensional representation. Phil Trans R Soc B. 2022;377:20210237.
25. Jaillard M, Lima L, Tournoud M, Mahé P, Van Belkum A, Lacroix V, et al. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. Didelot X, editor. PLoS Genet. 2018;14:e1007758.
26. Hoffman S, Podgurski A. Big bad data: law, public health, and biomedical databases. J Law Med Ethics. 2013;41:56–60.
27. Wang Q, Ma Y, Zhao K, Tian Y. A comprehensive survey of loss functions in machine learning. Ann Data Sci. 2022;9:187–212.
28. Stone M. Cross-Validatory Choice and Assessment of Statistical Predictions. J Royal Statistic Soc Series B (Methodological. 1974;36:111–47.
29. Bzdok D, Krzywinski M, Altman N. Machine learning: a primer. Nat Methods. 2017;14:1119–20.
30. Bashir D, Montañez GD, Sehra S, Segura PS, Lauw J. An Information-T. Cham: Springer International Publishing; 2020; 347–58. Available from: https://link.springer.com/10.1007/978-3-030-64984-5_27. [cited 2025 Sept 2].
31. Fix E, Hodges JL. Discriminatory analysis: Nonparametric discrimination: Consistency properties: (471672008–001). 1951 Available from: https://doi.apa.org/doi/10.1037/e471672008-001. [cited 2025 Sept 2].
32. Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory. 1967;13:21–7.
33. Yao Z, Ruzzo WL. A regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. BMC Bioinformatics. 2006;7:S11.
34. Mihelčić M, Šmuc T, Supek F. Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype. Sci Rep. 2019;9:19537.
35. Xu S. Bayesian naïve Bayes classifiers to text classification. J Inf Sci. 2018;44:48–59.
36. John GH, Langley P. Estimating Continuous Distributions in Bayesian Classifiers. arXiv; 2013 Available from: https://arxiv.org/abs/1302.4964. [cited 2025 Sept 2].
37. Webb GI. Naïve Bayes. In: Sammut C, Webb GI, editors. Encyclopedia of Machine Learning. Boston, MA: Springer US; 2011713–4. Available from: https://link.springer.com/10.1007/978-0-387-30164-8_576. [cited 2025 Sept 2].
38. Li F, Shen Y, Lv D, Lin J, Liu B, He F, et al. A bayesian classification model for discriminating common infectious diseases in Zhejiang province, China. Medicine. 2020;99:e19218.
39. Zhao Z, Cristian A, Rosen G. Keeping up with the genomes: efficient learning of our increasing knowledge of the tree of life. BMC Bioinformatics. 2020;21:412.
40. Sandberg R, Winberg G, Bränden C-I, Kaske A, Ernberg I, Cöster J. Capturing whole-genome characteristics in short sequences using a naïve Bayesian classifier. Genome Res. 2001;11:1404–9.
41. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4:e1000173 (Lewitter F, editor.).
42. McIntyre ABR, Ounit R, Afshinnekoo E, Prill RJ, Hénaff E, Alexander N, et al. Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. Genome Biol. 2017;18:182.
43. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20:273–97.
44. Tsirigos A. A sensitive, support-vector-machine method for the detection of horizontal gene transfers in viral, archaeal and bacterial genomes. Nucleic Acids Res. 2005;33:3699–707.
45. Weimann A, Mooren K, Frank J, Pope PB, Bremges A, McHardy AC. From Genomes to Phenotypes: Traitar, the Microbial Trait Analyzer. Segata N, editor. mSystems. 2016;1:e00101–16.
46. Belman S, Pesonen H, Croucher NJ, Bentley SD, Corander J. Estimating Between Country Migration in Pneumococcal Populations. Epidemiology; 2023. Available from: http://medrxiv.org/lookup/doi/10.1101/2023.11.15.23298520. [cited 2025 Sept 2].

47. Lupolova N, Dallman TJ, Holden NJ, Gally DL. Patchy promiscuity: machine learning applied to predict the host specificity of Salmonella enterica and Escherichia coli. Microbial Genomics. 2017;3. Available from: https://www.microbiologyresearch.org/content/journal/mgen/10.1099/mgen.0.000135. [cited 2025 Sept 2].
48. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1:81–106.
49. Li M, Xu H, Deng Y. Evidential decision tree based on belief entropy. Entropy. 2019;21:897.
50. Schrider DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. Trends Genet. 2018;34:301–12.
51. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
52. Statnikov A, Henaff M, Narendra V, Konganti K, Li Z, Yang L, et al. A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome. 2013;1:11.
53. Deneke C, Rentzsch R, Renard BY. Paprbag: a machine learning approach for the detection of novel pathogens from NGS data. Sci Rep. 2017;7:39194.
