npj | antimicrobials & resistance

Review

https://doi.org/10.1038/s44259-025-00085-4

How AI can help us beat AMR

Autumn Arnold[1,2,3,4], Stewart McLellan[1,2,3,4] & Jonathan M. Stokes[1,2,3]

Antimicrobial resistance (AMR) is an urgent public health threat. Advancements in artificial intelligence (AI) and increases in computational power have resulted in the adoption of AI for biological tasks. This review explores the application of AI in bacterial infection diagnostics, AMR surveillance, and antibiotic discovery. We summarize contemporary AI models applied to each of these domains, important considerations when applying AI across diverse tasks, and current limitations in the field.

The AMR crisis is one of the most pressing public health threats of this century[1]. The excessive application and incomplete use of antibiotics in clinical and agricultural settings have selected for bacteria that have evolved mechanisms to evade these molecules[2,3]. These resistance mechanisms disseminate rapidly through vertical and horizontal gene transfer (HGT), spreading through bacterial communities. The prevalence of multidrug-resistant isolates in clinical settings is rendering common infections more difficult to treat. Indeed, in 2019 AMR was associated with roughly 4.95 million deaths globally and is projected to result in 10 million deaths per year by 2050[1].

Global efforts to combat the AMR crisis via surveillance and new antibiotic discovery have resulted in the collection of datasets of diverse bacterial genomes, antibiotic susceptibility testing (AST), and chemical bioactivity screens. AI methods excel in handling large data volumes, making them a robust set of tools for extracting valuable insights from complex datasets. Machine learning (ML), a subset of AI, uses statistical algorithms to identify sophisticated relationships within datasets and extrapolate to new data. Deep learning (DL) methods, which are a subset of ML, employ neural networks to process data using layered, interconnected nodes. Such models are often trained on existing task-specific datasets and subsequently used to generalize to unseen data.

The ability to harness large datasets for AI-guided decision-making can revolutionize clinical diagnosis of bacterial infections, AMR surveillance, and antibiotic discovery[4,5]. For example, AI algorithms can learn from patient data in electronic health records (EHRs) to support real-time clinical decision-making[4]. Additionally, AI-guided surveillance systems can learn from genomic data across varying bacterial populations to uncover novel resistance mechanisms and enable public health authorities to implement targeted interventions and containment strategies. For drug discovery, AI has shown remarkable potential in identifying new antibiotic candidates by rapidly screening vast chemical libraries and predicting the efficacy and safety of potential candidates[6,7]. This expedites the discovery of new antibiotics and helps repurpose existing drugs to combat resistant bacteria. In this review, we explore the integration of AI across three critical domains essential for curbing the escalation of AMR: (1) clinical diagnostics, (2) AMR surveillance, and (3) antibiotic discovery. We offer a concise overview of recent AI advancements in these areas by highlighting key examples and exploring how they contribute to mitigating AMR threats and preserving public health. We further included a glossary (Table 1) containing brief descriptions of the ML terms and architectures discussed throughout this review.

AI in clinical diagnostics

Rapid diagnostic testing and treatment are essential for improving patient outcomes, particularly for cases of such diseases as bacterial sepsis, toxic shock syndrome, and bacterial meningitis, where early disease recognition, microbial identification, and AMR profile determination are essential[8–10]. The current gold standard methods used in clinical practice to diagnose bacterial infections include culture techniques for AST, nucleic acid-based tests (polymerase chain reaction [PCR] and 16S sequencing), and matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) for pathogen identification and AMR profiling when compared to known spectral databases.

Despite their widespread use, gold-standard methods have requirements that are sub-optimal in diagnostic situations where early detection is critical. For instance, bacterial culture methods are time-consuming, taking several days (or weeks in the case of tuberculosis) to yield results, and are ineffective at detecting viable but non-culturable cells[11]. PCR for bacterial identification and AMR profiling is constrained by the need for primers specific to known pathogens and resistance genes, which inherently limits detection to these predefined targets. Additionally, PCR assays may face inhibition challenges in complex clinical samples such as cerebrospinal fluid[12]. The effectiveness of MALDI-TOF MS is contingent on the comprehensiveness of reference databases, which may suffer from the underrepresentation of certain bacterial taxa. These issues primarily stem from limitations in data availability and collection, rather than the diagnostic methodologies themselves. However, AI methods offer advantages in streamlining diagnostic workflows, primarily by accelerating data analysis and enabling automated interpretation of results. This efficiency offers the potential to reduce turnaround times compared to traditional, labor-intensive clinical microbiology techniques.

[1]Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, ON, Canada. [2]Michael G. DeGroote Institute for Infectious Disease Research, McMaster University, Hamilton, ON, Canada. [3]David Braley Centre for Antibiotic Discovery, McMaster University, Hamilton, ON, Canada. [4]These authors contributed equally: Autumn Arnold, Stewart McLellan. e-mail: stokesjm@mcmaster.ca

Table 1 | Glossary of artificial intelligence terms and architectures discussed in this review

Term	Abbreviation	Description	Reference
Artificial intelligence	AI	A field of computer science focused on creating systems that can perform tasks requiring human intelligence.	
Machine learning	ML	A subset of AI focused on developing algorithms that enable computers to learn patterns and make decisions from data, improving their performance on a specific task through experience.	103
Deep learning	DL	A subset of machine learning that uses neural networks with multiple layers to learn hierarchical representations from large amounts of data, enabling the modeling of complex patterns.	104
Bidirectional long-short-term memory	Bi-LSTM	A network that processes sequential data in both the forward and backward directions to capture contextual information from both past and future states. In the context of EHRs, Bi-LSTMs process information from the forward (symptoms to diagnosis to treatment) and backward (treatment to diagnosis to symptoms) direction to capture temporal patterns.	105
Logistic regression	LR	A statistical model used for binary classification that estimates the probability of a binary outcome using a logistic function to transform a linear combination of input features into values between 0 and 1.	106
Random forest	RF	An ensemble of multiple decision trees that outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees.	107
Natural language processing	NLP	A field of AI that aims for machines to understand, interpret, and generate natural language in a useful and meaningful way.	108
Feed-forward neural network/multi-layer perceptron	FFNN/MLP	A neural network in which the flow of information is strictly forward, passing from the input layer, through one or more hidden layers, and lastly through an output layer.	109,110
Convolutional neural network	CNN	A neural network designed to process structured grid-like data such as images by using convolutional layers to learn spatial hierarchies or features.	111
Support vector machine	SVM	A supervised learning model that finds the optimal hyperplane to separate data into distinct classes with the maximum margin. This can be linear, or rely on kernel functions to project data into a higher dimensional space for more optimal separation, which is important for non-linear relationships.	112
Light gradient boosting machine	lightGBM	A gradient boosting framework comprised of an ensemble of histogram-based decision trees, in which the decision trees are built iteratively. Each decision tree is built using a leaf-wise (best node first) strategy to focus on the most impactful splits.	113
Linear regression		A statistical method for modeling the relationship between a dependent variable and one or more independent variables by fitting a linear equation to the observed data, where the goal is to minimize the difference between the predicted and actual values.	114
Linear support vector machine	LSVM	A type of SVM used to linearly separate data.	112
Graph neural network	GNN	A type of neural network that processes data structured as graphs, where nodes and edges represent entities and their relationships. This is achieved by iteratively aggregating and updating node features based on the graphs connectivity to learn both local and global representations.	67
Directed message passing neural network	D-MPNN	A type of GNN where messages are passed along edges in a fixed direction, capturing dependencies between nodes. This prevents unnecessary loops in the message passing process, which introduce noise into the representation.	115
Monte-Carlo tree search	MCTS	A search algorithm used for decision making, that builds a search tree by randomly simulating outcomes of actions and refining choices over time to find the best strategy, balancing exploration and exploitation. This strategy is useful in the exploration of larger chemical spaces, while retaining high prediction scores.	116
Variational autoencoder	VAE	A type of generative model that encodes input data into a latent space and then decodes samples from this latent space back to the original data distribution. This allows for the generation of new data with similar features to the training set.	117
Generative adversarial network	GAN	A generative model consisting of a generator and discriminator, which are trained simultaneously in a competitive setting. The generator creates realistic data while the discriminator attempts to distinguish between the real and generated samples, driving the generator to produces increasingly realistic outputs.	80
Junction tree variational autoencoder	JT-VAE	A type of VAE designed for molecular graph generation, which represents molecular structures as junction trees of chemically valid substructures, allowing for encoding and decoding of complex molecules while preserving their chemical validity.	76
Diffusion model		A class of generative models that learn to generate data by reversing a gradual noising process, where datapoints are progressively perturbed with noise (Gaussian) and the model is trained to iteratively denoise and recover the original data distribution.	118

AI for sepsis prediction

Timely sepsis recognition is crucial for early administration of antibiotics that significantly improve survival outcomes. For reference, each hour of delay in antibiotic treatment after sepsis onset increases mortality risk by 9%[13,14]. Bacterial sepsis, a life-threatening systemic response to infection, closely mimics non-infectious systemic inflammatory response syndrome, complicating rapid differentiation between septic and non-septic conditions[14,15]. The integration of AI applied to EHRs holds promise for enhancing the accurate diagnosis of sepsis[16]. These models learn the relationship between patient biomarkers and clinical data to distinguish bacterial sepsis from other systemic conditions, improving treatment decisions and reducing the risk of inappropriate drug use[17]. However, challenges exist in using EHR data, which often include difficult-to-interpret unstructured data (e.g., clinical notes) and structured data (e.g., lab results), which are measured at inconsistent times[18].

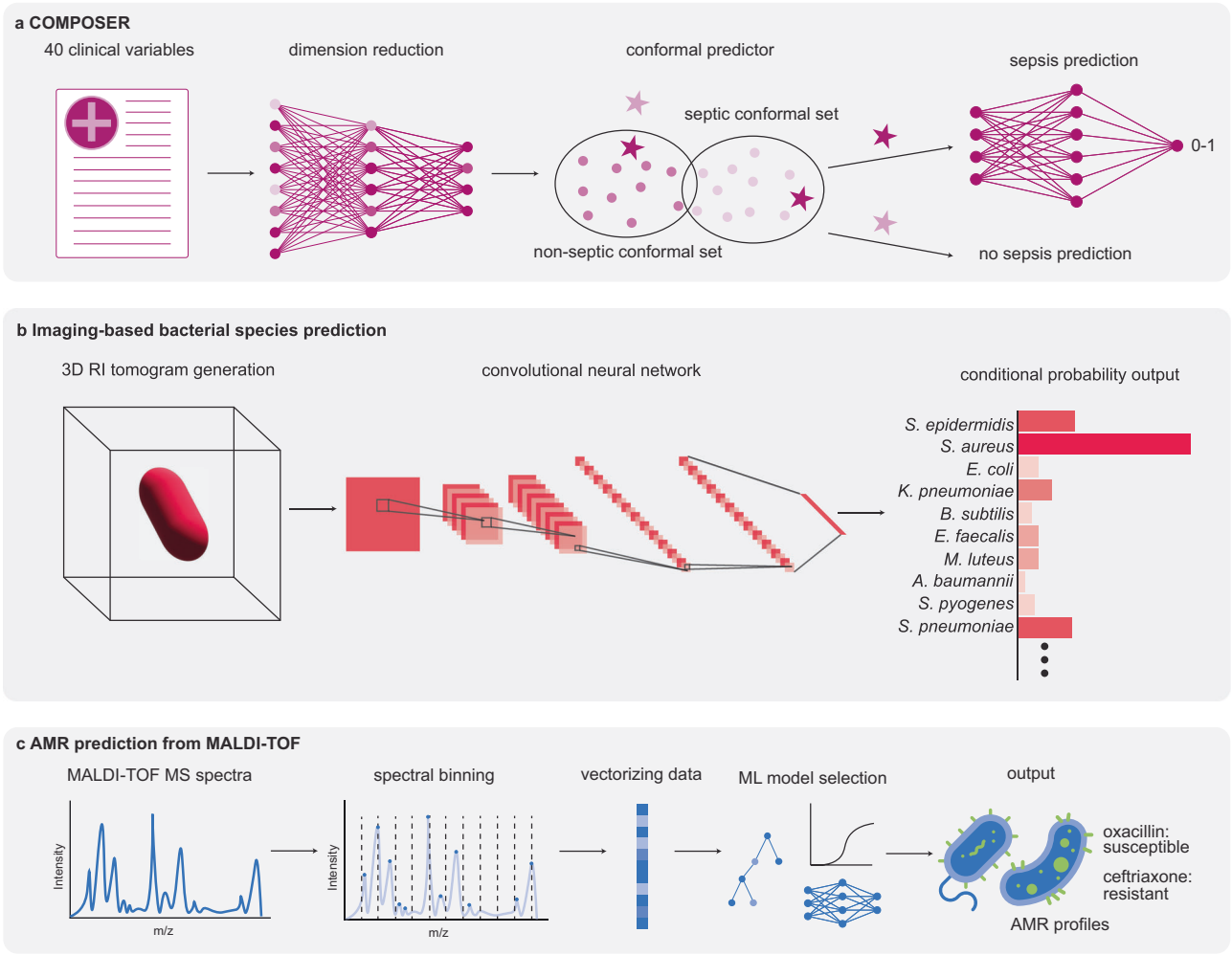


Fig. 1 | Contemporary AI models for bacterial infection diagnostics. a Conformal Multidimensional Prediction Of Sepsis Risk (COMPOSER) collects 40 clinical variables from EHRs, including dynamic variables such as vital signs and laboratory measurements, and demographic variables such as age and gender. This information is used for a weighted input layer that scales the value of a clinical variable dependent on the most recent measurement. This input layer is fed into a FFNN to reduce data dimensionality. The dimensionally reduced vector is fed into a conformal predictor that uses two conformal sets to quantify the conformity of new patient-level features to previously seen septic and non-septic examples. This enables the model to identify outliers not meeting the algorithm's conditions, assigning them to an indeterminate label class. If the data conforms, it is fed into a sepsis predictor (FFNN) that predicts the probability of sepsis from 0 to 1. **b** A 3D tomogram of a single cell or cell cluster is generated using 3D QPI tomography and assembly of 2D sinograms. The 3D tomogram is input into a CNN with four dense blocks containing 12, 24, 64, and 64 convolution operations which undergo batch normalization and average pooling between each block. The final output is a 19-dimensional vector that contains the conditional probability of the 3D RI tomogram being one of 19 BSI causing bacteria. **c** Model was trained on DRIAMS, which contains mass spectra and an AMR profile for each instance. A collected mass spectra is inputted into the model, which is then pre-processed and binned into 3 Da ranges from 2000 Da to 20,000 Da. These bins are vectorized, and inputted into the appropriate model (LR, lightBGM, or MLP) to generate a prediction of resistance to a certain antibiotic.

To address the issue of irregular time measurements from structured data in EHRs, Zhang et al. prioritized the most recent clinical measurements in EHRs to predict the risk of sepsis using a bidirectional long short-term memory (BiLSTM). BiLSTMs are designed to work with sequential data in forward and reverse directions by maintaining an internal memory state to retain contextual information from previous inputs[19]. They can integrate information from past and future events to make predictions across time steps[20]. Clinical features such as red blood cell count and body temperature are initially processed by an attention layer that dynamically weights their importance for prediction[20]. This improves the interpretability of model predictions, which is crucial for healthcare professionals to trust AI in clinical settings. These weighted features are concatenated and passed through an embedding layer with time encodings (the time at which the measurement was taken) to handle irregular timing intervals[20]. The embeddings are then fed into a BiLSTM, from which the outputs are aggregated and processed by a fully connected layer to yield the probability of sepsis[20]. Zhang et al. trained their model on data from ~180,000 patient EHR records from over 600 hospitals across the USA (Cerner Health Facts Database), and achieved an AUC of 0.94[20,21]. Importantly, this database contains a broad range of EHRs of different genders, ages, and demographics, which allows for robust predictive performance across these subpopulations, a pertinent feature that is absent in models that restrict their data to one hospital or healthcare system[20].

The use of unstructured EHR data, such as clinical text and images, has been challenging; it can be a time-consuming process to convert this information into a usable format for AI model training and inference[22]. To overcome this, the Sepsis Early Risk Assessment (SERA) algorithm was developed to integrate clinical notes with structured EHR data, enhancing sepsis risk prediction[17]. SERA's natural language processing component applies the latent Dirichlet allocation algorithm—a statistical model that identifies underlying topics within text by analyzing word co-occurrences and distribution across text[17]. This latent Dirichlet allocation algorithm was used to extract 100 common clinical topics from ~5000 de-identified EHRs comprised of ~115,000 clinical notes extracted from the Singapore Epic

System[17]. These clinical topics are assigned a weight that represents the strength or contribution of a topic in a given document, and combined with structured data for sepsis prediction (an ensemble of logistic regression [LR] and random forest [RF] models)[17]. Briefly, LR predicts the probability of a binary outcome by modeling the relationship between one or more independent variables and the target variable using a logistic (sigmoid) function, and RF models use multiple decision trees for classification tasks. If sepsis is not present at prediction time, SERA estimates the risk of onset within 4–48 h[17]. Incorporating topic mining from clinical notes significantly improved SERA's prediction accuracy relative to strictly using structured data, particularly 12–48 h before sepsis onset, outperforming standard tools like the modified early warning system and sequential organ failure assessment[17]. Sepsis prediction 12 h before onset using only structured data produced an AUC of 0.79, which is increased to 0.94 with the addition of unstructured clinical notes, enhancing the SERA's predictive capabilities[17]. However, this approach may limit the generalizability of SERA, since it may not be compatible with different languages and can vary based on the individual note-taking practices of different clinicians[17].

COMPOSER (COnformal Multidimensional Prediction Of SEpsis Risk) is a DL approach for early sepsis prediction, addressing generalizability issues such as data distribution shifts and missing data across hospital sites[23]. COMPOSER was trained on over 100,000 positive EHR records of sepsis and over 2,000,000 examples of non-septic patients[23]. COMPOSER's architecture consists of three modules: initially, it employs a feedforward neural network (FFNN) to generate representations from clinical and timing data, reducing discrepancies caused by varying hospital practices[23]. A conformal predictor that determines the probability distribution is then used to identify out-of-distribution samples by validating new patient data against established septic and non-septic patterns from the training set, enhancing model reliability by only predicting on patient data if the data distribution matches that of the training data[23]. The final module uses another FFNN, in which the final layer is a function that outputs a sepsis risk score between 0 and 1 (Fig. 1a)[23]. These three modules facilitate accurate sepsis predictions, achieving AUROC scores of 0.953 in intensive care units and 0.945 in emergency department settings[23]. COMPOSER's implementation in the UC San Diego Hospital System resulted in a 17% relative decrease in in-hospital mortality and a 10% increase in sepsis bundle compliance, showcasing significant potential in clinical settings[24]. However, as AI becomes more integrated into healthcare, potential issues like data distribution shifts, poor generalization to new populations, and discriminatory bias must be carefully considered. These factors can adversely affect patient outcomes and limit the utility of AI for underrepresented groups[25].

AI for bacterial identification

Spectroscopy and image-based techniques are emerging for rapidly detecting and identifying bacteria from minimal samples, sometimes bypassing the need for culturing. Spectroscopy methods generate spectral fingerprints, which are unique to each species, enabling their differentiation[26]. While imaging bacterial cells has long been a fundamental technique in microbiology, recent advancements in imaging technology and applicable AI methods have significantly enhanced the performance of bacterial identification.

Raman spectroscopy measures the energy shifts of scattered electrons relative to the energy of photons from a monochromatic light source, reflecting the chemical bonds within a sample. This generates a spectral fingerprint since the Raman shift correlates directly with the interacting molecules (which are cellular contents in this context)[27]. To interpret subcellular components, multiple Raman bands, each characterized by distinct positions, heights, and widths, are analyzed. These bands encompass thousands of variables based on wavenumbers, presenting a challenge due to their high dimensionality[28]. Lu et al. trained a convolutional neural network (CNN), which processes structured grid data through convolutional layers to capture hierarchical patterns to identify 14 microbial species. In this case, their training set was comprised of ~300 bacterial Ramanomes (~100 Raman spectra) per species, which were cultured under various growth conditions to account for intraspecies variability. Given the Ramanome for a query bacterium, the model will output the conditional probability for all 14 species, of which the highest probability is the predicted species[28]. The average accuracy across the 14 species was 95%, of which the highest accuracy for a single species was *E. coli* (99%), and the lowest accuracy was for *P. aeruginosa* (81%)[28]. Although this approach has not yet been implemented in clinical practice, it demonstrates significant potential to expedite the diagnosis of bacterial and fungal infections. The method's reliance on a single microbial cell and its adaptability to diverse growth conditions—including both rich and minimal media, as well as various growth phases—eliminates the need for extensive culturing. This robustness ensures reliable results regardless of the growth environment, offering a tool for rapid, culture-independent microbial identification.

Along with spectroscopy, CNNs have been used with imaging techniques that demonstrated rapid microbial identification by integrating three-dimensional (3D) quantitative phase imaging (QPI) with image classification[29]. This approach involves generating 3D refractive index (RI) tomograms by applying optical diffraction principles to combine sinograms from 2D QPI images[30]. The 3D RI tomograms are then classified into one of 19 species responsible for bloodstream infections using a CNN trained on ~9000 tomograms, split between each species[29]. The CNN output provides a conditional probability for each species based on the input tomogram and the training data distribution, achieving an 82.5% accuracy when predicting on an individual bacterial cell or cluster (Fig. 1b)[29].

Species identification using these techniques is comparable to MALDI-TOF MS, but it significantly reduces the number of bacteria required, from over 10^5 CFUs to single or few cells, thus eliminating long culturing times[31]. This reduction in culturing can expedite turnaround time, enabling earlier antibiotic treatment. However, despite these advancements, such imaging and spectroscopy techniques require rigorous standardization in sample preparation and necessitate robust databases to accurately differentiate species.

AI for AMR prediction

Rapidly identifying the causative microbe and their resistance profile is crucial for effective infection management, as early detection allows for targeted therapeutic strategies, leveraging knowledge of intrinsic resistance mechanisms and local epidemiological data. Yet, current culture-based methods can delay detailed resistance profiling by up to 72 h, potentially resulting in suboptimal antibiotic administration prior to receiving clinical microbiology lab results[32,33]. Accelerating resistance profiling substantially improves patient outcomes and abides by the goals of antimicrobial stewardship (providing optimal bacterial infection treatment while promoting sustainable use of antibiotics to prevent AMR) by minimizing broad-spectrum and last-resort antibiotic use[34].