54. Méric G, Mageiros L, Pensar J, Laabei M, Yahara K, Pascoe B, et al. Disease-associated genotypes of the commensal skin bacterium *Staphylococcus epidermidis*. Nat Commun. 2018;9:5034.
55. Mageiros L, Méric G, Bayliss SC, Pensar J, Pascoe B, Mourkas E, et al. Genome evolution and the emergence of pathogenicity in avian *Escherichia coli*. Nat Commun. 2021;12:765.
56. Chen ML, Doddi A, Royer J, Freschi L, Schito M, Ezewudo M, et al. Beyond multidrug resistance: leveraging rare variants with machine and statistical learning models in *Mycobacterium tuberculosis* resistance prediction. EBioMedicine. 2019;43:356–69.
57. Li Y, Metcalf BJ, Chochua S, Li Z, Gertz RE, Walker H, et al. Validation of β-lactam minimum inhibitory concentration predictions for pneumococcal isolates with newly encountered penicillin binding protein (PBP) sequences. BMC Genomics. 2017;18:621.
58. Arning N, Sheppard SK, Bayliss S, Clifton DA, Wilson DJ. Machine learning to predict the source of campylobacteriosis using whole genome data. PLoS Genet. 2021;17:e1009436 (Hughes D, editor.).
59. Pascoe B, Futcher G, Pensar J, Bayliss SC, Mourkas E, Calland JK, et al. Machine learning to attribute the source of *Campylobacter* infections in the United States: a retrospective analysis of national surveillance data. J Infect. 2024;89:106265.
60. Wheeler NE, Gardner PP, Barquist L. Machine learning identifies signatures of host adaptation in the bacterial pathogen *Salmonella enterica*. PLoS Genet. 2018;14:e1007333 (Didelot X, editor.).
61. Zhang S, Li S, Gu W, Den Bakker H, Boxrud D, Taylor A, et al. Zoonotic Source Attribution of *Salmonella enterica* Serotype Typhimurium Using Genomic Surveillance Data, United States. Emerg Infect Dis. 2019;25. Available from: http://wwwnc.cdc.gov/eid/article/25/1/18-0835_article.htm. [cited 2025 Sept 2].
62. Beavan AJS, Domingo-Sananes MR, McInerney JO. Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome. Proc Natl Acad Sci USA. 2024;121:e2304934120.
63. Mason L, Baxter J, Bartlett P, Frean M. Boosting Algorithms as Gradient Descent. Advances in Neural Information Processing Systems. MIT Press; 1999. Available from: https://proceedings.neurips.cc/paper/1999/hash/96a93ba89a5b5c6c226e49b88973f46e-Abstract.html.
64. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Statist. 2001;29. Available from: https://projecteuclid.org/journals/annals-of-statistics/volume-29/issue-5/Greedy-function-approximation-A-gradient-boosting-machine/10.1214/aos/1013203451.full. [cited 2025 Sept 2].
65. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc; 2017;3149–57 17.
66. Anahtar MN, Yang JH, Kanjilal S. Applications of Machine Learning to the Problem of Antimicrobial Resistance: an Emerging Model for Translational Research. McAdam AJ, editor. J Clin Microbiol. 2021;59:e01260–20.
67. Ramoneda J, Stallard-Olivera E, Hoffert M, Winfrey CC, Stadler M, Niño-García JP, et al. Building a genome-based understanding of bacterial pH preferences. Sci Adv. 2023;9:eadf8998.
68. Hopfield JJ. Neural networks and physical systems with emergent collective computational abilities. Proc Natl Acad Sci U S A. 1982;79:2554–8.
69. Sheehan S, Song YS. Deep Learning for Population Genetic Inference. Chen K, editor. PLoS Comput Biol. 2016;12:e1004845.
70. Li Y, Huang C, Ding L, Li Z, Pan Y, Gao X. Deep learning in bioinformatics: introduction, application, and perspective in the big data era. Methods. 2019;166:4–21.
71. Sejnowski TJ. The Deep Learning Revolution. The MIT Press; 2018 Available from: https://direct.mit.edu/books/book/4111/The-Deep-Learning-Revolution. [cited 2025 Sept 2].
72. Lugo L, Hernández EB. A recurrent neural network approach for whole genome bacteria identification. Appl Artif Intell. 2021;35:642–56.
73. Hasan MA, Lonardi S. Deeplyessential: a deep neural network for predicting essential genes in microbes. BMC Bioinformatics. 2020;21:367.
74. Assaf R, Xia F, Stevens R. Detecting operons in bacterial genomes via visual representation learning. Sci Rep. 2021;11:2124.
75. Wiatrak M, Weimann A, Dinan A, Brbić M, Floto RA. Sequence-based modelling of bacterial genomes enables accurate antibiotic resistance prediction. Microbiology; 2024 Available from: http://biorxiv.org/lookup/doi/10.1101/2024.01.03.574022. [cited 2025 Sept 2].
76. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. Neural Netw. 1989;2:359–66.
77. Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. Understanding deep learning requires rethinking generalization. arXiv; 2016. Available from: https://arxiv.org/abs/1611.03530. [cited 2025 Sept 2].
78. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in Neural Information Processing Systems. 2017;30.

79. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 2022;35:27730–44.
80. Holz HJ, Loew MH. Relative feature importance: A classifier-independent approach to feature selection. Machine Intelligence and Pattern Recognition. Elsevier; 1994;473–87. Available from: https://linkinghub.elsevier.com/retrieve/pii/B9780444818928500468. [cited 2025 Sept 2].
81. Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. Proc Natl Acad Sci USA. 2019;116:22071–80.
82. House of Commons Science, Innovation and Technology Committee. 2023. The governance of artificial intelligence: interim report. Ninth Report of Session 2022–23. HC1769. https://committees.parliament.uk/publications/41130/documents/205611/default/
83. Nielsen EM, Fussing V, Engberg J, Nielsen NL, Neimann J. Most *Campylobacter* subtypes from sporadic infections can be found in retail poultry products and food animals. Epidemiol Infect. 2006;134:758–67.
84. Garrett N, Devane ML, Hudson JA, Nicol C, Ball A, Klena JD, et al. Statistical comparison of *Campylobacter jejuni* subtypes from human cases and environmental sources: comparison of *Campylobacter* subtypes. J Appl Microbiol. 2007;103:2113–21.
85. Wilson DJ, Gabriel E, Leatherbarrow AJH, Cheesbrough J, Gee S, Bolton E, et al. Tracing the Source of Campylobacteriosis. Guttman DS, editor. PLoS Genet. 2008;4:e1000203.
86. Sheppard SK, Dallas JF, Strachan NJC, MacRae M, McCarthy ND, Wilson DJ, et al. *Campylobacter* genotyping to determine the source of human infection. Clin Infect Dis. 2009;48:1072–8.
87. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016;785–94. Available from: https://dl.acm.org/doi/10.1145/2939672.2939785. [cited 2025 Sept 2].
88. Mackay TFC. The genetic architecture of quantitative traits. Annu Rev Genet. 2001;35:303–39.
89. Peacock SJ, Moore CE, Justice A, Kantzanou M, Story L, Mackie K, et al. Virulent combinations of adhesin and toxin genes in natural populations of *Staphylococcus aureus*. Infect Immun. 2002;70:4987–96.
90. Astle W, Balding DJ. Population Structure and Cryptic Relatedness in Genetic Association Studies. Statist Sci. 2009;24. Available from: https://projecteuclid.org/journals/statistical-science/volume-24/issue-4/Population-Structure-and-Cryptic-Relatedness-in-Genetic-Association-Studies/10.1214/09-STS307.full. [cited 2025 Sept 2].
91. Price AL, Zaitlen NA, Reich D, Patterson N. New approaches to population stratification in genome-wide association studies. Nat Rev Genet. 2010;11:459–63.
92. Sheppard SK. Strain wars and the evolution of opportunistic pathogens. Curr Opin Microbiol. 2022;67:102138.
93. Pearl J. Causal inference in statistics: An overview. Statist Surv. 2009;3. Available from: https://projecteuclid.org/journals/statistics-surveys/volume-3/issue-none/Causal-inference-in-statistics-An-overview/10.1214/09-SS057.full. [cited 2025 Sept 2].
94. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. Causal associations between risk factors and common diseases inferred from GWAS summary data. Nat Commun. 2018;9:224.
95. Sheppard SK, Didelot X, Meric G, Torralbo A, Jolley KA, Kelly DJ, et al. Genome-wide association study identifies vitamin B_5 biosynthesis as a host specificity factor in *Campylobacter*. Proc Natl Acad Sci USA. 2013;110:11923–7.
96. Earle SG, Wu C-H, Charlesworth J, Stoesser N, Gordon NC, Walker TM, et al. Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol. 2016;1:16041.
97. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Stegle O, editor. Bioinformatics. 2018;34:4310–2.
98. Young BC, Earle SG, Soeng S, Sar P, Kumar V, Hor S, et al. Panton-valentine leucocidin is the key determinant of *Staphylococcus aureus* pyomyositis in a bacterial GWAS. Elife. 2019;8:e42486.
99. Earle SG, Lobanovska M, Lavender H, Tang C, Exley RM, Ramos-Sevillano E, et al. Genome-wide association studies reveal the role of polymorphisms affecting factor H binding protein expression in host invasion by Neisseria meningitidis. Nassif X, editor. PLoS Pathog. 2021;17:e1009992.