Whole-genome sequencing (WGS) has been employed to identify resistance markers from clinical isolates[35]. When used with AI methods, WGS has been further leveraged to detect genotypic markers linked to antibiotic resistance profiles[36–38]. Ren et al. sought to determine the best combination of encoding methods for WGS data and ML architectures to effectively predict AMR using bacterial genomic information[36]. WGS data for ~1000 *E. coli* strains resistant to ciprofloxacin, cefotaxime, ceftazidime, or gentamicin was aligned to wild-type *E. coli* to identify single-nucleotide polymorphisms (SNPs)[36]. The SNPs and resistance profiles were encoded using one-hot encoding (binary), label encoding (A, G, C, T, N; 1, 2, 3, 4, 0), or frequency matrix chaos game representation (FCGR), which converts SNP data into an image-like matrix[36]. Each encoding was applied to train four models: support vector machine (SVM; used to identify the optimal hyperplane for class separation), LR, RF, and a CNN, resulting in a total of twelve different architectures[36]. The twelve models were trained using a five-fold stratified cross-validation approach, incorporating upsampling to mitigate dataset imbalances[36]. The choice of encoding method did not significantly influence the AUROC for any model[36]. Notably, the RF consistently outperformed the other models in predicting resistance to all four

antibiotics, as evidenced by its superior AUROC values[36]. To further assess model robustness, an external validation was performed using a publicly available dataset of approximately 1500 *E. coli* strains. These results reinforced the findings, with the RF model using label encoding achieving the highest AUROC (0.95) for identifying *E. coli* strains resistant to ciprofloxacin among the twelve model configurations[36].

While genotypic changes, such as SNPs, can imply resistance phenotypes, such changes can be observed through phenotypic approaches directly. The Deep Antimicrobial Susceptibility Phenotyping platform uses CNNs to analyze micrographs for classifying bacteria as either susceptible or resistant to antibiotics[39]. Briefly, the CNN was trained on ~29,000 manually annotated *E. coli* cells segmented from 459 microscopy fields of view, stained with DAPI and Nile Red, and balanced across antibiotics for classification training. In this workflow, *E. coli* is treated with antibiotics (ciprofloxacin, gentamicin, co-amoxiclav, or rifampicin), inducing phenotypic changes in susceptible cells, while resistant cells retain their original morphology[39]. Cellular components are fluorescently stained—DAPI for the nucleoid and Nile red for the cell membrane—and imaged using fluorescence microscopy[39]. A CNN first isolates single cells from the image using Nile red staining as a guide, followed by a second CNN that classifies each cell as susceptible or resistant based on alterations in the nucleoid and membrane structure[39]. Separate models were trained for each antibiotic, achieving a minimum accuracy of 84%[39]. Although the total workflow can generate results within 30 min (the majority of which is sample preparation and antibiotic treatment), they still require culturing methods to obtain a sufficient number of cells for imaging[39].

Another phenotypic approach involves MALDI-TOF MS, widely adopted for rapid microbial identification and increasingly applied to assess antimicrobial susceptibility, thus improving clinical decision-making[40]. A key advancement in this space is the creation of the Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra (DRIAMS), which contains 303,195 mass spectra associated with 768,300 antimicrobial resistance (AMR) labels across 803 bacterial and fungal species[41]. This extensive dataset has been instrumental in training DL models for AMR prediction[42–45]. For instance, Weis et al. used the DRIAMS dataset, binning each mass spectrum into 3 Da intervals between 2000 and 20,000 Da, creating a 6000-dimensional vector concatenated with species and AMR labels[43]. These vectors were used to train various ML models including LR, lightGBM (gradient boosting framework built with decision trees), and multi-layer perceptron (MLP), to predict antibiotic susceptibility for specific pathogen-antibiotic pairs (Fig. 1c)[43]. Notably, the lightGBM models showed reasonable predictive performance, achieving AUROCs of 0.8 for oxacillin-resistant *Staphylococcus aureus* (MRSA) and 0.74 for ciprofloxacin-resistant *E. coli*[43].

In a retrospective analysis of 63 clinical cases from the same institution where the training data were derived, the integration of the ML-based AMR prediction model into clinical decision-making would have led to changes in the antibiotic regimen in nine cases[43]. Of these, the model correctly identified and recommended adjustments in eight cases, aligning with susceptibility profiles and showcasing an 89% success rate[43]. This highlights the model's potential to optimize antimicrobial therapy, facilitating targeted treatment while enhancing antimicrobial stewardship efforts. However, it should be noted that these models exhibited reduced predictive performance when applied to data from external sites, reflecting challenges in generalizing across different hospitals due to variations in microbial populations, resistance prevalence, and mass spectra parameters.

These observations underscore the potential of AI-enhanced diagnostics to improve and expedite AMR prediction, contributing to improved patient care. However, significant challenges remain, since these models must demonstrate exceptionally high performance to ensure reliable clinical implementation. False positives may result in the use of unnecessary last-resort treatments, whereas false negatives can result in persistent infections by allowing antibiotic-resistant bacteria to evade appropriate treatment. Additionally, a current limitation of these methods is the need for distinct models for each antibiotic-pathogen combination, restricting their applicability in cases where data for certain combinations are limited.

AMR surveillance

Monitoring the dissemination of AMR determinants and the emergence of novel resistance mechanisms is critical to develop evidence-based antibiotic stewardship guidelines. Indeed, it is important to adopt a One Health approach to the AMR response that considers the complex interactions between human health and the environment[34]. AMR surveillance in clinical settings can be used to advise healthcare professionals and monitor health outcomes from various infections[46]. Furthermore, monitoring AMR in agricultural settings can help ensure food safety[46]. Likewise, surveilling AMR in environmental settings (e.g., wastewater) can provide insight into transmission patterns and identify novel resistance genotypes/phenotypes that may eventually emerge in clinical settings[46].

AMR surveillance was classically achieved through phenotypic approaches, specifically AST[46]. Now, the collection of WGS data with associated AMR phenotypes from global surveillance efforts, as well as annotated AMR genes (ARGs) from databases such as CARD (Comprehensive Antibiotic Resistance Database) and ARDB (Antibiotic Resistance Genes Database), has created opportunities to apply computational strategies to predict AMR[47–49].

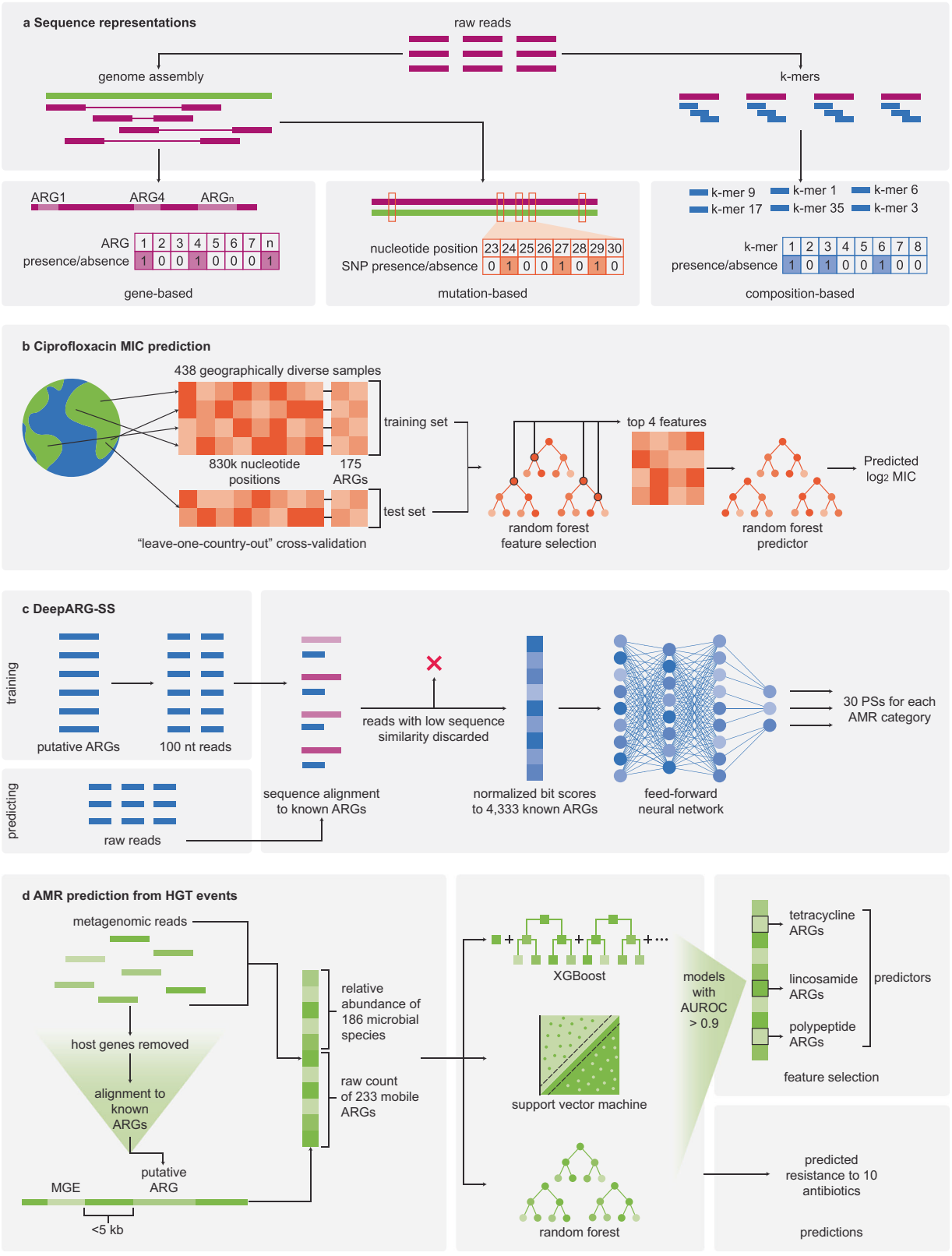
Rule-based algorithms for AMR surveillance work to identify ARGs based on sequence similarity cut-offs to known ARGs[50]. For example, ResFinder uses BLAST (Basic Local Alignment Search Tool) to align input sequences with the ResFinder database and calculate sequence identity and coverage[49]. While these methods rarely misclassify non-ARGs as ARGs, they struggle to identify novel ARGs with low sequence similarity to existing references[51].

AI methods can be more effective at generalizing to novel sequence space[51,52]. Indeed, an important benefit of leveraging AI for AMR surveillance is to identify novel resistance genes that may contribute to AMR phenotypes. For this reason, ML models such as decision-tree-based methods (e.g., RFs) and linear regression models are commonly employed for these tasks[51]. These models are considered more explainable than DL methods, since their decisions can be attributed to specific input features[51]. For reference, decision trees recursively separate data based on individual input features[51]. Linear regression models learn a coefficient for each input feature that represents the corresponding change in the output when that feature is changed. In contrast, DL models distribute information from one feature across many nodes, so learned parameters cannot easily be mapped to specific features.

Feature selection

AI applications in AMR surveillance primarily work to predict the AMR phenotype of a bacterium based on its genome. For these tasks, genomic data is commonly encoded as ARG presence/absence (based on alignment to AMR databases), mutation matrices (denoting presence/absence or type of mutations), or the presence/absence of k-mers (genomic subsequences of length k) (Fig. 2a)[51]. Considering only a small portion of the genome is linked to an observed AMR phenotype, it is crucial to select only the most relevant features to prevent inputting noise into the model. Indeed, by including a high proportion of irrelevant features during model training, particularly with high-dimensional data like sequence data, the model can overfit to the training data and generalize poorly to novel sequences[51].

Conventionally, features with the same value across all samples from the input dataset—for example, the absence of a certain ARG in all samples—are removed due to their lack of correlation to AMR phenotype[51]. Further feature selection can be performed using statistical evaluations such as pairwise association tests or by leveraging an explainable model type to determine feature importance[51]. Pairwise association tests use statistical methods (e.g., chi-squared test, ANOVA, mutual information) to calculate the correlation between each input feature and the resultant AMR phenotype. The most highly correlated features can be selected as input for model training[51]. Explainable ML models can also be used for feature selection. For



example, the magnitude of a coefficient in a LR model can represent feature importance, since it denotes the effect on the output when this feature is changed.

A 2018 study compared pairwise association and a linear SVM (LSVM) in selecting features to predict susceptible/resistant (S/R) phenotypes across *Mycobacterium tuberculosis* isolates based on genomes compiled from the Pathosystems Resource Integration Center (PATRIC) database[53,54]. Features were represented as the presence/absence of alleles present in a pan-genome (collection of all genes that exist in a species) for *M. tuberculosis* built from the compiled isolates[53]. Feature selection was performed using a grid search method wherein different cut-offs for absolute feature weights were iteratively tested to optimize model performance[53]. The most highly weighted

Fig. 2 | Application of AI to AMR surveillance. a Common methods of representing genetic information for antimicrobial resistance (AMR) prediction include gene-based, mutation-based, and composition-based encodings. Gene-based encodings denote the presence/absence of AMR genes (ARGs) identified through alignment to AMR databases. Mutation-based encodings represent presence/absence or type of mutation based on comparison to a reference genome. Composition-based methods denote the presence/absence of subsequences of a specified length (k-mers). **b** Pataki et al. trained a random forest (RF) model using geographically diverse whole-genome sequencing data to predict the minimum inhibitory concentration (MIC) of ciprofloxacin in *Escherichia coli*. Data were represented using a combination gene- and mutation-based encoding and were used to train an RF to perform feature selection based on each feature's contribution to reducing the Gini index, a measure of entropy across a decision tree. The top four features were used to train another RF to predict ciprofloxacin MIC in *E. coli* isolates of unknown country of origin. **c** DeepARG-SS takes raw sequencing data as input to predict 30 AMR categories. To simulate raw reads during model training, putative ARGs were broken into 100 nt reads. Reads are aligned to known ARGs and those with sequence homology below a user-defined cut-off are discarded. Reads are represented as their normalized bit scores to 4333 known ARGs and inputted into a multiclass FFNN which outputs a prediction score (PS) for each of the 30 AMR categories. **d** Baker et al. used metagenomic sequencing data collected from chicken farms across China to identify key predictors of AMR dissemination via HGT. Metagenomic reads were filtered for host genes and de novo assembled. Mobile-ARGs were identified based on alignment to known ARGs and distance from mobile genetic elements. For metagenomic samples that tested positive for *E. coli*, microbial species abundance and predicted-mobile ARG counts were quantified from the source chicken gut and used to train seven model types (three examples shown) to predict associated resistance against 10 antibiotics. Models with an area under the receiver-operating characteristic (AUROC) curve greater than 0.9 were used to select predictors.

features from models with an AUROC greater than 0.8 were selected[53]. This LSVM-based method was able to identify different classes of known ARGs, seven of which were not identifiable using pairwise association tests[53]. This is likely because pairwise association tests consider features individually, whereas the LSVM inherently accounts for the structure between features[53].

It is important to note that LSVMs linearly separate input data and therefore may not perform well on more complex tasks where data points cannot be optimally separated using a linear function[55]. To deal with more complex data, nonlinear SVMs apply a kernel function that transforms the dot product of each feature vector into a higher-dimensional space where it can be better separated by a linear decision boundary[55]. However, due to this kernel transformation, the "weights" in nonlinear SVMs, called Lagrange multipliers, represent the magnitude of the contribution that each transformed data point has on where the decision boundary was constructed[55]. Due to this transformation, Lagrange multipliers do not denote the importance of single features for a prediction and therefore cannot be used to represent feature importance.

Genomics-based AMR prediction

The amount of training data required to train a generalizable predictor depends on the pathogen and antibiotic. Larger and more diverse training datasets are typically required for pathogens with high genomic diversity and low penetrance ARGs to generalize effectively to novel sequences[51]. An important consideration for genomics-based AMR prediction is the imbalance of available genomic data across geographical regions[56]. Unfortunately, training data is rarely available from low- and middle-income countries (LMICs) where healthcare is less accessible, and AMR poses the greatest threat[56]. This is a major limitation of AI-based AMR surveillance since genotypes are geographically clustered. Indeed, a predictor trained on WGS data from England generalizes poorly to isolates from regions of Africa[56]. To maximize the predictive ability of AI models for AMR prediction in LMICs, a geographically diverse training set is required.

In 2020, Pataki et al. trained a RF using WGS data from five geographically diverse countries to predict antibiotic minimum inhibitory concentration (MIC; the lowest concentration of an antibiotic required to completely inhibit bacterial growth) against *E. coli* samples (Fig. 2b)[57]. While standard S/R definitions exist for common antibiotics, they vary based on context (clinical versus agricultural settings) and change over time[51]. Specifically, EUCAST (European Committee on Antimicrobial Susceptibility Testing) revised their definitions of susceptibility to include microbial strains that were initially classified as resistant[58]. MIC is therefore a more universal alternative to binary S/R classification.

Sequencing data for 704 *E. coli* samples and their ciprofloxacin MIC data were retrieved from the AMR Data Hub, 438 with known countries of origin for model training, and 266 of unknown origin held out as a test set[57]. Raw reads were mapped to a susceptible reference genome to identify mutations[57]. Across the 438 samples, 830,000 genomic positions with mutations were identified and used to create a mutation matrix, where no mutation was labeled "0", SNPs were labeled "1", and insertion/deletions were labeled "5"[57]. Additionally, ResFinder was used to identify 175 ARGs across samples[57]. A total of 175 features were concatenated to the mutation matrix denoting the presence or absence of the 175 ARGs identified by ResFinder, resulting in 830,175-dimensional input vectors[57]. These data were used to train a RF using a "leave-one-country-out" cross-validation scheme where individual models were trained on the data from four countries, and subsequently tested on data from the excluded country[57]. Feature importance was calculated for each input feature based on its contribution to reducing the Gini index, a measure of entropy across a decision tree[57]. The four most important features were selected to train another RF[57].

When employed to predict on the independent test set, the retrained RF received a mean absolute fold error (the mean error between the $\log_2$ true and predicted MIC values) of 0.883[57]. The model experienced more error predicting MICs between 8 and 64 μg/mL (resistant) than MICs below this range (susceptible) since most of the resistant samples in training had MICs of 32 μg/mL[57]. There were few training examples with MICs of 8, 16, or 64 μg/mL, leading the model to predict samples with these MICs to have an MIC of 32 μg/mL[57]. To improve model performance, more diverse examples of resistant isolates are clearly required for training.

Metagenomics-based AMR prediction

Metagenomic data encompasses all genetic information from environmental or clinical samples[59]. Metagenomics-based AMR prediction detects ARGs from phylogenetically complex samples, bypassing the need to isolate individual microbes. For example, DeepARG-SS (short sequences) (Fig. 2c) predicts 30 AMR categories (the antibiotic classes an isolate is resistant to) from short sequencing reads[60]. The authors compiled 10,602 genes from UNIPROT with high sequence similarity to 4333 ARGs from ARDB and CARD and labeled their AMR categories based on sequence-based clustering with these ARGs[60]. Labeled genes were broken into 100 nucleotide reads to simulate the short read data generated by next generation sequencing methods[60]. Each read was represented as a 4333-dimensional vector containing bit scores normalized between 0 and 1 to each known ARG from ARDB and CARD, and used to train a multiclass FFNN that outputs prediction scores for each of the 30 AMR categories[60].

DeepARG-SS was evaluated by predicting on a pseudo-metagenomic dataset constructed from 6,485,966 eukaryotic reads and 10,000 ARG reads from PATRIC[60]. Since DeepARG-SS aims to classify putative ARGs into categories, any reads with low sequence similarity to known ARGs are discarded prior to model employment[60]. Using a prediction score cut-off of 0.8, DeepARG-SS correctly predicted the AMR category of 9976 of the 10,000 ARG reads[60]. However, by using only eukaryotic sequences as negative examples for evaluation, the model could be identifying patterns unrelated to ARG classification, for example, GC content. While identifying ARGs from metagenomic samples containing eukaryotic genes is important, particularly in clinical settings, it is useful to include negative examples of prokaryotic sequences for evaluation to better simulate environmental

samples, and more accurately assess the model in the context of ARG classification.

Another benefit of using metagenomic data is the ability to study transmission factors that could contribute to the dissemination of AMR between species, such as HGT[59]. In 2023, Baker et al. trained a suite of ML models to predict S/R in *E. coli* samples resulting from HGT events in chicken farms across China (Fig. 2d)[61]. A total of 461 metagenomic samples were collected from chicken feces, feathers, carcasses, wastewater, soil, barn floors, and processing lines[61]. Raw sequencing reads were filtered for chicken DNA and de novo assembled[61]. The taxonomic classification of each sequence was determined using MetaPhlAn software[62]. ARGs were identified based on sequence homology to CARD and considered potentially mobile if they existed within 5 kb of a mobile genetic element, which were identified based on BLASTn (BLAST nucleotide) searches to ISfinder, a reference database for bacterial insertion sequences[61,63]. 170 of the metagenomic samples, which tested positive for *E. coli*, were cultured and tested against 10 antibiotics for S/R phenotype, which was used to train seven model types to predict resistance against each antibiotic[61]. Input data was represented as 419-dimensional vectors denoting the relative abundance of 186 gut microbiome species (inferred based on the output of MetaPhlAn) and raw counts of 233 mobile ARGs[61]. Models with an AUROC greater than 0.9 were used to identify the strongest predictors (mobile ARGs or gut microbial species) of resistance against each antibiotic[61].

Overall, the authors identified ten mobile ARGs as strong predictors of AMR in *E. coli* across these Chinese chicken farms[61]. Furthermore, the authors discovered that farms that use tetracyclines, lincosamides, and polypeptide antibiotics were shown to have the presence of diverse ARGs beyond those that confer resistance to these specified antibiotics, suggesting that the co-localization of ARGs may play an important role in AMR dissemination[61]. With enough validation, this method could potentially be applied to infer the basis of future AMR dissemination based on HGT events.