100. Green AG, Yoon CH, Chen ML, Ektefaie Y, Fina M, Freschi L, et al. A convolutional neural network highlights mutations relevant to antimicrobial resistance in *Mycobacterium tuberculosis*. Nat Commun. 2022;13:3817.
101. The CRyPTIC Consortium. Genome-wide association studies of global Mycobacterium tuberculosis resistance to 13 antimicrobials in 10,228 genomes identify new resistance mechanisms. Ladner J, editor. PLoS Biol. 2022;20:e3001755.
102. Mosquera-Rendón J, Moreno-Herrera CX, Robledo J, Hurtado-Páez U. Genome-wide association studies (GWAS) approaches for the detection of genetic variants associated with antibiotic resistance: a systematic review. Microorganisms. 2023;11:2866.
103. Didelot X, Bowden R, Wilson DJ, Peto TEA, Crook DW. Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 2012;13:601–12.
104. Walker TM, Cruz ALG, Peto TE, Smith EG, Esmail H, Crook DW. Tuberculosis is changing. Lancet Infect Dis. 2017;17:359–61.
105. Satta G, Lipman M, Smith GP, Arnold C, Kon OM, McHugh TD. *Mycobacterium tuberculosis* and whole-genome sequencing: how close are we to unleashing its full potential? Clin Microbiol Infect. 2018;24:604–9.
106. Jakobsdottir J, Gorin MB, Conley YP, Ferrell RE, Weeks DE. Interpretation of Genetic Association Studies: Markers with Replicated Highly Significant Odds Ratios May Be Poor Classifiers. Abecasis GR, editor. PLoS Genet. 2009;5:e1000337.
107. Yang Y, Niehaus KE, Walker TM, Iqbal Z, Walker AS, Wilson DJ, et al. Machine learning for classifying tuberculosis drug-resistance from DNA sequencing data. Birol I, editor. Bioinformatics. 2018;34:1666–71.
108. Kouchaki S, Yang Y, Walker TM, Sarah Walker A, Wilson DJ, Peto TEA, et al. Application of machine learning techniques to tuberculosis drug resistance analysis. Wren J, editor. Bioinformatics. 2019;35:2276–82.
109. Yang Y, Walker TM, Walker AS, Wilson DJ, Peto TEA, Crook DW, et al. DeepAMR for predicting co-occurrent resistance of *Mycobacterium tuberculosis*. Hancock J, editor. Bioinformatics. 2019;35:3240–9.

110. Gröschel MI, Owens M, Freschi L, Vargas R, Marin MG, Phelan J, et al. Gentb: A user-friendly genome-based predictor for tuberculosis resistance powered by machine learning. Genome Med. 2021;13:138.
111. The CRyPTIC Consortium and the 100,000 Genomes Project. Prediction of Susceptibility to First-Line Tuberculosis Drugs by DNA Sequencing. N Engl J Med. 2018;379:1403–15.
112. He G, Zheng Q, Shi J, Wu L, Huang B, Yang Y. Evaluation of WHO catalog of mutations and five WGS analysis tools for drug resistance prediction of *Mycobacterium tuberculosis* isolates from China. Georghiou SB, editor. Microbiol Spectr. 2024;12:e03341–23.
113. Ferrari E, Retico A, Bacciu D. Measuring the effects of confounders in medical supervised classification problems: the confounding index (CI). Artif Intell Med. 2020;103:101804.
114. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco California USA: ACM; 2016;1135–44. Available from: https://dl.acm.org/doi/10.1145/2939672.2939778. [cited 2025 Sept 2].
115. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. arXiv; 2017 Available from: https://arxiv.org/abs/1705.07874. [cited 2025 Sept 2].
116. Meyes R, Lu M, Waubert de Puiseau C, Meisen T. Ablation studies to uncover structure of learned representations in artificial neural networks. Proceedings of the International Conference on Artificial Intelligence (ICAI). Athens, Greece: CSREA Press; 2019 Available from: https://www.researchgate.net/publication/334871296_Ablation_Studies_to_Uncover_Structure_of_Learned_Representations_in_Artificial_Neural_Networks. [cited 2025 Sept 2].
117. Callaway E. How generative AI is building better antibodies. Nature. 2023;d41586–023–01516-w.
118. 118.Callaway E. 'ChatGPT for CRISPR' creates new gene-editing tools. Nature. 2024;629:272–272.
119. Tang X, Dai H, Knight E, Wu F, Li Y, Li T, et al. A survey of generative AI for de novo drug design: new frontiers in molecule and protein generation. Briefings in Bioinformatics. 2024;25:bbae338
120. Winnifrith A, Outeiral C, Hie BL. Generative artificial intelligence for de novo protein design. Current Opinion in Structural Biology. 2024;86:102794

Publisher's Note