Antibiotic discovery

No novel class of clinical antibiotics has been discovered since 1987[64]. This can be attributed to bottlenecks in conventional antibiotic discovery pipelines. Most clinical antibiotics originate from soil-dwelling microbes that produce antibacterial compounds to compete in their native ecological niches[64]. These were discovered by iteratively fractionating and testing microbial extracts for antimicrobial activity to isolate the associated bioactive agents[64]. However, this method was limited by the dereplication problem, wherein the same molecules were repeatedly discovered due to the physiological overlap among related organisms, hindering the ability to discover fundamentally novel antibiotic classes. The field then shifted to medicinal chemistry approaches by synthesizing analogs of existing antibiotics to retain efficacy while overcoming resistance, but this method was quickly hindered by the emergence of extended-spectrum resistance mechanisms[64]. Target-based high-throughput screening (HTS) for small molecule binding to purified protein targets showed limited success due to poor Gram-negative permeability, off-target effects, and a high propensity for the evolution of resistance in the single enzyme target[64]. While whole-cell HTS has recently been more promising, it is costly and time-consuming, with few hits advancing to clinical trials[64]. Furthermore, HTS libraries generally fail to capture the physicochemical properties inherent to antibacterial compounds. Specifically, Brown et al. observed that screening libraries tend to bias towards lipophilic compounds, resulting in low hit rates for Gram-negative pathogens where the passive diffusion of hydrophobic compounds is inhibited by the outer membrane[65]. Fortunately, AI techniques can now be leveraged to overcome bottlenecks across various domains of antibiotic discovery.

Virtual screening

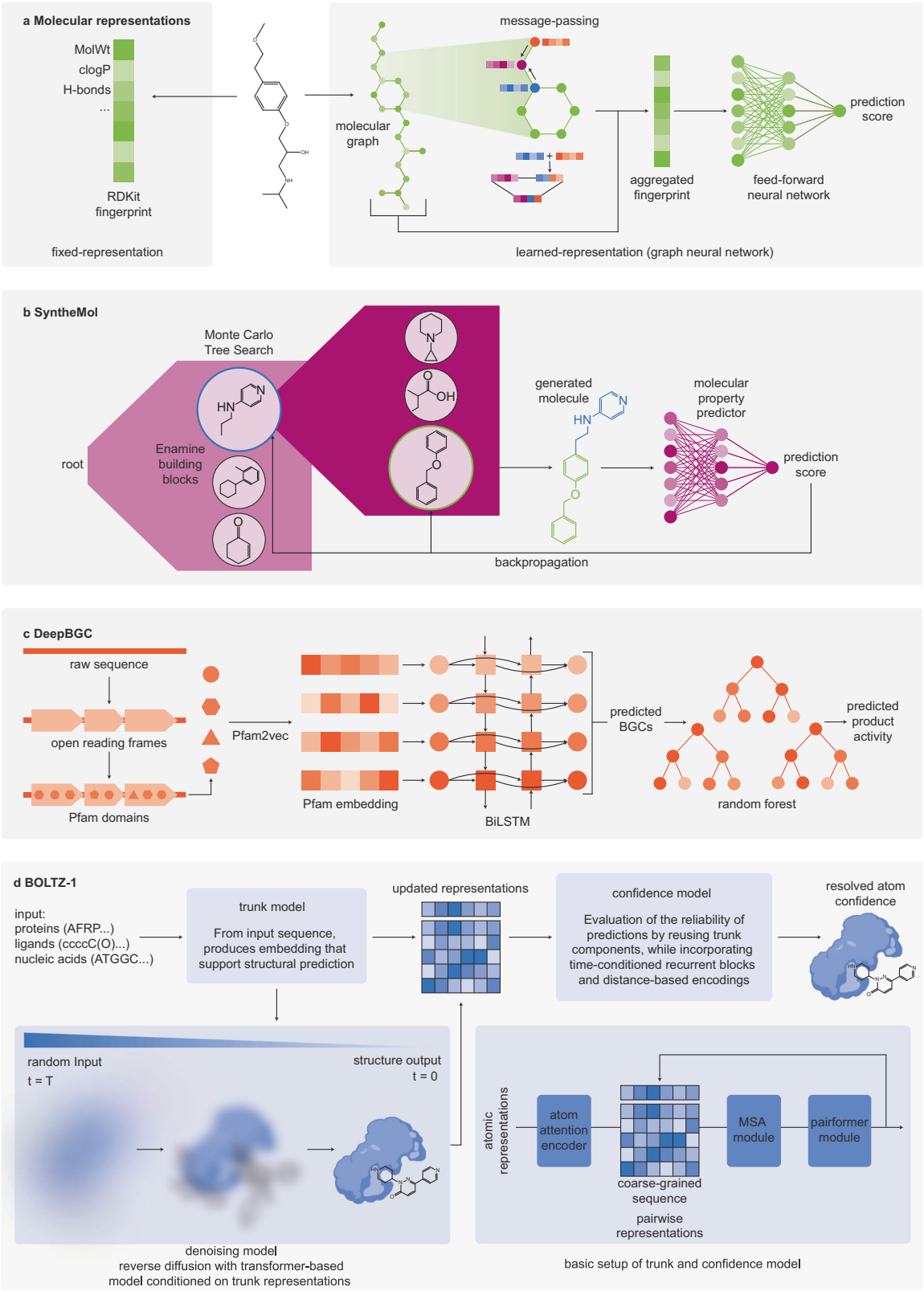
Virtual screening can be a less costly and less time-consuming alternative to conventional HTS that leverages ML techniques to predict novel molecules with a specified molecular property through searching vast in silico chemical repositories[66]. While conventional HTS campaigns are upper bound by a few million compounds, molecular property prediction models only require the screening of thousands to tens of thousands of compounds and can then be applied to virtually screen much larger chemical libraries of more than 10^8 compounds.

For property prediction tasks, molecules can be represented using fixed representations or learned representations (Fig. 3a). Fixed representations encode computable features of molecules, often chemical structure and physicochemical properties[66]. For example, Morgan fingerprints encode the presence and absence of all molecular substructures smaller or equal to a user-defined radius. RDKit fingerprints are 200-bit vectors containing calculated 2D physicochemical descriptors of a molecule (e.g., molecular weight, clogP, number of primary amines, etc.). In contrast, learned molecular representations are generated using an ML algorithm trained on task-specific data[67]. While learned representations can capture prediction-relevant information that is unable to be extracted using fixed representations, they are often less interpretable.

A leading method for learning molecular representations is through the application of graph neural networks (GNNs), which learn vector representations from the graph structures of molecules that are optimized for a specific task[67]. The nodes and edges of the graph represent the atoms and bonds of the molecule, respectively[67]. Atom features are initialized based on atomic properties (e.g., element, hybridization state), then updated via message-passing steps, where the features of neighboring atoms are aggregated and used to update the original atomic representations[67]. Over several iterations of message-passing (typically three to five), the vector representations of each atom capture more information about the interconnectivity of the molecule[67]. Once message-passing is complete, all updated atom vectors are aggregated to achieve a final representation of the entire molecule, which is used as input for an FFNN that outputs a prediction score for the inputted molecule[67].

Chemprop is a popular directed message-passing neural network (D-MPNN), a type of GNN that works on the features of directed edges (bonds) rather than atoms[68,69]. Chemprop has been applied to new antibiotic discovery by training models on bacterial growth inhibition datasets and subsequently predicting antibacterial activity of compounds from the Drug Repurposing Hub, the ZINC15 library, and Mcule purchasable database[70–72]. For these tasks, learned molecular representations from the D-MPNN were concatenated with RDKit fingerprints to include information about global molecular features before being inputted into the FFNN[70,71]. Molecules predicted to be antibacterial were filtered based on structural similarity to active compounds from the training set ("hits") to prioritize molecules that are structurally dissimilar to existing antibiotics and are more likely to evade current resistance mechanisms[70,71]. This general pipeline identified halicin, a broad-spectrum antibacterial, and abaucin, which has narrow-spectrum activity against *Acinetobacter baumannii*[70,71].

A major reason for the antibiotic resistance crisis is the lack of structurally novel antibiotics advancing through clinical trials[64]. Conventionally, antibiotic leads are prioritized primarily based on their potency against bacterial pathogens in vitro[66]. However, virtual screening methods can be applied to predict the ADMET (absorption, distribution, metabolism, excretion, toxicity) properties of compounds, thereby enabling the prioritization of compounds that are more likely to be successful in translational studies[7,66]. For example, ADMET-AI is a user-friendly online platform that contains two multi-task Chemprop-RDKit models (one classification, one regression) trained on the ADMET datasets from the Therapeutic Data Commons (TDC), an online repository to support ML model development for therapeutic tasks[7,73]. Examples of these datasets include clinical toxicity, solubility, and bioavailability. ADMET-AI takes SMILES (Simplified Molecular Input Line Entry System; string of characters denoting chemical structure) as input and outputs prediction scores or regression values for each TDC dataset[7]. To aid users in the interpretation of model outputs, ADMET-AI evaluates input molecules by comparing their ADMET predictions to those of 2579 approved drugs from DrugBank, calculating the



percentile rank of each molecule's predictions relative to the DrugBank dataset[7].

Molecular generation

Virtual screening is an efficient method to explore existing in silico chemical libraries, however, these models scale poorly to ultra-large chemical spaces[66]. Generative models can expand the existing chemical search space through the de novo design of compounds with specific properties—for example, antibacterial activity and drug-likeness[6,66]. While diverse architectures for molecular generation exist for drug discovery tasks, such as the discovery of antimalarial compounds and PI3Kγ inhibitors, the majority have yet to be applied to small-molecule antibiotic discovery[6].

Fig. 3 | AI-based methods for antibiotic discovery. a Molecules can be represented using fixed or learned representations. Fixed representations typically denote known structural features or calculated physicochemical features (e.g., RDKit features) of a compound. Learned representations encode unknown features of compounds for more complex tasks. An important learned representation method is the graph neural network, which learns encodings of molecules based on the aggregation of atom features that are iteratively updated through message-passing. **b** SyntheMol is a generative model for antibiotic discovery. SyntheMol employs a Monte-Carlo Tree Search (MCTS) algorithm to select molecular building blocks from the Enamine REAL Space based on predicted antibacterial activity and frequency of selection. Selected building blocks are pieced together to generate a molecule. The predicted antibacterial activity of the final molecule is backpropagated through the model to improve subsequent generations. **c** DeepBGC uses a bidirectional long-short-term memory (BiLSTM) model to predict novel BGCs in microbial genomes. Pfam domains are assigned to open reading frames (ORFs), then converted to Pfam word embeddings using Pfam2vec. Word embeddings are used as input for the BiLSTM which predicts whether each Pfam domain is part of a BGC. Consecutive highly predicted domains are considered BGCs. Predicted BGCs are then inputted into a random forest model to predict the bioactivity of the BGC products. **d** BOLTZ-1 is a diffusion-based model for complex biomolecular structure prediction that models the three-dimensional structures of proteins, nucleic acids, and small molecules in complex. Proteins are inputted as their amino acid sequence, nucleic acid structures as their nucleotide sequences, and small molecules as SMILES strings. BOLTZ-1 forms a multiple sequence alignment (MSA) and a pairwise residue matrix, which are updated simultaneously. These updated matrices are implemented into a diffusion model that works to predict the three-dimensional structure of the input sequences from a random configuration of the raw atom coordinates. The confidence model is fed information from the trunk model and each reverse diffusion step from the denoising model, outputting the resolved atom representation.

The synthesis of AI-designed compounds is a major limitation when deploying generative models for real-world drug discovery tasks[6]. Indeed, many generated molecules cannot be synthesized for experimental validation due to current limitations of organic synthesis. In 2024, Swanson et al. developed SyntheMol (Fig. 3b), a generative algorithm that uses a Monte-Carlo Tree Search (MCTS) to search a combinatorial space of molecular building blocks to generate predicted antibacterial compounds that are synthetically tractable[6]. For reference, MCTS is a simulation-based tree search where nodes are states and edges are actions leading to each state[74]. MCTS simulates edges based on random sampling and evaluates the outcome of each simulation[74]. The tree search is then expanded by selecting the most promising simulated action[74]. SyntheMol leverages this method to search through 132,479 molecular building blocks that can be combined into molecules containing two or three building blocks using 13 chemical reactions from the Enamine REadily AccessibLe (REAL) Space[6]. This enables the exploration of a chemical space of around 30 billion molecules[6,75].

Trained on growth inhibition data against *A. baumannii*, SyntheMol scores building blocks based on a balance of predicted activity against *A. baumannii* and how many times that building block had been previously selected[6]. After selecting an initial building block, SyntheMol then scores all synthetically compatible blocks and selects the highest scoring building block for combination[6]. The resulting molecule is then evaluated using a molecular property predictor, with the property prediction score then back propagated through the MCTS nodes to update the scores of each used building block[6]. SyntheMol was run for 20,000 iterations (rollouts), throughout which the model learned to generate molecules with predicted antibacterial activity against *A. baumannii*[6]. 58 structurally diverse generated molecules were synthesized and tested in vitro against *A. baumannii*[6]. Six compounds (>10% hit rate) had an MIC of 8 ug/mL or lower when combined with a subinhibitory concentration of the outer membrane permeabilizing agent SPR741[6].

Importantly, an array of generative architectures has been successfully employed for drug discovery tasks outside of the antibiotic space, including variational autoencoders (VAE), generative adversarial networks (GAN), and diffusion models[66]. These architectures are therefore well-suited for antibiotic discovery tasks in the near-term.

VAEs use two neural networks to encode and decode data from a low-dimensional, continuous latent space[76]. The decoder samples from specific regions of this latent space to generate new data with similar properties of the input data[76]. In the context of molecular generation, most prior work with VAEs used string-based representations of molecules, such as SMILES[77]. However, such approaches result in a high proportion of chemically invalid molecules[76]. This was superseded by the junction tree VAE (JT-VAE), which takes a graph and junction tree representation of a molecule and combines their vectors to generate molecules with desired properties[76,78]. While graph representations of molecules are constructed on an atom-by-atom basis, junction trees are constructed based on the connectivity of chemically valid subgraphs[76]. For example, a single node on a junction tree may represent an entire aromatic ring rather than an atom within that ring[76]. Combining junction trees with molecular graphs therefore enables the JT-VAE to generate chemically valid molecules[76]. Indeed, the JT-VAE reconstructed 100% valid molecules from the latent space[76]. The JT-VAE has been applied in combination with a property predictor to generate antimalarial compounds[79]. Two of these generated compounds were synthesized and shown to have nanomolar antimalarial activity in vitro[79].

GANs are another generative architecture that contain two neural networks: a generator and a discriminator. The generator works to generate molecules that are similar to a training set with a desired property, while the discriminator learns to discriminate between the generated molecules and those from the training set[80]. Once the generator is able to generate molecules that are indistinguishable from the training set by the discriminator, it can be employed to generate novel molecules with the desired property[80]. One such example, MolGAN, was trained to generate molecules with specific drug-like properties, including solubility, chemical validity, and quantitative estimate of drug likeness[81]. To optimize the generated molecules for these properties, MolGAN implements a third neural network that is architecturally identical to the discriminator, which informs the GAN of which molecules contain the desired properties during training through a reward function[81]. For example, generated molecules that are chemically invalid receive zero reward, prompting the GAN to generate molecules that are chemically valid. Indeed, 99.4% of the molecules generated by MolGAN were chemically valid[81]. However, applications of GANs for drug discovery efforts have yet to be validated in vitro.

Diffusion models transform random noise into structured molecular data through a two-phase process. Initially, a forward diffusion process systematically introduces Gaussian noise to a sample in a fixed manner. The model then learns to reverse this process, progressively denoising the sample in small steps until the original data is recovered[82]. This iterative refinement generates more valid molecular structures compared to VAEs and GANs, which decode latent representations in a single step[66]. As such, diffusion models have been employed for structure-based molecular generation, where compounds are designed to fit specific protein pockets[83]. For instance, using a one-shot generation framework (Gaussian noise undergoes reverse diffusion in one step) in which the protein pocket information is fixed, Huang et al. designed a small molecule that binds SARS-CoV-2 main protein in silico, which was informed by semantic (chemical binding) and geometric (3D spatial) information of the protein pocket[83]. It should be noted, however, that in vitro activity of the small molecule was not assessed.

Natural product discovery

Despite current challenges in natural product-based drug discovery, natural products remain a promising source of antibiotics[84,85]. They have been optimized through evolution to provide microbes with competitive advantages in complex polymicrobial environments[85,86]. The biosynthesis of microbial secondary metabolites is encoded in biosynthetic gene clusters (BGCs), co-localized genes that encode protein machinery for producing a

specific metabolite[87]. Genome mining has become a popular approach in contemporary natural product antibiotic discovery to identify novel BGCs that could produce promising molecules[88]. Computational approaches to genome mining allow for the exploration of species that cannot be cultured in the lab[86,89]. However, rule-based computational approaches to identify putative BGCs are based on sequence homology to known BGCs and are unable to generalize well to novel sequences[89]. AI-based approaches provide a promising avenue for the discovery of novel BGCs[89].

For example, DeepBGC (Fig. 3c) is a DL-based platform for predicting BGCs using a BiLSTM[89]. Assembled WGS reads are analyzed to identify open reading frames (ORFs). Within each ORF, conserved functional protein units known as Pfam (Protein families database) domains are detected[89]. Pfam domains are converted into 100-dimensional embeddings using Pfam2vec[89,90]. These embeddings are then inputted into a BiLSTM, which predicts whether the Pfam domain is part of a BGC[89]. Multiple predicted BGC domains that are adjacent are deemed putative BGCs[89].

DeepBGC was trained on 617 BGCs from the ClusterFinder training set, and 10,128 random gene clusters (non-BGCs)[89]. During training, BGC and non-BGC sequences are randomly combined into a continuous sequence to simulate assembled genomic data[89]. DeepBGC was able to outperform ClusterFinder, a state-of-the-art rule-based BGC predictor, in identifying BGC positions from whole (artificial) genomes (AUROC = 0.923 versus 0.847)[89,91]. Furthermore, when applied to identify novel BGCs using a leave-one-class-out validation scheme, DeepBGC achieved an AUROC of 0.946, surpassing the AUROC of 0.865 from ClusterFinder[89,91]. Predicted BGCs were further classified based on product activity using an RF[89]. This RF was trained on biosynthetic product class and their product activities compiled from the MIBiG database, including antibacterial activity, which was represented as a multi-class vector[89,92]. This modest training set containing 180 positive examples yielded an AUROC of only 0.61, indicating the need for a larger training set to improve performance; this is currently insufficient to provide accurate predictions of antibacterial activity[89].

Biomolecular structure prediction

Beyond hit compound identification, ML-based tools have been developed to aid in elucidating the mechanism of action (MOA) of potential lead molecules. For example, numerous ML-based tools have emerged to improve three-dimensional protein modeling, which can help streamline target elucidation[93–97]. Perhaps the most influential of these tools, AlphaFold2, predicts the three-dimensional structure of proteins based on amino acid sequence[93]. AlphaFold2 consists of two neural network-based modules: Evoformer and the Structure module[93]. Trained on structures from the PDB database, AlphaFold2 takes as input a primary amino acid sequence and generates a multiple-sequence alignment (MSA) with homologous protein sequences, and a pairwise residue matrix that encodes information between residues (e.g., distance, residue orientation)[93]. The MSA matrix reveals the co-evolution of certain residues, suggesting that they exist closely in three-dimensional space[93]. These matrices are concurrently updated through the Evoformer module to yield an embedding that combines structural and evolutionary information[93]. This joint embedding is then inputted into the Structure module, which predicts the position and rotation of each amino acid in the input sequence and outputs a three-dimensional structure[93]. AlphaFold2 significantly outperformed competing computational methods in the CASP14 (Critical Assessment of Structure Prediction 14) assessment, achieving near experimental accuracy in predicting protein structures that had not yet been deposited in the PDB[93,98]. Two AlphaFold2 developers, Demis Hassabis and John Jumper, were awarded the Nobel Prize in Chemistry in 2024 for their contributions to structural elucidation of proteins.

In 2024, Google DeepMind released AlphaFold3, a modified version of AlphaFold2 that models the structure of complexes containing combinations of biomolecules including proteins, nucleic acids, small molecules, and ions[96]. AlphaFold3 can therefore be leveraged as a docking algorithm for target-based drug discovery. The AlphaFold3 architecture was substantially modified from AlphaFold2[96]. The Evoformer module was replaced with a Pairformer module, which updates the pairwise residue matrix with less reliance on the MSA[96]. The Structure module was replaced with a diffusion model, a generative architecture that is trained to iteratively refine noisy sample data to generate clean data in a process called denoising[96,99]. In the context of AlphaFold3, the diffusion model works to denoise a random configuration of the atoms involved in a given biomolecular complex to generate a plausible three-dimensional structure for this complex[96]. The diffusion architecture allows AlphaFold3 to model diverse molecule types and capture important local stereochemistry, which is crucial for predicting intermolecular interactions[96]. Indeed, when evaluated on 428 protein-ligand structures omitted from the training dataset, AlphaFold3 outperformed traditional docking methods and RoseTTAFold, another neural network-based framework for biomolecular structure elucidation[96,100,101].

DeepMind initially did not release the code or model weights and restricted access to AlphaFold3, prompting the development of open-source alternatives[95,96]. One of these alternatives, BOLTZ-1 (Fig. 3d), leverages a similar architecture to achieve AlphaFold3-level accuracy[95]. The BOLTZ-1 architecture was based on AlphaFold3, with some advancements. Similar to AlphaFold3, BOLTZ-1 consists of a trunk model that updates a pairwise representation of the input data, a diffusion model that predicts three-dimensional biomolecular structures, and a confidence model that outputs confidence metrics for the structure prediction[95,96]. The most important distinction between AlphaFold3 and BOLTZ-1 is an improved confidence model which is compositionally identical to the trunk model, compared to the AlphaFold3 confidence model which contains only a Pairformer module[95,96]. The BOLTZ-1 confidence model is fed information from the trunk model, as well as from each reverse diffusion step from the denoising model[95]. These architectural advancements allow the BOLTZ-1 confidence predictions to be more robust for complex biomolecular structures[95]. BOLTZ-1 was benchmarked against CHAI-1, another AlphaFold3 replication of equal performance[95,97,102]. BOLTZ-1 outperformed CHAI-1 on the CASP15 dataset based on median all-atom local distance difference test, a measure of accuracy across all biomolecule types.

While the diffusion architecture allows for the modeling of mixed biomolecular complexes, generative models are prone to hallucinations[95,96]. Both AlphaFold3 and BOLTZ-1 suffer from instances of overlapping structures that disobey actual physical constraints, and the generation of plausible structures for unstructured regions of a biomolecule[95,96]. To address the latter, AlphaFold3 was further trained on AlphaFold-Multimer predicted protein structures, throughout which unstructured regions appear as long extended loops rather than incorrectly predicted structures[96].

Discussion

AI is not a panacea, only a set of (powerful) tools to support humans with domain expertise. Regardless of the sophistication of the model architecture, predictive ability is reliant on the data used to train the model. To adequately model complex chemical-biological systems, training data must be high-quality, diverse, and biochemically relevant[68]. While a lack of training data remains a limiting factor for many AI applications in biology, there has been a recent increase in public initiatives to support the development of AI tools for biological tasks. These resources aim to collect and organize robust, high-quality data for model training, and provide standardized benchmarks for model evaluation. Specifically for AMR tasks, TDC contains numerous datasets for infectious diseases and standardized benchmarks[73]. CARD 2023 provides annotated ARGs with standardized antibiotic resistance ontology terms that were condensed to 15 characters to facilitate use for AI efforts[48]. Moreover, there exists a reasonable amount of chemical data to support the implementation of AI for antibiotic discovery efforts.

The increasing availability of chemical screening data offers valuable resources for AI applications; however, it is imperative to critically assess the methodologies employed in collecting or generating this data to uncover potential limitations that may impede antibiotic development at later stages. For example, reliance solely on bacterial growth inhibition data to train models for predicting a compound's "antibiotic-likeness" incorporates

factors such as bacterial growth inhibition, permeability, and efflux capacity but fails to address other critical aspects such as solubility and human toxicity. This oversight can obstruct the antibiotic discovery process by omitting properties that are integral to the clinical development of novel antibacterials. Consequently, experimental in vivo toxicity testing is essential to identify chemical leads that retain antibacterial efficacy while avoiding mammalian cell toxicity.

Further, we see a lack of diverse data to train generalizable models for the widespread application of AI in clinical diagnostic and AMR surveillance settings, stemming from inconsistent phenotypic testing protocols and guidelines. To support bacterial infection diagnosis and global surveillance efforts, the field requires increased collaborative efforts that share WGS of AMR bacterial clinical isolates, standardized AST measurements for such isolates, as well as a comprehensive panel of clinical antibiotics to be tested to determine S/R that span an array of antibiotic classes, thereby compiling the diverse datasets required to train these more "patient-facing" models[51].

For AI to be deployed and trusted on the front lines of the AMR crisis, models must be accurate, interpretable, and cost reducing[59]. Diagnostic tools require rigorous experimental and clinical evaluation to be implemented as clinical decision support systems that aid healthcare providers but still require human oversight[59]. AI can only be ethically applied for real-time diagnosis in cases where it has been proven to perform better than current standards, a key example of which is sepsis prediction. AI models in AMR surveillance and antibiotic discovery have lower accuracy requirements since they don't directly impact patient treatment but rather guide broader efforts. However, this indirect application has contributed to a relative lack of experimental validation for these models, as their outputs are not immediately actionable in clinical settings. As such, there is a lack of experimental validation for AI-guided surveillance methods, and although we highlight antibiotic discovery methods that were supported by wet lab validation, many studies exist that solely use model test metrics to evaluate performance. Indeed, AI predictions must be experimentally validated going forward. The proper implementation of AI for AMR efforts therefore requires increased collaboration between computer scientists, microbiologists, clinicians, and policymakers.

Data availability

No datasets were generated or analysed during the current study.

Received: 26 September 2024; Accepted: 6 February 2025;
Published online: 13 March 2025

References

1. Antimicrobial Resistance Collaborators. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *Lancet* **399**, 629–655 (2022).
2. Walsh, T. R., Gales, A. C., Laxminarayan, R. & Dodd, P. C. Antimicrobial resistance: addressing a global threat to humanity. *PLoS Med.* **20**, e1004264 (2023).
3. Singh, B., Bhat, A. & Ravi, K. Antibiotics misuse and antimicrobial resistance development in agriculture: a global challenge. *Environ. Health* https://doi.org/10.1021/envhealth.4c00094 (2024).
4. Vandenberg, O. et al. Consolidation of clinical microbiology laboratories and introduction of transformative technologies. *Clin. Microbiol. Rev.* **33**, e00057 (2020).
5. Gupta, Y. D. & Bhandary, S. Artificial intelligence for understanding mechanisms of antimicrobial resistance and antimicrobial discovery (eds. Khanna, A. et al.) in *Artificial Intelligence and Machine Learning in Drug Design and Development* 117–156 (John Wiley & Sons, 2024).
6. Swanson, K. et al. Generative AI for designing and validating easily synthesizable and structurally novel antibiotics. *Nat. Mach. Intell.* **6**, 338–353 (2024).
7. Swanson, K. et al. ADMET-AI: a machine learning ADMET platform for evaluation of large-scale chemical libraries. *Bioinformatics* **40**, btae416 (2024).
8. Quagliarello, V. J. & Scheld, W. M. Treatment of bacterial meningitis. *N. Engl. J. Med.* **336**, 708–716 (1997).
9. Lappin, E. & Ferguson, A. J. Gram-positive toxic shock syndromes. *Lancet Infect. Dis.* **9**, 281–290 (2009).
10. Hotchkiss, R. S. & Karl, I. E. The pathophysiology and treatment of sepsis. *N. Engl. J. Med.* **348**, 138–150 (2003).
11. Lagier, J.-C. et al. Current and past strategies for bacterial culture in clinical microbiology. *Clin. Microbiol. Rev.* **28**, 208–236 (2015).
12. Aggarwal, D. et al. Clinical utility and cost-effectiveness of bacterial 16S rRNA and targeted PCR based diagnostic testing in a UK microbiology laboratory network. *Sci. Rep.* **10**, 7965 (2020).
13. Liu, V. X. et al. The timing of early antibiotics and hospital mortality in sepsis. *Am. J. Respir. Crit. Care Med.* **196**, 856–863 (2017).
14. Evans, L. et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Crit. Care Med.* **49**, e1063–e1143 (2021).
15. Rudd, K. E. et al. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *Lancet* **395**, 200–211 (2020).
16. Schinkel, M., Paranjape, K., Nannan Panday, R. S., Skyttberg, N. & Nanayakkara, P. W. B. Clinical applications of artificial intelligence in sepsis: a narrative review. *Comput. Biol. Med.* **115**, 103488 (2019).
17. Goh, K. H. et al. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nat. Commun.* **12**, 711 (2021).
18. Xiao, C., Choi, E. & Sun, J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J. Am. Med. Inform. Assoc.* **25**, 1419–1428 (2018).
19. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270 (2019).
20. Zhang, D. et al. An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns* **2**, 100196 (2021).
21. Texas Tech University Health Sciences Center. Cerner Health Facts Database. Clinical Research Data Warehouse https://www.ttuhsc.edu/biomedical-sciences/clinical-research-data-warehouse/default.aspx (2018).
22. Tayefi, M. et al. Challenges and opportunities beyond structured data in analysis of electronic health records. *Wiley Interdiscip. Rev. Comput. Stat.* **13**, e1549 (2021).
23. Shashikumar, S. P., Wardi, G., Malhotra, A. & Nemati, S. Artificial intelligence sepsis prediction algorithm learns to say "I don't know". *npj Digital Med.* **4**, 1–9 (2021).
24. Boussina, A. et al. Impact of a deep learning sepsis prediction model on quality of care and survival. *NPJ Digital Med.* **7**, 14 (2024).
25. Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G. & King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **17**, 195 (2019).
26. Morais, C. L. M., Lima, K. M. G., Singh, M. & Martin, F. L. Tutorial: multivariate classification for vibrational spectroscopy in biological samples. *Nat. Protoc.* **15**, 2143–2162 (2020).
27. Rebrosova, K. et al. Raman spectroscopy—a novel method for identification and characterization of microbes on a single-cell level in clinical settings. *Front. Cell. Infect. Microbiol.* **12**, 866463 (2022).
28. Lu, W., Chen, X., Wang, L., Li, H. & Fu, Y. V. Combination of an artificial intelligence approach and laser tweezers Raman spectroscopy for microbial identification. *Anal. Chem.* **92**, 6288–6296 (2020).
29. Kim, G. et al. Rapid species identification of pathogenic bacteria from a minute quantity exploiting three-dimensional quantitative phase imaging and artificial neural network. *Light Sci. Appl.* **11**, 190 (2022).

30. Wolf, E. Three-dimensional structure determination of semi-transparent objects from holographic data. *Opt. Commun.* **1**, 153–156 (1969).
31. Drancourt, M. Detection of microorganisms in blood specimens using matrix-assisted laser desorption ionization time-of-flight mass spectrometry: a review. *Clin. Microbiol. Infect.* **16**, 1620–1625 (2010).
32. Banerjee, R. et al. Randomized trial of rapid multiplex polymerase chain reaction-based blood culture identification and susceptibility testing. *Clin. Infect. Dis.* **61**, 1071–1080 (2015).
33. Kommedal, Ø., Aasen, J. L. & Lindemann, P. C. Genetic antimicrobial susceptibility testing in Gram-negative sepsis—impact on time to results in a routine laboratory. *APMIS* **124**, 603–610 (2016).
34. Majumder, M. A. A. et al. Antimicrobial stewardship: fighting antimicrobial resistance and protecting global public health. *Infect. Drug Resist.* **13**, 4713–4738 (2020).
35. Ellington, M. J. et al. The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee. *Clin. Microbiol. Infect.* **23**, 2–22 (2017).
36. Ren, Y. et al. Prediction of antimicrobial resistance based on whole-genome sequencing and machine learning. *Bioinformatics* **38**, 325–334 (2022).
37. Shi, J. et al. Antimicrobial resistance genetic factor identification from whole-genome sequence data using deep feature selection. *BMC Bioinforma.* **20**, 535 (2019).
38. Noman, S. M. et al. Machine learning techniques for antimicrobial resistance prediction of Pseudomonas aeruginosa from whole genome sequence data. *Comput. Intell. Neurosci.* **2023**, 5236168 (2023).
39. Zagajewski, A. et al. Deep learning and single-cell phenotyping for rapid antimicrobial susceptibility detection in Escherichia coli. *Commun. Biol.* **6**, 1164 (2023).
40. Osthoff, M. et al. Impact of MALDI-TOF-MS-based identification directly from positive blood cultures on patient management: a controlled clinical trial. *Clin. Microbiol. Infect.* **23**, 78–85 (2017).
41. Weis, C., Cuénod, A., Rieck, B., Borgwardt, K. & Egli, A. DRIAMS: Database of Resistance Information on Antimicrobials and MALDI-TOF Mass Spectra. https://doi.org/10.5061/dryad.bzkh1899q (2022).
42. Alegría Guajardo, C. E., López-Cortés, X. A. & Álvarez, S. H. Deep learning algorithm applied to bacteria recognition. in *2022 IEEE International Conference on Automation/XXV Congress of the Chilean Association of Automatic Control (ICA-ACCA)* 1–6 (IEEE, 2022).
43. Weis, C. et al. Direct antimicrobial resistance prediction from clinical MALDI-TOF mass spectra using machine learning. *Nat. Med.* **28**, 164–174 (2022).
44. De Waele, G., Menschaert, G., Vandamme, P. & Waegeman, W. Pre-trained Maldi Transformers improve MALDI-TOF MS-based prediction. *bioRxiv* https://doi.org/10.1101/2024.01.18.576189 (2024).
45. López-Cortés, X. A., Manríquez-Troncoso, J. M., Hernández-García, R. & Peralta, D. MSDeepAMR: antimicrobial resistance prediction based on deep neural networks and transfer learning. *Front. Microbiol.* **15**, 1361795 (2024).
46. McArthur, A. G. & Tsang, K. K. Antimicrobial resistance surveillance in the genomic age. *Ann. N. Y. Acad. Sci.* **1388**, 78–91 (2017).
47. Ajulo, S. & Awosile, B. Global antimicrobial resistance and use surveillance system (GLASS 2022): investigating the relationship between antimicrobial resistance and antimicrobial consumption data across the participating countries. *PLoS ONE* **19**, e0297921 (2024).
48. Alcock, B. P. et al. CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
49. Karp, B. E. et al. National antimicrobial resistance monitoring system: two decades of advancing public health through integrated surveillance of antimicrobial resistance. *Foodborne Pathog. Dis.* **14**, 545–557 (2017).
50. Florensa, A. F., Kaas, R. S., Clausen, P. T. L. C., Aytan-Aktug, D. & Aarestrup, F. M. ResFinder—an open online resource for identification of antimicrobial resistance genes in next-generation sequencing data and prediction of phenotypes from genotypes. *Microb. Genom.* **8**, 000748 (2022).
51. Kim, J. I. et al. Machine learning for antimicrobial resistance prediction: current practice, limitations, and clinical perspective. *Clin. Microbiol. Rev.* **35**, e0017921 (2022).
52. Anahtar, M. N., Yang, J. H. & Kanjilal, S. Applications of machine learning to the problem of antimicrobial resistance: an emerging model for translational research. *J. Clin. Microbiol.* **59**, e0126020 (2021).
53. Kavvas, E. S. et al. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nat. Commun.* **9**, 4306 (2018).
54. Davis, J. J. et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. *Nucleic Acids Res.* **48**, D606–D612 (2020).
55. Lin, S.-W., Lee, Z.-J., Chen, S.-C. & Tseng, T.-Y. Parameter determination of support vector machine and feature selection using simulated annealing approach. *Appl. Soft Comput.* **8**, 1505–1512 (2008).
56. Nsubuga, M., Galiwango, R., Jjingo, D. & Mboowa, G. Generalizability of machine learning in predicting antimicrobial resistance in E. coli: a multi-country case study in Africa. *BMC Genomics* **25**, 287 (2024).
57. Pataki, B. Á. et al. Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. *Sci. Rep.* **10**, 15026 (2020).
58. Nabal Díaz, S. G., Algara Robles, O. & García-Lechuz Moya, J. M. New definitions of susceptibility categories EUCAST 2019: clinic application. *Rev. Esp. Quimioter.* **35**, 84–88 (2022).
59. Wheeler, N. E. et al. Innovations in genomic antimicrobial resistance surveillance. *Lancet Microbe* **4**, e1063–e1070 (2023).
60. Arango-Argoty, G. et al. DeepARG: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome* **6**, 23 (2018).
61. Baker, M. et al. Machine learning and metagenomics reveal shared antimicrobial resistance profiles across multiple chicken farms and abattoirs in China. *Nat. Food* **4**, 707–720 (2023).
62. Blanco-Míguez, A. et al. Extending and improving metagenomic taxonomic profiling with uncharacterized species using MetaPhlAn 4. *Nat. Biotechnol.* **41**, 1633–1644 (2023).
63. Siguier, P., Perochon, J., Lestrade, L., Mahillon, J. & Chandler, M. ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **34**, D32–D36 (2006).
64. Brown, E. D. & Wright, G. D. Antibacterial drug discovery in the resistance era. *Nature* **529**, 336–343 (2016).
65. Brown, D. G., May-Dracka, T. L., Gagnon, M. M. & Tommasi, R. Trends and exceptions of physical properties on antibacterial activity for gram-positive and gram-negative pathogens. *J. Med. Chem.* **57**, 10144–10161 (2014).
66. Catacutan, D. B., Alexander, J., Arnold, A. & Stokes, J. M. Machine learning in preclinical drug discovery. *Nat. Chem. Biol.* **20**, 960–973 (2024).
67. Corso, G., Stark, H., Jegelka, S., Jaakkola, T. & Barzilay, R. Graph neural networks. *Nat. Rev. Methods Prim.* **4**, 1–13 (2024).

68. Yang, K. et al. Analyzing learned molecular representations for property prediction. *J. Chem. Inf. Model.* **59**, 3370–3388 (2019).
69. Heid, E. et al. Chemprop: a machine learning package for chemical property prediction. *J. Chem. Inf. Model.* **64**, 9–17 (2024).
70. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702.e13 (2020).
71. Liu, G. et al. Deep learning-guided discovery of an antibiotic targeting Acinetobacter baumannii. *Nat. Chem. Biol.* **19**, 1342–1350 (2023).
72. Corsello, S. M. et al. The Drug Repurposing Hub: a next-generation drug library and information resource. *Nat. Med.* **23**, 405–408 (2017).
73. Huang, K. et al. Artificial intelligence foundation for therapeutic science. *Nat. Chem. Biol.* **18**, 1033–1036 (2022).
74. Świechowski, M., Godlewski, K., Sawicki, B. & Mańdziuk, J. Monte Carlo Tree Search: a review of recent modifications and applications. *Artif. Intell. Rev.* **56**, 2497–2562 (2023).
75. Grygorenko, O. O. et al. Generating multibillion chemical space of readily accessible screening compounds. *iScience* **23**, 101681 (2020).
76. Jin, W., Barzilay, R. & Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *arXiv [cs.LG]* https://doi.org/10.48550/arXiv.1802.04364 (2018).
77. Gómez-Bombarelli, R. et al. Automatic chemical design using a data-driven continuous representation of molecules. *arXiv [cs.LG]* https://doi.org/10.48550/arXiv.1610.02415 (2016).
78. Kusner, M. J., Paige, B. & Hernández-Lobato, J. M. Grammar variational autoencoder. *arXiv [stat.ML]* https://doi.org/10.48550/arXiv.1703.01925 (2017).
79. Godinez, W. J. et al. Design of potent antimalarials with generative chemistry. *Nat. Mach. Intell.* **4**, 180–186 (2022).
80. Goodfellow, I. J. et al. Generative adversarial networks. *arXiv [stat.ML]* https://doi.org/10.48550/arXiv.1406.2661 (2014).
81. De Cao, N. & Kipf, T. MolGAN: an implicit generative model for small molecular graphs. *arXiv [stat.ML]* https://doi.org/10.48550/arXiv.1805.11973 (2018).
82. Cesaro, A., Bagheri, M., Torres, M., Wan, F. & de la Fuente-Nunez, C. Deep learning tools to accelerate antibiotic discovery. *Expert Opin. Drug Discov.* **18**, 1245–1257 (2023).
83. Huang, L. et al. A dual diffusion model enables 3D molecule generation and lead optimization based on target pockets. *Nat. Commun.* **15**, 2657 (2024).
84. Lewis, K. Recover the lost art of drug discovery. *Nature* **485**, 439–440 (2012).
85. Mullowney, M. W. et al. Artificial intelligence for natural product drug discovery. *Nat. Rev. Drug Discov.* **22**, 895–916 (2023).
86. Arnold, A., Alexander, J., Liu, G. & Stokes, J. M. Applications of machine learning in microbial natural product drug discovery. *Expert Opin. Drug Discov.* **18**, 1259–1272 (2023).
87. Medema, M. H. et al. Minimum information about a biosynthetic gene cluster. *Nat. Chem. Biol.* **11**, 625–631 (2015).
88. Yee, D. A. et al. Genome mining for unknown-unknown natural products. *Nat. Chem. Biol.* **19**, 633–640 (2023).
89. Hannigan, G. D. et al. A deep learning genome-mining strategy for biosynthetic gene cluster prediction. *Nucleic Acids Res.* **47**, e110 (2019).
90. Church, K. W. Word2Vec. *Nat. Lang. Eng.* **23**, 155–162 (2017).
91. Cimermancic, P. et al. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. *Cell* **158**, 412–421 (2014).
92. Terlouw, B. R. et al. MIBiG 3.0: a community-driven effort to annotate experimentally validated biosynthetic gene clusters. *Nucleic Acids Res.* **51**, D603–D610 (2023).
93. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
94. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
95. Wohlwend, J. et al. Boltz-1 democratizing biomolecular interaction modeling. *bioRxiv* https://doi.org/10.1101/2024.11.19.624167 (2024).
96. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
97. Discovery, C. et al. Chai-1: decoding the molecular interactions of life. *bioRxiv* https://doi.org/10.1101/2024.10.10.615955 (2024).
98. Pereira, J. et al. High-accuracy protein structure prediction in CASP14. *Proteins* **89**, 1687–1699 (2021).
99. Yim, J. et al. Diffusion models in protein structure and docking. *Wiley Interdiscip Rev. Comput. Mol. Sci.* **14**, e1711 (2024).
100. Lisanza, S. L. et al. Multistate and functional protein design using RoseTTAFold sequence space diffusion. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-024-02395-w (2024).
101. Eberhardt, J., Santos-Martins, D., Tillack, A. F. & Forli, S. AutoDock Vina 1.2.0: new docking methods, expanded force field, and Python bindings. *J. Chem. Inf. Model.* **61**, 3891–3898 (2021).
102. Kryshtafovych, A. et al. New prediction categories in CASP15. *Proteins* **91**, 1550–1557 (2023).
103. Bzdok, D., Krzywinski, M. & Altman, N. Points of significance: machine learning: a primer. *Nat. Methods* **14**, 1119–1120 (2017).
104. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
105. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process* **45**, 2673–2681 (1997).
106. Cox, D. R. The regression analysis of binary sequences. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **21**, 238–238 (1959).
107. Breiman, L. Random forests | machine learning. *Mach. Learn.* **45**, 5–32 (2001).
108. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.* **82**, 3713–3744 (2023).
109. Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychol. Rev.* **65**, 386–408 (1958).
110. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
111. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE Inst. Electr. Electron. Eng.* **86**, 2278–2324 (1998).
112. Cortes, C. & Vapnik, V. Support-vector networks. *Mach. Learn.* **20**, 273–297 (1995).
113. Ke, G. et al. LightGBM: a highly efficient gradient boosting decision tree. in *NeurIPS* (eds. Guyon, I. et al.) 3149–3157 (Curran Associates Inc., 2017).
114. Maulud, D. & Abdulazeez, A. M. A review on linear regression comprehensive in machine learning. *J. Appl. Sci. Technol. Trends* **1**, 140–147 (2020).
115. Gilmer, J., Schoenholz, S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. *ICML* **70**, 1263–1272 (2017).
116. Coulom, R. Efficient selectivity and backup operators in Monte-Carlo tree search. in *Computers and Games* (eds. van den Herik, H. J., Ciancarini, P. & Donkers, H. H. L. M.) 72–83 (Springer, 2007).
117. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. *arXiv [stat.ML]* https://doi.org/10.48550/arXiv.1312.6114 (2013).
118. Yang, L. et al. Diffusion models: a comprehensive survey of methods and applications. *ACM Comput. Surv.* **56**, 1–39 (2024).

Acknowledgements
This work was generously supported by the Canadian Institutes of Health Research, the Natural Sciences and Engineering Research Council of Canada, the Weston Family Foundation, and the David Braley Centre for Antibiotic Discovery. A.A. and S.M. are supported by Canadian Institutes of Health Research scholarships.

Author contributions
A.A., S.M., and J.M.S. wrote and edited the manuscript. A.A. and S.M. designed the figures.

Competing interests
J.M.S. is co-founder and CSO of Stoked Bio. All other authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Jonathan M. Stokes.